S

3-Selanyl-2-aminopropanoic Acid

▶ Selenocysteine

10Sa RNA

▶ TmRNA

Saccharomyces cerevisiae

▶ Cell Cycle, Budding Yeast

Saddle-Node Bifurcation

Tianshou Zhou
School of Mathematics and Computational Sciences,
Sun Yet-Sen University, Guangzhou, Guangdong,
China

Definition

In the mathematical area of bifurcation theory, a ▶ saddle-node bifurcation, tangential bifurcation, or fold bifurcation is a local bifurcation in which two fixed points (or equilibria) of a dynamical system collide and annihilate each other. The term “saddle-node bifurcation” is most often used in reference to continuous dynamical systems. In discrete dynamical systems, the same bifurcation is often instead called a fold bifurcation. Another name is blue skies bifurcation in reference to the sudden creation of two fixed points.

If the phase space is one-dimensional, one of the equilibrium points is unstable (the saddle), while the other is stable (the node).

The normal form of a saddle-node bifurcation is:

$$\frac{dx}{dt} = r + x^2$$

Here x is the state variable and r is the bifurcation parameter.

If $r < 0$ there are two equilibrium points, a stable equilibrium point at $-\sqrt{-r}$ and an unstable one at $+\sqrt{-r}$. At $r = 0$ (the bifurcation point) there is exactly one equilibrium point. At this point the fixed point is no longer hyperbolic. In this case the fixed point is called a saddle-node fixed point. If $r > 0$, then there are no equilibrium points.

A saddle-node bifurcation occurs in the consumer equation if the consumption term is changed from px to p, that is the consumption rate is constant and not in proportion to resource x.

Saddle-node bifurcations may be associated with hysteresis loops and catastrophes.

W. Dubitzky et al. (eds.), *Encyclopedia of Systems Biology*, DOI 10.1007/978-1-4419-9863-7,

S-Adenosylmethionine

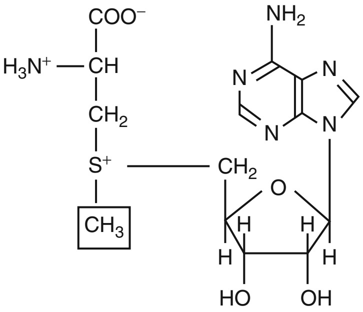
Gota Kawai
Department of Life and Environmental Sciences,
Chiba Institute of Technology, Narashino, Chiba,
Japan

Synonyms

SAM, adoMet

Definition

S-adenosylmethionine (Fig. 1) is synthesized from ATP and L-methionine and acts as a donor of methyl groups in prokaryote as well as eukaryote. By the transfer of a methyl group, *S*-adenosylmethionine is converted into *S*-adenosylhomocysteine, which is broken down to homocysteine and adenosine. Homocysteine is then converted to methionine. *S*-adenosylmethionine is also used for synthesis of cysteine as well as polyamines.



S-Adenosylmethionine, Fig. 1 Chemical structure of *S*-adenosylmethionine

Safety Assessment

▶ Pharmaceutical Toxicology, Application of Biosimulation

Safety Testing

▶ Pharmaceutical Toxicology, Application of Biosimulation

SAGA

Tetsuro Kokubo
Department of Supramolecular Biology, Graduate
School of Nanobioscience, Yokohama City
University, Yokohama, Kanagawa, Japan

Synonyms

PCAF; STAGA; TFTC (human)

Definition

SAGA (*S*pt-*A*da-*G*cn5-*a*cetyltransferase), which is orthologous to the mammalian TFTC, PCAF, and STAGA complexes, was identified originally as a HAT (*h*istone *a*cetyl*t*ransferase) for histone H3 (Baker and Grant 2007; Bhaumik 2011; Timmers and Tora 2005). This factor is a large protein complex (~1.8 MDa) containing 5 Tafs (Taf5, 6, 9, 10, and 12) shared by TFIID, as well as 16 other subunits (Ada1, Ada2, Ada3, Ada5, Gcn5, Spt3, Spt7, Spt8, Spt20, Sgf11, Sgf29, Sgf73, Ubp8, Sus1, Chd1, and Tra1). In TFIID, there are five Taf heterodimer pairs containing a HFD (*h*istone *f*old *d*omain): Taf4-Taf12, Taf6-Taf9, Taf3-Taf10, Taf8-Taf10, and Taf11-Taf13. Similarly, in SAGA, there are three HFD heterodimers, Ada1-Taf12, Taf6-Taf9, and Spt7-Taf10, and one intramolecular HFD dimer in Spt3 that resembles Taf11-Taf13 in TFIID.

SAGA functions as a co-activator by modifying histones, such as acetylation (H3) and/or deubiquitylation (H2B), as well as by delivering TBP to the TATA-containing core promoter. The HAT and DUB (*deubi*quitylase) enzyme activities are carried on the Gcn5-Ada2-Ada3 and Ubp8-Sgf11-Sus1 submodules, respectively (each underlined protein is a catalytic subunit of HAT and DUB). The TBP delivery function is carried out by Spt3 and Spt8, which show genetic and/or physical interaction with TBP.

Transcriptional activators recruit SAGA to promoters by binding to Tra1, the largest SAGA subunit (~400 KDa), which contains a catalytically inactive phosphatidylinositol 3-kinase domain. Once recruited by activators, SAGA is further stabilized on the promoter by recognizing chemical modifications of histones. For instance, acetylated and methylated histone H3 are recognized by the bromodomain of Gcn5 and chromodomain of Chd1, respectively.

A SAGA-related complex, such as SALSA (*S*AGA *al*tered, *S*pt8 *a*bsent) or SLIK (*S*AGA-*lik*e), contains Rtg2 instead of Spt8, and a processed form of Spt7 that lacks a carboxy-terminal region required for interaction with Spt8. Although SALSA/SLIK is suggested to be involved in a response to nitrogen starvation, it is still unclear whether these two similar complexes regulate the expression of different sets of genes in vivo.

Cross-References

▶ Transcription in Eukaryote

References

Baker SP, Grant PA (2007) The SAGA continues: expanding the cellular role of a transcriptional co-activator complex. Oncogene 26(37):5329–5340

Bhaumik SR (2011) Distinct regulatory mechanisms of eukaryotic transcriptional activation by SAGA and TFIID. Biochim Biophys Acta 1809(2):97–108

Timmers HT, Tora L (2005) SAGA unveiled. Trends Biochem Sci 30(1):7–10

SAM, AdoMet

▶ S-Adenosylmethionine

Sample Variability, Inter-Groups

Anyela Camargo[1] and Jan T. Kim[2]

[1]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, Ceredigion, UK

[2]School of Computing Sciences, University of East Anglia, Norwich, Norfolk, UK

Synonyms

Inter-class variability; Inter-sample variability

Definition

A central question in the Bioscience area – and related sciences – is the degree of variability between samples obtained from different populations. Inter-sample variability aims to determine how much variability should be ascribed to differences in experimental conditions (groups) (Jerrold 1974). As a simple but common example, an experiment may comprise two populations of samples, an untreated reference population, and a population perturbed by some treatment. A numeric quantity y is measured; y_r denotes measurements of sample from the reference population and y_p, measurements of a sample from the perturbed population. The problem could be represented in a linear equation for a general linear model:

$$\begin{aligned} y_r &= r + e \\ y_p &= r + x_p + e \end{aligned} \tag{1}$$

where r is the intercept, that is, the value of y when the system is unperturbed, x_p represents the effect of the experimental perturbation on y, and e is the error, comprising both biological variation and measurement error.

From Eq. 1, the inter-sample variability caused by the perturbation can be estimated as

$$s = |y_p - y_r| \tag{2}$$

Therefore, if $|y_p - y_r| \gg \sigma$ (where σ is a given threshold), there is a substantial difference between the reference yr and the perturbed class y_p, suggesting that x_p has an effect in the outcome. This is a typical problem in the drug discovery area when the effects of a drug over a target population are assessed. In such experiments, samples from subjects taking a placebo yr and samples from subjects taking the drug y_p are obtained and s is estimated. Note however, that the significance of s is proportional to the size of population sampled, the number or replicates from each experiment and the number of response variables being assessed. For large experiments, that is, high-throughput analysis, s should be put in the context of a multiple testing analysis to correct for the potential high number of false of positives that is the result of assessing hundreds of response variables.

However importantly, the degree of which there is an assumption of potential differentiation might and should vary according to the experimental settings used in the analysis and how complex the experiment is. For simple experiments, the assumption of non-equality might be computed straightforwardly according to Eq. 2. However, for more complex experiments such as microarray-based analyses, statistics estimates such as the one-way and two-way ANOVA tests need to be appropriately normalized and corrected for multiple testing.

References

Jerrold ZH (ed) (1974) Biostatistical analysis. Prentice Hall, Upeer Saddle River

Sample Variability, Intra-Groups

Anyela Camargo[1] and Jan T. Kim[2]
[1]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, Ceredigion, UK
[2]School of Computing Sciences, University of East Anglia, Norwich, Norfolk, UK

Synonyms

Intraclass variability; Intra-sample variability

Definition

A central question in the bioscience area – and related sciences – is the degree of variability in the samples obtained from similar experimental settings. This estimate can be used to identify outliers and problems in the handling of the experiment, to assess the suitability of the samples to represent a class and to check the quality of the data. The assessment of sample variability can be done by estimating various statistics that give an indication of the spread of the measurements around the center of the distribution (Jerrold 1974). For example, the coefficient of variation (CV) (1), the ratio of the standard deviation to the mean, is a simple but very informative metric that indicates the spread of a dataset as a proportion of its mean.

$$\mathrm{Cv} = \frac{\sigma}{|\mu|} \tag{1}$$

where σ is the population's standard deviation and μ is the population's mean.

Although the estimation of variability obeys the same principles regardless of the data, the selection of the estimate to measure variability and its interpretation should be given by the characteristics of the experiment. For experiments where no information on the distribution underlying intragroup variability is available, a Gaussian distribution may be assumed as a first approximation, and consequently the standard deviation may be used as a measure of variability.

However, it is important to notice that intragroup variability may be heterogeneous, i.e., spread or other characteristics may be different for individual groups.

As an elementary example, consider a scenario with four groups of patients, where the first three groups are comprised of patients with A, B, and C health conditions, respectively, and the fourth group contains patients who are unwell but no clear diagnosis is available. The fourth group very likely includes patients with several different diseases, and, therefore, intragroup variability can be expected to be elevated for that group.

The method that has been used to designate group membership to the samples also has various implications for intragroup variability. For example, multiple alternatives of group designation were considered and an optimal designation was chosen from among these.

As illustrated by the examples above, interpreting intragroup variability strongly depends on the group designation. Therefore, intragroup variation usually has to be interpreted relative to total variation and to intergroup variation.

References

Jerrold ZH (ed) (1974) Biostatistical analysis. Prentice Hall International, Upper Saddle River

Sampling

▶ Data Sampling

SBGN

Falk Schreiber[1,2] and Nicolas Le Novère[3]
[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), OT Gatersleben, Stadt Seeland, Germany
[2]Martin Luther University Halle–Wittenberg, Halle, Germany
[3]EMBL European Bioinformatics Institute and Babraham Institute, Cambridge, UK

Synonyms

Graphical representation of biological processes; Network visualization and exchange; Systems biology graphical notation

Definition

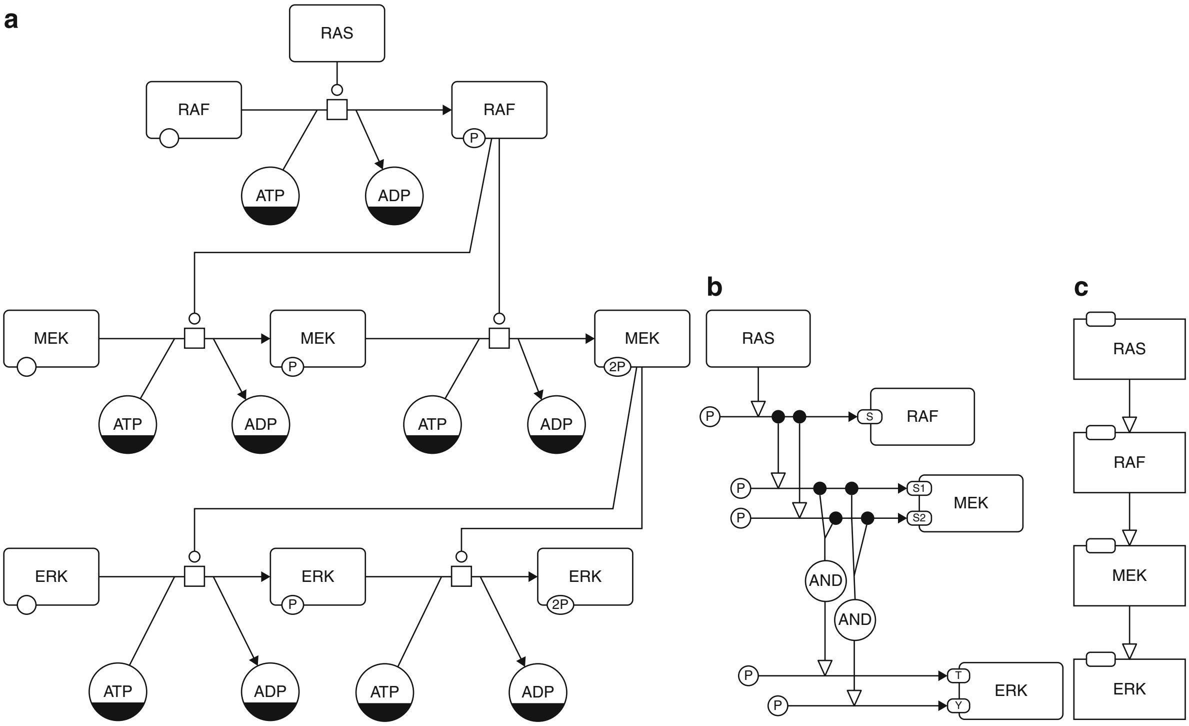
The Systems Biology Graphical Notation (SBGN) is a standard graphical format to represent unambiguously biological networks and cellular processes. SBGN allows to visualize complex biological knowledge, including gene regulation, protein interaction, signaling pathways, and metabolic networks. Information is represented using graphical objects (glyphs) organized in a (▶ Graph) structure and employs controlled vocabulary of the (▶ Systems Biology Ontology) to indicate certain types of information. To cover different levels of detail and to deal with alternative dimensions of biological knowledge, SBGN consists of three visual languages: process descriptions, entity relationships, and activity flows. SBGN focuses on the graphical and visualization facets of systems biology information and is complementary to exchange formats such as ▶ Systems Biology Markup Language (SBML) or BioPAX.

Characteristics

SBGN Languages

SBGN (Le Novère et al. 2009) encodes biological knowledge in a graphical form. Biological processes and networks can be represented and explored at different levels of detail and biological entities may be involved in a large number and different types of interactions. Representing all possible interactions and reactions is often not desired. To deal with this complexity and support different "views" of a biological system, SBGN provides three orthogonal and complementary languages. Each language conveys a certain level of detail and a specific part of the semantics of the underlying biological system. Each language comes with its advantages and weaknesses.

1. *Process descriptions* (PD). The focus of PD is the temporal dependencies of biological interactions and transformations in a network. PD represents networks of events which convert biological entities into other entities, change their states, or transport them to another location. Entities can be pools of simple chemicals, macromolecules, nucleic acid features (such as genes or promoters), and so on, represented by *entity pool nodes* (EPN). The transformation or transport of an entity is represented by



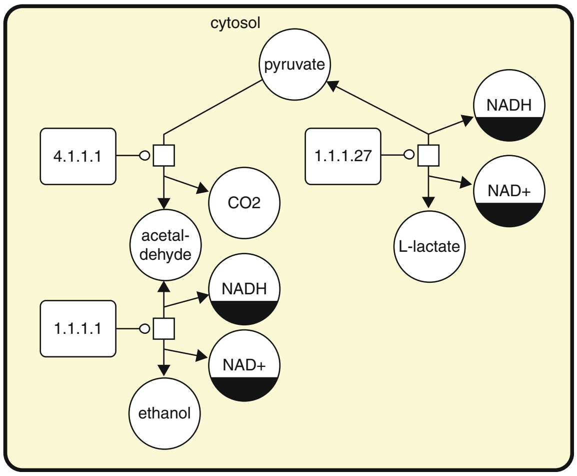
SBGN, Fig. 1 The same biological process represented with the SBGN languages PD, ER, and AF (from *left* to *right*)

a *process node* (PN). PD describes processes in a mechanistic manner, shows different states of an entity as different glyphs, and is best used to represent mechanistic and temporal aspects of a biological system. Figure 1a shows a biological system in PD. For more details about the PD language, one can read its technical specification (Moodie et al. 2011).

2. *Entity relationships* (ER). The focus of ER is the relationship in which entities are involved in the network, and their influences onto other entities. ER does not explicitly consider temporal aspects but shows for the biological system all possible relationships at once. *Entity nodes* (EN) represent entities that exist, each entity being represented only once in the map. *Relationships* are rules that decide whether an EN exists. ER describes relationships in a mechanistic manner and is best used to represent protein interactions and pathways which involve multistate or multicomponent entities. Figure 1b shows a biological system in ER. For more details about the ER language, one can read its technical specification (Le Novère et al. 2011).
3. *Activity flows* (AF). The focus of AF is the biological activity. In contrast to PD and ER, the representation can be ambiguous when it comes to the underlying mechanism. The biological activities are represented by *activity nodes*. *Modulation arcs* show the influence of activities onto other activities. Different activities of a biological entity may be represented separately. AF shows the sequential influence of activities and is best used to represent functional genomics and signaling pathways. Figure 1c shows a biological system in AF. For more details about the AF language, one can read its technical specification (Mi et al. 2009).

Structure of an SBGN Map

An SBGN map is composed of graphical objects (*glyphs*) which are connected following syntax, semantics, and layout *rules* defined in the SBGN specifications (Moodie et al. 2011; Le Novère et al. 2011; Mi et al. 2009). From a more technical point of view, an SBGN map is a (▶ Graph) consisting of *nodes* and *edges*.



SBGN, Fig. 2 A process description map showing the fermentation pathway

For a typical SBGN map, we might consider the SBGN process description (PD) language. Figure 2 displays a PD map of a biochemical pathway and presents typical elements of such a collection of biochemical processes: substrates and products, reactions, enzymes catalyzing reactions, and information concerning the location (the compartment where the pathway occurs). Relevant elements of PD to encode this information are *entity pool nodes* (EPNs) representing homogeneous pools of substrates, products, or effectors of processes; *process nodes* (PNs) representing the transition between EPNs; *connecting arcs* displaying the relationships between EPNs and PNs (such as "production" or "catalysis"); and *container nodes* combining sets of EPNs (such as "compartment"). Certain rules described in detail in the PD specification (Moodie et al. 2011) define possible connections between these elements. As shown in Fig. 2, simple chemicals (a type of EPN) can be connected to processes (a type of PN) using consumption and production (types of connecting arcs), thereby representing the biochemical reaction. Macromolecules (also EPNs) catalyze processes; this is shown by connecting these elements with connecting arcs of the type catalysis. The pathway occurs in the compartment cytosol. Finally, clone markers (black fillings of the lower part of an EPN) are used to deal with an entity which should be presented in the map more than once (for better graphical representation).

SBGN Supporting Tools and Libraries

Tool support is essential for the efficient use of SBGN. There are several tools which partly (only some languages) or completely (all languages) support SBGN. Some of them also provide additional functionality such as validation of SBGN maps, translation of maps from external sources, or layout of maps.

SBGN-ML is a computer-readable format of SBGN maps to support their exchange. Software tools can make use of SBGN-ML by employing LibSBGN (van Iersel et al. 2012), which is a software library for reading, writing, editing, and validating SBGN maps.

More information about the SBGN specifications, SBGN tools, and examples can be found under http://sbgn.org.

Cross-References

▶ Systems Biology Markup Language (SBML)

References

Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem M, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Hideya K, Li L, Matsuoka Y, Villèger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman T, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Schmidt E, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H (2009) The systems biology graphical notation. Nature Biotechnol 27:735–741

Le Novère N, Demir E, Mi H, Moodie S, Villeger A (2011) Systems biology graphical notation: entity relationship language level 1 (Version 1.2). Nature Preced. doi:10.1038/npre.2011.5902.1

Mi H, Schreiber F, Le Novère N, Moodie S, Sorokin A (2009) Systems biology graphical notation: activity flow language level 1. Nature Preced. doi:10.1038/npre.2009.3724.1

Moodie S, Le Novère N, Demir E, Mi H, Villeger A (2011) Systems biology graphical notation: process description language level 1. Nature Preced. doi:10.1038/npre.2011.3721.4

van Iersel MP, Villéger AC, Czauderna T, Boyd SE, Bergmann FT, Luna A, Demir E, Sorokin A, Dogrusoz U, Matsuoka Y, Funahashi A, Aladjem MI, Mi H, Moodie SL, Kitano H, Le Novère N, Schreiber F (2012) Software support for SBGN maps: SBGN-ML and LibSBGN. Bioinformatics 28(15): 2016–2021

SBML

▶ Systems Biology Markup Language (SBML)

SBO

▶ Systems Biology Ontology

SBPAX

▶ Systems Biology Pathway Exchange (SBPAX)

Scale-Free Network

▶ Biological Disease Mechanism Networks

Scale Integration

José Román Bilbao Castro
Supercomputing-Algorithms Group, University of
Almería, Almería, Spain

Definition

Processors are mainly designed by two components: micro-architecture and scale integration. Micro-architecture refers to a set of instructions that a processor are capable of processing (addition, subtraction, multiplication, etc.). The scale integration refers to the size of the elements composing the processor (the number of transistors). The larger the scale integration, the more transistors can be integrated in a single processor and the more complexity can be built on it. Currently, integration technologies on commercial processors allow nanometer-scale devices.

Cross-References

▶ Multicore Computing

Scale of Investigation

▶ Organism State, Lymphocyte

ScaRNA Databases

▶ Non-coding RNA Databases

Schemaless Databases

Steve R. Pettifer and Teresa K. Attwood
Faculty of Life Sciences and School of Computer
Science, The University of Manchester,
Manchester, UK

Synonyms

NoSQL databases

Definition

NoSQL databases are a class of structured datastore that do not use the traditional table structure of relational systems. They are particularly useful where "join" operators are not needed, and scale well for certain modern applications, such as serving web pages and streaming media.

Cross-References

▶ Data Integration and Visualization

Schizosaccharomyces pombe

▶ Cell Cycle, Fission Yeast

Schwarz Criterion

▶ Bayesian Information Criterion (BIC)

Schwarz Information Criterion (SIC)

▶ Bayesian Information Criterion (BIC)

Scientific Instrument

Jutta Schickore
Department of History and Philosophy of Science,
Indiana University, Bloomington, IN, USA

Definition

A (typically specifically designed) device to aid the investigation of nature, usually as part of a scientific experiment.

Characteristics

Scientific instruments have been an integral part of the scientific enterprise. Without them, we would know nothing about DNA and genes, atoms, electrons, and quasars. But despite their crucial role in science, philosophers of science have rarely discussed scientific instruments. For the most part of the twentieth century, philosophy of science focused on scientific theories and conceptual foundations of science, with a special emphasis on physics. Only in the late twentieth century, scientific instruments have become a theme for philosophy of science.

Philosophy of technology also deals with instruments, but this field has developed largely independently from philosophy of science. Reflections on instruments in philosophy of technology are usually much wider in scope and cover all kinds of tools, devices, and engineering feats from hammers to cable-stayed bridges. Philosophers of technology are concerned with the analysis of design processes, ethics of technology, engineering ethics, and the metaphysics of artifacts. By contrast, those few philosophers of science who have considered scientific instruments have focused more narrowly on the question of how knowledge is obtained through instruments, and how we may characterize and interpret the knowledge thus obtained.

In the following, I present some key general issues and questions related to scientific instruments that philosophers of science have raised, questions concerning the classification of instruments and their roles in science, the distinction between observables and unobservables, the relation between instrument and theory, and instrument-generated images.

Kinds of Instruments

Philosophers and historians of science have identified several functions scientific instruments may have in science, including the amplification of phenomena and events that are below the threshold of the human senses (e.g., microscopes and telescopes), the creation or construction of phenomena that do not occur in nature in the absence of the instrument (e.g., the air pump producing a vacuum), the (simplified, downsized) imitation of effects that occur in nature without human intervention (e.g., Leyden jar, Atwood's machine); registration (Geiger counters); and measuring or quantification of phenomena (e.g., astrolabes). Notably, scientific instruments may also be used for purposes other than the creation, study, and representation of phenomena, for instance, as demonstration tools, as "philosophical toys" for amusement and instruction (e.g., kaleidoscopes), and as routine measuring devices, e.g., in surveying. It is still an open question whether these classifications are exhaustive, whether the boundaries between the categories are clear-cut, and whether the epistemic functions of instruments differ for different categories.

Particularly with respect to recent science, analysts of science have acknowledged the difficulty, if not impossibility, of drawing the line between "science" and "technology." Large-scale projects in experimental science, such as genomics, are dependent on complex technologies such as automated high-throughput technologies for the rapid identification and analysis of a great number of samples of DNA, RNA, and other molecules. In these contexts, "scientific research" and "engineering" are largely indistinguishable. In science studies, the apparatus and devices that were conceived and developed by a community connected to both science and industry are called "research technologies." Research technologies such as radioactivity, liquid scintillation counters, mouse mutants, and plant genetics are flexible multipurpose instrument systems binding together universities, industries, instrument-making firms, the military, public and private research facilities, and metrological agencies (Joerges and Shinn 2001).

To identify the epistemic functions of scientific instruments, it is crucial not to consider scientific instruments in isolation but as part of and in the context of investigative settings. Such settings usually consist of a number of tools and devices that together form the material context for an investigation. To study mitosis in living cells, for instance, investigators need a high-power microscope as well as tools to prepare the samples, a device to keep the temperature at the appropriate level, a computer to process the data obtained by the microscope and to translate them into images, and so on. Together, these tools and devices provide the material conditions for the investigation to be carried out, but their epistemic functions will be different depending on the research situation. Microtomes and growth media, for instance, are usually considered unproblematic technical tools. But they may turn into obstacles for investigations in particular circumstances, and in these cases, they may even become driving forces for further research.

The Distinction Between Observable and Unobservable Phenomena

Recent biology and recent science more generally deal with phenomena and effects that are inaccessible to the unaided senses. Biologists, like scientists in other areas, draw on data and images produced by instruments in often highly complex experimental settings to tell us about such things as DNA, neurons, and mitosis. But what reasons do we have to assume that their theories are true and these processes and objects really exist? This is the topic of the philosophical debate about "realism" and "antirealism."

The portion of this debate that is related to instruments has centered on the question of the distinction between observables and unobservables. Intuitively, it seems obvious that there are things and processes that we can observe – tables, chairs, lightning – and other things and processes that are too small, to faint, or too distant for us to observe – mitosis, quasars, electrons. But, philosophers have asked, is it really possible to draw a clear-cut distinction between what is observable and what is unobservable? If so, is this distinction really philosophically significant? What is the status of those entities that can be detected with the aid of instruments but are inaccessible to the unaided senses?

The historical context for this debate is the view, advocated by some logical empiricists early in the twentieth century, that the things that our theories postulate are not physical things. Physical things are those that are accessible to our unaided senses, such as tables and trees, and we have good reason to believe that these things exist. However, since the only access to things and phenomena such as electrons and mitosis is provided through instruments, we do not have any reason to believe in their existence. Conceptions of electrons and mitosis are merely convenient and useful constructs that explain our data (say, the images produced by a camera or computer attached to a microscope).

During the second half of the twentieth century, philosophers of science have commented on and criticized this distinction in a variety of ways. In a classic article, Grover Maxwell argued that there is a continuum of instruments and things made accessible by these instruments: things we can see directly, things we can see by looking through a looking glass, by using a light microscope, by using an electron microscope, and so on. Any line we draw between "still observable" and "no longer observable" is arbitrary, and therefore, we cannot attach any ontological significance to it. It would be absurd to say that things and phenomena existed less and less (Maxwell 1962).

Bas van Fraassen, however, insists that it does make a difference *epistemologically* whether or not we can see things directly (van Fraassen 1980). He concedes that the line is, to an extent, arbitrary. But there are clear enough cases of "observable (in principle)" and "unobservable (in principle)" for the distinction to make sense: Observable are those things that we can see with the naked eye (including those things that we could see with the naked eye if we were close enough); unobservable are those things that we cannot see and will never be able to see with the naked eye, given our bodily makeup. For van Fraassen, it makes a difference in the epistemic status of our knowledge about things whether or not we can observe them directly. While we are perfectly justified to say that the things that we can *see directly* exist, we are in a more precarious, more difficult, less secure situation when we deal with entities of which our theories tell us that they exist, but which we can only detect (such as electrons). In this case, we need to be agnostic: They might exist, they might not, but *we will never know for sure*. The part of science that deals with unobservables cannot be regarded as securely established.

In his influential book *Representing and Intervening* of 1983, Ian Hacking famously refers to a biological instrument – the microscope – to advocate a radically different solution to the problem of the distinction between observables and unobservables, thereby giving the debate between realists and antirealists new impulses. Hacking argues that the distinction is not philosophically significant, and that the concept of "observation with instruments" is in fact misleading. A microscope is not a tool for observation. To use it, the investigators actively engage with the object under study – say, with ribosomes. Precisely because investigators are able to produce consistent results while manipulating and interfering with microscopic objects, they have good reason to assume that the subvisible things and phenomena they study really exist. Moreover, it is often possible to generate similar outcomes with different instruments. Hacking argues that if two physically different instruments produce similar results, we are justified in assuming that the phenomena and processes thus detected are real because it would be highly unlikely that those instruments coincidentally produced the same kind of artifact.

The "argument from coincidence" has been widely accepted. But philosophical discussions about instruments have been moving away from debates about realism and antirealism and toward discussions of criteria for ascertaining the reliability of empirical evidence. Several philosophers now interpret this argument as an argument for the validity of instrumentally generated results. Empirical information can be considered valid, the argument goes, if it can be reproduced through different instruments and independent confirmation can thus be obtained. Philosophers have probed the epistemic strength of this argument by reconstructing the logic underlying it and by analyzing the concept of "independent" confirmation.

Instruments, Theories, and Knowledge

One of the classic problems in philosophy of science more generally is the problem of "theory-ladenness" of observation (see "▶ Theory-Ladenness"). Empirical evidence is not neutral but imbued by the investigators' theoretical commitments. Theory tells us what observations to make, what observations are salient, and how to interpret the observations. The epistemological problem is that if empirical information is contaminated with theory, we do not have a solid empirical basis against which to test our theories. The entire project of "empirical" science is at stake. Scientific instruments pose a similar challenge because in most cases, they are used to investigate phenomena that are only accessible through the data produced by instruments. It thus seems that to obtain knowledge about the phenomena, one would need to know the theory of the instrument. The data produced by the instrument would therefore be contaminated with theory and could not serve as a solid foundation for theory appraisal.

In recent years, however, philosophers have developed a number of responses against this argument from theory-ladenness. Several philosophers have suggested that if the theory of an instrument is independent of the theory that is being investigated (which is often, but not always the case), then the theory-ladenness of the empirical investigation does not pose a problem. Empirical information, although in a sense theory-laden, does provide a solid enough empirical basis for the appraisal of the theory under consideration.

In his essay on microscopes, Hacking makes a number of additional points that are relevant for this discussion. For instance, he insists that it is not necessary to know the theory of the instrument to be able to use it. Most biologists are unable to explain the physical principles according to which a high-power microscope functions. They know how to use it reliably, but they do not know how it works. Moreover, there are examples from the history of science which illustrate that certain instruments have been successfully applied before any satisfying theory of their function became available.

Hacking also points out that "knowing how to use" a microscope is a skill. Such a skill has to be acquired through training, so the experienced scientist and the untrained person – whether a nonscientist or an expert from a different field – will respond differently if confronted with the same data. This is a common situation not only in microscopy but more generally, especially with those instruments that produce pictorial outcome. For philosophers, this raises the question of whether investigators with different levels of skill are really "seeing the same thing." Moreover they need to find ways to characterize the visual and tactile competence that are integral parts of the development and use of scientific instruments. However they respond to these challenges, it is clear that the ability to explain the physical principles of an instrument is not sufficient – and perhaps not even necessary – for the ability to use the instrument successfully. In this sense, the concept of theory-ladenness appears to be unhelpful for the philosophy of scientific instruments.

The most sustained recent revision of the traditional notion of theory-ladenness of scientific instruments is the conception of "thing knowledge" (Baird 2004). In Davis Baird's "materialist epistemology," the conception of theory-ladenness is completely reworked. The basic idea is that like propositions, instruments bear or encapsulate knowledge, indeed they *are* knowledge, but this knowledge cannot usefully be described as "theory." Instruments and theories are on a par. To put it differently: Epistemology, the philosophical conceptualization of knowledge, should comprise theoretical knowledge *and* scientific instruments, "thing knowledge." The roles that instruments have played in science have been analogous to theories in that instruments too are mediums in which to develop our understanding of nature. Moreover, while instruments and theories often develop together, instruments sometimes develop independent of theory development.

The conception of "thing knowledge" challenges the traditional explication of knowledge as "justified true belief" and as something that requires a knower. But precisely for this reason, it may help to tackle another epistemological challenge that philosophers have been slow to take up, namely, the epistemological problem arising from large-scale research technologies. In present science, we are confronted with big projects that involve many individuals, each of whom carries out a distinct task. None of these individuals will have enough knowledge and expertise to complete single-handedly the research and understand, justify, and defend its outcome. The epistemological problem is this: If we want to say that those individuals working with large-scale research technologies have knowledge or an understanding about what they are doing, we need to accept that justification sufficient for knowledge is something less than full epistemic control over the entire spectrum of justificatory arguments. Alternatively, we may want to accept that groups have knowledge. We could then say that the knower (the group) has the relevant knowledge to understand and justify what the group is doing (Hardwig 1985). But we would be hard-pressed to identify the group that comprises "the knower," particularly with respect to research technologies. The notion of "thing knowledge" may help to reconceptualize this complex of problems.

Visualization and Scientific Images

Many contemporary scientific instruments produce pictorial output. Instrument-generated images such as electron micrographs, x-ray images, and fMRIs have become ubiquitous and an integral part of biomedical research. Philosophers have been rather slow to appreciate the roles of these pictures because for the most part, traditional philosophy has privileged propositional knowledge. But there is now a growing body of work on visualization, visual thinking, and images in science, and the received view that pictorial information can simply be reduced to propositional knowledge has become questionable.

Visualization in biology poses special challenges to philosophy because current instrument-aided image-making procedures are highly indirect. Pictures produced by technologies such as fMRI often require computerization of the signals obtained from an interaction of an instrument with a biological specimen. Images that show macromolecules or neurons may be products of complex mathematical operations that transform data into pictures. There is now a growing body of philosophical work on visualization in biology, addressing questions concerning the reliability of instrument-generated images, their representational content, and their function as evidence in scientific arguments (e.g., Pauwels 2006).

Cross-References

▶ Theory-Ladenness

References

Baird D (2004) Thing knowledge: a philosophy of scientific instruments. University of California Press, Berkeley

Bogen J (2002) Epistemological custard pies from functional brain imaging. Philos Sci 69:S59–S71

Chalmers A (2003) The theory-dependence of the use of instruments in science. Philos Sci 70:493–509

Hacking I (1983) Representing and intervening. Cambridge University Press, Cambridge

Hardwig J (1985) Epistemic dependence. J Philos 82:335–349

Joerges B, Shinn T (eds) (2001) Instrumentation. Between science, state and industry. Kluwer, Dordrecht

Maxwell G (1962) The ontological status of theoretical entities. In: Feigl H, Maxwell G (eds) Scientific explanation, space, and time, vol III, Minnesota studies in the philosophy of science. University of Minnesota Press, Minneapolis

Pauwels L (ed) (2006) Visual cultures of science: rethinking representational practices in knowledge building and science communication. Dartmouth College Press, Hanover

van Fraassen BC (1980) The scientific image. Oxford University Press, Oxford

Score

▶ Scoring Function, Graph Alignment

Score Network Alignment

▶ Scoring Function, Graph Alignment

Scoring Function, Graph Alignment

Michal Kolář
Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic

Synonyms

Graph alignment; Score; Score network alignment

Definition

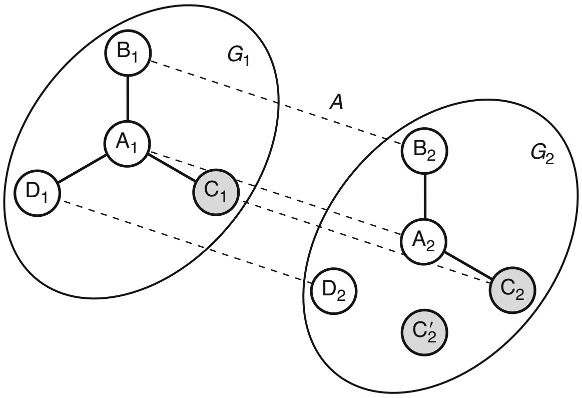
A scoring function of a graph alignment (▶ Graph Alignment, ▶ Protein Interaction Network) A evaluates the alignment quality by considering both topological (interaction) similarity of the aligned networks (▶ Protein-Protein Interaction Networks) G_1 and G_2 and similarity of their nodes. The two contributions of the score measure independently the agreement of the alignment A with the topology of the aligned networks and with the similarity of their nodes. The scoring function is maximized for the optimal graph alignment (▶ Graph Alignment, Protein Interaction Networks), which reproduces correctly the relationship between the nodes (and hence the links) of the networks.

The two contributions of the score represent independent pieces of biological information (▶ Information, Biological), thus the score decomposes into two parts, the node score (▶ Node Score, Graph Alignment) S_n and the link score (▶ Link Score, Graph Alignment) S_l (Kelley et al. 2003),

$$S = S_n + S_l. \qquad (1)$$

Consider Fig. 1. The alignment A of the graphs G_1 and G_2 pairs together proteins A_1A_2, B_1B_2, C_1C_2, and D_1D_2. The aligned pairs of proteins add to the node score (▶ Node Score, Graph Alignment) by a contribution, which depends on a pair-wise similarity of the proteins and rewards alignment of homologous proteins. Aligned links (A_1, B_1) and (A_2, B_2) and (A_1, C_1) and (A_2, C_2) add a positive contribution to the link score (▶ Link Score, Graph Alignment) as they are present in both networks. A mismatch between existence of the link (A_1, D_1) and absence of the link (A_2, D_2) contributes negatively to the link score (▶ Link Score, Graph Alignment). In a multiple graph alignment (▶ Graph Alignment, Protein Interaction Networks), the node score (▶ Node Score, Graph Alignment) consists of contributions from all equivalence classes of the alignment and the link score (▶ Link Score, Graph Alignment) assesses the total topological similarity of the aligned networks (e.g., the total number of matching and mismatching links).

Assume that a protein C_1 in Fig. 1 is homologous to both C_2 and C'_2. The existence of an interaction between (A_2, C_2) and absence of the interaction between (A_2, C_2) leads to alignment of C_1 and C_2. The topological part of the score, the link score



Scoring Function, Graph Alignment, Fig. 1 An illustration of a graph alignment (▶ Graph Alignment, Protein Interaction Networks) of two graphs G_1 and G_2. Each node represents a protein; a link stands for a protein–protein interaction. A graph alignment (▶ Graph Alignment, Protein Interaction Networks) *A* maps the proteins connected by dashed lines

(▶ Link Score, Graph Alignment), decides which of the two possible alignment partners is a better match to C_1. In this way, the graph alignment (▶ Graph Alignment, Protein Interaction Networks) may correctly identify an ortholog (▶ Orthologs) of C_1 between the ▶ Paralogs C_2 and C'_2. In the extreme case, when C_2 is not homologous to C_1 but its interaction pattern is highly similar to the pattern of C_1, the graph alignment (▶ Graph Alignment, Protein Interaction Networks) may prefer to align C_1 and C_2 rather than C_1 and its homolog C'_2. Then, the relative weight of the node score (▶ Node Score, Graph Alignment) (protein similarity) and the link score (▶ Link Score, Graph Alignment) (topological similarity) determines the partner of C_1 in the alignment *A*. Correct parameterization of the scoring function is essential (▶ Parameter Estimation, Graph Alignment) (Berg and Lässig 2006).

Cross-References

- ▶ Graph Alignment, Protein Interaction Networks
- ▶ Information, Biological
- ▶ Link Score, Graph Alignment
- ▶ Node Score, Graph Alignment
- ▶ Orthologs
- ▶ Paralogs
- ▶ Protein-Protein Interaction Networks

References

Berg J, Lässig M (2006) Cross-species analysis of biological networks by Bayesian alignment. Proc Natl Acad Sci 103(29):10967–10972

Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci 100(20):11394–11399

Screening Factor

Max Kistler
IHPST, Université Paris 1 Panthéon-Sorbonne,
Paris, France

Definition

A variable (or factor) C is called a screening factor with respect to variables A and B if and only if (1) there is a statistical correlation between A and B, so that, for example, $P(B|A) > P(B|\neg A)$, but (2) there is no such correlation if the probabilities are taken as conditional on factor C. In other words, if and only if C is a screening factor for the correlation between A and B, then conditionalizing with respect to C makes B probabilistically independent of A. In formulas, $P(B|A) > P(B|\neg A)$, but $P(B|A \,\&\, C) = P(B|\neg A \,\&\, C)$ and $P(B|A \,\&\, \neg C) = P(B|\neg A \,\&\, \neg C)$.

Search Engines with Faceted Search

Syed Toufeeq Ali-Ahmed
Department of BioMedical Informatics, Vanderbilt
University, Nashville, TN, USA

Synonyms

Faceted browsing; Faceted navigation; Parametric search

Definition

A faceted search system (or parametric search system) presents users with key value meta-data that is used for query refinement (Koren et al. 2008). By using facets (which are meta-data or class labels for entities such as genes or diseases), users can easily combine the hierarchies in various ways to refine and drill down the results for a given query; they do not have to learn custom query syntax or to restart their search from scratch after each refinement.

Characteristics

Faceted search combines faceted navigation with text search, allowing users to access (semi) structured content collections, thereby providing support for discovery and exploratory search, areas where conventional search falls short (Tunkelang 2009). Important design guidelines for faceted search interfaces focus on supporting flexible navigation, seamless integration with directed search, fluid alternation between refining and expanding, avoidance of empty results sets, and most importantly making users at ease by retaining a feeling of control and understanding of the entire search and navigation process (Hearst 1999, 2006).

An example of faceted search and navigation system in biomedical domain is BioEve discovery engine (Ahmed et al. 2010), which identifies hidden relationships between entities like drugs, diseases, and genes and highlights them, thereby allowing the researcher to not only navigate the literature, but also to see entities and the relations they are involved in immediately, without having to read the article text fully, thus providing another aspect of searching relevant articles.

BioFacets (Mahoui 2006) (another good example of faceted navigation system) is an integration system for biological databases that provides for researchers a common interface for querying through multiple online databases with biological facets as a mechanism to restrict the search criteria and to browse and refine the results. Biological facets such as gene information are user-defined features that researchers can use to provide a multifaceted description of database records.

References

Ahmed ST, Kanwar SP, Hakenberg J, Davulcu H (2010) BioEve: a discovery engine for life sciences literature. In Proceedings of 6th international symposium on bioinformatics research and applications (ISBRA10), Storrs, 2010

Hearst M (2006) Design recommendations for hierarchical faceted search interfaces. In ACM SIGIR workshop on faceted search, Seattle, 2006

Hearst MA (1999) User interfaces and visualization. In: Baeza-Yates R, Ribeiro-Neto B (eds) Modern information retrieval. ACM Press/Addison-Wesley Longman, Harlow, pp 257–323

Koren J, Zhang Y, Liu X (2008) Personalized interactive faceted search. In Proceeding of the 17th international conference on World Wide Web, Beijing, 2008, pp 477–486, ACM, New York

Mahoui M, Miled ZB, Godse A, Kulkarni H, Li N (2006) Biofacets faceted classification for biological information. In Proceedings of the 18th international conference on scientific and statistical database management, pp 225–234, IEEE Computer Society, Washington, DC

Tunkelang D (2009) Faceted search. Synthesis lectures on information concepts, retrieval, and services, vol 1(1). Morgan & Claypool Publishers, San Rafael, pp 1–80

Secondary Metabolite Production in Streptomyces

Claudio Avignone-Rossa[1], Andrzej M. Kierzek[2] and Michael E. Bushell[3]

[1]Department of Microbial and Cellular Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, Surrey, UK

[2]Division of Microbial Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, Surrey, UK

[3]Department of Microbial and Cellular Sciences, University of Surrey, Guildford, Surrey, UK

Definition

The members of the genus *Streptomyces* are considered to be the most important producers of bioactive molecules, such as antibiotics, immunosuppressors, antibacterials, antifungals, antitumoral, pesticides, etc. The broad chemical diversity of the products synthesized by *Streptomyces* is caused by the presence in their genome of a variety of metabolic pathways collectively known as *secondary metabolism*. These secondary metabolic routes present their highest activity when the microorganisms undergo a series of developmental changes associated to the formation of aerial hyphae (in solid cultures) or to the onset of the stationary phase of growth (in liquid cultures). The setoff of those physiological states, and therefore of the associated synthesis of secondary metabolites, is linked to the depletion of growth nutrients, and a decrease in growth rate may be the signal for triggering secondary metabolism (Bibb 2005).

The information obtained from the sequencing projects of *S. coelicolor* (Bentley et al. 2002), *S. avermitilis* (Ikeda et al. 2003), and *S. griseus* (Ohnishi et al. 2008) has revealed a very large and unexpected number of genes linked to secondary metabolism. A theoretical calculation indicates that the number of potential antimicrobial compounds from the genus can be estimated to be in the order of 100,000 (Watve et al. 2001). This is of extreme importance and makes the members of the genus highly valuable in drug discovery programs and in the development of bioprocess for the production of molecules of pharmaceutical interest.

Characteristics

A Bit of History

Due to their importance as sources of bioactive molecules, much is known about the physiology and genetics of the genus *Streptomyces*. The initial steps in the industrial production of antibiotics in the 1950s gave rise to comprehensive and exhaustive research projects devoted to the elucidation of antibiotic biosynthetic mechanisms. However, the design of the bioprocess for antibiotic production was immediately recognized to be a very difficult task. Product yields and productivities from wild-type strains are generally very low: The antibiotic titers obtained from natural isolates are generally below 10 mg/ml of culture broth, considered to be too low for cost-effective production processes. However, the interest shown by the pharmaceutical industry prompted researchers to study ways to increase yields. Initial advances were mostly due to the application of the technique of submerged fungal fermentation (developed for the production of citric acid by *Aspergillus*) to replace the surface cultures used in primitive antibiotic production. The methods developed for the production of penicillin by *Penicillium* spp. were later adapted for the production of streptomycin by *Streptomyces griseus* and other antibiotics. An interesting and very detailed account is given by Waksman (1951). Over the following decades, the application of classical strain screening and selection methods resulted in the development of antibiotic-producing strains with titers approximately 50 times higher than those found in wild-type strains. The use of physical- and chemical-mutation protocols allowed attaining titers as high as 7,000 mg/l, while the development of culture media also helped to increase yields, in particular through the use of complex medium components supplemented with precursors and elicitors of unknown function.

However, all the strategies used for yield improvement were empirical and typical examples of "wait and see" approaches, lacking any metabolic or biochemical rationale. The consequence of this was that a successful strategy for one species and product could be ineffective for others.

The Current Situation

Low profit margins and the development of generic drugs have made the production of antibiotics less important from the commercial point of view. In the last two decades, most of the major pharmaceutical companies have discontinued their production, and there are practically no new antibiotics in the pipeline (Anon 2010b). However, the rapid emergence of "superbugs," presenting multidrug resistance to the available antibiotics have prompted the warning that untreatable bacterial infections might be back, resembling the pre-antibiotic era (Stokowski 2010). Only two new antibiotics have been approved for use in the past few years, and the development of new drugs might not be able to match the rate of generation of resistance. A number of international programs directed to the development of new antibiotics have been launched, such as the initiative 10 × 20 ("Ten new antibiotics for 2020") sponsored by the Infectious Diseases Society of America (Anon 2010a) among others (Mossialos et al. 2009).

Why Modeling *Streptomyces*?

The physiology and biochemistry of the genus has been very well characterized, and much is understood about the genetics and regulation of antibiotic biosynthesis. This is the product of almost 60 years of studies of several antibiotic-producing *Streptomyces* species. Interestingly, pioneering work in the early 1960s was directed to the use of chemostats for the analysis of the physiology of antibiotic production (Bartlett and Gerhardt 1959; Sikyta et al. 1959, 1961), providing a large body of knowledge and expertise, helpful to overcome the problems associated to the complex life cycle of the microorganism. *Streptomyces* is a spore-forming saprophytic soil-dwelling filamentous bacterium. The spores in soil germinate to produce mycelium under favorable condition. Upon germination, the cells show complex vegetative hyphal growth, which enables the microbe to colonize the soil. Under conditions of high concentration of nutrients and energy, the hyphae are highly branched and excrete a number of extracellular enzymes in order to obtain the required nutrients for growth. In contrast, conditions of nutrient scarcity restrict cell growth and aerial mycelium is formed, eventually differentiating into spores and therefore restarting the cycle. This morphological change is accompanied by metabolic changes which result in the induction of secondary metabolism. The formation of spores, a semi-dormant stage in the life cycle of the bacterium, serves the double purpose of resisting unfavorable environments and spreading the organism.

This life cycle is observed during growth on solid substrate, either natural (soil) or artificial (agar cultures). However, in order to obtain commercially attractive titers, the production of antibiotics (as it was discussed above) should be performed in submerged liquid cultures. In these systems, the life cycle of *Streptomyces* is less complex but presents a series of problems for the design of suitable bioprocesses. The filamentous growth of the microorganism in liquid cultures may cause mass transfer limitations leading to insufficient dissolved oxygen concentrations, a problem which requires manipulation of the fermentation conditions (e.g., increasing agitation rate, etc.).

Arguably, the most important problem in liquid cultures of *Streptomyces* species lies in the complex metabolic and regulatory systems triggered when cells enter the stationary phase of growth, when the pathways of secondary metabolism become fully active. The complexity of these systems is reflected in the high degree of variability found in *Streptomyces* cultures. Minor differences in culture variables may cause large differences in the performance of the microorganism, generally affecting the yields and productivities of the desired antibiotic.

This problem can be tackled by using systems biology approaches. These can help to understand and elucidate the metabolic pathways involved in antibiotic biosynthetic, and to identify and alleviate metabolic limitations. Metabolic modeling brings a quantitative picture of metabolism and physiology, allowing the identification of biosynthetic routes, the elucidation of missing metabolic links, and the description and prediction of phenotypes. The analysis of models of metabolism generated from genome sequences allows for the discovery of novel metabolic activities and provides an interpretation of experimental data within a metabolic framework. In the particular case of antibiotics and other secondary metabolites, this can be used to explain the causes of the low yields and productivities observed. Importantly, this approach may assist in the design of metabolic engineering strategies and in the development of bioprocess strategies by predicting the effect of modifications in the growth medium or in culture conditions.

Metabolic Models for *Streptomyces*

The reconstruction of a metabolic network requires information from metabolism and genome sequences to build the *stoichiometric matrix* (Matrix containing the stoichiometric coefficients of the metabolites in the reactions of a metabolic network.). Details of this procedure can be found in the literature (Price et al. 2003, 2004; Durot et al. 2009). A number of reduced metabolic models have been published for *Streptomyces*: *S. lividans* (Bull-Daae and Ison 1998; Avignone-Rossa et al. 2002) and *S. coelicolor* (Kim et al. 2004) and *S. clavuligerus* (Kirk et al. 2000). After the publication of the genome sequence, it has been possible to build very good genome-scale metabolic network models for *S. coelicolor* (Borodina et al. 2005; Bushell et al. 2006b; Khannapho et al 2008; Alam et al. 2010) and *S. clavuligerus* (Bushell et al. 2006a).

Stoichiometric models are mathematical representations of metabolism describing quantitatively the flow of mass through a metabolic network, containing all the known and physiologically feasible biochemical reactions in the microorganism's metabolism. A stoichiometric model is based on mass balances around the metabolites: The sum of all the fluxes producing any given metabolite minus the sum of all fluxes consuming the same metabolite generates one equation for each metabolite. The set of all linear equations can be solved by assuming that all reactions are in a dynamic steady state: The fluxes of all metabolites are numerically balanced. Stoichiometric models have been used to analyze the metabolism of two important species, *Streptomyces coelicolor* and *Streptomyces clavuligerus* (Kirk et al. 2000; Bushell et al. 2006a, b; Khannapho et al. 2008) with the objective of rationally designing bioprocesses for antibiotic production. Using a *genome-scale metabolic network* (stoichiometric models built using the genome sequence, genomic annotation and literature information, comprising all intracellular enzymatic reactions in a cell, tissue or whole organism) of *Streptomyces coelicolor*, a series of reactions was identified from unrelated pathways that participate in an actinorhodin-producing subnetwork. Simulations using the model predicted that the antibiotic biosynthetic activity of this network can be influenced by modifying the composition of the culture medium. Those predictions were tested experimentally, and the observed effects of the designed media showed antibiotic production rates similar to those calculated in the simulations. In an unrelated study, a metabolic model for *Streptomyces clavuligerus* was used to analyze the metabolism of the microorganism growing under different carbon and energy sources. This approach allowed to identify metabolic limitations affecting antibiotic biosynthesis, and to redesign the growth medium to alleviate those limitations. The results of the simulations were experimentally confirmed using a combination of metabolomics, transcriptomics, and flux analysis, constituting a clear example of the application of metabolic modeling in systems biology for bioprocess design. The analysis identified regions of metabolism that at first glance may be considered unrelated, but that can be connected to form subnetworks with activities that can be enhanced by modifying the supply of precursors through a reformulation of the culture media feeds.

All the internal metabolic fluxes can be calculated by solving the system using measurements of input and output fluxes (i.e., extracellular fluxes). There are two approaches to be used, depending on the size of the stoichiometric network. One is *metabolic flux analysis* (MFA), where the unknown intracellular metabolic fluxes can be calculated mathematically using measurements of external fluxes. This is normally used in reduced systems, where a numerical solution can be found using the available experimental measurements of extracellular rates.

The second approach is *flux balance analysis*, and it is based in the use of linear programming (optimization) to explore possible values of unknown fluxes. Large stoichiometric networks can be constructed from genome sequence information (genome-scale metabolic networks, GSMNs). Typically, there is insufficient input-output information to obtain unique values for each and every intracellular reaction: These are underdetermined systems presenting infinite solutions, and in order to obtain a solution they are constrained with data obtained from real-life experiments (transcriptomics, proteomics or metabolomics, enzyme reversibility, thermodynamics data, etc.). Such constrained systems are solved defining an objective function (e.g., maximization/minimization of growth, nutrient uptake rates, production rates, etc.), and the solution yields a potential internal flux distribution that satisfies the system.

Metabolic Flux Analysis for Bioprocess Design and Optimization: Clavulanic Acid Production by *Streptomyces clavuligerus*

Clavulanic acid is an antibiotic produced by *Streptomyces clavuligerus* (Reading and Cole 1977), attracting commercial interest due to its generic status. The knowledge of the genetics, physiology, and biochemistry of the pathways involved in the biosynthesis of the molecule provides a good basis to explore systems biology approaches for the rational design of cultures to improve product yields in particular, and for the understanding of secondary metabolism in general.

Experimentally, the use of continuous cultures allows the analysis of metabolic flux distributions in vivo at (quasi-) steady-state conditions. These cultures permit the cultivation of microorganisms under tightly controlled growth conditions, and are ideal for the analysis of the influence of medium components, growth conditions, and genetic manipulations that can be assessed with any other variables held constant.

The production of clavulanic acid in cultures of *Streptomyces clavuligerus* grown at different growth rates, nutrient limitation, and culture media, was analyzed for antibiotic yields and production rates. Amino acids have diverse roles in *Streptomyces* metabolism, among them regulatory roles in nitrogen metabolism, stimulation of β-lactam antibiotic synthesis by lysine, or inhibitory activity of cephalosporin biosynthesis by alanine. Furthermore, some amino acids are precursors of some antibiotics, including clavulanic acid. Previous reports using metabolic flux analysis and metabolomics had shown that the nature of the growth-limiting nutrient affects the availability of precursors and the titers of clavulanic acid. One of the precursors of clavulanic acid derives from glyceraldehyde 3-phosphate, and the availability of this C3 precursor is limited by a high activity of the anaplerotic

Secondary Metabolite Production in Streptomyces, Table 1 Effect of growth rate on the clavulanic acid production rates and yields in chemostat cultures of *Streptomyces clavuligerus*. The corrected yield (Yield$_{\text{corrected}}$) takes into account the carbon supplied by glycerol and the amino acid feed

Biomass production rate g.l^{-1}.h^{-1}	Glycerol consumption rate mole.g^{-1}.h^{-1}	Clavulanic acid production rate mmole.g^{-1}.h^{-1}	Yield mmole. mole$_{\text{glycerol}}$$^{-1}$	Yield$_{\text{corrected}}$ mg.g$_{\text{carbon}}$$^{-1}$
0.036	4.84×10^{-3}	1.60×10^{-3}	0.33	0.88
0.052	6.45×10^{-3}	1.46×10^{-3}	0.23	0.61
0.071	1.00×10^{-2}	0.81×10^{-3}	0.08	0.21
0.088	2.34×10^{-2}	0.54×10^{-3}	0.02	0.05

Adapted from Kirk et al. (2000). Biotechnol. Lett. 22, 1803–1809

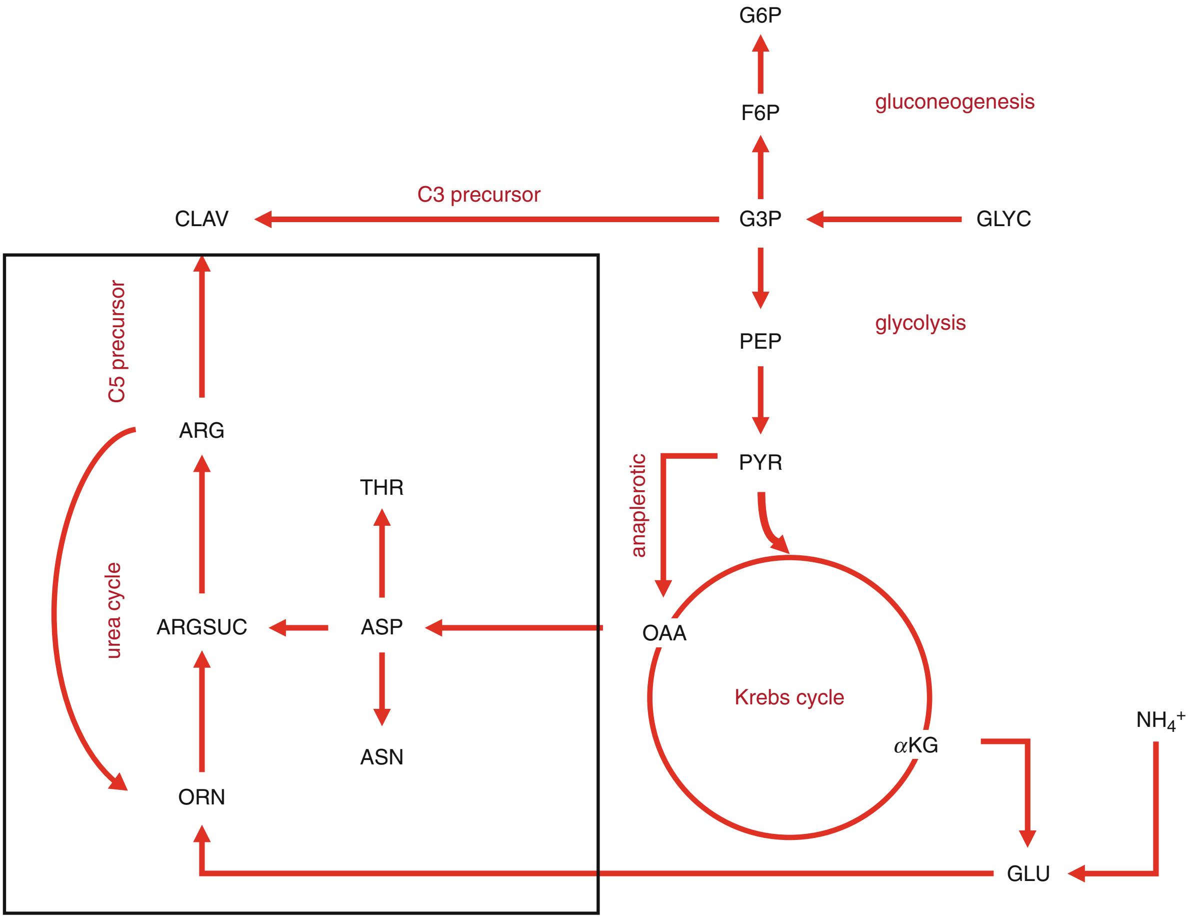
metabolism. In *S. clavuligerus*, the conversion of ornithine into arginine has been described, in a metabolic pathway akin to the urea cycle. Arginine is the C5 precursor in the synthesis of clavulanic acid. Intermediary metabolites and metabolic fluxes were identified which were highly correlated with antibiotic biosynthesis, and designed culture feeds alleviate limitation of C3 precursor supply, enhancing clavulanic acid production.

One of the main problems in the design of efficient bioprocesses for antibiotic production is that the carbon fluxes toward the final product (and through precursor intermediates such as amino acid pools) are very low compared to the metabolic fluxes through primary metabolism (Table 1). The analysis of the distribution of metabolic fluxes in antibiotic-producing cultures helps to identify intermediary metabolites whose production rates are critical to attainment of maximum yields. The carbon flux distribution through catabolic pathways depends on growth rate and nutrient availability: Increasing growth rates lead to increased flux through glycolysis and the pentose phosphate pathway. By applying MFA to a simplified metabolic network, it is possible to calculate intermediary metabolite fluxes and to estimate their influence on clavulanic acid production, with the ultimate goal of designing nutrient feeds enhancing clavulanic acid yield.

The metabolic network employed comprised the main primary metabolic pathways (glycolysis, Krebs cycle, pentose phosphate pathway), the routes for biosynthesis of precursors for biomass production and for biosynthesis of precursors for clavulanic acid, the routes for amino acid biosynthesis, the requirements for cofactors (ATP, NADH, NADPH), and an equation for the synthesis of biomass, derived from the biomass macromolecular composition determined experimentally. A simplified representation of the metabolic network is shown in Fig. 1.

In Table 1, the results obtained in chemostat cultures on *S. clavuligerus* grown at different growth (dilution) rates are shown. Clavulanic acid production rates and yields (expressed as amount of product formed per amount of substrate consumed) decrease with increasing growth rates, a result that is comparable with the behavior of the microorganism in solid cultures: The decrease in growth rate triggers the entry into stationary phase, and the secondary metabolism machinery becomes active. Exhaustion of substrates in soil, for instance, causes a decrease in growth rate and the onset of the events leading to the activation of secondary metabolism pathways.

Metabolic flux analysis of the system showed that the fluxes through the glycolytic pathway accounted for 92–98% of the carbon supplied by glycerol, the only carbon source, and that the flux through Krebs cycle represented approximately 50% of the total carbon input, while the activity of the pentose phosphate pathway activity remained very low, consistent with observed low biomass yield. The pentose phosphate pathway is the major source of biomass precursors, and a low biomass production rate would be reflected in low fluxes through the pentose phosphate pathway. The fluxes through the reactions of amino acid metabolism showed that approximately 5% of the carbon input flux goes through Asn, Asp, Thr, and Arg synthesis. Approximately 30% of the carbon input goes through Glu metabolism, while the flux through the urea cycle corresponds to approximately 15% of the carbon input flux. These values are relatively constant over the range of growth rates tested, probably reflecting the importance of glutamate in nitrogen metabolism and the major role played by the urea cycle in the metabolism of *S. clavuligerus*. The changes in fluxes toward the antibiotic do not correlate with fluxes through the urea cycle, supporting the previous findings that the urea cycle supplies



Secondary Metabolite Production in Streptomyces, Fig. 1 Schematic metabolic network for *Streptomyces clavuligerus*, showing the link between the synthesis of clavulanic acid and primary metabolism and amino acid biosynthesis. The box shows the area of amino acid metabolism highlighted by the results of metabolic flux analysis

nonlimiting amounts of the C5 precursor for clavulanic acid biosynthesis.

The synthesis of clavulanic acid shows a negative correlation to growth rate. An analysis was performed to determine the response of the fluxes through all the reactions in the metabolic network to growth rate (μ). All the reactions were ranked according to their correlation coefficient toward the synthesis of clavulanic acid (X_{clav}), using a cutoff value of 0.85. The most significant positive correlation was the flux for the synthesis of the C3 precursor glyceraldehyde 3-phosphate from 3-phosphoglycerate, while no significant correlations were found for reactions linked to C5 precursor. These results are consistent with an unlimited supply of Arg for clavulanic acid production under P-limitation.

The highest positive correlation is observed between the reaction producing glyceraldehyde 3-phosphate (the C3 precursor of clavulanic acid) from phosphoenolpyruvate. This reflects a high production rate of glyceraldehyde 3-phosphate, allowing increase of fluxes toward phosphoenolpyruvate and clavulanic acid under P-limitation. The reactions competing for precursors of G3P (i.e., X_{35}, X_{34}, X_2) show a strong negative correlation with the reaction for synthesis of clavulanic acid, X_{57} (Table 2).

Interestingly, the production of glutamate from glutamine (X_{42}) also shows a very strong negative correlation ($R = -0.99$) with clavulanic acid synthesis, as glutamate fuels reactions that compete with glyceraldehyde 3-phosphate synthesis by consuming phosphoenolpyruvate. This may indicate the need to

Secondary Metabolite Production in Streptomyces, Table 2 Metabolic fluxes showing highest correlation with clavulanic acid production rates ($r > 0.85$)

Reaction number	Reaction	Correlation coefficient with reaction 57
57	Clavulanic acid synthesis	
4	3-Phosphoglycerate + ADP + NAD → Glyceraldehyde 3-P + ATP + NADH	0.88
5	Glyceraldehyde 3-phosphate ↔ Phosphoenolpyruvate	0.85
34	Phosphoenolpyruvate + Erythrose 4-phosphate + NADPH + ATP + Glutamate → Phenylalanine + α-ketoglutarate + CO2 + ADP + NADP	−0.85
2	3-Phosphoglycerate + ADP ↔ Fructose 6-phosphate + ATP	−0.86
50	FADH + O_2 + ADP → ATP + FAD	−0.88
17	Ribulose 5-phosphate ↔ Xylose 5-phosphate	−0.89
35	Phosphoenolpyruvate + Erythrose 4-phosphate + NADPH + ATP + Glutamate + NAD → Tyrosine + α -ketoglutarate + CO_2 + ADP + NADP	−0.98
7	Pyr + NAD → Acetyl-CoA + NADH + ATP	−0.99
42	Aspartate + NAD ↔ UTP + ADP + Glutamate + NADH	−0.99

Secondary Metabolite Production in Streptomyces, Table 3 Effect of amino acid feed on the clavulanic acid production rates and yields in *Streptomyces clavuligerus* cultures grown at growth rate: 0.3 h^{-1}

Amino acid	Biomass production rate $g.l^{-1}.h^{-1}$	Glycerol consumption rate $mole.g^{-1}.h^{-1}$	Clavulanic acid production rate $mmole.g^{-1}.h^{-1}$	Yield mmole. $mole_{glycerol}{}^{-1}$	$Yield_{corrected}{}^{a}$ $mg_{clav}.g_{carbon}{}^{-1}$
None	0.036	4.84×10^{-3}	1.60×10^{-3}	0.33	1.81
Thr	0.046	5.27×10^{-3}	2.76×10^{-3}	0.52	2.86
Arg	0.052	4.30×10^{-3}	2.31×10^{-3}	0.54	2.94
Asp	0.038	4.95×10^{-3}	1.76×10^{-3}	0.36	1.94
Glu	0.039	4.41×10^{-3}	1.66×10^{-3}	0.38	2.04
Asn	0.046	3.44×10^{-3}	1.56×10^{-3}	0.45	2.48
Ile	0.031	5.16×10^{-3}	2.56×10^{-3}	0.50	2.71

[a] Yield value corrected for total carbon input (i.e., considering the consumption of the amino acids fed)

alleviate the limitations caused by the diversion of carbon from the central metabolic pathways toward the synthesis of amino acids through reactions competing with the synthesis of clavulanic acid precursors. These reactions are shown in a box in Fig. 1.

Feeds were designed based on the identity of the amino acids involved in the reactions with higher correlation coefficient, as discussed above. Previous results had shown that when amino acids are fed to cultures of *Streptomyces clavuligerus*, the intracellular pool sizes of the amino acid being fed increase, suggesting that fed amino acids block or inhibit their own biosynthesis. Therefore, it was assumed that the microorganism does not synthesize the amino acids fed, and therefore their biosynthetic reactions will carry a flux of 0. The growth rate was kept at 0.03 h^{-1}, and the amino acid concentration in the culture medium was 10 mM. The amino acids were undetectable in the culture supernatant, indicating that they have been completely consumed (Table 3).

As it can be seen, the amino acid feeds affect the antibiotic yields: compared to the control, Thr, Arg, and Ile promote yield increases of 58%, 64%, and 52%, respectively, whereas the increase in yield caused by Glu and Asp is in the order of 10% only. When MFA was applied to these results, variations in the flux distributions were also observed (Table 4). Amino acid feeds caused a reversal in X_2, which was offset by increases in X_{34}. As observed in the nonfed cultures, X_{42} showed a negative correlation with X_{57} ($R = -0.86$). A positive correlation ($R = 0.86$) was observed with the reaction converting pyruvate into

Secondary Metabolite Production in Streptomyces, Table 4 Values of fluxes through selected metabolic reactions calculated by application of metabolic flux analysis. The subindices refer to the reactions shown in Table 2. Values are expressed in C-mole to normalize to different substrate consumption rates

Amino acid	X_2 C-mole.l^{-1}.h^{-1}	X_7 C-mole.l^{-1}.h^{-1}	X_{34} C-mole.l^{-1}.h^{-1}	X_{42} C-mole.l^{-1}.h^{-1}	X_{57} C-mole.l^{-1}.h^{-1}
None	0.45	87.2	0.41	0.17	0.23
Thr	−3.98	87.8	0.35	1.12	0.25
Arg	−2.23	85.1	0.04	1.74	0.21
Asp	−8.49	82.5	0.64	1.70	0.16
Glu	−5.68	82.6	0.71	1.79	0.15
Asn	−9.45	84.5	0.64	1.84	0.14
Ile	−8.44	85.1	1.40	0.42	0.23

Secondary Metabolite Production in Streptomyces, Table 5 Effect of feeding combinations of amino acids on the clavulanic acid production rates and yields in *Streptomyces clavuligerus* cultures grown at growth rate: 0.3 h^{-1}

Amino acid	Biomass production rate g.l^{-1}.h^{-1}	Glycerol consumption rate mole.g^{-1}.h^{-1}	Clavulanic acid production rate mmole.g^{-1}.h^{-1}	Yield mmole. $mole_{glyc}{}^{-1}$	$Yield_{corrected}{}^{a}$ mg.$g_{carbon}{}^{-1}$
None	0.036	4.84×10^{-3}	1.60×10^{-3}	0.33	1.81
Arg/Asp/Thr	0.060	2.72×10^{-3}	6.68×10^{-3}	2.45	13.6
Arg/Asp/Thr/ Asn	0.049	2.39×10^{-3}	8.09×10^{-3}	3.38	18.7

[a] Yield value corrected for total carbon input (i.e., considering the consumption of the amino acids fed)

acetyl-CoA, indicating that a major effect of amino acids feeds is to cause an increase in the rate of entry of the latter into the Krebs cycle.

The Effect of Feeding a Combination of Amino Acids

As it was shown in the metabolic flux analysis discussed above, the biosynthesis of clavulanic acid depends on the supply of a C3-precursor, and it is potentially influenced by competing demands for intermediates by other primary metabolic routes such as glycolysis and the Krebs cycle. Therefore, feeds were designed based on combinations of amino acids derived from oxaloacetate.

The results in Table 5 show that a tenfold increase in antibiotic yields can be achieved by feeding a culture medium rationally designed by application of the results of an in silico analysis. This analysis showed that clavulanic acid synthesis is limited by the availability of the C3 precursor, and that the flux through amino acid biosynthesis may affect this availability. A way of relieving that limitation is by feeding those amino acids whose fluxes appear to limit clavulanic acid production. This resulted in yield increases of ca. 60%. The metabolic flux analysis results of these experiments showed that a combination of those amino acids promoted yield increases the yields up to ten times.

Use of a Genome-Scale Network for *Streptomyces clavuligerus*

The results showed above were obtained performing metabolic flux analysis using a small network. This calculation provides a unique solution, as the small size of the network generates an overdetermined system (A system of linear equations is overdetermined if the number of equations is larger than the number of unknowns. If the number of equations is smaller than the number of unknowns, the system is underdetermined.). Application of metabolic flux analysis to genome-scale metabolic networks is not possible, as the large systems of linear equations generated are necessarily underdetermined, and they do not have a unique solution. These systems must be solved using other approaches such as flux balance analysis or its variant, flux variability analysis, as discussed above.

We constructed a genome-scale network for *Streptomyces clavuligerus* using information from the genomes of *Streptomyces coelicolor* and *S. avermitilis*, combined with experimental data (transcriptome, proteome, metabolome, and enzymatic activities) and literature data for *Streptomyces clavuligerus*. The resulting network consisted of 724 metabolites, 837 reactions, 164 external metabolites, and 9 macromolecules.

We employed flux variability analysis (FVA) to explore the metabolic state of the system. In flux balance analysis, the fluxes computed are not necessarily unique; in FVA, the minimal and maximal admissible fluxes through each reaction are calculated. The objective function is constrained to its maximal value, minimizing and maximizing flux through each reaction in the system. The output is a range of reaction fluxes consistent with the maximal value of the objective function. The range of flux obtained is unique and define the internal flux distribution.

The 837 reactions were filtered according to the admissible flux ranges. Of these, 398 reactions were found to be nonessential to maximize clavulanic acid synthesis (their flux ranges = 0), while 439 reactions showed nonzero flux ranges (i.e., the difference between minimal and maximal fluxes satisfying the constraints is nonzero). The reactions showing the smallest variability (112) were considered to have the highest influence on clavulanic acid synthesis. Among those, the reactions involved in the synthesis of Thr, Arg, Asp, in agreement with the results obtained for the reduced metabolic network for *S. clavuligerus*.

Interestingly, FVA also showed that the biosynthesis of fatty acids may have a high influence in clavulanic acid synthesis. The results were used to analyze the effect of fatty acid feeds on clavulanic acid production. Simulations were run using fatty acid as carbon sources (long chain fatty acids: oleic, palmitic, linoleic, and linolenic acids) and with an active β-oxidation pathway. Their effect on clavulanic acid production was assessed by estimating the metabolic flux distribution when fatty acids are used as sole C-sources. The objective function was the maximization of clavulanic acid synthesis, and the results were compared to those obtained with glycerol as the C-source. The fatty acid biosynthesis pathways were considered to be inactive.

The theoretical yields were calculated and compared to the yields observed with glycerol as the sole C-source (Table 6). The values obtained showed that the expected yield with oleic acid is 29% higher than that obtained with glycerol, while the yield with palmitic and linoleic acid were 1% and 11% lower than those obtained with glycerol. Experimentally, cultures with oleic acid as the sole carbon source showed a yield of clavulanic acid 34% higher than the yield obtained using glycerol.

Secondary Metabolite Production in Streptomyces, Table 6 Theoretical and experimental yields of clavulanic acid in cultures of *Streptomyces clavuligerus* with fatty acids as carbon sources

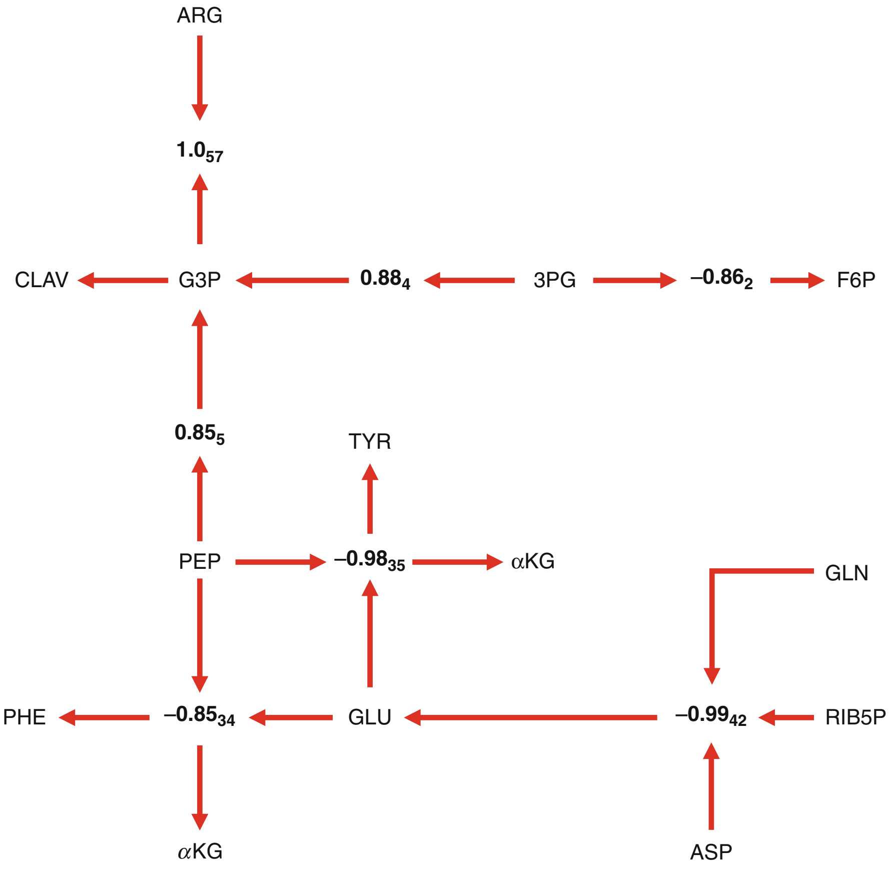
C-source	Theoretical yield	Experimental yield
Glycerol	1	0.87
Oleic	1.29	1.17
Palmitic	0.99	n.d.
Linoleic	0.89	n.d.

Conclusion

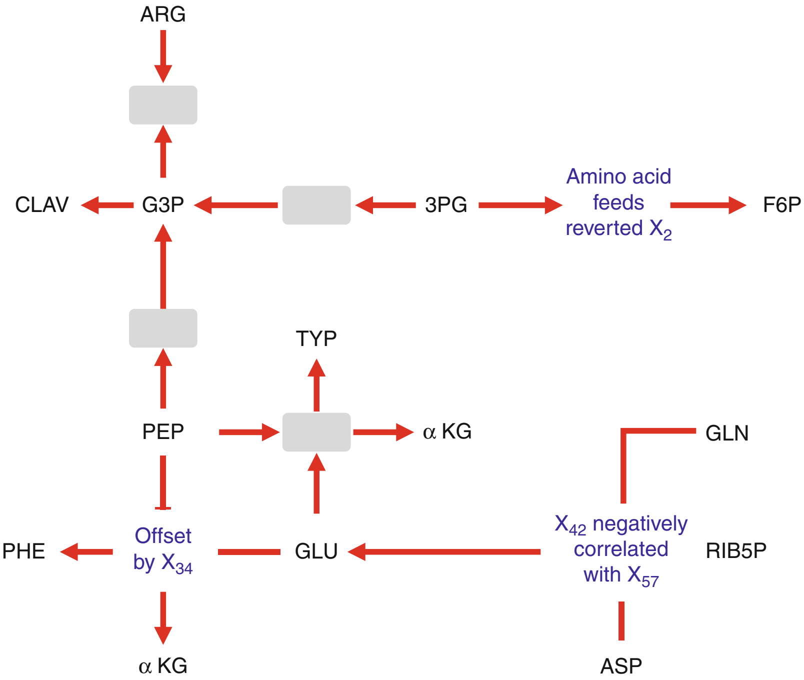
These results demonstrate that there is a major role for metabolic flux analysis and constraint-based flux balance modeling in the design of bioprocesses such as the production of antibiotics, in which the maximization of yields is desired.

While MFA can only be applied to small-scale metabolic networks, the results obtained are of extreme importance not only for the identification of possible metabolic bottlenecks but also for the design of culture media to alleviate those limitations. In the example discussed here, we identified metabolic limitations affecting antibiotic biosynthesis, and redesigned the growth medium to alleviate those limitations. The predictions were experimentally confirmed.

FBA and FVA can be applied to genome-scale metabolic networks, and both approaches provide distributions of metabolic fluxes that represent solutions to the system. The solution of FBA provides no unique single flux value, while FVA defines a feasible flux range for each individual reaction. However, the limits of each range are unique values, providing a set of parameters for quantifying the solution space. In our example, FVA was used to analyze the metabolism of *S. clavuligerus* growing under different carbon sources. FVA not only confirmed the results obtained with a small-scale network, but it also highlighted areas of metabolism from relatively disparate pathways that may be considered unrelated



Secondary Metabolite Production in Streptomyces, Fig. 2 Metabolic scheme showing the reactions that are highly correlated (correlation coefficients >0.85) to clavulanic acid biosynthesis. Subindices refer to the reactions numbers indicated in Table 2


ARG
CLAV
G3P
3PG
Amino acid feeds reverted X₂
F6P
TYP
PEP
α KG
GLN
PHE
Offset by X₃₄
GLU
X₄₂ negatively correlated with X₅₇
RIB5P
α KG
ASP

Secondary Metabolite Production in Streptomyces, Fig. 3 Metabolic scheme showing a metabolic interpretation of the results of metabolic flux analysis (e.g., fatty acid metabolism) but that affect the production of the desired molecule. The activity of those pathways can be altered by modifying the supply of precursors through a reformulation of the culture media feeds. The results allowed us to identify novel substrates promoting high yields of the antibiotic, and experimental observations using the designed media showed antibiotic yields similar to those predicted in the simulations.

References

Alam MT, Merlo ME, The STREAM Consortium, Hodgson DA, Wellington EMH, Takano E, Breitling R (2010) Metabolic modeling and analysis of the metabolic switch in Streptomyces coelicolor. BMC Genomics 11:202

Anon (2010a) The 10 × '20 initiative: pursuing a global commitment to develop 10 new antibacterial drugs by 2020. Clin Infect Dis 50:1081–1083

Anon (2010b) The urgent need: regenerating antibacterial drug discovery development. Report of the British Society for Antimicrobial Chemotherapy Initiative. http://www.bsac.org.uk/Resources/BSAC/TUN%20Report.pdf

Avignone Rossa C, White J, Kuiper A, Postma PW, Bibb M, Teixeira de Mattos MJ (2002) Carbon flux distribution in antibiotic-producing chemostat cultures of streptomyces lividans. Metabolic Engineering 4:138–150

Bartlett MC, Gerhardt P (1959) Continuous antibiotic fermentation. Design of a 20 litre, single-stage pilot plant and trials with two contrasting processes. J. Biochem. Microbiol. Technol. Eng. 1:359–377

Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D et al (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417:141–147

Bibb MJ (2005) Regulation of secondary metabolism in streptomycetes. Curr Opin Microbiol 8:208–215

Borodina I, Krabben P, Nielsen J (2005) Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. Genome Res 15:820–829

Bull-Daae E, Ison AP (1998) A simple structured model describing the growth of Streptomyces lividans. Biotechnology and Bioengineering 58(2–3): 263–266

Bushell ME, Kirk S, Zhao H, Avignone-Rossa CA (2006a) Manipulation of the physiology of clavulanic acid biosynthesis with the aid of metabolic flux analysis. Enz Microbial Technol 39:149–157

Bushell ME, Sequeira SIP, Khannapho ZH, Chater KF, Butler MJ, Kierzek AM, Avignone-Rossa CA (2006b) The use of genome scale metabolic flux variability analysis for process feed formulation based on an investigation of the effects of the zwf mutation on antibiotic production in *Streptomyces coelicolor*. Enz Microbial Technol 39:1347–1353

Durot M, Bourguignon PY, Schachter V (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol Rev 33:164–190

Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Omura S (2003) Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. Nat Biotechnol 21:526–531

Khannapho C, Zhao H, Bonde BK, Kierzek AM, Avignone-Rossa CA, Bushell ME (2008) Selection of objective function in genome scale flux balance analysis for process feed development in antibiotic production. Metabolic Eng 10:227–233

Kim HB, Smith CP, Micklefield J, Mavituna F (2004) Metabolic flux analysis for calcium dependent antibiotic (CDA) production in *Streptomyces coelicolor*. Metab Eng 6:313–325

Kirk S, Avignone-Rossa CA, Bushell ME (2000) Growth limiting substrate affects antibiotic production and associated metabolic fluxes in *Streptomyces clavuligerus*. Biotechnol Lett 22:1803–1809

Mossialos E, Morel C, Edwards S, Berenson J, Gemmill-Toyama M, Brogan D (2009) Policies and incentives for promoting innovation in antibiotic research. London School of Economics and Political Science, London. http://www.se2009.eu/polopoly_fs/1.16814!menu/standard/file/LSE-ABIF-Final.pdf

Ohnishi Y, Ishikawa J, Hara H, Suzuki H, Ikenoya M, Ikeda H, Yamashita A, Hattori M, Horinouchi S (2008) The genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. J Bacteriol 190:4050–4060

Price ND, Papin JA, Schilling CH, Palsson BO (2003) Genome-scale microbial in silico models: the constraints-based approach. Trends Biotechnol 21:162–169

Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886–897

Reading C, Cole M (1977) Clavulanic acid: a β-lactamase-inhibiting β-lactam from *Streptomyces clavuligerus*. Antimicrob Agent Chemother 11:852–857

Sikyta B, Doskoáil J, Kaŝparovĉ J (1959) Continuous streptomycin fermentation. J Biochem Microbiol Technol Eng 1:379–392

Sikyta B, Slezak J, Herold M (1961) Growth of *Streptomyces aureofaciens* in continuous culture. Appl Environ Microbiol 9:233

Stokowski LA (2010) Emerging antibiotics: will we have what we need?. Medscape Inf Dis. Available at: http://www.medscape.com/viewarticle/715971

Waksman S (1951) Streptomycin: isolation, properties, and utilization. J Hist Med Allied Sci VI:318–347

Watve MG, Tickoo R, Jog MM, Bhole BD (2001) How many antibiotics are produced by the genus Streptomyces?. Arch Microbiol. 176(5):386–390

Secondary Structure 2D Structure

► Protein Structure Metapredictors

Selected Reaction Monitoring

▶ Selective Reaction Monitoring

Selective Pressure

Philippe Huneman
Institut d'Histoire et de Philosophie (IHPST), des Sciences et des Techniques, Université Paris 1 Panthéon-Sorbonne, Paris, France

Definition

In a given environment, environmental parameters which impinge differently on different organisms' chances of reproduction and survival, according to the values of their heritable traits (e.g., growth rates of predators, scarcity of resources, heat, etc.). The selective pressures are not always known, especially in small populations.

Cross-References

▶ Explanation, Evolutionary

References

Lewens T (2009) The natures of selection. Brit J Phil Sci 61(2):1–21

Selective Reaction Monitoring

Stefanie Wienkoop
Department for Molecular Systems Biology, University of Vienna, Vienna, Austria

Synonyms

Absolute protein quantification; Multiple reaction monitoring (MRM); Selected reaction monitoring; Stable isotope dilution technique

Definition

▶ Selective Reaction Monitoring (SRM) is a mass-spectrometry-based technique for the absolute quantification of a targeted protein. Absolute quantification is enabled by spiking stable isotope labeled (heavy) target peptides (e.g., 13C and 15N) of known concentration into a complex sample. These stable isotope standard peptides are also referred to as signature or proteotypic peptides and need to be carefully selected prior to SRM. Due to a specific mass shift (label) the mass spectrometer monitors the signals of the standard and the native (unlabelled or light) target peptide simultaneously. After precursor ion selection of the target peptide(s), this ion(s) is then fragmented to yield product ions. A precursor/product pair is also referred to as a transition or reaction. In order to improve/ensure signal selectivity and sensitivity, the specificity of the reaction(s) is important. Given that these highly selective signals are in linear range (usually around four orders of magnitude) quantification in absolute terms can be calculated by comparing signals of standard peptide (known concentration) and native peptide.

Cross-References

▶ Mass Spectrometry, Proteomics, and Metabolomics
▶ Proteomics, Quantification-Unbiased and Target Approach

Selenocysteine

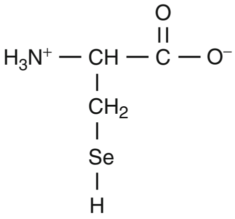
Taiichi Sakamoto[1] and Gota Kawai[2]
[1]Chiba Institute of Technology, Narashino, Japan
[2]Department of Life and Environmental Sciences, Chiba Institute of Technology, Narashino, Chiba, Japan

Synonyms

3-Selanyl-2-aminopropanoic acid; L-Selenocysteine

Definition

Selenocysteine (Fig. 1), which is the major biological form of the element selenium, has a structure similar to



Selenocysteine, Fig. 1 Selenocysteine

that of cysteine, but with an atom of selenium taking the place of the usual sulfur, forming a selenol group (Yuan et al. 2010).

Proteins that contain one or more selenocysteine residues are called selenoproteins (Alberts et al. 2008; Hüttenhofer and Böck 1998). The 21st amino acid is typically found in catalytic centers of selenoproteins where it plays a functionally essential role. Selenocysteine is incorporated into polypeptides to form selenoproteins through translation recoding. Selenocysteine is enzymatically produced from a serine attached to a special tRNA molecule that forms base pairs with the UGA codon, which is normally used as a stop codon (see ▶ Translational Control by cis RNA Elements, Bacteria). During translation, selenocysteinyl-tRNA^{Sec} is delivered to the ribosome by a specific translation factor that requires a characteristic stem-loop structure in the mRNA to recode an UGA from stop codon to selenocysteine sense codon.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2008) Molecular biology of the cell, 5th edn. Garland Science, New York

Hüttenhofer A, Böck A (1998) RNA structures involved in selenoprotein synthesis. In: Simons RW, Grunberg-Manago M (eds) RNA structure and function. Cold Spring Harbor Laboratory Press, New York

Yuan J, O'Donoghue P, Ambrogelly A, Gundllapalli S, Sherrer RL, Palioura S, Simonović M, Söll D (2010) Distinct genetic code expansion strategies for selenocysteine and pyrrolysine are reflected in different aminoacyl-tRNA formation systems. FEBS Letters 584:342–349

Self-Organization

Kepa Ruiz-Mirazo
Department Logic and Philosophy of Science, University of the Basque Country, Donostia – San Sebastián, Spain
Biophysics Unit (CSIC-UPV/EHU), University of the Basque Country, Leioa, Spain

Definition

Self-organization is a dynamic phenomenon in which a large number of individual units (molecules, cells, multicellular organisms) spontaneously generate a global, irreducible correlation that brings and holds them together, i.e., a collective pattern of order or behavior that involves all of those interacting units and cannot be explained just from their individual properties.

Characteristics

Although there are only a few fundamental forces or physical interactions in nature, the diversity of systems that we actually observe in it demonstrates that matter has found many different ways to come together. This "coming together" of material parts is classified as "self-organization" when it occurs spontaneously, without following some external goal, plan, or design, and when the process of association of – a necessarily big number of – those parts does not imply changing their intrinsic nature (otherwise, the system would be *transforming* into something else, not self-organizing).

The limits of this concept, in any case, are fuzzy, and various examples may challenge that classification criterion or, at least, be used ambivalently. For instance, some phase transitions (e.g., the transition from liquid water to ice) could be regarded as a self-organizing process. The transition occurs spontaneously (if a general condition of the system, like the temperature or pressure, changes); it involves many parts (water molecules), which find a new way of being correlated to the others; and it does not alter the intrinsic nature of those parts (H_2O) – even if their dynamic properties, for instance, do change. Someone could argue, however, that the system is not self-organizing but transforming into something else, i.e., that liquid water is *essentially* different from ice. A very different example would be chemical oscillations, also called "chemical clocks." Think, for instance, of a Belousov-Zhabotinsky (B-Z) reaction, in which some components are being converted into some others but in a cyclic way, so that the original reactants are also products of the process and, given the right combination of chemical activators and inhibitors of that process, periodic oscillations in the concentrations of the different parts of the system (or chemical waves, if there is also spatial diffusion) are observed. Here individual molecules are continuously changing their intrinsic nature but the system, as a whole, remains the same (i.e., it is made of the same type of molecules, even if the numbers/concentrations of each chemical species vary in time and space). Whereas the first case was clearly in the realm of physics, touching the bottom threshold of what is to be considered "self-organization," the second case, already within chemistry, brings us closer to the upper ceiling of it, in the sense that it faces the boundary with "self-producing" dynamics, which constitute the proper domain of metabolic systems. (Biological systems 'recruit' the self-organising properties of matter and make extensive use of them at different levels of complexity, but certainly go beyond self-organisation in a strong sense, for all living organisms are self-constructing agents (Ruiz-Mirazo et al. 2004)).

The previous two examples also illustrate a major distinction in the general domain of self-organizing phenomena in nature. Depending on whether the correlation among material parts consists in a dynamic and dissipative (i.e., far-from-equilibrium) pattern, like in the second case, or in a more stable, structural (equilibrium or quasi-equilibrium) configuration, like in the first, these types of processes are usually regarded as "self-organization" or "self-assembly." Certain phase transitions and, more commonly, the formation of supramolecular structures that do not involve covalent bonding (e.g., lipid membranes, micelles, polymer aggregation, polymer folding, etc.) are considered self-assembling processes (Lehn 1995). These are phenomena based on many weak interactions (van der Waals forces, hydrogen-bond formation, medium-range electrostatic/ionic forces, hydrophobic effect, etc.) acting at the same time among the different building blocks that make up the supramolecular structure(s), whose characteristic regularities (i.e., the emergent spatial and/or temporal periodicities) are roughly of the same order of magnitude than the dimensions and time scales of the actual building blocks. Again, if covalent bonds (or, more generally, chemical reactions) were included, then we would be facing the edge of what is an *organization* – versus a *transformation* – process.

The other type of phenomena, sometimes also called "dissipative structures" (Nicolis and Prigogine 1977), which occur just in far-from-equilibrium conditions and are sustainable only if there is a net flow of matter/energy through the system, constitute proper self-organization, according to the more standard and strict interpretation of the term. In these cases, fluctuations in the vicinity of bifurcation points, amplified by non-linear interactions among the material parts of the system are responsible for the collective dynamic behavior they get into. And the temporal and spatial scales of the collective patterns generated in this way are, at least, several – but typically more than several – orders of magnitude larger than those of the parts. The aforementioned B-Z reactions, whirls and hurricanes, or some complex convection patterns in liquids (e.g., Bénard cells) are the most frequently used examples, although there are many others in biology (e.g., chemical signaling, morphogenesis, ecosystem dynamics, etc.), ethology (e.g., social insects, flock movement, etc.), or cognitive sciences (neural network pattern generation).

One of the reasons why this second type of phenomena (nonequilibrium, dissipative structures) are considered to be more genuine "self-organization" is because they mark out a clearer distinction with systems traditionally dealt with in physics, by means of statistical mechanics or classical thermodynamics (Yates 1987). In contrast with self-assembly of closer-to-equilibrium structures, the link between local properties (microscopic description level) and global behavior (macroscopic or mesoscopic description levels) in dissipative structures is more intricate, and standard statistical methods have proved unsuitable to bridge that gap. In fact, there is not a well-established theory, so far, to explain or predict, starting from microscopic parameters, the key features of the macroscopic patterns generated under those far-from-equilibrium conditions. No state function of the system has been found (analogous to a thermodynamic 'free energy function') to be minimized or maximized along the process. The correlation among parts is understood as the result of the nonlinear amplification of a local fluctuation but, at the same time, is a necessary condition for the organizing phenomenon itself. Somehow, the collective-macroscopic pattern is both cause and effect of the individual-microscopic dynamics of the underlying units. Additional boundary conditions or external constraints are typically required to keep the system running, staying away from equilibrium (e.g., a thermal gradient, the input of material resources, etc.). But it is characteristic of this type of phenomena that, at least, one of the critical boundary conditions that bring the system about, and maintain it there, be endogenously created. And this is very difficult to express or make operational mathematically (technically speaking, it is a "non-holonomic" constraint). Hence the special relevance of the prefix "self-" in the term "self-organization": it refers to the spontaneous, inner generation of the collective pattern that immediately acts as a dynamic constraint on the behavior of the individual units, which, in turn, realize and reinforce the pattern.

The theoretical modeling or simulation of self-organizing phenomena has been tackled from many different standpoints and with diverse tools. Given the widely accepted difficulties of extending linear irreversible thermodynamics to far-from-equilibrium conditions (Nicolis and Prigogine 1977), a more fruitful way of analysis has been through a "dynamical systems" type of approach, i.e., handling a set of coupled differential equations (e.g., Turing's (1952) reaction–diffusion equations) that, depending on the specific boundary conditions of the system, become more or less complicated to integrate. In simple cases, global patterns of spatial and temporal order (e.g., oscillatory concentration profiles or spreading chemical waves) can be predicted and studied in detail (determining the stability of the different solutions, number of dynamic attractors of the system, bifurcation points, etc.). Nevertheless, as soon as the system becomes moderately complex (in terms of diversity of components, diversity of interactions, nonlinear reaction couplings, etc.), this type of treatment becomes impracticable. Although this has not prevented the application of the 'dynamic systems' framework to very intricate cases, like cognitive systems (Kelso 1995), managing to capture some of their basic constitutive and interactive features, too strong simplifications of the underlying processes were always involved in those approaches. Alternative attempts to overcome such intrinsic limitations in our understanding of self-organization have been made, for instance, through phenomenological-macroscopic approximations to the problem (Haken 1983), but their degree of success and explanatory power has been relatively modest.

However, the rise of "complex systems science" in the last decades (including here, in particular, the remarkable advances in network theory and agent-based modeling), together with the increasing power of computers and simulation methods, have provided wider possibilities to study self-organization phenomena. Most of the key ideas had already been conceived (earlier in cybernetics, systems theory, artificial intelligence, or physics itself) but they crystallized and have been further developed since the last part of the twentieth century. A surprisingly successful strategy has consisted in representing the system as a discrete and distributed set of units (nodes in a network, cells in a grid); each of these units shows a rather simple dynamic state (typically, a binary state: "on/off" or "up/down") and also simple – though often variable – connections with some others (connections according to which their individual states, at each subsequent time step, will be defined). Taking up this general approach (which may be implemented through numerous methods: random Boolean networks, cellular automata, neural networks, "spin-glass" (Ising-type) models, etc.), one can deal with systems of many components in a radically different way (as compared to standard "dynamical systems" or "statistic mechanics" approaches): namely, putting the emphasis on the degree of connectivity and the diverse types of interaction that the components of a system may present, and focusing on the effect that these interconnections have on its global properties.

One could think that, under such extremely simplifying assumptions for the individual dynamic behavior of each of the units, and by introducing just a few local rules of interaction among them, this type of strategy would have little applicability to real systems that exhibit self-organizing properties. Quite the contrary, it has proved successful across many disciplines, helping to explain key emergent features of regulatory genetic networks (Kauffman 1993), complex ecosystem dynamics (Solé and Bascompte 2006), or the behavior of social insects (Camacine et al. 2001). Furthermore, the fact that some collective properties of big ensembles of interacting units are, to a large extent, independent of the intricacies that each of the units actually involve has been interpreted as a sign of the potential universality of some organizational principles in nature. If this proves to be the case, and general principles of organization for complex, many-particle systems are eventually established (Bak 1996), the fundamental framework of other scientific theories will surely have to be reconsidered. In particular, the consequences that this would have for evolutionary theory and systems modelling in biology should be addressed.

More specifically, the self-organization paradigm is bound to play a central role in the unraveling of the origins of life, in bringing light to the still obscure transition from the realm of physics and chemistry toward biological complexity. Indeed, without the self-organizing capacities of matter, the appearance of life on an inert planet, like the primitive Earth, would seem completely unfeasible. Kant, for instance, who regarded matter as something essentially passive (with compounds aggregating into bigger structures just through random mechanical associations) considered that it was impossible to approach this problem scientifically (see comments in: Fry 2000, p. 181–182). But how would his thinking have changed, had he known about 'dissipative structures'? Wouldn't he have postulated 'self-organization' as the obvious bridge between the inert and the living? Despite the fact that, so far, most contributions supporting this claim in the field of origins of life have been theoretical (e.g., Kauffman's model of 'autocatalytic sets' – see, again, Kauffman 1993), the recent advances in the analysis of complex chemical mixtures, or so-called 'systems chemistry', through the use of microfluidics and dynamic combinatorial chemistry methods (Ludlow and Otto 2008), together with the increasing awareness that all living cells fundamentally depend on self-organising processes (Karsenti 2008), will surely open the way for important experimental landmarks, along those lines, in years to come.

The new approaches should take into account that biological organization is not only based on far-from-equilibrium processes but also on quasi-equilibrium supra- and macro- molecular structures, which *constrain* and yet enhance those processes. Therefore, the most promising avenues of research will be those that, keeping a 'bottom-up' perspective (i.e., taking relatively simple molecular units as the starting point), attempt to combine the dynamics of self-organization and self-assembly. Several interesting model systems that integrate these two different types of collective dynamic behaviour are currently under experimental exploration (e.g., oscillatory reactions taking place in micro-emulsions (Epstein & Vanag 2005) or chemically reacting 'self-propelled' oil droplets (Hanczyc et al. 2007)). However, they have not been designed following prebiotic-plausibility criteria for the components involved. Furthermore, the synergy between self-assembly and self-organization should be more specifically channelled, in this context, towards the implementation of minimally robust and autonomous proto-cellular systems.

The diverse (more or less robust and autonomous) forms of proto-cellular organization that such a line of research should bring about in the lab, adequately interpreted with the help of computer modeling and simulations, would surely contribute to develop a richer idea of how self-organizing dynamics may take place within self-assembling compartments and complex networks of interacting molecules. More specifically: an idea of self-organization conceived not just as a global, highly distributed process of generation of connectivity patterns, but as a process in which modularity must also emerge, together with functional and hierarchical relationships among the units (and intermediate modules) of the network. In this way, the transition from chemical feedback loops towards increasingly sophisticated mechanisms of regulation and control (not only within the system but also outwards, in its relationship with the environment) should also be illuminated. For all these certainly constitute distinctive features of the more complex types of organization observed throughout the biological world.

Acknowledgments Thanks to Alvaro Moreno for reading and commenting on a previous version of this entry, as well as to the *Ramón y Cajal Program (MICINN) for supporting my research position.*

Cross-References

- ▶ Agent-based Modeling
- ▶ Boolean Networks
- ▶ Cellular Automata
- ▶ Circadian Rhythm
- ▶ Closure, Causal
- ▶ Complex System

- ▶ Constraint
- ▶ Correlation Relationship
- ▶ Emergence
- ▶ Ensemble
- ▶ Epigenetics
- ▶ Feedback Regulation
- ▶ Holism
- ▶ Interlevel Causation
- ▶ Metabolic Networks, Evolution
- ▶ Modularity
- ▶ Ordering Parameter
- ▶ Pattern
- ▶ Reaction-Diffusion Equations
- ▶ Synchronization

References

Bak P (1996) How Nature Works: The Science of Self-Organized Criticality. Copernicus Books

Camazine S, Deneubourg J-L, Franks NR, Sneyd J, Theraulaz G, Bonabeau E (2001) Self-organization in biological systems. Princeton University Press, Princeton

Epstein IR, Vanag VK (2005) Complex patterns in reactive microemulsions: self-organized nanostructures? Chaos 15, 047510–1–7

Fry I (2000) The emergence of life on Earth: a historical and scientific overview. Rutgers University Press, London

Haken H (1983) Synergetics: nonequilibrium phase transitions and self-organization in physics, chemistry, and biology. Springer, Berlin

Hanczyc MM, Toyota T, Ikegami T, Packard N, Sugawara T (2007) Fatty acid chemistry at the oil-water interface: self-propelled oil droplets. JACS 129:9386–9391

Kauffman S (1993) The origins of order: self-organization and selection in evolution. Oxford University Press, Oxford

Karsenti E (2008) Self-organization in cell biology. A brief history. Nat Rev 9:255–262

Kelso JAS (1995) Dynamic patterns: the self-organization of brain and behavior. MIT Press, Cambridge, MA

Lehn J-M (1995) Supramolecular chemistry: concepts and perspectives. Wiley, New York

Ludlow RF, Otto S (2008) Systems chemistry. Chem Soc Rev 37:101–108

Nicolis G, Prigogine I (1977) Self-organization in non-equilibrium systems. Wiley, New York

Ruiz-Mirazo K, Peretó J, Moreno A (2004) A universal definition of life: autonomy and open-ended evolution. Orig Life Evol Biosph 34:323–346

Solé RV, Bascompte J (2006) Self-organization in complex ecosystems. Princeton University Press, Princeton

Turing AM (1952) The chemical basis of morphogenesis. Philos Trans R Soc Lond B 237:37–72

Yates FE (ed) (1987) Self-organizing systems. The emergence of order. Plenum Press, New York

Self-Regulation

- ▶ Regulation and Autoregulation

Self-Renewal

Steven D. Rhodes
School of Medicine, Indiana University, Indianapolis, IN, USA

Self-renewal is the property of a particular cell, such as a stem cell, to divide mitotically and replace itself while retaining potency – the capacity to give rise to multiple distinct cell types.

Cross-References

- ▶ Single Cell Assay, Mesenchymal Stem Cells

Self-Replication

Basilio Vescio
Department of Clinical and Experimental Medicine, Magne Graecia University of Catanzaro, Catanzaro, Italy

Definition

Self-replication is any behavior of a dynamical system that yields construction of an identical copy of the system itself. Self-replication can be observed in nature, where biological cells duplicate themselves by cell division. During cell division, DNA is replicated. Another example of self-replication is given by biological and computer viruses.

Von Neumann pioneered the research in the field of self-replicating systems and established its theoretical foundations (Von Neumann 1966). Replicators are nowadays categorized on the basis of the amount of support they require:

- Natural replicators have all or most of their design from nonhuman sources. Such systems include natural life forms.
- Autotrophic replicators can reproduce themselves "in the wild." They mine their own materials. It is conjectured that nonbiological autotrophic replicators could be designed by humans, and could easily accept specifications for human products.
- Self-reproductive systems are conjectured systems which would produce copies of themselves from industrial feedstocks such as metal bar and wire.
- Self-assembling systems assemble copies of themselves from finished, delivered parts. Simple examples of such systems have been demonstrated at the macroscale.

References

Von Neumann J (1966) Theory of self-reproducing automata. University of Illinois Press, Urbana/London

Self-Similarity

Basilio Vescio
Department of Clinical and Experimental Medicine,
Magna Graecia University of Catanzaro, Catanzaro,
Italy

Definition

In mathematics, a self-similar object is exactly or approximately similar to a part of itself (i.e., the whole has the same shape as one or more of the parts). Many objects in the real world, such as coastlines, are statistically self-similar: parts of them show the same statistical properties at many scales (Mandelbrot 1967). Self-similarity is a typical property of fractals.

References

Mandelbrot B (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. Science 156: 636–638

Semantic Frame Filling

▶ Template Filling, Text Mining

Semantic Integration

▶ Data Integration and Visualization

Semantic Web

Mark A. Musen
Stanford Center for Biomedical Informatics Research,
Stanford University, Stanford, CA, USA

Definition

A vision for computer interoperability in which intelligent agents communicate with one another over the World Wide Web. Knowledge is encoded using **ontologies** that are stored on the Web, and problem solvers (such as Web services) interpret the ontologies to generate intelligent behavior. Thus, the data available on the Web are stored in a manner that is suitable for processing by computers, rather than as text intended to be readable by humans. The World Wide Web Consortium views the Semantic Web as a "Web of data." The first step toward the Semantic Web has been work to make data available as linked data, represented in RDF.

Cross-References

▶ Protégé Ontology Editor

Semantic Web, Interoperability

Carole Goble, Sean Bechhofer and Katy Wolstencroft
School of Computer Science, University of Manchester, Manchester, UK

Definition

Systems Biology requires the integration and interpretation of many different types of data from many different sources. Understanding how these data can be compared and combined is a complex interoperability task that requires the understanding of both the structure and content of the different data sources, and an understanding of how these sources could be made to work together coherently. As the amount and complexity of biological data continues to increase, scale becomes an issue. There are now over 1,200 different public databases available to the life sciences (Galperin and Cochrane 2011). For biologists to benefit from this wealth of information these databases must interoperate and their content must effectively become an interlinked web of knowledge that can be easily navigated and searched. The Semantic Web is a means of facilitating this.

At the heart of the Semantic Web is the idea of publishing information as linked statements that are sufficiently well described that the information can be automatically processed. The original vision (Berners-Lee et al. 2001) was to augment the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other. This enables software to access the Web more intelligently and perform tasks on behalf of users. There were two primary motivations. The first was scale. The Web was principally intended for low-volume browsing and consumption by humans, whereas the Semantic Web would support machine processing to aid discovery, filtering, and the use of high volumes of data. The second was interoperability and integration as information spread across the Web needed to be exchanged and combined. Thus, the Semantic Web community has defined a collection of standards, languages, and principles for information interoperability as well as tools and technologies that implement them. These tools and technologies have been designed to be layered on top of the Web's preexisting infrastructure. However, they are sufficiently general to be used independently of the Web. Semantic Web tools and technologies are already being used in the Life Sciences for a wide range of data management tasks, such as: data annotation, mapping between data resources, data retrieval, knowledge management, and inferring new biological connections between processes. The Semantic Web concept is particularly well suited to communities where information encompasses a limited and defined domain, and where sharing data is a common necessity, such as in scientific research.

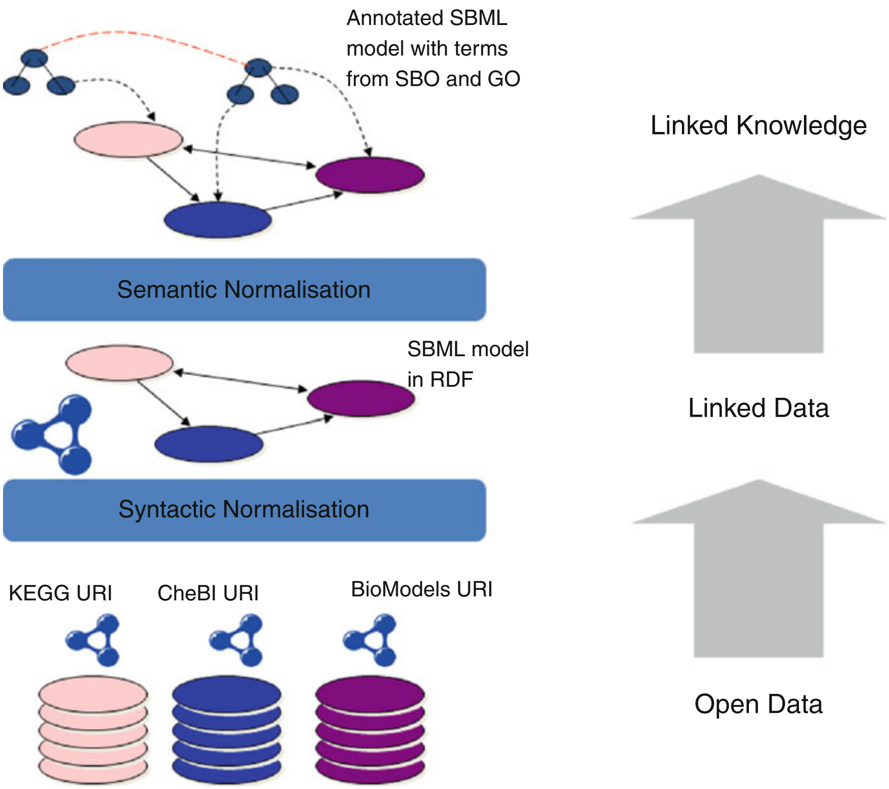
Characteristics

Semantic Web Infrastructure

Publishing Semantic Web data takes a layered approach and follows a set of conventions. Identity, structure and meaning are considered independently. One principle behind interoperability is to normalize incompatibilities through the adoption of common representations (see Fig. 1). Normalizing identities means that we can draw together information for an entity. Normalizing the structure of the data by using a common data model overcomes syntactic incompatibility. Normalizing the meaning of the data by using controlled vocabularies overcomes semantic incompatibilities. Thus, a number of standards have been defined, building on one another, to provide rich and shared descriptions using common data models for syntactic interoperability and shared controlled vocabularies for semantic interoperability.

Defining common data identity: The Internationalized Resource Identifier (IRI – a more general form of URI) is used to identify names and resources on the Web uniquely. The use of common identities by publishers of Semantic Web data is crucial to interoperability; for example, IRIs may identify data (sets), or terms in ontologies. Mapping services such as "sameAs" (http://sameas.org) – a part of the Semantic Web infrastructure – can identify entities with more than one IRI; a published "same-as" statement can then link them. Unlike URLs, IRIs are location independent and their authorities have obligations to guarantee

Semantic Web, Interoperability, Fig. 1 An example of the layers of Semantic Web technologies in use in Systems Biology



their long-term persistence. An example from systems biology is "Identifiers.org" (http://identifiers.org) developed and maintained by the MIRIAM standardization initiative.

Defining common data structure: The Extensible Markup Language (XML http://www.w3.org/XML) provides a common representation format for arbitrary data structures. Life Sciences, XML is widely used for specifying markup languages, such as SBML (Systems Biology Markup Language) (Hucka et al. 2003) or MAGE-ML (Microarray Gene Expression Markup Language) (Spellman et al. 2002). These descriptions standardize the metadata associated with a particular type of data so that every instance contains the same set of information. For instance, each SBML model has an author and a list of species and reactions, and each microarray experiment has an author (experimentalist) and descriptions of the hybridization and normalization methods used. These markup languages define a common data structure, but they do not describe relationships between metadata items. However, RDF and related standards enable these connections.

The Resource Description Framework (RDF) provides a common and flexible data model where data is organized into triples and represented as graphs. An RDF triple consists of a *subject*, *predicate*, and *object* (e.g., the *nucleus* [SUBJECT] is *part of* [PREDICATE] the *cell* [OBJECT]). The subject of one triple may be the object of another (and vice versa). There are many kinds of cells containing various organelles, for example; an RDF representation of that information will consist of a graph of objects (identified by IRIs), connected by links, which can further be linked to web pages and information from databases. A graph is a data structure consisting of a partially ordered set of edges (links) between nodes (entities). RDF triple stores can be queried through SPARQL (http://www.w3.org/TR/rdf-sparql-query), a query language that allows the matching of graph patterns (e.g., a SPARQL query could return all cell types that contain a nucleus). RDF is one of the core W3C Semantic Web standards for representing data on the web.

Defining common data meaning: RDF provides *syntactic* normalization but does not guarantee *semantic* interoperability. For this, we require agreement on the vocabularies that are used within resource descriptions. A shared *controlled vocabulary* or *ontology* constrains and standardizes the terminology used in descriptions and reduces semantic incompatibilities. For example, the same proteins in different databases can have many different names and identifiers (e.g., Glutamate carboxypeptidase 2 is also known as folate hydrolase 1, membrane glutamate carboxypeptidase, and by at least six other synonyms). However, if each database annotates proteins with common identifiers from UniProt (Magrane and Consortium 2011), or protein functions with Gene Ontology terms (Harris et al. 2004), it makes information about the same objects directly comparable and could allow inferences between related objects. RDF Schema (RDF(S)) and Simple Knowledge Organization System (SKOS) are additional W3C standards that support the description (and navigation) of such vocabularies by describing the relationships between them.

Vocabularies and terminologies can be further enriched by using more expressive, logic-based languages. The Web Ontology Language (OWL) allows for richer descriptions of the terms and concepts within a vocabulary. The term *ontology* is often used to refer to a collection of statements or axioms that provide a vocabulary structured using knowledge of the way in which its terms should be interpreted and how logical inferences should be made about them. Logical inference can be used for querying information sources, or to support ontology construction by managing hierarchies and identifying contradictions (Rubin et al. 2008). For example, the term ***species*** occurs in the Systems Biology Ontology and NCBI taxonomy, but the term has a different meaning in each context: in the NCBI taxonomy, it is a specific level in the taxonomy of living organisms; in the SBO, it refers to the elements of a biochemical reaction. The term is ambiguous, but in each context is specific and nonequivalent.

There are approximately 300 biological ontologies available through the BioPortal repository (http://bioportal.bioontology.org/). Many of these are available in OWL. For a review of Systems Biology and ontologies see (Courtot et al. 2011).

Semantic Web Content

The Semantic Web's infrastructure is independent of its *content*. To publish interoperable Semantic Web content requires that information providers agree upon common frameworks and common controlled vocabularies or ontologies for annotation. In the life sciences, there are many standards initiatives developing markup languages such as SBML and MAGE-ML; ontologies such as GO (the Gene Ontology) (Harriset al. 2004), CHEBI (Chemical Entities of Biological Interest) (Degtyarenko et al. 2008) and SBO (Systems Biology Ontology) (Le Novere 2006); and metadata specifications such as MIBBI (Minimum Information for Biological and Biomedical Investigations) (Taylor et al. 2008) and ISA (Investigation, Study, Assay) (Rocca-Serra et al. 2010). By basing these biological initiatives on technologies and expertise relating to the Semantic Web, such standards developers can use established tools and more easily establish interoperability with resources from other domains – particularly important in Systems Biology. Furthermore, scientists not only need to draw on resources from across the life sciences, but also from medical informatics, mathematical modeling, and other disparate domains. Following industry-wide standards allows this kind of interaction.

Practical Applications of the Semantic Web

The Semantic Web standards and technologies provide structural layers and guidelines for making use of the Semantic Web, but they can be employed in a number of different ways to gain added value. In the life sciences, they are used for a range of tasks from simple annotation of data to distributed knowledge discovery.

The Annotation Web. Here, the emphasis is on a separation of presentation from content, with annotations providing additional information about resources. This involves tagging or marking up web pages with assertions about their content (i.e., the original content is augmented with annotations). This approach was taken in Annotea (Kahan et al. 2001), an early W3C project aimed at supporting collaboration through the sharing of metadata, such as bookmarks or notes, and can be seen in recent initiatives such as the Annotation Ontology (Ciccarese et al. 2011) and the Open Annotation Collaboration (http://www.openannotation.org). Annotations can be embedded in pages using Microformats (http://microformats.org) and RDFa, making those pages both human and machine interpretable. Annotations can also be held separate from resources (requiring referencing mechanisms). SemanticSBML (Krause et al. 2010) is an example of the annotation web approach in Systems Biology. This resource allows the annotation of arbitrary data items with terms from a large collection of ontologies and vocabularies relating to Systems Biology. It was originally designed to help with the annotation of Systems Biology models, but is also an important tool for linking experimental data and models for comparison and validation.

The Data Web. Here the emphasis is on breaking down and combining silos of data (and their schemas) by publishing them in a common framework. Rather than a document-centric approach that considers publishing resources/documents along with annotations, the primary publication is of *data*, exported as RDF documents or accessed using the SPARQL *endpoint* protocol (a service supporting query of a triple store) that builds on top of the regular Web HTTP protocol. The Data Web approach has come to prominence recently through interest in *Open Data* and *Linked Data*.

Open Data are those made freely available for third parties to reuse and republish. There has been particular activity around the publishing of government and scientific data, including through the Science Commons. A number of such open data initiatives have adopted Semantic Web technologies.

Linked Data are those published according to guidelines and rules intended to facilitate their discovery, navigation, integration, and reuse (http://linkeddata.org). Linked Data require a common identifier scheme (e.g., IRIs) along with meaningfully described (typed) links between resources; in particular, the provision of links *between* datasets, using those identifiers. Linked Data are increasingly commonplace, with ever more data sources being exposed in the *Linked Data Cloud*. Examples of Linked Data publication include Bio2RDF (Belleau et al. 2008), Chem2Bio2RDF (Chen et al. 2010), and LinkedLifeData (http://linkedlifedata.com).

The Inference Web. Here, the emphasis is on querying or reasoning across concepts and data to generate new knowledge. In the inference web, data and concepts are already semantically linked and are used for formulating new hypotheses based on the *inference* of new relationships. The size and number of resources available to Systems Biology and the interdisciplinary nature of the work make it impossible for individual scientists to track all new research to catch what portion may be relevant. Querying Semantic Web resources allows connections to be made. Inference is also important in discovering possible inconsistencies in integrated data, acting as a quality control and guiding data curation.

The BioGateway (Antezana et al. 2009) provides the means to query a large number of systems biology ontologies and data sources represented in RDF. The uniform representation and shared semantics provides a rich environment for formulating new hypotheses and performing complex queries over multiple data sources. However, this approach involves converting native formats of relational databases and flat files into RDF and creating a central RDF store. While it is effective, there is a large overhead in the development and maintenance of such a resource. Systems that allow inferences over distributed, native RDF may prove more practical for long-term use, but this requires the provision of these resources in RDF. Pioneering projects like the BioGateway demonstrate the advantages of such approaches and encourage more resources to adopt RDF.

The value of reasoning over SBML models converted into OWL has been demonstrated by (Hoehndorf et al. 2011). Their framework integrates representations of in silico Systems Biology with those of in vivo biology as described by biomedical ontologies. An SBML Harvester automatically converts annotated SBML models from the BioModels Database into OWL, generating a knowledge base for complex biological queries that bridges levels of granularity and supports model verification. The reasoning revealed curation errors in published models. Similar work has been demonstrated by Lister, Pocock, and Wipat (2007).

The Semantic Web and the Life Sciences in the Future

This entry focuses on the current problems of data integration and interoperability from the point of view of knowledge discovery. This foundational work will support a range of applications; for example, to address concerns about the reproducibility of scientific results and the disconnection between published work and experimental data. If data featured in publications are readily obtainable from public repositories, the presentation of work can be more transparent. Several Semantic Web tools focus on this issue; for example, Utopia Documents (http://getutopia.com) uses the Bio2RDF Linked Data cloud (http://www.bio2rdf.org) to allow researchers to directly interact with data presented in scientific publications, transforming static representations of data into live content.

The Semantic Web is an active area of research in Computer Science and the Life Sciences provide an ideal test bed for many of the tools and ideas being developed. The Open PHACTS project (Open Pharmacological Concepts Triple Store), (http://www.openphacts.org) which is a European Innovative Medicines Initiative, is using Semantic Web approaches to make existing pharmaceutical knowledge available for linking, querying and reasoning across public and commercial domains. Open PHACTS, and other projects like it, are bringing the use of the Semantic Web into mainstream research. With the rise in popularity of RDF and Linked Data in particular, we are seeing a shift from promising prototypes and proofs of concept to large-scale adoption by life science service providers.

References

Antezana E, Blonde W et al (2009) BioGateway: a semantic systems biology tool for the life sciences. BMC Bioinformatics 10(Suppl 10):S11

Belleau F, Nolin M-A et al (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform 41(5):706–716

Berners-Lee T, Hendler J, Lassila O (2001) The Semantic Web, Scientific American, p. 29–37

Chen B, Dong X et al (2010) Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics 11:255

Ciccarese P, Ocana M et al (2011) An open annotation ontology for science on web 3.0. J Biomed Semantics 2(Suppl 2):S4

Courtot M, Juty N et al (2011) Controlled vocabularies and semantics in systems biology. Mol Syst Biol 7:543

Degtyarenko K, de Matos P et al (2008) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 36(Database issue):D344–D350

Galperin MY, Cochrane GR (2011) The 2011 nucleic acids research database issue and the online molecular biology database collection. Nucleic Acids Res 39(Database issue): D1–D6

Harris MA, Clark J et al (2004) The gene ontology (GO) database and informatics resource. Nucleic Acids Res 32-(Database issue):D258–D261

Hoehndorf R, Dumontier M et al (2011) Integrating systems biology models and biomedical ontologies. BMC Syst Biol 5:124

Hucka M, Finney A et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19(4):524–531

Kahan J, Koivunen MR et al (2001) Annotea: an open RD infrastructure for shared web annotations. In Proceedings of the Tenth International World Wide Web Conference, Hong Kong

Krause F, Uhlendorf J et al (2010) Annotation and merging of SBML models with semanticSBML. Bioinformatics 26(3):421–422

Le Novere N (2006) Model storage, exchange and integration. BMC Neurosci 7(Suppl 1):S11

Lister AL, Pocock M, Wipat A (2007) “Integration of constraints documented in SBML, SBO, and the SBML manual facilitates validation of biological models.” J Integr Bioinform 1 Oct 2007, IB07

Magrane M, Consortium U (2011) “UniProt knowledgebase: a hub of integrated protein data.” Database (Oxford) 2011: bar009.

Rocca-Serra P, Brandizi M et al (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics 26(18):2354–2356

Rubin DL, Shah NH et al (2008) Biomedical ontologies: a functional perspective. Brief Bioinform 9(1):75–90

Spellman PT, Miller M et al (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 3(9):RESEARCH0046

Taylor CF, Field D et al (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26(8):889–896

SemanticSBML

Franco duPreez
SysMO-DB team, University of Manchester,
Manchester Centre for Integrative Systems Biology,
Manchester, UK

Definition

SemanticSBML (http://www.semanticsbml.org) is an online service that facilitates comparison of biological models (Schulz et al. 2011). Its common uses include clustering groups of models, as well as alignment of models describing the same biological system. The similarity measure employed depends on the annotation that accompanies the model descriptions supplied in ▶ Systems Biology Markup Language (SBML) format. SemanticSBML also aids the process of annotating new models by employing automated searches of annotation resources using existing model identifiers, potentially returning a list of ▶ MIRIAM URNs for each of the species and reaction identifiers used in the model.

Cross-References

▶ JWS Online

References

Schulz M, Krause F, Le Novère N, Klipp E, Liebermeister W (2011) Retrieval, alignment, and clustering of computational models based on semantic annotations. Mol Syst Biol 7:512

Semidefinite Program

Frank Allgöwer, Jan Hasenauer and Steffen Waldherr
Institute for Systems Theory and Automatic Control,
University of Stuttgart, Stuttgart, Germany

Definition

A semidefinite program (SDP) is an optimization problem with a linear objective function and affine constraints with an optimization variable in the cone of positive semidefinite matrices. In mathematical terms, a semidefinite program can be written as

$$\begin{aligned} &\underset{x \epsilon \mathcal{S}^n}{\text{minimize}} \quad \operatorname{trace}(CX) \\ &\text{subject to trace } (A_i X) = b_i, \quad i = 1, \ldots, m, \\ &\qquad\qquad X \geq 0, \end{aligned}$$

where $\mathcal{S}^n$ denotes the space of symmetric $n \times n$ matrices, C and A_i, $i = 1, \ldots, m$ are symmetric $n \times n$ matrices, and b_i, $i = 1, \ldots, m$ are real numbers. The constraint $X \geq 0$ means that X is positive semidefinite, i.e., $v^T X v \geq 0$ for all vectors $u \in \mathbb{R}^n$.

Semidefinite programs are ▶ convex optimization problems and thus have a unique optimal objective function value. Many practical optimization problems can be formulated as semidefinite programs, making them amenable to efficient solution algorithms. The term *semidefinite programming* is commonly used to denote the activity of formulating and solving semidefinite programs. The standard algorithms to solve semidefinite programs are interior point methods.

Cross-References

▶ Model Falsification
▶ Semidefinite Programming

References

Vandenberghe L, Boyd S (1996) Semidefinite programming. SIAM Rev 38:49–95

Semidefinite Programming

▶ Model Falsification, Semidefinite Programming

Senescence

Sergio Moreno
Instituto de Biología Molecular y Celular del Cáncer,
CSIC/Universidad de Salamanca, Salamanca, Spain

Synonyms

Cellular aging

Definition

Permanent growth arrest in G1 that protects cells from different forms of stress and persistent hyperproliferative signals. Together with apoptosis, senescence provides a defense mechanism against aberrant proliferation. Defects in these processes are associated with tumorigenesis. During senescence, cells undergo morphological changes: they become larger and flattened, express a senescence marker, beta-galactosidase, and undergo changes in gene expression.

Cross-References

▶ CDK Inhibitors

Sensitivity

► True Positive Rate

Sensitivity Analysis

Alexandros Kiparissidis, Efstratios Pistikopoulos and Athanasios Mantalaris
Biological Systems Engineering Laboratory,
Department of Chemical Engineering, Centre for
Process Systems Engineering, Imperial College,
London, UK

Synonyms

Model analysis; Uncertainty analysis

Definition

Sensitivity analysis is the study of how sensitive the output of a model is to variation in the values of its input factors. SA can apportion the total variance observed in the model output to various sources of variation, should more than one be present (Saltelli et al. 2000). It can provide an indication of the structure of the model and identify the presence of abundant parameters and nonnecessary variables. It is considered to be an important step in model validation, and in conjunction with experimental work, it can increase confidence in a model.

Characteristics

Sensitivity analysis methods are commonly grouped in three main categories, namely:

- Screening methods
- Local methods
- Global methods

Screening methods are randomized, one-at-a-time numerical experiments, which aim to indicate the most important factors among the totality of model parameters. While screening methods involve computationally efficient algorithms, their use is limited to only preliminary results due to calculation of only first-order effects (i.e., effects the input factors have on the model output, without including their mutual interactions) and they inherently lack precision, especially when used on nonlinear models. Efforts to calculate higher-order effects, through screening methods, have been recorded in the literature, though these methods fall short either in terms of accuracy or computational time. Screening methods are usually preferred for large-scale models or for initial estimates as they are an economical analysis that provides a first view of the model's behavior. The basic concept behind screening methods is the experience that in most cases a small number of parameters tend to account for the majority of the variability in the model output. In exchange for the short computational time, one must sacrifice from the amount and quality of information he receives. Therefore, screening methods mostly give qualitative results and cannot provide an estimate of how much more important one factor is over the other.

Local- or derivative-based sensitivity analysis derives measures of importance by estimating the effects infinitesimal variations of each factor have on the model output, in the area of a predetermined nominal point. The product of local sensitivity analysis is derivatives of the variation of the model output with respect to the input variables. In essence, we get a map of slope information, thus gaining insight to the model's response near a certain operating point. This could provide useful information not only about the nature of the studied set of input factors but also for robustness of the process at the point under consideration. Model reduction and lumping can be based on results of local SA. Local methods are commonly used on steady-state models or on studies dealing with the stability of a nominal point. Consequently, local methods fail to capture large variations in the parameter set and can only account for small variations from the parameter nominal values.

Global sensitivity analysis (GSA) methods have the advantage of performing a full search of the parameter space, hence providing data independent of nominal points and are applicable to the whole range of the model's existence. Moreover, global methods apportion the total uncertainty in the model output to the various sources of variation, while all parameters are varied at the same time. GSA provides the most complete set of results and mapping of the system, being able to cope with nonlinearities and identify parameter interaction effects. The main drawback of GSA methods is their extensive computational cost for large models.

Cross-References

- ▶ Cell Cycle Models, Sensitivity Analysis
- ▶ Global Sensitivity Analysis

References

Saltelli A, Chan K, Scott EM (2000) Sensitivity analysis. Wiley, Chichester

Sensitivity and Robustness, Master Equation

Ruiqi Wang
Institute of Systems Biology, Shanghai University, Shanghai, China

Definition

Robustness characterizes the ability to maintain performance in the face of perturbations and is one of the essential features of cellular systems. Sensitivity characterizes the ability of living organisms to adequately react to certain stimulus. In deterministic modeling, robustness is usually quantified by calculating sensitivity, e.g., period and amplitude sensitivity in quantifying robustness of circadian rhythms.

Using an analogue of the classical sensitivity analysis, the parameter sensitivity can also be applied to master equations. In stochastic systems, the state is probability distribution and parameters affect it indirectly through a master equation. Therefore, sensitivity of master equation is given by the change of probability distribution upon changes in parameters.

References

Gunawan R, Cao Y, Petzold L, Doyle FJ III (2005) Sensitivity analysis of discrete stochastic systems. Biophys J 88: 2530–540

Sequence Motif

▶ Motif

Sequence Ontology

Karen Eilbeck[1,2] and Carson Holt[3]
[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
[2]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT, USA
[3]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA

Synonyms

SO

Definition

The Sequence Ontology (SO) (Eilbeck et al. 2005) is an ontology that controls the vocabulary classifying the contents of genomic sequence and how those parts of the genome relate to each other. It forms the basis of structuring and validating genomic annotations and provides the vocabulary for naming the genomic features in databases and file formats. The use of ontologies to type biological data provides two essential advantages for the management of large datasets. First, it provides a unification of the terminology used by different communities, ensuring interoperability and enabling comparative analyses. Second, it provides the relations linking classes within the ontology allowing the data to be computationally reasoned over; this simplifies quality control and nurtures scientific discovery.

Characteristics

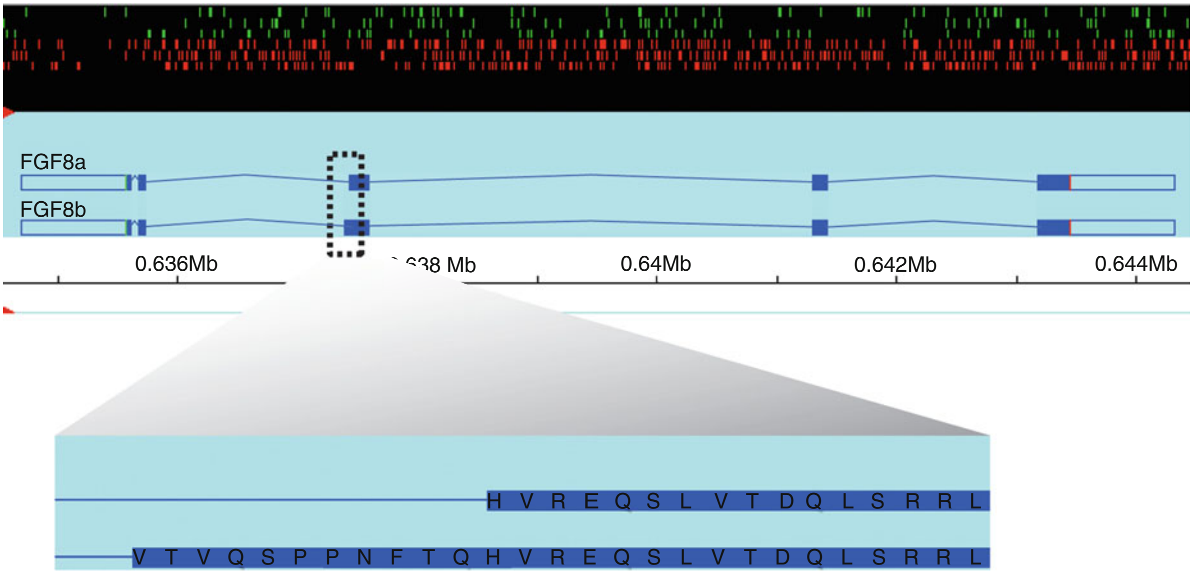
Genomic Annotation

Genomic annotation is the process whereby knowledge and evidence about parts of the genome are attached to the assembled sequence via genomic coordinates. A structural genomic annotation captures information and knowledge pertaining to the contents of a genome: the kinds of features, such as *exons* and *introns*, located within the sequence. The Sequence Ontology specifies the kinds of genomic features and how they relate to each other hierarchically and topologically. To illustrate, an *mRNA* is_a kind of *transcript*, whereas an *exon* is part_of a *transcript*. This is in contrast to functional genomic annotation whereby the function of a gene product, be it protein or RNA is annotated using functional descriptions such as that of the ▶ gene ontology, protein domains, or processes and pathways.

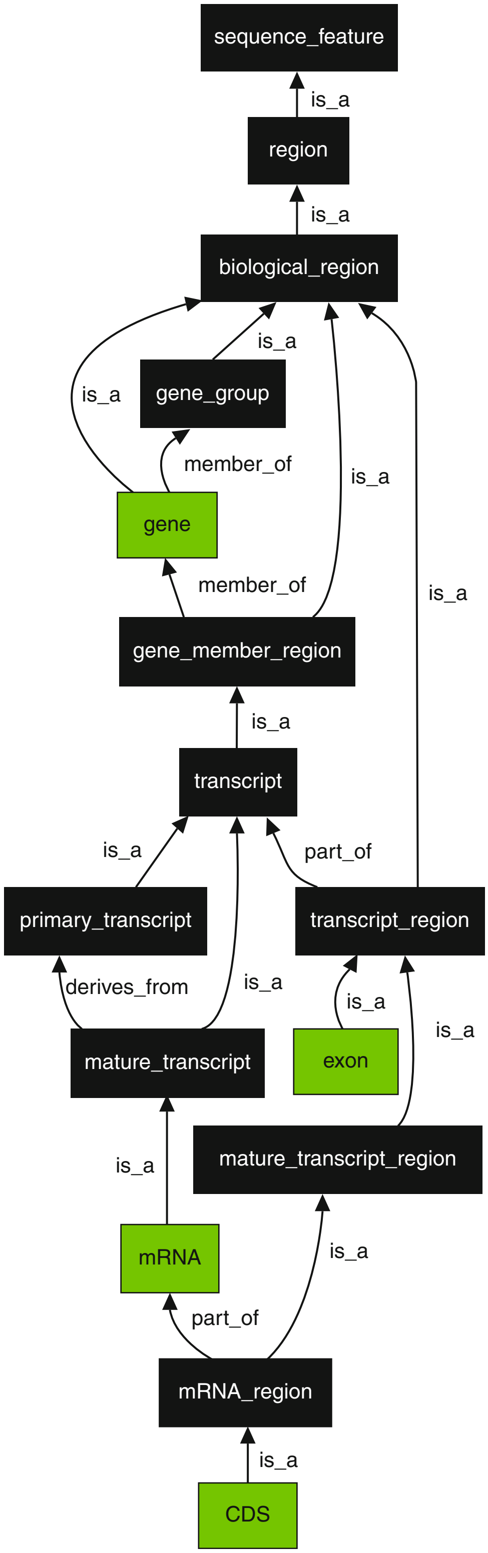
The structural annotation is the foundation upon which other knowledge is added. As illustrated in Fig. 1, the structural annotation provides two alternate transcripts that the functional annotation can be ascribed to. The annotations can be viewed using a genome browser such as GBrowse (Stein et al. 2002) or Apollo (Lewis et al. 2002). Structural genomic annotations form the basis of many molecular biology experiments. For example, at the simplest level, the design of PCR primers relies on the correct annotation of the feature coordinates such as the start and stop codons and the intron/exon boundaries.

What is in the Ontology?

Ontologies organize their terms into hierarchies and networks, this means that data labeled with terms drawn from an ontology (i.e., annotations) become substrates for computer-based logical inference, a necessity for making full use of very large data sets. The key feature distinguishing ontologies from controlled vocabularies is that ontologies use relationships to connect terms. The is_a and part_of relationships, for example, are commonly used to relate terms. Traversing the terms in the ontology via their relationships tells a user more about a given term. Following is_a relationships upward from a term to the root of the ontology, for example, defines what a term "is". Traversing the part_of relationships defines the composition of the genomic features, such as *introns* and *exons* form the parts of *transcripts*. This is how ontologies capture human knowledge in a computable format. Figure 2 shows a portion of the ontology used to describe the gene model represented in Fig. 1.


FGF8a
FGF8b
0.636Mb
0.638 Mb
0.64Mb
0.642Mb
0.644Mb
H V R E Q S L V T D Q L S R R L
V T V Q S P P N F T Q H V R E Q S L V T D Q L S R R L

Sequence Ontology, Fig. 1 Apollo produces a visual representation of FGF8 using the relationships among Sequence Ontology terms in the GFF3 file. The genomic sequence runs from left to right with transcripts displayed in *blue*. *Red* and *green* vertical bars in the dark panel represent start and stop codons in each reading frame. Zooming in on the region indicated by the *dotted line* shows how alternative splicing leads to the truncation of FGF8a in relation to FGF8b. The differences in the amino acid sequences due to the truncation are shown in the blowup of the same region



Sequence Ontology, Fig. 2 (continued)

Using SO to label the raw sequence data allows files and databases to be created that structure the genomic data, using the unified vocabulary and the defined relationships, thus removing ambiguity in the interpretation of the data. The ontology then provides the means to automatically validate the contents of an annotation. Validation is the process whereby an annotation is checked for mistakes. Statements made in the annotation such as *intron* part_of *mRNA* can be flagged as a problem because such a relation is not described in the ontology. This provides quality control of the data.

The Sequence Ontology is available as both released versions and frequent minor revisions made via a concurrent versioning system (CVS). SO is organized as a directed acyclic graph. It contains less than 2,000 terms falling into four areas: sequence features, sequence attributes, sequence variants, and sequence collections. The majority of terms are sequence features. These are the features that can be located within the sequence using genomic coordinates.

There are many kinds of features that can be annotated onto a genome sequence; the most obvious being the annotation of the location of biological features like genes onto the chromosomes. Gene models are composed of one or more transcript features which are in turn composed of a number of other features such as untranslated regions, coding regions, exons, introns, start codons, stop codons, and noncanonical splice sites. Other derived features such as the alignments made by sequence similarity searches may also be described and annotated using SO.

Sequence Ontology Resources

There are several ways in which to contribute to the development of SO. There is a term request tracker where term requests and updates can be made. There is also a mailing list for developers and those using SO to annotate data. The SO houses a wiki for the community to use for development. The SO is developed using OBO format (Day-Richter et al. 2007) and is nightly

Sequence Ontology, Fig. 2 A portion of the Sequence Ontology showing terms and relationships used in the annotation of the FGF8 gene model shown in Fig. 1. The *green boxes* are terms used in the annotation, and *black boxes* are terms in the ontology not explicitly used in this annotation. By traversing the relationships, it can be shown that both CDS and exon are legal parts of the mRNA

Sequence Ontology, Table 1 The URLs of Sequence Ontology related resources

SO resource	URL
SO website	http://www.sequenceontolgy.org
Term tracker	https://sourceforge.net/tracker/?group_id=72703&atid=810408
Mailing list	https://sourceforge.net/mailarchive/forum.php?forum_name=song-devel
Wiki	http://www.sequenceontology.org/wiki/index.php/Main_Page
Latest revision	http://song.cvs.sourceforge.net/viewvc/song/ontology/so.obo
OWL version	http://www.berkeleybop.org/ontologies/owl/SO
GFF3	http://www.sequenceontology.org/resources/gff3.html
GVF	http://www.sequenceontology.org/resources/gvf.html
SOBA	http://www.sequenceontology.org/cgi-bin/soba.cgi
GMOD	http://gmod.org

```
##gff-version 3
scaffold_238  x_trop  contig  1       1689823 .  .  .  ID=scaffold_238;Name=scaffold_238
scaffold_238  x_trop  gene    634698  644312  .  +  .  ID=FGF8;Name=FGF8
scaffold_238  x_trop  mRNA    634698  644312  .  +  .  ID=FGF8a;Parent=FGF8;Name=FGF8a
scaffold_238  x_trop  mRNA    634698  644312  .  +  .  ID=FGF8b;Parent=FGF8;Name=FGF8b
scaffold_238  x_trop  exon    634698  635597  .  +  .  ID=FGF8a:exon0;Parent=FGF8a,FGF8b
scaffold_238  x_trop  exon    635685  635721  .  +  .  ID=FGF8a:exon1;Parent=FGF8a,FGF8b
scaffold_238  x_trop  exon    637433  637580  .  +  .  ID=FGF8a:exon2;Parent=FGF8a
scaffold_238  x_trop  exon    641297  641403  .  +  .  ID=FGF8a:exon3;Parent=FGF8a,FGF8b
scaffold_238  x_trop  exon    643179  644312  .  +  .  ID=FGF8a:exon4;Parent=FGF8a,FGF8b
scaffold_238  x_trop  exon    637400  637580  .  +  .  ID=FGF8b:exon5;Parent=FGF8b
scaffold_238  x_trop  CDS     635566  635597  .  +  0  ID=FGF8a:cds;Parent=FGF8a
scaffold_238  x_trop  CDS     635685  635721  .  +  1  ID=FGF8a:cds;Parent=FGF8a
scaffold_238  x_trop  CDS     637433  637580  .  +  0  ID=FGF8a:cds;Parent=FGF8a
scaffold_238  x_trop  CDS     641297  641403  .  +  2  ID=FGF8a:cds;Parent=FGF8a
scaffold_238  x_trop  CDS     643179  643457  .  +  0  ID=FGF8a:cds;Parent=FGF8a
scaffold_238  x_trop  CDS     635566  635597  .  +  0  ID=FGF8b:cds;Parent=FGF8b
scaffold_238  x_trop  CDS     635685  635721  .  +  1  ID=FGF8b:cds;Parent=FGF8b
scaffold_238  x_trop  CDS     637400  637580  .  +  0  ID=FGF8b:cds;Parent=FGF8b
scaffold_238  x_trop  CDS     641297  641403  .  +  2  ID=FGF8b:cds;Parent=FGF8b
scaffold_238  x_trop  CDS     643179  643457  .  +  0  ID=FGF8b:cds;Parent=FGF8b
```

Sequence Ontology, Fig. 3 The structure of both the transcripts of FGF8 (FGF8a and FGF8b) are defined in Generic Feature Format 3. The feature type is defined using the Sequence Ontology terms in column 3 (in this example: gene, mRNA, exon, and CDS). The parent of a feature is defined in column 9 using the "Parent=" tag. The relationship among features must always be a valid "part_of" relationship as defined by the Sequence Ontology

converted to Ontology Web Language (OWL). These and other resources are documented in Table 1.

The SO is used to type genomic features in various file formats and databases. The GFF3 file format is commonly used for model organism structural genome annotation. This is a nine-column simple tab delimited format that specifies the features on a genomic sequence. As can be seen in Fig. 3, the SO is used to type the kind of feature being annotated. The relationships between features are captured in column 9 using the "Parent" tag value pair. The Generic Model Organism Database (GMOD) community has provided a comprehensive database schema that uses SO for the features. These annotations are the input of other GMOD tools such as GBrowse and Apollo. For re-sequencing projects and variant annotation, there is Genome Variation Format (GVF), which uses terms in the SO to type the kind of alteration observed, the feature that is altered and the effect of the alteration. This format is built upon the existing GFF3.

SOBA is a GMOD tool that provides statistical and ontological analysis of a genomic annotation. This tool is particularly useful for emerging genome projects that are beginning to generate automated genomic annotations. SOBA provides fast feedback about the features in the annotation such as genomic footprint and average length.

SO and Systems Biology

Structural genomic annotations are important in the context of systems biology as they provide the genomic foundation upon which more knowledge is added. To demonstrate, Fig. 1 shows two annotated splice forms of the gene FGF8 in *Xenopus tropicalis* (FGF8a and FGF8b). The functional products of these transcripts are responsible for different developmental processes and do not share the same function, therefore, knocking out one splice form leaves the other's function intact. FGF8a promotes posterior neural fate, and FGF8b affects early mesodermal development. The Sequence Ontology describes and separates the two transcript features allowing function, pathway, and location of expression to be annotated independently to the relevant transcript sequence. This is important for any systems biology analysis as the transcript functions are mutually exclusive, so treating the gene as a single entity would lead to false conclusions.

Cross-References

► Gene Ontology

References

Day-Richter J, Harris MA, Haendel M, Lewis S (2007) OBO-Edit–an ontology editor for biologists. Bioinformatics 23:2198–2200

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M (2005) The sequence ontology: a tool for the unification of genome annotations. Genome Biol 6:R44

Lewis SE, Searle SM, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby MA, Kaminker JS, Matthews BB, Prochnik SE, Smithy CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME (2002) Apollo: a sequence annotation editor. Genome Biology 3(12): RESEARCH0082

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12:1599–1610

Sequentially Rejective Bonferroni Test

► Holm's Method

Set3 Complex

► Set3C

Set3C

Tetsuro Kokubo
Department of Supramolecular Biology, Graduate School of Nanobioscience, Yokohama City University, Yokohama, Kanagawa, Japan

Synonyms

Set3 complex

Definition

The functions of Rpd3S and Rpd3L, two closely related yeast HDAC complexes, are described in the essay "Mechanisms of Transcriptional Activation and Repression." In yeast, there is another distinctive HDAC complex (Set3C) that contains two HDACs, Hos2p (class I) and Hst1p (class III), in addition to Set3p, Sif2p, Snt1p, YIL112w, and Cpr1p (Pijnappel et al. 2001). Phosphorylation of RNAPII CTD helps recruit Set1p-containing methyltransferase complex (Set1C/COMPASS) to the 5′-end of the coding region of active genes. Set1C generates dimethylated lysine 4 of histone H3 (H3K4me2), which then recruits Set3C to deacetylate histone H3 and H4. A recently proposed model predicts that like Rpd3S, Set3C is recruited co-transcriptionally by the phosphorylated CTD of RNAPII and subsequently activated upon binding to H3K4me2 (Govind et al. 2001). Set1C also catalyzes trimethylation of lysine 4 of histone H3 (H3K4me3) in regions of further upstream than those containing H3K4me2. Regions containing H3K4me3 are highly acetylated and occupied by fewer nucleosomes.

By contrast, the neighboring H3K4me2-rich regions just downstream are less acetylated due to the function of Set3C and are therefore occupied by more tightly associated nucleosomes. Set3C and Rpd3S appear to antagonize the positive function of HATs, e.g., SAGA and NuA4, which are also co-transcriptionally recruited to coding regions by the phosphorylated CTD of RNAPII. However, the details of the functional interactions between HDACs and HATs during transcriptional elongation still remain unclear.

Cross-References

▶ Mechanisms of Transcriptional Activation and Repression

References

Govind CK, Qiu H, Ginsburg DS, Ruan C, Hofmeyer K, Hu C, Swaminathan V, Workman JL, Li B, Hinnebusch AG (2010) Phosphorylated Pol II CTD recruits multiple HDACs, including Rpd3C(S), for methylation-dependent deacetylation of ORF nucleosomes. Mol Cell 39(2):234–46

Pijnappel WW, Schaft D, Roguev A, Shevchenko A, Tekotte H, Wilm M, Rigaut G, Séraphin B, Aasland R, Stewart AF (2001) The *S. cerevisiae* SET3 complex includes two histone deacetylases, Hos2 and Hst1, and is a meiotic-specific repressor of the sporulation gene program. Genes Dev 15(22):2991–3004

Sexual Selection

Philippe Huneman
Institut d’Histoire et de Philosophie (IHPST), des Sciences et des Techniques, Université Paris 1 Panthéon-Sorbonne, Paris, France

Definition

Selection for traits which impinge on chances of reproduction by giving an advantage either in the competition for mates or regarding the female choice. One distinguishes sexual selection from natural selection, even if they may not be two really distinct processes because the sexually selected traits as such may *prima facie* go against selection (e.g., the feathers of the peacock which increase its chances to get parasites and prevent him from running fast, etc.). This divergence has been accounted for by two main theories, the *runaway selection*, defended by Fisher (1915), which states that a slight preference of females for an arbitrary trait will push its values to a limit which can often differ from its optimal adaptive value; and the *handicap principle*, elaborated by Zahavi since the 1980s, which states that sexual characters are a costly signal of the genetic capacity of their bearer to face environmental demands. Their being costly makes them honest since they are too costly to fake. So females “have interest” to pick those males because they reliably signal that they have “good” genes. According to the handicap principle, peacocks develop long tails precisely because they display their high ability to run even though they carry such handicap, and also that they have few parasites because those are extremely visible on the symmetrical motives decorating their tails. Both runaway selection and handicap principle have been mathematically modeled.

Cross-References

▶ Explanation, Evolutionary

References

Fisher RA (1915) The evolution of sexual preference. Eugen Rev 7:184–192

Grafen A (1990) Biological signals as handicaps. J Theor Biol 144:517–546

Zahavi A, Zahavi A (1997) The handicap principle: a missing piece of Darwin’s puzzle. Oxford University Press, Oxford

Shannon Information

▶ Information

Shared Resources

▶ Translational Research, Fundamental Infrastructures

Shielding of Activation Domain

▶ Activation Domain Shielding

Short Hairpin RNA

Melissa L. Kemp
The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Definition

Short hairpin RNA is an RNA sequence with a hairpin structure that is expressed in a cell using a genetic vector carrying the sequence. Short (or *small*) hairpin RNA (shRNA) is cleaved by the Dicer complex to produce siRNA which can mediate mRNA degradation. ShRNA can be used to induce siRNA against specific target mRNAs.

Short ORF (sORF)

▶ UORF-mediated Translational Control in Eukaryotes

Short Tandem Repeats (STRs)

▶ Microsatellite Repeats

Shotgun Proteomics

▶ Protein Identification Analysis

Side-scattered Light (SSC)

Xiaojun Liu
Internal Medicine, The Second Hospital of Hebei Medical University, Shijiazhuang, Hebei, China

Definition

SSC is a special parameter of flow cytometry which can reflect the physical properties of a particle or cell examined by the flow cytometry. It is a part of deflected laser light by the specimen. The extent to which light scatters depends on the physical properties of a particle, namely, its size and internal complexity. Factors that affect light scattering are the cell's membrane, nucleus, and any granular material inside the cell. Cell shape and surface topography also contribute to the total light scattering.

Cross-References

▶ Cell Sorting

Sigma Cascade

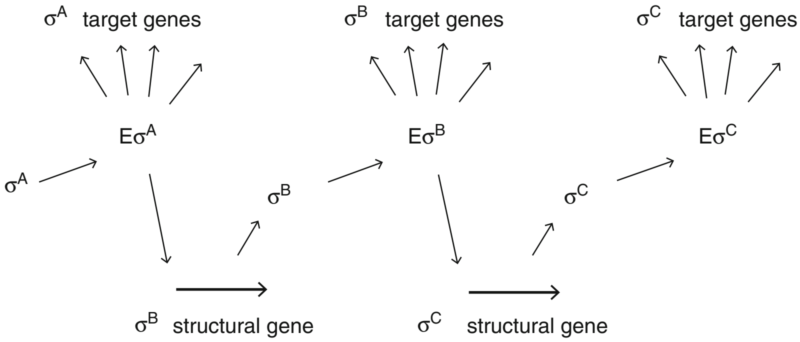
Kan Tanaka
Chemical Resources Laboratory, Tokyo Institute of Technology, Yokohama, Kanagawa, Japan

Synonyms

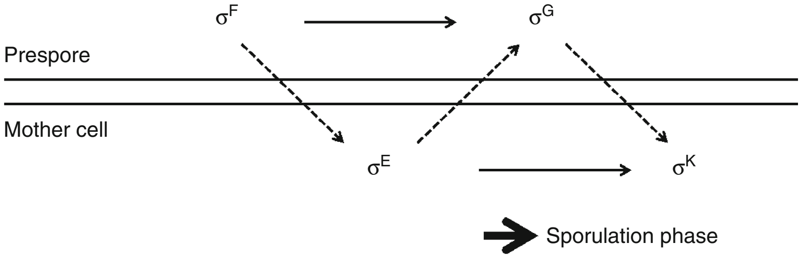
Bacterial transcriptional cascade

Definition

Sigma factor is a subunit of bacterial RNA polymerase essential for transcriptional initiation, which also determines the promoter recognition specificity of the RNA polymerase. Most bacteria possess multiple variants of sigma factors having distinct promoter



Sigma Cascade, Fig. 1 General scheme of sigma cascade. *E* indicates RNA polymerase core complex, and *Eσ* indicates RNA polymerase holoenzyme having ability for specific transcriptional initiation



Sigma Cascade, Fig. 2 Sigma cascade during *Bacillus* spore formation. *Arrows* indicate intracellular transcriptional activation, and *dotted arrows* indicate intercellular post-translational activation

specificities, which we call sigma heterogeneity, and thus bacterial cells are able to modify the RNA polymerase specificity in response to changing environmental or physiological situations by exchanging the sigma factor. Sequential transcriptional activation of a series of gene sets is frequently observed during various developmental processes both in prokaryotes and eukaryotes. Sigma cascade refers to a conceptual model that explains sequential transcriptional activation in bacteria: temporal expression or activation of one sigma factor results in the activation of another sigma factor, which results in activation of a set of genes in the next phase like a cascade (Fig. 1).

Characteristics

Sigma Heterogeneity

Purification and biochemical analyses of RNA polymerase from bacteria identified a general subunit composition of $\alpha_2\beta\beta'\sigma$, where the sigma (σ) subunit was defined as promoter recognition and initiation factor for transcription. While sigma factor was considered of unique molecular species at first, alternative sigma factors were later identified from most bacteria, and thus heterogeneity of sigma factor is a common characteristic of bacteria. However, the originally found sigma factor is still responsible for the major part of cellular transcription, and is often called the major or principal sigma factor. It is known that two hexamer sequences centered at -10 and -35 positions from the transcriptional start site are directly recognized by the cognate sigma factor, which indicates that substitution of sigma factor can change the promoter sequence specificity of the RNA polymerase. As a nomenclature for multiple sigma factor species, sigma factors are usually described with superscript for the molecular weight in kilodaltons as σ^{70} and σ^{38}, or just for respective names as σ^{A} and σ^{gp28}.

Bacteriophage Developments

Bacteriophages (or phages) are viruses that infect and propagate in bacteria cells as the host. After the phage infection, phage genes are orderly transcribed as conveniently divided into temporal classes, such as early, middle, and late classes. In *Bacillus subtilis* lytic phage SPO1 for example, early transcription begins immediately after the infection, and the transcription depends on the host RNA polymerase containing the principal sigma factor σ^{A}. After 5 min, phage genes of the middle class are turned on by the action of an early gene product protein, gp28 (gene product 28). The function of gp28 is now known as an alternative sigma factor (σ^{gp28}) that substitutes for the principal sigma factor σ^A, and enables the RNA polymerase to recognize the middle promoter classes. The middle class genes include genes 33 and 34, and these gene products (gp33 and gp34) synergistically direct transcription from the late class gene promoters as another sigma factor, $\sigma^{gp33\text{-}34}$. Thus, the sequential switching of promoter classes is well explained by a simple cascade of phage-encoded sigma factors. In addition, a simpler type of sigma factor cascade was found in *Escherichia coli* phage T4 development, where the middle gene product gp55 is transcribed by the host sigma factor and functions as an alternative sigma factor to activate the transcription of late class promoters.

Sporulation of *Bacillus subtilis*

B. subtilis is a gram-positive bacteria that differentiates a highly stress-resistant dormant spore. This differentiation is induced by nutritional limitation, and usually takes several hours for completion. During the vegetative growing phase, two sigma factors σ^A and σ^H are dominant in the cell. However, after onset of sporulation, at least five temporally defined classes of sporulation-specific gene expression occur, and a sigma cascade of $\sigma^H \rightarrow \sigma^F \rightarrow \sigma^E \rightarrow \sigma^G \rightarrow \sigma^K$ has been shown responsible as the underlying mechanism. It is of note that, in contrast to the sigma cascade in phage developments, sporulation is a differentiation process performed by two cells, prespore and mother cell. Thus, the overall sigma cascade is a composite of two sigma cascades, $\sigma^H \rightarrow \sigma^F \rightarrow \sigma^G$ and $\sigma^H \rightarrow \sigma^F \rightarrow \sigma^E \rightarrow \sigma^K$, tightly coupled by cell-cell communication across the boundary of the two cells (Fig. 2). For the sequential sigma factor activation, a number of regulatory mechanisms, such as transcriptional activation, binding of anti-σ factor, and even site-specific DNA recombination, have been found as the underlying mechanisms.

Examples in Other Bacteria

Sigma factor cascade has also been found in other bacterial systems, and thus the mechanism is likely general among the bacteria kingdom.

Cross-References

- ▶ Mechanisms of Transcriptional Activation and Repression
- ▶ Operon Theory
- ▶ Sigma Factor
- ▶ Transcription Factors and Transcriptional Apparatus in Bacteria
- ▶ Transcription in Bacteria

References

Fang FC (2005) Sigma cascades in prokaryotic regulatory networks. Proc Natl Acad Sci USA 102(14):4933–4934

Geiduschek EP, Kassavetis GA (2010) Transcription of T4 late genes. Virol J 7:288

Helmann JD, Chamberlin MJ (1988) Structure and function of bacterial sigma factors. Annu Rev Biochem 57:839–872

Hilbert DW, Piggot PJ (2004) Compartmentalization of gene expression during *Bacillus subtilis* spore formation. Microbiol Mol Biol Rev 68(2):234–262

Homerova D, Halgasova KJ (2008) Cascade of extracytoplasmic function sigma factors in *Mycobacterium tuberculosis*: identification of a σ^J-dependent promoter upstream of sigI. FEMS Microbiol Lett 280:120–126

Lonetto M, Gribskov M, Gross CA (1992) The sigma 70 family: sequence conservation and evolutionary relationships. J Bacteriol 174(12):3843–3849

Losick R, Pero J (1981) Cascades of sigma factors. Cell 25: 582–584

Mazurakova V, Sevcikova B, Rezuchova B, Kormanec J (2006) Cascade of sigma factors in Streptomycetes: identification of a new extracytoplasmic function sigma factor σ^J that is under the control of the stress-response sigma factor σ^H in *Streptomyces coelicolor* A3(2). Arch Microbiol 186:435–446

Stragier P, Losick R (1990) Cascades of sigma factors revisited. Mol Microbiol 4(11):1801–1806

Sigma Factor

Nobuo Shimamoto
Faculty of Life Sciences, Kyoto Sangyo University,
Kyoto, Japan

Definition

Sigma Factor is the initiation factor of bacterial transcription. A sigma factor is a protein subunit of bacterial RNA polymerase holoenzyme. The enzyme lacking it is called core enzyme, which is the form of RNA polymerase in elongation. Since it has an affinity for the non-template strand of its specific promoter, the binding of RNA polymerase to a promoter results in a formation of ▶ open complex. A bacterium has two or more sigma factors. In *Escherichia coli*, RpoD is the major sigma, and six minor sigma factors have been identified, RpoE, RpoF, RpoH, RpoN, RpoS, and FecI. RpoE holoenzyme transcribes the genes for preventing damages caused at an extremely high temperature, RpoF is responsible for the expression of flagella genes, and so on. Some sigma factors, such as *E. coli* RpoN, bind to a promoter and then the core enzyme binds to the factor complexed with DNA, similar to the way eukaryotic transcription is initiated.

Cross-References

▶ Transcription in Bacteria

Signal Recognition Particle

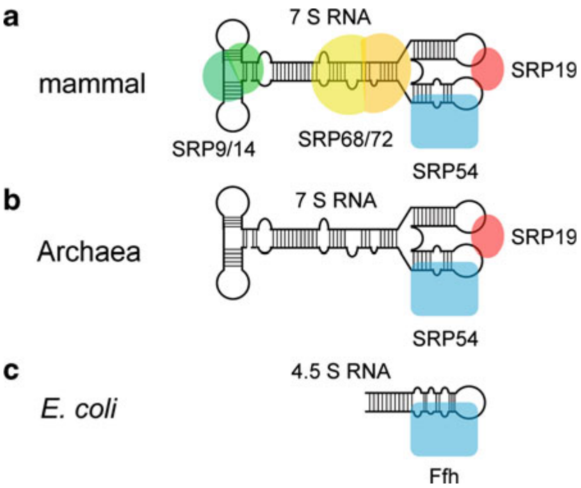
Taiichi Sakamoto[1] and Gota Kawai[2]
[1]Chiba Institute of Technology, Narashino, Japan
[2]Department of Life and Environmental Sciences, Chiba Institute of Technology, Narashino, Chiba, Japan

Synonyms

SRP

Definition

The signal recognition particle (SRP, Fig. 1) is an abundant cytosolic ribonucleoprotein that targets membrane proteins and secretory proteins to the endoplasmic reticulum (ER) in eukaryotes and the plasma membrane in prokaryotes (Alberts et al. 2008; Nagai et al. 2003). Although SRP is universally conserved in all organisms, its composition is varied. The mammalian SRP is composed of six distinct proteins (SRP9, SRP14, SRP19, SRP54, SRP68, and SRP72) bound to 7S RNA (about 300 nucleotides) and consists of two functional domains; the *Alu* and S domains. The SRP9-SRP14 heterodimer binds to one end of 7S RNA, forming the *Alu* domain, whereas SRP19, SRP54, SRP68, SRP72, and the remaining region of 7S RNA form the S domain. The SRP RNA of Archaea is similar in size and secondary structure to its mammalian counterpart, but homologues of only two mammalian SRP proteins, SRP54 and SRP19, have been identified in archaeal genomes. SRP of *Escherichia coli* is composed of 4.5S RNA (114 nucleotides) and one protein Ffh (Fifty-four homolog) (see ▶ Translational Control by Small RNAs, Bacteria).



Signal Recognition Particle, Fig. 1 Signal recognition particle of (**a**) mammal, (**b**) Archaea, and (**c**) *E. coli*

In eukaryotes, SRP recognizes the secretory signal or membrane-anchor sequences upon their emergence from the ribosomal exit tunnel. This interaction leads to the delay of protein synthesis known as “elongation arrest.” Translation resumes when the SRP-bound ribosome nascent chain complex (RNC) binds to the translocon in the ER membrane. This binding occurs via the interaction of SRP with its cognate SRP receptor that is located in close proximity to the translocon.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2008) Molecular biology of the cell, 5th edn. Garland Science, New York

Nagai K, Oubridge C, Kuglstatter A, Menichelli E, Isel C, Jovine L (2003) Structure, function and evolution of the signal recognition particle. EMBO J 22:3479–3485

Signal Transduction

▶ Signal Transduction Pathway

Signal Transduction Pathway

Yufei Huang
Picower Institute for Learning and Memory,
Massachusetts Institute of Technology, Cambridge,
MA, USA
Greehey Children's Cancer Research Institute,
University of Texas at Texas Health Science Center at
San Antonio, San Antonio, TX, USA
Department of Epidemiology and Biostatistics,
University of Texas at Texas Health Science Center at
San Antonio, San Antonio, TX, USA

Synonyms

Signal transduction; Signaling pathway

Definition

Signal transduction pathway refers to a set of chemical reactions in a cell that converts a mechanical/chemical stimulus into a specific cellular response (Campbell et al. 2003). These chemical reactions can take as short as milliseconds (for ion flux), or as long as days (for gene expression).

Cross-References

▶ Gene Regulation

Signaling Crosstalk

▶ Pathway Crosstalk

Signaling Network Resources

Annette A. Alcasabas
BioSyntha Technology, Welwyn Garden City, UK

Definition

Signaling network resources are online databases with information on molecules that participate in signal transduction pathways. These pathways can involve ligands, receptors, kinases, and other enzymes that modify both upstream and downstream molecules and transcription factors.

Characteristics

In terms of resources, there is overlap with databases that also include metabolic and transcriptional pathways. General resources such as the KEGG Pathway Database (http://www.genome.jp/kegg/pathway.html) contain information on signal transduction as well as other cellular pathways. However, there are also other resources that specialize on signal transduction and even those that further specialize on particular kinds of signaling.

Cell Signaling Resources

The Signaling Gateway (http://www.signaling-gateway.org; Saunders et al. 2007), hosted by UC San Diego in collaboration with other companies, is a free online database for signal transduction proteins. Its main feature is the Molecule Pages, a relational database that holds information on over 4,000 cell signaling proteins. Each protein is given a long description containing its protein family, its activity and transition states, its upstream and downstream interactors, and a graphical representation of the pathway it belongs to. There is highly structured data that can be interrogated using bioinformatics tools. The data is edited by invited experts, who are acknowledged in each molecule page, and peer-reviewed by the Nature Publishing Group. At the same time, peer-reviewed data is complemented regularly by information collected automatically from public data sources, sequence analyses, and other databases. The goal of the Molecule Pages is to have validated information that would be readily useful for the modeling of interactions and pathways. This resource also features Signaling Update, a weekly digest of the latest research on cell signaling.

Science Magazine's Database of Cell Signaling (http://stke.sciencemag.org/cm/) provides graphical representations of signaling pathways, called Connection Maps. A user can view both "canonical pathways" which are generalized representations of conserved pathways or "specific pathways" which could be specific signaling cascade in a particular organism. In addition, there is a separate section of the database for Pathway-Independent Component information.

Each component can be a protein or nonprotein molecule that participates in one or more pathways. Currently, there is information for over 1,800 components, this includes a description of how it is formed, transition states, the canonical and specific pathways it is part of, literature references and links to other databases. Information on both the Connection Maps and Pathway-Independent Components sections are entered by experts, called Pathway Authorities, their names appear on the pages they created. Access to this database is either by institutional license or by free registration.

Phosphorylation Network Resources

Many signals transduction pathways are comprised of one or more that modify the activity and/or localization of both upstream and downstream components by phosphorylation. In fact, at least half of all proteins in a eukaryotic cell undergo phosphorylation. Databases containing kinases, phosphatases, and phosphorylation have therefore emerged to assist signaling pathway studies.

Currently, the most extensive study of a eukaryote's kinome (kinase/phosphatase interactome) belongs to budding yeast *Saccharomyces cerevisiae*. Using mass-spectrometric analysis of yeast kinase and phosphatase complexes, Breitkreutz and colleagues (2010) identified over 1,844 interactions. This led to the creation of Yeast Kinome database (http://www.yeastkinome.org), which lists the interacting molecules of the organism's 130 protein kinases, 24 lipid and metabolic kinases, 38 protein phosphatases, 5 metabolic phosphatases, and their associated proteins. A complementary resource, PhosphoGRID (http://www.phosphogrid.org/; Stark et al. 2010) lists the phosphorylation sites within yeast proteins based on experimentally verified data.

In contrast to the smaller numbers in budding yeasts, humans have over 500 kinases and over 100 phosphatases. Resources that hold information on kinases in humans and other organisms include Kinase.com (http://www.kinase.com/) and a resource website from Cell Signaling Technology (http://www.cellsignal.com/reference/kinase/index.html).

To predict phosphorylation sites in protein sequences from other eukaryotes, one can use Scansite (http://scansite.mit.edu/) and Netphos (http://www.cbs.dtu.dk/services/NetPhos/).

PhosphoSitePlus (http://www.phosphosite.org) manually curates not only phosphorylation, but other post-transcriptional modifications on proteins from human, mouse, and rat species.

References

Breitkreutz A et al (2010) A global protein kinase and phosphatase interaction network in yeast. Science 328:1043. doi:10.1126/science.1176495

Saunders B et al (2007) The molecule pages. Nucleic Acids Res. doi:10.1093/nar/gkm907

Stark JB et al (2010) PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast Saccharomyces cerevisiae. Database. doi:10.1093/database/bap026

Signaling Pathway

▶ Signal Transduction Pathway

Signal-to-Noise Ratio

Tianshou Zhou
School of Mathematics and Computational Sciences,
Sun Yet-Sen University, Guangzhou, Guangdong,
China

Definition

Signal-to-noise ratio (often abbreviated SNR or S/N) is a measure used in science and engineering that compares the level of a desired signal to the level of background noise. It is defined as the ratio of signal power to the noise power. A ratio higher than 1:1 indicates more signal than noise. While SNR is commonly quoted for electrical signals, it can be applied to any form of signal (such as isotope levels in an ice core or biochemical signaling between cells).

Signal-to-noise ratio is sometimes used informally to refer to the ratio of useful information to false or irrelevant data in a conversation or exchange. For example, in online discussion forums and other online communities, off-topic posts and spam are regarded as "noise" that interferes with the "signal" of appropriate discussion.

Signal-to-noise ratio is defined as the power ratio between a signal (meaningful information) and the background noise (unwanted signal):

$$\mathrm{SNR} = \frac{P_{signal}}{P_{noise}}$$

where P is average power. Both signal and noise power must be measured at the same or equivalent points in a system, and within the same system bandwidth. If the signal and the noise are measured across the same impedance, then the SNR can be obtained by calculating the square of the amplitude ratio:

$$\mathrm{SNR} = \frac{P_{signal}}{P_{noise}} = \left(\frac{A_{signal}}{A_{noise}}\right)^2$$

where A is root mean square (RMS) amplitude (e.g., RMS voltage). Because many signals have a very wide dynamic range, SNRs are often expressed using the logarithmic decibel scale. In decibels, the SNR is defined as

$$\mathrm{SNR}_{dB} = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) = P_{signal,dB} - P_{noise,dB}$$

which may equivalently be written using amplitude ratios as:

$$\mathrm{SNR}_{dB} = 10\log_{10}\left(\frac{A_{signal}}{A_{noise}}\right)^2 = 20\log_{10}\left(\frac{A_{signal}}{A_{noise}}\right)$$

The concepts of signal-to-noise ratio and dynamic range are closely related. Dynamic range measures the ratio between the strongest undistorted signal on a channel and the minimum discernable signal, which for most purposes is the noise level. SNR measures the ratio between an arbitrary signal level (not necessarily the most powerful signal possible) and noise. Measuring signal-to-noise ratios requires the selection of a representative or reference signal. In audio engineering, the reference signal is usually a sine wave at a standardized nominal or alignment level, such as 1 KHz at +4 dBu (1.228 V_{RMS}).

SNR is usually taken to indicate an average signal-to-noise ratio, as it is possible that (near) instantaneous signal-to-noise ratios will be considerably different. The concept can be understood as normalizing the noise level to 1 (0 dB) and measuring how far the signal "stands out."

Silent Chromatin and Active Chromatin

▶ Heterochromatin and Euchromatin

Silicon Cell

Franco du Preez
SysMO-DB team, Manchester Centre for Integrative Systems Biology, University of Manchester, Manchester, UK

Definition

The long-term goal of the Silicon Cell Initiative (http://www.siliconcell.net) is the computation of life at the cellular level using mathematical models based on the complete genomic, transcriptomic, proteomic, metabolomic, and cell-physiomic information (Westerhoff et al. 2003), which is becoming increasingly available with the advent of high-throughput biology. Silicon Cell models are based on the properties of the macromolecules comprising the cell and should be real and based on experimental determinations of those properties themselves. Such experiments typically correspond to in vitro enzyme kinetics, or to in vivo determinations of the kinetic properties of the individual macromolecules. This distinguishes a silicon cell model from many existing "phenomenological" models, wherein the parameters are fitted to replicate the behavior of the living cell.

A number of silicon cell models for parts of cellular networks have been published and are available from the ▶ JWS Online model repository, from where they can be simulated directly (http://jjj.mib.ac.uk). The aim is to link such models to ultimately model complete cellular networks (Snoep and Westerhoff 2004; Snoep 2005).

Cross-References

▶ JWS Online

References

Snoep JL (2005) The silicon cell initiative: working towards a detailed kinetic description at the cellular level. Curr Opin Biotechnol 16:336–343

Snoep JL, Westerhoff HV (2004) The silicon cell initiative. Curr Genomics 5:687–697

Westerhoff HV, Bruggeman F, Hofmeyr JH, Snoep JL (2003) Attractive models: how to make the silicon cell relevant and dynamic. Comp Funct Genomics 4(1):155–158

SIM

► Canonical Network Motifs

Simple Conditional Analysis (SCA)

Marie I. Kaiser and Andreas Hüttemann
Department of Philosophy, University of Cologne,
Cologne, Germany

Definition

Let Ds stand for system s having the disposition D, that is, s being disposed to M (manifestation) provided enabling conditions E obtain. According to the simple conditional analysis (SCA), the necessary and sufficient conditions for s having D can be symbolized as follows:

$$\mathrm{Ds} \leftrightarrow (\mathrm{Es} \rightarrow \mathrm{Ms})$$

which is to be read as: s has Disposition D if and only if: If s were confronted with E then s would manifest M (Choi and Fara 2012).

Cross-References

► Disposition

References

Choi S, Fara M (2009) Dispositions. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (Spring 2012 Edition), http://plato.stanford.edu/archives/spr2012/entries/dispositions/

Simple Object Access Protocol

Richard G. Côté
European Molecular Biology Laboratory, European
Bioinformatics Institute (EBI), Hinxton,
Cambridge, UK

Synonyms

SOAP

Definition

The Simple Object Access Protocol (SOAP) is a lightweight protocol intended for exchanging structured information in a decentralized, distributed environment.

It uses XML technologies to define an extensible messaging framework providing a message construct that can be exchanged over a variety of underlying protocols.

The framework has been designed to be independent of any particular programming model and other implementation-specific semantics.

Cross-References

► Ontology Lookup Service for Controlled Vocabularies and Data Annotation

References

SOAP Version 1.2 Part 1: Messaging framework (Second Edition), W3C Recommendation 27 Apr 2007. http://www.w3.org/TR/soap12-part1/

Simple Sequence Repeats (SSRs)

▶ Microsatellite Repeats

Simplification

▶ Reduction

Simulated Annealing

Feng-Sheng Wang and Li-Hsunan Chen
Department of Chemical Engineering, National Chung Cheng University, Chiayi, Taiwan

Definition

Simulated annealing (SA) is a heuristic algorithm, which mimics certain thermodynamic principles of producing an ideal crystal, in order to achieve a global optimal solution (Sharda et al. 2003; Zhilinskas and Žilinskas 2008). The algorithm is used for simulating thermal moves of molecules at a certain temperature. This temperature is the crucial parameter of SA that influences both reliability and efficiency of optimization. The annealing process in metallurgy can make particles arrange themselves in the position with minima potential as the temperature is slowly decreased. The algorithm starts with a solution and moves to one of its neighbors in each iteration, randomly. The probabilities for the uphill moves, which are the moves toward a worse neighbor, and the downhill moves, which are the moves toward a better neighbor, to succeed are different and the temperature, remaining iteration times, decreases with the uphill probability. Slowly decrease the temperature until the particle is trapped in the minimum potential, the minimum solution can be found.

Characteristics

Pseudocode:

```
$X = X_0$
$e = E(X)$
$X_{best} = X$; $e_{best} = e$
$T = T_0$
Do
    $X_{new}$ = random neighbor($X$)
    $e_{new} = E(X_{new})$
    if P($e$, $e_{new}$, $T$) > random() then
        // P($e$, $e_{new}$, $T$) is the probability for the particle to move from $X$ to $X_{new}$
        //where P($e$, $e_{new}$, $T$) > P($e$, $e_{new}$, $T'$) when $e < e_{new}$ and $T > T'$
        $X = X_{new}$; $e = e_{new}$
    If $e_{new} < e_{best}$ then
        $X_{best} = X_{new}$; $e_{best} = e_{new}$
    T = T-1
While the maximum iteration is not attained.
```

References

Sharda R, Voß S, Woodruff DL, Fink A (2003) Optimization software class libraries. Springer, Berlin, pp 91–94

Zhilinskas A, Žilinskas A (2008) Stochastic global optimization. Springer, New York, pp 103–124

Simulated Evolution

▶ Artificial Evolution

Simulation

▶ Lymphocyte Dynamics and Repertoires, Modeling

Single Cell Assay, Hematopoietic Stem Cell

Karl Staser[1] and Feng-Chun Yang[2]
[1]Departments of Pediatrics and Biochemistry, Indiana University School of Medicine, Indianapolis, IN, USA
[2]Departments of Pediatrics, Herman B Wells Center for Pediatric Research, School of Medicine, Indiana University, Indianapolis, IN, USA

Synonyms

Competitive repopulation; Single hematopoietic stem cell transplantation

Definition

The definitive single cell assay for a hematopoietic stem cell (HSC) is long-term multi-lineage reconstitution in transplanted primary and secondary hosts. In single cell experimental designs, putative hematopoietic stem cells are isolated using antibody-based detection methods. These methods include antibody-conjugated magnetic beads and multi-parameter fluorescence-activated cell sorting (FACS). Current methodologies and phenotypic knowledge permit the isolation of single cells each with a ~50% chance of engendering long-term multi-lineage hematopoiesis.

Characteristics

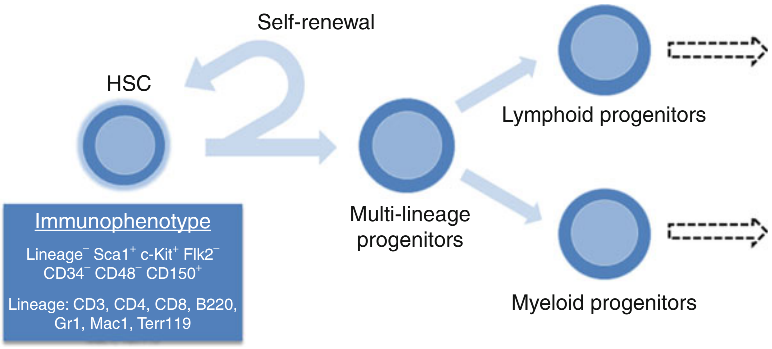
The Hematopoietic Stem Cell

A single hematopoietic stem cell can both self-renew and differentiate to generate all blood and immune effector cells (Czechowicz and Weissman 2010; Weissman and Shizuru 2008). A single transplanted hematopoietic stem cell can hypothetically restore the hematopoietic system of a radioablated or otherwise marrow-conditioned mouse for its lifetime. Similarly, HSCs from these transplanted mice can give rise to all blood and immune effector cells in secondary recipients, definitively demonstrating the principal of stem cell self-renewal and multi-lineage reconstitution (Fig. 1) (Spangrude et al. 1988).

The HSC Immunophenotype

Over the past 60 years, investigators have utilized both tissue culture- and transplantation-based assays to pinpoint the HSC phenotype with increasing accuracy. As shown in long-term transplantation studies, the HSC can now be detected, isolated, and transplanted with about 50% reliability on the single cell level (Kiel et al. 2005). The actual accuracy may exceed this number, as radioablation-based transplantation assays are subject to experimental error and engraftment failure at multiple levels, including unknown consequences of ex vivo HSC manipulation prior to intravenous injection.

Current methods of characterization and isolation rely principally upon the binding and detection of cell


Self-renewal
HSC
Immunophenotype
Lineage$^-$ Sca1$^+$ c-Kit$^+$ Flk2$^-$ CD34$^-$ CD48$^-$ CD150$^+$
Lineage: CD3, CD4, CD8, B220, Gr1, Mac1, Terr119
Multi-lineage progenitors
Lymphoid progenitors
Myeloid progenitors

Single Cell Assay, Hematopoietic Stem Cell, Fig. 1 Basic schema and immunophenotype of the HSC, showing the principal of self-renewal and differentiation (Adapted, with modifications, from (Weissman and Shizuru 2008)) surface proteins with antibody-conjugated magnetic beads and fluorophores. Accordingly, the phenotypic detection of HSCs relies on known expression patterns of certain antigens and, in complementary methods, dye uptake and efflux (e.g., Hoechst staining) (Darzynkiewicz et al. 2004; Weissman and Shizuru 2008).

Phenotypically, the HSC expresses the c-Kit receptor tyrosine kinase (c-Kit) and the stem cell antigen 1 (Sca1) while expressing low or undetectable levels of multiple mature hematopoietic cell proteins – known as the lineage negative/low (lin$^{-/lo}$) fraction (Czechowicz and Weissman 2010; Spangrude et al. 1988; Weissman and Shizuru 2008). These lineage markers include T-cell markers CD3, CD4, and CD8; B-cell marker B220; erythroid progenitor marker Ter119; and myeloid markers Gr-1 and Mac-1 (also known as Ly6C/G and CD11b, respectively). Using antibodies to these markers, the investigator can identify the LSK fraction – lin$^{-/lo}$, Sca1^{+}, and c-Kit^{+} – an enriched population containing a high frequency of hematopoietic stem and progenitor cells.

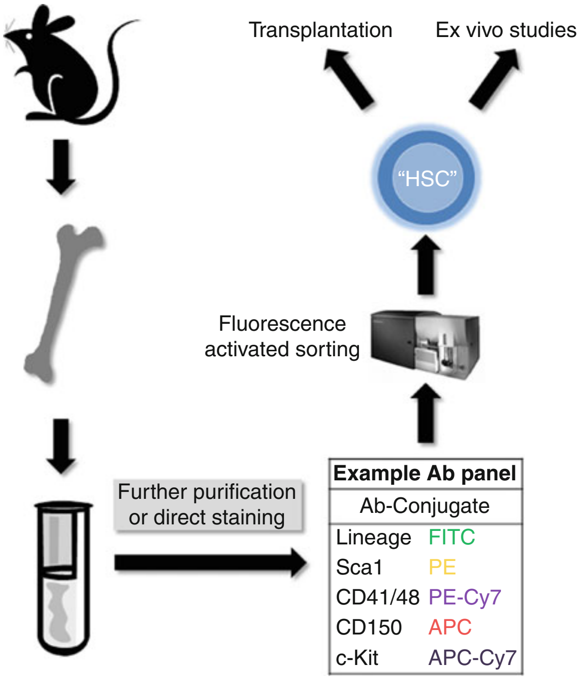
Further phenotypic refinement provides a greater purity and, thus, a greater chance of a single cell being an HSC. Long-term repopulating murine HSCs have been described as being negative for both Fetal liver kinase 2 (Flk-2) and CD34 (Christensen and Weissman 2001; Osawa et al. 1996). Moreover, microarray transcriptional profiling of HSC-enriched populations have helped to reveal novel and potentially simplifying identifier proteins. Specifically, the expression pattern of the signaling lymphocyte activation molecule (SLAM) family proteins, CD150, CD48, and CD41, may distinguish HSCs (Kiel et al. 2005). In fact, CD150^{+} CD41/48^{-} LSK (SLAM-LSK) cells demonstrate approximately a 50% chance of long-term multi-lineage hematopoietic reconstitution when transplanted as single cells. Moreover, in rigorous detection and isolation protocols, some investigators have been able to achieve high levels of HSC purity using *only* the SLAM markers.

Here, we will use the SLAM-LSK cell as the prototype for single HSC isolation and assay. However, other detection methods exist, including the use of CD34, Flk-2, and Hoechst staining, and these phenotypical markers may identify "true" yet heterogeneous populations of HSCs and/or overlap in expression within the putative HSC population. (For example Kiel et al. (2005) found that nearly all CD150^{+} CD41/48^{-} cells also expressed Sca1 and c-Kit.) Importantly, independent research groups must develop and validate a detection and isolation technique before initiating large studies, as many variables affect purity and yield. A detailed discussion of modern flow cytometry-based methods of HSC immunophenotyping can be found in (Challen et al. 2009).

HSC Isolation

Bone marrow contains the greatest numbers of HSCs and, in mice, is the readiest HSC source. Marrow isolated from the long bones (femur, tibia, iliac) of a wild-type adult male mouse yields about 50–100 million mononuclear cells, of which about 0.005–0.01% are SLAM-LSK cells (as derived from (Kiel et al. 2005)). Thus, 2,500–10,000 SLAM-LSK cells can be isolated from the leg bone marrow of the adult male mouse, a yield varying with the strain of the mouse, the dissection/purification techniques employed, and, of course, any pre-dissection interventions performed.

In a typical experiment, the investigator first isolates whole bone marrow from the long bones either by flushing the shaft of the bone with fluid and a needle or by crushing the whole bone, enzymatically processing the tissue, and purifying mononuclear cells by density gradient. After mononuclear cell isolation, the investigator may choose to further purify the cell population, such as by using anti-c-Kit antibody-conjugated magnetic beads, which enriches the HSC frequency approximately 20-fold compared to non-purified mononuclear cells. Next, the investigator adds distinct antibody-conjugated fluorophores. These antibodies detect different cell surface proteins which identify the putative HSC. Fluorophore-marked cells are analyzed and sorted using an appropriate flow cytometer (e.g., Becton Dickinson FACS Aria or iCyt Reflection). These machines are capable of plating a single cell into each well of a 96-well plate. In the example given here, the investigator would use flow cytometry software to identify the lin$^{-/lo}$, Sca1^{+}, c-Kit^{+}, CD150^{+}, and CD41/48^{-} fraction, which the machine could then sort as a single cell into the well of a 96-well plate (or other vessel). Figure 2 demonstrates a general schema for this procedure. These single cells may then be assayed in transplantation or ex vivo studies.


Transplantation
Ex vivo studies
"HSC"
Fluorescence activated sorting
Further purification or direct staining

Example Ab panel	
Ab-Conjugate	
Lineage	FITC
Sca1	PE
CD41/48	PE-Cy7
CD150	APC
c-Kit	APC-Cy7

Single Cell Assay, Hematopoietic Stem Cell, Fig. 2 Overview of HSC detection and isolation. Bone marrow is isolated, purified, and labeled with anti-mouse antibody-conjugated fluorophores. An example five-color antibody panel demonstrates a detection technique for SLAM-LSK cells (lin$^{-/lo}$, Sca1^{+}, c-Kit^{+}, CD150^{+}, and CD41/48^{-}). Using an appropriately equipped flow cytometer, single cells can be sorted into a 96-well plate and subsequently transplanted or studied in cell culture

Single HSC Transplantation

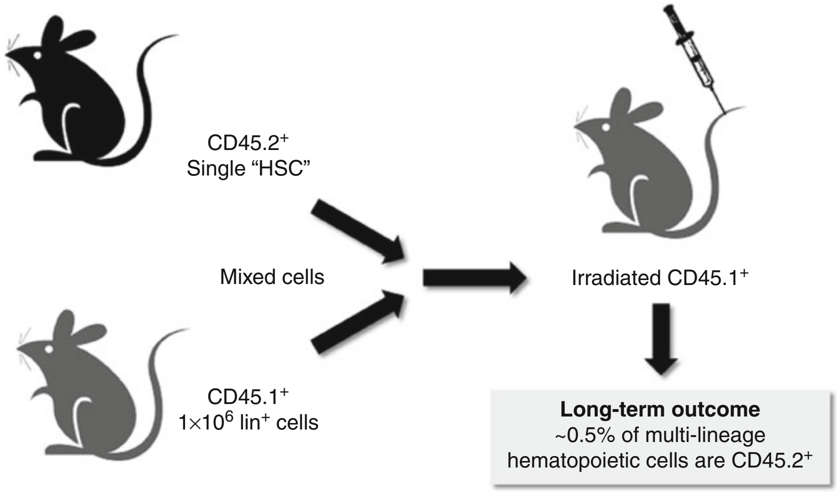
Singly sorted, putative HSCs can be transplanted into radioablated or otherwise conditioned mice. Essentially, the single cell transplant assesses the putative HSC's potential to find the bone marrow niche, engraft, and restore long-term multi-lineage hematopoiesis. Thus, the investigator may have various intentions with the single cell assay, from improving and validating the phenotypic definition of an HSC to testing the hematopoietic potential of a wild-type versus a genetically modified cell.

Cells to be transplanted are suspended in saline and then delivered by intravenous injection. Transplantation validation requires that the donor cells possess a distinguishing feature from the recipient/host hematopoietic cells. In many experimental designs, investigators use congenic mouse strains. These strains differ in the hematopoietic expression of a single antigenic epitope, easily distinguished on flow cytometry with antibody-conjugated fluorophores. For example, an investigator may isolate an HSC from a C57BL/6 strain mouse, whose mononuclear cells all express the CD45.2 surface protein, and transplant this HSC into a radioablated BoyJ strain mouse, whose mononuclear cells all express the CD45.1 surface protein. After transplantation, the investigator can isolate mononuclear cells from peripheral blood and stain them with a mixture of CD45.1, CD45.2, and other fluorophore-conjugated antibodies. In this way, the investigator can use flow cytometry to easily and precisely identify donor (CD45.2^{+}) versus host (CD45.1^{+}) hematopoietic cells.

Because the radioablated mouse lacks a healthy hematopoietic system, the mouse requires a co-injection of non-ablated mononuclear cells to support its immune functions. In the competitive repopulation design, the single, putative HSC (e.g., CD45.2^{+} SLAM-LSK cell) is transplanted along with a varying number of a host-type (e.g., CD45.1^{+}) non-HSC fraction (e.g., CD150^{-} or lin^{+}). While these cells will help support the mouse's immunity following radioablation and transplantation, the cells should not engender long-term hematopoiesis. Thus, if the single isolated cell is truly an HSC, a fraction (~0.5%+) of the circulating lymphoid and myeloid cells in the transplanted mouse should be CD45.2^{+} (in this example) 4–6 months following transplantation. This finding would indicate that the single CD45.2^{+} cell has given rise to many (~0.5%+ of billions) multi-lineage (myeloid/lymphoid) for a long period of time (4–6 months), fulfilling most criteria of an HSC. The remaining hematopoiesis derives from host stem cells escaping lethal radioablation or from contaminating stem cells contained within the putatively stem cell-depleted competitor fraction. Figure 3 presents this competitive repopulation schema. (See Kiel et al. (2005) or Christensen and Weissman (2001) for example experimental details.)

Secondary Transplantation

After 4–6 months of hematopoietic reconstitution in the primary recipient, definitive validation of the self-renewing HSC can be affirmed by secondary transplantation. In a basic experimental design, ~2 × 10^{6} mononuclear cells are isolated from the chimeric primary recipient and transplanted without purification into a radioablated host-type secondary recipient (e.g., CD45.1^{+}). After 4 months, circulating lymphoid and myeloid cell chimerism should roughly reflect the chimerism of the original host (e.g., ~0.5–10% CD45.2^{+}),



Single Cell Assay, Hematopoietic Stem Cell, Fig. 3 [Competitive repopulation]. One putative HSC expressing CD45.2 is mixed with one million CD45.1$^+$ non-HSCs (support cells) and injected intravenously into a radioablated mouse with a CD45.1$^+$ hematopoietic system. After 4–6 months, ~0.5%+ of the circulating lymphoid and myeloid cells should be CD45.2$^+$ (if the single cell was, indeed, an HSC), indicating long-term, multi-lineage hematopoietic reconstitution. Lineage reconstitution can be assessed by flow cytometric detection of the simultaneous expression of CD45.2 and lymphoid/myeloid markers such as CD3,4,8; B220; Gr-1; Mac-1; and Ter119

indicating successful engraftment, self-renewal, and multi-lineage commitment of the singly isolated and transplanted HSC.

Ex Vivo Studies

As an alternative to transplantation, the investigator may wish to perform any number of studies with single or low numbers of HSCs in culture. For example, sorted HSCs can be stimulated with fetal bovine serum and stem cell factor to induce proliferation, allowing the investigator to test different conditions on HSC physiology. However, HSCs inevitably lose their "stemness," differentiating away from a cell capable of in vivo, multi-lineage, long-term hematopoietic reconstitution. Thus, several research groups are currently investigating various strategies (e.g., gene delivery, chemical inhibition, culture methods) to support long-term HSC self-renewal ex vivo.

Cross-References

- ▶ Competitive Repopulation
- ▶ Fluorescence-activated Cell Sorting
- ▶ Single Cell Assay, Hematopoietic Stem Cell
- ▶ Self-Renewal

References

Challen GA, Boles N, Lin KK, Goodell MA (2009) Mouse hematopoietic stem cell identification and analysis. Cytom A 75:14–24

Christensen JL, Weissman IL (2001) Flk-2 is a marker in hematopoietic stem cell differentiation: a simple method to isolate long-term stem cells. Proc Natl Acad Sci USA 98:14541–14546

Czechowicz A, Weissman IL (2010) Purified hematopoietic stem cell transplantation: the next generation of blood and immune replacement. Immunol Allergy Clin North Am 30:159–171

Darzynkiewicz Z, Juan G, Srour EF (2004) Differential staining of DNA and RNA. Curr Protoc Cytom *Chapter 7*, Unit 7 3

Kiel MJ, Yilmaz OH, Iwashita T, Terhorst C, Morrison SJ (2005) SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. Cell 121:1109–1121

Osawa M, Hanada K, Hamada H, Nakauchi H (1996) Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. Science 273:242–245

Spangrude GJ, Heimfeld S, Weissman IL (1988) Purification and characterization of mouse hematopoietic stem cells. Science 241:58–62

Weissman IL, Shizuru JA (2008) The origins of the identification and isolation of hematopoietic stem cells, and their capability to induce donor-specific transplantation tolerance and treat autoimmune diseases. Blood 112:3543–3553

Single Cell Assay, Mesenchymal Stem Cells

Steven D. Rhodes[1] and Feng-Chun Yang[2]
[1]School of Medicine, Indiana University, Indianapolis, IN, USA
[2]Departments of Pediatrics, Herman B Wells Center for Pediatric Research, School of Medicine, Indiana University, Indianapolis, IN, USA

Synonyms

Colony-forming fibroblastic cells; Mesenchymal stem progenitor cells (MSPCs); Multipotent marrow stromal cells; Single-cell culture

Definition

Single-cell assays (► Single Cell Experiments) have utility in a variety of fields, particularly in the area of stem cell biology where stem cell ► differentiation potency and ► self-renewal potentials need to be assessed on a unicellular basis. Single stem cells are typically isolated using a technology known as ► fluorescence-activated cell sorting (FACS) to separate individual cells of a particular type based on the expression level of specific cell surface antigens or markers. The proliferation and Differentiation potency of single stem cells can be examined in vitro and in vivo by a variety of experiments, including colony formation in liquid or semisolid media, differentiation in the presence of cytokines and other growth factors, and, for hematopoietic stem cells, single-cell transplantation to irradiated host animals such as mice (► Single Cell Assay, Hematopoietic Stem Cell) (Ema et al. 2006).

► Mesenchymal stem cells (MSCs) are a specialized type of adult stem cell that reside in multiple tissues throughout the body, including the bone marrow, peripheral blood, fetal liver, and lung. Like other stem cells, MSCs retain the properties of both self-renewal and multipotency, the potential to differentiate into multiple cell types depending on the extracellular environment and growth factors present. Mesenchymal stem cells exhibit the capacity to differentiate into a variety of cell lineages, including ► osteoblasts, ► chondrocytes, ► adipocytes, connective tissue stromal cells, muscle, lung, gut epithelium, and neurons (Uccelli et al. 2008).

While adult stem cell types such as hematopoietic and neuronal stem cells have been well characterized by the expression of specific cell surface markers (► Fluorescent Markers), mesenchymal stem cells express a number of nonspecific cell surface antigens making them difficult to define at a single cell level. This issue has resulted in debate among stem cell biologists regarding the true nature of mesenchymal stem cells. To resolve this question, single cell assays are being utilized to validate the capacity for self-renewal and multipotency of putative mesenchymal stem cells defined by various combinations of cell surface markers.

Characteristics

Mesenchymal Stem Cell Isolation

Mesenchymal stem cells are most commonly harvested from mononuclear cells of the bone marrow, although they also reside within a number of other tissues, including cord blood, amniotic fluid, fetal liver, periosteum, and adipose tissue (Chamberlain et al. 2007). Bone marrow mononuclear cells are isolated from the buffy coat fraction following density gradient centrifugation to eliminate nonnucleated cells, such as anuclear erythrocytes and polymorphonuclear neutrophils. These cells are suspended in medium supplemented with fetal bovine serum and allowed to adhere to tissue culture dishes in a 37°C, 5% CO_2 incubator. Adherent bone marrow mononuclear cells contain mesenchymal stem cells as well as a number of other cell types, including fibroblasts, hematopoietic progenitor cells, macrophages, endothelial cells, and adipocytes. With repeated passaging in culture, many contaminating lineages die off under certain culture conditions or are washed away. Further enrichment of MSC cultures can also be achieved using deprivational media or frequent medium changes to facilitate the removal of non-MSC lineages (Soleimani and Nadri 2009).

The issue of defining a true MSC population within a culture of MSC-like cells has been a topic of controversy among stem cell biologists over recent years. Although MSCs express a variety of cell surface markers (► Fluorescent Markers), these antigens are neither unique nor specific to MSCs. Currently, MSCs in tissue culture are defined by the expression level of a panel of cell surface markers. MSCs derived from the bone marrow commonly express CD29, CD44, CD49a-f, CD51, CD73, CD105, CD106, CD166, and Stro1. They are also characterized by their lack of expression of CD11b, CD14, and CD45, which are all markers of the hematopoietic lineage as well as the endothelial cell marker CD31 (Chamberlain et al. 2007). Unfortunately, the expression level of MSC cell surface antigens varies significantly across the species of origin, the tissue source from which they are derived, and even cell culture conditions. Therefore, true MSCs are difficult to define by cell markers alone, but rather, functional assays are required to assess their self-renewal and multipotency. Such experiments are most informative when they are performed at the level of the single cell (▶ Single Cell Experiments).

Single-Cell Assays of Mesenchymal Stem Cell Function

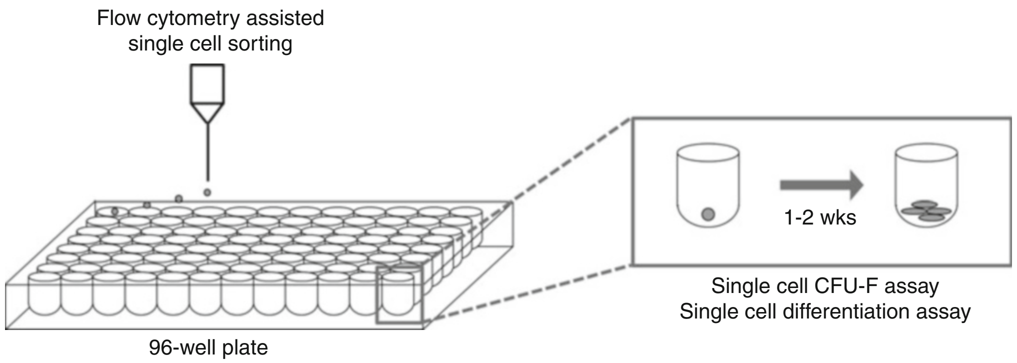
Like other adult stem cells, MSCs are characterized by their capacity to self-renew (▶ Self-Renewal) and give rise to multiple terminally differentiated cell types (▶ Differentiation Potency). MSCs are capable of forming mesodermal cell types such as ▶ osteoblasts, ▶ adipocytes, ▶ chondrocytes, muscle, and connective tissue stromal cells. Recently, they have also been reported to transdifferentiate to tissues of the endoderm such as gut epithelium and lung tissue, as well as neurons of the neuroectoderm (Uccelli et al. 2008). Techniques involving in vitro single-cell assays of mesenchymal stem cell differentiation are discussed in the following paragraphs.

After isolation and culture of MSCs from tissue such as the bone marrow, multicolor ▶ Fluorescence-activated cell sorting (FACS) is commonly utilized to isolate individual MSC-like cells based on the expression level of an array of cell surface markers (▶ Fluorescent Markers). Modern day flow cytometers are now capable of sorting single cells into 96 well plates selected via a panel of surface antigens (▶ Flow Cytometry, ▶ Cell Sorting). Besides flow cytometry, another commonly used method for single-cell harvesting and transfer is micromanipulation. In particular, automated micromanipulators offer advantages for high throughput applications, including speed and precise cell positioning; however, unless coupled to a flow cytometer, this technology lacks the high specificity achieved by selecting individual cells with a signature of fluorescent markers.

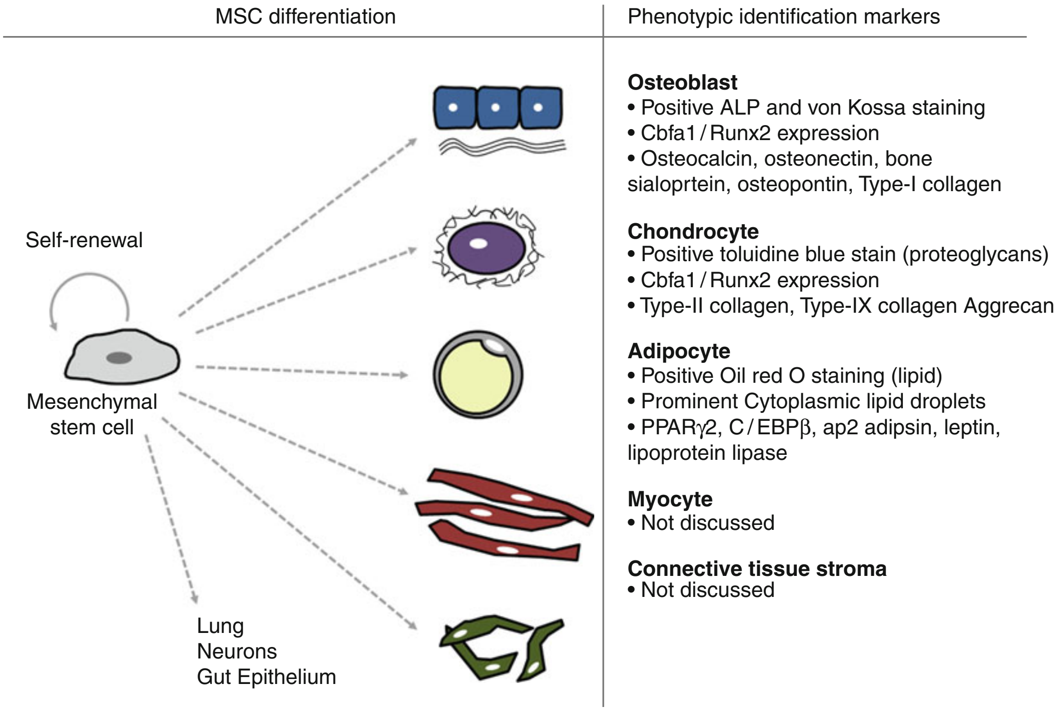
After sorting single ▶ mesenchymal stem cells into 96 well plates, in vitro single cell assays can now be performed. One property common to all stem cells is their capacity for ▶ self-renewal. For mesenchymal stem cells, this is traditionally assessed by the ability to form colonies after replating on plastic tissue culture dishes at low density. This is termed the colony-forming units (CFU) assay in which the number of individual colonies is counted after plating MSCs at subconfluent densities for 2 weeks of culture. The traditional CFU assay, however, is limited in that MSCs from one colony may be lifted and replated elsewhere, giving rise to multiple colonies. In contrast, the single-cell CFU (sc-CFU) assay circumvents this issue because it ensures that any and all colonies formed in a particular well belong to a single MSC. As such, sc-CFU assays are particularly useful in determining the true percentage of stem cells with high self-renewing capacity in various populations of MSC-like cells defined by cell surface marker expression panels (▶ Fluorescent Markers) (Fig. 1, Pochampally 2008).

Differentiation potency, or the capacity to give rise to multiple cell lineages, is another fundamental characteristic of stem cells. As noted previously, MSCs can differentiate into multiple mesodermal cell types, including ▶ osteoblasts, ▶ chondrocytes, ▶ adipocytes, connective tissue stromal cells, and muscle. Transdifferentiation to lineages of the endoderm (gut epithelium, lung, etc.) and neuroectoderm (neurons) has also been reported but will not be discussed in detail here.

Osteoblast (▶ Osteoblasts) differentiation of MSCs can be achieved by the addition of osteotropic factors to the culture media such as ascorbic acid, dexamethasone, β-glycerophosphate, and bone morphogenetic protein (BMP). After 5–7 days of culture, the osteoblast differentiation potential of MSCs can be evaluated by a variety of phenotypic markers at the cellular and molecular level. Mature osteoblasts stain positively for alkaline phosphatase (ALP). Von kossa staining may also be performed to check for calcium deposits in the extraceullar matrix. On a molecular level, terminally differentiated osteoblasts have been shown to express the critical ▶ transcription factor Cbfa1/Runx2, as well as an array of other markers, including osteocalcin, osteonectin, bone sialoprotein, osteopontin, and Type-I collagen, which can be assessed at the mRNA level by ▶ Quantitative



Single Cell Assay, Mesenchymal Stem Cells, Fig. 1 Single-cells mesenchymal stem cells are sorted into 96 well plates via fluorescence-activated cell sorting (FACS) utilizing a flow cytometer. In vitro single-cell assays can subsequently be performed. These include the single-cell CFU (sc-CFU) assay, a metric for self-renewal capacity, as well as differentiation assays to assess differentiation potency


MSC differentiation
Phenotypic identification markers
Self-renewal
Mesenchymal
stem cell
Lung
Neurons
Gut Epithelium
Osteoblast
• Positive ALP and von Kossa staining
• Cbfa1 / Runx2 expression
• Osteocalcin, osteonectin, bone sialoprtein, osteopontin, Type-I collagen
Chondrocyte
• Positive toluidine blue stain (proteoglycans)
• Cbfa1 / Runx2 expression
• Type-II collagen, Type-IX collagen Aggrecan
Adipocyte
• Positive Oil red O staining (lipid)
• Prominent Cytoplasmic lipid droplets
• PPARγ2, C / EBPβ, ap2 adipsin, leptin, lipoprotein lipase
Myocyte
• Not discussed
Connective tissue stroma
• Not discussed

Single Cell Assay, Mesenchymal Stem Cells, Fig. 2 Mesenchymal stem cells retain the capacity for self-renewal as well as the property of multipotency. MSCs have been shown to differentiate to an array of mesodermal cell types, including osteoblasts, chondrocytes, adipocytes, connective tissue stromal cells, and muscle. Transdifferentiation to certain ectodermal cell types, such as neurons, and endodermal cell lineages, such as gut epithelium and lung, has also been reported

Real-time Polymerase Chain Reaction (qRT-PCR) (▶ Gene Expression) (Minguell et al. 2001).

MSCs can differentiate toward the chondrocyte (▶ Chondrocytes) lineage with the addition of ascorbic acid plus growth factors such as transforming growth factor-beta 1 (TGF-β1) or TGF-β3. Chondrocytes can be defined phenotypically by the expression of Cbfa1/Runx2, in addition to Type-II collagen, Type-IX collagen, and Aggrecan mRNA. Histologically, mature chondrocytes stain positively with toluidine blue, signifying an abundance of proteoglycans within the extracellular matrix (Minguell et al. 2001).

Adipocytes can also be derived from mesenchymal stem cells with combinations of factors such as dexamethasone, isobutilmethylxanthine, indomethacin, and insulin. A classic methodology for identifying adipocytes in cell culture is positive Oil red O staining, indicating the accumulation of lipid droplets within the cytoplasm. Molecular markers for adipocytes are also available, and they include peroxisome proliferator-activated receptor gamma 2 (PPARγ2), C/EBPβ, aP2, adipsin, leptin, and lipoprotein lipase (Fig. 2, Minguell et al. 2001).

Single-cell differentiation assays (▶ Single Cell Experiments) for MSCs hold distinct advantages over traditional bulk culture approaches because they allow the investigator to accurately determine the percentage of MSC-like cells that retain true multipotency. Examining the proportion of true multipotent stem cells selected using various combinations cell surface markers can better inform our phenotypic definition of a mesenchymal stem cell. In turn, this will allow for improved selection of purified MSC populations for in vivo experimentation and future therapeutic applications in human patients.

Limitations

A major limitation of single-cell assays for mesenchymal stem cells is the lack of in vivo experiments for assessing MSC function on a single-cell basis. By contrast, hematopoietic stem cell (HSC) function can be evaluated on a cell-by-cell basis both in liquid culture and semisolid media as well as by the transplantation of single HSCs into a lethally irradiated host animal (▶ Single Cell Assay, Hematopoietic Stem Cell). For mesenchymal stem cells, comparable experiments would be difficult to perform in vivo due to the inherent difficulty of ablating prexisting MSCs in the host animal without incurring irreparable damage to the hematopoietic system or other cell lineages.

Cross-References

▶ Adipocytes
▶ Cell Sorting
▶ Chondrocytes
▶ Differentiation Potency
▶ Flow Cytometry
▶ Fluorescence-Activated Cell Sorting
▶ Fluorescent Markers
▶ Gene Expression
▶ Mesenchymal Stem Cells
▶ Osteoblasts
▶ Quantitative Real-time Polymerase Chain Reaction
▶ Self-Renewal
▶ Single Cell Assay, Hematopoietic Stem Cell
▶ Single Cell Experiments
▶ Transcription
▶ Transcription Factor

References

Chamberlain G, Fox J, Ashton B, Middleton J (2007) Concise review: mesenchymal stem cells: their phenotype, differentiation capacity, immunological features, and potential for homing. Stem Cells 25:2739–2749

Ema H, Morita Y, Yamazaki S, Matsubara A, Seita J, Tadokoro Y, Kondo H, Takano H, Nakauchi H (2006) Adult mouse hematopoietic stem cells: purification and single-cell assays. Nat Protoc 1:2979–2987

Minguell JJ, Erices A, Conget P (2001) Mesenchymal stem cells. Exp Biol Med (Maywood) 226:507–520

Pochampally R (2008) Colony forming unit assays for MSCs. Methods Mol Biol 449:83–91

Soleimani M, Nadri S (2009) A protocol for isolation and culture of mesenchymal stem cells from mouse bone marrow. Nat Protoc 4:102–106

Uccelli A, Moretta L, Pistoia V (2008) Mesenchymal stem cells in health and disease. Nat Rev Immunol 8:726–736

Single Cell Experiments

Feng-Chun Yang
Departments of Pediatrics, Herman B Wells Center for Pediatric Research, School of Medicine, Indiana University, Indianapolis, IN, USA

Definition

The stem cell is an unspecialized cell that can self-renew and give rise to specialized cell lineages, such as a blood cell. Stem cells are found in all multicellular organisms. Stem cells can now be grown and can differentiate into specialized cell types with characteristics consistent with cells of various tissues such as blood, muscles, or nerves in in vitro cultures. Up-to-now, highly plastic adult stem cells are broadly used in clinical therapies. Highly purified and characterized stem cells from mice have opened up exceedingly rich fields of basic research with a variety of clinical potential. Many of the techniques used in the study of stem cell biology have become more standardized nowadays. In this entry, we provide information on up-to-date tools for studying stem cells. We wish the information in this entry will help accelerate studies in the stem cell field.

Single Stem Cell Studies

A stem cell is characterized by its ability to self-renew, high proliferative capability, and multilineage differentiation potential. Stem cells can be found in multicellular organisms. The research in the stem cell field was initiated by pioneers Ernest A. McCulloch and James E. Till in the early 1960s (McCulloch and Till 1960).

There are two major types of stem cells: embryonic stem cells and adult stem cells. Embryonic stem cells are isolated from the inner cell mass of a blastocyst. During development, embryonic stem cells have the capability to differentiate into all of the specialized embryonic tissues. An adult stem cell is considered to be an undifferentiated cell that can be found in adult tissues/organs. An adult stem cell can renew itself and differentiate to yield some or all of the major specialized cell types of the tissue or organ. Adult stem cells can maintain the normal turnover of regenerative organs, such as blood, skin, liver, or intestinal tissues. These adult stem cells can also repair systems for the body, replenishing specialized cells.

Hematopoietic stem cells (HSCs) are multipotent stem cells that give rise to all blood cell types including lymphoid cells (T cells, B cells, nature killer cells), myeloid cells (monocytes, macrophages, neutrophils, eosinophils, erythrocytes, dendritic cells, megakaryocytes, and platelets). In the late 1980s, stem cell biologists were able to purify the hematopoietic stem cells, which opened a new era for (1) better understanding the stem cell biology, (2) improving stem cell transplantation, (3) achieving better understanding of stem cell diseases such as leukemia and myeloma, and (4) promoting regenerative medicine. Umbilical cord blood is obtained from the umbilical cord at the time of childbirth, after the cord has been detached from the newborn (Cairo and Wagner 1997; Broxmeyer and Smith 2009). Scientists are able to collect stem cells from cord blood including hematopoietic stem cells. The placenta is another source of hematopoietic stem cells. The placenta contains up to ten times more stem cells than cord blood (Cairo and Wagner 1997).

Hematopoietic stem cell transplantation is a procedure of transplanting pluripotent stem cells into recipients. The stem cells can be derived from either bone marrow or umbilical cord blood. Nowadays, stem cell transplantation has become a common medical treatment procedure for people with diseases of the blood, bone marrow, inflammation, or certain cancers. However, hematopoietic stem cell transplantation remains risky as this procedure has many possible complications. Thus, it has been reserved for patients with life-threatening diseases (Tyndall et al. 1999; Burt et al. 2008).

Induced pluripotent stem cells (iPS cells) are a type of pluripotent stem cells derived from a non-pluripotent cell, generally an adult somatic cell by artificial overexpression of specific genes. Similar to ES and adult stem cells, iPS cells express certain stem cell genes and proteins. iPS cells also have similar chromatin methylation patterns and doubling time as ES cells and adult stem cells. Most importantly, these iPS cells form embryonic bodies and have the potential of differentiation into multilineages under certain conditions (Takahashi and Yamanaka 2006).

Basic research of stem cells in laboratories enables scientists to understand the essential properties of stem cells. In fact, scientists have already used stem cells in the laboratory for drug screening and also for the development of model systems to understand normal growth and to identify the causes of birth defects. However, the heterogeneity of stem cells from stem cell cultures has been a significant obstacle in obtaining a homogeneous population for understanding the stem cell physiology and conducting directed differentiation protocols. Studies of single cell/homogeneous populations can eliminate the variability hindering population studies, thereby providing a better understanding of stem cell physiology. Single cell analysis with our technique will yield valuable insight into the process of self-renewal, proliferation, lineage commitment, and differentiation. It will also provide a platform for the systematic discovery of lineage specific markers useful in refining directed differentiation process.

In this entry, we will present methods/approaches for studying the stem cells at single cell level. Flow cytometry is a technique for scoring and examining cells by suspending them in a stream of fluid and passing them by an electronic detection apparatus. This allows scientists to perform multiparametric analysis of the physical and/or chemical characteristics of up to thousands of particles per second. Nowadays, flow cytometry is routinely used in the diagnosis of health disorders. Dr. Liu provides detailed information regarding the flow cytometry in this entry. The fluorescence-activated cell sorter is a machine that can rapidly separate the cells in a suspension on the basis of size and the color of their fluorescence. The cells will not be damaged by the process. In contrast, the percent viability of the sorted cells can be higher than that in the original suspension due to the setting of the machine to ignore droplets containing dead cells. Fluorescent marker is a molecule, like a protein, which is covalently attached by a fluorophore to selectively bind to a functional group of the target for detection. The most commonly used fluorescent molecules are antibodies. As detailed in the section of Fluorescent Markers described by Dr. He, cell marker is a specified protein on the surface of every cell, named receptor, which can selectively bind or adhere to other signaling molecules. The biological uniqueness of the receptors and chemical properties of certain compounds was used to mark cells. Cluster of differentiation (CD) molecules are markers on the cell surface, which can be recognized by specific sets of antibodies. Cluster of differentiation system can be used to identify the cell type, stage of differentiation, and activity of a cell.

Live cell imaging is the study of living cells using images acquired from microscopy. This method provides a tool to better view biological function through the study of cellular dynamics. This technology has become increasingly and widely accessible for scientists to produce pivotal publications in cell biology, developmental biology, cancer biology, and many other related biomedical research laboratories.

Mesenchymal stem cells (MSCs) are adult stem cells that reside in multiple tissues throughout the body including the bone marrow, peripheral blood, and fetal liver and lung. MSCs retain the properties of both self-renewal and pluripotency, the potential to differentiate into multiple cell types depending on the extracellular environment and growth factors present. MSCs exhibit the capacity to differentiate into a variety of cell lineages including osteoblasts (bone), chondrocytes (cartilage), adipocytes (fat), keratinocytes (skin), fibroblasts (connective tissue), and muscle. In this entry, Dr. Rhodes describes the up-to-date protocols for study of the MSC biologic functions, including MSC isolation and MSC functions at single cell level.

Clonal cultures are likely the closest technique for single cell in vivo studies of hematopoietic stem cells. In addition, hematopoietic stem cell transplantation is a traditional protocol that has been used to ascertain that a single hematopoietic stem cell contained in donor cells is transplanted into a lethally irradiated host with a high probability. The clonal expansion of this stem cell can then be observed over time by monitoring the percent donor-type cells in blood as the host is reconstituted. The resulting time series is defined as the repopulation kinetic of the hematopoietic stem cells. Dr. Staser will discuss these methodologies in this entry.

Spectroscopic methods are the main tools of modern chemistry for the identification of molecular structures. These methods play a vital part in many areas of science. Spectroscopic methods allow scientists to identify compounds, to confirm molecular structures, to determine and monitor reactions, and to control the purity of compounds. Spectroscopy can provide information as to how the electromagnetic spectrum interacts with matter. Spectroscopic methods use the entire electromagnetic spectrum: from X-rays (with a wavelength of 0.1 nm) to radio waves (with a wavelength of 1,000 m). Spectroscopic methods are critical in organic chemistry and are required as part of chemistry courses in all universities. Dr. Wu describes these methods in this entry.

It is our intention that the entry "▶ Single Cell Experiments" provides readers detailed and up-to-date information for single stem cell research. We hope you find "▶ Single Cell Experiments" instructive and useful.

Summary

This entry provides up-to-date, yet standard study methods for the research of single stem cells. It is our hope that this entry will help accelerate studies in the stem cell field. Our sincere thanks goes to all of the contributors and those in the stem cell field that enlarged our thinking and provided new views/tools to further understand this promising stem cell.

References

Broxmeyer HE, Smith FO (2009) Cord blood hematopoietic cell transplantation. In: Appelbaum FR, Forman SJ, Negrin RS, Blume KG (eds) Thomas' hematopoietic cell transplantation, 4th edn. Wiley-Blackwell, West Sussex

Burt RK, Loh Y, Pearce W et al (2008) Clinical applications of blood-derived and marrow-derived stem cells for nonmalignant diseases. J Am Med Assoc 299:925–936

Cairo MS, Wagner JE (1997) Placental and/or umbilical cord blood: an alternative source of hematopoietic stem cells for transplantation. Blood 90:4665–4678

McCulloch EA, Till JE (1960) The radiation sensitivity of normal mouse bone marrow cells, determined by quantitative marrow transplantation into irradiated mice. Radiat Res 13:115–125

Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126:663–676

Tyndall A, Fassas A, Passweg J et al (1999) Autologous haematopoietic stem cell transplants for autoimmune disease – feasibility and transplant-related mortality. Autoimmune disease and lymphoma working parties of the European group for blood and marrow transplantation, the European league against rheumatism and the international stem cell project for autoimmune disease. Bone Marrow Transplant 24:729–734

Single Hematopoietic Stem Cell Transplantation

► Single Cell Assay, Hematopoietic Stem Cell

Single Nucleotide Polymorphisms

Vani Brahmachari and Shruti Jain
Dr. B. R. Ambedkar Center for Biomedical Research,
University of Delhi, Delhi, India

Synonyms

Point mutations; SNPs

Definition

It is a variation that can be mapped to a single locus, as it results from the substitution of one nucleotide by another. Genetic polymorphism is distinguished from mutation based on its higher frequency of occurrence in the population indicating that the variation in the gene sequence is not causing a drastic effect relative to the normal sequence, which is seen in a much larger number of individuals in the population. But, in principle, point mutations are single base changes in the gene sequence. All SNPs may not bring about a change in the amino acid sequence of the protein coded by the gene because of the redundancy in the genetic code, which is defined as the synonymous substitution, or they can bring about a change in the amino acid sequence of the protein causing missense or they can prematurely terminate the protein synthesis because of generating a nonsense mutation. They allow for high-throughput genotyping. SNPs bring about subtle changes in the activity of gene/its product, which can have considerable effect under conditions of environmental challenges like exposure to ► xenobiotics.

Cross-References

► Epigenetics, Drug Discovery
► Xenobiotics

Single Round Assay

Nobuo Shimamoto
Faculty of Life Sciences, Kyoto Sangyo University,
Kyoto, Japan

Definition

Single Round Assay of transcription is the assay using DNA only in a single round of transcription. A promoter complex is formed during preincubation and heparin is added together with substrate nucleoside triphosphates (NTPs) to prevent free RNA polymerase from binding to DNA.

Cross-References

▶ Transcription in Bacteria

Single-Cell Culture

▶ Single Cell Assay, Mesenchymal Stem Cells

Single-Cell Time-Lapse Microscopy

▶ Cell Cycle Analysis, Live-Cell Imaging

Single-Gene Disorders

▶ Mendelian Traits

Single-Input Module

Guangxu Jin
Systems Medicine and Bioengineering,
Bioengineering and Bioinformatics Program, The
Methodist Hospital Research Institute, Weill Medical
College, Cornell University, Houston, TX, USA

Synonyms

Single-input module motif, SIM

Definition

Single-input module (SIM) was found in the *E. coli* transcriptional regulation network (Shen-Orr et al. 2002). A single transcription factor, X, regulates a set of operons $Z_1, \ldots, Z_n$. X is usually autoregulatory. All regulations are of the same sign (positive or negative). No other transcription factor regulates the operons. Mathematical modeling suggests that SIM can show a detailed temporal program of expression resulting from differences in the activation thresholds of the different genes. Built into this design is a pattern in which the first gene activated is the last one to be deactivated. The SIM motif is found in systems of genes that function stoichiometrically to form a protein assembly (such as flagella) or a metabolic pathway (such as amino-acid biosynthesis).

References

Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31:64–68

Single-Input Module Motif, SIM

▶ Single-Input Module

Single-Nucleotide Polymorphism Database

▶ dbSNP

Singular Value Decomposition (SVD)

▶ Principal Component Analysis (PCA)

Site-directed Mutagenesis

Animesh Bhattacharya
Department of Dermatology, Venereology and
Allergology, University of Leipzig, Leipzig, Germany

Synonyms

Oligonucleotide-directed mutagenesis; Site-specific mutagenesis

Definition

In molecular biology the method called site-directed mutagenesis is used to induce a mutation at a defined locus (site) in a DNA sequence.

Site-Specific Mutagenesis

► Site-directed Mutagenesis

Size Checkpoint

► Cell Cycle, Cell Size Regulation

Size Control

► Cell Cycle, Cell Size Regulation

Skp1/Cul1/F-Box Containing Complex (SCF)

Sergio Moreno
Instituto de Biología Molecular y Celular del Cáncer,
CSIC/Universidad de Salamanca, Salamanca, Spain

Definition

SCF is a multiprotein E3 ubiquitin ligase complex that is involved in the ubiquitylation of proteins for their degradation by the proteasome. In the SCF complex, the F-box component recognizes specific targets for destruction by the proteasome. F-box protein Skp2 specifies degradation of CDK inhibitors p21$^{\mathrm{Cip1}}$, p27$^{\mathrm{Kip1}}$, and p57$^{\mathrm{Kip2}}$ at G1/S. In mitosis, F-box protein βTrCP marks the degradation of mitotic proteins, such as Emi1 and Wee1.

Cross-References

► CDK Inhibitors

Slot Filling

► Template Filling, Text Mining

Slow-Fast Dynamics

Jinzhi Lei
Zhou Pei-Yuan Center for Applied Mathematics,
Tsinghua University of Beijing, Beijing, China

Definition

Slow-fast dynamics is a class of dynamical systems that are characterized by several different timescales (Berglund and Gentz 2006).

There are many examples of biological systems in which such differential timescales are well separated, for instance:

- The 24 h cycles of ► circadian rhythms that are closely associated with the fast process of the expressions of clock genes
- The fast transportation of extracellular single molecules and the slow embryo growing
- Different timescale reactions in a genetic network, including protein-protein interaction, ► transcription, translation, and molecule transportation

The feature of the separation of timescales allows us to model the system by slow-fast ordinary differential equations of form

$$\begin{aligned} \varepsilon \frac{dx}{dt} &= f(x, y), \\ \frac{dy}{dt} &= g(x, y), \end{aligned} \qquad (1)$$

where ε is a small parameter. Here x contains the fast variables, and y the slow ones.

Equivalently, the system can be written in fast time $s = t/\varepsilon$ as

$$\begin{aligned}\frac{dx}{ds} &= f(x,y),\\ \frac{dy}{ds} &= \varepsilon g(x,y).\end{aligned} \tag{2}$$

In many situations, we can reduce the slow-fast dynamics by decoupling different timescales variables. As the variable y varies slowly in time, the dynamics of the fast variables in (2) can be considered as a parameter-dependent ordinary differential equation

$$\frac{dx}{dt} = f(x,\lambda). \tag{3}$$

If (3) admits an asymptotically stable equilibrium point $x(\lambda)$ for each value of λ, the fast variables can be eliminated, at least locally, by a projection onto the set of equilibria, called the *slow manifold*. This yields the effective *reduced system*

$$\frac{dy}{dt} = g(x(y), y) \tag{4}$$

for the slow dynamics, which is simpler to analyze than the full system. For instance, switches between different (parts of) slow manifolds (toggle switch), relaxation oscillations in which periodic motions of fast and slow phase alternate, etc.

Adding noise to a slow-fast dynamics will add one or several new timescales to the dynamics, namely, the metastable lifetimes (or Kramer's time). The dynamics will depend in an essential way on the relative values of the deterministic system's intrinsic timescales, and the Kramer's time that is introduced by noise (Berglund and Gentz 2006).

References

Berglund N, Gentz B (2006) Noise-induced phenomena in slow-fast dynamical systems: a sample-paths approach. Springer, London

Small GTPases

Sebastian Mana-Capelli and Dannel McCollum
Department of Molecular Genetics and Microbiology,
University of Massachusetts, Worcester, MA, USA

Synonyms

Monomeric GTPases

Definition

Small GTPases are a GTP-binding superfamily of proteins ubiquitous among eukaryotic cells. They typically range between 21 and 25 KDa in size and can be classified into the Ras, Rho/Rac, Rab, Arf, and Ran subfamilies based on conserved domains that determine their function. For example, members of the Rho family of proteins are usually involved in the regulation of cytoskeleton dynamics, including cell morphology, motility, and cytokinesis. Differently, Ran GTPases usually regulate nuclear import/export dynamics and mitotic spindle assembly (Alberts et al. 2008).

Small GTPases are considered molecular switches that cycle between an active GTP-bound state and an inactive GDP-bound state. Hydrolysis of GTP to GDP and the exchange back to GTP is not typically performed by the intrinsic activity of the GTPase but by interactions with catalytic binding partners. Proteins that catalyze the hydrolysis of GTP into GDP are GTPase-activating proteins (GAPs), and proteins that help GTPases to release GDP in exchange for GTP are GTP exchange factors (GEFs). Control over the cellular localization and activity of GAPs and GEFs results in spatial and temporal regulation of small GTPase. In addition, guanine nucleotide dissociation inhibitors, or GDIs, can sequester small GTPases in their GDP-bound state and prevent them from interaction with their corresponding GEF (Alberts et al. 2008).

Cross-References

▶ Cytokinesis

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2008) Molecular biology of the cell, 5th edn. Garland Science, New York, pp 178–181, 708–710, 895–895, 926–931

Small Interfering RNA

Melissa L. Kemp
The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Definition

Small interfering RNAs are short 19–25-nucleotide long double-stranded RNA molecules that are endogenously generated by processing natural or synthetic precursors. Small interfering RNAs (siRNAs) can associate with the multiprotein complex RNA-induced silencing complex (RISC) which separates the two strands of the siRNA and retains one of them. The RISC-incorporated strand of siRNA then associates with complementary mRNA sequences in the cytoplasm and mediates degradation of the mRNA.

Small Molecule

Riza Theresa Batista-Navarro
National Centre for Text Mining, Manchester Interdisciplinary Biocentre, Manchester, UK

Synonyms

Small-molecule drug

Definition

A small molecule is a molecule with a low molecular weight, as opposed to a biological macromolecule (e.g., therapeutic proteins). Small molecules comprise majority of the pharmaceutical agents (Osbourn 2007).

A small molecule has a single, fixed chemical formula. Unlike biological macromolecules which are degraded due to their protein nature, small-molecule drugs are metabolized. Their low molecular weight makes it possible for them to be taken orally and allows them to diffuse easily (Osbourn 2007).

Cross-References

▶ Natural Product Resources

References

Osbourn JK (2007) Biological macromolecules. In: Taylor JB, Triggle DJ (eds) Comprehensive medicinal chemistry II. Elsevier, Oxford, pp 431–447

Small Protein B

▶ SmpB

Small-Molecule Drug

▶ Small Molecule

Small-World Property

Hendrik Mehlhorn[1] and Falk Schreiber[1,2]
[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Stadt Seeland, Germany
[2]Martin Luther University Halle–Wittenberg, Halle, Germany

Definition

A ▶ graph G = (V, E) has a small-world property if it has a high ▶ clustering coefficient and a small ▶ characteristic path length. A high ▶ clustering coefficient represents a local connectivity property, typically resulting in a high number of cliques and near-cliques, which denote subnetworks comprising edges between all or almost all vertices. A small ▶ characteristic path length represents a global reachability property and roughly behaves logarithmic to the number of ▶ graph vertices.

Characteristics

Properties

The high ▶ clustering coefficient in small-world networks points to the importance of dense local interconnections and cliquishness. In the case of biological networks, these represent functional modules such as a set of proteins achieving a common function, a set of genes participating in a common cellular process, or local interconnections within metabolic pathways. The small ▶ characteristic path length in small-world networks points to the importance of short paths between any two vertices. In the case of biological networks, these afford fast information flow in gene regulatory and signal transduction networks as well as fast reaction paths in metabolic networks. Both properties frequently appear jointly in biological networks and significantly discriminate their topology from random ▶ graphs, thus the topology of biological networks is most likely the result of selection pressure (Wagner and Fell 2001; Tong et al. 2004; Yook et al. 2004).

Other networks which have been shown to have the small-world property are the World Wide Web, energy grids, and social networks. ▶ Graphs, which model spatial or temporal proximity, are less likely to have the small-world property. Reasons are that there is simply no short way between remote places and no short time interval between distant time points.

Scale-Freeness

A small ▶ characteristic path length of biological networks arises from a ▶ graph property named scale-freeness. A ▶ graph is scale free if its connectivity (also called degree) distribution follows a power-law. The probability that a vertex is adjacent to k other vertices follows $P(k) \sim ck^{-\gamma}$, where c is a normalization constant and $\gamma \in [2, 3]$. Thus, the degree distributions of multiple scale-free ▶ graphs of different sizes look similar after rescaling the degree axis. There are also results indicating that the degree distribution of many biological networks is not exactly scale free, but rather follows a truncated power-law (Khanin and Wit 2006). However, scale-free ▶ graphs have been shown to be ultrasmall denoting a ▶ characteristic path length which behaves like $\log \log |V|$ with $|V|$ representing the number of vertices in the ▶ graph (Cohen and Havlin 2003). The scale-freeness property results in many weakly connected vertices and only few strongly connected vertices. The latter vertices are also named ▶ hubs and are considered to be responsible for the small ▶ characteristic path length. At the same time, the scale-freeness property results in a great error-robustness of biological networks, because random errors are more likely to effect a weakly connected vertex then a strongly connected vertex. For example, in case of the protein-protein interaction network of yeast it has been shown that a mutation of a ▶ hub protein is more likely to be lethal than a mutation of a non-▶ hub protein (Jeong et al. 2001).

Models for Small-World Networks

The first model for small-world networks was proposed by Watts and Strogatz and is called the Watts-Strogatz model (Watts and Strogatz 1998). The starting point is a ▶ graph which is composed of a cycle of vertices, each vertex being connected to the next k ($k \geq 2$) vertices on the cycle. For each vertex and each edge connected to this vertex, the edge gets rewired with a specified probability p by reconnecting the opposite end of the edge to a vertex chosen equiprobable from all vertices, such that no two edges connect the same vertex pair and no edge starts and ends at the same vertex. Rewired edges are also called shortcuts, because these probably reduce the number of edges in the shortest path between several vertices. Depending on p, the ▶ clustering coefficient and the ▶ characteristic path length change dramatically. However, there is a certain range of p resulting in a high ▶ clustering coefficient and a small ▶ characteristic path length yielding the small-world property. Another popular model was proposed independently by Monasson (Monasson 1999) and Newman and Watts (Newman and Watts 1999). Instead of rewiring edges to create shortcuts, new edges are added between all vertex pairs with a specified probability p. This model has the advantage that the ▶ graph always stays connected resulting in a finite ▶ characteristic path length. Both models have been examined in the literature of mathematics and physics (Newman 2000). Although the proposed models are capable to reproduce the small-world property, other properties of biological networks such as scale-freeness and network motifs are lacking.

Cross-References

- ▶ Characteristic Path Length
- ▶ Clustering Coefficient
- ▶ Graph
- ▶ Hub

References

Cohen R, Havlin S (2003) Scale-free networks are ultrasmall. Phys Rev Lett 90(5):058701

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411(6833):41–42

Khanin R, Wit E (2006) How scale-free are biological networks. J Comput Biol 13(3):810–818

Monasson R (1999) Diffusion, localization and dispersion relations on "small-world" lattices. Eur Phys J B 12:555–567

Newman MEJ (2000) Models of the small world: A review. eprint arXiv:cond-mat/0001118

Newman MEJ, Watts DJ (1999) Renormalization group analysis of the small-world network model. Phys Lett A 263(4–6): 341–346

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Ménard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C (2004) Global mapping of the yeast genetic interaction network. Science 303(5659):808–813

Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc R Soc B: Biol Sci 268(1478): 1803–1810

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442

Yook SHH, Oltvai ZN, Barabási ALL (2004) Functional and topological characterization of protein interaction networks. Proteomics 4(4):928–942

SmpB

Tatsuhiko Someya[1], Nobukazu Nameki[2] and
Gota Kawai[3]
[1]University of Tsukuba, Tsukuba, Japan
[2]Gunma University, Kiryu, Japan
[3]Department of Life and Environmental Sciences,
Chiba Institute of Technology, Narashino, Chiba,
Japan

Synonyms

Small protein B

Definition

A small protein B (SmpB), which is highly conserved protein among all bacteria and some organelle, is an important factor in the *trans*-translation (Shpanchenko et al. 2010). SmpB binds to the 3′ end of the D-loop in the tRNA-like domain of tmRNA. The main core of SmpB (residues 1–133 for *Thermus thermophilus*) consists of an oligonucleotide-binding fold (OB fold) with a central β-barrel and three flanking α-helices. The poorly structured C-terminal tail of the protein (about 20 residues), rich in basic residues, plays a critical role in tmRNA tagged and function (Shpanchenko et al. 2010; Moore and Sauer 2007).

References

Moore SD, Sauer RT (2007) The tmRNA system for translational surveillance and ribosome rescue. Annu Rev Biochem 76:101–124

Shpanchenko OV, Golovin AV, Bugaeva EY, Isaksson LA, Dontsova OA (2010) Structural aspects of *trans*-translation. IUBMB Life 62:120–124

Smoluchowski Equation

▶ Stochastic Processes, Fokker-Planck Equation

SnoRNA Databases

▶ Non-coding RNA Databases

SNPedia

Jingky Lozano-Kühne
Department of Public Health, University of Oxford,
Oxford, UK

Definition

SNPedia (pronounced "snipedia") is an online database of ▶ Single Nucleotide Polymorphisms (SNPs) organized in a ▶ wiki-format to allow sharing of information about genetic variations. The Website was launched in 2006 by geneticist Greg Lennon and bioinformatician Michael Cariaso to help optimize the application of the Human Genome project to practical living and to realize the relevance of understanding genetic variations in humans (Cariaso and Lennon 2011). SNPedia can be accessed at the site http://www.snpedia.com.

Characteristics

Description

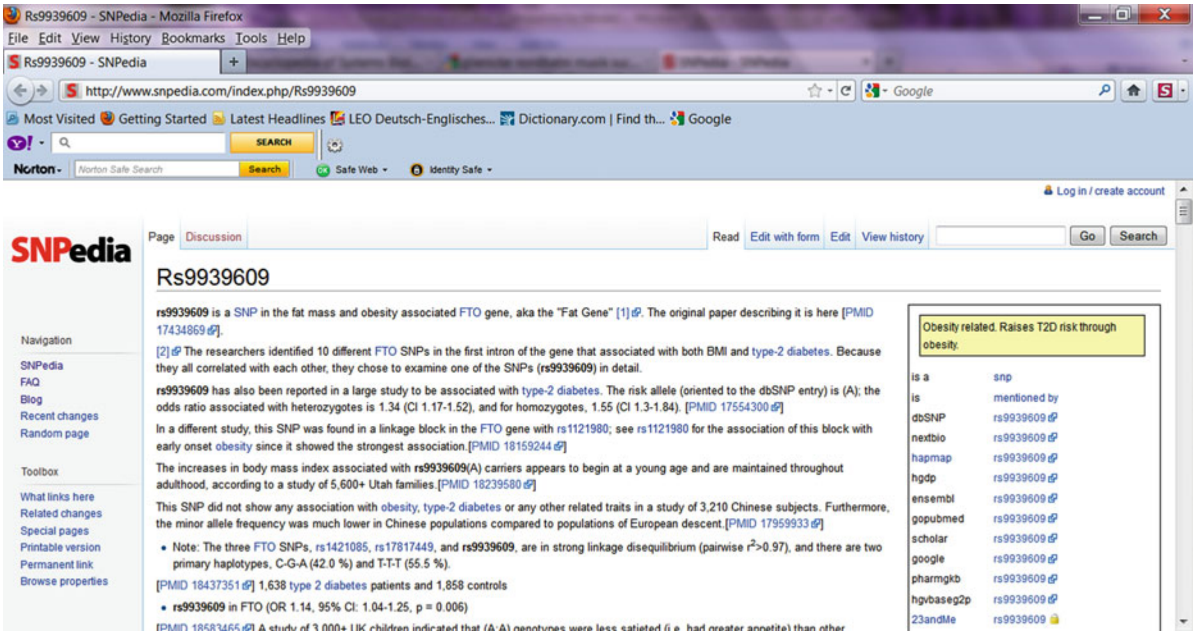
SNPedia is a free online platform that serves as a repository of information about SNPs and their roles in different health conditions based on published studies. It utilizes information from various public databases and includes interpretations of the SNPs' importance (Check Hayden 2008). SNPedia also contains users' contributed data and information on SNPs that are present on commercially available microarray (▶ DNA Microarrays) chips enabling comparison of different microarray platforms. Researchers in systems biology needing SNPs data will find more than 24,000 SNPs (as of November 2011) in the SNPedia database.

Beyond being a simple online database, SNPedia is a semantic wiki which has the ability to identify information and relationships of information between Web pages. It is powered by Media Wiki and the Semantic Media Wiki software (Cariaso and Lennon 2011). Web users and researchers can easily search or query information in a semantic wiki page. One can search the SNPedia Website by SNP, gene, medical condition, or other related topics such as medicines. Being a wiki, SNPedia allows communication between users and enables users to contribute and edit information on the Website. The contents of the Website may be freely copied, quoted, reused, and adapted by anyone in accordance with the Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License. SNPedia is associated with Prometease, a freeware computer program developed by the same SNPedia team that enables users to compare their genotype information against the SNPedia database.

Contents

Each SNP in SNPedia has a Web page with description, links to research publications as well as microarray information and links to personal genomics Websites. SNPs are usually identified by their "rs" numbers following the nomenclature of the National Center for Biotechnology Information (NCBI). For example, rs9939609 is an SNP included in SNPedia and is located in the fat mass and obesity-associated (FTO) gene or also known as the "Fat Gene" (Frayling et al. 2007). The description of the SNP in SNPedia includes its gene, chromosome location, and information (if any) of its association with human traits such as eye color, response to drugs, disease susceptibility, and other traits. The scientific evidence that supports the SNP's association with a human trait is provided as a link to PubMed (▶ MEDLINE and PubMed) abstracts or other peer-reviewed publications. The Web page for each SNP also contains links to related pages within SNPedia and other databases such as ▶ dbSNP, ▶ Ensembl, and ▶ HapMap. Data on the SNP's genotype frequencies in the HapMap population are presented as graphs in the SNP Web page. In addition, one can also connect to Internet search engines and genotyping Websites through links within the Web page. SNPedia's data and links may also provide additional supporting information in interpreting results of personal genotyping tests. A screenshot of a sample SNP Web page from the Website is shown in Fig. 1.

An SNP reported to SNPedia will automatically be connected to its neighboring SNPs. Its presence on any known commercial microarrays (▶ DNA Microarrays) will also be identified once it is included in the database. Knowing the neighboring SNPs might help in genetic testing in case the SNP of interest does not exist on any known microarrays. If an SNP is on a microarray, it can be used as a surrogate marker to test for its neighbor which is not on the microarray (Cariaso 2007).


Rs9939609 - SNPedia - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Rs9939609 - SNPedia
http://www.snpedia.com/index.php/Rs9939609
Google
Most Visited Getting Started Latest Headlines LEO Deutsch-Englisches... Dictionary.com | Find th... Google
SEARCH
Norton - Norton Safe Search Search Safe Web Identity Safe
Log in / create account
SNPedia
Page Discussion
Read Edit with form Edit View history Go Search
Navigation
SNPedia
FAQ
Blog
Recent changes
Random page
Toolbox
What links here
Related changes
Special pages
Printable version
Permanent link
Browse properties

Rs9939609

rs9939609 is a SNP in the fat mass and obesity associated FTO gene, aka the "Fat Gene" [1]. The original paper describing it is here [PMID 17434869].

[2] The researchers identified 10 different FTO SNPs in the first intron of the gene that associated with both BMI and type-2 diabetes. Because they all correlated with each other, they chose to examine one of the SNPs (**rs9939609**) in detail.

rs9939609 has also been reported in a large study to be associated with type-2 diabetes. The risk allele (oriented to the dbSNP entry) is (A); the odds ratio associated with heterozygotes is 1.34 (CI 1.17-1.52), and for homozygotes, 1.55 (CI 1.3-1.84). [PMID 17554300]

In a different study, this SNP was found in a linkage block in the FTO gene with rs1121980; see rs1121980 for the association of this block with early onset obesity since it showed the strongest association.[PMID 18159244]

The increases in body mass index associated with **rs9939609**(A) carriers appears to begin at a young age and are maintained throughout adulthood, according to a study of 5,600+ Utah families.[PMID 18239580]

This SNP did not show any association with obesity, type-2 diabetes or any other related traits in a study of 3,210 Chinese subjects. Furthermore, the minor allele frequency was much lower in Chinese populations compared to populations of European descent.[PMID 17959933]

- Note: The three FTO SNPs, rs1421085, rs17817449, and **rs9939609**, are in strong linkage disequilibrium (pairwise r^2>0.97), and there are two primary haplotypes, C-G-A (42.0 %) and T-T-T (55.5 %).

[PMID 18437351] 1,638 type 2 diabetes patients and 1,858 controls

- **rs9939609** in FTO (OR 1.14, 95% CI: 1.04-1.25, p = 0.006)

[PMID 18583465] A study of 3,000+ UK children indicated that (A;A) genotypes were less satiated (i.e. had greater appetite) than other

Obesity related. Raises T2D risk through obesity.

is a	snp
is	mentioned by
dbSNP	rs9939609
nextbio	rs9939609
hapmap	rs9939609
hgdp	rs9939609
ensembl	rs9939609
gopubmed	rs9939609
scholar	rs9939609
google	rs9939609
pharmgkb	rs9939609
hgvbaseg2p	rs9939609
23andMe	rs9939609

SNPedia, Fig. 1 A screenshot of a sample SNP page from SNPedia Website

SNPedia also provides a possibility for users to discuss about a particular SNP in the online chat room. A tab for discussion is provided on every SNP page in SNPedia for this purpose. Aside from being a discussion venue, for some users the chat room can also become a venue for genetic counseling.

Limitations and Issues

There are limitations to SNPedia's contents and there are issues concerning the privacy of personal information shared within the site. The general privacy policies associated with a wiki environment also applies to SNPedia. This means that anyone may edit the publicly editable pages of the Website and is identified publicly as an editor. As a wiki, SNPedia will continually grow and accumulate more information. However, the completeness and accuracy of the content of the wiki cannot be fully guaranteed.

Cross-References

▶ dbSNP
▶ DNA Microarrays
▶ Ensembl
▶ HapMap
▶ MEDLINE and PubMed
▶ Wiki

References

Cariaso M, Lennon G (2011) SNPedia. World Wide Web URL: http://www.SNPedia.com/. Accessed 1 May 2011

Check Hayden E (2008) How to get the most from a gene test. Nature 456(7218):11

Cariaso M (2007) SNPedia: a wiki for personal genomics. Bio-IT World, 17 Dec 2007

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316(5826):889–894, Epub 2007 Apr 12

SNPs

▶ Single Nucleotide Polymorphisms

SO

▶ Sequence Ontology

SOAP

▶ Simple Object Access Protocol

Social Epistemology

Martin Carrier
Department of Philosophy, Bielefeld University,
Bielefeld, Germany

Definition

Social epistemology analyzes the gain of knowledge as a social practice. It is understood in contrast to the traditional conception of the individual researcher, acting in isolation. It studies the impact of social interactions on the production and assessment of assertions. Social epistemology emphasizes that social factors affect the credibility of expert judgment and testimony and that scientific claims are tested and assessed within the pertinent scientific community following social rules.

Cross-References

▶ Non-empirical Values

Soft X-Ray Microscopy

Xiaohua Wu
Department of Pediatrics, Herman B Wells Center for
Pediatric Research, Indiana University School
of Medicine, Indianapolis, IN, USA

Definition

Soft x-ray microscopy provides a unique set of capabilities in between those of visible light and electron microscopy.

Cross-References

▶ Spectroscopy and Spectromicroscopy

Solid Tissue

▶ Biomarkers, Protein Expression

Solvent Accessible Surface

▶ Mathematical Morphology for Protein Surface Modeling

Solvent Excluded Surface

▶ Mathematical Morphology for Protein Surface Modeling

Somatic Gene Rearrangements

▶ Immune Repertoire Diversity

Somatic Mutations

Vito Quaranta
The Vanderbilt-Ingram Cancer Center, Nashville,
TN, USA

Definition

Mutations arising in non-germline cells not transmissible to offspring.

Cross-References

▶ Cell Cycle, Cancer Cell Cycle and Oncogene Addiction

Source Control

▶ Distributed Version Control System (DVCS)

Sources of Variability

▶ Experimental Design, Variability

Spatiotemporal Pattern Formation

Andreas Deutsch
Center for Information Services and High Performance Computing (ZIH), Technical University Dresden, Dresden, Germany

Definition

Spatiotemporal pattern formation is the process of pattern development in space (e.g., spots or spirals) or in time (e.g., oscillations). It can be observed in a large variety of systems in nature (Haken 2004; Mikhailov and Loskutov 1996; Murray 2002). Examples include oscillating chemical reactions (e.g., Belousov Zhabotinskii reaction), spiral waves in excitable media (e.g., cyclic CAMP waves during dictyostelium discoideum aggregation), and reaction-diffusion systems (e.g., Turing patterns).

References

Haken H (2004) Synergetics. Introduction and advanced topics. Springer, Berlin
Mikhailov A, Loskutov A (1996) Foundations of synergetics II: chaos and noise. Springer, Berlin
Murray JD (2002) Mathematical biology. Springer, Berlin

Special Sciences

Max Kistler
IHPST, Université Paris 1 Panthéon-Sorbonne, Paris, France

Definition

Special sciences are those sciences whose domain of application is not universal. Thus, the class of special sciences is complementary relative to fundamental physics. Any object existing in space-time falls in the domain of fundamental physics, whereas special sciences deal only with a more or less extended part of those objects. Molecular biology is a special science insofar as diamonds or stars do not belong to its domain of application. The concept of special science has become well known since Fodor's (1974) argument that special sciences are not reducible to fundamental physics.

Cross-References

▶ Causality

References

Fodor JA (1974) Special sciences. In: Representations, repr. as Chapter 5, 1981. MIT Press, Cambridge, MA, pp 127–145

Specialized Metabolic Component Databases

Orland Gonzalez[1] and Alberto Sanguino[2]
[1]Institute for Bioinformatics, Ludwig-Maximilians-University Munich, Munich, Germany
[2]Specialty Division for Systems Biotechnology, Technical University Munich, Garching, Germany

Synonyms

Enzyme-ligand interaction databases; Metabolomics databases; Organism-specific metabolic databases;

Reaction kinetics databases; Thermodynamic databases for enzyme-catalyzed reactions

Definition

In addition to databases that focus on particular components of metabolism, such as metabolites (e.g., KEGG Glycan, ChEBI) or enzymes (e.g., BRENDA), even more specialized resources exist that are devoted to specific aspects of these components. Examples of these are databases that deal with the thermodynamics or kinetics of enzyme-catalyzed reactions, the ligand binding of enzymes, reference spectra for metabolomics experiments, or organism-specific metabolite data (e.g., clinical data).

Specialized Metabolic Component Databases, Table 1 Protein–ligand interaction databases

Database	URL[a]
Relibase	relibase.ccdc.cam.ac.uk
Ligand	ligand-expo.rcsb.org
MSDsite	www.ebi.ac.uk/pdbe-site/pdbemotif/
eF-Site	ef-site.hgc.jp
BindingMOAD	bindingmoad.org
PDBbind	pdbbind.org
AffinDB	agklebe.de/affinity
BindingDB	bindingdb.org
Het-PDB	hetpdbnavi.nagahama-i-bio.ac.jp
BIND	bind.ca
LigAsite	www.bigre.ulb.ac.be/Users/benoit/LigASite/
PROCOGNATE	www.ebi.ac.uk/thornton-srv/databases/procognate/
CREDO	http://www-cryst.bioc.cam.ac.uk/databases/credo

[a] Compiled on March 9, 2011

Characteristics

Enzyme–Ligand Interactions

The molecular recognition of small molecules is critical to many processes in a living cell. In the case of enzymes, this determines substrate specificity and, accordingly, catalytic activity. One of the most useful resources for this type of information (i.e., protein–ligand interactions) is the worldwide protein data bank (PDB) (Berman et al. 2000). The structural data contained therein provides direct evidence for atomic interactions between protein residues and their binding partners. Other databases that deal with protein-ligand interactions are listed in Table 1. Most of them derive their data, at least in part, from the PDB.

One problem with using the PDB directly is that the ligands are frequently noncognate (i.e., not biological/natural). This is because nonphysiological binding partners are often employed simply to aid crystallization. In the case of enzymes, these would be inhibitors, such as transition state analogues (Bashton and Thornton 2009). For this reason, some databases, such as PROCOGNATE, BindingMOAD, and LigAsite, take steps to ensure the biological significance of the observed ligands (e.g., by comparison with reactions defined in KEGG). Indeed, PROCOGNATE, as the name suggests, specializes in making this distinction.

Some of the databases mentioned in Table 1 supplement their protein-ligand interaction data with relevant information. For example, PROCOGNATE and BIND further identify domains. This is important because domains often possess conserved activity, including, potentially, ligand binding. Other examples of databases that supplement their interaction data are PDBbind, BindingDB, and AffinDB, all of which supply binding affinities curated from the literature. A recent review of databases that deal with protein-ligand interactions was made by Bashton and Thornton (2009).

Thermodynamics and Kinetics

Thermodynamic data on enzyme-catalyzed reactions is important for a number of modeling and analysis frameworks used in systems biology. For instance, ▶ metabolic flux analysis (MFA), which is a constraint-based technique used to predict the distribution of fluxes through metabolic networks, uses thermodynamic information to further reduce the range of feasible states that a network can reach (e.g., reaction reversibilities). Furthermore, thermodynamic data can partially replace kinetic information when building dynamic models or performing sensitivity analysis (Goldberg et al. 2004, see ▶ Metabolic Control Analysis).

The largest online collection of thermodynamics data on enzyme-catalyzed reactions can currently be found in the NIST standard reference database, TECRDB (Goldberg et al. 2004). In particular, it provides apparent equilibrium constants and calorimetrically determined molar enthalpies of reaction that were curated from the literature. Each entry in the database contains the appropriate citation, the name of the enzyme used (including its Enzyme Commission number), the experimental conditions under which the data was obtained (e.g., buffers and cofactors), and occasionally some commentary. In addition, TECRDB also provides a subjective evaluation, using a simple scoring scheme from A (high quality) to D (low quality), based on the method of measurement used, the number of data points determined, and the extent to which the effects of temperature, pH, and ionic strength were investigated. Other thermodynamic properties, such as species formation free energies, enthalpies, entropies, and heat capacities, can be derived or estimated from the primary data provided.

In contrast to thermodynamics, which provides information on the equilibrium conditions of products after a reaction takes place, kinetics is concerned with the rate of a reaction and, accordingly, how fast equilibrium is reached. A database that collects extensive kinetic information is the BRENDA (Braunschweig Enzyme Database) enzyme information system (Scheer et al. 2011). Among many other things, it provides experimentally determined K_M (Michaelis constant) and k_{cat} (turnover number) values for substrates or cofactors of enzymes from a wide range of organisms. For molecules that act as inhibitors, it gives the inhibition constant K_i and the half-maximal inhibitory concentration IC_{50}. Kinetic parameters are invaluable for the construction of dynamic models of metabolism (see ▶ Kinetic Modeling and Simulation; ▶ Pathway Modeling, Metabolic and ▶ Ordinary Differential Equation (ODE), Model).

Another database that curates kinetic parameters from the literature is SABIO-RK (Rojas et al. 2007). In contrast to BRENDA, which provides many other types of information related to enzymes, SABIO-RK is devoted to reaction kinetics and was developed specifically with the modeling community in mind. In addition to kinetic constants, the database specifies the type of kinetic law associated with a reaction, if it was defined in the original source. Moreover, SABIO-RK data is structured to facilitate comparison of kinetic parameters obtained under different experimental conditions or from different organisms, tissues, etc. The web-based interface of the database supports export in SBML format (Hucka et al. 2003) (see ▶ Systems Biology Markup Language (SBML)), allowing facile import into simulation and modeling programs that support the standard.

Organism-Specific Metabolism

The human metabolome database (HMDB) is a comprehensive metabolomics resource that contains chemical, physical, as well as clinical data for small molecule metabolites found in the human body (Wishart et al. 2009). Specific data held by HMDB include reference NMR, GC-MS, and MS-MS spectra (MS and NMR), which are invaluable for metabolomics experiments (other databases that provide similar data are BMRB, MMCD, and MassBank), and information on normal and abnormal concentrations of metabolites in a number of biofluids (e.g., blood, cerebrospinal fluid, urine, etc.). Although metabolite records are linked to KEGG pathways, HMDB also provides its own hand-drawn metabolic pathway maps. These are unlike most other online metabolic diagrams in that they are specific to human metabolism, explicitly show the subcellular compartments where specific reactions are known to take place, and allow direct visualization of the chemical structures of participating molecules.

Extensively curated, organism-specific metabolic databases similar to HMDB are currently only available for a few model organisms (e.g., BioCyc; see ▶ Metabolic Networks, Databases). Nevertheless, alternative resources exist in the form of genome-scale metabolic network reconstructions (Reed et al. 2006; Feist et al. 2009; see ▶ Metabolic Networks, Reconstruction). These reconstructed networks provide explicit relationships between the genes of an organism, the enzymes that they encode, and the reactions that occur within a cell. A lot of them are manually curated, and provide citations to relevant literature (e.g., experimental support for the existence of a reaction or pathway). Although reconstruction efforts have focused primarily on microbes (e.g., *E. coli*), high-quality networks are already available for a number of higher organisms, including human, yeast, and *Arabidopsis*. The BiGG database (Schellenberger et al. 2010) attempts to integrate published metabolic reconstructions into one resource using a standard nomenclature.

Cross-References

▶ KEGG Pathway Database
▶ Kinetic Modeling and Simulation
▶ Metabolic Control Analysis

- ▶ Metabolic Flux Analysis
- ▶ Metabolic Networks, Databases
- ▶ Metabolic Networks, Reconstruction
- ▶ Ordinary Differential Equation (ODE), Model
- ▶ Pathway Modeling, Metabolic
- ▶ Systems Biology Markup Language (SBML)

References

Bashton M, Thornton JM (2009) Domain-ligand mapping for enzymes. J Mol Recognit 23:194–208

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN et al (2000) The protein data bank. Nucleic Acids Res 28:235–242

Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol 7:129–143

Goldberg RN, Tewari YB, Bhat TN (2004) Thermodynamics of enzyme-catalyzed reactions – a database for quantitative biochemistry. Bioinformatics 20:2874–2877

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531

Reed JL, Famili I, Thiele I, Palsson BO (2006) Towards multidimensional genome annotation. Nat Rev Genet 7:130–141

Rojas I, Golebiewski M, Kania R, Krebs O, Mir S et al (2007) SABIO-RK: a database for biochemical reactions and their kinetics. BMC Syst Biol 1(Suppl 1):26

Scheer M, Grote A, Chang A, Schomburg I, Munaretto C et al (2011) BRENDA, the enzyme information system in 2011, Nucleic acids research. Nucleic Acids Res 39:D670–D676

Schellenberger J, Park JO, Conrad TM, Palsson BO (2010) BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11:213

Wishart DS, Knox C, Guo AC, Eisner R, Young N et al (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37:D603–D610

Specific Response

Jinzhi Lei
Zhou Pei-Yuan Center for Applied Mathematics,
Tsinghua University of Beijing, Beijing, China

Definition

In interconnected signal transduction pathways, specific response means clearly defined or identified output in response to input signal.

Different cellular signal transduction pathways are often interconnected, and therefore, undesirable cross talk between pathways exists. Specificity and fidelity are two properties that all pathways in a network must possess in order to avoid paradoxical situation where one pathway activates another pathway's output, or responds to another pathway's input, more than its own (Komarova et al. 2005).

In a network with two pathways (X and Y), specificity of a pathway is the ratio of its authentic output to its spurious output (Bardwell et al. 2007; Komarova et al. 2005):

$$S_X = \frac{X_{\text{out}}|X_{\text{in}}}{Y_{\text{out}}|X_{\text{in}}}.$$

The specificity S_X is infinite if the pathway X does not affect outputs from the pathway Y. If $S_X < 1$, it means that the signal for X is actually promoting the output of pathway Y more than its own output. Fidelity of a pathway is defined as its output when given an authentic signal divided by its output in response to a spurious signal (Bardwell et al. 2007; Komarova et al. 2005):

$$F_X = \frac{X_{\text{out}}|X_{\text{in}}}{X_{\text{out}}|Y_{\text{in}}}.$$

A pathway with fidelity larger than 1 (i.e., $F > 1$) means that it is activated by its authentic signal than by others. In contrast, if a pathway has fidelity less than 1 (i.e., $F < 1$), it is activated by another pathway's signal than by its own.

Network specificity is defined as the product of the pathway specificities (the network fidelity, the product of the pathway fidelities, is always equal to network specificity).

Mutual specificity (mutual fidelity) of a network means a property that all pathways in the network have specificity (fidelity) greater than 1.

References

Bardwell L, Zou X, Nie Q, Komarova NL (2007) Mathematical models of specificity in cell signaling. Biophys J 92:3425–3441

Komarova NL, Zou X, Nie Q, Bardwell L (2005) A theoretical framework for specificity in cell signaling. Mol Syst Biol 1:2005.0023

Spectral Count

Stefanie Wienkoop
Department for Molecular Systems Biology,
University of Vienna, Vienna, Austria

Definition

For ▶ mass-spectrometry-based comprehensive quantitative ▶ proteome analysis, the ▶ spectral count emerged as one of the most simple and even so most efficient methods in systems biology. Initially, it has been defined as the sum of all peptide fragment spectra leading to the identification of a protein (Liu et al. 2004). Thus, a changing spectral count is correlated with the relative change of protein abundance between different samples. Due to the development of high precursor mass accuracy mass spectrometer it consequently became possible to define the spectral count as the sum of all fragment spectra of the same precursor mass of a peptide. This technique is also known as “mass accuracy precursor alignment” (MAPA; Hoehenwarter et al. 2008). MAPA allows for the analysis of a relative abundance change of peptides in response to experimental perturbation, which may be detected independent of identification. This means, the spectral count now enables database (genomic sequence information)-independent proteome quantification. One major advantage of this finding is the recognition of unknowns such as post-translational modifications and splice variations not detectable via a database-dependent search. Another important advantage of the MAPA-based spectral count technique is the ability to generate vast data matrices (comparison of hundreds of samples) in a very short time (minutes). This is not possible with any other method so far.

Cross-References

▶ Mass Spectrometry, Proteomics, and Metabolomics
▶ Proteomics, Quantification-Unbiased and Target Approach

References

Hoehenwarter W, van Dongen JT, Wienkoop S, Steinfath M, Hummel J, Erban A, Sulpice R, Regierer B, Kopka J, Geigenberger P, Weckwerth W (2008) A rapid approach for phenotype-screening and database independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. Proteomics 8:4214–4225

Liu H, Sadygov RG, Yates JR 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76:4193–4201

Spectrometry

▶ Spectroscopy and Spectromicroscopy

Spectromicroscope

▶ Spectroscopy and Spectromicroscopy

Spectromicroscopy

Xiaohua Wu
Department of Pediatrics, Herman B Wells Center for Pediatric Research, Indiana University School of Medicine, Indianapolis, IN, USA

Definition

Spectromicroscopy is a combination of two well-established concepts, spectroscopy and microscopy, in an attempt to reach the ultimate goal of obtaining element-specific electronic and magnetic information on the atomic and molecular scale.

Cross-References

▶ Spectroscopy and Spectromicroscopy

Spectroscopy

Xiaohua Wu
Department of Pediatrics, Herman B Wells Center for
Pediatric Research, Indiana University School of
Medicine, Indianapolis, USA

Definition

Spectroscopy was originally the study of the interaction between radiation and matter as a function of wavelength (λ). Recently, the definition has been expanded to include the study of the interactions between particles such as electrons, protons, and ions, as well as their interaction with other particles as a function of their collision energy.

Cross-References

▶ Spectroscopy and Spectromicroscopy

Spectroscopy and Spectromicroscopy

Xiaohua Wu
Department of Pediatrics, Herman B Wells Center
for Pediatric Research, Indiana University School
of Medicine, Indianapolis, IN, USA

Synonyms

Microspectroscopy; Spectrometry; Spectromicroscope

Definition

Spectroscopy was originally the study of the interaction between radiation and matter as a function of wavelength (λ). Spectroscopy was initially referred to the use of visible light dispersed according to its wavelength. More recently, the definition has been expanded to include the study of the interactions between particles such as electrons, protons, and ions, as well as their interaction with other particles as a function of their collision energy. Spectroscopic analysis has been crucial in the development of the most fundamental theories in physics, including quantum mechanics, the special and general theories of relativity, and quantum electrodynamics. Spectroscopy, as applied to high-energy collisions, has been a key tool in developing scientific understanding not only of the electromagnetic force but also of the strong and weak nuclear forces.

Spectromicroscopy is a combination of two well-established concepts, spectroscopy and microscopy, in an attempt to reach the ultimate goal of obtaining element-specific electronic and magnetic information on the atomic and molecular scale.

Characteristics

Common Types of Spectromicroscopy

Absorption Spectroscopy

Absorption spectroscopy is a technique in which the power of a beam of light measured before and after interaction with a sample is compared. Specific absorption techniques tend to be referred to by the wavelength of radiation measured such as ultraviolet, infrared, or microwave absorption spectroscopy. Absorption occurs when the energy of the photons matches the energy difference between two states of the material.

Fluorescence Spectroscopy

Fluorescence spectroscopy uses higher-energy photons to excite a sample, which will then emit lower energy photons. This technique has become popular for its biochemical and medical applications, and can be used for confocal microscopy, fluorescence resonance energy transfer, and fluorescence lifetime imaging.

X-ray Spectroscopy

When X-rays of sufficient frequency (energy) interact with a substance, inner shell electrons in the atom are excited to outer empty orbitals, or they may be removed completely, ionizing the atom. The inner shell “hole” will then be filled by electrons from outer orbitals. The energy available in this de-excitation process is emitted as radiation (fluorescence) or will remove other less-bound electrons from the atom (Auger effect). The absorption or emission frequencies (energies) are characteristic of the specific atom. In addition, for a specific atom, small frequency (energy) variations that are characteristic of the chemical bonding occur. With a suitable apparatus, these characteristic X-ray frequencies or

Auger electron energies can be measured. X-ray absorption and emission spectroscopy is used in chemistry and material sciences to determine elemental composition and chemical bonding.

Flame

Liquid solution samples are aspirated into a burner or nebulizer/burner combination, desolvated, atomized, and sometimes excited to a higher-energy electronic state.

Visible and Ultraviolet

Many atoms emit or absorb visible light. In order to obtain a fine line spectrum, the atoms must be in a gas phase. The spectrum is studied in absorption or emission. Visible absorption spectroscopy is often combined with UV absorption spectroscopy in UV/Vis spectroscopy. All atoms absorb in the Ultraviolet (UV) region because these photons are energetic enough to excite outer electrons. If the frequency is high enough, photoionization takes place. UV spectroscopy is also used in quantifying protein and DNA concentration as well as the ratio of protein to DNA concentration in a solution.

Infrared Spectroscopy

Infrared spectroscopy offers the possibility to measure different types of interatomic bond vibrations at different frequencies. Especially in organic chemistry the analysis of IR absorption spectra shows what types of bonds are present in the sample. It is also an important method for analyzing polymers and constituents like fillers, pigments, and plasticizers.

Near-infrared Spectroscopy

The near-infrared NIR range, immediately beyond the visible wavelength range, is especially important for practical applications because of the much greater penetration depth of NIR radiation into the sample than in the case of mid-IR spectroscopy range. This allows also large samples to be measured in each scan by NIR spectroscopy, and is currently employed for many practical applications such as: rapid grain analysis, medical diagnosis pharmaceuticals/medicines, biotechnology, genomics analysis, proteomic analysis, interactomics research, inline textile monitoring, food analysis and chemical imaging/hyperspectral imaging of intact organisms, plastics, textiles, insect detection, forensic lab application, crime detection, and various military applications. Interpretation of Near-Infrared Spectra is important for chemical identification.

Raman Spectroscopy

Raman spectroscopy is based on the absorption of photons of a specific frequency followed by scattering at a higher or lower frequency. The modification of the scattered photons results from the incident photons either gaining energy from or losing energy to the vibrational and rotational motion of the molecule. The resulting “fingerprints” are an aid to analysis.

Coherent Anti-Stokes Raman spectroscopy (CARS)

CARS is a recent technique that has high sensitivity and powerful applications for in vivo spectroscopy and imaging.

Nuclear Magnetic Resonance

Nuclear magnetic resonance spectroscopy analyzes the magnetic properties of certain atomic nuclei to determine different electronic local environments of hydrogen, carbon, or other atoms in an organic compound or other compound. This is used to help determine the structure of the compound.

Photoemission Mössbauer

Transmission or conversion-electron (CEMS) modes of Mössbauer spectroscopy probe the properties of specific isotope nuclei in different atomic environments by analyzing the resonant absorption of characteristic energy gamma-rays known as the Mössbauer effect.

Spectromicroscopy

Fluorescence microscope is an optical microscope used to study properties of organic or inorganic substances using the phenomena of fluorescence and phosphorescence instead of, or in addition to, reflection and absorption.

In most cases, a component of interest in the specimen can be labeled specifically with a fluorescent molecule called a fluorophore. The specimen is illuminated with light of a specific wavelength (or wavelengths) which is absorbed by the fluorophores, causing them to emit light of longer wavelengths (i.e., of a different color than the absorbed light). The illumination light is separated from the much weaker emitted fluorescence through the use of a spectral emission filter. Typical components of a fluorescence microscope are the light source (xenon arc lamp or mercury-vapor lamp), the excitation filter, the dichroic mirror (or dichromatic beam splitter), and the emission filter. The filters and the dichroic are chosen to match the spectral excitation and emission characteristics of the fluorophore used to label the specimen. In this manner, the distribution of a single fluorophore (color) is imaged at a time. Multicolor images of several types of fluorophores must be composed by combining several single-color images.

Most fluorescence microscopes in use are epifluorescence microscopes (i.e., excitation and observation of the fluorescence are from above (epi–) the specimen). These microscopes have become an important part in the field of biology, opening the doors for more advanced microscope designs, such as the confocal microscope.

Epifluorescence microscopy is a method of fluorescence microscopy that is widely used in life sciences. The excitatory light is passed from above (or, for inverted microscopes, from below), through the objective lens and then onto the specimen instead of passing it first through the specimen. The fluorescence in the specimen gives rise to emitted light which is focused to the detector by the same objective that is used for the excitation. Since most of the excitatory light is transmitted through the specimen, only reflected excitatory light reaches the objective together with the emitted light and this method therefore gives an improved signal-to-noise ratio. An additional filter between the objective and the detector can filter out the remaining excitation light from fluorescent light. A common use in biology is to apply fluorescent or fluorochrome stains to the specimen in order to image distributions of proteins or other molecules of interest.

Confocal and Multiphoton Microscopy

The most widely used technologies for in vivo microscopy of tissues are confocal and two- or multiphoton (2P/MP) microscopy. In contrast to conventional microscopy that requires thin tissue sections, confocal and 2P/MP microscopy can achieve up to diffraction-limited resolution when imaging virtual tissue sections in intact tissue volumes, which is an aspect of these methods referred to as "tissue sectioning" or "optical sectioning." In particular, laser-scanning confocal microscopy scans a focused laser beam inside the specimen and uses a pinhole to reject photons that arrive to the detector from out-of-focus areas; these have been typically scattered multiple times and contribute to image blurring. Although the pinhole rejects a large part of the photons, sufficient signal from the focal point can be detected using high-intensity light sources and sensitive detectors. Two-dimensional "tissue sections" are formed by scanning the focused beam over a plane in the sample imaged and piecing together information from each individual focal area. As information is collected only from the laser focal spot at each time point, single element detectors such as photomultiplier tubes are used. By focusing the beam at different tissue depths, three-dimensional images can also be generated. Typical imaging is restricted to depths of a few MFPs owing to diminishing confocal signal with increasing depth, primarily because of scattering. Whereas 2P/MP microscopy also uses laser-scanning principles, focused femtosecond laser pulses are used for illumination. By concentrating the beam energy in space (focusing) and in time (ultrafast pulses) substantial signal can be generated based on 2P/MP absorption but only within the spatially confined area of the focus point. All fluorescence photons generated therefore come from a highly localized volume. By collecting light generated from scanning a laser beam over an area of interest, one can piece together two- or three-dimensional images as in confocal microscopy. In this case, the photons collected have generally been scattered multiple times because no pinhole is used. In two-photon microscopy, two near-infrared photons (for example, 900 nm) can excite a fluorochrome in the visible spectrum (for example, 450 nm). By collecting all the available light and by using near-infrared excitation light, which is attenuated less than the visible light used for excitation in confocal microscopy, higher penetration can be achieved in two-photon compared to confocal microscopy. Typical two-photon setups usually achieve worse resolution than that of confocal microscopes because the diffraction-limited focal spots widen as the illumination wavelength increases. When expressed in terms of tissue penetration depth in physical units (millimeters), the depth of two-photon microscopy (which operates in the near-infrared spectral region) is reported as 2–3 times deeper than confocal microscopy (which operates in the visible range). However, this difference is not markedly different when expressed in MFP terms because the MFP is longer for near-infrared than for visible light.

Together, confocal and 2P/MP microscopy have been used extensively for in vivo imaging of fluorescent proteins, probes, or dyes to investigate structure, function, and molecular events as they occur in unperturbed environments. Two-photon imaging currently defines the upper limit of penetration depth in diffraction-limited microscopy, achieving depths of about half a TMFP.

Soft x-ray Microscopy

Soft x-ray microscopy provides a unique set of capabilities in-between those of visible light and electron microscopy. It has long been recognized that nature provides a "water window" spectral region between the *K* shell x-ray absorption edges of carbon ($\sim$290 eV) and oxygen ($\sim$540 eV), where organic materials show strong absorption and phase contrast, while water is relatively nonabsorbing. This enables imaging of specimens that are several microns thick with high intrinsic contrast using x-rays with a wavelength of 2–4 nm. In recent years, the promise of high-resolution imaging has been realized, thanks to advances in x-ray sources, focusing optics, and specimen-preparation methods, and several microscopes can now image biological specimens at about 30-nm resolution. New developments are adding the capability for labeling, for spectroscopic mapping of various biochemical components of cells and tissues at high spatial resolution, and for mapping of trace (low-Z) elements at submicron resolution with unprecedented sensitivity using higher-energy x-rays. Soft x-rays have a photon energy between about 100 and 1,000 eV, corresponding to very short wavelengths. At these photon energies, there is essentially no inelastic or plural elastic scattering, making quantitative analysis of images especially favorable. It is possible to image single cells through water layers up to $\sim$10 μm thick, but the required photon exposure for imaging at 30-nm resolution leads to cells receiving a radiation dose of 108–1,010 rads, ruling out repeated imaging of live specimens. Some particularly robust specimens can be imaged with no special treatment, but other specimens show immediate morphological changes and/or mass loss.

Raman Spectromicroscopy

Raman spectromicroscopy offers several advantages for microscopic analysis. Since it is a scattering technique, specimens do not need to be fixed or sectioned. Raman spectra can be collected from a very small volume ($<$1 μm in diameter); these spectra allow the identification of species present in that volume. Water does not generally interfere with Raman spectral analysis. Thus, Raman spectroscopy is suitable for the microscopic examination of minerals, materials such as polymers and ceramics, cells, and proteins. A Raman microscope begins with a standard optical microscope, and adds an excitation laser, a monochromator, and a sensitive detector (such as a charge-coupled device (CCD), or photomultiplier tube (PMT)). FT-Raman has also been used with microscopes.

In direct imaging, the whole field of view is examined for scattering over a small range of wavenumbers (Raman shifts). For instance, a wavenumber characteristic for cholesterol could be used to record the distribution of cholesterol within a cell culture.

The other approach is hyperspectral imaging or chemical imaging, in which 1,000 of Raman spectra are acquired from all over the field of view. The data can then be used to generate images showing the location and amount of different components. Taking the cell culture example, a hyperspectral image could show the distribution of cholesterol, as well as proteins, nucleic acids, and fatty acids. Sophisticated signal- and image-processing techniques can be used to ignore the presence of water, culture media, buffers, and other interferents.

By using Raman microspectroscopy, in vivo time- and space-resolved Raman spectra of microscopic regions of samples can be measured. As a result, the fluorescence of water, media, and buffers can be removed. Consequently in vivo time- and space-resolved Raman spectroscopy is suitable to examine proteins, cells, and organs.

Raman microscopy for biological and medical specimens generally uses near-infrared (NIR) lasers. This reduces the risk of damaging the specimen by applying higher-energy wavelengths. However, the intensity of NIR Raman is low, and most detectors required very long collection times. Recently, more sensitive detectors have become available, making the technique better suited to general use. Raman microscopy of inorganic specimens, such as rocks and ceramics and polymers, can use a broader range of excitation wavelengths.

Fourier Transform Infrared Microspectroscopy (FT-IR)

Fourier transform IR (FT-IR) microspectroscopy, in which the spectrometer is coupled to a light microscope, first introduced in the 1940s, and applied to bone in the 1980s, enabled investigators to examine spectra at discrete points within thin sections of tissues. FTIR spectroscopy is a long-established and invaluable technique, which is based on the principle that molecules absorb mid-IR radiation, yielding richly structured IR absorption spectra. The use of apertures in single point detection methods with a global source allows routine spatial resolutions of the order of 100 µm. Enhanced spatial resolution can be achieved using a synchrotron radiation source, which is order of magnitudes brighter in the IR range. This increased brightness greatly improves the signal-to-noise ratio and allows very high-quality spectra to be acquired. This point-by-point mapping has been widely applied to the analysis of polymers and to a variety of tissues and individual cells to learn more about spatial variation in tissue and cellular composition. It has also been used to identify pathologic inclusions such as foreign matter and unusual soft tissue calcifications.

References

Ade H (1998) Experimental methods in the physical sciences, vol 32. Academic, USA

Breusegem SY, Levi M, Barry NP (2006) Fluorescence correlation spectroscopy and fluorescence lifetime imaging microscopy. Nephron Exp Nephrol 103:e41–e49

Ellis DI, Goodacre R (2006) Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. Analyst 131(8):875–885

Grude O, Hammiche A, Pollock H, Bentley AJ, Walsh MJ, Martin FL, Fullwood NJ (2007) Near-field photothermal microspectroscopy for adult stem-cell identification and characterization. J Microsc 228(Pt 3):366–372

Jacobsen C (1999) Soft x-ray microscopy. Trends Cell Biol 9:44–47

McKellar ARW (2010) High-resolution infrared spectroscopy with synchrotron sources. J Mol Spectrosc 262:1–10

Ntziachristos V (2010) Going deeper than microscopy: the optical imaging frontier in biology. Nat Meth 7(8):603

Visible and Ultraviolet Spectroscopy, spectroscopynow.com/coi/cda/detail

Walsh MJ, German MJ, Singh M, Pollock HM, Hammiche A, Kyrgiou M, Stringfellow HF, Paraskevaidis E, Martin-Hirsch PL, Martin FL (2007) IR microspectroscopy: potential applications in cervical cancer screening. Cancer Lett 246:1–11

Spindle Checkpoint

▶ Cell Cycle Checkpoints

Spindle Pole Body

Sebastian Mana-Capelli and Dannel McCollum
Department of Molecular Genetics and Microbiology,
University of Massachusetts, Worcester, MA, USA

Synonyms

Microtubule organizing center (or MTOC); Yeast centrosome

Definition

The yeast spindle pole body (or SPB) is the analogous organelle to the metazoan centrosome. The SPB is the major nucleation site for microtubules in mitosis and forms the ends of the mitotic spindle. As in animal cells, γ-TUBULIN localizes to the MTOC and directs microtubule nucleation. Fission yeast have multiple MTOCs during interphase in addition to the SPB; therefore, SPBs only play a major role in microtubule nucleation during mitotic spindle formation in this organism. SPB architecture differs from the one of the metazoan centrosome. Instead of the orthogonally arranged pair of centrioles surrounded by pericentrosomal material, yeast SPBs are multilayered structures that remain associated to the nuclear membrane at all times. Interestingly, the characteristics of the SPB cycle greatly differ between budding and fission yeast. The SPB of budding yeast remains embedded in the nuclear envelope throughout the cell cycle and divides in G1 by de novo formation of a SPB satellite (Yoder et al. 2003). In contrast, fission yeast SPB divide through a semiconservative mechanism that conceptually resembles the one of higher eukaryotes. Fission yeast SPBs are also attached to the nuclear membrane and only ingress into it during spindle formation (West et al. 1997).

Cross-References

▶ Cytokinesis

References

West DR, Morphew RR, Oakley DM, McIntosh JRBR (1997) The spindle pole body of *Schizosaccharomyces pombe* enters and leaves the nuclear envelope as the cell cycle proceeds. Mol Biol Cell 8:1461–1479

Yoder TJ, Pearson CG, Bloom K, Davis TN (2003) The *Saccharomyces cerevisiae* spindle pole body is a dynamic structure. Mol Biol Cell 14:3494–3505

SRNA Databases

▶ Non-coding RNA Databases

SRP

▶ Signal Recognition Particle

SSA

▶ Stochastic Simulation Algorithm

SsrA

▶ TmRNA

S-System

Rui-Sheng Wang
Department of Physics, Pennsylvania State University,
University Park, PA, USA

Synonyms

Canonical nonlinear modeling; Power-law formalism

Definition

S-system is a quantitative nonlinear model based on power-law functions. It is characterized by a good compromise between approximate accuracy and mathematical flexibility. In systems biology, S-system is a specific kind of ordinary differential equations with a simple canonical form. It has a rich structure capable of capturing complex dynamics of many biochemical systems, such as gene regulatory networks, signal transduction networks, and metabolic networks. Especially, S-system models have been widely used to infer gene regulatory networks from time-course microarray data.

Given n genes of interest, the S-system for modeling a gene regulatory network is described by:

$$\frac{dx_i}{dt} = \alpha_i \prod_{j=1}^{n} x_j^{g_{i,j}} - \beta_i \prod_{j=1}^{n} x_j^{h_{i,j}} \qquad (1)$$

where x_i is the mRNA concentration of gene i, α_i and β_i are the positive rate constants. $g_{i,j}$ and $h_{i,j}$ are the exponential parameters called kinetic orders. $g_{i,j} > 0$ indicates that gene j activates the expression of gene i, and $g_{i,j} < 0$ indicates that gene j inhibits the expression of gene i. $h_{i,j}$ has the opposite effects on controlling gene expression compared to $g_{i,j}$. In the inference of gene regulatory networks by using S-systems, a main task is to estimate the parameters α_i, β_i, $g_{i,j}$, and $h_{i,j}$ based on experimental microarray data.

Cross-References

▶ Linear Model

References

Savageau MA (1991) 20 years of S-systems. In: Voit EO (ed) Canonical nonlinear modeling: S-system approach to understand complexity. van Nostrand Reinhold, New York

Voit EO (2000) Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists. Cambridge University Press, Cambridge

Stability

Tianshou Zhou
School of Mathematics and Computational Sciences,
Sun Yet-Sen University, Guangzhou, Guangdong,
China

Definition

The *stability* of an orbit of a dynamical system characterizes whether nearby (i.e., perturbed) orbits will remain in a neighborhood of that orbit or be repelled away from it. *Asymptotic stability* additionally characterizes attraction of nearby orbits to this orbit in the long time limit. The distinct concept of structural stability is treated elsewhere, and concerns changes in the family of all solutions due to perturbations to the functions defining the dynamical system.

Characteristics

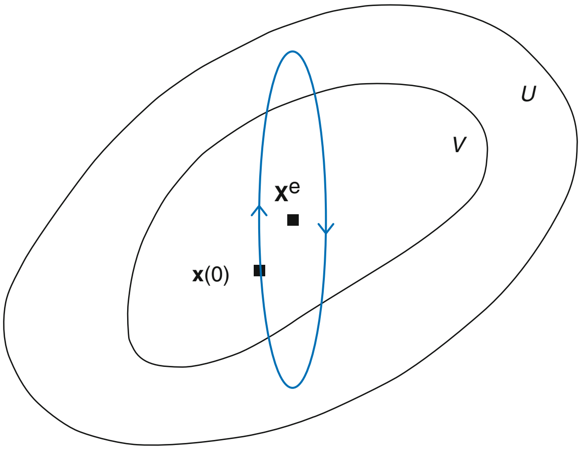
Equilibrium Point

Consider a set of coupled autonomous ordinary differential equations (ODEs) that can describe the time evolution of key components in a gene regulatory network, written in vector notation as:

$$\frac{dx}{dt} = \dot{x} = F(x), x \in R^n \tag{1}$$

$$\text{where } F(x) = \begin{pmatrix} F_1(x_1, x_2, \cdots, x_n) \\ F_2(x_1, x_2, \cdots, x_n) \\ \vdots \\ F_n(x_1, x_2, \cdots, x_n) \end{pmatrix}, \; x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Denote by $x(t)$ a solution of Eq. 1 satisfying initial conditions: $x_0 = x(0)$. Equilibria (sometimes called equilibrium points or fixed points or steady states) x^e are special constant solutions $x(t) = x^e$, where $F(x^e) = 0$ or $F_j(x_1^e, x_2^e, \cdots, x_n^e) = 0$, $j = 1, 2, \cdots, n$. For the system of a gene regulatory network, $F(x)$ can often be expressed as $F(x) = f(x) - g(x)$, where $f(x)$ represents the production part whereas $g(x)$ does the degradation part, so the equilibrium x^e satisfies $f(x^e) = g(x^e)$.



Stability, Fig. 1 Lyapunov stability

Below, we first treat the stability of equilibria and stability analysis, and then mention extensions to the stability of more general solutions.

Stability of an Equilibrium

Lyapunov Stability

x^e is a stable equilibrium if for every neighborhood U of x^e, there is a neighborhood $V \subseteq U$ of x^e such that every solution $x(t)$ starting in V, i.e., the solution starting from $x_0 \in V$, remains in U for all $t \geq 0$. Note that $x(t)$ need not approach x^e.

x^e is unstable if it is not stable.

Asymptotic Stability

An equilibrium x^e is asymptotically stable if it is Lyapunov stable and additionally V can be chosen so that $|x(t) - x^e| \to 0$ as $t \to \infty$ for all $x_0 \in V$.

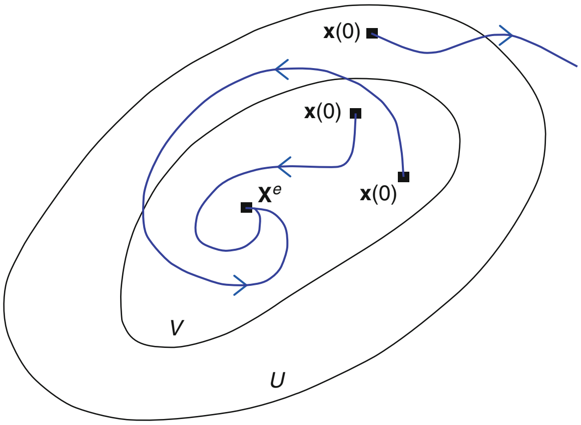
An equilibrium that is Lyapunov stable but not asymptotically stable is sometimes called neutrally stable. See Figs. 1 and 2 for illustrations.

Note: Lyapunov stability and asymptotic stability are of local stability.

Stable Region

Definition: If there is a region V of the equilibrium x^e such that a solution of Eq. 1 starting from every point x_0 in V, denoted by $x(t)$, is convergent, i.e., $|x(t) - x^e| \to 0$ as $t \to \infty$, then V is called a stable region or an attracting basin of x^e.

Note 1: The boundary of the stable region of an equilibrium is possibly complicated and is not easy to be determined.



Stability, Fig. 2 Asymptotic stability

Note 2: If V equates just the definition region of the system Eq. 1, then we say that x^e is globally stable.

Note 3: Completely similarly, we can give definitions for the stable region of an orbit and global stability.

Stability Analysis

Linearization

Suppose that $x = x^e$ is an equilibrium, implying that if $x(0) = x^e$, then $x(t) \equiv x^e$. For a small perturbation of x^e, denoted by $\xi(t) = (\xi_1, \xi_2, \cdots, \xi_n)$ with $|\xi(t)| \ll 1$, let $x(t) = x^e + \xi(t)$. Substitute the expression of $x(t)$ into both sides of Eq. 1 and expand function f in a multivariable, vector-valued Taylor series (we assume that f is sufficiently differentiable so that Taylor's Theorem with remainder applies to each component) to obtain

$$\begin{aligned}\dot{x}^e + \dot{\xi} &= F(x^e + \xi)\\ &= F(x^e) + DF(x^e)\xi + O\left(|\xi|^2\right) \qquad (2)\end{aligned}$$

where $DF(x^e)$ denotes the $n \times n$ Jacobian matrix of partial derivatives $\left[\partial F_i / \partial x_j\right]$, evaluated at the equilibrium x^e, and $O\left(|\xi|^2\right)$ denotes terms of quadratic and higher order in the components $\xi_1, \xi_2, \cdots, \xi_n$. Thus, for small enough $|\xi|$, the first term $DF(x^e)\xi$ dominates. Taking into account that $\dot{x}^e$ and $F(x^e)$ vanish and ignoring the small term $O\left(|\xi|^2\right)$, we obtain the linear system:

$$\dot{\xi} = DF(x^e)\xi \qquad (3)$$

which is called the linearization of Eq. 1 at x^e. It can be solved by standard methods (Boyce and DiPrima 1997).

The general solution $\xi(t)$ of Eq. 3 is determined by the eigenvalues and eigenvectors of the Jacobian matrix $DF(x^e)$. However, we are usually concerned with qualitative properties rather than complete solutions. In particular in studying stability, we want to know whether the size (norm) of solutions grows, stays constant, or shrinks as $t \to \infty$. This can be answered just by examining the eigenvalues.

Recall that if λ is a real eigenvalue with the eigenvector υ, then there is a solution to the linearized equation of the form: $\xi(t) = e^{\lambda t}\upsilon$; if $\lambda = \alpha \pm i\beta$ is a complex conjugate pair with eigenvectors $\upsilon = u \pm iw$ (u and w are real), then $\xi_1(t) = e^{\alpha t}(u \cos \beta t - w \sin \beta t)$ and $\xi_2(t) = e^{\alpha t}(u \sin \beta t + w \cos \beta t)$ are two linearly independent solutions. In both cases, the real part of λ (almost) determines stability. Since any solution of the linearized equation can be written as the linear superposition of terms of these forms (except for the case of multiple eigenvalues), we can deduce the following:

- If all eigenvalues of $DF(x^e)$ have strictly negative real parts, then $|\xi(t)| \to 0$ as $t \to \infty$ for all solutions.
- If at least one eigenvalue of $DF(x^e)$ has positive real parts, then there is a solution with $|\xi(t)| \to +\infty$ as $t \to \infty$.
- If some pairs of complex conjugate eigenvalues have zero real parts with distinct imaginary parts, then the corresponding solutions for $|\xi(t)| \to +\infty$ oscillate and neither decay nor grow as $t \to \infty$.

Note 1: The eigenvalues of the linearization are preserved under (smooth) changes of coordinates (Arnold 1973).

Note 2: When multiple eigenvalues exist and there are not enough linearly independent eigenvectors to span R^n, solutions behave like $|\xi(t)| \sim t^k e^{\lambda t}$, so that they still decay for sufficiently long times If $\lambda < 0$ and grow if $\lambda > 0$.

Note 3: The form $t^k e^{\lambda t}$ implies that *transient growth* occurs over initial times even if $\lambda < 0$.

This can also occur in the case of distinct eigenvalues. See Trefethen and Embree (2005) for more on this, but consider the example

$$\begin{cases}\dot{\xi}_1 = -2\xi_1 + \alpha\xi_2\\ \dot{\xi}_2 = -\xi_2\end{cases} \qquad (4)$$

for large $|\alpha|$. This system has eigenvalues -1 and -2. However, taking $\xi_1(0) = 0$ and $\xi_2(0) = 1$, the first coordinate $\xi_1(t) = \alpha(e^{-t} - e^{-2t})$ initially grows from zero to a maximum value of $\alpha/4$. For sufficiently large α, the growth of ξ_1 will initially overwhelm the decay of $\xi_2 = e^{-t}$ so that the trajectory transiently moves farther from the fixed point before approaching it as $t \to \infty$. This also illustrates the need for the two neighborhoods U and V in the definitions of stability.

This motivates us to introduce the following concept.

Hyperbolic Equilibria

Definition: x^e is a *hyperbolic* or non-degenerate equilibrium if all the eigenvalues of $DF(x^e)$ have nonzero real parts.

Equipped with the linear analysis sketched above, and recognizing that the remainder terms ignored in passing from Eqs. 2 to 3 can be made as small as we wish by selecting a sufficiently small neighborhood of x^e, we can determine the stability of *hyperbolic* equilibria from their linearization:

Proposition: If x^e *is an equilibrium of* $\dot{x} = F(x)$ *and all the eigenvalues of the Jacobian matrix* $DF(x^e)$ *have strictly negative real parts, then* x^e *is exponentially (and hence asymptotically) stable. If at least one eigenvalue has strictly positive real part, then* x^e *is unstable.*

Moreover, the Hartman-Grobman Theorem says that the full nonlinear system Eq. 1 is topologically equivalent to the linearized system Eq. 3 in a small neighborhood of a hyperbolic equilibrium.

Borrowing from fluid mechanics, we say that if all nearby solutions approach an equilibrium (e.g., all eigenvalues have negative real parts), it is a *sink*; if all nearby solutions recede from it, it is a *source*, and if some approach and some recede, it is a *saddle point*. When the equilibrium is surrounded by nested closed orbits, we call it a *center*.

Degenerate Equilibria

One might hope to claim that Lyapunov stability (per the definition above) holds even if (some) eigenvalues have zero real part, but the following counterexamples demonstrate that this is not the case:

Example 1.
Consider

$$\dot{x} = \alpha x^3, \alpha \neq 0 \tag{5}$$

Here $x = 0$ is the equilibrium and the linearization at 0 is

$$\dot{\xi} = \left(3\alpha x^2\big|_{x=0}\right)\xi = 0 \tag{6}$$

with solution $\xi(t) = \xi(0) = cons\tan t$, so certainly $x = 0$ is Lyapunov stable for Eq. 6, but not asymptotically stable.

The exact solution of the nonlinear ODE Eq. 5 may be found by separating variables:

$$\int_{x(0)}^{x(t)} \frac{dx}{x^3} = \int \alpha dt \Rightarrow x(t) = \frac{x(0)}{\sqrt{1 - 2\alpha x^2(0)t}}$$

We therefore deduce that

$|x(t)| \to \infty$ as $t \to \frac{1}{2\alpha x^2(0)}$ if $\alpha > 0$ (blowup! Instability)
$|x(t)| \to 0$ as $t \to \infty$ if $\alpha < 0$ (asymptotic stability)

The linearized system Eq. 6 is *degenerate* and the nonlinear "remainder terms," ignored in our linearized analysis, determine the outcome in this case. Here it is obvious, at least in retrospect, that ignoring these terms is perilous, since while they are indeed $O(\xi^2)$ (in fact, $O(\xi^2)$), the linear $O(\xi)$ term is identically zero!

Example 2.
Consider the two-dimensional system:

$$\begin{cases} \dot{x} = y + \alpha(x^2 + y^2)x \\ \dot{y} = -x + \alpha(x^2 + y^2)y \end{cases}$$

Note that the linearization is simply a harmonic oscillator with eigenvalues $\pm i$. Is the equilibrium $(x, y) = (0, 0)$ of this system stable or unstable? To answer this, it is convenient to transform to polar coordinates $x = r\cos\theta$, $y = r\sin\theta$, which gives the uncoupled system:

$$\dot{r} = \alpha r^3, \ \dot{\theta} = -1$$

The first equation is as in the example above, so we conclude: $\alpha > 0 \Rightarrow$ unstable; $\alpha = 0 \Rightarrow$ stable; $\alpha < 0 \Rightarrow$ asymptotically stable. The linearization gives no information if $\alpha = 0$.

How can we prove stability in such degenerate cases, in which one or more eigenvalues has zero real part? One method requires construction of a function, often called a Lyapunov function, which remains constant, or decreases, along solutions. For mechanical systems, the total (kinetic plus potential) energy is often a good candidate. This allows one to prove stability and even asymptotic stability in certain cases, via describe Lyapunov's second method or direct method:

Theorem (Hirsch et al. (2004))*: Suppose that* $dx/dt = \dot{x} = F(x)$ *has an isolated equilibrium at* $x = 0$ *(without loss of generality one can move an equilibrium* x^e*to* 0 *by letting* $y = x - x^e$*). If there exists a differentiable function* $V(x)$*, which is positive definite in a neighborhood of* 0 *(in the sense that* $V(0) = 0$ *and* $V(x) > 0$ *for* $x \neq 0$*) and for which* $dV/dt = \nabla V \cdot F$*is negative definite on some domain* D *containing* 0*, then* 0 *is asymptotically stable. If* dV/dt *is negative semidefinite (*i.e. $dV/dt = 0$ *is allowed), then* 0 *is Lyapunov stable.*

Stability of General Orbits

Definitions

The notions of stability may be generalized to nonconstant orbits (periodic, quasiperiodic, or nonperiodic) of ODEs.

First, we give some definitions and notation. Let $\gamma_t(t) = x(t)$, given the initial value $x(0) = x$ Then, the (forward) *orbit* is the set of all values that this trajectory obtains: $\gamma(x) = \{\gamma_t(x) | t \geq 0\}$. Next, we have the following:

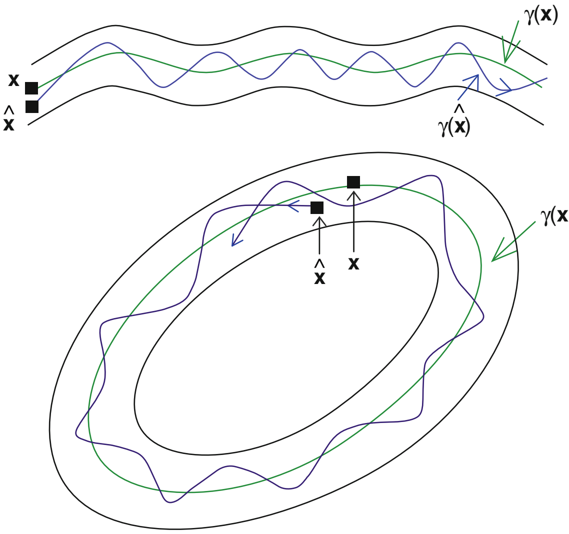
Definition: Two orbits $\gamma(x)$ and $\gamma(\hat{x})$ are *ε-close* if there is a reparameterization of time (a smooth, monotonic function) $\hat{t}(t)$ such that $\left|\gamma_t(x) - \gamma_{\hat{t}(t)}(\hat{x})\right| < \varepsilon$ for all $t \geq 0$.

We say that an orbit is orbitally stable if all orbits with nearby initial points remain close in this sense:

Definition: An orbit $\gamma(x)$ is *orbitally stable* if, for any $\varepsilon > 0$, there is a neighborhood V of x so that, for all $\hat{x}$ in V, $\gamma(x)$ and $\gamma(\hat{x})$ are ε-close.

Definition: If additionally V may be chosen so that, for all $\hat{x} \in V$, there exists a constant $\tau(\hat{x})$ so that $\left|\gamma_t(x) - \gamma_{\hat{t}-\tau(\hat{x})}(\hat{x})\right| \to 0$ as $t \to \infty$. Then $\gamma_t(x)$ is *asymptotically stable*.

See Fig. 3, which show (a segment of) the orbit $\gamma(x)$ as well as a neighboring orbit $\gamma(\hat{x})$. The black lines indicate the boundary of an ε neighborhood of $\gamma(x)$.



Stability, Fig. 3 Orbital stability

Floquet Theory

Floquet theory (Chicone 1999; Floquet (1883)) is a branch of the theory of ordinary differential equations relating to the class of solutions to linear differential equations of the form

$$\dot{x} = A(t)x \tag{7}$$

with $A(t)$ being a continuous periodic function with period T.

The main theorem of Floquet theory, *Floquet's theorem*, due to Gaston Floquet (1883), gives a canonical form for each fundamental matrix solution of this common linear system. It gives a coordinate change $y = Q^{-1}(t)x$ with $Q(t + 2T) = Q(t)$ that transforms the periodic system to a traditional linear system with constant, real coefficients.

In solid-state physics, the analogous result (generalized to three dimensions) is known as Bloch's theorem.

Note that the solutions of the linear differential equation form a vector space. A matrix $\Phi(t)$ is called a fundamental matrix solution if all columns are linearly independent solutions. A matrix $\Psi(t)$ is called a principal fundamental matrix solution if all columns are linearly independent solutions and there exists t_0 such that $\Psi(t_0)$ is the identity. A principal fundamental matrix can be constructed from a fundamental matrix using $\Psi(t) = \Phi(t)\Phi^{-1}(t_0)$. The solution of the linear differential equation with the initial condition

$x(0) = x_0$ is $x(t) = \Phi(t)\Phi^{-1}(0)x_0$, where $\Phi(t)$ is any fundamental matrix solution.

Floquet's Theorem If $\Phi(t)$ is a fundamental matrix solution of the periodic system $\dot{x} = A(t)x$, with $A(t)$ a periodic function with period T, then for all $t \in R$,

$$\Phi(t+T) = \Phi(t)\Phi^{-1}(0)\Phi(T) \tag{8}$$

In addition, for each matrix B (possibly complex) such that $e^{TB} = \Phi^{-1}(0)\Phi(T)$, there is a periodic (period T) matrix function $t \mapsto P(t)$ such that $\Phi(t) = P(t)e^{tB}$ for all $t \in R$. Also, there is a *real* matrix S and a *real* periodic (period $2T$) matrix function $t \mapsto Q(t)$ such that $\Phi(t) = Q(t)e^{tS}$ for all $t \in R$.

Consequences and Applications This mapping $\Phi(t) = Q(t)e^{tS}$ gives rise to a time-dependent change of coordinates ($y = Q^{-1}(t)x$), under which our original system becomes a linear system with real constant coefficients $\dot{y} = Sy$. Since $Q(t)$ is continuous and periodic, it must be bounded. Thus, the stability of the zero solution for $y(t)$ and $x(t)$ is determined by the eigenvalues of S.

The representation $\Phi(t) = P(t)e^{tB}$ is called a *Floquet normal form* for the fundamental matrix $\Phi(t)$.

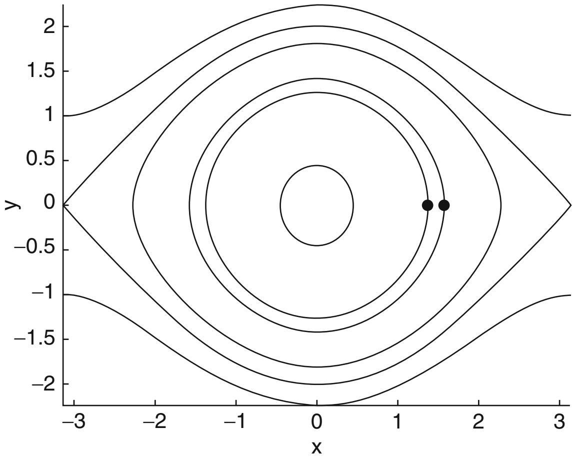
The eigenvalues of e^{TB} are called the characteristic multipliers of the system. They are also the eigenvalues of the (linear) Poincaré maps $x(t) \to x(t+T)$. A *Floquet exponent* (sometimes called a characteristic exponent) is a complex μ such that $e^{\mu T}$ is a characteristic multiplier of the system. Notice that Floquet exponents are not unique, since $e^{(\mu+(2\pi ik/T))T} = e^{\mu T}$, where k is an integer. The real parts of the Floquet exponents are called Lyapunov exponents. The zero solution is asymptotically stable if all Lyapunov exponents are negative, Lyapunov stable if the Lyapunov exponents are nonpositive and unstable otherwise.

- Floquet theory is very important for the study of dynamical systems.
- Floquet theory shows stability in Hill's equation (introduced by George William Hill) approximating the motion of the moon as a harmonic oscillator in a periodic gravitational field.

Examples

Example 1: The nonlinear pendulum

Consider the pendulum equations



Stability, Fig. 4 Orbital stability for the nonlinear pendulum

$$\begin{cases} \dot{x} = y \\ \dot{y} = -\sin x \end{cases}$$

Orbits lie on the energy level sets shown in Fig. 4. Neighboring orbits have different periods. However, the two orbits animated in the figure are ε-close, as the corresponding trajectories remain close under a reparameterization of time (under which their periods would become equal). As this is true for all orbits in a neighborhood of either of the animated trajectories, they are both *orbitally stable*. In fact, all orbits are orbitally stable for this system, except for the saddle points and their connections.

Example 2: Linear flows on the tori

The flow on the two tori

$$\begin{cases} \dot{\theta}_1 = 0 \\ \dot{\theta}_2 = \sin(\theta_1) \end{cases}$$

is similar to the pendulum example above: Here, *all* orbits are orbitally stable, as their neighbors are ε-close under reparameterization of time.

However, upon adding a third coordinate with constant velocity

$$\begin{cases} \dot{\theta}_1 = 0 \\ \dot{\theta}_2 = \sin(\theta_1) \\ \dot{\theta}_3 = 1 \end{cases}$$

the situation changes dramatically. Consider two neighboring orbits with slightly different initial values of θ_1. These two orbits are linear flows on invariant two tori with different, fixed values of θ_1. Generically, the two flows are irrational, so that each orbit is dense on its two tori. Therefore, the two orbits are close as *sets*. However, time cannot be reparameterized so that the orbits will be ε-close under the definition above, because the flows have different slopes.

Cross-References

▸ Lyapunov Stability

References

Arnold VI (1973) Ordinary differential equations. MIT Press, Cambridge, MA

Boyce WE, DiPrima RC (1997) Elementary differential equations and boundary value problems. Wiley, New York

Chicone C (1999) Ordinary differential equations with applications. Springer, New York

Floquet G (1883) Sur les équations différentielles linéaires à coefficients périodiques. Ann École Norm Sup 12:47–88

Hirsch MW, Smale S, Devaney RL (2004) Differential equations, dynamical systems and an introduction to chaos. Academic Press/Elsevier, San Diego

Trefethen LN, Embree M (2005) Spectra and pseudospectra: the behavior of nonnormal matrices and operators. Princeton University Press, Princeton

Stability, States and Regions

Tianshou Zhou
School of Mathematics and Computational Sciences, Sun Yet-Sen University, Guangzhou, Guangdong, China

Definition

The stable steady state, stable region, and local stability are three common subjects of nonlinear dynamical systems including gene regulatory systems [1–3]. There is some correlation among the three conceptions. The stable steady state means that the steady state of the dynamical system of interest is independent of the time and that an arbitrary trajectory starting at a small neighborhood of the steady state finally tends to the steady state in the limit of the time. A stable region represents the attracting basin of the steady state of interest. Local stability means that the steady state of interest is stable in the Lyapunov sense. When one studies the qualitative properties of orbits of a dynamical system, existence, stability, and attracting domain of steady states of this system are first concerned.

Characteristics

Consider a set of coupled autonomous ordinary differential equations (ODEs) that can describe the time evolution of key components in a gene regulatory network, written in vector notation as:

$$\frac{dx}{dt} = F(x), x \in R^n \qquad (1)$$

where

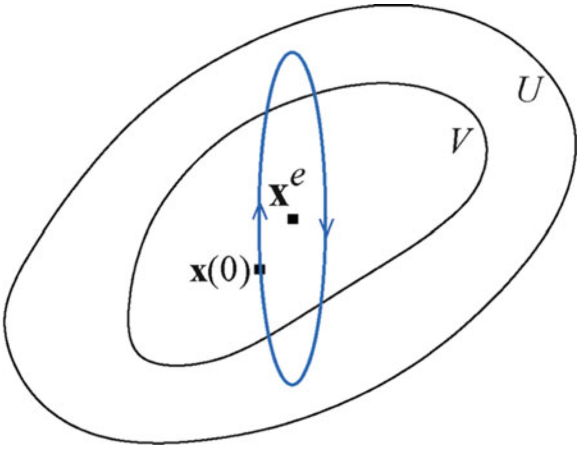
$$F(x) = \begin{pmatrix} F_1(x_1, x_2, \cdots, x_n) \\ F_2(x_1, x_2, \cdots, x_n) \\ \vdots \\ F_n(x_1, x_2, \cdots, x_n) \end{pmatrix}, x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Denote by $x(t)$ a solution to this equation satisfying initial conditions: $x_0 = x(0)$. Equilibria (sometimes called equilibrium points or fixed points or steady states) x^e are special constant solutions $x(t) = x^e$, where $F(x^e) = 0$ or $F_j\left(x_1^e, x_2^e, \cdots, x_n^e\right) = 0, j = 1, 2, \cdots, n$. For the system of a gene regulatory network, $F(x)$ can often be expressed as $F(x) = f(x) - g(x)$, where $f(x)$ represents the production part whereas $g(x)$ does the degradation part, so the equilibrium x^e satisfies $f(x^e) = g(x^e)$.

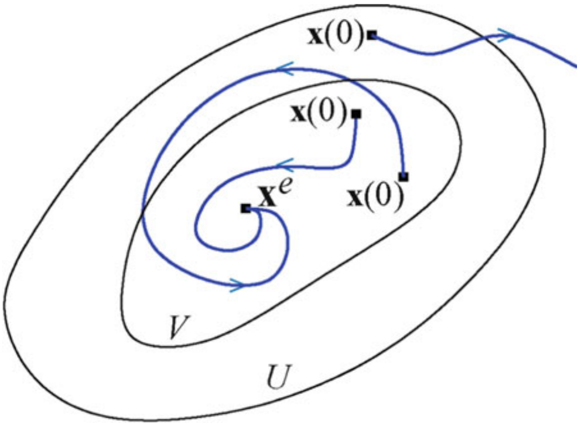
The stability of equilibria mainly includes Lyapunov stability and asymptotic stability.

Lyapunov stability: x^e is a stable equilibrium if for every neighborhood U of x^e, there is a neighborhood $V \subseteq U$ of x^e such that every solution $x(t)$ starting in V, i.e., the solution starting from $x_0 \in V$, remains in U for all $t \geq 0$. Note that $x(t)$ need not approach x^e. x^e is not stable if it is unstable.

Asymptotic stability: An equilibrium x^e is asymptotically stable if it is Lyapunov stable and additionally



Stability, States and Regions, Fig. 1 Lyapunov stability



Stability, States and Regions, Fig. 2 Asymptotic stability

V can be chosen so that $|x(t) - x^e| \to 0$ as $t \to \infty$ for all $x_0 \in V$.

An equilibrium that is Lyapunov stable but not asymptotically stable is sometimes called neutrally stable. See Figs. 1 and 2 for illustrations.

Note: Lyapunov stability and asymptotic stability are of local stability.

Stable region: If there is a region V of the equilibrium x^e such that a solution of (1) starting from every point x_0 in V, denoted by $x(t)$, is convergent, i.e., $|x(t) - x^e| \to 0$ as $t \to \infty$, then V is called a stable region or an attracting basin of x^e.

Note 1: The boundary of the stable region of an equilibrium is possibly complicated and is not easy to be determined.

Note 2: If V equates just the definition region of the system (1), then we say that x^e is globally stable.

Note 3: Completely similarly, we can give definitions for the stable region of an orbit and global stability.

Below, we introduce stability analysis.

Linearization: Suppose that $x = x^e$ is an equilibrium, implying that if $x(0) = x^e$, then $x(t) \equiv x^e$. For a small perturbation of x^e, denoted by $\xi(t) = (\xi_1, \xi_2, \cdots, \xi_n)$ with $|\xi(t)| \ll 1$, let $x(t) = x^e + \xi(t)$. Substitute the expression of $x(t)$ into both sides of (1) and expand function f in a multivariable, vector-valued Taylor series (we assume that f is sufficiently differentiable so that Taylor's Theorem with remainder applies to each component) to obtain

$$\begin{aligned} \dot{x^e} + \dot{\xi} &= F(x^e + \xi) \\ &= F(x^e) + DF(x^e)\xi + O\left(|\xi|^2\right) \end{aligned} \tag{2}$$

where $DF(x^e)$ denotes the $n \times n$ Jacobian matrix of partial derivatives $[\partial F_i / \partial x_j]$, evaluated at the equilibrium x^e, and $O\left(|\xi|^2\right)$ denotes terms of quadratic and higher order in the components $\xi_1, \xi_2, \cdots, \xi_n$. Thus, for small enough $|\xi|$, the first term $DF(x^e)\xi$ dominates. Taking into account that $\dot{x^e}$ and $F(x^e)$ vanish and ignoring the small term $O\left(|\xi|^2\right)$, we obtain the linear system:

$$\dot{\xi} = DF(x^e)\xi \tag{3}$$

which is called the linearization of (1) at x^e. It can be solved by standard methods.

The general solution $\xi(t)$ of (3) is determined by the eigenvalues and eigenvectors of the Jacobian matrix $DF(x^e)$. However, we are usually concerned with qualitative properties rather than complete solutions. In particular, in studying stability, we want to know whether the size (norm) of solutions grows, stays constant, or shrinks as $t \to \infty$. This can be answered just by examining the eigenvalues.

Recall that if λ is a real eigenvalue with the eigenvector υ, then there is a solution to the linearized equation of the form: $\xi(t) = e^{\lambda t}\upsilon$; if $\lambda = \alpha \pm i\beta$ is a complex-conjugate pair with eigenvectors $\upsilon = u \pm iw$ (u and w are real), then $\xi_1(t) = e^{\alpha t}(u \cos \beta t - w \sin \beta t)$ and $\xi_2(t) = e^{\alpha t}(u \sin \beta t + w \cos \beta t)$ are two linearly independent solutions. In both cases, the real part of λ

(almost) determines stability. Since any solution of the linearized equation can be written as the linear superposition of terms of these forms (except for the case of multiple eigenvalues), we can deduce the following:

- If all eigenvalues of $DF(x^e)$ have strictly negative real parts, then $|\xi(t)| \to 0$ as $t \to \infty$ for all solutions.
- If at least one eigenvalue of $DF(x^e)$ has positive real parts, then there is a solution with $|\xi(t)| \to +\infty$ as $t \to \infty$.
- If some pairs of complex-conjugate eigenvalues have zero real parts with distinct imaginary parts, then the corresponding solutions for $|\xi(t)| \to +\infty$ oscillate and neither decay nor grow as $t \to \infty$.

Note 1: The eigenvalues of the linearization are preserved under (smooth) changes of coordinates.

Note 2: When multiple eigenvalues exist and there are not enough linearly independent eigenvectors to span R^n, solutions behave like $|\xi(t)| \sim t^k e^{\lambda t}$, so that they still decay for sufficiently long times if $\lambda < 0$ and grow if $\lambda > 0$.

Note 3: The form $t^k e^{\lambda t}$implies that *transient growth* occurs over initial times even if $\lambda < 0$.

Consider the example

$$\begin{cases} \dot{\xi}_1 = -2\xi_1 + \alpha\xi_2 \\ \dot{\xi}_2 = -\xi_2 \end{cases} \tag{4}$$

for large $|\alpha|$. This system has eigenvalues -1 and -2. However, taking $\xi_1(0) = 0$ and $\xi_2(0) = 1$, the first coordinate $\xi_1(t) = \alpha(e^{-t} - e^{-2t})$ initially grows from zero to a maximum value of $\alpha/4$. For sufficiently large α, the growth of ξ_1 will initially overwhelm the decay of $\xi_2 = e^{-t}$so that the trajectory transiently moves farther from the fixed point before approaching it as $t \to \infty$. This also illustrates the need for the two neighborhoods U and Vin the definitions of stability.

This motivates us to introduce the following concept.

Hyperbolic equilibria: x^e is a *hyperbolic* or nondegenerate equilibrium if all the eigenvalues of $DF(x^e)$ have nonzero real parts.

Equipped with the linear analysis sketched above, and recognizing that the remainder terms ignored in passing from Eqn. (2) to (3) can be made as small as we wish by selecting a sufficiently small neighborhood of x^e, we can determine the stability of *hyperbolic* equilibria from their linearization:

Proposition. *If x^e is an equilibrium of $\dot{x} = F(x)$ and all the eigenvalues of the Jacobian matrix $DF(x^e)$ have strictly negative real parts, then x^e is exponentially (and hence asymptotically) stable. If at least one eigenvalue has strictly positive real part, then x^e is unstable.*

Moreover, the Hartman-Grobman Theorem says that the full nonlinear system (1) is topologically equivalent to the linearized system (3) in a small neighborhood of a hyperbolic equilibrium.

Borrowing from fluid mechanics, we say that if all nearby solutions approach an equilibrium (e.g., all eigenvalues have negative real parts), it is a *sink*; if all nearby solutions recede from it, it is a *source*, and if some approach and some recede, it is a *saddle point*. When the equilibrium is surrounded by nested closed orbits, we call it a *center*.

Degenerate Equilibria: One might hope to claim that Lyapunov stability (per the definition above) holds even if (some) eigenvalues have zero real part, but the following counterexamples demonstrate that this is not the case:

Example 1. Consider

$$\dot{x} = \alpha x^3, \alpha \neq 0 \tag{5}$$

Here $x = 0$ is the equilibrium and the linearization at 0 is

$$\dot{\xi} = \left(3\alpha x^2\big|_{x=0}\right)\xi = 0 \tag{6}$$

with solution $\xi(t) = \xi(0) = cons\tan t$, so certainly $x = 0$ is Lyapunov stable for (6), but not asymptotically stable. The exact solution of the nonlinear ODE (5) may be found by separating variables:

$$\int_{x(0)}^{x(t)} \frac{dx}{x^3} = \int \alpha dt \Rightarrow x(t) = \frac{x(0)}{\sqrt{1 - 2\alpha x^2(0)t}}$$

We therefore deduce that

$|x(t)| \to \infty$ as $t \to \frac{1}{2\alpha x^2(0)}$ if $\alpha > 0$ (blowup! Instability)

$|x(t)| \to 0$ as $t \to \infty$ if $\alpha < 0$ (asymptotic stability)

The linearized system (6) is *degenerate* and the nonlinear "remainder terms", ignored in our linearized analysis, determine the outcome in this case. Here it is obvious, at least in retrospect, that ignoring these terms is perilous, since while they are indeed $O(\xi^2)$ (in fact, $O(\xi^2)$), the linear $O(\xi)$ term is identically zero!

Example 2. Consider the two-dimensional system:

$$\begin{cases} \dot{x} = y + \alpha(x^2 + y^2)x \\ \dot{y} = -x + \alpha(x^2 + y^2)y \end{cases}$$

Note that the linearization is simply a harmonic oscillator with eigenvalues $\pm i$. Is the equilibrium $(x, y) = (0, 0)$ of this system stable or unstable? To answer this, it is convenient to transform to polar coordinates $x = r\cos\theta$, $y = r\sin\theta$, which gives the uncoupled system:

$$\dot{r} = \alpha r^3 \dot{\theta} = -1$$

The first equation is as in the example above, so we conclude: $\alpha > 0 \Rightarrow$ unstable; $\alpha = 0 \Rightarrow$ stable; $\alpha < 0 \Rightarrow$ asymptotically stable. The linearization gives no information if $\alpha = 0$.

How can we prove stability in such degenerate cases, in which one or more eigenvalues has zero real part? One method requires construction of a function, often called Lyapunov function, which remains constant, or decreases, along solutions. For mechanical systems, the total (kinetic plus potential) energy is often a good candidate. This allows one to prove stability and even asymptotic stability in certain cases, via describe Lyapunov's second method or direct method:

Theorem. *Suppose that $dx/dt = \dot{x} = F(x)$ has an isolated equilibrium at $x = 0$ (without loss of generality one can move an equilibrium $x^e$to 0 by letting $y = x - x^e$). If there exists a differentiable function $V(x)$, which is positive definite in a neighborhood of 0 (in the sense that $V(0) = 0$ and $V(x) > 0$ for $x \neq 0$) and for which $dV/dt = \nabla V \cdot F$is negative definite on some domain D containing 0, then 0is asymptotically stable. If $dV/dt$is negative semidefinite* (i.e., *$dV/dt = 0$is allowed), then 0 is Lyapunov stable.*

References

Arnold VI (1973) Ordinary differential equations. MIT Press, Cambridge, MA

Boyce WE, DiPrima RC (1997) Elementary differential equations and boundary value problems. Wiley, New York

Hirsch MW, Smale S, Devaney RL (2004) Differential equations, Dynamical systems and an introduction to chaos. Academic Press/Elsevier, San Diego

Stable Isotope Dilution Technique

▶ Selective Reaction Monitoring

STAGA

▶ SAGA

Stalk of RNAPII

Tetsuro Kokubo
Department of Supramolecular Biology, Graduate School of Nanobioscience, Yokohama City University, Yokohama, Kanagawa, Japan

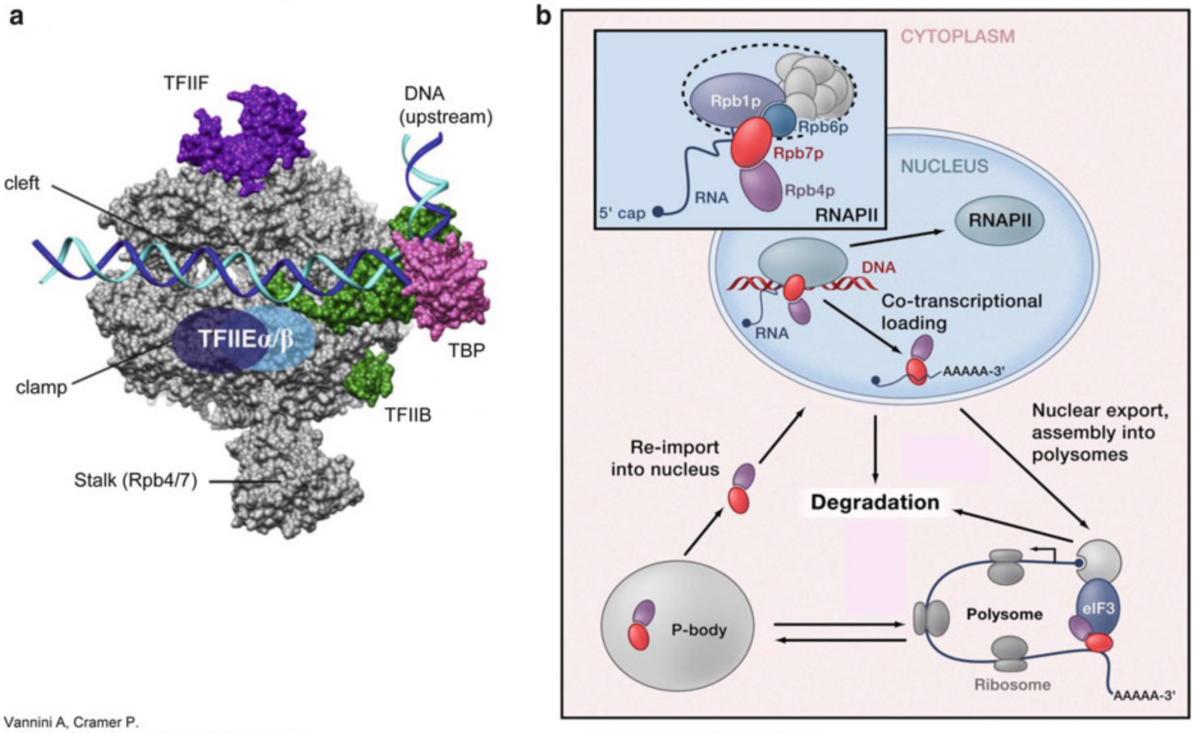
Synonyms

Rpb4/7 heterodimer

Definition

Among the twelve subunits of RNAPII, Rpb4 and Rpb7 possess intriguing properties. These two subunits form a "stalk"-like heterodimer (Rpb4/7) near the RNA exit site in RNAPII (Fig. 1) (Harel-Sharvit et al. 2010; Preker et al. 2010; Vannini et al. 2012). Rpb4/7 attaches to RNAPII via interaction with Rpb1 and Rpb6 and is readily dissociated from RNAPII especially under growing conditions. Importantly, RNAPII lacking Rpb4/7 (RNAPII-ΔRpb4/7) can be incorporated into PIC but is deficient in transcription initiation in vitro. X-ray structural analyses showed that the clamp domain of RNAPII-ΔRpb4/7 takes on an "open" conformation, whereas that of RNAPII containing Rpb4/7 takes on a "closed" conformation. In addition, the clamp domain of RNAPII-ΔRpb4/7 containing a DNA:RNA hybrid (i.e., in an elongation state)

Multiple functions of the stalk-like structure (Rpb4/7) in RNAPII



Vannini A, Cramer P.
Conservation between the RNA polymerase I, II,
and III transcription initiation machineries.
Mol Cell. 2012 Feb 24;45(4):439-46.

Pascal Preker, Torben Heick Jensen
Translation by Remote Control
Cell, Volume 143, Issue 4, 12 November 2010, Pages 501-502

Stalk of RNAPII, Fig. 1 Multiple functions of the stalk-like structure (Rpb4/7) in RNAPII, (**a**) Structure of the PIC. The stalk-like structure comprising Rpb4 and Rpb7 (Rpb4/7) is located at a position close to the mRNA exit site of RNAPII. (**b**) The Rpb4/7 heterodimer is transferred from RNAPII to mRNA in the nucleus after which it regulates reactions in the cytoplasm such as export, translation, and degradation. The P-body (processing body) is a cellular compartment where mRNA is either degraded or stored in an inactive form

takes on a "closed" conformation. Thus, Rpb4/7 may play a central role in the conformational change of the clamp domain that likely occurs at an early stage of transcription initiation.

Rpb4/7 is important not only for transcription initiation but also for some post-initiation steps, such as elongation, termination, and poly (A) + addition. Furthermore, Rpb4/7 shuttles between the nucleus and cytoplasm, and functions in the cytoplasm independently of RNAPII. Remarkably, Rpb4/7 is transferred from RNAPII to mRNA in the nucleus after which it regulates the export, translation, and/or degradation of mRNA. Importantly, the physical attachment of Rpb4/7 to RNAPII in the nucleus is essential for its cytoplasmic function, indicating a requirement for the proper activation of Rpb4/7 during the transcription process, before the transfer to mRNA. Altogether, Rpb4/7 coordinates the fate of mRNA from the birth (transcription initiation) to death (degradation). Although stalk structures are also present in RNAPI and III (Vannini et al. 2012), it is currently unknown whether or not they have a similar "coordinator" function.

Cross-References

▶ Transcription Initiation in Eukaryote

References

Harel-Sharvit L, Eldad N, Haimovich G, Barkai O, Duek L, Choder M (2010) RNA polymerase II subunits link transcription and mRNA decay to translation. Cell 143(4):552–563

Preker P, Jensen TH (2010) Translation by remote control. Cell 143(4):501–502

Vannini A, Cramer P (2012) Conservation between the RNA polymerase I, II, and III transcription initiation machineries. Mol Cell 45(4):439–446

Standard Setting

▶ Learning, Attribute-Value

Standards for Reporting Enzymology Data

▶ STRENDA

Start

▶ Cell Cycle Transition, Principles of Restriction Point

State

▶ Global State, Boolean Model
▶ Local State, Boolean Model

State Synchronization

Tianshou Zhou
School of Mathematics and Computational Sciences,
Sun Yet-Sen University, Guangzhou,
Guangdong, China

Definition

Partial synchronization is also a specific type of synchronization. In this synchronization, the synchronization error between the corresponding state variables of any two oscillators in the population finally tends to zero. For example, for a system of N coupled units governed by the following set of differential equations

$$\frac{dx_i}{dt} = F(x_1, x_2, \cdots, x_N) + C_i(x_1, x_2, \cdots, x_N), \text{every } x_i \in R^n$$

where

$$F(x) = \begin{pmatrix} F_1(x_1, x_2, \cdots, x_N) \\ F_2(x_1, x_2, \cdots, x_N) \\ \vdots \\ F_n(x_1, x_2, \cdots, x_N) \end{pmatrix},$$

$$C_i(x) = \begin{pmatrix} C_{i1}(x_1, x_2, \cdots, x_N) \\ C_{i2}(x_1, x_2, \cdots, x_N) \\ \vdots \\ C_{in}(x_1, x_2, \cdots, x_N) \end{pmatrix},$$

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in} \end{pmatrix}, i = 1, 2, \cdots, N \quad \text{and}$$

$C_i(x_1, x_2, \cdots, x_N)$ represents the coupling term. Define the synchronization error as

$$E(t) = \frac{1}{N} \sum_{i,j=1}^{N} \left| x_i(t) - x_j(t) \right|$$

If $\lim_{t \to +\infty} E(t) = 0$, then we say that the N subsystems achieve state synchronization.

Statistical Experimental Design

▶ Designing Experiments for Sound Statistical Inference

Statistical Methods in Systems Biology

Eugene Kolker
Bioinformatics and High-throughput Analysis
Laboratory, Seattle Children's Research Institute,
Seattle, WA, USA
Departments of Biomedical Informatics and Medical
Education and Pediatrics, University of Washington,
Seattle, WA, USA

Definition and Description

Systems (integrative, high-throughput) biology studies cover vast range of experimental approaches, computational algorithms, and data types. Therefore to analyze and interpret these studies, it is imperative that the correct statistical methods be applied. Systems biology studies involve modern high-throughput data technologies and complex experimental designs, and these characteristics involve many complex issues and require innovation in statistical methodology. Important issues related to systems biology studies include accounting for diverse sources of variability, multiple hypothesis testing, and adapting statistical methods to advance biological technologies such as microarrays, next-gen sequencing, and mass spectrometry.

Modern high-throughput technologies generate large volumes of diverse data that can be used to address biological questions. The broad spectrum of research problems in systems biology includes identifying genes, proteins, and metabolites primary to a given condition/disease/inference state; comparing expression profiles across different populations; modeling chemical and biological processes and their effect on expression changes; assigning functions to newly identified entities; and much more. Both the complexity and the volume of the data generated to answer these questions require advanced, computationally intensive statistical methods to process and extract knowledge.

Unfortunately, the generation of large data sets can easily lead to GOBS – the Generation of BS (*bovine scatus*; K. H. Nealson as quoted in Holzman and Kolker 2003). Time and resources can be wasted unless proper thought is given to the experimental design before the experiment is started. "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of." (Fisher 1938).

Data collection is a rather complicated, multistage process. The duration and complexity of data acquisition introduces bias and sizable technical variation. A successful experiment, therefore, requires a careful design in order to reduce technical and instrumental variation and increase the signal-to-noise ratio.

Consequently, the experimental design is reflected in statistical analysis models. Such models must consider the large number of sources that may generate technical variability. As with any other field, a random sample of experimental units has to be chosen which is pertinent to the hypothesis of interest. In systems biology, the experiments are often costly and time consuming. Consequently, investigators typically have a fairly small number of samples from which they derive a large number of hypotheses, a problem known in statistics as p>>n. For example, proteomics experiments identify expression levels for thousands of proteins but are usually constrained to a handful of experimental samples.

Researchers are often interested in building a prediction model using a number of expression values, p, to predict the output of interest based on sample of size, *n*. For example, researchers try to identify genes that can accurately predict a disease status or severity level. Standard algorithms for prediction models (forward selection, backward elimination, stepwise regression) require $p < n$. When the sample size is less than the number of expression values, the potential is higher that a random variable could discriminate the two data sets. Thus, identifying genes of interest that can reliably distinguish cases from controls poses serious methodological, statistical, and computational challenges.

In systems biology, data analysis relies on a complex combination of different methods. One illustrative example is a "standard" proteomics experiment that compares protein expression levels in two groups of subjects. Before any cross-experimental comparisons can be performed, the protein expression levels must be calculated for each experiment, a process requiring numerous steps. Peptide identification is accomplished by searching collected spectra against a target database. False discovery rates are corresponded with peptide identifications using either a parametric modeling of the score distribution or a search against a decoy database of reshuffled spectra. Next, the peptide identifications need to be combined to determine the protein content of the sample.

The accuracy of protein identifications is characterized by the false discovery rate (FDR) which is typically reported on the global level. In numerous proteomics and other systems biology studies, however, the local FDR is more useful as it measures error rate for each individual protein. The local FDR can be used to select a set of proteins for further analysis ensuring that selected proteins were identified with desired certainty. The error measurements can also be directly incorporated into the analyses.

Also different proteins can contain identical peptides, so the proteins cannot necessarily be differentiated from each other based on identified peptides. Computational models are calibrated to properly assign peptides to proteins. The number of identified peptides for each protein is correlated with protein concentration and, as such, can be used as proxy measure in hypothesis testing. Having calculated protein expression levels, cross-experimental comparisons can now be performed. To compare protein expression levels between the two populations, the number of identified peptides can be normalized within experiments to adjust for systematic bias. For each protein, one computes a test statistic (using F-test, (moderated) *t*-test, etc.) and a corresponding p-value (using distributional assumptions or by constructing a permutation distribution). Since the test is performed simultaneously on a large number of proteins, the p-values are adjusted to maintain the required type I error rate. Often, this expression analysis is followed by studying the pathways and protein interactions that belong to the same pathway.

Functional annotation is another major research challenge in systems biology. Here, the goal is to assign biological functions to genes and proteins. Large volumes of data make manual curation impractical. Automated methods rely on identifying a protein/gene, or a group of proteins/genes, with a sequence most similar to the protein/gene of interest and whose function is known and propagating this function to uncharacterized protein/gene. However, even the simplest methods to estimate sequence similarity and group proteins present an immense computational challenge. One can reduce the computational complexity of the problem by developing simple, yet sensitive, scoring rules to classify uncharacterized proteins. Clustering methods based on hierarchical agglomeration or Hidden Markov models are prominently used to tackle the problem of functional annotation.

To meet the challenge of big data in modern life sciences, the Data-Enabled Life Sciences Alliance (DELSA Global, "delsaglobal.org") was formed. The alliance is an ecosystem of stakeholders: scientists, statisticians, data analysts, computer scientists, educators, funding agencies and others working together to enable a much needed paradigm shift to translate data to knowledge to action through collective innovation (Kolker et al. 2012). Big data not only significantly enable our research capabilities, but also introduce major challenges, so called the 5 Vs of big data: volume, veracity, velocity, variety, and value (Higdon et al. 2013). Both life sciences in general and systems biology in particular demand special attention to the variety, value and veracity of big data (for details, see Higdon et al. 2013).

The Statistical Methods section is intended as a reference to the fundamental statistical methods and approaches used in systems biology. The section reviews classical approaches to statistical inference and hypothesis testing and provides a background on standard tests and models such as *t*-test, analysis of variance, linear regression, generalized linear models, and others. It also discusses methods for multiple hypothesis testing and describes statistical approaches to gene and protein expression data analysis.

References

Fisher RA (1938) Indian Statistical congress. Cambridge University

Higdon R et al (2013) Unraveling the complexities of life sciences data. Big Data J 1(1):BD17-23. http://www.liebertpub.com/overview/big-data/611/

Holzman T, Kolker E (2003) Statistical analysis of global gene expression data: Some practical considerations. Curr Opin Biotechnol 15(1):52–57

Kolker E, Stewart E, Ozdemir V (2012) Opportunities and challenges for the life sciences community. OMICS: J Integrative Biol 16:138–147

Statistical Modeling

▶ Modeling Formalisms, Lymphocyte Dynamics and Repertoires

Statistical Sampling

▶ Data Sampling

Statistics Model

▶ Probabilistic Model-based Transcription Regulatory Network Construction

Steady State

Ruiqi Wang
Institute of Systems Biology, Shanghai University,
Shanghai, China

Synonyms

Equilibrium

Definition

In deterministic differential equations, a steady state for a system means that there is no net change in the number or concentration of molecules in the system (Guckenheimer and Holmes 1983).

Cross-References

▶ Life Span, Turnover, Residence Time
▶ Lymphocyte Population Kinetics

References

Guckenheimer J, Holmes P (1983) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Springer, New York

Steady-State Probability Distribution

▶ Equilibrium Probability Distribution

Stem Body

▶ Midbody

Stem Cell Cenes

▶ Stem Cell Networks

Stem Cell Networks

Eric Werner
Department of Physiology, Anatomy and Genetics,
University of Oxford, Oxford, UK
Department of Computer Science, University of
Oxford, Oxford, UK
Oxford Advanced Research Foundation, Fort Myers,
FL, USA

Synonyms

Developmental control networks; Figurate networks; Geometric networks; Stem cell Cenes

Definition

A stem cell network is a developmental control network (▶ Developmental Control Networks), also called a ▶ cene, that contains one or more linear self-regenerating loops resulting in endless cell proliferation. The nodes in the network are cell control states. The branches denote cell division with each daughter cell entering the control state at the end of its branch.

Characteristics

Stem cell networks are a kind of developmental control network (▶ Developmental Control Networks) (Werner 2011a) with a special topology consisting of one or more linear self-regenerating control loops. Stem cell networks can be normal or cancerous. The difference between normal stem cell networks and cancer stem cell networks (▶ Cancer Networks) depends on the location of the stem cell network in the global developmental control network (▶ Cenome) as well as the cell types the stem cell network generates (Werner 2011a, b).

Stem Cell Network Types

Stem cell networks come in various basic forms: Linear Networks, Meta-Stem Cell Networks, and the more general k-th order Stem Cell Networks. These are also called geometric networks or figurate networks because their proliferation properties are related to the geometric or figurate numbers and the coefficients of Pascal's Triangle (Werner 2011b). In summary, we have

- Linear stem cell networks or 1st-order stem cell networks G1
- Meta-stem cell networks or 2nd-order stem cell networks G2
- k-th order stem cell geometric or figurate networks Gk

Stem cell networks are further divided into deterministic or stochastic ▶ developmental control networks, and they also may or may not involve cell signaling (Werner 2011b).

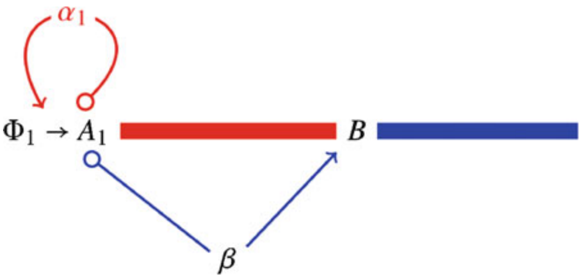
- Deterministic
- Stochastic
- Signaling

A stem cell is a self-regenerating cell that generates other terminal progenitor cells. Depending on the network these terminal cells may stochastically dedifferentiate to stem cells (Werner 2011b).

Linear Stem Cell Networks

Linear stem cell networks G_1 are 1st-order geometric networks that contain one self-regenerating loop and one link to a terminal developmental network (▶ Developmental Control Networks) (Fig. 1).

For one loop after $n > 0$ synchronous divisions, we get a one-dimensional structure where its length gives the number of cells:



Stem Cell Networks, Fig. 1 A linear, 1st-order stem cell network. It generates cells B if the condition Φ_1 is met (Werner 2011b)

$$Cells(n,1) = 1 + Lin(n) = 1 + n$$

$$= \sum_{i=0}^{1} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} \qquad (1)$$

where $Lin(n) = n$

Meta-Stem Cell Networks

Meta stem cell generate stem cells. A meta-stem cell network is a 2nd order stem cell network that contains two linked linear loops (Werner 2011b).

Meta-stem cells proliferation is related to the triangular numbers and the area of a two-dimensional triangle.

$$Tri(n) = \frac{n^2 + n}{2} = \frac{n(n+1)}{2} \qquad (2)$$

Thus, for two loops add the triangular number to give the area of a two-dimensional triangle (Werner 2011b). When $n > 0$ (Fig. 2):

$$Cells(n,2) = 1 + Lin(n) + Tri(n-1) \qquad (3)$$

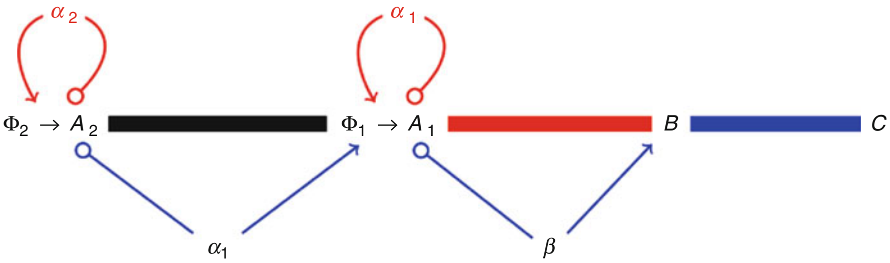
$$= 1 + n + \frac{n(n-1)}{2} \qquad (4)$$

$$= \sum_{i=0}^{2} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} \qquad (5)$$

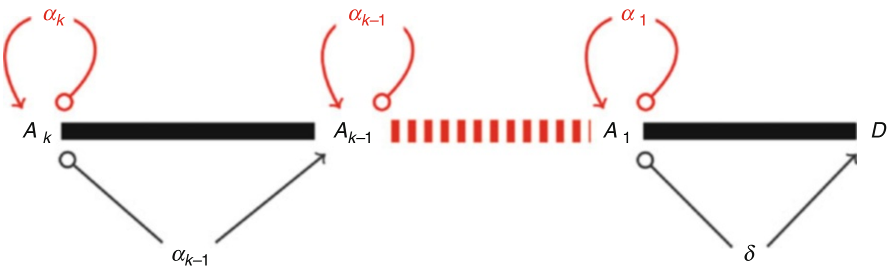
Meta-Meta-Stem Cell Networks

For three loops add the tetrahedral number to the above to give the volume of three-dimensional pyramid (tetrahedron) (Werner 2011b). When $n > 0$:

Stem Cell Networks, Fig. 2 A meta-stem cell network is a 2nd-order geometric network. It consists of a 2nd-order loop at A_2 linked to a 1st-order loop at A_1 (Werner 2011b)



Stem Cell Networks, Fig. 3 A k-th order stem cell network (Werner 2011b). The network contains k loops at control states $A_k \ldots A_1$ and ends in a terminal developmental network D



$$Cells(n,3) = 1 + Lin(n) + Tri(n-1) + Tet(n-2) \tag{6}$$

$$= 1 + n + \frac{n(n-1)}{2} + \frac{n(n-1)(n-2)}{6} \tag{7}$$

$$= \sum_{i=0}^{3} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \binom{n}{3} \tag{8}$$

Where the tetrahedral number is defined as:

$$Tet(n) = \frac{n(n+1)(n+2)}{6} \tag{9}$$

Geometric or Figurate Networks

Because their proliferative properties are related to the geometric or figurate numbers and the coefficients of Pascal's Triangle, simple stem cell networks are also called geometric or figurate networks. A k-th order stem cell network is a k-th order geometric network G_k that contains k loops ending with one link to a terminal developmental network (▶ Developmental Control Networks) (Fig. 3).

For k loops and $n > 0$, sum the sequence of numbers through $\frac{n!}{k!(n-k)!}$ to give the volume of a k-dimensional pyramid (Werner 2011b):

$$Cells(n,k) = 1 + Lin(n) + Tri(n-1) + Tet(n-2) + \ldots + \binom{n}{i} \tag{10}$$

$$= 1 + n + \frac{n(n-1)}{2} + \frac{n(n-1)(n-2)}{6} + \ldots + \frac{n!}{k!(n-k)!} \tag{11}$$

$$= \sum_{i=0}^{k} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{k-1} + \binom{n}{k} \tag{12}$$

So, in general, we have:

$$Cells(n,k) = \begin{cases} 1 & \textit{if } n = 0 \\ \sum_{i=0}^{k} \frac{n!}{i!(n-i)!} = \sum_{i=0}^{k} \binom{n}{i} & \textit{otherwise} \end{cases} \tag{13}$$

When $n > 0$ we have:

$$Cells(n,k) = \sum_{i=0}^{k} \binom{n}{i} \tag{14}$$

Limited Exponential Growth with Geometric Networks

When the number of rounds of division n is less than the number of loops k, that is, $n \leq k$, then a geometric stem network exhibits exponential growth since the following holds:

$$\sum_{i=0}^{n} \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n-1} + \binom{n}{n} = 2^n \tag{15}$$

Subnetworks and Tangential Networks in Stem Cell Networks

Any loop can include a path that is a subpath of another developmental network (▶ Developmental Control Networks). This means the loop has further developmental progeny generated by that subpath. Furthermore, a stem cell network can link to a complex progenitor network that generates complex multicellular structures.

The Stem Cell Hierarchy and Cancer Stem Cells

Higher-order stem cell networks generate a hierarchy of stem cell types. This stem cell network hierarchy imposes a hierarchy on metastases resulting from cancer ▶ stem cell networks (▶ Cancer Networks).

Cross-References

▶ Cancer Networks
▶ Cene
▶ Developmental Control Networks

References

Werner E (2011a) On programs and genomes, arXiv:1110.5265v1 [q-bio.OT]. http://arxiv.org/abs/1110.5265

Werner E (2011b) Cancer networks: a general theoretical and computational framework for understanding cancer, arXiv:1110.5865v1 [q-bio.MN]. http://arxiv.org/abs/1110.5865v1

Stem-Loop

Gota Kawai
Department of Life and Environmental Sciences,
Chiba Institute of Technology, Narashino, Chiba,
Japan

Synonyms

Hairpin; Hairpin loop

Definition

Stem-loop is an essential unit of the structure of single stranded RNA or DNA. A stem-loop consists of a stem, double helix, and a loop which links the stem. The length of the loop is typically 3–8. Tetraloops, stem-loops with four nucleotides in the loop, are frequently found in RNA. Especially, tetraloop RNAs with a loop sequence of GNRA, UNCG, and CUNG (N is any nucleotide, R is a purine nucleotide) are commonly found (Nowakowski and Tinoco 1999). For DNA, triloops with GNA sequences in the loop are found to be stable (Yoshizawa et al. 1997).

References

Nowakowski J, Tinoco I Jr (1999) RNA structure in solution. In: Neidle S (ed) Oxford handbook of nucleic acid structure. Oxford Science, New York

Yoshizawa S, Kawai G, Watanabe K, Miura K, Hirao I (1997) GNA trinucleotide loop sequences producing extraordinarily stable DNA minihairpins. Biochemistry 36:4761–4767

Stem-Loop Structure

▶ Hairpin Structure

Stepwise Assembly Pathway

▶ PIC Assembly Pathways

Stimulus Conditions

▶ Enabling Conditions

Stochastic Chemical Kinetics

▶ Chemical Master Equation

Stochastic Differential Equation

▶ Stochastic Processes, Fokker-Planck Equation

Stochastic Effects in Metabolic Networks

Andrea Rocco and Andrzej M. Kierzek
Division of Microbial Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, Surrey, UK

Synonyms

Noise in metabolic networks; Noise in metabolic pathways; Random fluctuations in metabolic networks; Random fluctuations in metabolic pathways; Stochastic fluctuations in metabolic pathways

Definition

Stochastic fluctuations in metabolic networks are random differences between individual cells in metabolic flux distribution (▶ Metabolic Flux Analysis), the metabolite concentrations, or both. The origin of such stochastic fluctuations may vary. Stochastic fluctuations at the enzymatic level may originate because of the low copy numbers of the enzyme, or low numbers of molecules participating in expression of its gene, or because of other mechanisms. Metabolites are also affected by stochasticity, propagated from enzyme fluctuations, or related to their own low copy numbers. Finally, stochastic effects on both enzymes and metabolites can be expected in presence of fluctuating control parameters, such as temperature or pH levels.

Characteristics

Understanding that inherent randomness of chemical reactions constituting molecular machinery of the living cell can result in different phenotypes of individual, genetically identical cells is a major basic discovery resulting from application of systems biology approach. Chemical reactions are inherently random because they result from the reactive collision between molecules moving randomly within reaction volume. In the case of bulk chemical reactions, very large numbers of reactant molecules result in a negligible variance in the concentration of products. However, biochemical reactions occur in very small volumes of cells where numbers of reactant molecules can become very small resulting in considerable variance in the amounts of products. This inherent noise in biochemical processes has been demonstrated to result in phenotypic variability of single cells with identical genotypes.

Propagation of Gene Expression Noise to the Level of Metabolic Networks

It is widely appreciated that transcription and translation machinery may involve molecules that occur in low copy numbers. Lactose repressor occurring on average in 10 copies per cell of *Escherichia coli* is a classical example. Therefore, the heterogeneity of gene expression in individual cells has been long recognized in theoretical studies and confirmed by numerous experiments including protein molecule tracking in individual cells of *E. coli*. As metabolites and metabolic enzymes are present in the cell usually at much higher copy numbers metabolic networks are usually assumed to be deterministic and their possible stochasticity is frequently ignored. However, stochasticity of gene regulatory networks implies that transcription and translation rates of genes encoding enzymes may be very different in individual cells leading to different metabolic flux distributions.

Puchalka and Kierzek (2004) studied the propagation of gene expression noise to the level of metabolic pathways by large-scale stochastic kinetic simulation. They studied the model involving glycolysis, and glucose, lactose, and glycerol transport. Metabolites, proteins, transcripts, and DNA elements (promoters, open reading frames) were represented as molecular species in stochastic kinetic simulation. In a hybrid algorithm integrating Gillespie's exact stochastic simulation and a tau leap method no explicit distinction has been made between gene regulatory and metabolic network; all processes where described as molecular species taking part in the reactions. It was possible to estimate kinetic parameters of the model from quantitative experiments performed on this very well studied system. Computer simulation of a diauxic shift from glucose to the mixture of lactose and glycerol showed that individual cells adopted one of two distinct metabolic flux distributions where either lactose or glycerol, but not both, was used as a carbon source where the glucose was depleted from the medium. This behavior was caused by the propagation of gene expression noise through the network of negative and positive feedbacks involved in the diauxic shift response. Delay in the negative feedback loop of cAMP, CRP, and adenylate cyclase gene resulted in an "overshooting" behavior leading to the burst of cAMP amount. Transcription of catabolic operons was switched on during this time by increased amounts of cAMP, CRP complex. Depending on the random delay in activation of lactose operon transcription either lactose or glycerol promoter was activated first. Activation of either of the two operons was reinforced by the positive feedback through increased transport of the inducer. Full activation of the catabolic operon resulted in the increased availability of carbon source, increased levels of PEP and removal of hunger signal. This prevented other catabolic operon from activation. Thus propagation of stochastic fluctuations in catabolic operon expression resulted in two populations of cells using either lactose or glycerol, but not both as a carbon source. According to simulation lactose was preferred carbon source in a sense that more cells were growing on lactose than glycerol, but significant population of cells growing on glycerol was still present. Moreover, the state of individual cell was epigenetically inherited as after cell division two daughter cells inherited elevated levels of one of the two permeases thus maintaining the carbon source phenotype. In conclusion propagation of gene expression noise through complex networks of negative and positive regulatory feedback may result in epigenetically inherited heterogeneity of metabolic flux distribution in single cells.

Intrinsic and Extrinsic Noise

Understanding propagation of noise across different modules and layers of regulation requires a systemic description. Two classes of stochastic fluctuations are relevant, usually referred to as intrinsic and extrinsic fluctuations (▶ Noise, Intrinsic and Extrinsic). Both types of fluctuations can affect both enzymes and metabolites, even though their effects may be different.

Intrinsic Noise in Metabolic Networks

A natural way of studying intrinsic noise of metabolic networks is to extend metabolic control analysis to the case when noise is present (Rocco 2009; Bruggeman et al. 2009; Kim and Sauro 2010).

In Bruggeman et al. (2009), a mathematical framework is presented which aims at studying noise propagation and management in metabolic networks. The cases analyzed here refer to hierarchical networks, and the authors extend typical results of metabolic control analysis, such as those concerning the role of feedback loops, or the presence of cascades, to the case when noise is present, and propagates across the network.

Along similar lines is the work by Kim and Sauro (2010), who define a stochastic metabolic control analysis, based on the introduction of stochastic sensitivities for mean and covariance values of concentrations and fluxes, and derive the corresponding summation theorems.

Extrinsic Noise in Metabolic Networks

Extrinsic fluctuations have also been studied in metabolic networks, both in small modules and at the systemic level. In particular Samoilov et al. (2005) have analyzed the effect of extrinsic noise in enzymatic futile cycles, and found that bistable oscillatory behavior emerges in the stochastic system, whereas the corresponding deterministic dynamics is monostable. Changes in the stability properties of the lac operon are also predicted in Ochab-Marcinek (2010). These findings show how the effect of extrinsic noise can be highly counterintuitive and implies in general a change in position, number, and stability properties of the steady states of the system (Horsthemke and Lefever 2006).

At the systemic level, Rocco (2009) describes the effect of extrinsic noise within the framework of metabolic control analysis (MCA). Noise is assumed to affect any control parameter of the system $\boldsymbol{\mu}$, with an intensity ε, in general small. In presence of noise, the natural extension of the deterministic metabolite concentrations x is assumed to be given by the values of the concentrations that maximize the resulting stationary probability distribution, $x^{(m)}$:

$$x \rightarrow x^{(m)} \qquad \text{when noise is present} \qquad (1)$$

Then it can be proven (Rocco 2009) that the summation theorems of standard MCA for both the concentrations $x^{(m)}$ and the corresponding fluxes $J^{(m)}$ still hold true, namely:

$$\sum_{j=1}^{M} C_{\tilde{v}_j}^{x_i^{(m)}} = 0, \qquad \forall i = 1, \ldots, N \qquad (2)$$

and

$$\sum_{j=1}^{M} C_{\tilde{v}_j}^{J_i^{(m)}} = 1, \qquad \forall i = 1, \ldots, N \qquad (3)$$

In Eqs. 2 and 3, N represents the number of metabolites in the network, M is the number of reactions, and the control coefficients C are defined as:

$$C_{\tilde{v}}^{x_i^{(m)}} = \frac{\partial \ln x_i^{(m)}}{\partial \ln \tilde{v}_j} \qquad \text{and} \qquad C_{\tilde{v}_j}^{J_i^{(m)}} = \frac{\partial \ln J_i^{(m)}}{\partial \ln \tilde{v}_j} \qquad (4)$$

Here $\tilde{v}_j$ is the j-th rate when noise is present, and can be computed as

$$\tilde{v}_j = v_j - \sum_{k=1}^{N} \sum_{l=1}^{R_j} \varepsilon_l \left[\frac{\partial}{\partial x_k} \left(\frac{\partial v_j}{\partial \mu_l} \right) \right] \left[\frac{\partial v_j}{\partial \mu_l} \right] S_{kj} \qquad (5)$$

where the v_j is the deterministic reaction rate, μ_l is l-th of the R_j parameters present in v_j, and ε_l is the intensity of the noise affecting the parameter μ_l, S_{kj} is the deterministic stoichiometry matrix.

Analogous results can be obtained for connectivity theorems and partitioned response. This shows that even though the mathematical structure of MCA is preserved when extrinsic noise is present, control coefficients, as much as response and elasticity coefficients, acquire an explicit dependency on the noise intensity. Noise therefore can be interpreted as a control mechanism for the whole network.

This extension of standard MCA describes how stochastic perturbations originating locally at the level of single enzymes can propagate up to the systemic level and affect global variables.

References

Bruggeman FJ, Bluthgen N, Westerhoff H (2009) Noise management by molecular networks. PLoS Comp Biol 5: e1000506, 11 pp

Horsthemke W, Lefever R (2006) Noise-induced transitions: theory and applications in physics, chemistry, and biology. Springer, Berlin

Kim KH, Sauro HM (2010) Sensitivity summation theorems for stochastic biochemical reaction systems. Math Biosc 226:109–119

Ochab-Marcinek A (2010) Extrinsic noise passing through a Michaelis-Menten reaction: a universal response of a genetic switch. J Theor Biol 263:510–520

Puchalka J, Kierzek AM (2004) Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. Biophys J 86:1357–1372

Rocco A (2009) Stochastic control of metabolic pathways. Phys Biol 6:016002, 12 pp

Samoilov M, Plyasunov S, Arkin AP (2005) Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. PNAS 102:2310–2315

Stochastic Fluctuations in Metabolic Pathways

► Stochastic Effects in Metabolic Networks

Stochastic Modeling of Translation Elongation and Termination

M. Carmen Romano[1,2] and Ian Stansfield[2]
[1]Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen, UK
[2]Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK

Synonyms

Diffusion driven lattice-gas model for translation; One-dimensional lattice modeling of translation;

Total asymmetric exclusion process (TASEP) translation modeling; Translation modeling

Definition

Stochastic modeling of translation refers to mathematical models (▶ Mathematical Model, Model Theory) of the protein synthesis process that represent the movement of ribosomes along the mRNA as a stochastic process. Ribosome movement along an mRNA is governed by the activity of protein translation factors, and also by the delivery of transfer RNAs cognate to the codon at any given ribosomal position. In a stochastic model of this process, ribosome movement codon-by codon along the mRNA is not described by a deterministic rate, but instead by a probability. Such models are likely to better represent in vivo translation systems.

Characteristics

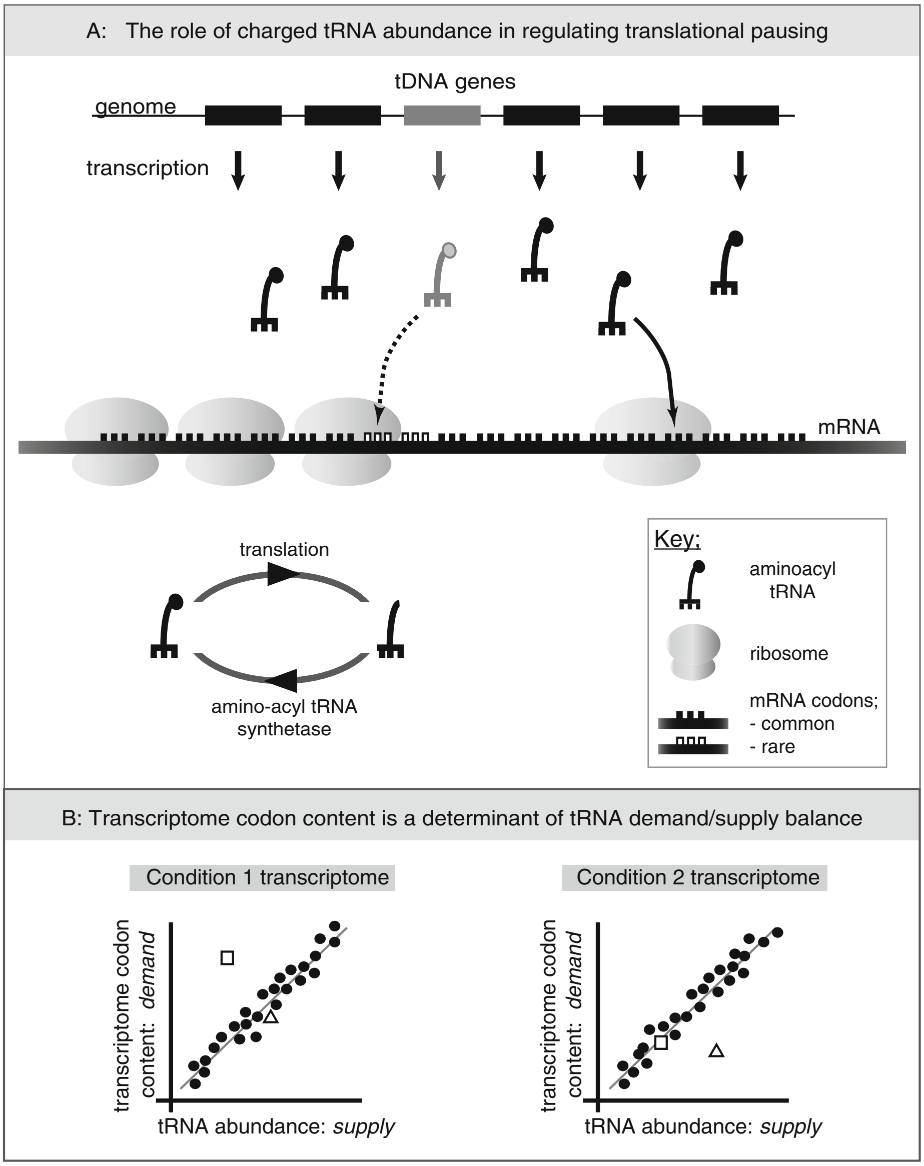
Translation Elongation and Termination

Cellular protein synthesis involves ribosomal translation of an mRNA, typically considered a three-stage process comprising ▶ translation initiation, ▶ translation elongation and ▶ translation termination (Kapp and Lorsch 2004). During the initiation step, small ribosomal subunits join the mRNA and locate the AUG initiation codon that begins the open reading frame. Following large ribosomal subunit joining, translation elongation then directs sequential addition of amino acids in response to each successive codon. New translation initiation events are continually occurring on each mRNA, loading the mRNA with multiple ribosomes, a so-called polysomal mRNA (Fig. 1). Each ▶ translation elongation cycle begins with delivery of charged transfer RNA (amino-acyl tRNA) delivery to the vacant ribosome acceptor site. The rate of delivery is proportional to the relative concentration of each tRNA species. The subsequent peptidyl-transferase reaction incorporates the newly-delivered amino acid into the growing polypeptide. Following this, a translocation reaction moves the ribosome forward to bring a new codon into the A-site, initiating the search for a new cognate tRNA. Following a series of elongation cycles, ribosomal encounter of an in-frame stop codon triggers the binding of a protein ▶ release factor and a consequent ▶ translation termination event.

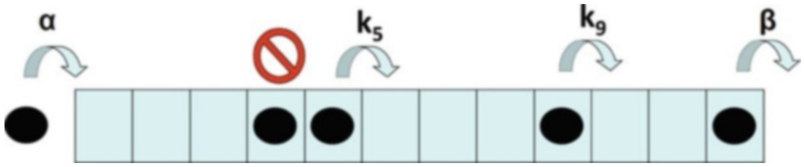
TRNA Abundance and Codon Bias

As amino acid delivery agents, tRNAs control the stochastic progression of ribosomes along the mRNA. The availability of each tRNA type and the codon composition of any given mRNA together regulate that stochastic movement. TRNA availability is dictated by the molecular abundance of that tRNA, and by what proportion of its population is charged with amino acid. Following delivery of the charged, aminoacyl tRNA to the ribosome, the tRNA is eventually released uncharged, having delivered its amino acid. It must subsequently be recharged by aminoacyl tRNA synthetase enzymes (Fig. 1a). The abundance of a particular isoacceptor tRNA population must therefore be matched by the catalytic capacity of its counterpart tRNA synthetase population (Fig. 1).

There is good evidence that tRNA abundance is determined by the gene copy number of their encoding tDNA gene, which in baker's yeast can vary from 1 to 16 for different tRNAs (Fig. 1). TRNA gene copy number is in turn matched to relative usage of codons in the ▶ transcriptome. In many microorganisms codon usage is subjected to bias (Plotkin and Kudla 2011). In codon-biased genomes, each codon within an amino acid family that is translated by a distinct tRNA will be used with different frequencies, with usage in extreme cases differing by as much as 30-fold within a codon family. Because tRNA abundance is matched to codon usage, tRNAs cognate for infrequently used codons are correspondingly rare. The encounter of such a rare codon by a ribosome during translation will produce an extended search for the right tRNA, and a stochastic ribosomal pause (Fig. 1). Stochastic variations in tRNA delivery and the consequential queuing ribosome interactions on an mRNA lattice are of central importance in defining the proteome (▶ Proteomics) that will be encoded by any given ▶ transcriptome (Gingold and Pilpel 2011). The requirement to understand and predict how efficiently ribosomes will translate any given mRNA sequence (▶ Post-transcriptional Regulation) has motivated the development of stochastic models of translation elongation.


A: The role of charged tRNA abundance in regulating translational pausing
tDNA genes
genome
transcription
mRNA
translation
amino-acyl tRNA synthetase
Key;
aminoacyl tRNA
ribosome
mRNA codons;
- common
- rare
B: Transcriptome codon content is a determinant of tRNA demand/supply balance
Condition 1 transcriptome
Condition 2 transcriptome
transcriptome codon content: demand
tRNA abundance: supply
transcriptome codon content: demand
tRNA abundance: supply

Stochastic Modeling of Translation Elongation and Termination, Fig. 1 Translation elongation and the regulation of ribosome transit along an mRNA. *Panel* **a**; Codon sequence and tRNA abundance dictate ribosome progression along mRNA. TRNA abundances are influenced by their gene copy number, and codon usage is matched to tRNA abundance. Rare codons (found in codon-biased genomes) generate a stochastic ribosomal pause while the corresponding low abundance tRNA is encountered. Following an elongation event, uncharged tRNAs are released from the ribosome to be recharged with amino acid by the corresponding aminoacyl tRNA synthetase. *Panel* **b**; Charged tRNA availability is influenced by tRNA demand/supply ratios. TRNA supply is determined tRNA abundance (see *panel* **a**), and by tRNA synthetase activity. Demand is dictated by the content of any given codon type within the transcriptome, the latter a response to environment. The supply/demand ratio for any given tRNA type may thus be balanced (*closed circles* and *open triangle*, transcriptome 1) or imbalanced (*open square*, transcriptome 1). With the induction of a new transcriptome, a different tRNA type may exhibit supply/demand imbalance (*open triangle*, transcriptome 2), altering the stochastic transit of ribosome populations along mRNAs, potentially directing changes in the translational efficiency of any given mRNA



Stochastic Modeling of Translation Elongation and Termination, Fig. 2 Sketch of the TASEP model, with the lattice representing the mRNA and the particles representing the ribosomes. The ribosome at codon $i = 4$ cannot advance, since the next codon is occupied

Stochastic Models and Statistical Physics; General Description of Motion of Ribosomes on a mRNA

Deterministic models have been widely used to model a broad range of biological systems. However, the complexity of cellular translation is better represented using stochastic models, for a number of reasons. First, the individual motion of one ribosome is a stochastic process, since it relies on the arrival of the correct transfer RNA molecule, which can be assumed to perform a random diffusive motion in the cytoplasm. Stochastic models take explicitly into account fluctuations due to the random movement of the molecules in the medium, enabling a more accurate description. Second, since several ribosomes are translating the same mRNA at the same time, we are confronted with a many-particle system. Differential equations are inadequate tools to describe such systems, which can be better represented by statistical physics (Krapivsky et al. 2010). Last, since new ribosomes can bind the mRNA as soon as the first nucleotides of the mRNA are unoccupied, the system is subjected to a continuous flow of particles, permitting the theory of nonequilibrium statistical physics to be readily applied.

TASEP-based Models to Describe Translation

Within nonequilibrium statistical physics, the so-called Totally Asymmetric Exclusion Process (TASEP) has been widely applied to the study of processes in which biological particles move along a track, for example in translation, motion of molecular motors, transcription, surface growth, and traffic (Krapivsky et al. 2010). It describes a one-dimensional lattice consisting of N sites, along which particles hop in one fixed direction with hopping rate k (▶ Lattice-Gas Cellular Automaton Models). The interaction among the particles is exclusion; particles cannot overlap. The particles hop onto the first site of the lattice (left boundary) with initiation rate α, then hop from site i of the lattice to site $i + 1$ with hoping rate k_i, only if site $i + 1$ is empty, and hop off the lattice at the last site (right boundary) with termination rate β. In the context of translation elongation, the one-dimensional lattice represents the mRNA molecule, the sites of the lattice represent the codons, and the particles represent the ribosomes (Fig. 2). The rate α is the initiation rate with which new ribosomes bind the ORF, starting the translation, and β describes the rate with which ribosomes detach from the lattice at the stop codon, releasing the completed protein into the cytoplasm. In a first approximation, assuming that the ribosomes occupy one single codon, occupation numbers n_i with $i = 1,\ldots, N$ are typically introduced, such that either $n_i = 0$ if codon i is empty or $n_i = 1$ if codon i is occupied. The particle density or average occupancy at codon i is then given by $i_f = \langle n_i \rangle_t$, where $\langle \cdots \rangle$ denotes time average. The system is characterized by the density of ribosomes averaged over the whole mRNA $\bar{\rho} = N^{-1} \sum_i \rho_i$ and the current J of ribosomes through the mRNA or the number of ribosomes per unit time which detach from the mRNA at the stop codon. Hence, the current J of ribosomes corresponds to the protein production rate of the underlying mRNA. In the simplest case in which all codons are equal, the hopping rates do not depend on the codon index, that is, $k_i = k$ for $i = 1,\ldots, N - 1$. Additionally, neglecting correlations between the occupation probabilities of different sites (▶ Mean-Field Approximation), one obtains the following master equations (▶ Master Equation):

$$\begin{aligned} \tfrac{d\rho_1}{dt} &= \alpha(1-\rho_1) - k\rho_1(1-\rho_2) \\ \tfrac{d\rho_i}{dt} &= k\rho_{i-1}(1-\rho_i) - k\rho_i(1-\rho_{i+1}) \\ \tfrac{d\rho_N}{dt} &= k\rho_{N-1}(1-\rho_N) - \beta\rho_N \end{aligned} \qquad (1)$$

Assuming that we reach steady state, that is, $d\rho_i/dt = 0$ for all i, one identifies the current of particles: $J = k\rho_i(1 - \rho_{i+1})$. Notice that since particles cannot detach from the lattice until they reach the stop codon, the current is conserved through the lattice. Depending on the relative magnitude of the parameters α, β, and k, it can be shown that the system can exist in three different phases: low density (LD), where the current is limited by the initiation rate α, high density (HD), where the rate limiting factor is the termination rate β, and the maximal current phase (MC), where the hopping rate k limits the current of the process. These phases are characterized as follows:

$$\begin{aligned}
\text{LD}: &\quad J=\alpha\left(1-\frac{\alpha}{k}\right), \quad \rho=\frac{\alpha}{k}, \quad \text{if } \alpha<\beta, \alpha\leq\frac{k}{2}\\
\text{HD}: &\quad J=\beta\left(1-\frac{\beta}{k}\right), \quad \rho=1-\frac{\beta}{k}, \quad \text{if } \beta<\alpha, \beta\leq\frac{k}{2}\\
\text{MC}: &\quad J=\frac{k}{4}, \quad \rho=\frac{1}{2}, \quad \text{if } \alpha,\beta>\frac{k}{2}
\end{aligned}$$

Here ρ denotes the density in the bulk of the lattice. Due to edge effects, i_f deviates from ρ near $i = 1$ and $i = N$, but if the lattice is large enough, $\bar{\rho} \approx \rho$. Moreover, a mixed LD-HD phase is found when $\alpha = \beta$ and $\alpha, \beta \leq k/2$, which is also denoted as shock phase (SP). The transition between the LD and HD phase is of first order in J, whereas between LD and MC, and HD and MC is of second order (Krapivsky et al. 2010).

Computer Simulation

In order to simulate the TASEP computationally, one can use either a discrete time simulation approach (▶ Monte Carlo Simulation) or a continuous time simulation, following, for example, the Gillespie algorithm (▶ Gillespie Stochastic Simulation). In the typical discrete time simulation, one chooses a Monte-Carlo time step Δt, in which the particles on the lattice are updated according to the rules of the TASEP. A site i of the lattice is picked at random, with $i = 0,\ldots, N$. If $1 \leq i \leq N - 1$, and if the site is occupied, the particle in it has probability $k\Delta t$ of hopping to the right, given that the next site is empty. If $i = 0$, a new particle can hop onto the lattice with probability $\alpha\Delta t$, given the first site is empty. If $i = N$ and that site is occupied, then the particle can hop off the lattice with probability $\beta\Delta t$. The other possibility is to apply a continuous time simulation algorithm, in which both the particle and the time at which the next update will happen are chosen according to a probability distribution which can be derived using probabilistic arguments (Gillespie 1977).

Inhomogenous Lattice: Different tRNAs are in Different Concentrations

As explained above, different tRNAs are present in the cytoplasm in different concentrations. Therefore, when a ribosome is on a codon whose corresponding tRNA has a low concentration, it will take the ribosome a longer time to find it, and therefore, to proceed to the next codon. In terms of the lattice model, it means that the hopping rate depends on the site of the lattice; the mRNA is then represented as a lattice with hopping rates k_i, $i = 1,\ldots, N$. In a first approximation, the hopping rates k_i can be estimated according to the abundances of the corresponding tRNAs. However, the abundances of all 41 different species of tRNAs are not known. Therefore, gene copy numbers of the tRNAs are typically used as predictors for their abundances, since tRNA gene copy numbers have been found to strongly correlate with their abundances. To understand the effect of slow codons, lattices with one single slow codon in the center were first studied (Kolomeisky 1998). The lattice can be then modeled as two separate sub-lattices connected across the slow site; an effective termination rate is assigned to the left sub-lattice, and an effective initiation rate to the right sub-lattice. Then, since the current is conserved across the slow site, mean-field expressions for the current and bulk density can be obtained for the four possible phases of the combined system: LD/LD, HD/HD, MC/MC, and HD/LD. This last combined phase is also referred to as queueing phase (QP). Studies including multiple slow codons show that not only the total number of slow codons plays a role, but its configuration; the current is maximally reduced when slow codons are clustered as tight as possible (Chou and Lakatos 2004). Moreover, a comprehensive study using configurations of slow codons from real mRNA sequences from *Saccharomyces cerevisiae* has shown that mRNAs can be classified into two main types, according to the type of transition they are subjected to when the initiation rate α is increased: abrupt transition (type-I) and smooth transition (type-II). Importantly, sequences classified as type-II share a common biological function, such as ribosomal proteins. This indicates that as the initiation rate α changes (which is linked to the availability of ribosomes in the cytoplasm), the translation rate of proteins is regulated (▶ Gene Regulation) in a different way according to the need of the cell for that protein (Romano et al. 2009).

Summary

In summary, lattice-gas models of translation are powerful analytical tools, since they condense the underlying essential biological mechanisms using a few simple rules, but can reproduce rich and complex behaviors. Lattice-gas models are likely to be increasingly used to understand and predict the complex relationship between the codon composition of an mRNA, and its translational efficiency.

Cross-References

- ▶ Ergodicity
- ▶ Gene Regulation
- ▶ Gillespie Stochastic Simulation
- ▶ Lattice-Gas Cellular Automaton Models
- ▶ Markov Process
- ▶ Master Equation
- ▶ Mathematical Model, Model Theory
- ▶ Mean-Field Approximation
- ▶ Monte Carlo Simulation
- ▶ Release Factor, Translation
- ▶ Translation Elongation
- ▶ Translation Termination
- ▶ Translational Control

References

Chou T, Lakatos G (2004) Clustered bottlenecks in mRNA translation and protein synthesis. Phys Rev Lett 93:1981011–1981014

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 8:2340–2361

Gingold H, Pilpel Y (2011) Determinants of translation efficiency and accuracy. Mol Syst Biol 7:481

Kapp LD, Lorsch JR (2004) The molecular mechanics of eukaryotic translation. Annu Rev Biochem 73:657–704

Kolomeisky AB (1998) Asymmetric simple exclusion model with local inhomogeneity. J Phys A Math Gen 31:1153–1164

Krapivsky PL, Redner S, Ben-Naim E (2010) A kinetic view of statistical physics. Cambridge University Press, Cambridge

Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12:32–42

Romano MC, Thiel M, Stansfield I, Grebogi C (2009) Queueing phase transition: theory of translation. Phys Rev Lett 102:1981041–1981044

Stochastic Neural Network

Zhong-Yuan Zhang
School of Statistics, Central University of Finance and Economics, Beijing, China

Definition

Though noise is considered as the by-product of life activities, it often plays important roles in life activities, for example, gene regulatory networks are inherently noisy. Furthermore, noise contamination in experimental data is inevitable due to the limitations of technology. Hence, stochastic model conforms to reality better than the classical ones.

Stochastic neural network model introduces stochastic processes to describe the biochemical process in the gene regulatory networks; in other words, given the expression state vector $u(t)$ where $u_i(t)$ is the expression value of gene i at time t, the model predicts the state vector $u(t+\Delta t)$ at the next time point $t+\Delta t$ as follows:

$$u_i(t+\Delta t) = u_i(t) + P(\Delta t m_i f(x)) - P(\Delta t d_i u_i(t)),$$

where $P(\lambda)$ is a random variable following certain distribution, $f(x)$ is the sigmoid function of the sum of $u'_i s$ received weighted inputs, d_i is the degradation rate of gene i, and m_i is the maximal expression value of gene i.

Cross-References

- ▶ Identification of Gene Regulatory Networks, Neural Networks

References

Tian T, Burrage K (2003) Stochastic neural network models for gene regulatory networks. In: CEC 2003: the 2003 congress on evolutionary computation. Proceedings, Canberra, 8–12 Dec 2003. IEEE Computer Society, Piscataway, USA, pp 162–169

Stochastic pi-Calculus

Alida Palmisano[1] and Corrado Priami[2]
[1]Department of Biological Sciences and Department of Computer Science, Department of Biological Sciences Virginia Tech, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[2]Microsoft Research-University of Trento Centre for Computational and Systems Biology and DISI, University of Trento, Povo, Trento, Italy

Synonyms

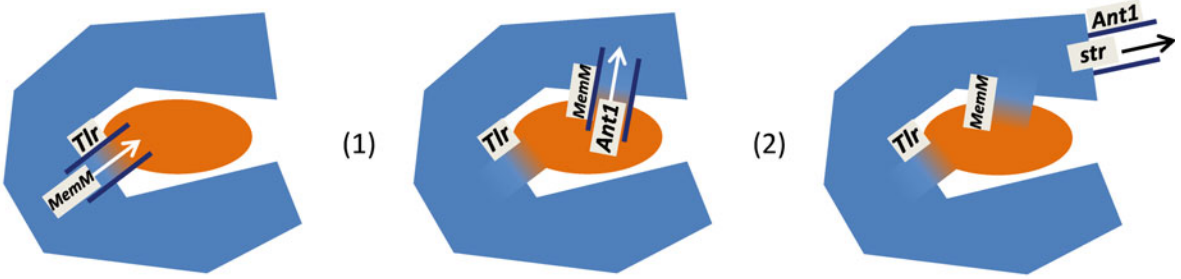
Biochemical pi-calculus; Stochastic π-calculus

Definition

Stochastic pi-calculus (Priami 1995) is the first process algebra used to represent biological systems (Priami et al. 2001). Molecules are modeled as processes, and molecular complexes are rendered by parallel compositions of processes sharing private names. Movements between complexes and formations of new complexes are represented as transmissions of private names. Once a complex is formed, its components interact by communicating on complementary sites.

Moreover, since the calculus is stochastic, the behavior of biological system can also be described and analyzed quantitatively. Two simulators for the biochemical stochastic pi-calculus exist that implement the direct method of the ▶ stochastic simulation algorithm (▶ BioSPI).

In Guerriero et al. (2009), the authors present a simple and nice example of a biological system modeled with biochemical pi-calculus that we want to report here for showing the language on a specific case study (we refer the reader to Priami et al. [2001] for a formal description of the language).



System specification:

```
              SYS ::= MACROPHAGE | VIRUS | TCELL1 | TCELL2
              MACROPHAGE ::= (νMemM)(Tlr<MemM>.MemM(a).! a<str>)
              VIRUS ::= (νMemV)(Tlr(y).y<Ant1>)
```

System evolution:

```
              MACROPHAGE | VIRUS ::= (νMemM)(Tlr<MemM>.Mem(a).! a<str>) | (νMemV)(Tlr(y).y<Ant1>)
                (1) (νMemM)(MemM(a).! a<str>) | (νMemV)(MemM<Ant1>)
                (2) (νMemM)(! Ant1<str>) | nil
```

Stochastic pi-Calculus, Fig. 1 The biochemical pi-calculus code fragment that specifies the antigen presentation phase. The global system SYS is given by the parallel composition of four processes: VIRUS (*orange oval* in the upper part of the figure), MACROPHAGE (*blue c-shape* in the upper part of the figure), TCELL1, and TCELL2. The code only presents the specifications of the first two elements. Here we just sketch the intuition of the behavior of the subsystem given by MACROPHAGE | VIRUS. The restriction (*ν*) on top of each component stands for its enclosing membrane. The macrophage phagocytizes the virus by means of a communication on the public channel *Tlr*. Operationally, this communication involves the output action Tlr<MemM> and its complementary input action Tlr(y). Its effect is twofold: (1) the restricted name MemM becomes a private communication channel of both MACROPHAGE and VIRUS (thus modeling the engulfment of the virus); (2) the name y in VIRUS is renamed into MemM (modeling the adaptation of the internal machinery of the macrophage to start the lysis). The subsequent communication over the channel MemM is such that Ant1 is transmitted to MACROPHAGE, which in turn can make Ant1 available to the lymphocytes T (either TCELL1 or TCELL2) by means of the last action Ant1<str>. The *bang* operator, "!", allows us to model infinite behavior meaning that the output signal is continuously sent and therefore MACROPHAGE can activate many TCells expressing Ant1

The example comes from the biology of the immune system, and it is relative to the activation of the lymphocyte T helper, which are eukaryote cells belonging to our immune system. They play a central role by controlling many specific defense strategies. Lymphocytes are normally inactive but they can be activated by macrophages. Macrophages are cells that engulf a virus (phagocytosis). When this happens, the virus is degraded into fragments (digestion or lysis), and a molecule, the so-called antigen, is displayed on the surface of the macrophage (presentation or mating). The antigen may be recognized by a specific lymphocyte T helper, and this in turn activates the mechanisms of immune reply, a response specific to the recognized virus (Fig. 1).

Associating rates to the different reactions (i.e., communication channels) of the model described above creates a model that can be analyzed quantitatively through stochastic simulations.

Other interesting applications of biochemical pi-calculus on real biological scenarios include gene regulatory and metabolic networks, autoreactive lymphocyte recruitment, cell cycle (▶ Cell Cycle Modeling, Process Algebra).

Cross-References

▶ Cell Cycle Modeling, Process Algebra

References

Guerriero ML, Prandi D, Priami C, Quaglia P (2009) Process calculi abstractions for biology. In: Condon A, Harel D (eds) Process calculi abstractions for biology in algorithmic bioprocesses. Springer, Heidelberg

Priami C (1995) Stochastic pi-calculus. Comput J 38(6): 578–589

Priami C, Regev A, Silverman W, Shapiro E (2001) Application of a stochastic name-passing calculus to representation and simulation of molecular processes. Inf Process Lett 80(1):25–31

Stochastic pi-Calculus Simulator

▶ BioSPI

Stochastic Processes, Fokker-Planck Equation

Hong Qian[1] and Hao Ge[2]
[1]Department of Applied Mathematics, University of Washington, Seattle, WA, USA
[2]School of Mathematical Sciences and Centre for Computational Systems Biology, Fudan University, Shanghai, China

Synonyms

Diffusion approximation to chemical master equation; Diffusion processes; Kolmogorov forward equation; Smoluchowski equation; Stochastic differential equation

Definition

The Fokker–Planck equation describes the time evolution of the probability density function of the position of a particle that follows a stochastic differential equation. It is assumed that the sample trajectories of the particle are continuous functions of time; but they are nowhere differentiable with respect to time. It is a generalization of the diffusion equation with the presence of a drift force field. It is named after A. Fokker and M. Planck; It is also known as the Kolmogorov forward equation, named after A. Kolmogorov. The first use of the Fokker–Planck equation was for the statistical description of Brownian motion of a particle in a fluid, independently, by A. Einstein and M. von Smoluchowski. Diffusion motion can be considered as a limiting case of biased random walk.

In one spatial dimension x, the Fokker–Planck equation for a diffusion process with drift $B(x, t)$ and diffusion $A(x, t)$ is:

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}[B(x,t)p(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[A(x,t)p(x,t)] \tag{1}$$

If drift $B(x, t)$ and diffusion $A(x, t)$ do not depend on time t, then it is called time-homogeneous (or simply homogeneous) Fokker–Planck equation. In this case, let $p(y, t|x, 0)$ be the conditional probability density observing y at time t, starting from x at time 0, then

B and A are determined by the rates of growth of the mean value and the variance:

$$\begin{aligned} B(x) &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{-\infty}^{+\infty}(y-x)p(y,t|x,0)dy,\\ A(x) &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{-\infty}^{+\infty}(y-x)^2 p(y,t|x,0)dy \end{aligned} \tag{2}$$

If the diffusion $A(x,\ t)\ =\ 0$, then the equation is reduced to a simple ordinary differential equation with a smooth trajectory:

$$\frac{dx}{dt} = B(x,t) \tag{3}$$

More generally, the time-dependent probability distribution may depend on a set of N macrovariables x_i. The general form of the Fokker–Planck equation is then:

$$\begin{aligned} \frac{\partial}{\partial t}p(x,t) = &-\sum_{i=1}^{N}\frac{\partial}{\partial x_i}[B_i(x,t)p(x,t)]\\ &+\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\partial^2}{\partial x_i\partial x_j}[A_{ij}(x,t)p(x,t)] \end{aligned} \tag{4}$$

where B is the drift vector and A the diffusion tensor; the latter results from the presence of a particular type of stochastic effect: the white noise.

Any stochastic process can be characterized by two very different types of mathematics: either by its random sample trajectories or by the probability distribution as a function of time. The Fokker–Planck equation is the latter: It gives the time-dependent probability density function for the random trajectories described by stochastic differential equations.

Characteristics

We shall only consider time-homogeneous case. According to the mass conservation law, which can be applied to the probability:

$$\frac{\partial}{\partial t}p(x,t) = -\nabla\cdot J(x,t), \tag{5}$$

where $\nabla \bullet J(x,t) = \sum_{i=1}^{N}\frac{\partial}{\partial x_i}J_i(x,t)$ is the divergence of the vector field $J(x,t)$, then

$$J_i(x,t) = B_i(x)p(x,t) - \frac{1}{2}\sum_{j=1}^{N}\frac{\partial}{\partial x_j}A_{ij}(x)p(x,t) \tag{6}$$

In case the distribution $p(x,\ t)$ at long times approach a stationary distribution $p_{ss}(x)$, we must have time-independent J_{ss} as well. The J_{ss} is not zero in general. This gives rise to the classification of equilibrium stationary state (or steady state) with $J_{ss} = 0$, and nonequilibrium steady state (NESS) which has nonzero J_{ss}.

For equilibrium steady state, $J_{ss} = 0$ leads to:

$$B_i(x)p_{eq}(x) = \frac{1}{2}\sum_{j=1}^{N}\frac{\partial}{\partial x_j}A_{ij}(x)p_{eq}(x) \tag{7}$$

This can be written as:

$$\begin{aligned} \tilde{B}_i(x)p_{eq}(x) &= \frac{1}{2}\sum_{j=1}^{N}A_{ij}(x)\frac{\partial}{\partial x_j}p_{eq}(x),\\ \tilde{B}_i(x) &= B_i(x) - \sum_{j=1}^{N}\frac{\partial A_{ij}(x)}{\partial x_j} \end{aligned} \tag{8}$$

We then see that the drift $B(x)$ and diffusion $A(x)$, when the latter is nonsingular, satisfy a condition:

$$2\sum_{i=1}^{N}\left(A^{-1}(x)\right)_{ji}\tilde{B}_i(x) = \frac{\partial}{\partial x_j}\ln p_{eq}(x) \tag{9}$$

The vector field on the left-hand side has a potential function. In the case of one-dimensional system:

$$p_{eq}(x) \propto \exp\left(\int^{x}\frac{2\tilde{B}(y)}{A(y)}dy\right) \tag{10}$$

In many cases, because of the gradient nature of the drift, the stationary distribution for an equilibrium steady state can be obtained even for high-dimensional systems. However, for nonequilibrium steady state, this is a much more difficult task because of the presence of nonzero J_{ss}, which makes the problem nonlocal. Nongradient vector field has a circular component which is intimately related to the J_{ss}.

Some Applications of Fokker–Planck Equations

Beyond the classic work of Einstein and Smoluchowski, H.A. Kramers was the first one who used a Fokker–Planck equation to study discrete chemical reactions in terms of diffusion in a potential function with an activation barrier, and derived his famous rate formula. This work laid the foundation for the physical basis of chemical reactions in condensed phase.

When the potential force is harmonic, the Fokker–Planck equation yields the simple Gaussian Markov process also known as Ornstein–Uhlenbeck process (Wax 1954). See an earlier, extensive review by R.F. Fox. A major application of the multidimensional Fokker–Planck equation is in the theory for polymer dynamics (Doi and Edwards 1986).

D. Shoup and A. Szabo later developed a unified treatment of diffusion controlled chemical reaction studied by Smoluchowski and transition-state controlled reaction studied by Kramers.

Diffusion Approximation to Chemical Master Equation

Traditionally, a Fokker–Planck equation, or its corresponding stochastic differential equation, is developed for macroscopic continuous motion of dynamic variables with fluctuations. Such an equation is used to explain small fluctuations around a deterministic mean dynamics of a macroscopic system. Recently, a different derivation of the Fokker–Planck equation from a mesoscopic nonequilibrium thermodynamic theory has been developed by Reguera et al. 2005. In systems biology, however, there is a very different origin for Fokker–Planck equations. They are the macroscopic, continuous limit of stochastic population dynamics. The population can be number of molecules in biochemical reaction systems, or number of cells in an organism, or number of individual in an ecological setting, etc. These stochastic dynamics are known as birth-and-death processes in the theory of probability. The process takes nonnegative integer values. When the number of individual is sufficiently large, the discrete increments can be approximated by a continuous variables $X(t)$, in real numbers.

The chemical master equation is the stochastic theory for discrete biochemical reaction systems in a finite volume. If in a biochemical system, the number of all the components, $X(t) = (X_1(t), X_2(t), \ldots, X_N(t))$, are very large compared to 1, we can regard the components of $X(t)$ as real numbers. In the discrete model, let $a_j(\mathrm{x})$ be the transition rate for reaction j, changing the number of molecules from $x = (x_1, x_2, \ldots, x_N)$ to $x + v_j = (x_1 + v_{j1}, x_2 + v_{j2}, \ldots, x_N + v_{jN})$. Then one can carry out multivariate Taylor's expansion for the functions $f_j(x) \equiv a_j(x)P(x, t)$ if they are analytic in the real variable x:

$$f_j(x - v_j) = f_j(x) + \sum_{|m|\geqslant 1} \prod_{i=1}^{N} \frac{(-1)^{m_i}}{m_i!} \left(\frac{\partial}{\partial x_i}\right)^{m_i} f_j(x)\ v_{ji}^{m_i} \tag{11}$$

Here $m = (m_1, \ldots, m_N) \in Z^N$, denoting the nonnegative, N-dimensional integer space (lattice). $|m| = m_1 + m_2 + \ldots + m_N$. Substituting Eq. 11 into the chemical master equation, we obtain the Kramers–Moyal expansion:

$$\frac{\partial}{\partial t}P(x,t) = \sum_{|m|\geqslant 1} \prod_{i=1}^{N} \frac{(-1)^{m_i}}{m_i!} \left(\frac{\partial}{\partial x_i}\right)^{m_i} (A^m(x)P(x,t)) \tag{12}$$

where

$$A^m(x) = \sum_{j=1}^{M} v_{j1}^{m_1} v_{j2}^{m_2} \cdots v_{jN}^{m_N} a_j(x). \tag{13}$$

If one truncates the right-hand side at $|m| = 2$, we obtain the Fokker–Plank approximation (or diffusion approximation) of the chemical master equation:

$$\begin{aligned} \frac{\partial}{\partial t}p(x,t) = &-\sum_{i=1}^{N} \frac{\partial}{\partial x_i}[B_i(x)p(x,t)] \\ &+ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial^2}{\partial x_i \partial x_j}[A_{ij}(x)p(x,t)] \end{aligned} \tag{14}$$

Where

$$B_i(x) = \sum_{j=1}^{M} v_{ji}a_j(x),\ \ A_{ij}(x) = \sum_{k=1}^{M} v_{ki}v_{kj}a_k(x). \tag{15}$$

The corresponding stochastic differential equations associated with this chemical Fokker–Plank equation is called the chemical Langevin equation.

A critique on the diffusion approximation. For many chemical reaction systems described by the Law of Mass Action and nonlinear kinetic equations, there are multiple steady states. For reaction systems with multiple steady states, the Fokker–Planck equation is not a faithful global approximation to the chemical master equation. Simple, explicit examples are known to show the diffusion approximation gives wrong stationary distribution with respect to two steady states, represented as two local maxima in the distribution function. On the other hand, van Kampen has shown that the Kramers–Moyal expansion is not a cogent method of approximation. The more cogent system-size expansion method developed by van Kampen is based on conditional variance of a dynamic process and it yields a time-inhomogeneous Fokker–Planck equation.

However, these methods are both local rather than global; they do not allow for the determination of global features such as multistability and the decay of metastable states, i.e., barrier crossing. Note that while the system-size expansion is mathematically more cogent, it still does not give a global view of multistability since it could only be applied in a single basin of attractor for the deterministic system depending on the initial state.

What can be guaranteed, from both approaches, is a faithful fluctuation description near a given steady state which is not absorbing. In other words, a Gaussian approximation. Note this is precisely the static and dynamic fluctuation theory developed by A. Einstein and L. Onsager, respectively. However, if one is interested in the relative importance (i.e., stationary probability) among different possible steady states and their transition kinetics, a more careful analysis is required. J. Keizer showed a deterministic unstable fixed point can have probability 1 according to the CME; and P. Hanggi et al. showed the relative probability of two stable steady states could be inverted in the stationary solution from the approximated Fokker–Planck equation for some range of the parameters. These observations were later termed "Keizer's paradox"(Vellela and Qian 2007). C. Knessel et al. pointed out that the mean first passage times between different states can differ by many orders of magnitude depending on which approach is used. F. Baras et al. used the microscopic simulation to demonstrate the failure of the diffusion approximation approach, but showed the chemical master equation to be in excellent agreement with the simulations.

The Keizer's paradox is due to the following origin: As an approximation to the CME, the Fokker–Planck equation is expected to be valid near a solution to the deterministic dynamics. But dynamic going uphill is impossible in the deterministic dynamics. Therefore, the Fokker–Planck equation gives poor approximation for any processes involves uphill dynamics. In fact, with multiple stable steady states, moving away from one has a very small probability; but with probability 1 it will occur. The catch is the time it takes is usually astronomically long.

Cross-References

- ▶ Chemical Master Equation
- ▶ Stochastic Differential Equation

References

Baras F, Malek-Mansour M, Pearson JE (1996) Microscopic simulation of chemical bistability in homogeneous systems. J Chem Phys 105:8257–8261

Doi M, Edwards SF (1986) The theory of polymer dynamics. Clarendon Press, Oxford

Fox RF (1978) Gaussian stochastic processes in physics. Phys Rep 48:180–283

Gillespie DT (2002) The chemical Langevin and Fokker-Planck equations for the reversible isomerization reaction. J Phys Chem A 106:5063–5071

Hanggi P, Grabert H, Talkner P, Thomas H (1984) Bistable systems: master equation versus Fokker-Planck modeling. Phys Rev A 29:371–378

Knessel C, Mangel M, Matkowsky BJ, Schuss ZC (1984) Tier: solution of Kramers-Moyals equations for problems in chemical physics. J Chem Phys 81:1285–1293

Kramers HA (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. Physica 7:284–304

Nicolis G, Lefever R (1977) Comment on the kinetic potential and the Maxwell construction in non-equilibrium chemical phase transitions. Phys Lett A 62:469

Onsager L, Machlup S (1953) Fluctuatios and irreversible processes. Phys Rev 91:1505–1512

Qian H (2006) Open-system nonequilibrium steady-state: statistical thermodynamics, fluctuations and chemical oscillations. J Phys Chem B 110:15063–15074 (Feature Article)

Qian H (2011) Nonlinear stochastic dynamics of mesoscopic homogeneous biochemical reactions systems – An analytical theory. Nonlinearity 24:R19–R49 (Invited Article)

Reguera D, Rubi JM, Vilar JMG (2005) The mesoscopic dynamics of thermodynamic systems. J Phys Chem B 109:21502–21515 (Feature Article)

Risken H (1989) The Fokker–Planck equation: methods of solutions and applications, 2nd edn, Springer series in synergetics. Springer, Berlin

Shoup D, Szabo A (1982) Role of diffusion in ligand binding to macromolecules and cell-bound receptors. Biophys J 40:33–39

van Kampen NG (1961) A power series expansion of the master equations. Can J Phys 39:551–567

van Kampen NG (1981) Stochastic processes in physics and chemistry. North-Holland, Amsterdam

Vellela M, Qian H (2007) A quasistationary analysis of a stochastic chemical reaction: Keizer's paradox. Bull Math Biol 69:1727–1746

Wax N (1954) Selected papers on noise and stochastic processes. Dover, New York

Stochastic Resonance

Tianshou Zhou
School of Mathematics and Computational Sciences,
Sun Yet-Sen University, Guangzhou, Guangdong,
China

Definition

Stochastic resonance (SR) is observed when noise added to a system changes the system's behavior in some fashion. More technically, SR occurs if the signal-to-noise ratio of a nonlinear system or device increases for moderate values of noise intensity. It often occurs in bistable systems or in systems with a sensory threshold and when the input signal to the system is "subthreshold." For lower noise intensities, the signal does not cause the device to cross threshold, so little signal is passed through it. For large noise intensities, the output is dominated by the noise, also leading to a low signal-to-noise ratio. For moderate intensities, the noise allows the signal to reach threshold, but the noise intensity is not so large as to swamp it. Thus, a plot of signal-to-noise ratio as a function of noise intensity shows a "∩" shape.

Strictly speaking, stochastic resonance occurs in bistable systems, when a small periodic (sinusoidal) force is applied together with a large wide band stochastic force (noise). The system response is driven by the combination of the two forces that compete/cooperate to make the system switch between the two stable states. The degree of order is related to the amount of periodic function that it shows in the system response. When the periodic force is chosen small enough in order to not make the system response switch, the presence of a non-negligible noise is required for it to happen. When the noise is small, very few switches occur, mainly at random with no significant periodicity in the system response. When the noise is very strong, a large number of switches occur for each period of the sinusoid and the system response does not show remarkable periodicity. Between these two conditions, there exists an optimal value of the noise that cooperatively concurs with the periodic forcing in order to make almost exactly one switch per period (a maximum in the signal-to-noise ratio).

Such a favorable condition is quantitatively determined by the matching of two time scales: the period of the sinusoid (the deterministic time scale) and the Kramers rate (i.e., the inverse of the average switch rate induced by the sole noise: the stochastic time scale). This is the term "stochastic resonance."

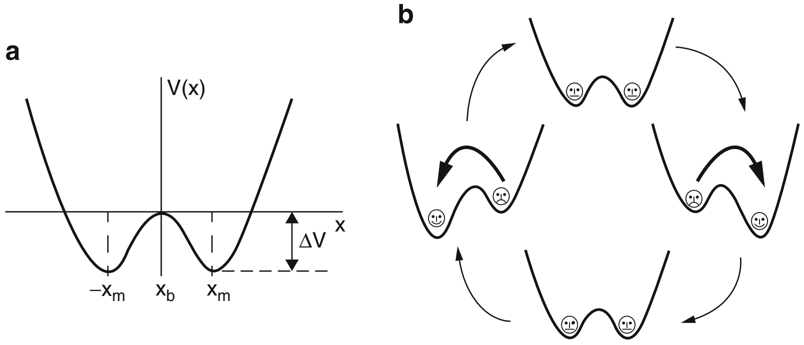
Stochastic resonance was discovered and proposed for the first time in 1981 to explain the periodic recurrence of ice ages. Since then the same principle has been applied in a wide variety of systems. Nowadays stochastic resonance is commonly invoked when noise and nonlinearity concur to determine an increase of order in the system response.

The mechanism of stochastic resonance is simple to explain. Consider a heavily damped particle of mass m and viscous friction γ, moving in a symmetric double-well potential $V(x)$ (see Fig. 1a). The particle is subject to fluctuational forces that are, for example, induced by coupling to a heat bath. Such a model is archetypal for investigations in reaction-rate theory. The fluctuational forces cause transitions between the neighboring potential wells with a rate given by the famous Kramers rate, that is,

$$r_K = \frac{\omega_0 \omega_b}{2\pi\gamma} \exp\left(-\frac{\Delta V}{D}\right) \tag{1}$$

with $\omega_0^2 = V''(x_m)/m$ being the squared angular frequency of the potential in the potential minima at $\pm x_m$, and $\omega_b^2 = |V''(x_b)/m|$ the squared angular frequency at the top of the barrier, located at x_b; ΔV is the

Stochastic Resonance, Fig. 1 Stochastic resonance in a symmetric double well: (**a**) sketch of the double-well potential; (**b**) description for the cyclic variation



height of the potential barrier separating the two minima. The noise strength $D = k_B T$ is related to the temperature T.

If we apply a weak periodic forcing to the particle, the double-well potential is tilted asymmetrically up and down, periodically raising and lowering the potential barrier, as shown in Fig. 1b. Although the periodic forcing is too weak to let the particle roll periodically from one potential well into the other one, noise-induced hopping between the potential wells can become synchronized with the weak periodic forcing. This statistical synchronization takes place when the average waiting time $T_K(D) = 1/r_K$ between two noise-induced interwell transitions is comparable with *half* the period T_Ω of the periodic forcing. This yields the *time-scale matching condition* for stochastic resonance, that is,

$$2T_K(D) = T_\Omega \qquad (2)$$

In short, stochastic resonance in a symmetric double-well potential manifests itself by a synchronization of activated hopping events between the potential minima with the weak periodic forcing (Gammaitoni et al. 1989). For a given period of the forcing T_Ω, the time-scale matching condition can be fulfilled by tuning the noise level $D_{\max}$ to the value determined by Eq. 2.

In order to better understand some definitions, let us consider the overdamped motion of a Brownian particle in a bistable potential in the presence of noise and periodic forcing:

$$\frac{dx(t)}{dt} = \dot{x}(t) = -V'(x) + A_0 \cos(\Omega t) + \xi(t) \qquad (3)$$

where $V(x)$ denotes the reflection-symmetric quartic potential:

$$V(x) = -\frac{a}{2}x^2 + \frac{b}{4}x^4 \qquad (4)$$

$\xi(t)$ denotes a zero mean, Gaussian white noise with autocorrelation function:

$$\langle \xi(t)\xi(0) \rangle = 2D\delta(t) \qquad (5)$$

and intensity D. The potential $V(x)$ is bistable with minima located at $\pm x_m$.

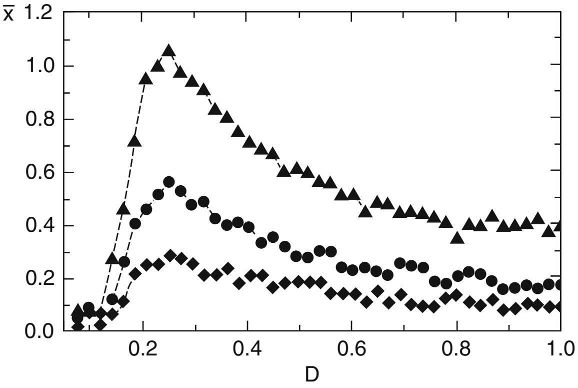
Characteristics

Periodic Response

For convenience, we choose the phase of the periodic driving $\varphi = 0$, that is, the input signal reads explicitly $A(t) = A_0 \cos(\Omega t)$ (see Eq. 3). The mean value $\langle x(t)|x_0, t_0\rangle$ is obtained by averaging the inhomogeneous process $x(t)$ with initial conditions $x_0 = x(t_0)$ over the ensemble of the noise realizations. Asymptotically ($t_0 \to -\infty$), the memory of the initial conditions gets lost and $\langle x(t)|x_0, t_0\rangle$ becomes a periodic function of time, that is, $\langle x(t)\rangle_{as} = \langle x(t + T_\Omega)\rangle_{as}$ with $T_\Omega = 2\pi/\Omega$. For small amplitudes, the response of the system to the periodic input signal can be written as $\langle x(t)\rangle_{as} = \bar{x}\cos(\Omega t - \bar{\phi})$ with amplitude $\bar{x}$ and a phase lag $\bar{\phi}$. The relationship between $\bar{x}$ and D is called response curve (refer Fig. 2).

Signal-to-Noise Ratio

Signal-to-noise ratio (often abbreviated *SNR* or *S/N*) is a measure used in science and engineering to quantify



Stochastic Resonance, Fig. 2 Response curves for several input signals with different amplitudes

how much a signal has been corrupted by noise. It is defined as the ratio of signal power to the noise power corrupting the signal. A ratio higher than 1:1 indicates more signal than noise. While SNR is commonly quoted for electrical signals, it can be applied to any form of signal (such as isotope levels in an ice core or biochemical signaling between cells).

In less technical terms, signal-to-noise ratio compares the level of a desired signal (such as music) to the level of background noise. The higher the ratio, the less obtrusive the background noise is. "Signal-to-noise ratio" is sometimes used informally to refer to the ratio of useful information to false or irrelevant data in a conversation or exchange. For example, in online discussion forums and other online communities, off-topic posts and spam are regarded as "noise" that interferes with the "signal" of appropriate discussion.

Signal-to-noise ratio is defined as the power ratio between a signal (meaningful information) and the background noise (unwanted signal):

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{6}$$

where P is average power. Both signal and noise power must be measured at the same or equivalent points in a system, and within the same system bandwidth. If the signal and the noise are measured across the same impedance, then the SNR can be obtained by calculating the square of the amplitude ratio:

$$SNR = \frac{P_{signal}}{P_{noise}} = \left(\frac{A_{signal}}{A_{noise}}\right)^2 \tag{7}$$

where A is root mean square (RMS) amplitude (e.g., RMS voltage). Because many signals have a very wide dynamic range, SNRs are often expressed using the logarithmic decibel scale. In decibels, the SNR is defined as

$$SNR_{dB} = 10 \log_{10} \left(\frac{P_{signal}}{P_{noise}}\right) = P_{signal,dB} - P_{noise,dB} \tag{8}$$

which may equivalently be written as

$$SNR_{dB} = 10 \ \log_{10} \left(\frac{A_{signal}}{A_{noise}}\right)^2 = 20 \ \log_{10} \left(\frac{A_{signal}}{A_{noise}}\right) \tag{9}$$

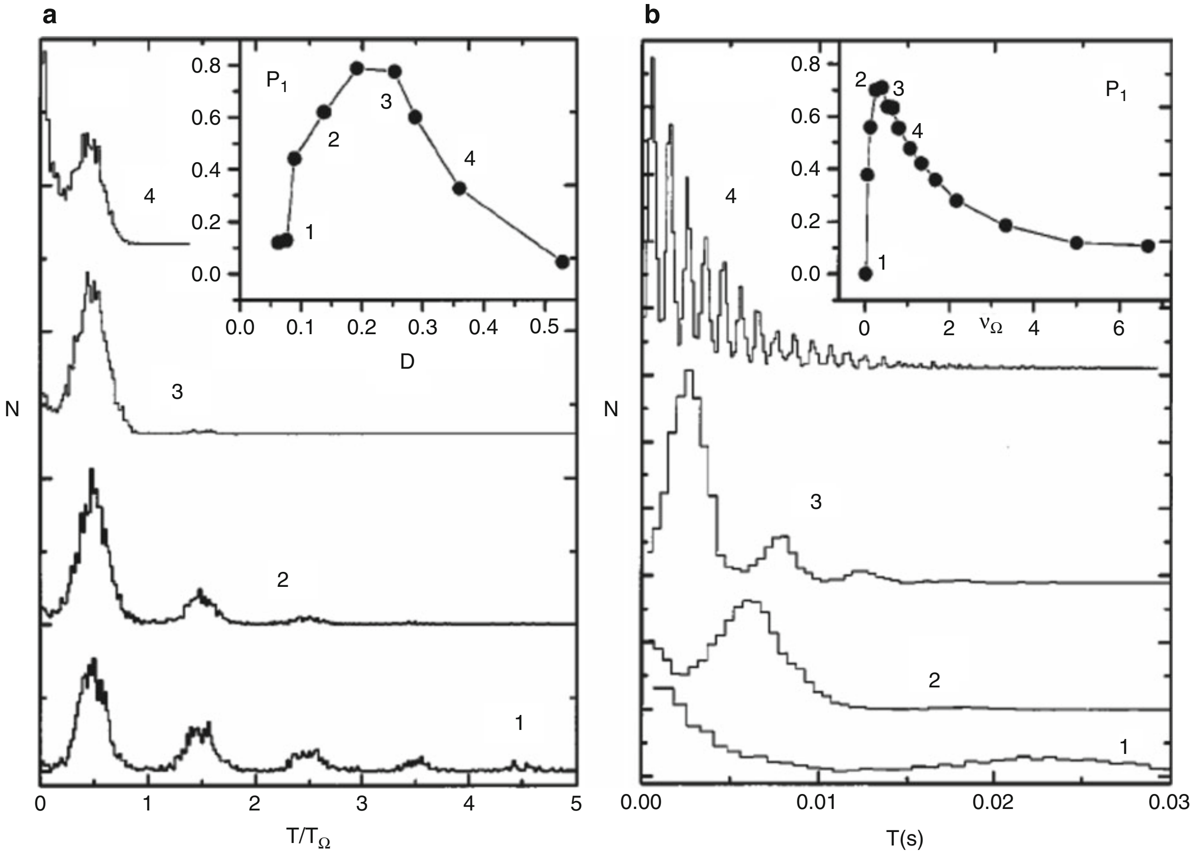
The concepts of signal-to-noise ratio and dynamic range are closely related. Dynamic range measures the ratio between the greatest undistorted signal on a channel and the smallest detectable signal, which for most purposes is the noise level. SNR measures the ratio between an arbitrary signal level (not necessarily the most powerful signal possible) and noise. Measuring signal-to-noise ratios requires the selection of a representative or *reference* signal. In audio engineering, the reference signal is usually a sine wave at a standardized nominal or alignment level, such as 1 kHz at +4 dBu (1.228 V_{RMS}).

SNR is usually taken to indicate an *average* signal-to-noise ratio, as it is possible that (near) instantaneous signal-to-noise ratios will be considerably different. The concept can be understood as normalizing the noise level to 1 (0 dB) and measuring how far the signal "stands out."

An alternative definition of SNR is as the reciprocal of the coefficient of variation, that is, the ratio of mean to standard deviation of a signal or measurement:

$$SNR = \frac{\mu}{\sigma} \tag{10}$$

where μ is the signal mean or expected value and σ is the standard deviation of the noise, or an estimate thereof. Notice that such an alternative definition is only useful for variables that are always positive (such as photon counts and luminance). Thus it is commonly used in image processing, where the SNR of an image is usually calculated as the ratio of the



Stochastic Resonance, Fig. 3 Residence-time distributions $N(T)$ for the symmetric bistable system: (**a**) increasing D (from below) with Ω held constant; inset: the strength P_1 of the first peak of $N(T)$ vs D; (**b**) increasing Ω (from below) with D held constant; inset: P_1 vs the forcing frequency

mean pixel value to the standard deviation of the pixel values over a given neighborhood. Sometimes SNR is defined as the square of the alternative definition above.

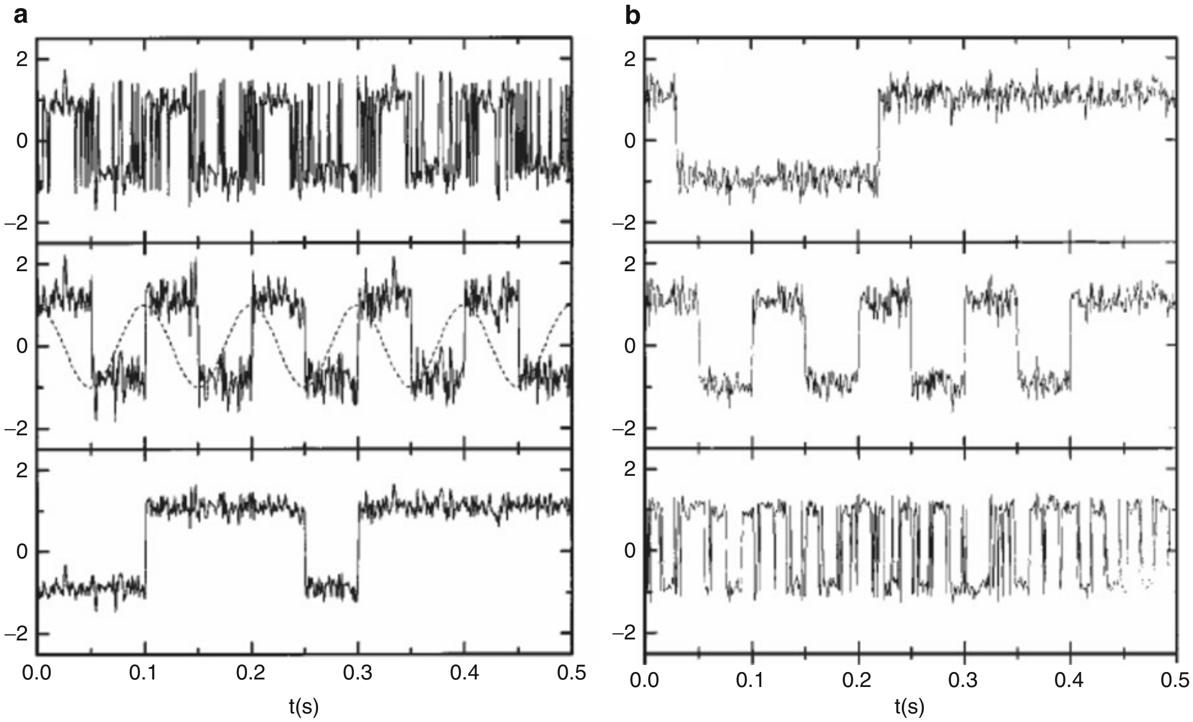
The *Rose criterion* (named after Albert Rose) states that an SNR of at least 5 is needed to be able to distinguish image features at 100% certainty. An SNR less than 5 means less than 100% certainty in identifying image details.

Yet another alternative, very specific and distinct definition of SNR is employed to characterize sensitivity of imaging systems; see signal-to-noise ratio (imaging).

Related measures are the "contrast ratio" and the "contrast-to-noise ratio."

Residence-Time Distribution

A deeper understanding of the mechanism of stochastic resonance in a bistable system can be gained by mapping the continuous stochastic process $x(t)$ (the system output signal) into a *stochastic point process* $\{t_i\}$. The symmetric signal $x(t)$ is converted into a point process by setting two crossing levels, for instance at $x_\pm = \pm c$ with $0 \leq c \leq x_m$. On sampling the signal $x(t)$ with an appropriate time base, the times t_i are determined as follows: Data acquisition is triggered at time $t_0 = 0$ when $x(t)$ crosses, say, x_- with negative time derivative $(x(0) = c, \dot{x}(0) < 0)$; t_1 is the subsequent time when $x(t)$ first crosses x_+ with positive derivative $(x(t_1) = c,\ \dot{x}(t_1) > 0)$; t_2 is the time when $x(t)$ switches back to negative values by



Stochastic Resonance, Fig. 4 Coherent switch in the symmetric bistable system of Eq. 3: (**a**) varying the noise intensity D with Ω held constant; (**b**) effect of varying Ω with D held constant

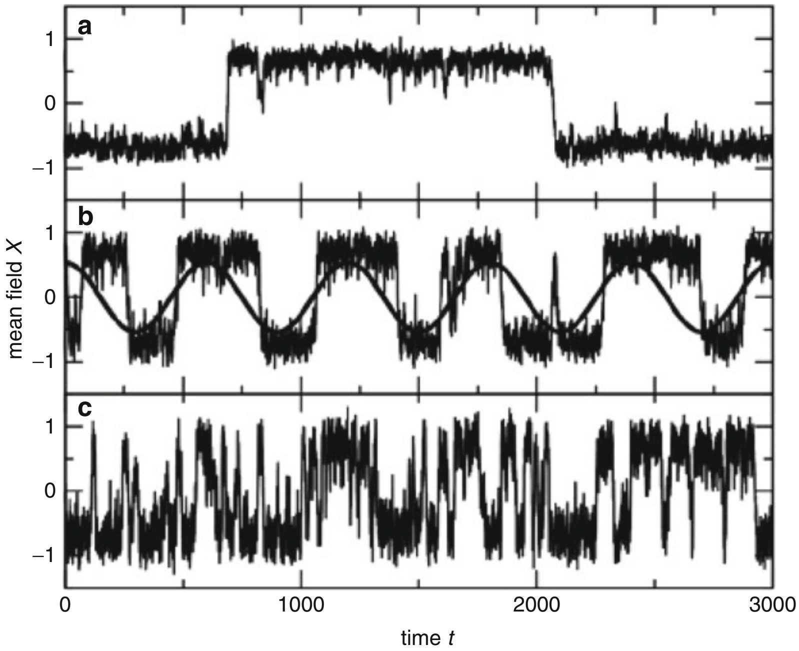
recrossing x_- with negative derivative, and so on. The quantities $T(i) = t_i - t_{i-1}$ represent the residence times between two subsequent switching events. For simplicity, we set $c = x_m$. The statistical properties of the stochastic point process $\{t_i\}$ are the subject of intricate theorems of probability theory. In particular, no systematic way is known to find the distribution of threshold crossing times. An exception is the symmetric bistable system: Here, the *long* intervals T of consecutive crossings obey Poissonian statistics with an exponential distribution:

$$N(T) = (1/T_K)\exp(-T/T_K) \qquad (11)$$

The distribution Eq. 11 is important for the forthcoming discussion, because it describes to a good approximation the first-passage time distribution between the potential minima in unmodulated bistable systems (Fig. 3).

Coherent Switch

Among the various patterns of regulation associated with nonlinear kinetics, bistability, a system-level property that even relatively simple signaling networks have the potential to produce, allows a graded signal to be turned into a discontinuous evolution of the system along several possible distinct signaling pathways which can be either reversible or irreversible. A system is termed bistable if it can switch between two distinct stable steady states but cannot rest in intermediate states under the excitation of external stimuli (e.g., noise). Biological examples of bistable systems include the λ phage lysis-lysogeny switch, several mitogen-activated protein kinase (MAPK) cascades in animal cells, and cell cycle regulatory CI circuits in *Xenopus* and *Saccharomyces cerevisiae*. Usually, bistable systems in the biological context are thought of as those involved in the generation of switch-like biochemical responses, the establishment of cell cycle oscillations and mutually exclusive cell cycle phases, the production of self-sustaining biochemical “memories” of transient stimuli, or the rapid lateral propagation of receptor tyrosine kinase activation. In spite of their simple dynamic behaviors, bistable systems are building blocks of larger



Stochastic Resonance, Fig. 5 Time evolution of the mean field of model Eq. 12 for three different system sizes

regulatory elements: genetic networks and signaling cascades.

Usually, coherent switch means that for a bistable system, some stochastic fluctuations (or noise) induce switching between two stable states. In order to understand the basic mechanism of stochastic switch, we consider a double-well potential system in the presence of noise and periodic forcing (see Eq. 3). Numerical simulations show that the appropriate noise strength can induce optimal switches between two steady states (refer Fig. 4).

Except for noise-induced stochastic resonance, coupling can noticeably enhance the stochastic resonance effect. In fact, the influence of spatial coupling in the SR scenario is revealed in the effect of the system size. For example, consider the following system consisting of globally coupling arrays of bistable elements:

$$\frac{dx_i}{dt} = \dot{x}_i = x_i - x_i^3 + \frac{\varepsilon}{N}\sum_{j=1}^{N}(x_j - x_i) + A\cos\omega t + \xi_i(t) \qquad (12)$$

where the noise $\xi_i(t)$ is assumed as Gaussian white noise with correlation given by $\langle \xi_i(t)\xi_j(t')\rangle = 2\sigma^2\delta_{ij}\delta(t-t')$, ε represents coupling strength, and N denotes the system size. Figure 5 shows the system size–induced stochastic resonance.

References

Gammaitoni L, Hänggi P, Jung P, Marchesoni F (1998) Stochastic resonance. Rev Mod Phys 70(1):223–287

Sagués F, Sancho JM, García-Ojalvo J (2007) Spatiotemporal order out of noise. Rev Mod Phys 79(3):829–882

Stochastic Simulation Algorithm

Alida Palmisano[1] and Corrado Priami[2]
[1]Department of Biological Sciences and Department of Computer Science, Department of Biological Sciences Virginia Tech, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[2]Microsoft Research-University of Trento Centre for Computational and Systems Biology and DISI, University of Trento, Povo, Trento, Italy

Synonyms

Doob–Gillespie algorithm; Gillespie algorithm; SSA

Definition

The term "simulation" is generally used to indicate the calculation of the system's dynamics over time, given an initial specific system configuration; for biological systems the initial configuration corresponds usually to the initial concentration of molecules. Biological systems can be simulated in different ways using different algorithms depending on the assumptions made about the underlying kinetics. Once the kinetics have been specified, these systems can be used directly to construct full dynamic simulations of the system behavior on a computer.

Chemical stochastic systems are usually represented by a chemical master equation (CME) that describes the time evolution of the probability distribution of the discrete molecule quantities (expressed by natural numbers). This evolution is a continuous time Markov chain (CTMC), of which any possible realizations can be generated through the Monte Carlo sampling methods. The most famous of these methods for coupled chemical reactions is the SSA algorithm of Gillespie (Gillespie 1976, 1977).

Gillespie designed an efficient way to simulate a trajectory of a set of coupled chemical reactions. The algorithm he proposed is exact with respect to the underlying principles behind the CME; it simulates a jump Markov process and is based on the assumption that two events take place at the same time with zero probability.

Gillespie proposes two mathematically equivalent methods: the direct method (DM) and the first reaction method (FRM). The algorithm is computationally expensive, so many modifications and adaptations exist: the efficient next reaction method by Gibson and Bruck (2000) that achieves a significant reduction in complexity with respect to the Gillespie algorithms, tau-leaping, and hybrid techniques where reactants in abundance are modeled with deterministic behavior. The price to be paid when those more efficient techniques are used is that the exactitude of the theory behind the algorithm as it connects to the master equation is generally compromised, but they offer reasonable realizations for greatly improved timescales.

A summarized description of the steps of FRM algorithm follows (for the other methods, see the papers cited in the references).

General idea: at each step a random putative reaction time is calculated for each reaction and the one with the shortest time is chosen and executed.

1. Initialize the number of molecules for each species and the initial time $= 0$.
2. Calculate the propensity value a_i for each $i \in \{1, \ldots, m\}$ (where m is the total number of reactions in the system). Propensities represent the probability that a reaction R_j occurs in the next infinitesimal time interval and depend on the amount of reactants that are present in the system in that time instant).
3. For each $i \in \{1, \ldots, m\}$ generate a putative time τ_i in accordance with an exponential distribution of parameter a_i.
4. Let τ_μ and μ be the fastest time and the corresponding reaction.
5. Update the number of molecules to reflect the execution of μ.
6. Set $t = t + \tau_\mu$.
7. Go back to Step 2 unless the number of all reactants is zero or the simulation time has been exceeded.

This method is used heavily in computational systems biology at the basis of almost all the stochastic simulators implemented for modeling biological systems with different computational languages. See ▶ Cell Cycle Modeling, Process Algebra for an example of a biological relevant case study, modeled with different process algebra approaches and simulated with SSA.

Cross-References

▶ Cell Cycle Modeling, Process Algebra

References

Gibson MA, Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. J Phys Chem 104:1876–1889

Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Phys Chem 22:403–434

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81(25):2340–2361

Stochastic Simulation Algorithms

▶ Mass Action Stochastic Kinetics

Stochastic Simulation Algorithms (SSAs)

► Stochastic Simulation Methods

Stochastic Simulation Methods

Ruiqi Wang
Institute of System Biology, Shanghai University,
Shanghai, China

Synonyms

Stochastic simulation algorithms (SSAs)

Definition

Gene expression is a stochastic process. The noise may come in two ways: intrinsic and extrinsic. Such noises are believed to play especially important roles when species are present at low-copy numbers. The stochastic modeling framework grasps the essence of the stochastic collision of biochemical components. However, most stochastic models are not analytically or numerically solvable in any but the simplest cases. Therefore, one has to resort Monte Carlo type simulations. Stochastic simulation methods have become an invaluable tool to study temporal dynamics of biomolecular systems. In contrast to deterministic approach based on ordinary differential equations, they can capture effects that occur due to the underlying discreteness of the systems and random fluctuations in molecular numbers. Various stochastic, approximate stochastic, and hybrid simulation methods have been proposed.

Monte Carlo simulation produces a random walk through the possible states of the system. In other words, instead of calculating the probability distribution, the approach simulates the time evolution of a particular trajectory, starting at a given initial state. Stochastic simulation algorithms (SSAs) can roughly be divided into exact, approximate, or hybrid strategies, depending on whether or not they introduce approximations or combine different approaches into one calculation scheme.

Generally, the SSA first constructs numerical realizations and then averages the results of many realizations. When performing simulations, the next reaction and the time of its occurrence need to be determined, e.g., through the Gillespie's direct method. The goal of stochastic simulation is then to describe the evolution of the state from some given initial state.

The most widely used SSA was developed by Gillespie in 1976. The Gillespie's SSA describes the state with discrete number of chemical molecules involved, and models its time evolution as a jump Markov chain with discrete steps. The Gillespie's algorithm numerically simulates individual occurrences of reactions. Some simulation trajectories are required for accurately capturing the probabilistic nature of the transient behavior of a system.

Exact stochastic methods include direct method and the first reaction method proposed by Gillespie and next reaction method proposed by Gibson and Bruck. Exact stochastic methods explicitly simulate each reaction event in the system, thus having time complexity approximately proportional to the overall number of particles present in the system. Therefore, they are slow for large systems. These exact methods are mathematically equivalent but differ in how they calculate the so-called reaction probability density function.

The major drawback of exact stochastic methods is computational cost because the SSA simulates each individual reaction event. Therefore, various approximate simulation methods that sacrifice an acceptable amount of accuracy in order to speed up the simulation have been developed. The proposed methods often involve a grouping of reaction events, i.e., they permit more than one reaction events per step, e.g., τ-leap method and Langevin method.

In 2001, Gillespie developed an approximate stochastic simulation method named τ-leap method to accelerate the stochastic simulation procedure. This method allows a sensible trade-off between accuracy and speed. It avoids simulation of every individual reaction event. Instead, it leaps in steps of length τ containing many single reaction events. Each timestep τ has to fulfill the so-called leap condition: It must be small enough so that no significant change in the propensities occurs during $[t, t + \tau]$. Since many single reaction events can be leaped over when τ is large enough, the simulation can be much faster. How to choose an appropriate τ depends on the trade-off between accuracy of the simulation and computation time. Besides the method proposed by Gillespie, a number of variants and extensions of the τ-leap method have been developed. For example, unbiased τ-leap method has been proposed to correct the bias in τ-leaping.

As a reaction occurs more frequently, it becomes more continuous, with the fluctuations in the rates of the reaction becoming Gaussian, via the central limit theorem. One may then approximate the fast reactions using the chemical Langevin equation (CLE), which is equivalent to an ordinary differential equation (ODE) with an additive Gaussian term whose variance is calculated from the reaction kinetics. In other words, when a system possesses a macroscopically infinitesimal time scale in the sense that during any time increment dt on that scale, all the reaction channels fire much more than once and none of the propensity functions change appreciably, its dynamics can be well approximated by Langevin equations. In particular, when the number of each species is large, Langevin equations can well describe the dynamics of cellular systems. The CLEs for all species form a system of stochastic differential equations (SDEs). One can then use continuous-stochastic techniques to simulate cellular dynamics.

Hybrid methods aim to combine different approaches into one calculation scheme. The essential idea is to partition reactions or model species into two or more groups: e.g., a group of low-copy number species and a group of high-copy number species, and then treat them in different ways. For further details, see the entry on hybrid simulation strategies.

Cross-References

- ▶ Hybrid Simulation Strategies
- ▶ Langevin Equation
- ▶ Law of Mass Action
- ▶ Master Equation
- ▶ Stochastic Simulation Algorithms

References

Chen L, Wang R, Li C, Aihara K (2010) Modeling biomolecular networks in cells: structures and dynamics. Springer, London

Gibson MA, Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. J Phys Chem A 104:1876–89

Gillespie DT (1976) General method for numerically simulating stochastic time evolution of coupled chemical-reactions. J Comput Phys 22:403–434

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81:2340–2361

Gillespie DT (2000) The chemical Langevin equation. J Chem Phys 113:297–305

Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. J Chem Phys 115:1716–33

van Kampen NG (1981) Stochastic processes in physics and chemistry. Elsiever, Amsterdam

Wilkinson DJ (2006) Stochastic modelling for systems biology. Chapman & Hall/CRC, Boca Raton

Xu Z, Cai X (2008) Unbiased τ-leap methods for stochastic simulation of chemically reacting systems. J Chem Phys 128:154112

Stochastic Switch

Tianshou Zhou
School of Mathematics and Computational Sciences,
Sun Yet-Sen University, Guangzhou, Guangdong,
China

Definition

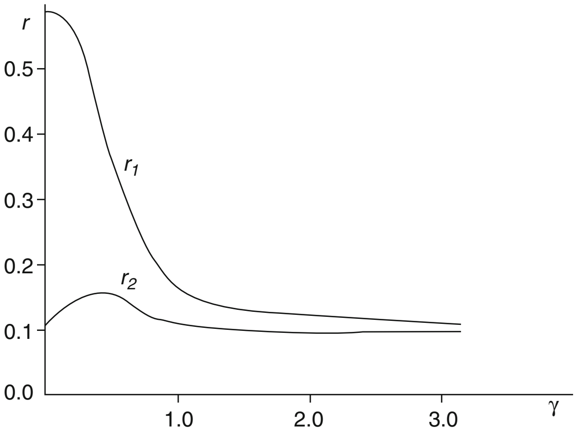
Stochastic switch means that random forces induce switching between two stable states of a dynamical system. The common stochastic switching is the one occurring in a bistable system subjected to an external stochastic force.

Stochastic Synchronization

Tianshou Zhou
School of Mathematics and Computational Sciences,
Sun Yet-Sen University, Guangzhou, Guangdong,
China

Definition

Stochastic synchronization is a type of synchronization in the sense of statistics. The fundamental phenomenon of synchronization occurs in nonlinear self-sustained oscillators subjected to a periodic force or coupled with each other. This phenomenon manifests itself in locking



Stochastic Synchronization, Fig. 1 Dependencies of the MSF in the subsystems on the coupling strength for the parameter values $\alpha = 0.5$, $\beta = 1.0$, and $D = 0.1$.

or suppressing of the natural frequency of the oscillator by periodic force. Synchronization-like phenomena can occur in stochastic bistable systems which have no natural frequency at all. Stochastic bistable system possesses a noise-controlled mean switching frequency (MSF) between metastable states being an analogy of natural frequency. The *stochastic* synchronization reveals locking of the mean switching frequency by external periodic force. The same phenomenon can be observed in coupled stochastic bistable systems.

For coupled stochastic bistable systems, the stochastic differential equations have the form

$$\frac{dx}{dt} = \alpha x - x^3 + \gamma(y - x) + \sqrt{2D}\xi_1(t)$$
$$\frac{dy}{dt} = \beta - y^3 + \gamma(x - y) + \sqrt{2D}\xi_2(t)$$

where α and β are parameters characterizing the barrier heights in the subsystems, $\xi_{1,2}(t)$ are statistically independent white Gaussian noises. The parameter γ refers to the strength of coupling in the system. In the decoupled case ($\gamma = 0$) the stochastic processes in the subsystems are statistically independent with different mean switching frequencies (MSFs), determined by the parameters α and β. However, with the increase of the coupling strength γ some kind of coherence can be observed and the MSFs in the subsystems tend to coincide. In the following figure, the dependence of the MSF in the subsystems versus γ is shown for the fixed values of D, α, and β. As it is seen from the figure, the MSFs in the subsystems draw closer to one another when the strength of coupling is increased. Such behavior of partial frequencies is indeed typical for the phenomenon of synchronization of coupled classical self-sustained oscillators (Fig. 1).

Stochastic Variable

▶ Random Variable

Stochastic π-Calculus

▶ Stochastic pi-Calculus

Stock Center

Sabina Leonelli
ESRC Centre for Genomics in Society, University of Exeter, Exeter, Devon, UK

Definition

A stock center is an institution responsible for the collection, storage, and distribution of several strains of a specific group of organisms. Notable examples are seed banks and stock centres for model organisms (Rosenthal et al 2002).

Cross-References

▶ Model Organism

References

Rosenthal N, Ashburner M (2002) Taking stock of our models: the function and future of stock centres. Nat Rev Genet 3:711–717

Stoichiometric Mass Balance Analysis

▶ Conservation Analysis

Stoichiometric Matrix

Osbaldo Resendis-Antonio
Center for Genomic Sciences-UNAM, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

Definition

The stoichiometric matrix (***S***) is a matrix that contains information about all the metabolic transformations included in a metabolic reconstruction for a microorganisms. In practice, it is formed from the stoichiometric coefficients of each reaction that comprise the metabolic reconstruction, which commonly are integer numbers. These numbers are organized in such a way that the columns identify chemical reactions and the rows correspond to metabolic compounds. For instances, if we have a biological system conformed by metabolites *A, B* and *C*, which can transform among them following this set of charge and mass balance metabolic reactions:

$$A + 2B \rightarrow C$$
$$C \rightarrow A$$
$$A \rightarrow B$$

The stoichiometric matrix is written as:

$$S = \begin{bmatrix} -1 & 1 & -1 \\ -2 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix}$$

It is important to stress that the stoichiometric matrix is an essential component for proceeding to a variety of computational analysis in genome scale metabolic reconstructions. Among these computational analyses, mainly the stoichiometric matrix plays an important role in calculating the flux production and degradation for each one of the metabolites included in a metabolic reconstruction. This balance is important as it determines the dynamic profile of metabolic concentrations by solving the equation:

$$\frac{d\mathbf{x}}{dt} = S \cdot \mathbf{v}$$

Here $\mathbf{x} = (\mathrm{x}_1, \ldots \mathrm{x}_\mathrm{m})^\mathbf{T}$ and $\mathbf{v} = (\mathrm{v}_1, \ldots, \mathrm{v}_\mathrm{n})^\mathbf{T}$ are column vectors representing all the metabolic concentrations and all the metabolic fluxes included in the reconstruction respectively. Notably, in this contextual scheme, the analysis of the column, row, null, and left-null space of ***S*** allow us to explore the dynamical behavior and the feasible steady-state phenotype of a metabolic network (Palsson 2006).

Cross-References

- ▶ Constraint-based Modeling
- ▶ Dynamic Metabolic Networks, k-Cone
- ▶ k-Cone Space

References

Palsson BO (2006) Systems biology: properties of reconstructed networks. Cambridge University Press, Cambridge

Storey Tibshirani Method

Winston Haynes
Seattle Children's Research Institute, Seattle, WA, USA

Definition

The Storey Tibshirani method calculates a specific False Discovery Rate (FDR) for each feature in a multiple hypothesis test. Specifically motivated by microarrays, the Storey Tibshirani method was developed for experiments with a large number of features (individual genes) being tested on the same statistical hypotheses (significant differential expression).

The Storey Tibshirani method emphasizes an important difference between FDR and the false positive rate: whereas the false positive rate is the rate that null features are identified as significant, the FDR is the rate that features identified as significant are null. The *p*-value and *q*-value are feature specific values for the false positive rate and FDR, respectively.

Given a list of *p*-values, the Storey Tibshirani method determines *q*-values for each feature.

The method assumes an accurate calculation of these *p*-values. As a foundation, the Storey Tibshirani method establishes a general definition that the FDR is the number of false positive features divided by the total number of significant features.

The Storey Tibshirani method approximates the proportion of features which are truly null, π_0. The calculation of π_0 is derived from a parameter estimation that relies on the assumption that null *p*-values will be uniformly distributed. The proportion of null features is calculated from the density of values in the uniformly distributed section of *p*-values.

The number of false positives is approximated as the proportion of features which are truly null, π_0, times the total number of features, *m*, times the *p*-value, *t*. The total number of significant features is the number of *p*-values in the list which are less than or equal to *t*. So, the formula for the *q*-value is:

$$\frac{\pi_0 \cdot m \cdot t}{count(p_i \leq t)}$$

References

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. PNAS 100:9440–9445

STRENDA

Carsten Kettner
Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Frankfurt am Main, Germany

Synonyms

Standards for reporting enzymology data

Definition

STRENDA is both an initiative inaugurated and funded by the Beilstein-Institut since 2004 and a working group (Commission) that aims at the improvement of the quality of reporting functional enzymology data (Kettner and Hicks 2005). The STRENDA initiative and working group aim to establish standards for reporting enzyme data to allow a full understanding of the conditions under which they were obtained. The hope is that such standards will become required by the major scientific journals and that they will be fully documented in those databases, such as BRENDA and SABIO-RK related to organisms and enzyme groups that compile enzyme activity and kinetic data (▶ Data Integration). The way toward these standards is being paved by the compilation of guidelines which are published on the STRENDA project website (http://www.beilstein-institut.de/en/projects/strenda/guidelines/) and which are already recommended by 28 major biochemistry journals. These guidelines are divided in two parts: part one is the Level 1 A checklist that samples the information that is required for a complete description of an experiment (▶ Experimental Design, Variability). This information allows a quality check on the data and ensures their value to others. The second part is called Level 1 B checklist that is concerned with the description of the enzyme activity data (▶ Biological Activity). In principle, this is the minimum information to describe the experimental results (Apweiler et al. 2005).

Characteristics

Background

The modern experimental techniques facilitate the generation of huge amounts of protein structure and enzyme activity data. Technical advances increase the accuracy of both the recording and analysis of data which leads, consequently, to large data sets which are published in the scientific journals and collected in databases such as BRENDA, SwissProt, PDB, and other electronic repositories (▶ Data Integration). Since experiments (▶ Experiment) on enzyme characterizations are carried out under individually defined and laboratory-specific conditions and experimental designs (▶ Experimental Design, Variability) depend on the given experimental know-how, methods, and technical equipment available, raw data for the same enzyme from different labs are normally not comparable (Kettner and Hicks 2005). Consequently, functional enzyme databases do not provide definite values of pH and temperature optima, transition rates of reaction kinetics, K_m, K_D, and K_i of molecules that act as activating and inhibiting substances, and instead only numbered data within relatively wide ranges can be found. For example, K_m values from the literature (as stored, e.g., in BRENDA) have been measured at pH values between 3 and >10 and at temperatures between 0 °C and more than 100 °C. However, these ranges of values are neither suitable for making direct comparisons between two enzymes nor is it possible to model even sections of physiological pathways due to the inaccuracy and the broad statistical mean variations. Therefore, the experimental conditions need to be clearly and fully stated to avoid misinterpretations of laboratory findings when data move between researchers whose laboratories employ individual methods. Additionally, a clear statement on the materials used and methods applied is essential for the successful integration of experimental and theoretical biology which include in silico analysis and representations of metabolic systems (▶ Kinetic Modeling and Simulation; ▶ Synthetic Models and Methods) (Stelling et al. 2002; Klipp et al. 2007; LeNovere et al. 2007).

Aims

STRENDA proposes uniform assay standards (▶ Biological Assay) of data for single enzymes and groups of enzymes. The Commission is aware that the conditions under which an enzyme operates will depend on the organism and organelle in which it occurs. To take an extreme example, the physiological temperature at which an enzyme operates in a mammal may have little relevance to the behavior of the corresponding enzyme in a hyperthermophile. Additionally, the use of very different assay conditions for assaying the forward and reverse reactions catalyzed by the same enzyme may mean that valuable thermodynamic data are lost.

STRENDA develops an electronic data-submission form (STRENDA E-Form) that incorporates the STRENDA guidelines for reporting enzymology data. This E-Form is a functional data acquisition system that is intended to serve as a portal (i) to support both authors and journals as an assessment tool on the compliance with the STRENDA guidelines with an emphasis on providing information comprehesively rather than defining acceptance criteria, (ii) to store entered functional data along with the experimental conditions in a data base that will be publicably accessible. However, neither this form nor the guidelines are intended to create a substitute for the review process. The prototype and later the productive data acquisition system can be accessed at http://www.beilstein-institut.de/en/projects/strenda/e-form/. The STRENDA Commission hopes to work with the community to develop an E-form that can be integrated into the publication practices of the community (Apweiler et al. 2010) (▶ Data Integration).

The STRENDA Commission is aware of the fact that any recommendation on the standardization of experimental conditions requires broad discussions within the scientific community to gain acceptance of the guidelines. Therefore, the Beilstein-Institut organizes a symposium called “Experimental Standard Conditions of Enzyme Characterizations” every 2 years to provide a platform for the exchange of ideas and link to the scientific community. Experts from all fields of experimental, theoretical, and bioinformatics enzymology and metabolic network investigation present and discuss new results, approaches, and methodologies as well as pitfalls and problems of data generation and reproduction. Suggestions from the STRENDA Commission form the basis of subsequent discussions in following symposia where they were are improved, rejected, or replaced by alternatives.

The Commission is open for the cooperations with other standardization initiatives in pertinent subjects.

The STRENDA Guidelines

After extensive discussions with the scientific community on recent ▶ Experimental Standard Conditions of Enzyme Characterizations (ESCEC) Symposia, STRENDA proposes the STRENDA guidelines for supporting authors, referees, and editors to improve the quality of scientific data publication.

These guidelines (version 1.6) are divided in two checklists A and B. List A guides through the determination of those data required for materials and methods sections of publications and includes the description of the identity of enzyme, the assay conditions (▶ Biological Assay), and methodologies, preparation of the enzyme(s), and additional details. List B supports the description of the experimental results comprising the determination of reaction rates and the dimensions of kinetic parameters, the proper use of units and correct terminology, as well as the identification of inhibition and activation parameters.

The STRENDA guidelines have been approved by NC-IUBMB in 2005. Additionally, the guidelines are recommended to be considered by authors when reporting kinetic data by 28 biochemistry journals amongst them are:

- ACS Biochemistry
- The Journal of Biological Chemistry
- Archives in Biochemistry and Biophysics
- Biochemical and Biophysical Research Communications
- BBA (all nine sections)
- FEBS Journal
- Nature Chemical Biology
- Proceedings of the National Academy of Sciences, USA

The recommendation to refer to the MIBBI portal (Taylor et al. 2008) for prescriptive checklists for reporting research data is made by the following publishers and journals:

Publishers:

- BioMedCentral (e.g., BMC Bioinformatics, BMC Biochemistry, BMC Biology, BMC Systems Biology, etc.)
- PLoS (e.g., PLoS One, PLoS Biology, PLoS Medicine, etc.)

Journals:

- OMICS: A Journal for Integrative Biology

The Commission is concerned with the development of an electronic submission tool (STRENDA E-Form) for enzymology data to support authors and journals to report this data in compliance with the STRENDA Guidelines since several years (Apweiler et al. 2005). After an evaluation process, STRENDA and the Beilstein-Institut decide on the final implementation of E-Form which could lead to development of a database for the deposition of protein-function data (Apweiler et al. 2010).

Cross-References

- ▶ Biological Activity
- ▶ Biological Assay
- ▶ Data Integration
- ▶ ESCEC
- ▶ Experiment
- ▶ Experimental Design, Variability
- ▶ Kinetic Modeling and Simulation
- ▶ Synthetic Models and Methods

References

Apweiler R, Cornish-Bowden A, Hofmeyr J-HS, Kettner C, Leyh TS, Schomburg D, Tipton K (2005) The importance of uniformity in reporting protein-function data. TiBS 30(1):11–12

Apweiler R, Armstrong R, Bairoch A, Cornish-Bowden A, Halling PJ, Hofmeyer J-HS, Kettner C, Leyh TS, Rohwer J, Schomburg D, Steinbeck C, Tipton K (2010) A large-scale protein-function database. Nat Chem Biol 6:785

Kettner C, Hicks MG (2005) The dilemma of modern functional enzymology. Curr Enzyme Inhib 1:171–181

Klipp E, Liebermeister R, Helbig A, Kowald A, Schaber J (2007) Systems biology standards – the community speaks. Nat Biotechnol 25:390–391

LeNovere N, Courtot M, Laibe C (2007) Adding semantics in kinetic models of biochemical pathways. In: Hicks MG, Kettner C (eds) *Proceedings of the 2nd international Beilstein symposium on ESCEC*, Logos-Verlag, Berlin, pp 137–153

Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED (2002) Metabolic network structure determines key aspects of functionality and regulation. Nature 420:190–193

Taylor CF et al (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26(8):889–896

Stroma

Mary Helen Barcellos-Hoff
Department of Radiation Oncology and Cell Biology,
New York University School of Medicine, New York,
NY, USA

Synonyms

Connective tissue; Microenvironment

Definition

Stroma is the non-parenchyma compartment in which vessels, nerves, and migratory immune and inflammatory cells reside. Stroma consists of mostly mesenchymal cells, such as fibroblasts, which produce the interstitial extracellular matrix (ECM), pericytes adjacent to blood vessels, and adipocytes.

Characteristics

All organs are composed of the parenchyma, i.e., the cells that perform the function of the organ, and the stroma, a supporting or connective tissue composed mainly of ▶ fibroblasts, and in some organs, adipocytes. The connective tissues essentially hold the cells of the body together. These tissues form a framework, or matrix, for the body. Systemic tissues that include the vasculature network, immune cells, and peripheral nerves reside within the stroma. The stroma provides support for the architecture of the parenchyma, e.g., epithelial lobules, ducts, and glands, by producing the interstitial ECM which has abundant elastin and collagen type I, one of the most abundant proteins in animals.

The adult stroma is derived from the fetal mesenchyme, which is both instructive and inductive for parenchyma. The early work of developmental biologists established that epithelia are "specified" via inductive interactions with stroma (Kratochwil 1969). During development, the epithelium and stroma acquire more differentiated phenotypes as a result of multiple and often reciprocal hormonally regulated interactions. Tissue recombination studies in vivo have been useful in elucidating these relationships. An example is the studies using the mouse mammary gland to demonstrate that interactions between the embryonic epithelial and dense fibroblast stroma determine the ability of the epithelia to interact with the fatty stroma (Sakakura et al. 1982). If the epithelium does not come into contact with mesenchymal cells in the postnatal period, ductal morphogenesis fails to occur. The tissue-specific pattern of ductal branching is also dictated by stromal signals. Adipose stroma is required since epithelial growth occurs only in an adipose stroma, while adipose itself is dependent upon the presence of epithelium for inducing the changes in glycogen metabolism that accompany pregnancy and lactation. Thus, both fibroblast and adipose stroma are intimately involved in the development and differentiation of the epithelium.

While all organs have stroma, not all stroma are the same. Some organs have dense connective tissue; a good example is bone and cartilage, in which the ▶ extracellular matrix is highly specialized for rigidity and strength. Loose connective tissue acts as a partition between tissue elements. Adipose tissue is made of mostly adipocytes, but adipocytes can be dispersed as they are in the bone marrow, which consists of a reticular stroma defined by a highly organized network of collagen fibrils.

The vascular and lymphatic endothelium is supported by specialized stromal cells called pericytes. Pericytes provide both structural scaffolding and communicate with endothelial cells by direct physical contact and paracrine signaling pathways. Gap junctions between the cytoplasm of pericytes and endothelial cells enable the exchange of ions and small molecules.

Stroma often contains myofibroblasts that exhibit smooth muscle phenotypic properties characterized by the expression of most of the smooth muscle markers with a significant degree of heterogeneity in smooth muscle protein expression. The phenotypic transition of fibroblasts to myofibroblasts is an example of the plasticity of the differentiated cell phenotype. Myofibroblasts are contractile and are induced during healing to facilitate closure of wounds and are also prevalent in fibrotic pathologies. A special myofibroblast is induced in cancer that exhibits a mixed phenotype designated as cancer-associated fibroblasts (CAF). These fibroblasts may be considered together as "activated," in contrast to the relatively quiescent state of most fibroblasts in normal tissue (Olumi et al. 1999).

Immune cells, including macrophages, dendritic cells, and mast cells, reside in the stromal tissue compartment and contribute to the composition of the microenvironment. The adult stroma is also subject to modification by inflammation, during which fibroblasts are responsive to cytokines and participate in a reciprocal action that is important for the resolution of acute inflammation. Chronic inflammation can induce a wound-like (▶ Wound Response) stromal response that in turn promotes persistent inflammation. Connective tissue disease is any disease in which connective tissue is a primary target of pathology. Many connective tissue diseases feature abnormal immune system activity with inflammation in tissues as a result of an immune system that is directed against one's own body tissues, often accompanied by degradation or loss of ECM.

Stroma is also an important source of growth factors, cytokines, and regulatory signals that include IGF-1, SDF-1, FGF-2, and TGF-β. Together, diverse stromal cells, soluble cytokines, and the insoluble ECM form a dynamic network of information that mediates tissue function. A reactive stroma, characterized by a shift in either resident or recruited cell types, participates in wound healing, inflammation, disease pathology, and cancer progression (Rowley 1998).

Cross-References

- ▶ Extracellular Matrix
- ▶ Fibroblasts
- ▶ Wound Response

References

Kratochwil K (1969) Organ specificity in mesenchymal induction demonstrated in the embryonic development of the mammary gland of the mouse. Dev Biol 20:46–71

Olumi AF, Grossfeld GD, Hayward SW, Carroll PR, Tlsty TD, Cunha GR (1999) Carcinoma-associated fibroblasts direct tumor progression of initiated human prostatic epithelium. Cancer Res 59(19):5002–5011

Rowley DR (1998) What might a stromal response mean to prostate cancer progression? Cancer Metastasis Rev 17(4):411–419

Sakakura T, Sakagami Y, Nishizuka Y (1982) Dual origin of mesenchymal tissues participating in mouse mammary gland embryogenesis. Dev Biol 91:202–207

Structural and Practical Identifiability Analysis

Andreas Raue[1] and Jens Timmer[1,2,3]
[1]Institute for Physics, University of Freiburg, Freiburg, Germany
[2]BIOSS Centre for Biological Signalling Studies and Freiburg Institute for Advanced Studies (FRIAS), Freiburg, Germany
[3]Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden

Introduction

In systems biology, mathematical models of the dynamics of cellular processes promise to yield new insights into the underlying biology and their systems' properties (Becker et al. 2010). Often, models contain parameters such as reaction rate constants, amount of molecular compounds, and detection sensitivities, or measurement backgrounds are involved that are unknown or known only with large uncertainty (Bachmann et al. 2011). Before a model can be used for prediction, the unknown parameters have to be estimated by comparing model output to experimental data. For a realistic assessment of the accuracy of model predictions, it is important that uncertainties in the experimental data and in prior model assumptions are propagated correctly via the estimated parameters to the model predictions. The processes are usually nonlinear, high-dimensional and time-resolved experimental data of the processes are sparse. Therefore, parameter estimation faces the challenges of structural and practical nonidentifiability of the parameters (Raue et al. 2009). ▶ Identifiability indicates whether a parameter can be inferred from the experimental data (Walter 1987). Nonidentifiability of the model parameters reduces the predictive power of a model. The results of an identifiability analysis can be used for designing new experiments that resolve nonidentifiabilities (Raue et al. 2010).

Definition

In the context of signaling networks (▶ Metabolic and Signaling Networks), ordinary differential equation (ODE) systems are frequently used to investigate the dynamic properties of pathway components and their transient modifications (Swameye et al. 2003). This assumes that diffusion is fast compared to reaction rates and cell volume. Intrinsic stochasticity can be neglected if the copy number of proteins is sufficiently large. The model equations

$$\dot{\mathrm{x}}(t,\theta) = \mathrm{f}(\mathrm{x}(t,\theta),\mathrm{u}(t),\theta) \tag{1}$$

$$\mathrm{y}(t_i,\theta) = \mathrm{g}(\mathrm{x}(t_i,\theta),\theta) + \varepsilon_i \tag{2}$$

describe via the ODE system (Eq. 1) the dynamics of n species x such as concentrations of proteins in different phosphorylation states (▶ Partial Differntial Equations, Numerical Methods and Simulations). Their dynamical behavior may depend on an input function u(t) such as an external treatment with ligands and model parameters $\theta = \{\theta_1 \ldots \theta_l\}$ such as rate constants or initial conditions of (Eq. 1). The species are mapped to m model observables y via an observation function g in (Eq. 2). The model observables are the quantities accessible by experiments measured at times t_i. They may depend on additional parameters such as scaling or offset parameters included in θ. Often, only a subset or combinations of the modeled species are accessible by experiments, meaning that $m < n$. The distribution of the measurement noise $\epsilon_{ki} \sim N\left(0, \sigma_{ki}^2\right)$ is assumed to be known.

Commonly, many model parameters θ are unknown and have to be estimated from experimental data (▶ Parameter Estimation; ▶ Optimization and Parameter Estimation, Genetic Algorithms; ▶ Grid Computing, Parameter Estimation for Ordinary Differential Equations). The agreement of experimental data $y_k^\dagger(t_i)$ with the observables predicted by the model $y_k(t_i, \theta)$ for parameters θ is measured by an objective function, commonly the weighted sum of squared residuals

$$\chi^2(\theta) = \sum_{k=1}^{m} \sum_{i=1}^{d_k} \frac{1}{\sigma_{ki}^2} \left(y_k^\dagger(t_i) - y_k(t_i, \theta)\right)^2 \qquad (3)$$

where d_k denotes the number of data points for each observable $k = 1 \ldots m$, measured at time points t_i with $i = 1 \ldots d_k$. σ_{ki} are the corresponding measurement errors that are assumed to be known. The parameters can be estimated by finding the parameter values $\hat{\theta}$ that minimize $\chi^2(\theta)$. For normally distributed measurement noise, this corresponds to ▶ maximum likelihood estimation (see, e.g., Seber and Wild (2003)). Therefore, $\chi^2(\theta)$ will be called likelihood in the following.

The key point is that it is not sufficient to rely on the mere estimated parameter values and the predictions corresponding to these values. It is important to consider the uncertainties in the parameter estimation procedure: from measurement uncertainties, to parameter uncertainties and possibly non-▶ identifiability, to uncertainties in the model predictions. Uncertainties in the parameter estimates are usually described by ▶ confidence intervals.

Characteristics

An approach for ▶ identifiability analysis utilizing the profile likelihood

$$\chi_{PL}^2(\theta_i) = \min_{\theta_{j \neq i}} \left[\chi^2(\theta)\right] \qquad (4)$$

was proposed by Raue et al. (2009). The idea of the approach is to detect flatness of the likelihood by exploring the parameter space for each parameter in the direction of least increase in $\chi^2(\theta)$. Therefore, for each parameter θ_i, individually a section along the minimum of the objective function with respect to all of the other parameters $\theta_{j \neq i}$ is computed. At the same time, the profile likelihood enables calculation of likelihood-based confidence intervals (Murphy and van der Vaart 2000).

Structural Nonidentifiability

A structural nonidentifiability arises from the model structure only and is independent of the amount and quality of experimental data (see Walter (1987)). Assuming ideal measurements, with arbitrarily many and perfectly chosen measurement time points t_i and absence of measurement errors $\varepsilon_i = 0$, the crucial question is whether the model parameters θ are uniquely estimable from the model observables y(t_i,θ).

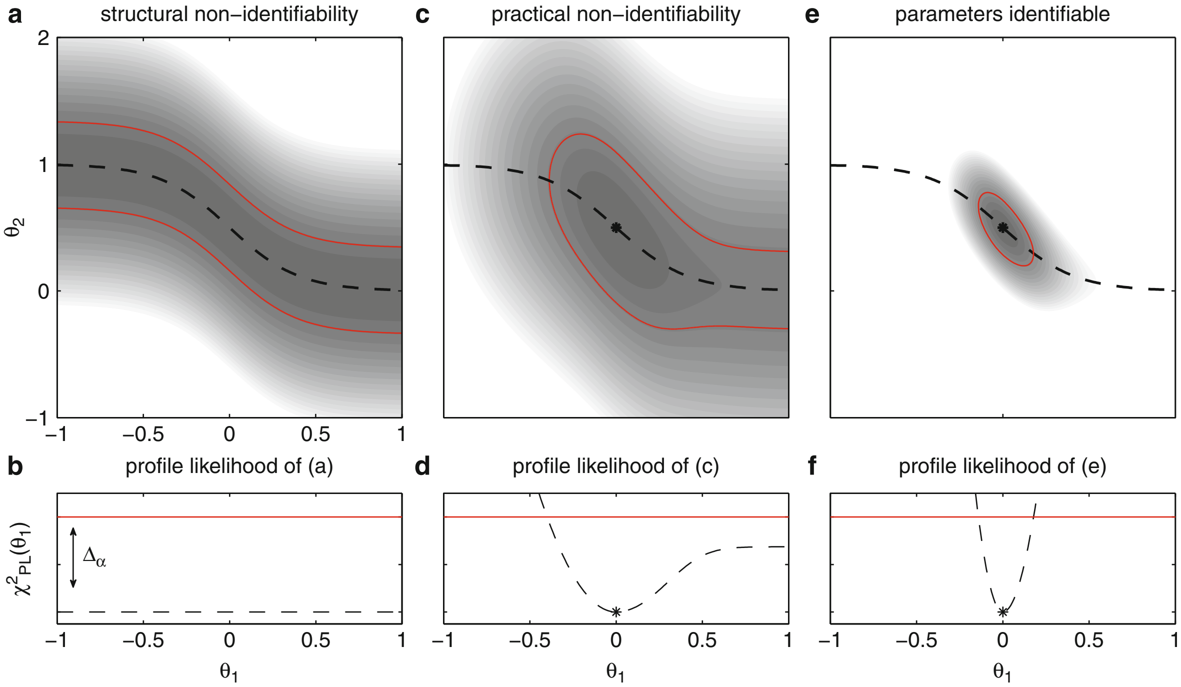
The analytical solution of y(t_i,θ) may contain an ambiguous parameterization with respect to θ, arising from an insufficient mapping function g in (Eq. 2) that is characterized by functional relations h(θ_{sub}) $= 0$ of a subset of parameters $\theta_{sub} \subset \theta$. In terms of likelihood, a structural nonidentifiability manifests as iso-χ^2 manifold

$$\{\theta | \mathrm{h}(\theta_{sub}) = 0\} \quad \Rightarrow \quad \chi^2(\theta) = const. \qquad (5)$$

For a two-dimensional parameter space, a structural nonidentifiability can be visualized by a perfectly flat valley that is infinitely extended along the corresponding functional relation, as illustrated in Fig. 1a by the dashed line. Correspondingly, this can be detected by a flat line of the profile likelihood for each parameter of θ_{sub} (see Fig. 1b). Consequently, structural nonidentifiable parameters are not uniquely identified by measurements of y(t_i,θ), and confidence intervals of $\theta_i \in \theta_{sub}$ are infinite. A parameter is structurally identifiable if a unique minimum of $\chi^2(\theta)$ with respect to θ_i exists (see Fig. 1c–f).

Practical Nonidentifiability

A parameter that is structurally identifiable may still be practically nonidentifiable. This can arise due to insufficient amount and quality of experimental data or inappropriately chosen measurement time points. It manifests in a confidence interval that is infinite,



Structural and Practical Identifiability Analysis, Fig. 1 Assessing identifiability of parameter θ_1 from the profile likelihood $\chi^2_{PL}(\theta_1)$. Contour lines in (**a**, **c**, **e**) shaded from *black* to *white* correspond to low, respectively, high values of $\chi^2(\theta)$. *Highlighted contour lines* indicate the threshold Δ_α utilized to asses likelihood-based confidence intervals, and *asterisk* corresponds to the optimal parameters $\hat{\theta}$. *Dashed lines* in (**b**, **d**, **f**) indicate the profile likelihood of θ_1 and its corresponding trace in (**a**, **c**, **e**)

although the likelihood has a unique minimum for this parameter. Confidence intervals can be defined by a threshold Δ_α in the likelihood. This threshold defines a confidence region

$$\left\{\theta|\chi^2(\theta)-\chi^2(\hat{\theta})<\Delta_\alpha\right\} \quad \text{with} \quad \Delta_\alpha = Q\left(\chi^2_{df}, 1-\alpha\right) \tag{6}$$

whose borders represent likelihood-based confidence intervals (Meeker and Escobar 1995). The threshold Δ_α is the 1–α quantile of the χ^2_{df}-distribution. The choice of *df* yields confidence intervals that hold jointly for *df* number of parameters (Press et al. 1990); often, $df = 1$ is desired.

A parameter is practically nonidentifiable if the likelihood-based confidence region (Eq. 6) is infinitely extended in the direction of θ_i indicated by the likelihood staying below a desired Δ_α (Raue et al. 2009). Similar to structural nonidentifiability, the flattening out of the likelihood can continue along a functional relation. For a two-dimensional parameter space, a practical nonidentifiability can be visualized as a relatively flat valley, which is infinitely extended, as seen in Fig. 1c. This can be detected by the corresponding profile likelihood in Fig. 1d, indicating that the height distance of the valley bottom to the lowest point at $\hat{\theta}$ never excesses Δ_α. By designing new experiments that increase the amount and quality of measured data and/or adjust the choice of measurement time points t_i, a practical nonidentifiability will ultimately be remediated, yielding finite confidence intervals (see Fig. 1e–f).

Experimental Design

Structural nonidentifiability is independent of the accuracy of experimental data. Therefore, it cannot be resolved by increasing the amount and quality of existing measurements. The only remedy is a qualitatively new measurement which alters the mapping function g in (Eq. 2), usually by increasing the number of observed species. For practical nonidentifiability, increasing the amount and quality of existing measurements may be sufficient but is often not very efficient.

To plan new experiments (▶ Optimal Experimental Design) that efficiently resolve nonidentifiability problems affecting θ_i, the set of trajectories along the profile likelihood $\chi^2_{PL}(\theta_i)$ should be investigated (Raue et al. 2009). Spots of large variability of the trajectories reveal where the uncertainty of θ_i has high impact. Additional measurements at these spots are promising candidates to resolve both structural and practical nonidentifiabilities and narrow confidence intervals efficiently. Furthermore, the amplitude of variability of the trajectories at these spots allows assessment of the necessary measurement precision to provide adequate data.

Cross-References

- ▶ Confidence Intervals
- ▶ Grid Computing, Parameter Estimation for Ordinary Differential Equations
- ▶ Identifiability
- ▶ Maximum Likelihood Estimation
- ▶ Metabolic and Signaling Networks
- ▶ Optimal Experimental Design
- ▶ Optimization and Parameter Estimation, Genetic Algorithms
- ▶ Parameter Estimation
- ▶ Partial Differntial Equations, Numerical Methods and Simulations

References

Bachmann J, Raue A, Schilling M, TimmerJ KU (2011) Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. Mol Syst Biol 7(516). doi:10.1038/msb.2011.50

Becker V, Schilling M, Bachmann J, Baumann U, Raue A, Maiwald T, Timmer J, Klingmueller U (2010) Covering a broad dynamic range: information processing at the erythropoietin receptor. Science 328(5984):1404–1408

Meeker WQ, Escobar LA (1995) Teaching about approximate confidence regions based on maximum likelihood estimation. The Am Stat 49(1):48–53

Murphy SA, van der Vaart AW (2000) On profile likelihood. J Am Stat Assoc 95(450):449–485

Press WH, Teukolsky SLA, Flannery BNP, Vetterling WMT (1990) Numerical recipes: Fortran. Cambridge University Press, Cambridge

Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, Timmer J (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics 25(15):1923–1929

Raue A, Becker V, Klingmüller U, Timmer J (2010) Identifiability and observability analysis for experimental design in non-linear dynamical models. Chaos 20(4):045105

Seber GAF, Wild CJ (2003) Nonlinear regression. Wiley, New York

Swameye I, Müller TG, Timmer J, Sandra O, Klingmüller U (2003) Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. Proc Natl Acad Sci 100(3):1028–1033

Walter E (1987) Identifiability of parametric models. Pergamon Press, New York

Structural Immunoinformatics

Shoba Ranganathan
Department of Chemistry and Biomolecular Sciences and ARC Center of Excellence in Bioinformatics, Macquarie University, Sydney, NSW, Australia
Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Synonyms

Immunoinformatics; Structure-based immunoinformatics

Definition

Structural immunoinformatics is the study of Immune system using computer-aided biotechnological (bioinformatics) tools and x-ray crystal structures of immune system components (Khan et al. 2009).

Cross-References

- ▶ Major Histocompatibility Complex (MHC)
- ▶ TR Recognition of MHC-Peptide Complexes

References

Khan JM, Tong JC, Ranganathan S (2009) Structural immunoinformatics: understanding MHC-peptide-TR binding. In: Davies MN, Ranganathan S, Flower DR (eds) Bioinformatics for immunomics, vol 3, Immunomics reviews series. Springer, New York, pp 77–94

Structural Motif

▶ Motif

Structure Type

▶ IMGT-ONTOLOGY, StructureType

Structure-based Immunoinformatics

▶ Structural Immunoinformatics

Structured Terminologies

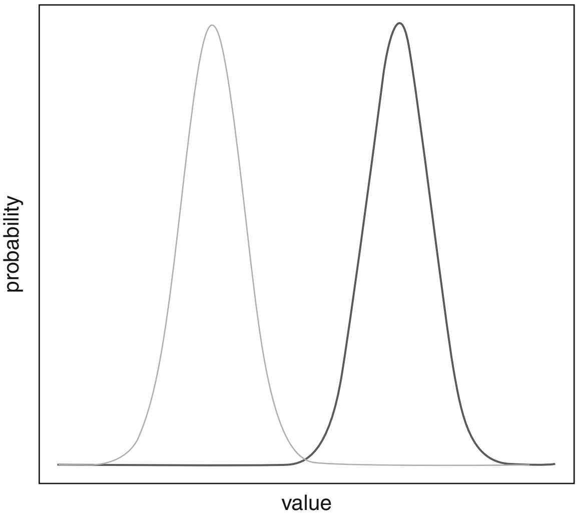
▶ Ontologies

Student's t-Test

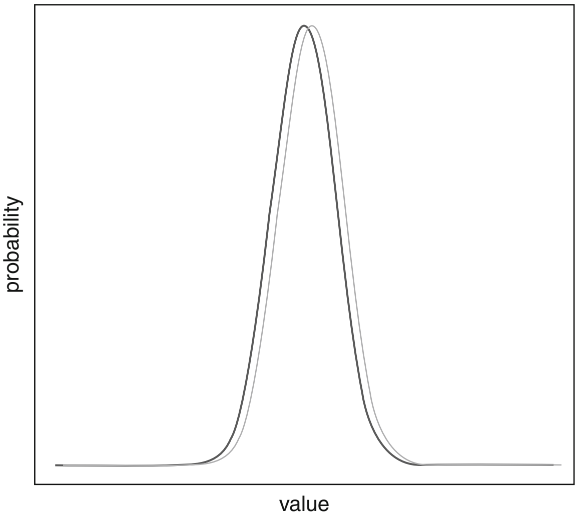
Winston Haynes
Seattle Children's Research Institute,
Seattle, WA, USA

Definition

The Student's *t*-test determines whether two populations express a significant or nonsignificant difference between population means. The Student's *t*-test places emphasis on controlling for sample size. A significant difference, seen in Fig. 1, is distinguished from a nonsignificant difference, seen in Fig. 2, by the properties of the normal distributions characterized by the data.



Student's t-Test, Fig. 1 *Significantly different populations*. The *dark-* and *light-gray lines* represent the distributions of two different populations. Since these distributions have minimal overlap, they appear to be significantly different. (Parameters: *dark-gray* mean = 3, standard deviation = 1; *light-gray* mean = −3, standard deviation = 1)



Student's t-Test, Fig. 2 *Nonsignificantly different populations*. The *dark-* and *light-gray* lines represent the distributions of two different populations. These distributions are nearly identical and, thus, are not significantly different. (Parameters: dark-gray mean = 0, standard deviation = 1; light-gray mean = 0.25, standard deviation = 1)

Properties

Null Hypothesis

The null hypothesis for the Student's *t*-test is that there is no difference in the means of two populations. Thus, rejection of the null hypothesis asserts a statistically significant difference between the population means.

Requirements

The Student's *t*-test distinguishes between exactly two population sets. Typically, one control condition is compared to a test condition. Measurements of the two populations must be in the same units.

Since *t*-test calculations rely on mean and standard deviation values from a normal distribution, the Student's *t*-test requires that both populations are reasonably approximated by a normal distribution. Further, the variance between the two populations must be roughly equal. If there exists unequal variance in populations, then the *t*-test for unequal variances, ► Wilcoxon Rank Sum Test, or Welch's *t*-test may be used [Ruxton 2006].

t Score Calculation

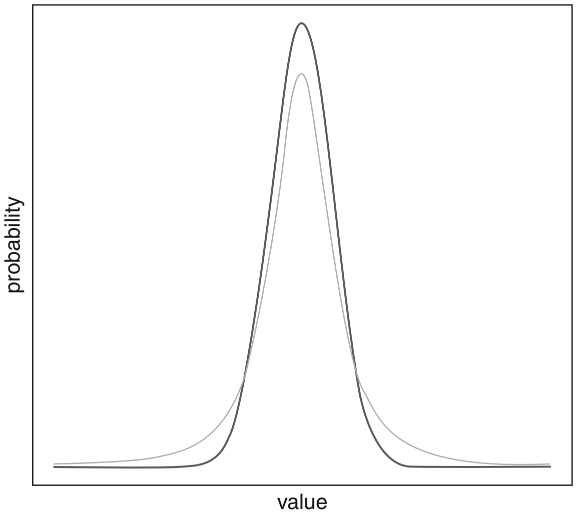
The *t* score represents the difference between sample means divided by the standard error.

$$t = \frac{\text{Mean difference}}{\text{Standard error}}$$

Standard error decreases as variance decreases and sample size increases. Accordingly, a lower standard error indicates more confidence in the answer. So, a high *t* score indicates that there is a significant difference in the means and a high confidence in the difference. To assess the significance of a *t* score, the score must be compared to the *t* distribution, discussed below.

t Distribution

Though similar to the bell shape of the normal distribution, the *t* distribution is characterized by a distinct ► Probability Distribution. The *t* distribution is parameterized by the degrees of freedom for the data. In Fig. 3, the normal- and *t* distributions are drawn in light- and dark-gray, respectively. The *t* distribution has a wider base than the normal distribution, indicating that a lower percentage of *t* scores lie near the mean than in a normal distribution.



Student's t-Test, Fig. 3 *Normal distribution vs. t distribution*. The normal- and *t* distributions are drawn in *light-* and *dark-gray*, respectively. (Parameters: normal distribution mean = 0, standard deviation = 1; *t* distribution degrees of freedom = 2)

The *t* distribution is dependent on the degrees of freedom. In the case of the Student's *t*-test, the degrees of freedom is the total sample size of both populations minus two. As the degrees of freedom increase, the *t* distribution increasingly favors the mean. Many textbooks and online sources contain tables where degrees of freedom and confidence intervals are used to look up threshold values for *t* scores. Once a threshold *t* score is determined, significance of the *t* score can be determined. If the *t* score is greater than the threshold, the difference between the populations is significant up to the selected confidence interval. Otherwise, there is no significant difference between the two populations.

Formalized *t* Score Calculation

Note: In practice, most statistical software and spreadsheet applications can be used to perform a student's *t*-test.

To calculate the *t* score for two means μ_1 and μ_2 with variances s_1 and s_2, respectively, and sample sizes of n_1 and n_2, respectively, the formula is represented as:

$$t = \frac{\mu_1 - \mu_2}{s_p^2 \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where the pooled variance, $s_p{}^2$ is calculated as:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Example

We performed a test to determine whether our new chemical significantly increased cell size. For testing purposes, we measured both a control population, C and a test population, T one hour after nontreatment and treatment, respectively. Our results for C and T are $\{35, 43, 40, 38, 36, 40\}$ and $\{52, 47, 39, 43, 41, 48\}$, respectively.

From this dataset, we used our spreadsheet to calculate the t score as 2.73. Since there are 12 data points, the distribution has 10 degrees of freedom. From the t distribution table, we found that a 95% confidence interval with 10 degrees of freedom has a t threshold of 1.812, so we can say with 95% confidence that the difference between control and test was significant. Our trials require even more precision, so we look at the 99% confidence interval and find a threshold of 2.764. Unfortunately, our t score is less than 2.764, so we cannot say with 99% confidence that the difference is significant.

References

Ruxton GD (2006) The unequal variance t-test is an underused alternative to Student’s *t*-test and the Mann–Whitney *U* test. Behavior Ecol 17(2):688–690

Sokal RR, Rohlf FJ (1995) Biometry: the principles and practice of statistics in biological research. W.H. Freeman, New York

Zar JH (1999) Biostatistical analysis. Prentice Hall, Upper Saddle River, NJ

Subgraph Patterns

▶ Canonical Network Motifs

Subset Surprisology

▶ Subset Surprisology and Toponomics

Subset Surprisology and Toponomics

Andreas Dress
University of Bielefeld, Bielefeld, Germany
Key Laboratory for Computational Biology (PICB),
Shanghai Institutes for Biological Sciences (SIBS),
Chinese Academy of Science and Max Planck Society
(CAS–MPG) Partner Institute, Shanghai, China
Science Center of “infinity3 GmbH”, Bielefeld,
Germany

Definition

In subset surprisology, one studies the stochastic behavior of the cardinality of the intersection of collections of subsets of a given finite set of cardinality n. More specifically, given three natural numbers n, a, k and a family of natural numbers a_0, a_1,. . ., a_k, one investigates the asymptotic behavior of the total number

$$A_{n|a_0,a_1,\ldots,a_k}(a) := \binom{n}{a} \sum_{a'=a}^{n} \binom{n-a}{a'-a} (-1)^{a'-a} \prod_{i=0}^{k} \binom{n-a'}{a_i-a'} \tag{1}$$

of families of subsets $A_0, A_1, \ldots, A_k$ of $\{1, \ldots, n\}$ for which $\#A_i = a_i$ holds for all $i = 0,\ldots, k$ whose intersection $\cap_{i=0,\ldots,k} A_i$ has cardinality a. This is applied in particular to given collections of real-valued maps $f_0, \ldots, f_k$ defined on the set $\{1, \ldots, n\}$ to search for *thresholds*, that is, real numbers $T_0, T_1, \ldots, T_k \in \mathbf{R}$, for which the intersection $\cap_{i=0,\ldots,k}\{j \in \{1, \ldots, n\} : f_i(j) \geq T_i\}$ is – relative to the size of the individual sets $A_i(T_i) := \{j \in \{1, \ldots, n\} : f_i(j) \geq T_i\}$ – surprisingly large (or small).

Characteristics

The Formal Setup

In (Dress et al. 2004), it was shown that, given n, k, and $a_0, a_1, \ldots, a_k$ as above, the associated probability distribution

$$p_{n|a_0,\ldots,a_k} = \left(p_{n|a_0,\ldots,a_k}(a)\right)_{a \in \mathbf{N}_0}$$

defined on the set $\mathbf{N}_0$ of nonnegative integers by

$$p_{n|a_0,\ldots,a_k}(a) := \frac{A_{n|a_0,\ldots,a_k}(a)}{\prod_{i=0}^{k}\binom{n}{a_i}} \quad (a \in N_0)$$

converge, with $n \to \infty$, toward the *Poisson distribution*

$$\mathrm{poiss}_\alpha(a) := \frac{\alpha^a}{a!}\exp(-\alpha)$$

for some fixed $\alpha \in \mathbf{R}_{>0}$, provided the numbers a_i $(i = 0, \ldots, k)$ are assumed to converge with n to infinity in such a way that the conditions

$$a_i \leq n \quad \text{and} \quad \lim_{n\to\infty}\frac{\prod_{i=0}^{k} a_i}{n^k} = \alpha$$

are satisfied (More specifically, the proof revealed that the alternating signs in the expressions (1) for $A_{n|a_0,\ldots,a_k}(a)$ resulting from the standard exclusion-inclusion principle correspond exactly to the alternating signs in the power series expression for $\exp(-\alpha) = \sum_{s=0}^{\infty}\frac{(-1)^s\alpha^s}{s!}$ when n turns to infinity.).

In consequence, given any $k + 1$ subsets $A_0, A_1,\ldots, A_k$ of $\{1,\ldots, n\}$ of cardinality $a_0, a_1, \ldots, a_k$, respectively, whose intersection has cardinality a, one may put $\alpha = \alpha(A_0, A_1, \ldots, A_k) := \frac{\prod_{i=0}^{k} a_i}{n^k}$ and then approximate, for example, the probability $Q_+(a_0, \ldots, a_k)$ of finding by chance a collection of subsets of $\{1, \ldots, n\}$ of cardinality $a_0, a_1,\ldots, a_k$ whose intersection has cardinality at least a by the sum $\mathrm{poiss}_\alpha(\geq a) := \sum_{s=a}^{\infty}\frac{\alpha^s}{s!}\exp(-\alpha)$.

And, given a collection of real-valued maps $f_0,\ldots,f_k$ defined on the set $\{1, \ldots, n\}$ as above, one may use these approximations to quickly identify thresholds $T_0, T_1,\ldots, T_k \in \mathbf{R}$, for which the intersection $A(T_0, T_1, \ldots, T_k) := \cap_{i=0,\ldots,k}\{j \in \{1, \ldots, n\} : f_i(j) \geq T_i\}$ is – relative to the size $a_i(T_i)$ of the individual sets $A_i(T_i) := \{j \in \{1, \ldots, n\} : f_i(j) \geq T_i\}$ – surprisingly large (or small) as, putting $\alpha = \alpha(T_0, T_1, \ldots, T_k) := \frac{\prod_{i=0}^{k} a_i(T_i)}{n^k}$ and $a = a(T_0, T_1, \ldots, T_k) := |A(T_0, T_1, \ldots, T_k)|$, one can easily search for local maxima (or minima) of the function $R_+(T_0, \ldots, T_k) := -\ln(\mathrm{poiss}_\alpha(\geq a))$.

Applications in Toponomics

The recently developed fluorescence-microscopy technique called *Multi-Epitope Ligand Cartography/TIS* (cf. *Toponomics*) allows medical researchers to measure – for any given tissue (or blood) sample, any position j within that sample, and any given family of protein(-epitope)s $P_0, P_1,\ldots, P_k$ – the (relative) abundance $f_i(j)$ of protein P_i $(i = 0,\ldots, k)$ at position j.

It has been established (see, for instance, Schubert 2002, 2003; Schubert et al. 2006, 2009) that

- biologically relevant interactions between the proteins in question will show up in *surprisingly* large (or small) numbers of "areas" (i.e., just single positions j or – just as well – appropriately chosen collections of nearby positions) within the sample for which the measurements $f_i(j)$ simultaneously exceed some appropriately chosen thresholds T_i,
- by considering "areas" rather than just single positions, *topologically* defined linkages between proteins can be detected directly at the cellular site of their biological (inter)action (see also ▶ *Topology and Toponomics*),
- from the detection of such linkages, it will be straightforward to identify those proteins and "pathways" that are linked to, for example, disease-specific or any other biologically important functional states of a cell.

Direct functional linkage analysis in cells thus complements standard "large-scale protein-expression profiling" techniques that are all based upon homogenization of cell tissue and subsequent extraction of proteins from the homogenate. Actually, it might even be superior to such techniques that are all based on the *destruction* of the specific modes of spatial protein arrangements and the resulting topologically determined functional protein hierarchies (cf. Schubert 2002, 2003; Schubert et al. 2006).

A Typical Result

As reported in (Dress et al. 2004), this strategy was applied to various examples involving 2–18 proteins, and it was observed for instance that, in one case considering just a pair of two proteins, a local maximum $R(T_0, T_1) = 36.7368$ for T_0, T_1 values with $a_0(T_0)/n = 0.0088$ and $a_1(T_1)/n = 0.0088$ was attained while, for T_0, T_1 values with $a_0(T_0)/n = 0.141$ and $a_1(T_1)/n = 0.106$, the value of $R(T_0, T_1)$ was inconspicuously low.

This has to be interpreted as follows: Looking at the threshold $T_0 := T_0(14, 1\%)$ for which protein P_0 shows an intensity above T_0 at exactly 14, 1% of all pixels and the threshold $T_1 := T_1(10, 6\%)$ for which protein P_1 shows an intensity above T_1 at exactly 10, 6% of all pixels, no conspicuous local interaction between P_0 and P_1 can be detected. In contrast, looking at the corresponding thresholds $T_0' := T_0(0, 88\%)$ and $T_1' := T_1(0, 88\%)$, a surprisingly high local interaction between P_0 and P_1 could be observed. That is, while at already reasonably high levels of abundance, P_0 and P_1 do not seem to influence each other, this is completely reversed at really high levels of abundance.

This underlines that direct functional linkage analysis based on toponomics data can detect rare, but perhaps crucial protein-interaction patterns that are not observable by standard proteomics techniques.

Related Matters

It is worth noting that related approaches based on Kullback-Leibler divergence (see, e.g., Wikipedia on Kullback-Leibler Divergence) have been worked out in (Barysenka et al. 2009, 2010, 2011), and that approaches based on copula theory (see Wikipedia on copula theory) are studied in (Barysenka in preparation).

And it is also worth noting that subset surprisology (as well as the methods described in (Barysenka et al. 2009, 2010, 2011) and (Barysenka in preparation)) can just as well be applied to other image stacks that arise, say, in *multispectral* or any other kind of *multichannel imaging* (cf., e.g., Wikipedia on multi spectral imaging).

Cross-References

- ▶ Fluorescence Microscopy
- ▶ Poisson Regression
- ▶ Proteomics
- ▶ Topology and Toponomics
- ▶ Toponomics

References

Barysenka A, Dress A, Schubert W (2009) An information-theoretical approach to medical image segmentation. In: ICISE'09 proceedings of the first IEEE international conference on information science and engineering. IEEE Computer Society, Washington, DC, pp 3592–3595

Barysenka A, Dress A, Schubert W (2011) A comparative method for analysing toponome image stacks. East Asian J Appl Math 1:35–48

Barysenka A, Dress A, Schubert W (2010) An information theoretic thresholding method for detecting protein colocalizations in stacks of fluorescence images. J Biotechnol 149:127–131

Barysenka A. Copula-based methods of exploring statistical dependence between fluorescent markers, Ph.D. thesis, in preparation

Dress A, Lokot T, Pustyl'nikov LD, Schubert W (2004) Poisson numbers and poisson distributions in subset surprisology. Ann Comb 8:473–485

Schubert W (2002) Polymyositis, topological proteomics technology and paradigm for cell invasion dynamics. J Theor Med 4:75–84

Schubert W (2003) Topological proteomics, toponomics, MELK-technology. Adv Biochem Eng Biotechnol 83:189–209, http://www.ncbi.nlm.nih.gov/pubmed/12934931

Schubert W, Bonnekoh B, Pommer A, Philipsen L, Böckelmann R, Malykh Y, Gollnick H, Friedenberger M, Bode M, Dress A (2006) Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. Nature Biotechnol 24:1270–1278

Schubert W, Gieseler A, Krusche A, Hillert R (2009) Toponome mapping in prostate cancer: detection of 2000 cell surface protein clusters in a single tissue section and cell type specific annotation by using a three symbol code. J Proteome Res 8:2696–2707

Wikipedia on Kullback-Leibler divergence. http://en.wikipedia.org/wiki/Kullback-Leibler_divergence

Wikipedia on copula theory. http://en.wikipedia.org/wiki/Copula_(statistics)

Wikipedia on multi-spectral imaging. http://en.wikipedia.org/wiki/Multi-spectral_image

Subunit Vaccine

Gajendra Raghava
Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, Chandigarh, India

Definition

This is the minimum component from the pathogen required to generate host immune response. This may be a purified protein. It does not include whole pathogen (dead or alive) in the vaccine formulation thus posses less risk of adverse reactions than whole organism–based vaccines.

Summation Theorem

Emma Saavedra and Rafael Moreno-Sánchez
Department of Biochemistry, National Institute for Cardiology “Ignacio Chávez”, Mexico City, Mexico

Definition

It is the sum of all the control coefficients for flux (summation theorem for flux control coefficients) and for metabolite concentrations (summation theorem for concentration control coefficients). In the case of flux control coefficients the sum has to add up a value of 1 whereas for that of concentration control coefficients it has to be zero.

Cross-References

▶ Metabolic Control Theory

Superorganism

▶ Metaorganism

Supervenience

Ulrich Krohs
Department of Philosophy, University of Münster, Münster, Germany

Definition

Supervenience is a relation between lower-level properties of a system and higher-level properties, where the further determine the latter.

Let A and B be classes of properties. B-properties supervene on A-properties if and only if there cannot be any change in B without change in A. In other words, no change in the class of supervenient (higher level) properties can occur without change in the class of subvenient (lower level) properties, while, on the other hand, not any change in the subvenient class is accompanied by a change in the supervenient class (McLaughlin and Bennett 2011). For example, according to neuronal determinism, there cannot be a change in mental properties without at least a minute change on the neuronal level. There are, nevertheless, many neuronal activities that do not change any mental state.

The supervenience relation is quiet about the foundation of this coupling, in particular about its metaphysical background. It is thus weaker than any concept of ▶ emergence. It does neither rule out nor require a reduction relation to hold between both classes of properties (▶ Reduction). Ontological and conceptual reduction in the strict sense, on the other hand, requires a supervenience relation to hold.

Cross-References

▶ Emergence

References

McLaughlin B, Bennett K (2011) Supervenience. In: Zalta EN (ed) The Stanford encyclopedia of philosophy. http://plato.stanford.edu/archives/win2011/entries/supervenience/. Accessed on 15 Sep, 2012

Surrogate Endpoint

Weida Tong and Donna L. Mendrick
Division of Systems Biology, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, USA

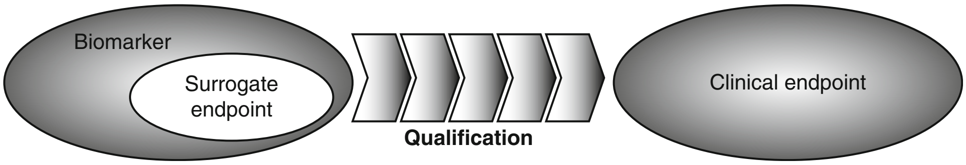
Synonyms

Surrogate markers

Definition

Surrogate endpoints (Fig. 1) are a subset of biomarkers, which are intended to be used as a substitute

Disclaimer: The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration.



Surrogate Endpoint, Fig. 1 Path diagram illustrating the biomarker-surrogate-clinical endpoint relationship and qualification

for a clinically meaningful endpoint. Characterization of a biomarker as a surrogate endpoint requires it to be "reasonably likely, based on epidemiologic, therapeutic, pathophysiologic or other evidence, to predict clinical benefit." (The Food and Drug Modernization Act of 1997). The term "validation" is discouraged for use in linking biomarkers as surrogate endpoints to clinical endpoints (Biomarkers Definitions Working Group 2001). This is largely due to the limitation for generalization in use of biomarkers as clinical endpoints in a clinical trial, which are often dependent on the specification of the therapeutic intervention, the characteristics of the population and disease state, and the statistics applied. Although the term "evaluation" is recommended for determining surrogate endpoint status (Biomarkers Definitions Working Group 2001), the term "qualification" is more commonly used to date to describe a process of linking a biomarker to a meaningful biological event.

Cross-References

▶ Biomarkers

References

Biomarkers Definitions Working Group (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 69:89–95

The Food and Drug Modernization Act of 1997. Title 21 Code of Federal Regulations Part 314 Subpart H Section 314.500

Surrogate Markers

▶ Surrogate Endpoint

Survival Analysis

Shuangge Ma
Yale School of Public Health, Yale University,
New Haven, CT, USA

Synonyms

Duration analysis; Lifetime data analysis

Definition

Survival analysis involves analyzing longitudinal data on the occurrence of events. In biomedical studies, events may include death, injury, onset of illness, recovery from illness, and transition above or below a clinical threshold of a meaningful continuous marker. The main objectives of survival analysis include (a) estimation of time to event for a group of individuals, for example, time to second heart attack for MI (myocardial infarction) patients; (b) comparison of time to event between two or more groups, for example, treated versus placebo MI patients in a randomized controlled clinical trial; and (c) assessment of the relationship between variables and time to event, for example, whether weight, insulin resistance, and cholesterol influence survival time of MI patients.

In survival analysis, commonly used measurements include (a) survival function, which is the probability that the time of event is later than a specified time; (b) hazard function, which is the event rate at a specified time conditional on survival until this point; (c) expected survival time, which is the expected value of the time remaining until event at a specified time, and others.

Survival analysis can be complicated due to the nonnegative nature of time to event data and, more importantly, censoring and truncation. Existing survival models include parametric (for example exponential model, Weibull model), semiparametric (e.g., Cox proportional hazards model, additive risk model, accelerated failure time model), and nonparametric models. Estimation and inference in survival analysis can be based on the maximum likelihood theories and (generalized) estimating equations. Other commonly used estimation and inference techniques are martingales and empirical processes techniques.

References

Collett D (2003) Modeling survival data in medical research, 2nd edn. Chapman & Hall/CRC, Boca Raton

Klein JP, Moeschberger ML (2010) Survival analysis: techniques for censored and truncated data. Springer, New York

Kosorok MR (2009) Introduction to empirical processes and semiparametric inference. Springer, New York

Survival Analysis, Fundamental Statistical Techniques

Shuangge Ma
Yale School of Public Health, Yale University,
New Haven, CT, USA

Synonyms

Duration analysis; Lifetime data analysis

Definition

Denote T as the event time, which is a nonnegative random variable having a certain probability distribution. The main objectives of survival analysis are to (a) estimate the time to event for a group of individuals; (b) compare the time to event between two or more groups; and (c) assess the relationship between covariates (risk factors) and time to event.

Characteristics

Main Measurements

For the random variable T, denote $f(T)$ as the probability density function. Other commonly used measures include

- Distribution function $F(t) = P(T \leq t)$ for $t \geq 0$; Survival function $S(t) = P(T > t) = 1 - F(t)$;
- Mean survival time $\mu = E(T)$; Median survival time $m = \max\{t : S(t) \leq 0.5\}$;
- Mean residual time at a specific time $t_0 : mrl(t_0) = E(T - t_0|T \geq t_0)$;
- Hazard function $\lambda(t) = \lim_{h \to 0} \frac{P(t \leq T < t+h|T \geq t)}{h}$, which is the instantaneous rate of failure at time t, given that an individual is alive at time t. Cumulative hazard function $\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t)$.

Commonly Adopted Statistical Models

Commonly adopted survival models, as in other regression analysis, can be classified as parametric, semiparametric, or nonparametric.

- Examples of commonly used parametric models include (a) exponential model, where $\lambda(t) = \lambda$, i.e., a constant hazard; (b) Weibull model, where $\lambda(t) = \alpha\lambda t^{\alpha-1}$. Depending on the value of parameter α, the hazard function may increase or decrease over time; and (c) Gamma model, where $S(t) = 1 - I(\lambda t, \beta)$ and $I(t, \beta) = \int_0^t \frac{u^{\beta-1}\exp(-u)}{\Gamma(\beta)}du$. More parametric models can be found in Klein and Moeschberger (2010).
- Examples of commonly used semiparametric models include (a) the Cox proportional hazards model. Denote X as the length-d covariate. Under the Cox model, $\lambda(t|X) = \lambda_0(t)\exp(\beta' X)$. Here, β is the length-d regression coefficient, β' is the transpose of β, and $\lambda_0(t)$ is the unspecified nonparametric baseline hazard; (b) the additive risk model, where $\lambda(t|X) = \lambda_0(t) + \beta' X$; and (c) the accelerated failure time model, where $\log(T|X) = \alpha + \beta' X + \varepsilon$. Here, the logarithm transformation can be replaced by other known, monotone increasing functions, α is the unknown intercept, and ε is the random error.
- With nonparametric models, the forms of the density (or distribution, survival, hazard) functions are left unspecified. Sometimes it is assumed that those functions satisfy certain properties. For example, it has been assumed that the hazard function is monotone or continuously differentiable.

Censoring and Truncation

Survival analysis is often more complicated than other types of regression analysis because of censoring and truncation.

The most commonly encountered censoring is right censoring, where the event time is only known to be after some time point. Right censoring occurs for subjects who have not experienced the event of interest when the follow-up ends. Other forms of censoring include left censoring and interval censoring. Left censoring occurs when the event of interest has already happened at the observation time, but it is not known exactly when. Interval censoring occurs when the event of interest is only known to happen in a finite time interval. It is possible that a study cohort is composed of subjects with different types of censorings.

Truncation happens when subjects with event times less than some threshold are not observed at all. Note that truncation is different from left censoring. For a left-censored datatum, we know the subject exists, whereas for a truncated datum, we may be completely unaware of the subject.

Maximum Likelihood Estimation and Inference

In survival analysis, the most commonly used estimation and inference techniques are based on the likelihood function. Assume that the observations are independent given the parameters. When constructing the likelihood function, we partition the data into four categories: uncensored, left censored, right censored, and interval censored. Denote θ as the generic unknown parameter. Then the likelihood function has the form

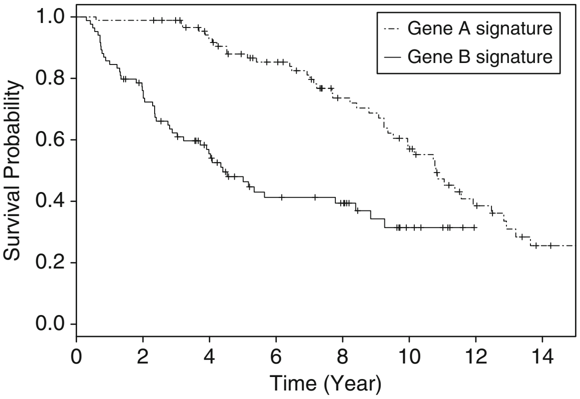
$$L(\theta) = \prod_{i \in uncensored} P(T = T_i|\theta) \times \prod_{i \in left\ censored} P(T \leq T_i|\theta) \times \prod_{i \in right\ censored} P(T > T_i|\theta) \times \prod_{i \in \text{interval censored}} P(T_{i,l} < T \leq T_{i,r}|\theta)$$

- For an uncensored datum, T_i is the actual event time. We have $P(T = T_i|\theta) = f(T_i|\theta)$;
- For a left-censored datum, the event time is known to be less than or equal to T_i. We have $P(T \leq T_i|\theta) = F(T_i|\theta) = 1 - S(T_i|\theta)$;
- For a right-censored datum, the event time is known to be greater than T_i. We have $P(T > T_i|\theta) = 1 - F(T_i|\theta) = S(T_i|\theta)$;
- For an interval-censored datum, the event time is known to be less than or equal to $T_{i,r}$ but greater than $T_{i,l}$. We have $P(T_{i,l} < T \leq T_{i,r}|\theta) = S(T_{i,l}|\theta) - S(T_{i,r}|\theta)$.

Once the likelihood function is properly constructed, most likelihood-based estimation and inference techniques are applicable.

Kaplan–Meier Estimator

The Kaplan–Meier (KM) estimator is a nonparametric estimate of the survival function. It can be used to measure the fraction of subjects surviving for a certain amount of time, for example, after treatment.



A representative plot of the KM estimator is shown above. In this plot, subjects with gene B signature die much more quickly than those with gene A signature. After 6 years, more than 80% of the subjects with gene A signature are still alive, whereas only 40% of the subjects with gene B subjects are alive.

Consider N subjects with the observed event times $t_1 \leq t_2 \leq \ldots \leq t_N$. Corresponding to each t_i is n_i, the number "at risk" just prior to time t_i, and d_i, the number of deaths at time t_i. The KM estimator of $S(t)$ is $\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$. When there is no censoring, n_i is the number of survivors just prior to time t_i. With censoring, n_i is the number of survivors less the number of losses (censored cases). For the KM estimator, the most commonly used variance estimator is the Greenwood's formula $v\hat{a}r(\hat{S}(t)) = \hat{S}^2(t) \sum_{t_i<t} \frac{d_i}{n_i(n_i - d_i)}$.

Logrank Statistic

Quite often researchers need to compare the survival functions of two or more groups. The logrank statistic, also called the Mantel–Cox statistic, can be used for such a purpose. It is a nonparametric statistic applicable to uncensored and right-censored data.

Consider two groups. Let $j = 1...J$ be the distinct times of observed events in either group. For each time j, let $N_{1,j}$ and $N_{2,j}$ be the number of subjects at risk in groups 1 and 2, respectively. Let $N_j = N_{1,j} + N_{2,j}$. Let $O_{1,j}$ and $O_{2,j}$ be the observed number of events in the groups, respectively, at time j. Define $O_j = O_{1,j} + O_{2,j}$. The logrank statistic is defined as $Z = \frac{\sum_{j=1}^{J}(O_{1,j}-E_{1,j})}{\sqrt{\sum_{j=1}^{J} V_j}}$. Here $E_{1,j} = O_j \frac{N_{1,j}}{N_j}$ and $V_j = \frac{O_j(N_{1,j}/N_j)(1-N_{1,j}/N_j)(N_j-O_j)}{N_j-1}$.

If the two groups have the same survival functions, the logrank statistic is approximately standard normal. If the hazard ratio is λ, there are a total of N subjects, ρ is the probability a subject in either group will eventually have an event (so that $N\rho$ is the expected number of events at the time of the analysis), and the proportion of subjects in each group is 50%, then the logrank statistic is approximately normal with mean $\log(\lambda)\sqrt{\frac{N\rho}{4}}$ and variance 1.

More Estimation and Inference Techniques

Although likelihood-based techniques are applicable to most estimation and inference problems encountered in survival analysis, they are not applicable to all or not necessarily the most convenient techniques. A family of techniques extensively used are martingale techniques. They are built on the observation that the observed event process less its expectation forms a martingale. Another family of techniques are empirical processes techniques, which are especially powerful with semiparametric models.

Advanced Survival Analysis

- Competing risk. In some situations, the endpoint may consist of several distinct events of interest and the eventual failure may be attributed to one event exclusively to the others. Under such a competing risk situation, both the cause-specific and overall hazard (survival, distribution) functions need to be modeled.
- Improper survival function. A survival function S is improper if $S(+\infty) > 0$. An improper survival function is used to describe a nonhomogeneous cohort with a subgroup that will never experience the event of interest. Corresponding survival models have been referred to as "cure rate" or "immune" models.
- Correlated observations, which may arise from, for example, family-based studies. Beyond adopting the aforementioned models for each individual observations, it is also necessary to model the correlation among subjects using, for example, frailties.

References

Fleming TR, Harrington DP (1991) Counting processes and survival analysis. Wiley, New York

Klein JP, Moeschberger ML (2010) Survival analysis: techniques for censored and truncated data. Springer, New York

Kosorok MR (2009) Introduction to empirical processes and semiparametric inference. Springer, New York

Lawless JF (2002) Statistical models and methods for lifetime data. Wiley, Hoboken

Sun J (2009) The statistical analysis of interval-censored failure time data. Springer, New York

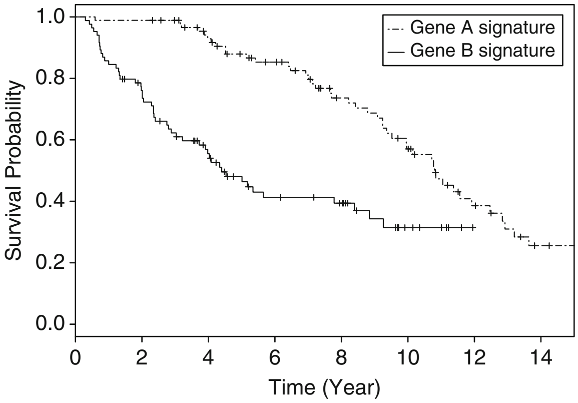
Survival Curve

Shuangge Ma
Yale School of Public Health, Yale University,
New Haven, CT, USA

Definition

A survival curve is a statistical picture of the survival experience of a group of subjects in the form of a graph showing the percentage surviving versus time.

In prognosis studies, survival curves are commonly used to compare the survival experience of subjects in different groups (for example, subjects with gene-A signature versus subjects with gene-B signature).



With a survival curve, the vertical (Y) axis gives the percentage of subjects surviving. The horizontal (X) axis gives the time after the start of the observation or experiment. Although expected survival curves can be smooth, those computed from real data (for example, the Kaplan–Meier estimators) are usually step functions.

Cross-References

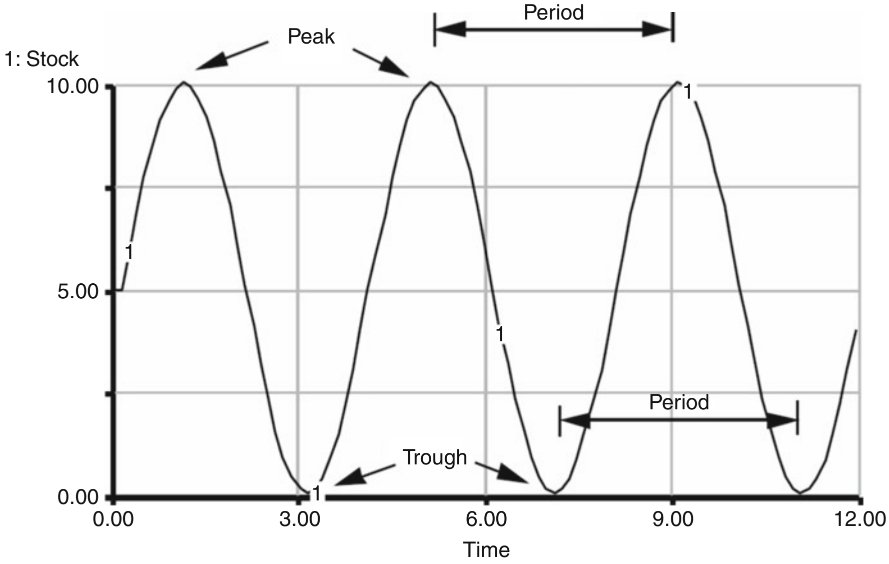
- ► Survival Analysis, Fundamental Statistical Techniques

Sustained Oscillation

Tianshou Zhou
School of Mathematics and Computational Sciences,
Sun Yet-Sen University, Guangzhou, Guangdong,
China

Definition

Sustained oscillation is a specific type of oscillation where each cycle of the oscillation is identical to the previous one. In control theory, sustained oscillation means continued oscillation due to insufficient attenuation in the feedback path. In physics, sustained oscillation means oscillation in which forces outside the system, but controlled by the system, maintain a periodic oscillation of the system at a period or frequency that is nearly the natural period of the system. Figure 1 is an example of sustained oscillation, showing the behavior of a stock exhibiting sustained oscillation over the course of 12 time units.



Sustained Oscillation, Fig. 1 Period, peak, and trough in sustained oscillation

Switch

▶ Toggle Switch, Switching Network

Switch Rate

Tianshou Zhou
School of Mathematics and Computational Sciences,
Sun Yet-Sen University, Guangzhou, Guangdong,
China

Definition

Switch rate refers to the speed at which a limit switch opens or closes a set of contacts after initial actuation.

Switching Function

▶ Boolean Function

Switch-Like Response

▶ Ultrasensitivity

Symbolic Model

Eberhard O. Voit
The Wallace H. Coulter Department of Biomedical
Engineering, Georgia Institute of Technology and
Emory University, Atlanta, GA, USA

Definition

A pathway system is often first conceptualized and diagrammed and the diagram is subsequently translated into equations. These equations are initially *symbolic*, because values for their parameters have not yet been specified.

Cross-References

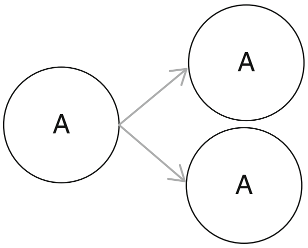
▶ Forward and Inverse Parameter Estimation for Metabolic Models

Symmetric Cell Division

Heiko Enderling
Center of Cancer Systems Biology, St. Elizabeth's
Medical Center - CBR 115D, Tufts University School
of Medicine, Boston, MA, USA

Definition

A symmetric cell division yields two daughter cells with equivalent cellular fate, i.e., a cell of type A gives rise to two daughter cells of type A.



Cross-References

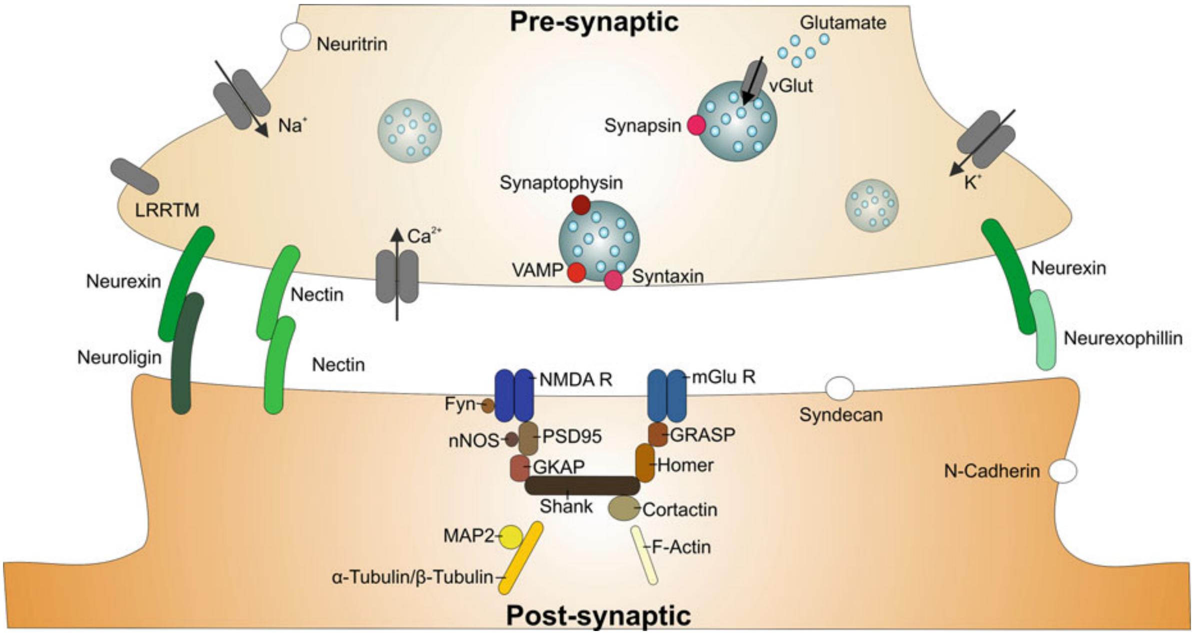
▶ Cancer Stem Cell Kinetics

Synaptic Proteins

Anne Gieseler
Molecular Pattern Recognition Research
(MPRR) Group, Otto-von-Guericke-University
Magdeburg Medical Faculty, Magdeburg, Germany

Definition

Synaptic proteins exert their function at the synapse – a highly specialized structure on the surface of nerve cells. The human brain contains around 10^{15} synapses,



Synaptic Proteins, Fig. 1 *Scheme of prominent proteins in a synapse*. The named proteins belong to different synaptic functions: presynaptic ion homeostasis (*gray*); vesicle-mediated neurotransmitter release (*red*); transsynaptic cell adhesion (*green*); postsynaptic neurotransmitter receptors (*blue*); proteins of the PSD (*brown*); and cell-cytoskeleton-related proteins (*yellow*)

which connect approximately 10^{12} neurons (Pocklington et al. 2006). This entry focuses on the so-called chemical synapse which represents the most frequent synaptic type in vertebrates transmitting electrochemical impulses from one neuron to another. This synaptic activity is fundamental for neurobiological processes such as learning, memory (Kandel and Schwartz 1982), and development of the nervous system (Changeux and Danchin 1976). The molecular machinery underlying synaptic function consists of a network of hundreds, if not thousands of distinct synaptic proteins (Yoshimura et al. 2004). Hence, synaptic proteins can be defined as a collection of highly specialized molecules, assembled as local molecular networks, enabling the synapse to process and transmit interneuronal information.

Characteristics

Synaptic transmission involves the release of neurotransmitters from presynaptic neurons (Fig. 1, “presynaptic,” upper part of the figure) and their binding by specific ion channels located at the surface membrane of postsynaptic neurons (Fig. 1, “postsynaptic” lower part of the figure). The presynaptic part is specifically characterized by presence of small vesicles. The vesicles contain a high abundance of synaptic vesicle proteins/peptides, such as components of the free synaptic fusion and retrieval machinery (e.g., SNARE proteins) (Südhof and Rothman 2009), and other proteins potentially involved in regulating the functional and structural dynamics of the nerve terminal (Coughenour et al. 2004). In postsynaptic membrane preparations isolated from homogenized brain tissue, the number of biochemically identified proteins has been reported to range from around 100 to more than 1,000. These proteins can be classified into distinct functional groups: organizers/cytoskeletal scaffold proteins; transporters and channels; sensors and signal transduction proteins; priming, docking and fusion apparatus; endocytotic and recycling machinery; components of energy supply; and linkers between the presynaptic and postsynaptic membranes (Peng et al. 2004; Cheng et al. 2006). Note that Fig. 1 depicts some prominent protein classes. The whole postsynaptic region, including the postsynaptic membrane and the corresponding submembrane structure (e.g., postsynaptic density – PSD, ribosomes, etc.), is one of the most complex and well-organized subcellular structure, or, macromolecular machine, in evolutionary biology. It appears to have the capacity to function in a semiautonomous manner.

Although a large number of synaptic proteins have been identified, it is important to note that not all of these proteins are present in each synapse, indicating a large range of functional diversity based on differential assembly of synaptic proteins (▶ synaptic toponome). Moreover, the kind and characteristics of the postsynaptic electrical activity of the postsynaptic membrane depends on the type and combination of receptor channels present in that membrane. This combination determines many fundamental properties, such as (1) reversion of the postsynaptic electrical potential, (2) the action potential threshold voltage, (3) ionic permeability of the ion channels, as well as, (4) the concentrations of the ions inside and outside of the neuronal cell surface membrane. Altogether these components determine whether interneuronal communication is stimulated or inhibited (excitatory or inhibitory synapse, respectively) (Xu 2011).

The strength of signaling between the pre- and postsynaptic neurons is based on the coordinated interaction by the different protein components. Synaptic plasticity is regulated by changes in the amount of receptors of the postsynaptic membrane; changes in the shape and size of dendritic spines; post-translational modification of PSD components; and modulation kinetics of the synthesis and degradation of proteins (Xu 2011). Integration of these processes leads to long-lasting changes in synaptic function and neuronal networks underlying learning-related plasticity, memory, and information processing in the nervous system of multicellular organisms.

The detailed analysis of synaptic plasticity based on the myriads of interactions of the synaptic protein network is a fundamental future challenge toward understanding the brain and the detailing of selective and efficient therapies against the so-called synaptic disorders, such as Alzheimer's, and Parkinson's disease (Reddy et al. 2010).

Cross-References

▶ Synaptic Toponome

References

Changeux JP, Danchin A (1976) Selective stabilization of developing synapses as a mechanism for the specification of neuronal networks. Nature 264(5588):705–712

Cheng D, Hoogenraad CC, Rush J, Ramm E, Schlager MA, Duong DM, Xu P, Wijayawardana SR, Hanfelt J, Nakagawa T, Sheng M, Peng J (2006) Relative and absolute quantification of postsynaptic density proteome isolated from rat forebrain and cerebellum. Mol Cell Proteomics 5(6):1158–1170

Coughenour HD, Spaulding RS, Thompson CM (2004) The synaptic vesicle proteome: a comparative study in membrane protein identification. Proteomics 4(10):3141–3155

Kandel ER, Schwartz JH (1982) Molecular biology of learning: modulation of transmitter release. Science 218(4571):433–443

Peng J, Kim MJ, Cheng D, Duong DM, Gygi SP, Sheng MJ (2004) Semiquantitative proteomic analysis of rat forebrain postsynaptic density fractions by mass spectrometry. Biol Chem 279(20):21003–21011

Pocklington AJ, Armstrong JD, Grant SGN (2006) Organization of brain-complexity-synapse proteome form and function. Brief Funct Genomic Proteomic 5:66–73

Reddy PH, Manczak M, Mao P, Calkins MJ, Reddy AP, Shirendeb U (2010) Amyloid-beta and mitochondria in aging and Alzheimer's disease: implications for synaptic damage and cognitive decline. J Alzheimers Dis 20(Suppl 2):499–512

Südhof TC, Rothman JE (2009) Membrane fusion: grappling with SNARE and SM proteins. Science 323(5913):474–477

Xu W (2011) PSD-95-like membrane associated guanylate kinases (PSD-MAGUKs) and synaptic plasticity. Curr Opin Neurobiol 21(2):306–312

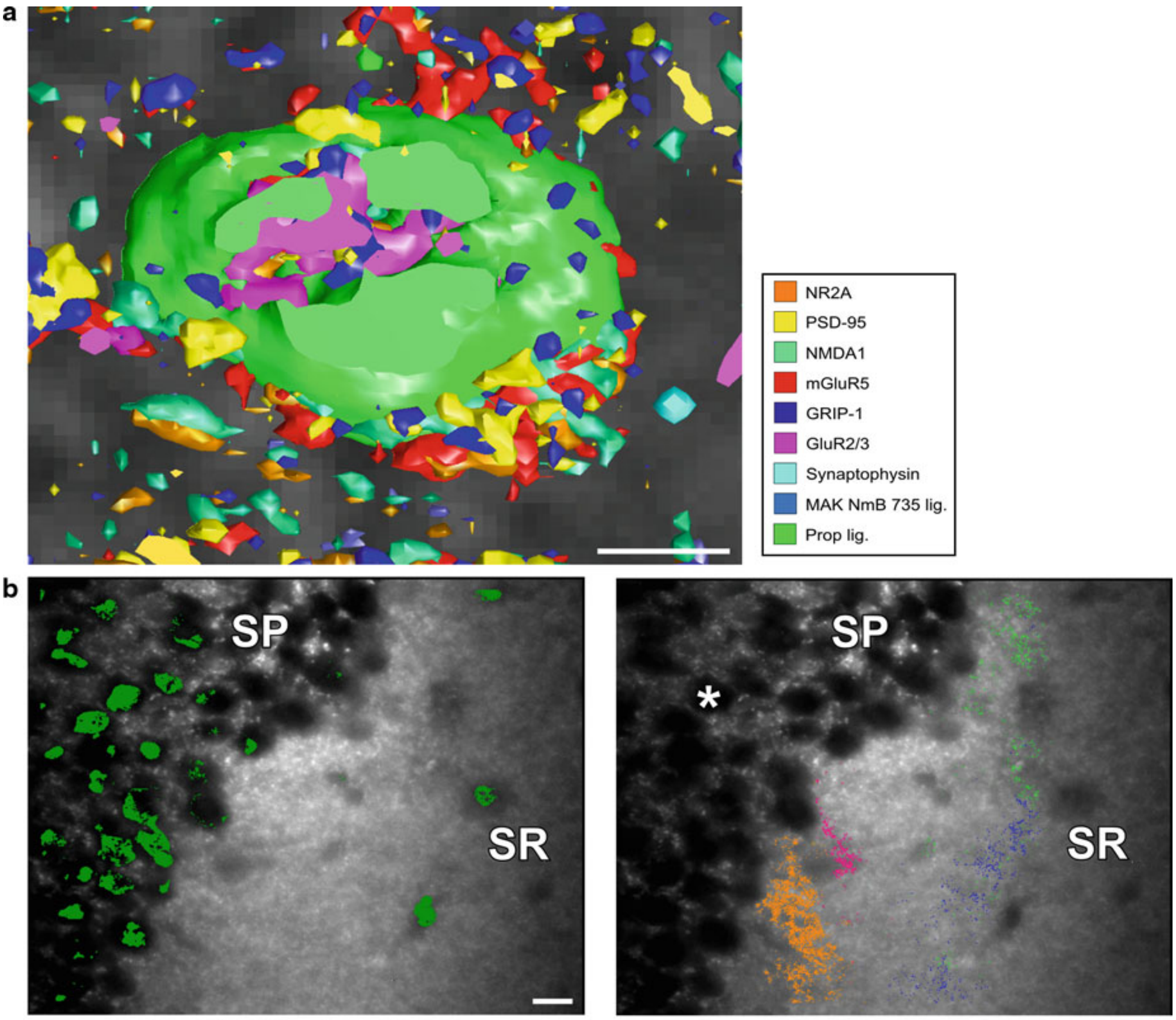
Yoshimura Y, Yamauchi Y, Shinkawa T, Taoka M, Donai H, Takahashi N, Isobe T, Yamauchi T (2004) Molecular constituents of the postsynaptic density fraction revealed by proteomic analysis using multidimensional liquid chromatography-tandem mass spectrometry. J Neurochem 88(3):759–768

Synaptic Toponome

Anne Gieseler
Molecular Pattern Recognition Research (MPRR) Group, Otto-von-Guericke-University Magdeburg Medical Faculty, Magdeburg, Germany

Definition

The term "toponome" describes one of the functional levels of the cell, such as genome, transcriptome, and proteome, and can be quantitatively described and deciphered by means of the ▶ TIS technology.



CMP No.	ApoE	CAT	ConA	GAP43	GluR1	GluR2	GluR5	GluR2/3	Grip1c	Interα	mGluR5	NEFHp	NeuN	Neurofl	NMDA	nNOS	NR2A	NR2B	Prop	Synaps	Synphys	Ubiqu	Freq.
1 🟧	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3.379
2 🟩	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	1	887
3 🟦	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	1	819
4 🟥	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	567

Synaptic Toponome, Fig. 1 *Exemplified synaptic toponome in the central nervous system.* (**a**) 3D imaging of the toponome of a single neuron inside rat spinal cord: colocalization map of seven surface-rendered signals of postsynaptic proteins on the surface of a nerve cell body (*red*, *yellow*, and *blue* colors). Note that these colors highlight regions in which these seven proteins are differentially assembled as multi-protein clusters; cell nucleus is marked in *solid green* for reasons of orientation (propidium iodide signal). This example illustrates that these postsynaptic toponome structures can be analyzed at high subcellular spatial resolution. Scale bar 10 μm. (**b**) 2D imaging of a fraction of the synaptic toponome present in a murine brain hippocampus tissue section fluorescently labeled for the synaptic marker protein synaptophysin (*light gray*) (5 μm thickness, area CA3). Four distinct mutually exclusive synaptic regions are shown (on the *right*, different colors), each of which expresses a distinct synaptic protein assembly in the stratum radiatum (SR) of hippocampus-CA3. Colors indicate specific protein colocation and anti-colocation codes (combinatorial molecular phenotyes, CMPs, unique to these highlighted subregions). Note on the *left* of the figure precisely the same visual field illustrating for orientation the location of neuronal cell bodies accumulating in the stratum pyramidale (SP), as indicated by their prominent stain for nuclear DNA (propidium iodide signal, corresponding to *round dark areas* spared for synaptic fluorescence signals, *light gray*: *asterisk* in the image on the *right*). The subregion-specific distinct CMPs (four distinct colors) characterizing the four distinct synaptic toponome fractions are illustrated in the color-decoding list (protein colocation and anti-colocation code, 1/0). Scale bar 12 μm (After Schubert et al. 2006; Bode et al. 2008)

It encompasses the topology of all proteins, protein complexes, and protein networks in a subcellular structure (Schubert et al. 2006). The ▶ synaptic toponome is defined as the entirety of protein networks of the synapse, in which proteins and protein clusters physically interact to form complexes and structures with a given spatial localization and function (Fig. 1) (Bode et al. 2008).

Understanding the toponome of the synapse, which controls normal and disease-related pathways of interneuronal communication, will provide access to the mechanisms underlying synaptic functions in health and disease. A novel approach to detect and map the functional molecular networks of synaptic proteins is the functional super-resolution (fSR) microscopy TIS (Schubert 2003; Schubert et al. 2011), a technology based on cyclical fluorescence imaging of proteins of morphologically cells and tissues. It can co-map thousands of distinct multi-protein clusters associated with synapses in brain tissue sections. Resulting toponome maps have revealed the existence of higher-order rules for the spatial organization of the synaptic toponome: synapses expressing the identical toponome are grouped together to define new functional regions inside known brain areas (Fig. 1). Exact description of these regions and their functional interaction is held to lead to understanding the functional compartmentalization of distinct neuronal qualities.

Cross-References

▶ Synaptic Proteins
▶ TIS Robot

References

Bode M, Irmler M, Friedenberger M, May C, Jung K, Stephan C, Meyer HE, Lach C, Hillert R, Krusche A, Beckers J, Marcus K, Schubert W (2008) Interlocking transcriptomics, proteomics and toponomics technologies for brain tissue analysis in murine hippocampus. Proteomics 8(6):1170–1178

Schubert W (2003) Topological proteomics, toponomics, MELK-technology. Adv Biochem Eng Biotechnol 83:189–209

Schubert W, Bonnekoh B, Pommer AJ, Philipsen L, Boeckelmann R, Maliykh J, Gollnick H, Friedenberger M, Bode M, Dress A (2006) Analyzing proteometopology and function by automated multidimensional fluorescence microscopy. Nat Biotechnol 24:1270–1278

Schubert W, Gieseler A, Krusche A, Serocka P, Hillert R (2011) Next-generation biomarkers based on 100-parameter functional super-resolution microscopy TIS. N Biotechnol, doi: 10.1016/j.nbt.2011.12.004

Synchronization

Xiaojuan Sun and Jinzhi Lei
Zhou Pei-Yuan Center for Applied Mathematics,
Tsinghua University of Beijing, Beijing, China

Definition

The word synchronization has originated from a Greek root (*σ&grupsilon;ν*: syn = the same, common and *χρόνος*: chronos = time), which means “to share the common time” (Boccaletti et al. 2002). Colloquially, synchronization is often referred to as an adjustment of the rhythms of oscillating objects due to their weak interactions. On the contrary, asynchrony (or desynchronization) refers to those situations that are not synchronized or uncoordinated in time.

The most well-known example of synchronization, now known as the ▶ synchronized oscillations, was discovered by Huygens in the seventeenth century in which two pendulum clocks hang on the same wooden beam. Many new types of synchronization have been found in the last 30 years, such as partial synchronization, complete synchronization, phase synchronization, general synchronization, lag synchronization, and ▶ synchronized switching, etc.

In the field of physiology, there are many synchronization phenomena (Glass 2001), for example, the synchronizing of circadian oscillators to the light–dark cycle. There are two possible mechanisms that can induce synchronization in multicellular systems: ▶ cell communication and common extracellular signaling.

Characteristics

Mathematical Formulations for Synchronization

Consider a system of N individuals whose dynamics is described by an ▶ Ordinary Differential Equation (ODE), Model

$$\frac{dx_i}{dt} = f(x_i), \quad (i = 1, \ldots, N). \tag{1}$$

Here x_i is either a scalar or vector that represents the state of the i'th individual. When there are interactions among the individuals, the system dynamics can be modeled as

$$\frac{dx_i}{dt} = f(x_i) + g_i(x, t), \quad (i = 1, \ldots, N). \tag{2}$$

Here $x = (x_1, \ldots, x_n)$, and $g_i(x,t)$ measures the interconnections among the individuals. The above system is said to be synchronized when all individuals (starting from different initial states) converge to the same dynamical process. Mathematically, there exists a common dynamical process $\varphi(t)$ such that

$$\lim_{t\to\infty} \|x_i(t) - \varphi(t)\| = 0 \tag{3}$$

for all i.

A simple example of synchronization described above is the ▶ Kuramoto model. Kuramoto model is one of the most representative models of coupled phase oscillators that displays a large variety of synchronization patterns (Acebrón et al. 2005). The Kuramoto model consists of a population of N all-to-all coupled phase oscillators $\theta_i(t)$ whose natural frequencies ω_i distribute with a probability density $g(\omega)$. The dynamics are described with

$$\dot{\theta}_i = \omega_i + \frac{K}{N}\sum_{j=1}^{N} \sin(\theta_j - \theta_i), \quad i = 1, \ldots, N. \tag{4}$$

In the Kuramoto model as described by (4), collective dynamics of the whole population is measured by the macroscopic complex *order parameter*

$$r(t)e^{i\phi(t)} = \frac{1}{N}\sum_{j=1}^{N} e^{i\theta_j(t)}. \tag{5}$$

Here the module $0 \le r(t) \le 1$ measures the phase coherence among all individuals and $\phi(t)$ is the average phase. Using the order parameter, we can rewrite (4) as

$$\dot{\theta}_i = \omega_i + Kr\sum_{j=1}^{N} \sin(\phi - \theta_i), \quad i = 1, \ldots, N. \tag{6}$$

The synchronizability of the oscillators is measured by the limit $R = \lim_{t\to\infty} r(t)$, which depends on the coupling strength K. When $K = 0$, all oscillators move incoherently and therefore $R = 0$. On the other hand, in the case of strong coupling that $K \to \infty$, all oscillators become synchronized to their average phase $\theta_i \approx \phi$, and hence (6) implies $R = 1$, which gives complete synchronization. For intermediate couplings, $K_c < K < \infty$ (K_c is the critical value of coupling strong to have synchronization), a part of oscillators are phase locked ($\dot{\theta}_i = 0$), and some oscillators are rotating out of synchrony with the locked oscillators. This gives the state of partial synchronization with $0 < R < 1$ (Acebrón et al. 2005).

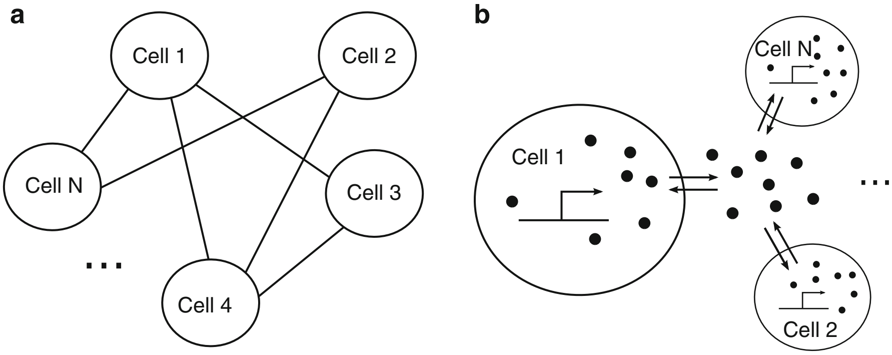
In the above definition, (3) is one of many mathematical descriptions of the word "synchronization," each of which represents different physical meaning of synchronization. For extend discussions, refer Brown and Kocarev (2000), Boccaletti et al. (2002), Arenas et al. (2008), and references therein.

Cell Communication and Synchronization

The above simple example of the Kuramoto model shows that coupling between oscillators is essential to have synchronization. In biological systems, ▶ cell communications among different cells through signaling molecules provide an important way of coupling. In this way, each cell acts individually in accordance to its intrinsic gene regulatory network and meanwhile communicates with other cells to form a complex network. Cell communications appear in two forms that are described by different mathematical formulations and induce synchronization through different mechanisms.

The first form of cell communication is direct cell-to-cell interconnection through various ways, such as ion channel conductivity, synaptic release of neurotransmitters, transcytosis, etc. Neural networks are well-known examples of this type in which neurons interact with each other to form a network by either ionic current or chemical synaptic transmission. Therefore, all involved cells connect with each other to form a complex network (Fig. 1a). The network dynamics can be described by a mathematical formulation of form

$$\dot{x}_i = f(x_i) + \sum_{j=1}^{N} K_{ij}\, h_{ij}(x_i, x_j), \quad (i = 1, \ldots, N). \tag{7}$$



Synchronization, Fig.1 Cell communication. (**a**) Direct cell-to-cell interconnection. (**b**) Secretion of diffusible signal molecules

Here x_i (usually a vector) represents the state of the i'th cell, K_{ij} equals 1 or 0 indicating the connectedness of the j'th to the i'th cells, and $h_{ij}(x_i, x_j)$ are functions describing the interactions between two cells. In this system, a cell does not affect other cells equally. Instead, the way how cells connect with each other is essential for the network dynamics. Synchronizability of a network is known to be associated with many characteristics of the network structure, such as the average shortest path length, the betweenness centrality, the clustering coefficient, and the degree correlations, etc. (Arenas et al. 2008). Nevertheless, knowledge for relationship between network topologies and synchronizability is still far from complete.

The other form of cell communication is communicating through indirect diffusible signal molecules (Fig. 1b). Each cell synthesizes and secretes signal molecules that can diffuse in extracellular matrix, and are endocytosed by other cells to regulate the gene regulatory networks in these cells. Through this form of communication, the extracellular environment serves as buffer chamber that averages out the concentration of signal molecules in each individual cell, and therefore drives the cell to display synchronization behavior (Danino et al. 2010). Mathematical description of population dynamics can be given by equations of form

$$\begin{aligned} \dot{x}_i &= f(x_i, s_i) \\ \dot{s}_i &= g(x_i, s_i) + D(s - s_i) \quad (i = 1, \ldots, N). \\ \dot{s} &= -\sum_{i=1}^{N} D(s - s_i) \end{aligned} \tag{8}$$

Here (x_i, s_i) represent the state of the i'th cells, with s_i the concentration of signal molecules, s stands for the concentration of extracellular signal molecules, D is the diffuse coefficient across cell membrane.

Synchronization Induced by Extracellular Stimuli

In additional to cell communication, extracellular stimuli is an alternative way to induce synchronization among cell populations. Synchronization of circadian clocks to the light–dark cycle is a familiar example of this type.

For a population of cells each of which has intrinsic clock, common external periodic stimuli can induce, enhance, or ruin collective rhythms (Zhou et al. 2007). This effect becomes significant if the external stimuli are resonant with the intrinsic clocks.

Extracellular fluctuations in environment can also induce synchronization. Stochasticity has been known to play important role in the dynamics of ▶ gene regulatory network. There are two sources of noise, including intrinsic noise that is inherent to the system and extrinsic noise that comes from fluctuations external to the system (▶ Noise, Intrinsic and Extrinsic). The two types of noise contribute oppositely in synchronization. In a population of cells, intrinsic noise (or intracellular noise) inherent to each individual cell tends to break the synchronization and induces cell-to-cell variances, while the extracellular noise that is common to all cells can induce collective dynamics and stochastically synchronize the population (Nakao et al. 2005; Zhou et al. 2005).

In a gene regulatory network with ▶ bistability, noise perturbation has been known to induce ▶ toggle switch. Furthermore, extracellular noise can induce ▶ synchronized switching among a population (Wang et al. 2007).

Cross-References

▶ Synchronous Model

References

Acebrón JA, Bonilla LL, Pérez-Vicente CJ, Ritort F, Spigler R (2005) The Kuramoto model: a simple paradigm for synchronization phenomena. Rev Mod Phys 77:137–185

Arenas A, Díaz-Guilera A, Kurths J, Moreno Y, Zhou CS (2008) Synchronization in complex networks. Phys Rep 469:93–153

Boccaletti A, Kurths J, Osipov G, Valladares DL, Zhou CS (2002) The synchronization of chaotic systems. Phys Rep 366:1–101

Brown R, Kocarev L (2000) A unifying definition of synchronization for dynamical systems. Chaos 10:344–349

Danino T, Mondragón-Palomino O, Tsimring L, Hasty J (2010) A synchronized quorum of genetic clocks. Nature 463:326–330

Glass L (2001) Synchronization and rhythmic processes in physiology. Nature 410:277–284

Nakao H, Arai K, Nagai K, Tsubo Y, Kuramoto Y (2005) Synchrony of limit-cycle oscillators induced by random external impulse. Phys Rev Lett 72:026220

Wang JW, Zhang JJ, Yuan ZJ, Zhou TS (2007) Noise-induced switches in network systems of the genetic toggle switch. BMC Syst Biol 1:50

Zhou TS, Chen L, Aihara K (2005) Molecular communication through stochastic synchronization induced by extracellular fluctuations. Phys Rev Lett 95:178103

Zhou TS, Zhang JJ, Yuan ZJ, Xu AL (2007) External stimuli mediate collective rhythms: artificial control strategies. PLoS One 2:e231

Synchronization of Oscillators

► Synchronization Oscillation

Synchronization Oscillation

Xiaojuan Sun
Zhou Pei-Yuan Center for Applied Mathematics,
Tsinghua university of Beijing, Beijing, China

Synonyms

Synchronization of oscillators

Definition

Oscillation is referred to be a repetitive variation in magnitude or position in a regular manner around a central point and is often described through the changing of phase with time. Synchronized oscillation is a phenomenon that a group of individual oscillators vary simultaneously from the central point, that is, phases of different oscillators share the common dependence with time. Synchronization of oscillators is thought to be a process of adjusting the rhythms of many oscillations due to (weak) coupling or external forcing (Pikovsky et al. 2001). The phenomenon of synchronized oscillation was first observed by Huygens in the seventeenth century. He found that two weakly coupled pendulum clocks (hanging on the same wooden beam) become synchronized in phase.

References

Pikovsky A, Rosenblum M, Kurths J (2001) Synchronization–a unified approach to nonlinear science. Cambridge University Press, Cambridge

Synchronization Switching

Xiaojuan Sun
Zhou Pei-Yuan Center for Applied Mathematics,
Tsinghua University of Beijing, Beijing, China

Definition

Synchronization switching is a phenomenon of collective behavior of ► toggle switches that all subsystems switch between the two stable steady states in a synchronous manner. In the case of cell behavior, synchronization switching can be induced by extracellular stimuli (Wang et al. 2007). This is to be distinct from incoherent switches of individual cells that are induced by intracellular driving forces.

References

Wang JW, Zhang JJ, Yuan ZJ, Zhou TS (2007) Noise-induced switches in network systems of the genetic toggle switch. BMC Syst Biol 1:50

Synchronous Model

Xi Chen, Wai-Ki Ching and Nam-Kiu Tsing
Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, University of Hong Kong, Hong Kong, China

Synonyms

Synchronization

Definition

There are two types of Boolean Network (BN) models for modeling genetic regulatory networks: synchronous model and asynchronous model, depending on whether or not the states of nodes (genes) are updated synchronously. In a synchronous model, all the states are updated synchronously in accordance with the functions assigned to them (see, e.g., Shmulevich et al. (2002)).

The following is an example of a BN having two genes with the truth table given in Table 1.

Suppose the current state is (0, 1). Then in the next step, these two nodes will be updated synchronously:

$$f^{(1)}(0,1) = 1 \quad \text{and} \quad f^{(2)}(0,1) = 0$$

where the first gene transforms from 0 to 1, and the second gene transforms from 1 to 0. Hence, the state of the network in the next step is (1, 0).

Synchronous Model, Table 1 The truth table

State	$v_1(t)$	$v_2(t)$	$f^{(1)}$	$f^{(2)}$
1	0	0	1	1
2	0	1	1	0
3	1	0	0	0
4	1	1	0	0

References

Shmulevich I, Dougherty E, Kim S, Zhang W (2002) From Boolean to probabilistic boolean networks as models of genetic regulatory networks. Proceedings of the IEEE 90:1778–1792

Syntactic Analysis

▶ Text Parsing

Synthetic Biology, Predictability and Reliability

Jinzhi Lei and Xiaojuan Sun
Zhou Pei-Yuan Center for Applied Mathematics, Tsinghua University of Beijing, Beijing, China

Definition

Synthetic biology is a new field of biological research that brings engineers and biologists together to design and build novel biomolecular components, networks, and pathways and to use these constructs to rewrite and reprogram organisms (Khalil and Collins 2010). In this field, synthetic biologists engineer complex artificial biological systems in order to investigate the natural biological phenomena and to induce various applications (Andranantoandro et al. 2006).

To achieve predictability and ▶ reliability for the synthetic biological systems, classical engineering strategies have to be extended to take into account the inherent characteristics of biological devices and modules, such as ▶ robustness, ▶ adaptation, ▶ specific response, etc. It has been shown that in many biological systems, it is the topological structure of gene networks that is responsible to achieve these properties, instead of the fine tuning of reaction rates. Such well-designed structures can be achieved by evolution of ▶ gene regulation.

Characteristics

Robustness

Robustness is a property that allows a system to maintain its functions despite both external and internal perturbations. The property of robustness is a fundamental feature of complex evolvable systems (Kitano 2004). Phenotypes of biological systems are robust against both mutations between generations and fluctuations from both internal and external origins during a single generation.

Robustness is a ubiquitously observed property of biological systems. For example, fate decision behavior of λ phage – either lysis or lysogeny – is robust against point mutations in the promoter region. Bacterial chemotaxis is highly sensitive to environmental changes over a broad dynamic range and is independent to ligand concentration. In embryo development, gradients of many morphogens are robust against gene mutations and changes in boundary concentration levels to enable reliable pattern formation.

Robustness is an important principle for designing biological circuitries: biological circuits are robustly designed such that their essential functions are nearly independent to biochemical parameters that might vary from cell to cell (Alon 2005).

In ▶ gene regulatory networks, there are specific architectural features that are known to be responsible for robustness. Furthermore, these features might be universal to many robust and evolvable complex systems. For example, system controls, modularity, decoupling, and redundancy are known to be basic mechanisms to provide robustness to the system (Kitano 2004).

System control consists of negative and positive feedbacks to attain a robust dynamic response. ▶ Negative feedback is a principle mode of control that enables robust response (or ▶ adaptation) to perturbations. ▶ Positive feedback contributes to robustness by amplifying the stimuli. It often produces ▶ bistability so that the activation level of downstream pathway clearly distinguishes from non-stimulated states and both states can be maintained under perturbations.

Modularity is an effective mechanism for containing perturbations or damage locally in order to minimize their effects on the whole system. Modules are widely observed in biological systems, which constitute semiautonomous entities that show dense internal functional connections but loose connections with environment. A single cell is an obvious example of a module that constitutes multicellular systems. Modules are often organized hierarchically.

Decoupling is a mechanism similar to modularity that isolates low-level variation from high-level functionalities.

Finally, the simplest strategy to ensure robustness is to provide multiple ways to achieve a specific function. In this way, failures in a specific component can be rescued by others.

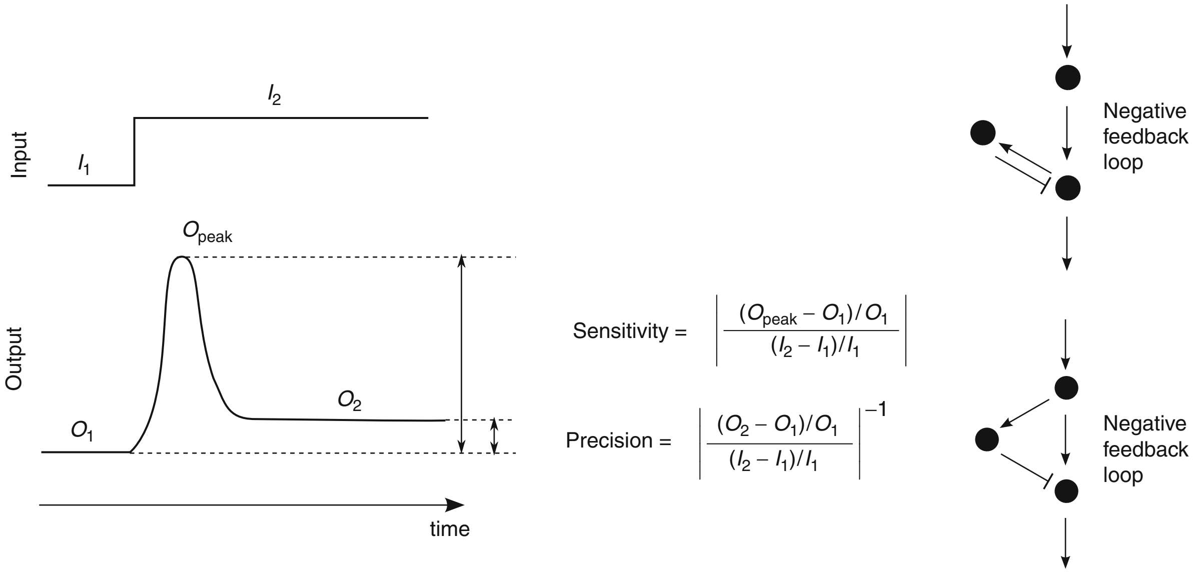
In biological systems, the above mechanisms are organized in a coherent architecture so that they are effective at the level of organisms.

Biological systems are evolved to be robust against certain perturbations but extremely fragile to unexpected perturbations. This robust yet fragile trade-off is fundamental to complex dynamic systems.

Adaptation

Many signaling systems show ▶ adaptation – the ability to reset their output to the original levels after a transient response to a stimulus. A mathematical definition of adaptation is given by two characteristic quantities: the circuit's sensitivity to input change and the precision of adaptation (Fig. 1) (Ma et al. 2009). Networks with perfect adaptation display both sensitivity (large peak output) and precision (the output returns exactly to the prestimulation levels) (Ma et al. 2009; Artyukhin et al. 2009). Examples of perfect or near-perfect adaptation include many important biological processes, such as light sensing, osmoresponse, calcium regulation, and bacteria chemotaxis.

Despite the large amount of possible topologies, there are very limited number of gene networks that can achieve perfect adaptation. Among all minimal framework of three-node topologies in which one node is for receiving inputs, one node for transmitting output, and one regulatory node, only two of them have perfect adaptation: ▶ negative feedback loops with a buffering node and incoherent ▶ feed-forward loops with a proportioner node (Fig. 1) (Ma et al. 2009). The regulatory node therein plays essential role in these two topologies. In negative feedback loops, the regulatory node is a buffer that integrates the difference between


Input
I_1
I_2
Output
O_{peak}
O_1
O_2
time
Sensitivity = $\left|\frac{(O_{peak} - O_1)/O_1}{(I_2 - I_1)/I_1}\right|$
Precision = $\left|\frac{(O_2 - O_1)/O_1}{(I_2 - I_1)/I_1}\right|^{-1}$
Negative feedback loop
Negative feedback loop

Synthetic Biology, Predictability and Reliability, Fig. 1 Network topologies for perfect adaptation (Ma et al. 2009)

network response and steady-state output. In feedforward loops, the regulatory node negatively regulates output so that it is proportional to the input. These results provide a rule for how to robustly engineer biological circuits capable of achieving adaptation.

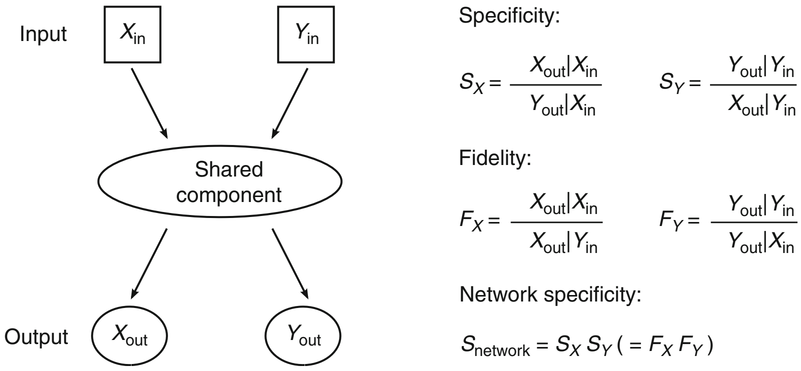
Specificity in Cell Signaling

Different cellular signal transduction pathways are often interconnected so that pathways cross-talk to each other. Therefore, different signals are often transmitted by common components, yet evoke distinct outcomes. Special strategies are needed to ensure that the specificity in cell response is maintained between different signal transduction pathways sharing similar (or identical) components, particularly when this occurs in a same cell.

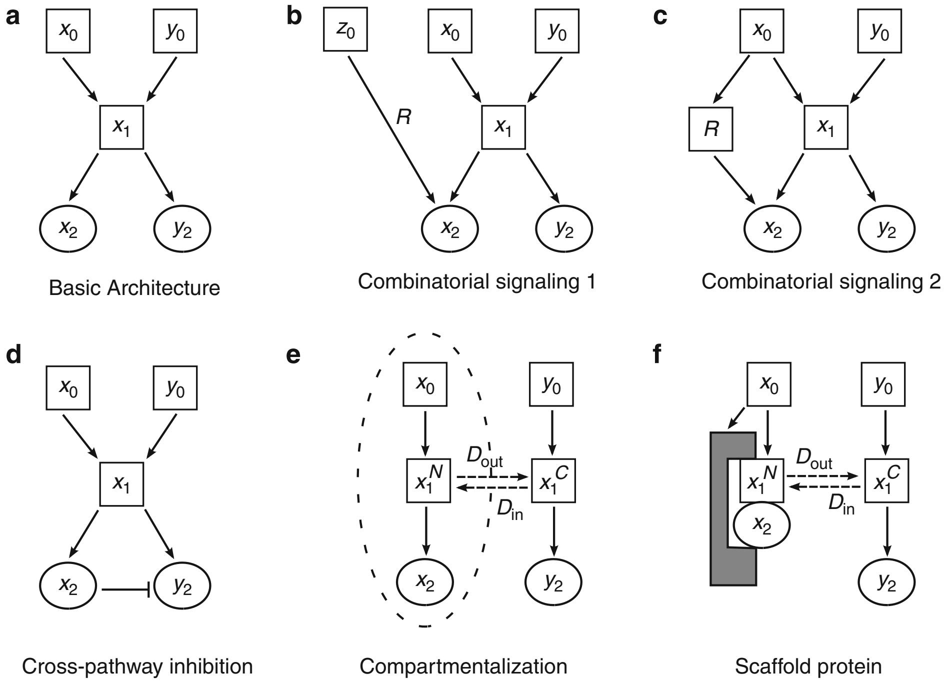
A framework of analyzing the specificity in cell signaling was proposed in Komarova et al. (2005) and Bardwell et al. (2007). Two properties, *specificity* and *fidelity*, are important for a network with cross-talk pathways to have specific response. Here the specificity of a pathway is measured by the ratio of its authentic output to its spurious output, and the fidelity of a pathway is defined as its output in response to an authentic signal divided by its output in response to a spurious signal (Fig. 2). The network specificity is then defined as the product of the all pathway specificities. In a network with specific response to different inputs, all pathways should have both specificity and fidelity greater than 1. If otherwise, paradoxical situations would occur so that one pathway activates another pathway's output or responds to another pathway's input more than its own (Komarova et al. 2005).

Special characteristics have to be emerged so that a gene network possesses specificity signaling. The simplest network with shared components, as shown by the "basic architecture" in Fig. 3a, cannot achieve specificity and fidelity. Fig. 3b–f shows several insulating mechanisms – combinatorial signaling, cross-pathway inhibition, compartmentalization, and the selective activation of scaffold proteins – that are found in nature and have been proposed to promote specificity (Bardwell et al. 2007).

In combinatorial signaling, the simultaneous action of two or more different signals is required to induce a response, so that the output of a pathway is determined by the combination of signals acting on a network. In such networks, the downstream component (x_2 in Fig. 3b–c) acts as a molecular "AND gate" or "coincidence detector" that integrates two separate inputs. In both types of combinatorial signaling, specificity of a network is inversely proportional to the amount of leakage k_{leak}, the basal activity of R when X is off. Thus, it is possible to obtain arbitrarily high levels of network specificity by making k_{leak} small (Bardwell et al. 2007).


Input
X_{in}
Y_{in}
Shared component
Output
X_{out}
Y_{out}
Specificity:
$S_X = \frac{X_{out}|X_{in}}{Y_{out}|X_{in}}$
$S_Y = \frac{Y_{out}|Y_{in}}{X_{out}|Y_{in}}$
Fidelity:
$F_X = \frac{X_{out}|X_{in}}{X_{out}|Y_{in}}$
$F_Y = \frac{Y_{out}|Y_{in}}{Y_{out}|X_{in}}$
Network specificity:
$S_{network} = S_X S_Y \,(= F_X F_Y)$

Synthetic Biology, Predictability and Reliability, Fig. 2 Specificity in cell signaling (Komarova et al. 2005)



Synthetic Biology, Predictability and Reliability, Fig. 3 Signaling network with shared components (Bardwell et al. 2007). (**a**) The "basic architecture"; (**b**) combinatorial signaling with independent and parallel inputs; (**c**) combinatorial signaling via the branching and reintegration; (**d**) cross-pathway inhibition; (**e**) compartmentalization; (**f**) the action of a scaffold protein

Cross-pathway inhibition (Fig. 3d) occurs when a downstream component inhibits the other one. An example of this type has been found in the mating pathway of yeast. In the cross-pathway inhibition as shown in Fig. 3d, the specificity of the network is larger than 1 only when the signal strength for the pathway *X* is stronger than that of the pathway *Y*. The requirement of fidelity requires additional conditions on both the relative strength and duration of the input signals (Bardwell et al. 2007).

Compartmentalization and the action of scaffold proteins are two simple insulating mechanisms to achieve specificity and fidelity by separating the pathways. In compartmentalization, different pathways are localized to different compartments or regions of a cell. Scaffold proteins can bind to two or more consecutively actin components of a pathway and accelerate the reaction rates between them. In fact, scaffolds can create the equivalent "microcompartments" by binding to multiple components of a given pathway. When deactivation rates are fast compared to exchange rates, a network with either compartmentalization or the action of scaffold proteins is able to have both specificity and fidelity (Komarova et al. 2005). Such networks reduce to the "basic architecture" when the exchange rates tend to infinity.

Evolution of Gene Regulation

Natural gene regulatory networks are robustly designed such that various cellular functions are carried out reliably despite the complex environment in and outside the cell. How could this be? Are there principles of natural designs that can help us to construct predictable and reliable circuits? Natural gene regulation networks are results of billions of years of evolution that random changes are induced and survival changes are selected. Evolution has played an important role in the selection of these networks (Perez and Groisman 2009).

Each individual has a genome and transcription factors from which its gene regulatory networks are derived. Random changes in genes or transcription factors may alter the genome and regulatory circuits. The resulting networks after change determine the gene expression patterns, cell behaviors, and hence fitness of the individual. The changes with better fitness in a changing environment will survive under Darwinian selection. Further, the resulting gene network has to be evolvable when environment continues to change. With this protocol, long-term evolution of complex gene regulatory networks in a changing environment can lead to a robust design and to carrying out cell functions reliably. Being hard to observe the evolution in laboratory, computational simulations have shown that evolution is able to generate gene networks that are robust with respect to both noise perturbations and mutations (Kaneko 2007).

References

Alon U (2005) An introduction to systems biology–design principles of biological circuits. CRC, Boca Raton

Andranantoandro E, Basu S, Karig DK, Weiss R (2006) Synthetic biology: new engineering rules for an emerging discipline. Mol Syst Biol 2:2006.0028

Artyukhin A, Wu L, Altschuler S (2009) Only two ways to achieve perfection. Cell 138:619–671

Bardwell L, Zou X, Nie Q, Komarova N (2007) Mathematical models of specificity in cell signaling. Biophys J 92:3425–3441

Kaneko K (2007) Evolution of robustness to noise and mutation in gene expression dynamics. PLoS One 2(5):e434

Khalil AS, Collins JJ (2010) Synthetic biology: applications come of age. Nat Rev Genet 11:367–379

Kitano H (2004) Biological robustness. Nat Rev Genet 5:826–837

Komarova N, Zou X, Nie Q, Bardwell L (2005) A theoretical framework for specificity in cell signaling. Mol Syst Biol 1:2005.0023

Ma W, Trusina A, El-Samad H, Lim W, Tang C (2009) Defining network topologies that can achieve biochemical adaptation. Cell 138:760–773

Perez JC, Groisman EA (2009) Evolution of transcriptional regulatory circuits in bacteria. Cell 138:233–244

Synthetic Models and Methods

C. Anthony Hunt[1], Andrew D. Gewitz[2] and
Tai Ning Lam[1]
[1]Department of Bioengineering and Therapeutic
Sciences, University of California, San Francisco,
CA, USA
[2]Institute for Computational and Mathematical
Engineering and School of Medicine, Stanford
University, Stanford, CA, USA

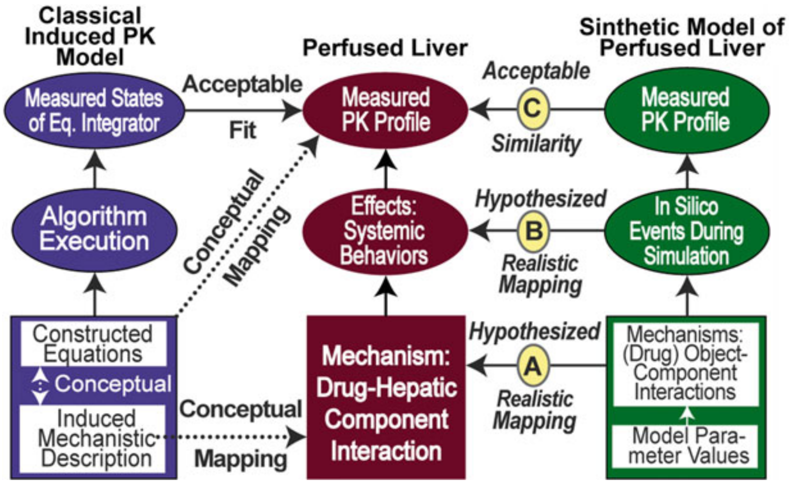
Definition

A synthetic (biomimetic) model (SM) is constructed from extant, autonomous software components whose existence and purpose are independent of the underlying model they comprise. It combines these elements in a systematic manner to form a coherent whole. A simulation, which is an executed instance of an SM, generates the model's behavior according to a prespecified computational design. Synthetic methods (Hunt et al. 2006) were invented to tease apart the underlying dynamics of complex systems (▶ Complex System; ▶ Complexity), in contrast to inductive models (▶ Induction) and related methods, which target prediction of the average behavior of systems in a continuous manner and are less concerned with the dynamics of individual model components.

Characteristics

Fundamentally Different Models

Successful development of new, safe, and effective drugs and treatment protocols requires deeper insight


Classical Induced PK Model
Perfused Liver
Sinthetic Model of Perfused Liver
Measured States of Eq. Integrator
Acceptable
Fit
Measured PK Profile
Acceptable
C
Similarity
Measured PK Profile
Algorithm Execution
Conceptual Mapping
Effects: Systemic Behaviors
Hypothesized
B
Realistic Mapping
In Silico Events During Simulation
Constructed Equations
Conceptual
Induced Mechanistic Description
Conceptual
Mapping
Mechanism: Drug-Hepatic Component Interaction
Hypothesized
A
Realistic Mapping
Mechanisms: (Drug) Object-Component Interactions
Model Para-meter Values

Synthetic Models and Methods, Fig. 1 Contrasting synthetic, wet-lab, and familiar inductive pharmacokinetic models

into the mechanisms involved, at all biological levels. Because the systems are complex and dynamic, we need new classes of simulation models, calibrated to current mechanistic knowledge (▶ Knowledge), to facilitate experimentation to discover plausible treatment outcomes and enable exploration of the origins of emergent phenomena as they unfold.

As an example of describing differences in model types, we consider hepatic drug disposition. Relationships between three different model types are illustrated in Fig. 1. Measures of drug loss from perfusate during liver perfusion (center) provide a classical pharmacokinetic (PK) profile (▶ Pharmacokinetic Modeling). During perfusion, hepatic components interact with transiting drug changing the drug's concentration-time profile. The leftmost diagram illustrates creating and fitting a PK model to the profile data. The researcher identifies patterns in the data in both an exploratory and a statistical fashion. An abstract, albeit idealized, hypothesized mechanism is formulated, thus establishing a conceptual mapping between this abstract description and hepatic cellular mechanisms. One or more PK equation models are selected to describe data patterns. There is also a conceptual mapping from description to equations. Software is executed to simulate equation output, enabling a quantitative mapping from simulation output to PK data. Metrics specify the goodness of fit.

In contrast, in the rightmost model, a mechanistic description is specified; it is similar, but not identical, to the one in the leftmost diagram. Software components are designed, coded, verified, and connected, guided by the mechanistic specifications. The end product is a collection of micro-mechanisms to be rendered in software. A concrete, realistic mapping (A) exists from in silico components and how they fit together to (1) hepatic physiological and microanatomical details and (2) drug interactions between components. Dynamics observed during simulations map to (B) corresponding dynamics (hypothesized to occur) within the liver. Simulation metrics provide a PK profile that is intended to mimic the liver perfusion PK profile. Quantitative metrics establish similarity between the two outflow profiles (C).

The rightmost model of Fig. 1 is an extant hypothesis: the components (objects) will illuminate a traceable mechanism upon simulation execution, a consequence of which will be emergent phenomena (e.g., as response following xenobiotic exposure). If similarities between the resulting simulated system (systems, modeling) and the referent system meet some prespecified criterion, the simulation stands as an abstract, plausible, mechanistic theory about events that may have occurred during wet-lab experiments (▶ Experiment).

Agent-Based Methods

The mechanisms that generate pharmacological phenomena are consequences of components at multiple levels of detail interacting in a complex manner with drug compounds. Simulation of such behavior may be achieved by adopting discrete-event modeling and simulation (M&S) methods (▶ Agent-based Models, Discrete Models and Mathematics) (Fishman 2001) in which component interactions can proceed according to stochastically defined rules (▶ Rule-based Methods) (Ullah and Wolkenhauer 2010).

Some biological components subject to wet-lab analysis possess a degree of spatial organization and are both semi-modular and quasi-autonomous. Synthetic models must be capable of exhibiting these same attributes. Component quasi-autonomy, coupled with realistic, spatially organized, biomimetic mechanisms, can be achieved using agent-based (agent-based modeling) and agent-oriented methods (An et al. 2009; Hunt et al. 2009), a discrete-event M&S method based on the object-oriented programming paradigm. The quasi-autonomous, decision-making software objects, called agents, can map to an organism, an organ, a tissue subsection, a cell, and/or a subcellular process. Other components, such as compounds (biologics/ xenobiotics), may themselves be represented as objects. Agents follow sets of rules that govern their actions and interactions. A biomimetic agent will have its own agenda, can schedule its own actions, and can dynamically change its operating logic. Agent-based SMs possess advantages when the modeler wishes to understand and simulate phenomena produced by systems of interacting components, and that makes them useful for gaining deeper insight into pharmacological phenomena within different individuals. An important characteristic of these models is that they yield an understanding of the mechanisms that generate disease-related phenomena and how compounds and treatment protocols influence these mechanisms by altering pharmacological phenomena.

Agents can be either atomic or composite. Increasing levels of organization define the system's granularity (▶ Granularity), or the extent to which the system is subdivided, and in which the smallest components are considered "atomic" (an atomic object contains no components of its own). Indeed, the more finely grained the system, the greater level of biological detail a given model wishes to examine. Agents, both atomic and higher-level, are designed to be inherently modular, and can be replaced (as distinct from being subdivided) with more finely grained components that exhibit similar behaviors. Components can be hierarchically nested, allowing the use of SMs to discover plausible bidirectional relations necessary for hypothesizing, instantiating, and in silico testing of multiscale genotype-phenotype relationships. Though in practice, a greater degree of nesting implies the need for more components and interactions, SMs should be just finely grained enough to produce targeted phenomena and achieve the simulation's specified uses.

Precise stoichiometric knowledge of component-compound interactions is rarely if ever available. An advantage of discrete-event methods is that both knowledge and ignorance (uncertainties) can be represented concurrently and simulated across a wide spectrum. The effective stoichiometry of interactions involving compounds can be represented at almost any convenient granularity level below that of the targeted phenomena, but the mappings from objects representing compounds to their referent molecules are not one-to-one. The presence of a compound can be represented as a property of a space or as mobile objects (Hunt et al. 2009). Mobile objects representing compounds can map to an arbitrary number of molecules. An important feature of the synthetic approach, from a pharmaceutical sciences perspective, is that each mobile object carries a list of physicochemical properties (PCPs) along with bioactivity attributes (the chemical entity is a CYP 2C9 substrate, etc.). In that way, SMs can accommodate any number of different compounds concurrently, which of course is ideal for studying and exploring drug-drug interactions (Lam and Hunt 2009). A component empowered to interact can use PCP information to adjust how it interacts.

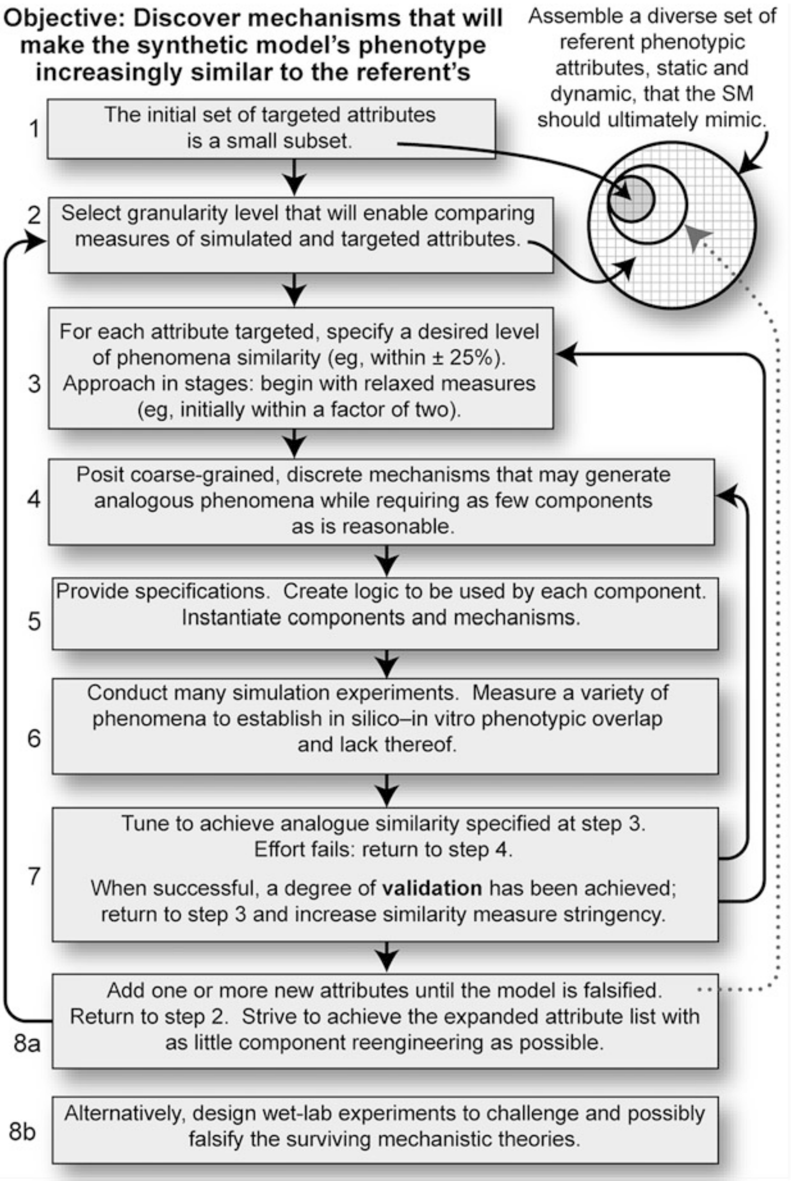
Parameterizations

Early in SM development, micro-mechanistic knowledge is insufficient to parameterize component-compound interactions a priori using PCPs. Micro-mechanism logic must be tuned for the first several compounds. As the set of compounds enlarges, inductive modeling methods (e.g., ▶ Ordinary Differential Equation (ODE), Model; ▶ Partial Differential Equation (PDE), Models) can be used to establish quantitative mappings from patterns in PCPs to patterns in parameter values of tuned component-compound interactions. Such mappings will be the synthetic model's counterpart to a structure-activity relationship. In subsequent rounds of SM refinement, the new knowledge contained in that relationship can be used, in some cases automatically, to provide an initial SM parameterization for the next chemical entity to be studied. Simulations using those parameterizations will stand as crude predictions of the new compound's attributes (Yan et al. 2008).

Iterative Refinement

The stages in scientific M&S are illustrated on the right side of Fig. 2. The micro-mechanisms form a hypothesis to be executed in silico and to yield output data. When the data fails to achieve a prespecified measure of similarity with referent data, the mechanisms are rejected as being plausible representations of wet-lab counterpart (Lam and Hunt 2010), and the cycle begins anew.

The iterative refinement protocol in Fig. 2 facilitates parsimony, which is important when building SMs that are expected to become increasingly complex. The protocol facilitates generating multiple


Objective: Discover mechanisms that will make the synthetic model's phenotype increasingly similar to the referent's
Assemble a diverse set of referent phenotypic attributes, static and dynamic, that the SM should ultimately mimic.
1
The initial set of targeted attributes is a small subset.
2
Select granularity level that will enable comparing measures of simulated and targeted attributes.
3
For each attribute targeted, specify a desired level of phenomena similarity (eg, within ± 25%). Approach in stages: begin with relaxed measures (eg, initially within a factor of two).
4
Posit coarse-grained, discrete mechanisms that may generate analogous phenomena while requiring as few components as is reasonable.
5
Provide specifications. Create logic to be used by each component. Instantiate components and mechanisms.
6
Conduct many simulation experiments. Measure a variety of phenomena to establish in silico–in vitro phenotypic overlap and lack thereof.
7
Tune to achieve analogue similarity specified at step 3. Effort fails: return to step 4.
When successful, a degree of validation has been achieved; return to step 3 and increase similarity measure stringency.
8a
Add one or more new attributes until the model is falsified. Return to step 2. Strive to achieve the expanded attribute list with as little component reengineering as possible.
8b
Alternatively, design wet-lab experiments to challenge and possibly falsify the surviving mechanistic theories.

Synthetic Models and Methods, Fig. 2 An iterative refinement protocol used to improve synthetic models mechanistic hypotheses and then eliminating the least plausible through experimentation.

The iterative refinement protocol shown in Fig. 2 is core to the scientific use of SMs. When faced with the task of building a scientifically relevant, multi-attribute SM in the face of significant gaps in knowledge, parameterizations, and model components must strike a balance between too many and too few. Doing so can be complicated by the fact that a validated, parsimonious, multi-attribute SM will be over-mechanized ("over-parameterized"; ▶ Overfitting) *for any one attribute*. Too many components and parameters can imply redundancy or a lack of generality; too few can make the SM useless for researching multi-attribute pharmacological phenomena. SMs are ideal for discovering mechanistic explanations in the form of relationships between components. However, because of the uncertainties reflected in poorly resolved model parameters and mappings to the referent system, they lack the precise predictive power of mathematical models. A SM such as that on the right side of Fig. 1 can be used to make predictions (quantitative or qualitative), for example, about where and how multiple "compounds" administered together may effectively interact.

Knowledge Embodiments

SMs have the potential to evolve into executable representations of what we know (or hypothesize) about biological systems during pharmacological exposure: they become executable biological (Fisher and Henzinger 2007) knowledge embodiments that provide concrete instances of that knowledge (right side of Fig. 1) rather than computational descriptions of conceptual representations (left side of Fig. 1). During simulation, a synthetic model demonstrates when, how, and where our knowledge matches or fails to match details of the referent system.

Such systems will represent the current best theory for how different pharmacological phenomena emerge within different individuals (Hunt and Ropella 2011). Adjusting (tuning) an SM to represent, for example, a normal rat liver in one in silico experiment, a diseased rat liver in another (as in Park et al. 2010), and a human liver in another will be relatively straightforward because uncertainty can be preserved. Automatable cross-validation of component functions can specify which features to tune and by how much. One may take copies of the same model and tune each separately to reflect differences in measured, patient-specific attributes.

Cross-References

- ▶ Agent-based Modeling
- ▶ Agent-based Models, Discrete Models and Mathematics
- ▶ Complex System
- ▶ Complexity
- ▶ Experiment
- ▶ Induction
- ▶ Knowledge
- ▶ Lattice-Gas Cellular Automaton Models
- ▶ Ordinary Differential Equation (ODE), Model
- ▶ Overfitting
- ▶ Partial Differential Equation (PDE), Models
- ▶ Partial Differntial Equations, Numerical Methods and Simulations
- ▶ Pharmacokinetic Modeling
- ▶ Rule-based Methods

References

An G, Mi Q, Dutta-Moscato J, Vodovotz Y (2009) Agent-based models in translational systems biology. Wiley Interdiscip Rev Syst Biol Med 1:159–171

Fisher J, Henzinger TA (2007) Executable cell biology. Nat Biotechnol 25:1239–1249

Fishman GS (2001) Discrete-event simulation: modeling, programming, and analysis. Springer, New York

Hunt CA, Ropella GEP (2011) Moving beyond in silico tools to in silico science in support of drug development research. Drug Develop Res. doi:10.1002/ddr.20412, Published online: 16 Dec 2010

Hunt CA, Ropella GE, Lam TN, Tang J, Kim SH, Engelberg JA, Sheikh-Bahaei S (2009) At the biological modeling and simulation frontier. Pharm Res 26:2369–2400

Lam TN, Hunt CA (2009) Discovering plausible mechanistic details of hepatic drug interactions. Drug Metab Dispos 37:237–246

Lam TN, Hunt CA (2010) Mechanistic insight from in silico pharmacokinetic experiments: roles of P-glycoprotein, Cyp3A4 enzymes, and microenvironments. J Pharmacol Exp Ther 332:398–412

Park S, Kim SH, Ropella GE, Roberts MS, Hunt CA (2010) Tracing multiscale mechanisms of drug disposition in normal and diseased livers. J Pharmacol Exp Ther 334:124–136

Ullah M, Wolkenhauer O (2010) Stochastic approaches in systems biology. Wiley Interdiscip Rev Syst Biol Med 2: m385–m397

Yan L, Sheihk-Bahaei S, Park S, Ropella GE, Hunt CA (2008) Predictions of hepatic disposition properties using a mechanistically realistic, physiologically based model. Drug Metab Dispos 36:759–768

SysMO

Franco du Preez
SysMO-DB team, Manchester Centre for Integrative
Systems Biology, University of Manchester,
Manchester, UK

Definition

SysMO (http://www.sysmo.net) is a Pan-European consortium studying the Systems Biology of Microorganisms, containing over 100 research groups working on eight projects, generating a wide range of data including transcriptomics, proteomics, metabolomics, and reaction kinetics. The goal of the SyMO projects is to combine experimental data and mathematical models for the description of dynamic molecular processes. SysMO was started in 2006 and will run up to 2014 in two funding periods.

System Identification in Signal Transduction, Experimental Perturbations

Gaurav Dwivedi, Nnenna A. Finn, Eberhard O. Voit
and Melissa L. Kemp
The Wallace H. Coulter Department of Biomedical
Engineering, Georgia Institute of Technology and
Emory University, Atlanta, GA, USA

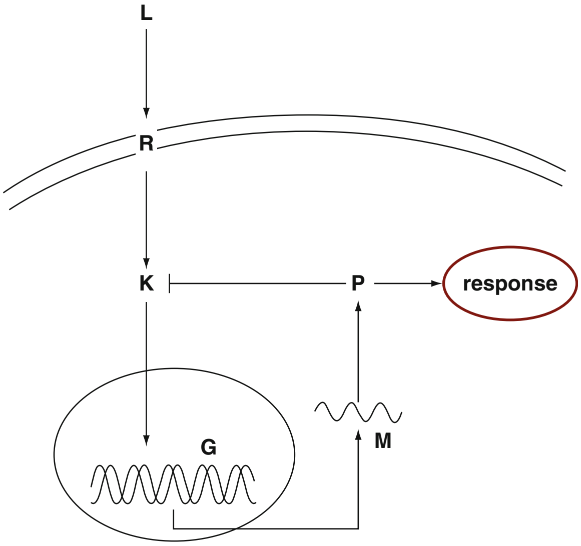
Definition

This essay discusses specific types of experiments that are useful for the quantitative characterization of signaling systems. Since signal transduction often involves genes and proteins; distinctly different experimental perturbations can shed light on various aspects of the signaling system.

Characteristics

Creating Experimental Perturbations

An important goal in systems biology is to construct, as completely as possible, abstract representations of biological systems that can be analyzed with computational techniques. The resulting *biological network models* define the system in a graphical or mathematical manner that contains information about the interaction partners of an entity and the strengths of these interactions (Alon 2003). Cellular signaling pathways comprise a particularly interesting type of biological network. These networks relay environmental signals to the interior of the cell and initiate appropriate responses. Figure 1 illustrates a generic signaling network. An environmental cue in the form of ligand L activates the cell surface receptor R. The activated receptor protein often undergoes a conformational change, which activates the protein K, which in turn induces the transcription of gene G into the corresponding mRNA M. M is translated into protein P, which triggers an appropriate response to the extracellular signal. The response could take the form of targeted gene expression, cell proliferation, apoptosis, or some other physiological event. P also acts to inhibit or inactivate protein K, thereby turning off the signal over time. It is clear that this system, as is typical of signal response systems, is comprised of functional elements at the genome (G), transcriptome (M), and proteome (R, K, P) levels. This multi-scale operation is



System Identification in Signal Transduction, Experimental Perturbations, Fig. 1 Generic signal transduction system. Ligand L binds to cell surface receptor R and activates it. Active R activates K, which induces expression of P. P triggers a response and at the same time suppresses K significant, because it allows experimental perturbations that may be introduced into the signaling system at the genome, transcriptome, or proteome level.

Experimentally introduced perturbations can be used as the basis for inferring the features of a biological system. To this end, the same or similar perturbations are simulated in corresponding quantitative models, which make predictions regarding the expected responses of the natural system. Sophisticated computer-aided comparisons between experimental and simulation results together with iterated analyses eventually lead to insights into the true structure of the system.

Experimental perturbations can be applied in different ways. Gene-knockout techniques and mutagenesis can be applied to perturb the system at the level of DNA (G in Fig. 1). By interfering with the gene sequence, these methods result in altered gene expression and a corresponding loss of function. *RNA interference* using synthetic *siRNA* or expressed *shRNA* constructs, can perturb cellular systems at the transcript level by “knocking down” (i.e., silencing) genes and thereby reducing the number of mRNA copies in the cell (M) (Leung and Whittaker 2005). Both approaches ultimately attenuate the concentration of the targeted protein (R, K, P). Finally, perturbations at the protein level can be created with stimulating agents, such as natural receptor *ligands* (L) or *agonists* that activate the signal. Opposite perturbations are possible with *chemical inhibitors* and *antagonists* that suppress receptor activation.

Kinetic Models of the System

The most common way of quantitatively describing the kinetics of a signal transduction pathway is through a system of ordinary differential equations (ODEs) (Aldridge et al. 2006). Multiple mathematical formats are available to describe the temporal progress of individual biochemical reactions and can be chosen as deemed fit. Commonly used models include mass action, Michaelis–Menten and power-law kinetics (Voit et al. 2000). Irrespective of the particular model choice, the ODE system must be parameterized for further analysis. This process consists of determining numerical values for quantities representing the initial concentrations of the species in the model and for quantities determining the strengths of interactions and the speed of the biochemical reactions. For example, consider the Michaelis–Menten description (Eq. 1) of a unimolecular elementary reaction where substrate S is converted into product P.

$$S \rightarrow P$$
$$-\frac{d[S]}{dt} = \frac{d[P]}{dt} = \frac{v_m[S]}{K_m + [S]} \tag{1}$$

$[S]$ and $[P]$ represent the concentrations of the biochemical species S and P, respectively, as functions of time. To solve Eq. 1, the quantities v_m and K_m need to be specified along with the values of $[S]$ and $[P]$ at the beginning of the experiment, usually for time 0. This specification is accomplished through a statistical evaluation of specific experiments or from literature information. An ODE system made up of multiple reaction rates requires similar quantities, namely, rate and interaction parameters for individual reactions and the initial values of all the participating biochemical species (Voit et al. 2000). Thus, an ODE model can be written for the system in Fig. 1 using the same principles for all interactions. As a variation to including all components explicitly, the model may be simplified by making some species implicit. For example, the steps leading from the activation of transcription by K to the production of protein P could be simplified by making gene G and mRNA M *implicit*. The production of P would then follow directly from K. It would be modeled with a more complicated function and possibly a time delay, and the species G and M would no longer be visible in the model.

Implementing Perturbations in the Model

It is straightforward to simulate experimental manipulations with the parameterized model, if all components are explicitly included. Because all components and interactions have their unique representation in the model, an experimental perturbation applied at the level of a gene, RNA, or protein is directly implemented in the corresponding component or process of the model. If a component is only implicitly accounted for in the model, the perturbation can still be modeled, but the procedure is not quite as straightforward.

A stable gene-knockout disturbs the functionality of a gene sequence and leads to the absence of functional protein. This perturbation is easily modeled by setting the initial values and production rates of mRNA and protein to zero. Mutagenesis can result in a protein that is completely nonfunctional with respect to its natural role or exhibits altered reactivity. A nonfunctional protein can be modeled as in the case of a knockout by setting the mRNA and protein initial values and production rates to zero. By contrast, perturbation through mutagenesis resulting in a protein with altered reactivity is modeled by adjusting the rate parameters of the reactions in which the protein participates. A perturbation achieved with a transgenesis experiment can enhance the level of an existing protein or even introduce a new protein into the system. The former situation is easy to resolve since only the production rate and the initial value of the existing protein is adjusted. In the latter case, new reactions may come into play, and these can change the structure of the system. Such changes require that entirely new species are introduced and that additional rate and interaction parameters are determined.

A knockdown experiment targeting the transcripts of a gene attenuates the protein level without completely abolishing it. In a kinetic model that explicitly includes mRNA, a changed protein level is easily implemented by lowering the initial value and enhancing the decay rate of mRNA induced by RNA interference. If mRNA is only implicit in the model, the effect of knockdown can be modeled by decreasing the initial value and production rate parameters of the protein to a level commensurate with the extent of the knockdown.

Perturbations can be introduced at the level of proteins to enhance or suppress a signaling pathway. Small chemical molecules can activate a signaling pathway through the receptor or even bypass the receptor and activate signaling directly. For example, agonist molecules can activate receptor-induced signaling when the natural ligand is not present. By contrast, an antagonist molecule may suppress receptor-induced signaling in the pathway. The direct effects of agonist (antagonist) molecules can be modeled by taking into account the enhanced (attenuated) rate of receptor activation in response to these molecules. Small chemical molecules can also be used to inhibit the activities of enzymes in the signaling pathway, and their effects can be translated into the kinetic model by decreasing the catalytic rates of the target enzymes. In all these cases, an experimental perturbation can be translated to the kinetic model by altering the corresponding initial conditions or rate parameters or by slightly adjusting the structure of the system.

Experimentally introduced perturbations can sometimes have unexpected off-target effects. For example, a small-molecule inhibitor with a known target enzyme may also have lower yet significant specificity for other proteins in the system, hence creating additional variation (Davies et al. 2000). Other perturbation measures, such as the introduction of an exogenous transgene that reacts with existing species in the network, could introduce new reactions or alter the existing parameters of the model. These newly introduced or altered parameters may not always be simple to determine and can complicate the problem of parameter estimation instead of simplifying it.

Data obtained from perturbation experiments can be valuable in several ways. Importantly, they can help establish a hierarchical relationship between the elements of the signaling network. Knocking out protein K in Fig. 1 would prevent synthesis of P even in the presence of a stimulatory signal from ligand L. This result, for example, establishes that P lies downstream of K in the network. An inhibitor of enzyme P would shut off the inhibition of K in Fig. 1. Measurements of K under perturbed and unperturbed conditions provide additional data that can be used to estimate rate parameters in the model. Perturbation data can be further used to validate a candidate kinetic model of the system. For example, protein K could be measured over time after knocking down enzyme P using RNAi. Without using this experimental result to train the model, the model could then be queried by simulating knocked-down levels of P. Finally, perturbation experiments can be used with a computational model to unravel regulator principles operating in signal transduction systems.

Time course measurements on species in the signaling network obtained under different perturbation conditions require more effort but provide a much richer data set that can aid the process of parameter estimation and system identification (Chou et al. 2009). Once the signaling system is reliably identified, simulating the effects of many additional perturbations generates hypothetical drug targets and informs experimental design for biotechnology and pharmaceutical applications (Schoeberl et al. 2009).

Cross-References

- ▶ Data Integration
- ▶ Dynamical Systems Theory, Delay Differential Equations

▶ Gene Regulatory Networks
▶ Metabolic and Signaling Networks
▶ Ordinary Differential Equation (ODE), Model
▶ Pathway Modeling, Metabolic

References

Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK (2006) Physicochemical modelling of cell signalling pathways. Nat Cell Biol 8:1195–1203

Alon U (2003) Biological networks: the tinkerer as an engineer. Science 301:1866–1867

Chou I-C, Voit EO (2009) Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Math Biosci 219:57–83

Davies SP, Reddy H, Caivano M, Cohen P (2000) Specificity and mechanism of action of some commonly used protein kinase inhibitors. Biochem J 351:95–105

Leung RK, Whittaker PA (2005) RNA interference: from gene silencing to gene-specific therapeutics. Pharmacol Ther 107:222–239

Schoeberl B, Pace EA, Fitzgerald JB, Harms BD, Xu L, Nie L, Linggi B, Kalra A, Paragas V, Bukhalid R, Grantcharova V, Kohli N, West KA, Leszczyniecka M, Feldhaus MJ, Kudla AJ, Nielsen UB (2009) Therapeutically targeting ErbB3: a key node in ligand-induced activation of the ErbB receptor-PI3K axis. Sci Signal 2:ra31

Voit EO (2000) Computational analysis of biochemical systems. A practical guide for biochemists and molecular biologists. Cambridge University Press, Cambridge, xii + 530 pp

System-Immanent Conditions

▶ Intrinsicality

Systems Biology

▶ Biomarker Discovery, Knowledge Base

Systems Biology Applications in Drug Discovery

Nagasuma Chandra
Department of Biochemistry, Indian Institute of Science, Bangalore, India

Definition

It is perhaps no exaggeration to say that systems biology is transforming the way we understand biological processes in health and disease. Its influence on drug discovery is not in the least surprising, marking an important paradigm shift in drug discovery science (Kitano 2002; Butcher et al. 2004; Hood and Perlmutter 2004). Several factors such as availability of publicly accessible databases containing genome sequences, functional and structural data of macromolecules, high-throughput experimental profiling, protein-protein interactions and pathway models, as well as adaptation and application of computational methods for efficient data mining and modeling, have all been directly contributing to this paradigm shift. Systems approach is really not new to the study of diseases or medicine. There has been sufficient emphasis in literature on the merit of wholistic approaches and phenotypic medicine. Heavy reliance on whole animal models and need for rigorous clinical trials stand out as evidence to such thinking. Yet, there is a remarkable difference with those approaches and the emerging systems biology approaches. While conventional approaches have surely benefitted from systems philosophy, comprehension of the “system” is at best only implicit in these. In fact the system is most often more of a “black box,” which only facilitates a systems output as a “readout,” but does not tell us why or how, such an output results. Current approaches to systems biology, on the other hand, adopt a bottom-up strategy and reconstruct the system brick by brick, thus facilitating an understanding of the mechanistic basis of individual molecular events, leading to the ability to simulate different scenarios, thus enabling predictions. It must be noted that systems biology approaches thus also differ from the theoretical “spherical cow” type of abstractions. Data describing various complex real-life phenomena, in the form of multilevel “*omics*” data on various fronts are increasingly becoming available, making such realistic systems level modeling feasible and in fact a necessity.

One of the main challenges is to build complete models with high enough resolution accounting for presence of all components, interactions, and reactions such that variation in genotypes at the molecular level can be mapped onto phenotypic differences. For example, it would be desirable to understand not merely a relation between a specific gene mutation to disease susceptibility or disease prognosis, but to reckon why that mutation should result in such a phenotype. There is still a paucity of quantitative or experimentally validated data required for model building. Lack of established protocols for high-throughput validation of the predictions is yet another problem. Notwithstanding these problems, the benefits of using systems level analyses far outweigh the current limitations. One of the biggest advantages is that predictions of the outcomes of a variety of scenarios can be made through simulations. In addition, systems level modeling enables dissection of the roles of individual components and their interactions with other components in the system.

Available methods for modeling complex networks span different abstraction levels and capture such complex cellular behavior through mathematical equations describing molecular interactions. Methods such as kinetic modeling using ODEs, stochastic modeling, flux balance analysis, and metabolic control analysis have been developed to model metabolic and signaling pathways (Orth et al. 2010). Kinetic modeling methods have been used extensively for specific pathways where detailed reaction rate information is available, while the reaction flux-based methods, although semi-quantitative in nature, cater to study of genome-scale metabolic networks and measure the relative fluxes of individual reactions for a given set of optimization criteria. Flux variability analysis enables the study of the range of concentrations, a particular reaction flux could adopt. Networks involving molecular interactions and influences at a genome scale have been studied using graph-theoretical methods and help in providing a molecular connectivity map in the cell. Simulations using synthetically constructed models enable dissection of the role of individual elements and underlying dynamics of complex systems. Higher levels of abstraction, where an entire process is described in a couple of equations, enable generalization of different phenomena with similar characteristics and thus help in providing a bird's-eye view of the system and find common patterns in related systems. An urgent need presently is to generate data and develop scalable quantitative predictive models, built with the foundations of biochemical knowledge. While genome sequence data is available for many organisms, quantitative data reflecting reaction velocities, interactions strengths, and temporal dynamics is relatively scarce, making it difficult to perform accurate quantitative predictions on several systems. Systematic experiments to obtain quantitative measurements of the components and the parameters that govern cellular behavior will be very useful for this purpose.

Once obtained, a systems view of the disease, can be used for addressing a variety of questions such as (a) identifying optimal strategies for treating a given disease; (b) identifying which proteins would serve as ideal drug targets to achieve the desired strategies (Rappuoli and Aderem 2011); (c) identifying which targets are druggable; (d) identifying which targets would be best suited for particular disease states such as actively dividing bacteria in an infectious disease; (e) understanding effects of a given drug, which translates to understanding pharmacodynamics and pharmacokinetics and hence ranking drug candidates; and (f) repurposing drugs used for other pathological indications. Besides these, the additional advantages of systems biology come from its ability to address important but difficult aspects of drug discovery such as polypharmacology, emergence of drug resistance, drug safety, and personalized medicine (Boran and Iyengar 2010). The entries in this section illustrate many of these aspects.

The study of disease through the integration of clinical, morphological, quantitative, and molecular parameters into detailed networks that can be studied using well-established mathematical techniques can be described as systems pathology. Systems models are developed that explain pathophysiological processes in their entirety and generate testable hypotheses. Similarly, systems pharmacology is emerging as a discipline in its own right, which addresses drug action including pharmacokinetics, pharmacodynamics at a systems level. A wide variation is seen in the response to treatment in different patients both with respect to beneficial effects as well as adverse reactions. Pharmacogenomics, another emerging discipline, seeks to address molecular basis of such variability at a genome scale and has the potential to lead to personalized medicine in the future. A systems perspective enables a study of the outcome of drug treatment as a nonlinear function of multiple events of molecular recognition and biochemical and biophysical reactions involving drug transporters, metabolizers, and intended targets. A further degree of complexity is brought about by epigenetic modifications of these classes of genes. Thus, besides the genetic variations, it is important to study epigenetic variations between individuals and in populations.

Complex biological processes are represented as networks in terms of specific interactions among thousands of molecules. Genome-scale molecular networks have been constructed for a variety of organisms (Lee et al. 2009) and are known to be significantly different from random networks and exhibit specific properties in terms of their structure and organization. Specific rewiring of the interactions is known to occur in disease, making a network perspective important to identify and characterize a disease, elucidate disease mechanisms, and identify important biomarkers and drug targets.

Experimental data such as genome-wide gene expression profiles have been incorporated onto networks to derive "response networks" specific to certain conditions such as in the case host responses to an infection, disease progression, or drug treatment (Forst 2006; Ideker and Krogan 2012). Variations in a network between two states can be mapped onto specific modules representing a closer set of connected reactions such as in a pathway or a process. Such modules often contain characteristic features to represent a signature of the disease. Topological analysis and comparison of relevant networks can help in identification of such signature patterns and further link it to the functional categories of the individual components. Several examples of such studies sometimes termed as gene or protein enrichment analysis of networks are available in literature (Glaab et al. 2012) where specific genes in a module are mapped to known pathways and gene ontology terms and interpreted in terms of which pathway or functional category is overrepresented in a module.

Discovery of new drugs for diseases such as cancer require an understanding of the complex nonlinear interactions among the hundreds of molecular components contributing to cancer pathogenesis. Alterations in networks between health and disease provide significant clues as to the mechanisms that could be targeted with drugs. Networks then, rather than individual genes or proteins, need to be targeted for drug discovery. An example of a systems biology approach has been described for melanoma, where knowledge of signaling networks with feedback loops and redundant pathways has led to exploring targets for possible combination drug therapy. An example for glioblastoma is described in the drug discovery section. In the case of Parkinson's disease, a neurodegenerative movement disorder, the dynamics and interdependence of the disease pathways, their temporal order, synergy, and regulation has been addressed through systems biology–based dynamic modeling and predictive analyses supported by experimental data. Such knowledge will ultimately help in design of better optimized drugs and more importantly reduce the risk of failure in the discovery pipeline.

Networks among different patients with the same disease could explain differences in drug susceptibility and emergence of drug resistance. A genome-scale protein-protein interaction network has been studied in a deadly bacterial pathogen to predict possible molecular pathways through which drug resistance could be triggered (Raman and Chandra 2008). Such knowledge lead to a new concept in drug discovery that targets could perhaps be prioritized and picked based on their potential to trigger drug resistance.

Systems approaches are being applied to study several infectious diseases. In the case of malaria, availability of *omics* data has led to systems studies of both the causative parasite *Plasmodium falciparum* as well as the mosquito vector it uses for transmitting malaria. Roles of molecular features corresponding to active growth, starvation, and environmental stress response of the parasite have been identified, as also those in the vector such as specific pattern recognition receptors important for its longevity and hence for the disease. These studies have led to several insights for the rational identification of drug targets for the design of novel antimalarial agents. Systems biology methods have been applied to understanding *Mycobacterium tuberculosis*, the pathogen responsible for the deadly disease of tuberculosis, through an integration of proteome, reactome, and interactome. A multilevel analysis involving study of genome-scale metabolic networks, protein-protein interaction networks integrated with the study of three-dimensional structures of proteins at a proteome scale has led to the prediction of a list of high-confidence drug targets in the pathogen (Raman et al. 2008). A systems approach to understanding host-pathogen interactions is being applied for viral infections as well, where models are developed to study the influence of virus infection on cellular signaling pathways, roles of specific viral and host genes in different steps of the viral life cycle, and disease initiation and the host response to viral infection.

Biology is increasingly tending to be a data-driven science, which serves as an excellent precursor for systems biology. It is almost automatic that these approaches will have a significant impact in the area of drug discovery. Predictions based on sound physical models and tacit rule-based models have long been used in product development and safety testing in various engineering disciplines, such as aerospace engineering and electronic circuit design. It can be envisaged that predictive approaches is expected to increase significantly in the coming years in drug discovery too. The extent of *omics*-scale data and the advances in systems technologies to enable comprehension of such large complex data in the form of meaningful biological models are promising to help in this process. The power of systems biology methods is such that it may become possible in the not-too distant future, that a disease could get diagnosed in a clinical setting and characterized at the systems level with precise genotype and phenotype definitions, leading all the way up to predictive quantitative titrations of the available remedies and finally personalized prescriptions.

References

Boran D, Iyengar R (2010) Systems approaches to polypharmacology and drug discovery. Curr Opin Drug Discov Dev 13(3):297–309

Butcher EC, Berg EL et al (2004) Systems biology in drug discovery. Nat Biotechnol 22(10):1253–1259

Forst CV (2006) Host-pathogen systems biology. Drug Discov Today 11:220–227

Glaab E, Baudot A et al (2012) EnrichNet: network-based gene set enrichment analysis. Bioinformatics 28(18): i451–i457

Hood L, Perlmutter RM (2004) The impact of systems approaches on biological problems in drug discovery. Nat Biotechnol 22(10):1215–1217

Ideker T, Krogan NJ (2012) Differential network biology. Mol Syst Biol 8:565

Kitano H (2002) Systems biology: a brief overview. Science 295(5560):1662–1664

Lee DS, Burd H et al (2009) Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple *Staphylococcus aureus* genomes identify novel antimicrobial drug targets. J Bacteriol 191(12):4015–4024

Orth JD, Thiele I et al (2010) What is flux balance analysis? Nat Biotechnol 28(3):245–248

Raman K, Chandra N (2008) *Mycobacterium tuberculosis* interactome analysis unravels potential pathways to drug resistance. BMC Microbiology 2008(2):109

Raman K, Yeturu K et al (2008) TargetTB: a target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. BMC Syst Biol 2:109

Rappuoli R, Aderem A (2011) A 2020 vision for vaccines against HIV, tuberculosis and malaria. Nature 473(7348):463–469

Systems Biology Graphical Notation

► SBGN

Systems Biology Markup Language (SBML)

Michael Hucka
Computing and Mathematical Sciences,
California Institute of Technology, Pasadena,
CA, USA

Synonyms

SBML

Definition

SBML (the Systems Biology Markup Language) is a representation format, based upon XML, used for communicating and storing computational models of biological processes (Hucka et al. 2003). SBML can represent many different classes of biological phenomena, including metabolic networks, cell-signaling pathways, regulatory networks, disease models, and many others. It does not attempt to be a universal language for quantitative models; rather, SBML's purpose is to serve as a franca lingua for exchanging the essential aspects of a computational model between software systems and databases. It is intended to be used by software and not written by humans directly, although its text-based nature makes it reasonably comprehensible for situations such as debugging and software development, when direct access is useful.

Characteristics

There are many aspects to SBML and its correct use. The following paragraphs can only summarize some notable points; elaborations and further details are available in the SBML specification documents (SBML Team 2011).

General Principles of SBML

SBML has three main purposes:

1. Enable modelers to use multiple software tools without having to rewrite models to conform to every tool's idiosyncratic file format
2. Enable models to be shared and published in a form that other researchers can use even when working with different software environments
3. Ensure the survival of models beyond the lifetime of the software used to create them

The most common and basic form of an SBML model consists of entities (called *species* in SBML) acted upon by processes (called *reactions* in SBML). An important principle is that models are decomposed into explicitly labeled constituent elements, the set of which resembles a verbose rendition of an explicit set of equations. This set of equations can represent both chemical reaction equations (if the model uses reactions) and equations derived from other concepts (again, if the model uses them). The SBML representation deliberately does not express a model directly as (for instance) a set of differential equations or other specific interpretation of the model. This explicit, framework-agnostic decomposition makes it easier for a software tool to interpret the model and translate the SBML form into whatever internal form the tool actually uses.

A software system can read an SBML model description and translate it into its own internal format for model analysis. For example, a software system might provide the ability to simulate the model by constructing differential equations and then perform numerical time integration on the equations to explore the model's dynamic behavior. Or, alternatively, a package might construct a discrete stochastic representation of the model and use a Monte Carlo simulation method such as the Gillespie algorithm.

Another important feature of SBML is that every entity can have machine-readable annotations attached to it. These annotations can be used to express relationships between the entities in a given model and entities in external resources such as online databases. A good example of the value of this is in BioModels Database (Li et al. 2010), where every model is annotated and linked to relevant data resources such as publications, databases of compounds and pathways, controlled vocabularies, and more. With annotations, a model becomes more than simply a rendition of a mathematical construct – it becomes a semantically enriched framework for communicating knowledge.

SBML is sometimes incorrectly assumed to be limited in scope only to biochemical network models because the original publications and early software focused on this domain. In reality, although the central features of SBML are indeed oriented toward representing "reaction-like" processes that act on some entities to generate new or different amounts of other entities, this same formalism serves analogously for many other types of processes; moreover, SBML also supports the direct expression of mathematical formulas and discontinuous events apart from reaction processes, allowing SBML to represent much more than only biochemical reactions.

Structure of SBML

SBML allows models of arbitrary complexity to be represented. Each type of component in a model is described using a specific type of SBML data structure that organizes the relevant information. The data structures determine how the resulting model is encoded in XML. The main data structures in SBML Level 3 Version 1 are the following:

- *Function definition*: A named mathematical function that may be used throughout the rest of a model.
- *Unit definition*: A named definition of a new unit of measurement. Named units can be used in the expression of quantities in a model.
- *Compartment*: A well-stirred container of finite size where species may be located. Compartments may or may not represent actual physical structures.
- *Species*: A pool of entities of the same kind located in a compartment and participating in reactions (processes). In biochemical network models, common examples of species include ions, proteins, and other molecules; however, in practice, an SBML species can be any kind of entity characterizable in terms of an amount.
- *Parameter*: A quantity with a symbolic name. In SBML, the term "parameter" is used in a generic sense to refer to named quantities regardless of whether they are constants or variables in a model. SBML provides the ability to define parameters that are global to a model as well as parameters that are local to a single reaction.
- *Initial Assignment*: A mathematical expression used to determine the initial conditions of a model. This type of object can only be used to define how the value of a variable can be calculated from other values and variables at the start of simulated time.

- *Rule*: A mathematical expression added to the set of equations constructed based on the reactions defined in a model. Rules can be used to define how a variable's value can be calculated from other variables, or used to define the rate of change of a variable. The set of rules in a model can be used with the reaction rate equations to determine the behavior of the model with respect to time.
- *Constraint*: A means of detecting out-of-bounds conditions during a dynamical simulation and optionally issuing diagnostic messages. Constraints are defined by an arbitrary mathematical expression computing a true/false value from model variables, parameters, and constants. An SBML constraint applies at all instants of simulated time; however, the set of constraints in the model should not be used to determine the behavior of the model with respect to time.
- *Reaction*: A statement describing some transformation, transport, or binding process that can change the amount of one or more species. For example, a reaction may describe how certain entities (reactants) are transformed into certain other entities (products). Reactions have associated kinetic rate expressions describing the speed at which reactions occur.
- *Event*: A statement describing an instantaneous, discontinuous change in one or more variables of any type (species, compartment, parameter, etc.) when a triggering condition is satisfied.

Mathematical formulas are represented using a subset of MathML. The SBML specification defines the MathML operators allowed in formulas in SBML, as well as how the identifiers of the various constructs like species and compartment objects are linked with MathML formulas.

Annotations

The constructs in SBML have only limited mathematical semantics. They have no predefined biological or biochemical semantics, and though a human could make inferences when inspecting a given model, software programs are not as competent in that regard. For software, the intended meaning of each model component needs to be made explicit and in a machine-readable form. SBML defines two separate systematic ways of adding annotations to any component of a model.

The first type of annotation takes the form of references to terms taken from the Systems Biology Ontology (SBO; Le Novère 2006). This set of controlled vocabularies provides terms for identifying such things as common reaction rate expressions, common participant types and roles in reactions, common parameter types and their roles in rate expressions, common modeling frameworks (e.g., "continuous," "discrete," etc.), and common types of biochemical species and reactions. By adding references to SBO terms to components of an SBML model, a software tool can provide additional details using independent, shared vocabularies that can enable other software tools to recognize precisely what the component is meant to represent. For example, if a given reaction in a model has an SBO attribute referencing term SBO:0000049 (which corresponds to "first-order irreversible mass-action kinetics, continuous framework" in SBO), then regardless of the identifier and name given to the reaction in the model, a software tool can look up the SBO term to inform users that the reaction is a first-order irreversible mass-action reaction.

The second type of annotation in SBML is more flexible and wider in scope. Its syntax consists of a structured subelement (the "annotation" subelement) that can be attached to any component in a model and can be used to define a relationship between the SBML component being annotated and the annotation content. The content can be either history information (date created, date modified, author, contact info, etc.) or references to external resources. The external resource can be anything – an entry in an online database, a publication, a part of another model, a term in an ontology, etc. The format of this kind of extended annotation in SBML follows the MIRIAM guidelines. Each annotation is a triplet consisting of (1) a data type, (2) an identifier, and (3) an optional qualifier. The data type is a unique controlled description of the type of the data in annotation and should be recorded as a Uniform Resource Name (URN); the identifier refers/points to a specific datum in whatever source is identified by the data type; and the qualifier serves to refine the nature of the relationship between the model component being annotated and the referred-to datum. Examples of common qualifiers include "is version of," "has part," etc.

SBML Evolution and Growth

The development of SBML is stratified in order to organize architectural changes and versioning. Major editions of SBML are termed *Levels* and represent substantial changes to the composition and structure of the language. Models defined in lower Levels of SBML can always be represented in higher Levels, though some translation may be necessary. The converse (from higher Level to lower Level) is sometimes also possible, though not guaranteed. The Levels remain distinct; a valid SBML Level 1 document is not a valid SBML Level 2 document. Minor revisions of SBML are termed *Versions* and constitute changes within a level to correct, adjust, and refine language features. Finally, specification documents inevitably require minor editorial changes as its users discover errors and ambiguities. Such problems are corrected in new *Releases* of a given SBML specification.

The latest generation of SBML, which is Level 3, is modular in the sense of having a defined core set of features and optional packages adding features on top of the core. This modular approach means that models can declare which feature-sets they use, and likewise, software tools can declare which packages they support. It also means that the development of SBML Level 3 can proceed in a modular fashion. The development process for Level 3 is designed around this concept.

SBML Level 3 package development is today an ongoing activity, with packages being created to extend SBML in many areas that its core functionality does not directly support. Examples include models whose species have structure and/or state variables, models with spatially nonhomogeneous compartments and spatially dependent processes, and models in which species and processes refer to qualitative entities and processes rather than quantitative ones.

SBML Development Process

SBML uses a community-oriented development approach. For example, technical decisions are made by a group of volunteer editors, with major decisions made as much as possible using electronic voting by the whole SBML community. Much of the development process is defined in a written document (made available on the SBML.org website) that provides guidelines for various aspects of the overall management of SBML development and the SBML community. The following are some of the features of the process:

- The SBML community is organized into the SBML Forum, the SBML Editors, and the SBML Team. The SBML Forum consists of all members of the community who subscribe to the sbml-discuss mailing list, with the list membership acting as a kind of basic voter registration mechanism. The SBML Editors are volunteers who are sufficiently interested in SBML and its continued development that they are willing to spend time in the development, writing, and correction of SBML specification documents. There are five SBML Editors at any given time; they are elected by a majority vote from among the SBML Forum, and they serve 3-year terms, with reelection being possible but consecutive terms being disallowed. The SBML Team are members who are employed to work on SBML-related activities. Their tasks include maintaining the resources that support the SBML community and SBML development in general; developing critical software such as libraries and online facilities; and organizing events and other similar activities.
- Discussions are held publicly as much as possible, usually on the sbml-discuss mailing list and in biannual face-to-face meetings. The public discussions and archives improve transparency, provide a public record of arguments and reasoning, and stimulate the broader community.
- Consensus is sought as much as possible. In situations where a decision appears to have no obvious right or wrong answer on technical grounds alone, the SBML Editors may initiate a public vote on the matter. These votes are typically conducted using an electronic voting system, with the topic and call-for-votes announced on the sbml-discuss@caltech.edu mailing list.

Strengths and Weaknesses of SBML

A few notable strengths of the SBML approach are (1) the use of explicit constructs for representing different facets of a model, (2) the relatively limited number of constructs, and (3) the community-driven development approach. A few notable weaknesses of SBML are (1) the seeming focus on biochemical reaction-based processes and (2) the introduction of syntactic differences between versions.

The use of explicit constructs means that the different aspects of a model in SBML are labeled and characterized explicitly. This facilitates consistent software interpretation of models, because the important features of a model are made explicit – an interpreter does not have to guess at the meaning behind, say, a set of equations (which ones stand for reactions? which ones are other relationships?), and various features such as function definitions are provided in a consistent (if limited) syntax. Moreover, this makes it more straightforward to take the same model and express it in any of a variety of different mathematical frameworks.

A second strength, the relatively limited number of constructs in SBML, means that it is less work for a software developer to implement tools for working with the format. SBML could have been designed with, for example, a deeper hierarchy of data types, but this was rejected purposefully to limit the complexity of implementations. (However, in fairness, this is not to say that SBML is very simple; there are still quite a few constructs and nuances in their meaning).

Finally, a third strength of SBML is, as mentioned above, the support and involvement of the community in its development and adoption. Specifications and technical decisions are made collectively by a small set of SBML Editors in collaboration with the whole SBML community, and electronic voting is used to reach community-wide consensus on important decisions.

Among the notable weaknesses, the first is SBML's seeming focus on reaction-based processes. The objects and terms in SBML (such as its "Species" and "Reaction" objects) are admittedly couched in biochemical reaction terms, reflecting SBML's origins and history. Many potential users assume SBML is limited to models of this type, but in reality, the same concepts can easily be used in other domains. In hindsight, it is now clear that more neutral terms could have been chosen.

A second weakness is the number of changes between versions within an SBML Level. The changes reflect hard-won experience by the SBML community and especially software developers, so from one perspective they are an understandable consequence of evolution and improvement. However, the changes make it admittedly more difficult for software developers to support all versions of SBML.

Relationships to Other Standardization Efforts

SBML has proven useful for software tools to exchange computational models. Still, by itself it does not provide a complete framework for reproducible modeling. Several related efforts now exist to standardize additional aspects of model exchange.

- *The Systems Biology Ontology* (*SBO*). As mentioned above, SBO provides a set of controlled vocabularies that can be used to annotate a model to make its mathematical semantics more precise (Le Novère 2006).
- *The Minimum Information Requested in the Annotation of biochemical Models* (*MIRIAM*). SBML defines a syntax for how to encode annotations in a MIRIAM-compliant way, but MIRIAM is itself a separate standardization effort applicable to any encoding format, not just SBML. It defines a basic and straightforward annotation scheme. It is also backed by a software resource, the *MIRIAM Services*, that provides a variety of tools for address resolution of references to resources on the Internet.
- *The Simulation Experiment Description Markup Language* (*SED-ML*). This XML-based format for encoding simulation experiments provides a tool-independent way of defining the model(s) to be used, the experimental task(s) to be run, and the result(s) to be produced (Waltemath et al. 2011).

Besides these, there also exist standardization activities for closely related topics. Many have separate sections in this encyclopedia: the *Systems Biology Graphical Notation* (SBGN; Le Novère et al. 2009), the *Biological Pathway Exchange* (BioPAX) *language* (Demir et al. 2010), *NeuroML*, and *CellML*.

Resources for SBML

For software developers who seek to implement support for SBML in their software, as well as modelers working with SBML files, there are many relevant resources available. The following paragraphs summarize some that may be especially useful.

Software

In addition to the hundreds of SBML-aware software systems for biological modeling and other purposes, two classes of important software resources for SBML are software libraries and validation tools.

Developers interested in supporting SBML in their software are encouraged to consider the use of the free, open-source software libraries libSBML (Bornstein et al. 2008) and JSBML (Dräger et al. 2011). LibSBML is written in ISO C and C++ and provides language interfaces for C, C#, C++, Java, MATLAB, Octave, Python, Perl, and Ruby. LibSBML is supported on Linux, Windows, and Mac OS X, and is distributed in both source-code form and as precompiled, ready-to-install libraries. Among its many features are support for all releases of SBML, unit checking and dimensional analysis, full validation of SBML, and APIs for working with mathematical formulas, annotations, and handling of compressed files. The JSBML library is a pure-Java implementation similar in its API to libSBML; at the time of this writing, it is relatively young and so does not offer as many features as libSBML, but may be more convenient to use for developers who cannot use the Java Native Interface employed by libSBML to provide its Java layer.

A free, online validation system is provided by SBML.org. Users can interact with it directly through a Web-based user interface or through a network programming interface. It provides the ability to upload a model and have it analyzed for conformance to the SBML specifications. The validation system cannot report whether the model is correct (i.e., the models' behavior may be wrong or pointless), but at least it can determine syntactic validity and consistency of the SBML file.

Documentation

The specification documents that define SBML are freely available from SBML.org. In addition, a list of Frequently Asked Questions and Answers is available, along with other help documents and code samples.

The libSBML library described above comes with extensive documentation for many of the supported programming languages. It also comes with sample programs to help developers get started.

Other Resources

There is an Internet MIME type defined for SBML, defined by RFC 3823 (Kovitz 2004).

SBML.org provides a web forum interface to the several SBML-related mailing lists. The list archives contain many years' worth of discussions about SBML, making this a helpful resource when first starting out programming with SBML.

Cross-References

- ▶ Biological Network Model
- ▶ Biological System Model
- ▶ CellML
- ▶ Collaborative and Distributed Biomedical Applications
- ▶ Data Integration
- ▶ Databases for Kinetic Models
- ▶ Information, Biological
- ▶ Interaction Networks
- ▶ Interoperability
- ▶ Knowledge Representation
- ▶ Mathematical Model, Model Theory
- ▶ Metabolic and Signaling Networks
- ▶ Metabolic Flux Analysis
- ▶ Metabolic Model
- ▶ Metabolic Networks
- ▶ Metabolic Networks, Databases
- ▶ Metabolic Networks, Structure and Dynamics
- ▶ Metabolic Pathway Analysis
- ▶ MIRIAM
- ▶ MIRIAM URI
- ▶ Model Repositories
- ▶ Pathway Modeling, Metabolic
- ▶ Pathway, Functional Units
- ▶ SBGN
- ▶ Semantic Web, Interoperability
- ▶ Specialized Metabolic Component Databases
- ▶ Web Service
- ▶ XML
- ▶ XML Schema

References

Bornstein BJ, Keating SM, Jouraku A, Hucka M (2008) LibSBML: an API library for SBML. Bioinformatics 24:880–8801

Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I et al (2010) The BioPAX community standard for pathway data sharing. Nat Biotechnol 28:935–942

Dräger A, Rodriguez N, Dumousseau M, Dörr A, Wrzodek C, Le Novère N, Zell A, Hucka M (2011) JSBML: a flexible Java library for working with SBML. Bioinformatics 27:2167–2178

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531

Kovitz B (2004) MIME media type for the systems biology markup language (SBML). RFC 3823, Accessed June, 2011 http://www.rfc-editor.org/rfc/rfc3823.txt

Le Novère N (2006) Model storage, exchange and integration. BMC Neurosci 7:S11–S11

Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A et al (2009) The systems biology graphical notation. Nat Biotechnol 27:735–741

Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V et al (2010) BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol 4:92

SBML Team (2011) SBML specification documents. Accessed June, 2011 http://sbml.org/Documents/Specifications

Waltemath D, Adams R, Beard DA, Bergmann FT, Bhalla US, Britten R et al (2011) Minimum information about a simulation experiment (MIASE). PLoS Comput Biol 7: e1001122

Systems Biology of Viral Pathogens

▶ Systems Virology

Systems Biology of Virus–Host Interactions

▶ Systems Virology

Systems Biology Ontology

Nick Juty and Nicolas le Novère
EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK

Synonyms

SBO

Definition

As the field of computational modeling flourished, it became clear that there was a need to provide a means of supplementing model data with additional information to clarify or specify the semantic content of computational models. Furthermore, this additional semantic information needed to be of a standard form to facilitate interoperability and exchange between different model-encoding formats. Orthogonal, structured controlled vocabularies, comprised of commonly used modeling terms and concepts, were created to meet the requirements of the SBML (▶ Systems Biology Markup Language (SBML)), community (Courtot et al. 2011).

The ontology currently consists of seven orthogonal vocabularies that cover the following: “participant role,” to describe component roles, such as “substrate”; “systems description parameter,” to describe various parameters and constants, such as the “Michaelis constant”; “mathematical expression,” to ascribe particular calculus associations between model parameters and variables, such as “mass action rate law”; “modeling framework,” to specify the approach or assumptions made in model creation, such as “logical framework”; “physical entity representation,” to define the type of the component within the model, such as “macromolecule”; “occurring entity representation,” to define processes that take place, such as “transport reaction,” and “metadata representation,” to specify the different types of metadata that may be incorporated within a model, such as “database cross reference.”

A mechanism to directly incorporate SBO terms within SBML models has been available since Level 2 Version 2, using the attribute “sboTerm,” and is described in the SBML specification. It is also directly linked to ▶ SBGN, where each glyph is associated with a specific SBO term, facilitating the conversion between SBML and the graphical SBGN representation.

Systems Biology Ontology is a member of the Open Biomedical Ontology effort (OBO; Smith et al. 2007), and hence committed to ontology development by a community-prescribed set of principles. The ontology can be browsed and downloaded in several formats at http://www.ebi.ac.uk/sbo, and is accessible programmatically via Web Services.

References

Courtot M, Juty N, Knüpfer C, Waltemath D, Zhukova A, Dräger A, Dumontier M, Finney A, Golebiewski M, Hastings J, Hoops S, Keating S, Kell DB, Kerrien S, Lawson J, Lister A, Lu J, Machne R, Mendes P, Pocock M, Rodriguez N, Villeger A, Wilkinson DJ, Wimalaratne S, Laibe C, Hucka M, Le Novère N (2011) Controlled vocabularies and semantics in systems biology. Mol Syst Biol 7:543

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, The OBI Consortium, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah N, Whetzel PL, Lewis S (2007) The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25:1251–1255

Systems Biology Pathway Exchange (SBPAX)

Oliver Ruebenacker
Center for Cell Analysis and Modeling, University of Connecticut Health Center, Farmington, CT, USA

Synonyms

SBPAX

Definition

Systems Biology Pathway Exchange (SBPAX) is an ontology and data format designed to store and organize systems biology knowledge data related to biological pathways, quantitative modeling, and the relationships between these two. Originally developed as an extension to BioPAX Level 2, it is now being developed as an extension to BioPAX Level 3 (the current BioPAX version) and proposed as an addition to the upcoming BioPAX Level 4. SBPAX enables broad interoperation between BioPAX and other systems biology modeling formats by creating hierarchies into which data, terms, and data objects from other formats can be inserted. While technically, SBPAX can incorporate any controlled vocabulary, some are better suited for this purpose than others. The preferred choice is the Systems Biology Ontology (SBO). Other suitable sources of terms would be the Systems Biology Markup Language (SBML), the Virtual Cell Markup Language (VCML), CellML, and MathML.

Characteristics

By serving as a common container for all these formats, SBPAX enables cross-format integration of biological data and knowledge. Integration is achieved by converting, mapping, or merging data across formats. In general, it is not possible to directly convert, map, or merge across formats, because there is no one-to-one correspondence between the elements of one format and the elements of the other format. But it is always possible to map one-to-one from a source format to SBPAX. Once mapped to SBPAX, the data can be reorganized and additional data be added as needed, and automatic derivations can be employed. Since the target format also maps one-to-one to SBPAX, it is clear what data needs to be added to export the data one-to-one to the target format, and as soon as that data has been added, the export can be performed. Without SBPAX, integrating two formats is a difficult process usually involving a set of complex rules that are formulated based on both formats and the possible relationships between the two. With SBPAX, integration consists of two components: the trivial one-to-one mappings between SBPAX and other formats, and the essential algorithms which now can be formulated entirely within SBPAX. SBPAX also serves as a container for all additional information added during the integration process, where it can be reused whenever data has been modified and needs to be reintegrated.

Current Status

As a response to the release of BioPAX Level 3 in 2009, a new version of SBPAX, called SBPAX3, is under development and has been heavily revised compared to the SBPAX version accompanying BioPAX Level 2: Features that were essentially adopted into BioPAX were removed from SBPAX and new features have been added reflecting new priorities.

The new focus of SBPAX3 is the inclusion of the kind of systems biology data provided by pathway databases that so far is not supported by BioPAX. For this purpose, SBPAX supports the addition of controlled vocabulary and quantitative values, and allows these to be arranged into hierarchies. The preferred choice for the controlled vocabulary is the Systems Biology Ontology (SBO), but other vocabularies can be used as well, such as the Systems Biology Markup Language (SBML), the Virtual Cell Markup Language (VCML), CellML, or MathML. If a database provides a type of data for which no established controlled vocabulary exists, it may make sense to provide their own terms.

The primary base class in SBPAX3 is the systems biology entity, which represents anything that can be characterized by one or more systems biology terms. A systems biology term is a term taken from a controlled vocabulary dedicated to systems biology, as explained above. A systems biology entity can have

Systems Biology Pathway Exchange (SBPAX), Table 1 Possible subentity relationships in SBPAX3, with examples. Column 1 shows a category for entities that can have a subentity, and column 2 shows what category that subentity would be. Column 3 shows an example for such an entity, and column 4 shows an example for the respective subentity

Entity type	Subentity type	Entity example	Subentity example
Material object	Component	Hemoglobin	Hb subunit, heme group
Process	Partial process	A → B → C	A → B
Object	Property	Conductor	Conductance
Entity	Mathematical description	Reaction	Rate law
Mathematical expression	Partial expression	Rate law	Rate law parameter
Indexable	Index	Michaelis constant for substrate ATP	ATP

any number of other systems biology entities as subentities, which represent a part, aspect, attribute, or specialization of an entity. Examples of subentities are listed in Table 1.

An overview over SBPAX3 classes and properties can be seen in Fig. 1. An important subclass of the systems biology entity is the BioPAX entity, and all its subclasses, such as pathways, interactions, physical entities, and genes. This allows any entity in BioPAX to be characterized by systems biology terms and subentities. For example, a BioPAX interaction can be described by the SBO term for redox reaction or can have as a subentity one of the many rate laws specified by SBO.

Another subclass of the systems biology entity is the systems biology measurable, which represents any measurable quantity relevant to systems biology, characterized by a number and a unit. The number can be any floating-point number and the unit is taken from Unit of Measurement Expressions (UOME), another proposed standard to specify units by controlled vocabulary, derivation from more basic units, and a growing list of currently hundreds of predefined units. SBO has large classes of quantitative systems description parameters, including conductance, dissociation constant, pressure, half-life, or Michaelis constant, so any of these quantities can be attached to any interaction, pathway, physical entity, or gene.

Since any systems biology entity can have subentities, hierarchies can be built. A simple application is this: An interaction can have, as a subentity, one of the many rate laws included in SBO, and the rate law (e.g., Michaelis-Menten rate law) can have subentities the appropriate parameters (e.g., specific activity and Michaelis constant), which are also included in SBO.

Finally, systems biology entity has a subclass systems biology state, which represents the state of a system, and can be used to group entities that belong to the same system state. For example, quantities that are measured together, such as an equilibrium constant and the temperature, would best be described as subentities to a common state.

Figure 2 shows how a quantitative value (the dissociation constant), a systems biology measurable with number and unit, is added to a systems biology entity, which in this case is a BioPAX complex formation. The meaning of the quantity is described by the appropriate SBO term (SBO 0000282 for dissociation constant). The complex formation can also be described by an SBO term.

Figure 3 is slightly more involves example. Here, two quantities are characterized by the same SBO term (conductance), but indexed by two different subentities (calcium and sodium), representing the differing conductance of an ion channel for two different ions.

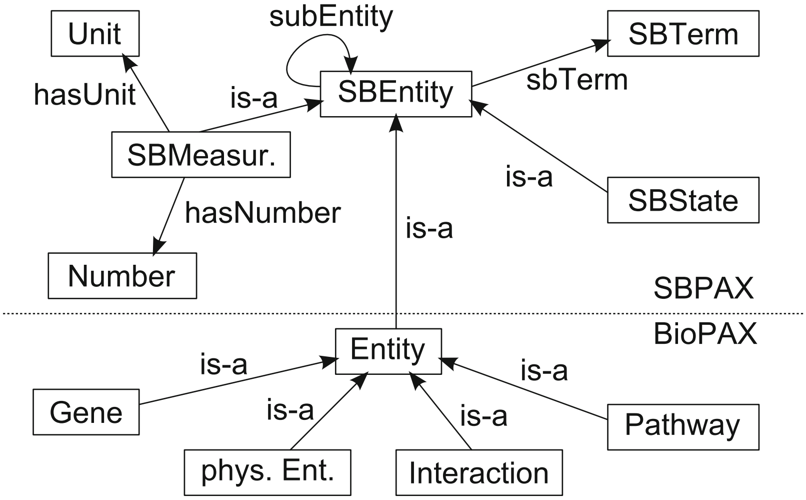
Figure 4 shows a practical example for model building: A catalysis process is an entity, the kinetics (Michaelis-Menten with two substrates) is its subentity, and the parameters are subentities of the kinetics.

Implementation

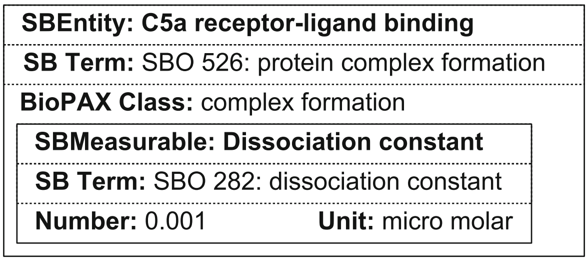
SBPAX3 with examples and documentation is available as for download from the SBPAX website (www.sbpax.org). Signaling Gateway Molecule Pages has implemented preliminary SBPAX export. MetaCyc and EcoCyc have declared to implement SBPAX export within months. VCell plans to implement within the coming weeks or months import of SBPAX3 to help build and annotate models, including by providing kinetic parameters. A validator is also being developed.

Background

When SBPAX was originally introduced in 2008, the latest version of BioPAX was Level 2. To integrate BioPAX Level 2 or earlier with SBML, it was



Systems Biology Pathway Exchange (SBPAX), Fig. 1 An overview over classes (*boxes*) and properties (*arrows*) of SBPAX3. Here, "is-a" means "is subclass of." The dashed line separates SBPAX3 classes form BioPAX classes



SBEntity: C5a receptor-ligand binding

SB Term: SBO 526: protein complex formation

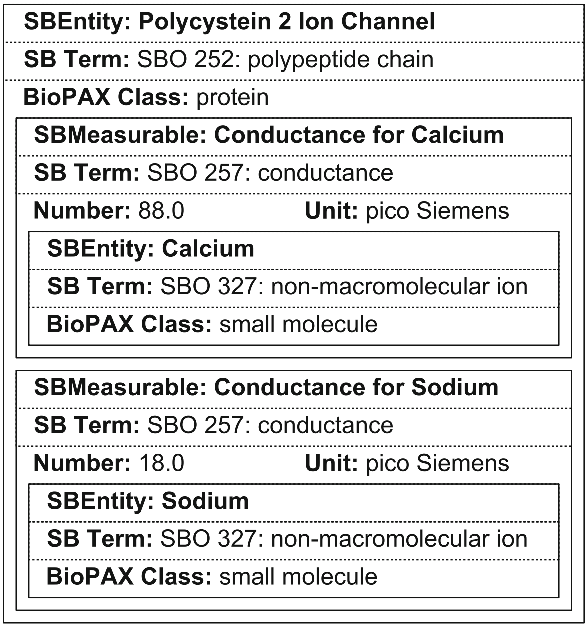
BioPAX Class: complex formation

SBMeasurable: Dissociation constant

SB Term: SBO 282: dissociation constant

Number: 0.001 **Unit:** micro molar

Systems Biology Pathway Exchange (SBPAX), Fig. 2 A simple example of SBPAX3 usage to add a quantity to an entity. A solid *box* refers to a systems biology entity. A *box* nested inside another represents a subentity



SBEntity: Polycystein 2 Ion Channel

SB Term: SBO 252: polypeptide chain

BioPAX Class: protein

SBMeasurable: Conductance for Calcium

SB Term: SBO 257: conductance

Number: 88.0 **Unit:** pico Siemens

SBEntity: Calcium

SB Term: SBO 327: non-macromolecular ion

BioPAX Class: small molecule

SBMeasurable: Conductance for Sodium

SB Term: SBO 257: conductance

Number: 18.0 **Unit:** pico Siemens

SBEntity: Sodium

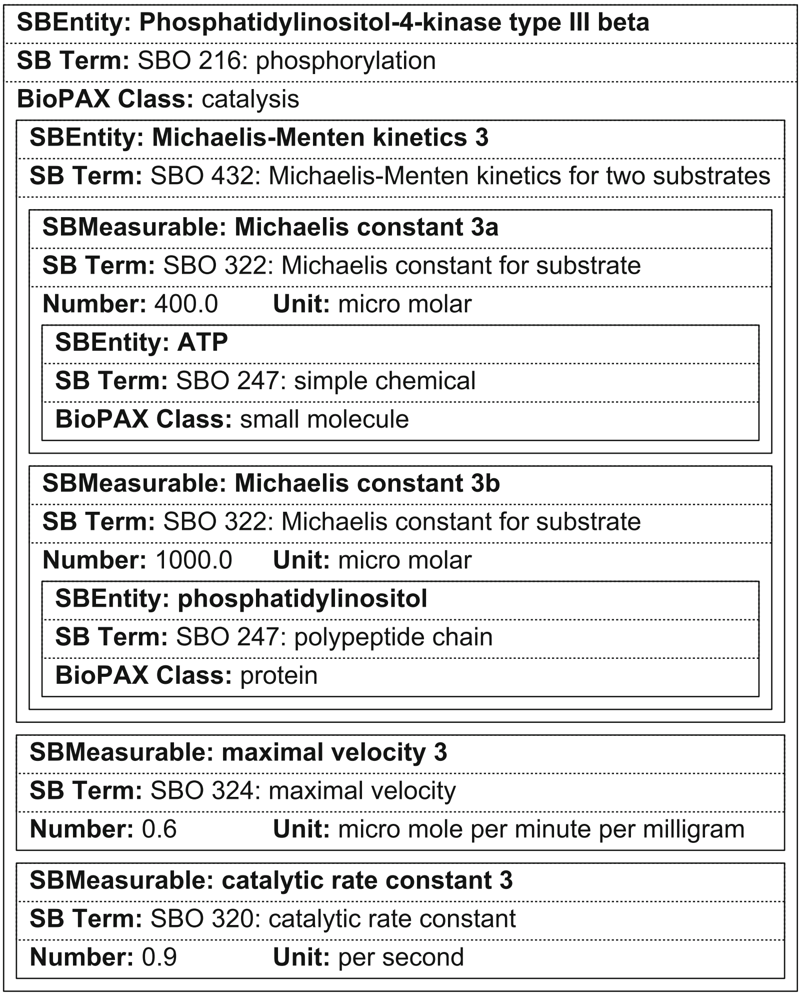
SB Term: SBO 327: non-macromolecular ion

BioPAX Class: small molecule

Systems Biology Pathway Exchange (SBPAX), Fig. 3 An ion channel has a conductance that depends on the type of ion (e.g., calcium versus sodium)

necessary to extract from BioPAX the information needed to construct species in SBML, which are defined as pools of indistinguishable entities at a specific location, and which are a central concept in SBML and quantitative modeling in general. The problem is that BioPAX has no term for species. Instead, BioPAX uses the term "physical entities" as collectives regardless of locations and in the case of sequence-based molecules such as proteins, DNA, and RNA, regardless of chemical modifications that leave the sequence intact. Locations and modifications were specified individually for each interaction in which a physical entity participated, and the difficulty was that in BioPAX Level 2, sequence features were used to represent modifications, but not every sequence feature was a modification. It was further not clear, whether the absence of a sequence feature signified that a modification was absent, or merely that it was not relevant for the interaction.

SBPAX facilitated BioPAX-SBML integration by providing both the term "substance," which corresponds to the BioPAX physical entity, and the term "species," which corresponds to the same term in SBML and is


SBEntity: Phosphatidylinositol-4-kinase type III beta
SB Term: SBO 216: phosphorylation
BioPAX Class: catalysis
SBEntity: Michaelis-Menten kinetics 3
SB Term: SBO 432: Michaelis-Menten kinetics for two substrates
SBMeasurable: Michaelis constant 3a
SB Term: SBO 322: Michaelis constant for substrate
Number: 400.0 Unit: micro molar
SBEntity: ATP
SB Term: SBO 247: simple chemical
BioPAX Class: small molecule
SBMeasurable: Michaelis constant 3b
SB Term: SBO 322: Michaelis constant for substrate
Number: 1000.0 Unit: micro molar
SBEntity: phosphatidylinositol
SB Term: SBO 247: polypeptide chain
BioPAX Class: protein
SBMeasurable: maximal velocity 3
SB Term: SBO 324: maximal velocity
Number: 0.6 Unit: micro mole per minute per milligram
SBMeasurable: catalytic rate constant 3
SB Term: SBO 320: catalytic rate constant
Number: 0.9 Unit: per second

Systems Biology Pathway Exchange (SBPAX), Fig. 4 Michaelis-Menten kinetics with two substrates

a substance with a location. This allows to specify all relevant relationships between a set of BioPAX data and the corresponding set of SBML data.

A typical work flow would be as follows: BioPAX data (or those part of it that are related to quantitative modeling) is converted to SBPAX data. Then, additional data is gradually acquired and added to the SBPAX data. Finally, the extended SBPAX data is converted one-to-one to SBML. There is no one-to-one relationship between the BioPAX and the SBML elements, but some BioPAX objects have a one-to-one relationship with SBPAX objects which have relationships to SBPAX objects that have a one-to-one relationship with some SBML elements. Should the BioPAX input data be modified or extended, some of the data that was added to the SBPAX can still be used to convert again from BioPAX to SBML.

SBPAX was implemented as the native format of the Systems Biology Linker (SyBiL), an application that imported BioPAX and exported SBML (and allowed to save intermediate results as SBPAX). This functionality was then integrated into the Virtual Cell.

In SBPAX, a substance can be a subset of another substance, and the absence of a feature can be specified, as well as basic set operations. This means we can easily specify a substance with a feature, a substance without that feature, and the union set of both, where the presence of the feature is optional. The intersection of a substance with one feature and the same substance but with another feature is to be understood to be that substance with both features.

In 2009, BioPAX Level 3 was released which incorporated several improvements previously introduced by SBPAX: Physical entities were redefined to correspond to what SBML and SBPAX call species; newly introduced physical entity references corresponds to a restricted version of SBPAX's substance; sequence features were clearly divided into modification features and other features; the absence of a feature can now be explicitly expressed, as can be subset relationships among physical entities, as well as among physical entity references.

Cross-References

- ▶ CellML
- ▶ RDFS
- ▶ Systems Biology Markup Language (SBML)
- ▶ Systems Biology Ontology
- ▶ Virtual Cell (VCell) Modeling and Analysis Platform
- ▶ Web Ontology Language (OWL)

References

Blinov ML, Ruebenacker O (2009) Integrating BioPAX pathway knowledge with SBML models. IET Syst Biol 3(5):317–328

Blinov ML, Ruebenacker O, Moraru II (2008) Complexity and modularity of intracellular networks – a systematic approach for modeling and simulation. IET Syst Biol 2(5):363–368

Blinov ML, Ruebenacker O, Schaff JC, Moraru II (2010) Modeling without borders: creating and annotating VCell models using the web. Lect Notes Comput Sci, Vol 6053

Ruebenacker O, Moraru II, Schaff JC, Blinov ML (2007) Kinetic modeling using BioPAX ontology. Proceedings of the IEEE international conference on bioinformatics and biomedicine, pp 339–348

Systems Biology Resources

Eduardo Mendoza
Department of Computer Science, University of the Philippines Diliman, Quezon City, Philippines

Introduction

Since the advent of high-throughput technologies toward the end of the twentieth century, the life sciences are experiencing decades of "data deluge" and "information explosion." Respected researchers in fact speak of "data-driven science" as a new era of science, citing the life sciences as the newest example (Hey 2010). This scenario highlights the need for appropriate resources and tools to help life scientists extract the knowledge useful for their research.

This need is perhaps even greater in the field of Systems Biology: it not only displays the typical dynamics of an emerging field but also an unusually broad multidisciplinary character largely due to its highly integrative approach. The field may be viewed currently as a highly dynamic "network of disciplines" (at least seven according to the Institute of Systems Biology in Seattle) and simultaneously tightly coupling computational and experimental efforts (a novelty in most of the life sciences!). These factors together with the field's high impact potential –as expressed in the concept of "Systems Medicine" (EU Workshop Report 2010) – explain why even after more than a decade since its renaissance in 2000, there is still a lot of variation in the defining the field's scope (Mendoza 2009). A section "Resources" has hence to balance covering sufficient ground and at the same time be selective to be useful for the audience of the Encyclopedia.

Section Structure

This section focuses primarily on databases as resources for Systems Biology and discusses other forms such as portals, wikis, and specific community-oriented information systems only briefly in this overview. It should be noted however that some of the systems considered have a hybrid character and include digital library features.

Hierarchical Network as Paradigm

Though some authors of Systems Biology textbooks restrict the concept of Systems Biology resources to databases directly needed for modeling (e.g., Klipp et al. 2009), this section is structured according to the predominant network paradigm for computationally representing biological systems. The majority of the entries (around 80%) are dedicated to resources for the major molecular components (genes, proteins, metabolites) and the interactions between them in terms of important subnetworks (gene regulatory, signaling, and metabolic). In one entry, the transition from the "molecular" to the "systemic" via the concept of modularity is discussed both in physical terms (organelles) and computational (functional modules) terms. In the remaining entries, a few important aspects of higher levels of organization are addressed in terms of sample topics. In particular, in view of the very broad scope of resources for "organs" and "organ systems" (the whole realm of physiology), only pointers to the work of the IUPS Physiome Project and related activities are provided later in this overview.

Genome and Transcriptome Resources

The first five entries deal with resources for the study of DNA, genes, and genomes. Overviews of the resources at the three most important sites (NCBI, EBI, and DDBJ) are provided to orient users, in particular the beginners, in these rich collections. The following entry on viral genome resources at NCBI provides an example of the use of these resources. The next three entries deal with resources for three related topics of intensive current research: single nucleotide polymorphisms (SNIPs), coupling of genomic and environmental data (metagenomics), and important genome resource tools. A glimpse into the world of transcriptomes is provided by the entry on resources for non-coding RNAs.

Proteome Resources

Three entries describe important resources for proteins and proteomes. The first entry discusses in detail several proteome databases, while the next two handle information on parts of protein structure (domains) and related functions on the one hand, and then groups of proteins (protein complexes) and relation functions on the other. Such (physical) protein complexes find their computational analogues in functional modules (discussed in the entry on ▶ Organelle and Functional Module Resources).

Metabolome Resources

After an overview entry on metabolism resources, an entry on special metabolic components is provided. These include resources on organism-specific metabolism, information on thermodynamics and kinetics and enzyme–ligand interactions. The third entry discusses resources for natural products, which are metabolites produced by living organisms. The importance of these substances is highlighted by the fact that, to this day, natural products and their derivatives are the basis of the majority of available drugs (Li and Vederas 2009).

Resources on Pathways and Networks

Three entries on pathways and networks follow, each corresponding to an important subnetwork in the cell. The overview entry on ▶ transcriptional and epigenetic regulation resources provides insights into genetic and epigenetic regulatory networks. The entry on ▶ signaling network resources is then complemented with a description of databases on metabolic networks which are general (in particular, not organism specific). The final entry of this series addresses resources on functional subunits of cells, both physical (organelles) and computational (functional modules). Resources on organelles need to explicitly consider spatial aspects, for example, location, which are very important particularly for proteins. Functional modules are often represented as subnetworks and, in this sense, are partly considered in resources for networks. On the other hand, specific resources for such functional subnetworks are just beginning to be established and discussed in the entry.

Model Organism and Disease Resources

A comprehensive discussion of resources of higher levels of organisms and their interactions with environmental factors is well beyond the scope of this section. However, overviews of information on model organisms and diseases are useful as examples for the kind of information available. Entries on information concerned with two important diseases, cancer and metabolic diseases, are also provided to give a flavor of such resources.

Model Repositories

The final entry describes the repositories available for models of biological systems. Such models cover small functional units, like network motifs, which corresponds to basic "circuits") over mid-sized functional modules to large subnetworks over several levels of hierarchy (an example would be the human circadian system which consists of a "master clock" in the brain, responding to light input and interacting with clocks in various organs in different parts of the body via a multitude of signaling mechanisms. In this sense, the models in such resources select and attempt to integrate the data and information from the previous resources to answer biological questions posed by the researchers.

Further Systems Biology Resources

Cross References to Other Encyclopedia Sections

Various sections discuss topics closely related to specific entries in this section in more detail. These include, along with their respective entry(s), the following:

- ► Transcriptional and Epigenetic Regulation
  Gene regulatory networks: modeling, reconstruction, and analysis
  Systems biology in epigenetics and post-translational regulation
- Metabolism Resources, General Metabolic Network Resources, ► Specialized Metabolic Component Databases
  Metabolic networks
- ► Metabolic Diseases Resources
  Systems biology of diabetes and beta cells
- ► Natural Products Resources
  Systems Biology applications in drug discovery
- ► Model Repositories
  Systems Biology model databases

The following sections are directly relevant for the discussion of resources in general:

Standards, guidelines, and infrastructure
► Text mining in systems biology
► Ontologies and controlled vocabularies

Note also that tools relevant to each section are discussed within that section.

Special Journals, Issues, and Books

One of the best references for Systems Biology resources is the Database Issue of the journal *Nucleic Acids Research* (*NAR*), annually published in January. It contains short descriptions of new databases related to molecular biology as well as updates on important systems like those from MCBI, EBI, and DDBJ. For example, the 2011 edition has descriptions of 96 new online databases and updates on 83 previously introduced systems (Galperin and Cochrane 2011). In 2009, as an additional service, the Molecular Biology Database Collection (described in more detail in the following subsection) was introduced and now includes 1,300 databases. Due to the high demand for publication of information on such resources, NAR also introduced a new open access journal *DATABASE: The Journal of Biological Databases and Curation* (http://database.oxfordjournals.org/).

Specialized journals of course constitute another important form of Systems Biology resource and the growing number with "Systems Biology" or "Biosystems" in their title reflect the emergence of the field since 2000 as the paradigm of life sciences research in the early twenty-first century. Among these journals are *Molecular Systems Biology, BMC Systems Biology, IET Systems Biology (formerly IEE Proceedings Systems Biology), Systems and Synthetic Biology, Biosystems, and Molecular Biosystems.* Also given the growing trend to integrative approaches (including data integration), many journals on Computational Biology and Bioinformatics include a large number of papers on Systems Biology. The best known examples are *PloS Computational Biology*, *Bioinformatics* (with a special section on "► Systems Biology" in each issue), and *BMC Bioinformatics.*

In the same period, a bonanza of books on Systems Biology have been published and most of these have chapters on resources. As an example, the textbook of Klipp et al. 2009 discusses three kinds of databases: pathway databases, databases of kinetic data, and model databases. However, the book entitled *Bioinformatics and Systems Biology: Collaborative Research and Resources* by Frederick Markus (2010) is particularly valuable since it has specifically pulled together information on Bioinformatics Systems Biology research networks and communities and resources they have generated, particularly in the context of European Union research programs.

Collection of Databases

The NAR Molecular Biology Database Collection (http://www.oxfordjournals.org/nar/database/a/) stands out as a source of information for Systems Biology databases because of its structured format, breadth, and strong community support, resulting in a well-maintained up-to-date resource. Entry information consists of a NAR Collection ID, the system's URL, the email address of a contact person, a one-paragraph summary of the database's features, one or more NAR categories to which the system belongs (http://www.oxfordjournals.org/nar/database/c/) and a link to the most current NAR Database Issue description of the system. The list of currently 1,300 entries can be viewed alphabetically or by NAR category.

Further lists include the "Reactome Resource Guide," which can be found on the Reactome database's Wiki (http://wiki.reactome.org/index.php/Reactome_Resource_Guide) and Wikipedia's bio databases list (http://en.wikipedia.org/wiki/List_of_biological_databases). Both lists have only around 100 entries, the former provides only one-sentence descriptions and the system's URL while the latter categorizes the entries and provides their names and links.

Portals

The differentiation between a portal and community-oriented systems discussed in the following subsection is based on the audience addressed – a Systems Biology portal strives to address the wide audience of persons with general interest in the field and the majority of its information is available to all who visit the website. A good example of this approach is the Systems Biology Institute's Portal for Systems Biology http://www.systems-biology.org/ initiated by one of the field's pioneers, Hiroake Kitano (2002). A community-oriented information system on the other hand typically is more focused on specific topics and provides resources related to these topics. Such an information system also typically has more sophisticated access and security features. However, the difference is not black and white since, although most of the resources are accessible only to the specific community, such systems also have portal-like features for "outreach" such as the public resources available at the Institute of Systems Biology in Seattle (http://www.systemsbiology.org/Public_Resources), founded also by another of the field's pioneers, Leroy Hood. The websites of various Systems Biology research centers and groups hence offer resources like educational material, software, and even data sets. In general, the quality of resources available from the latter is better, both due to stronger focus and higher demands on the site's maintenance.

In addition to particular institutions, several journals have initiated portals for Systems Biology. Examples of these are *BioMed Central*'s Systems Biology Gateway http://www.biomedcentral.com/gateways/systemsbiology and *Nature*'s Systems Biology portal: http://www.nature.com/sysbio/index.html.

Several region-oriented portals have been established by regional "communities of interest," a recent example is the "Council for Systems Biology in Boston" (http://www.csb2.org/), which "builds local, regional, and national links between academic and industrial laboratories active in the areas of systems and computational biology." Again the information quality is quite dependent on the community's level of activity: while the "San Diego Consortium for Systems Biology," founded in 2005, transformed into the "San Diego Center for Systems Biology" (http://sdcsb.org/), an NIH-funded center and offering a unique resource called "CircadianServer" (http://circadian.salk.edu/), others like the Munich Systems Biology Forum (http://www.msbf.mpg.de/) has practically ceased to exist and the information on its portal is quite outdated. Portals maintained by individuals or small groups generally are not well maintained.

http://www.biochemweb.org/systems.shtml

Specific Community-oriented Information Systems

A recent development in Systems Biology is the establishment of networks of Systems Biology research centers by national funding authorities. In most cases, the mandate of these networks of research centers includes an "outreach" component and leads to very informative portals: a good example is given by the portal of the National Centers for Systems Biology http://www.systemscenters.org/, with resources covering databases, software tools, and via *Biositemaps* available experimental technology. Similar networks have been established in Switzerland (SystemsX, http://www.systemsx.ch/), Germany (FORSyS Centers, http://www.forsys.net/joomla/index.php?lang=en), and the UK (BBSRC Systems Biology Centres, http://www.bbsrc.ac.uk/organisation/institutes/systems-biology-centres.aspx).

The information systems of international research networks are also a good source of specific resources in Systems Biology. Interesting examples are provided by:

- Current Physiome projects, in particular the IUPS Physiome Project (http://www.physiome.org.nz/)

and the Virtual Physiological Human (http://www.vph-noe.eu/), where resources regarding organs and organ systems (particularly on the computational side) are bundled.

- Clusters formed from related projects of European Union Framework Programs, such as the NanoSafety Cluster, which addresses important technological and social questions raised by growing use of nanoparticles in daily life and biomedical applications ("Nanomedicine") and the resulting Bionanointeractions. An important specific resource is, for example, the comprehensive compendium of European nanosafety projects under http://www.nanosafetycluster.eu/home/european-nanosafety-cluster-compendium.html.
- EUCLIS (EUCLOCK Information System,), which has evolved from an information for chronobiology researchers in a large 5-year project to an information infrastructure for the worldwide community of chronobiologists, as evidenced by the decision of the two largest professional societies, the Society for Research on Biological Rhythms (SRBR) and the European Biological Rhythms Society (EBRS) collaborating to fund the maintenance of the system beyond the termination of the EUCLOCK project. A unique resource available in EUCLIS is ChronoCollections, which contain scanned papers and research notes of early pioneers of the field which are otherwise not easily accessible.

Related Resources and Perspectives

The broadly integrative approach of Systems Biology has led to its tremendously growing scope and dooms any attempt at comprehensiveness, particularly in describing the resources available. The entries in this section cover only some important aspects and focus on the most important form: online databases. In this overview, examples of other forms of resources such as portals and community information systems (which are also partly digital libraries) were briefly discussed.

An important perspective is the importance of establishing community-based standards regarding the "core structure" of biological databases. Such standards are urgently needed to facilitate the development and ease the maintenance and management of the growing number of systems worldwide. While this can only succeed through a joint effort of scientists and partners from industry (who will provide better platforms), an initial effort to specify such a "core structure" has been started: in Gaudet et al. (2011), a proposal for "BioDBCore" is sketched and a working group encompassing representatives from a wide range of existing life sciences resources, but remains open to all interested parties concurring its goals which include maximizing the consistency and interoperability of resources, the promotion of adoption of syntactic and semantic standards, and provision of guidance for users in evaluating the scope and relevance of a esource. An initial list of attributes for a BioDBCore checklist is available (Table 1 in Gaudet et al. 2011).

References

EU FP7 (2010) Report on the workshop "from systems biology to systems medicine", Brussels

Galperin MY, Cochrane GR (2011) The 2011 *nucleic acids research* database issue and the online molecular biology database collection. Nucleic Acids Research 19(Database Issue):D1–D6

Gaudet P, Bairoch A, Field D, Sansone SA, Taylor C et al (2011) Towards BioDBCore: a community-defined information specification for biological databases. Nucleic Acids Research 19(Database Issue):D7–D10

Hey T (2010) Data-driven scientific computing. In: Proceedings of the 10th international conference on systems biology, Edinburgh, 10–15 Oct 2010

Kitano H (2002) Systems biology: a brief overview. Science 295:1663–1664

Klipp E, Liebermeister W, Wierling C, Kowald A, Lehrach H, Herwig R (2009) Systems biology. A textbook. Wiley, Weinhei

Li JWH, Vederas JC (2009) Drug discovery and natural products: end of an Era or an endless frontier? Science 325(10):161–165

Markus F (2010) Bioinformatics and systems biology: collaborative research and resources. Springer, Berlin

Mendoza ER (2009) Systems biology: its past, present and potential. Phil Sci Lett 2(1):16–34. http://www.philsciletters.org/

Systems Biomedicine

▶ Systems Pharmacology

Systems Histopathology

▶ Systems Pathology

Systems Immunology

Sudipto Saha
School of Medicine, Center for Proteomics and
Bioinformatics, Case Western Reserve University,
Cleveland, OH, USA

Introduction

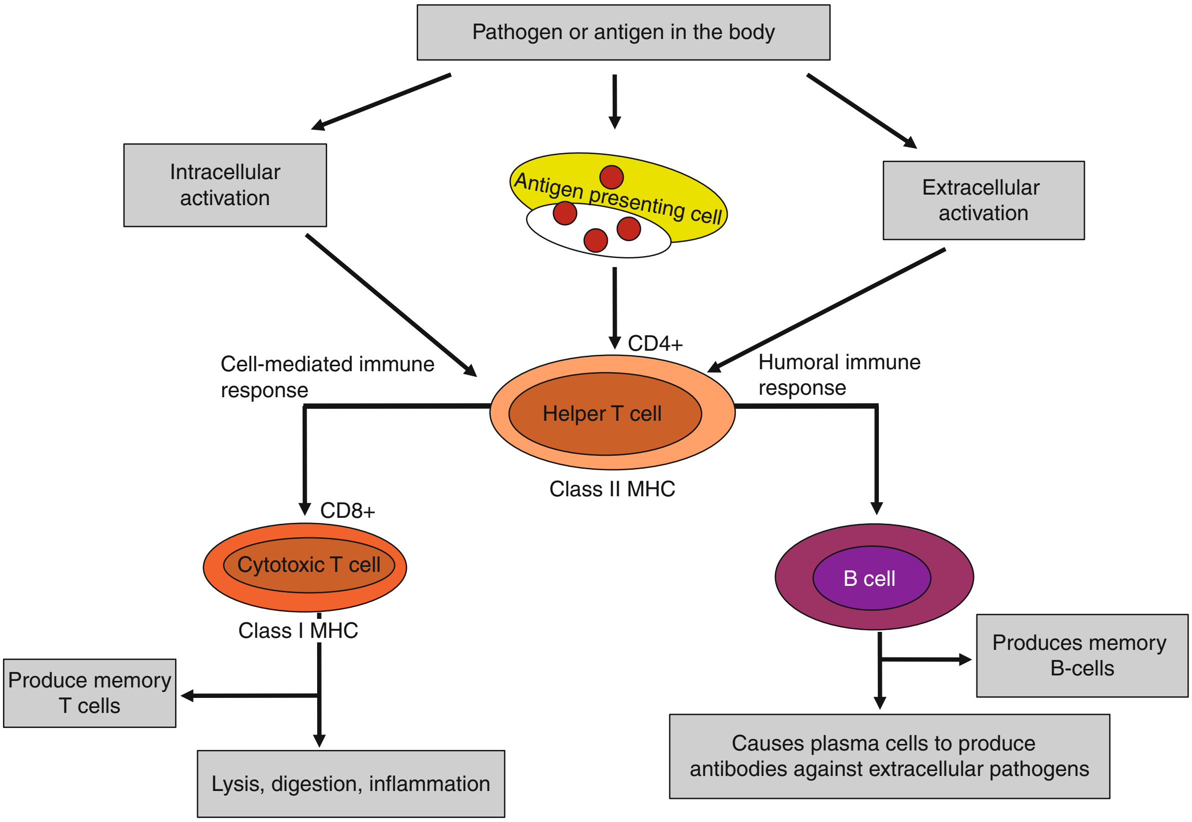
Systems immunology is a novel means for studying, analyzing, and understanding complex immune systems using a systems approach. This is an interdisciplinary approach that uses high-throughput technologies and computational methods that can be applied to identify a global map of complex interactions between cell–cell, cell–environment, protein–protein, and protein–DNA interactions. Currently, DNA microarrays, next generation sequencing, and modern mass spectrometry are used to define and monitor all components of the immune system. The overall goal of this approach is to generate a hypothesis, identify new biological rules, and predict the behavior of biological systems. Traditional approaches to studying immune regulation is primarily based on reductionist approaches of molecular biology. They offer a limited view of the complex immune system since there are so many different types of host cells and genes perturbed by the entry of a pathogen. To make things more complex, the pathogens occasionally use alternate virulence factors and manipulate the host's systems to kill the host cells or survive inside the host cells for prolonged intervals of time. Systems approaches have been used to study the role of innate and adaptive immune systems, host-pathogen interactions, and lymphocyte dynamics after stimulation by a pathogen, and development of drugs or vaccines. In this introductory chapter, brief overviews of the host immune system, pathogens, and systems approaches used in immunology are presented, concluding with challenges and caveats of using systems approaches.

Systems Immunology, Table 1 Innate and adaptive immunity

	Innate	Adaptive
Evolution		
Time scale	Ancient	Relatively new
Phylum	Plants, invertebrates, and vertebrates	Vertebrates
Characteristics		
Specificity	For structures shared by group of related microbes termed pathogen-associated molecular patterns (PAMPs)	For antigens of microbes and for nonmicrobial antigens
Diversity	Limited; germline-encoded	Very large; receptors are produced by somatic recombination of gene segments
Action time	Immediate activation of effectors	Delayed activation of effectors
Memory	None	Yes
Nonreactivity to self	Yes	Yes
Components		
Physical and chemical barriers	Skin, mucosal epithelia; antimicrobial chemicals and peptides	Lymphocytes in epithelia; antibodies secreted at epithelial surfaces
Blood proteins	Complement	Antibodies
Cells	Phagocytes (macrophages, neutrophils), natural killer cells	Lymphocytes (B- and T-cells)

Immune System

The immune system is the body's defense mechanism and protects it against disease. In vertebrates, it consists of two different types: innate and adaptive immunity. Table 1 describes the characteristics and components of innate and adaptive immunity. The innate immune response is the first defense barrier against invading pathogens and relies on a limited number of germ line–encoded receptors that recognize pathogen-associated molecular patterns (PAMPs) of microbial pathogens, but not the host. Recognition of these molecular structures allows the immune system to distinguish infectious nonself from noninfectious self. Toll-like receptors (TLRs) present in phagocytes (macrophages, dendritic cells), and natural killer cells play a major role in pathogen recognition. There are 10 TLRs in humans, 13 in mice, and 222 in sea urchins, which are evolved to recognize PAMPs from fungi, bacteria, viruses, and parasites. Stimulation of toll-like receptors by PAMPs leads to the activation of



Systems Immunology, Fig. 1 *The adaptive immune response*. There are two branches of the adaptive immune response: humoral immunity and cell-mediated immunity. Generally, for extracellular and intracellular pathogens, humoral immunity, and cell-mediated immune responses protect the host, respectively

signaling pathways that result in the induction of antimicrobial genes, inflammatory cytokines, and maturation of dendritic cells to induce costimulatory molecules and increased antigen-presenting capacity. Innate immunity helps to direct adaptive immune responses to eliminate the encountered pathogens and establish long-lasting protective immunity against them (Janeway and Medzhitov 2002).

The principal features of adaptive immunity are specificity and generation of immunologic memory. There are two branches of the adaptive immune response: humoral immunity and cell-mediated immunity (Bonilla and Oettgen 2010). Humoral immunity involves the transformation of B cells into plasma cells that produce and secrete antibodies to a specific antigen. Antibodies (or immunoglobulin, Ig) are large Y-shaped proteins used by the immune system to identify and neutralize foreign objects. In mammals, there are five types of antibodies: IgA, IgD, IgE, IgG, and IgM, differing in biological properties. Each has evolved to handle different kinds of antigens. Cell-mediated immunity is dependent upon T lymphocytes which are sensitized by first exposure to a specific antigen. Subsequent exposure stimulates the release of a group of substances known as lymphokines, such as interferon, and interleukins as well as direct killing by cytotoxic T lymphocytes (Fig. 1).

Pathogens

A pathogen is a biological agent such as a virus, bacteria, fungi, and protist that cause disease to its host. Some notable pathogenic viruses are Human rhinoviruses (HRVs), Human T-lymphotropic virus type-1 (HTLV-1), and Human Immunodeficiency virus (HIV). HRVs are thought to be the cause of more than half of all acute upper respiratory tract infections (common cold) worldwide. HTLV-1 is the etiological agent of an aggressive leukemia, called adult T-cell leukemia/lymphoma (ATL), and inflammatory disorders, including arthritis and dermatitis. Transmission of HTLV-1 occurs through transfer of infected cells from mother to child during breast-feeding, via sexual intercourse, and through exposure of infected blood products or sharing of needles and syringes. Measurement of lymphocyte dynamics in HTLV-1 infected subjects shows that T-cell proliferation is increased compared to controls (Boxus and Willems 2009). HIV is a lentivirus (a member of the retrovirus family) that causes acquired immune deficiency syndrome (AIDS), a condition in humans in which the immune system begins to fail, leading to life-threatening opportunistic infections. Transmission of HIV is similar to the HTLV-1 virus, such as, breast milk, unprotected sex, and contaminated needles. An advanced phase of this disease is marked with depletion of CD4+ T-cells that leads to acquired immune deficiency syndrome. Highly active antiretroviral therapy (HAART) has greatly reduced HIV plasma viremia, which results in increased CD4+ T-cell counts (Moir et al. 2010). The most common bacterial disease is tuberculosis, caused by the bacterium *Mycobacterium tuberculosis* $H_{37}Rv$. Its success fully relies on its ability to utilize macrophages for its replication (Meena and Rajni 2010). There is a wide diversity of pathogenic bacteria species, and there is even an enormous diversity of virulence genes in strains of the same species (Rosenberg 2005). Pathogenic bacteria contribute to other globally important diseases, such as pneumonia, food borne illnesses, and infections such as tetanus, typhoid fever, diphtheria, and syphilis. Bacteria can often be killed by antibiotics. Fungi are common problems in the immunocompetent population as the causative agents of skin, nail, or yeast infections. Some eukaryotic organisms, such as protists and helminths, cause disease. One of the epidemic diseases caused by protists in the genus Plasmodium is malaria.

Systems Approaches in Immunology

The systems approach is an imperative means to study complex immune systems, where interaction occurs between different cell types, and there is variation within a cell due to the large number of receptors on the cell membrane. Broadly speaking, systems immunology is focused on immunoinformatics, antigen processing and presentation, host-pathogen interactions, modeling of the immune system, and vaccinomics.

Immunoinformatics

The field of immunology generated huge amount of data, comprising of pathogen's antigen or epitopes, host antibodies, and T-cell receptors and are manually curated and maintained in databases. The Marie-Paule Lefranc group maintains *IMGT®, the international ImMunoGeneTics information system®* http://www.imgt.org, created in 1989, which consists of web *resources of sequence and structure* databases and *interactive tools*. This is an integrated knowledge resource specialized in immunoglobulins (IG) or antibodies, T-cell receptors (TR), major histocompatibility complex (MHC) of human, and other vertebrate species, and in the immunoglobulin superfamily (IgSF), MHC superfamily (MhcSF) and related proteins of the immune system (RPI) of vertebrates and invertebrates (Ehrenmann et al. 2010). The Raghava group from the institute of microbial technology, India, maintains an immunoinformatics database of pathogen antigens, B- and T-cell epitopes and MHC binders available at http://www.imtech.res.in/raghava/antigendb (Ansari et al. 2010). The National Institute of Allergy and Infectious Diseases maintains the immune epitope database and analysis resource (IEDB) available at http://www.immuneepitope.org/, which contains data related to antibody and T-cell epitopes for humans, nonhuman primates, rodents, and other animal species (Vita et al. 2010). The B-cell epitopes and available MHC binders curated data were used to develop prediction methods using machine learning techniques, such as artificial neural networks and support vector machine. There are online resources for prediction of B-cell epitopes (Saha and Raghava 2006), and MHC class I and II binders (Bhasin and Raghava 2004; 2006).

Antigen Processing and Presentation

Antigen-presenting cells are comprised of dendritic cells (DCs) and macrophages, they internalize bacteria, fungi, and parasites sensed by toll-like receptors (TLRs) into phagosomes, where proteolytic processing of microbial antigens (Ags) produces antigenic peptides that are subsequently presented by class I major histocompatibility complex (MHC I) to CD8+ T-cells and class II MHC (MHC II) molecules to CD4+ T-cells. The regulation of phagosomal internalization and processing seems to be controlled by TLRs in bacteria. During the process of phagocytosis, the host delivers antimicrobial properties that include the generation of toxic reactive oxygen species and antimicrobial peptides. Antimicrobial peptides (also called host defense peptides) and toll-like receptors are evolutionarily conserved components of the innate immune response and are found in both vertebrates and invertebrates. Thus, antigen processing and presentation contribute to both the innate host defense and adaptive immune response to microbes by MHC I and II restricted to T-cells (Ramachandra et al. 2009).

Host–Pathogen Interactions

Host–pathogen interactions (HPIs) are a multilevel problem ranging from molecular interactions to cell–cell communications. HPIs may occur at five different levels, e.g., at the entry level where host receptors like toll receptors bind to PAMPs, at the virulence level, at the nutrition level, at the critical host-cell pathway level, and at the immune evasion level. This interspecies interaction plays a crucial role in initiating infection in a host-pathogen system. The HPIs may induce programmed host-cell death pathways which involve a delicate balance between the host's defensive responses and a pathogen's virulence mechanism. Microbial pathogens have evolved many strategies to modulate the host system and facilitate replication and proliferation inside host cells (Lamkanfi and Dixit 2010). Current computational methods to study HPIs focus on topological connections based on network biology which uses the graph structure of protein–protein, protein–DNA, and protein–small chemical interaction data. There are few simple mathematical and computational network models developed that study the process of signaling through immune system receptors. Knowledge of the host-pathogen system enables accelerated drug development such as successful antibiotics against chronic *Chlamydia* infection (Forst 2006).

Modeling of the Immune System

The immune response to pathogen infection stimulates lymphocytes such as B- and T-cells and other factors like inflammatory cytokines, and forms a complex system that necessitates the use of systems approaches for its analysis. Experimental methods to study lymphocyte population dynamics after stimulation by a pathogen are mainly based on lymphocyte labeling and quantifying cell division as function of time, rates of accumulation, rates of proliferation, or rates of replacement of labeled cells. These methods are tedious and costly, whereas computational modeling requires less time and cost, aids in experiments that are not possible in a laboratory setting, and finally generates novel insights and hypotheses for further research. Another important aspect of computational methods is once the models are constructed and validated, they can be perturbed in different ways which enables the exploration of many possibilities. There are numerous mathematical approaches, including nonlinear dynamics, differential equations, and agent-based modeling (ABM). ABM is a mathematical approach used in simulating immune systems in discrete time and space and uses logical rules learned by experimental outcomes. ABM is useful in characterizing properties of immune systems and is well suited for addressing key challenges in immunology (Chavali et al. 2008).

Vaccinomics

Vaccinomics deals with vaccine analysis, which includes *in silico* epitope vaccine design and novel evaluation of vaccine efficacy. Traditionally, the killed/inactivated vaccine (e.g., cholera) and live/attenuated vaccine (e.g., oral polio vaccine) are successfully used in many individuals; however, there are a few limitations. The components of inactivated whole organism may be toxic and are responsible for side effects, whereas there is a potential risk of virulence in attenuated vaccine. Modern vaccine includes subunit vaccines (e.g., using the surface antigen of the hepatitis B virus) and epitope vaccines, where T-cell and B-cell epitopes of the antigen are used. The identification of potential vaccine candidate antigens and their epitopes is a challenging job. Experimental approaches include epitope mapping which are expensive and rigorous. Computational methods can help the development of *in silico* epitope vaccines by narrowing down the potential antigen and identification of B-cell and T-cell epitopes. However, a major challenge is the evaluation of vaccine efficacy in a short time period. The current clinical trial format is lengthy (four clinical phase trials) and, thus, it involves many resources and other factors. Blood samples provide a snapshot of the immune status, and gene or protein expression signatures were used to identify biomarkers of vaccine efficacy from clinical phase I volunteers. A systems approach was used to identify early gene signatures that predicted immune responses in humans vaccinated with the yellow fever vaccine (Querec et al. 2009). The three most common factors that correlate protection are the magnitude of the antigen-specific antibody titers, the functional signature of the T-cell response, and the functional signature of vaccine-induced innate immunity (Pulendran et al. 2010). Another area that could benefit from a systems approach is the identification of new adjuvants. The adjuvants are molecules used to improve the vaccine efficacy. The functional signatures of the innate immunity (e.g., TLRs) may be used to screen adjuvants. A great variety of adjuvants are available; however, alum described by Glenny in 1926 was only globally licensed for human use.

Challenges and Pitfalls of Systems Approaches

Despite the promise of the systems approach, there are still challenges in the real understanding of complex immune systems. In immunoinformatics, prediction methods of MHC binders are based on machine learning and fail to predict with high accuracy to unknown experimental or blind datasets (Gowthaman and Agrewala 2008). When modeling the immune system, it is often difficult to locate key model parameters, including rate constants, in published literature. In these cases, parameters need to be estimated, either through additional laboratory experiments or theoretical approaches. In addition, theoretical models raise the challenge of experimental validation, and it is risky to rely on these models for vaccine or drug development decisions (Forst 2006). In vaccinomics, researchers have recently been using high-throughput gene expression microarray data which provide a global picture of the biological response to a vaccine. There are cases of genes that are statistically differentially expressed, yet are of no consequence to the biological response to the activation because there are redundancies of gene functions during evolution. Still, there is a challenge to identify true casual relationships between genes differentially expressed and the stimulated immune response. Recently, emphasis has been given to pathway and network analyses because they use prior knowledge into data analysis. In the future, we have to get beyond colorful heat maps and graphical protein–protein interaction networks to an understanding of the biological significance of the molecular signatures or biomarkers of the vaccine discovery process (Pulendran et al. 2010). Inspite of these short comings, systems immunology promises to offer a new paradigm in vaccine discovery. Still, there is a paramount interest in the rational design of future vaccines against HIV, malaria, and tuberculosis.

Systems Immunology, Table 2 Systems approaches in immunology

Systems immunology approaches	Topic	References
Immunoinformatics	Databases of antigens, B-cell and T-cell epitopes, MHC I and II binders	Ansari et al. 2010; Ehrenmann et al. 2010; Vita et al. 2010
Antigen processing and presentation	The effect of proteosome on shaping T-cell epitopes	Ramachandra et al. 2009
Host–pathogen interactions	Gene regulation and signal transduction of host cells by pathogens	Lamkanfi and Dixit 2010
Modeling of the immune system	Mathematical modeling of lymphocytes against infection	Chavali et al. 2008
Vaccinomics	Vaccinomics deals with vaccine analysis	Pulendran et al. 2010

Summary

The immune system is the body's defense against pathogens and other invaders. The pathogen can be any biological agent such as a virus, bacteria, fungi, or protist that causes disease to its host. The first body defense after stimulation by a pathogen comes from the innate immune system, which in turn activates the adaptive response. Primary features of adaptive immunity are specificity and generation of immunologic memory. There are two branches of the adaptive immune response: humoral immunity and cell-mediated immunity. Humoral immunity is activated by extracellular pathogens and intracellular pathogens activate the cell-mediated immune response in order to protect the host. The pathogen manipulates the host immune system to promote infection. Thus, there is a need to implement systems approaches which comprises of high-throughput experimental and computational methods for better understanding of the immune system. Systems approaches in immunology are briefly summarized in Table 2. These approaches allow us to better understand the complex communications between the host cell–cell and cell–environment which can further be used to develop potential drugs or vaccines.

Cross-References

- ▶ Adaptive Immunity
- ▶ Adjuvants
- ▶ Epitope
- ▶ Immunoinformatics
- ▶ Innate Immunity
- ▶ Major Histocompatibility Complex (MHC)
- ▶ Vaccinomics

References

Ansari HR, Flower DR, Raghava GP (2010) AntigenDB: an immunoinformatics database of pathogen antigens. Nucleic Acids Res 38:D847–D853

Bhasin M, Raghava GP (2004) SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. Bioinformatics 20:421–423

Bhasin M, Raghava GPS (2006) A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. J Biosci 32:31–42

Bonilla FA, Oettgen HC (2010) Adaptive immunity. J Allergy Clin Immunol 125(2 Suppl 2):S33–S40

Boxus M, Willems L (2009) Mechanisms of HTLV-1 persistence and transformation. Br J Cancer 101:1497–1501

Chavali AK, Gianchandani EP, Tung KS, Lawrence MB, Peirce SM, Papin JA (2008) Characterizing emergent properties of immunological systems with multi-cellular rule-based computational modeling. Trends Immunol 29:589–599

Ehrenmann F, Kaas Q, Lefranc MP (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC. Nucleic Acids Res 38:D301–D307

Forst CV (2006) Host-pathogen systems biology. Drug Discov Today 11:220–227

Gowthaman U, Agrewala JN (2008) In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion. J Proteome Res 7:154–163

Janeway CA Jr, Medzhitov R (2002) Innate immune recognition. Annu Rev Immunol 20:197–216

Lamkanfi M, Dixit VM (2010) Manipulation of host cell death pathways during microbial infections. Cell Host Microbe 8:44–54

Meena LS, Rajni (2010) Survival mechanisms of pathogenic Mycobacterium tuberculosis H37Rv. FEBS J 277:2416–2427

Moir S, Chun TW, Fauci AS (2010) Pathogenic mechanisms of HIV disease. Annu Rev Pathol 6:223–248, PMID: 21034222

Pulendran B, Li S, Nakaya HI (2010) Systems vaccinology. Immunity 33:516–529

Querec TD, Akondy RS, Lee EK, Cao W, Nakaya HI, Teuwen D, Pirani A, Gernert K, Deng J, Marzolf B, Kennedy K, Wu H, Bennouna S, Oluoch H, Miller J, Vencio RZ, Mulligan M, Aderem A, Ahmed R, Pulendran B (2009) Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol 10:116–125

Ramachandra L, Simmons D, Harding CV (2009) MHC molecules and microbial antigen processing in phagosomes. Curr Opin Immunol 21:98–104

Rosenberg E (2005) The diversity of bacterial pathogenicity mechanisms. Genome Biol 6:320

Saha S, Raghava GP (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 65(1):40–48

Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B (2010) The immune epitope database 2.0. Nucleic Acids Res 38:D854–D862

Systems Immunology, Adaptive Immune Response to HIV Infection

Elizabeth Yohannes[1] and Mark R. Chance[2]

[1]Center for Proteomics and Bioinformatics, School of Medicine, Case Western Reserve University, Cleveland, OH, USA

[2]Center for Proteomics and Bioinformatics and Department of Genetics, School of Medicine, Case Western Reserve University, Cleveland, OH, USA

Synonyms

Host adaptive immune response to HIV infection; Systems network in HIV

Definition

Systems immunology, under the umbrella of system biology, provides integrated approaches at different level of complexity to define the underlying mechanisms of adaptive immune responses to HIV infection. With the advent of high-throughput data acquisition platforms coupled with advances in computational biology, systems immunology approaches to HIV infection investigate molecular signatures of adaptive immunity in response to HIV infection, evaluate and establish pathways and/or networks linking biomolecules to perturbations related to HIV infection, and generate testable hypotheses. Thus, it offers better platforms to define networks of molecular modules that are unique to HIV infection, activation, replication, and pathogenesis. Elucidation of these molecular modules should eventually lead to defining better treatment options to HIV infection.

Characteristics

Adaptive Immune Response

Our immune system has evolved over millions of years and includes highly complex innate and more specific adaptive immune responses. While both responses are unique there is a strong interplay between the two. They are part of a single, coordinated response of the host mounted against microbial infection or any other foreign biological mater in the host body. In fact, the defining feature of an adaptive immune response is its specific, inducible reaction to pathogens. It focuses and drives the effectiveness of innate immunity mechanisms to the levels not present in lower organisms (Kindt et al. 2007).

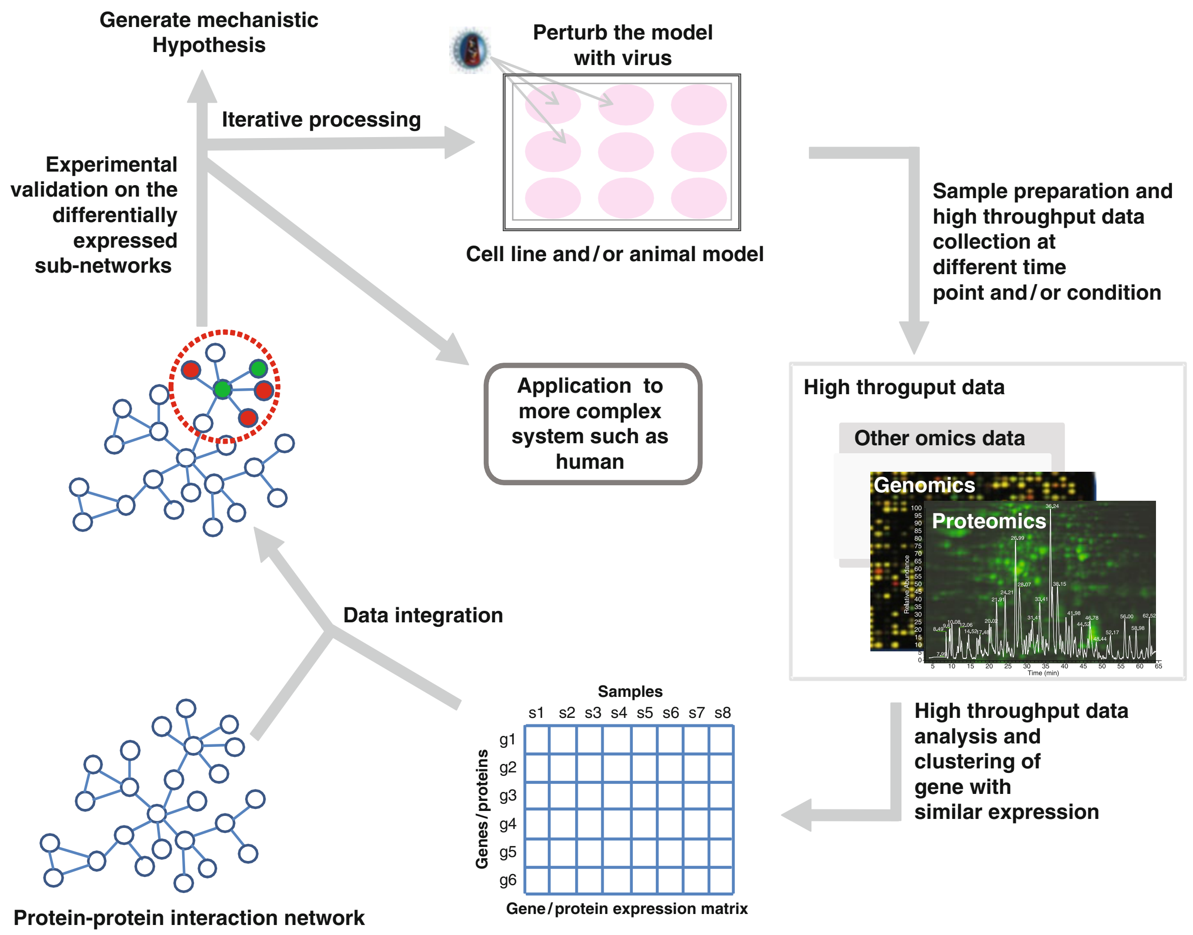
Human Immunodeficiency Virus (HIV)

HIV is a member of the retrovirus family which causes acquired immune deficiency syndrome (AIDS). It is a persistent pathogen and evolves its own means to survive and replicate in the host's hostile environment. HIV-1 not only manages to escape the first line of host defense mechanisms (innate responses) but is also able to hijack the adaptive immune defense mechanisms by infecting specifically the very cells (CD4 and CCR5 and/or CXR4 receptor expressing cells) necessary to activate both B-cell and cytotoxic T-cell immune responses, sequestering itself in privileged cells including long-lived memory T-cells and encoding and expressing proteins that restrict, redirect, or modify various protective functions of the host immune system (Nathanson et al. 2007). In addition, viral sequence diversification during the spreading infection allows the virus to escape or tolerate adaptive immune responses.

Enormous progresses have been made over the past decades defining the individual components of HIV-1 and the host cell at the molecular level. This progress has led to developing highly active antiretroviral therapies (HAART) that have sharply altered the course of HIV infection and progression to AIDS allowing patients to live longer and with greatly improved quality of life. Thus far, chemotherapy has been the most successful intervention for HIV-1. However, we still have no cure from HIV and/or no HIV-1 preventative vaccines has shown efficacy, in spite of considerable efforts. As a result, HIV, a generation after its identification, still has major health and socioeconomic impacts, particularly in developing countries. There is no clear path from the discovery and characterization of the molecular components of HIV-1 to the development of vaccines. Part of the solution lies in understanding the complex interactions between HIV-1 and the host in a systems immunology context.

Approaches to HIV-1 Infection

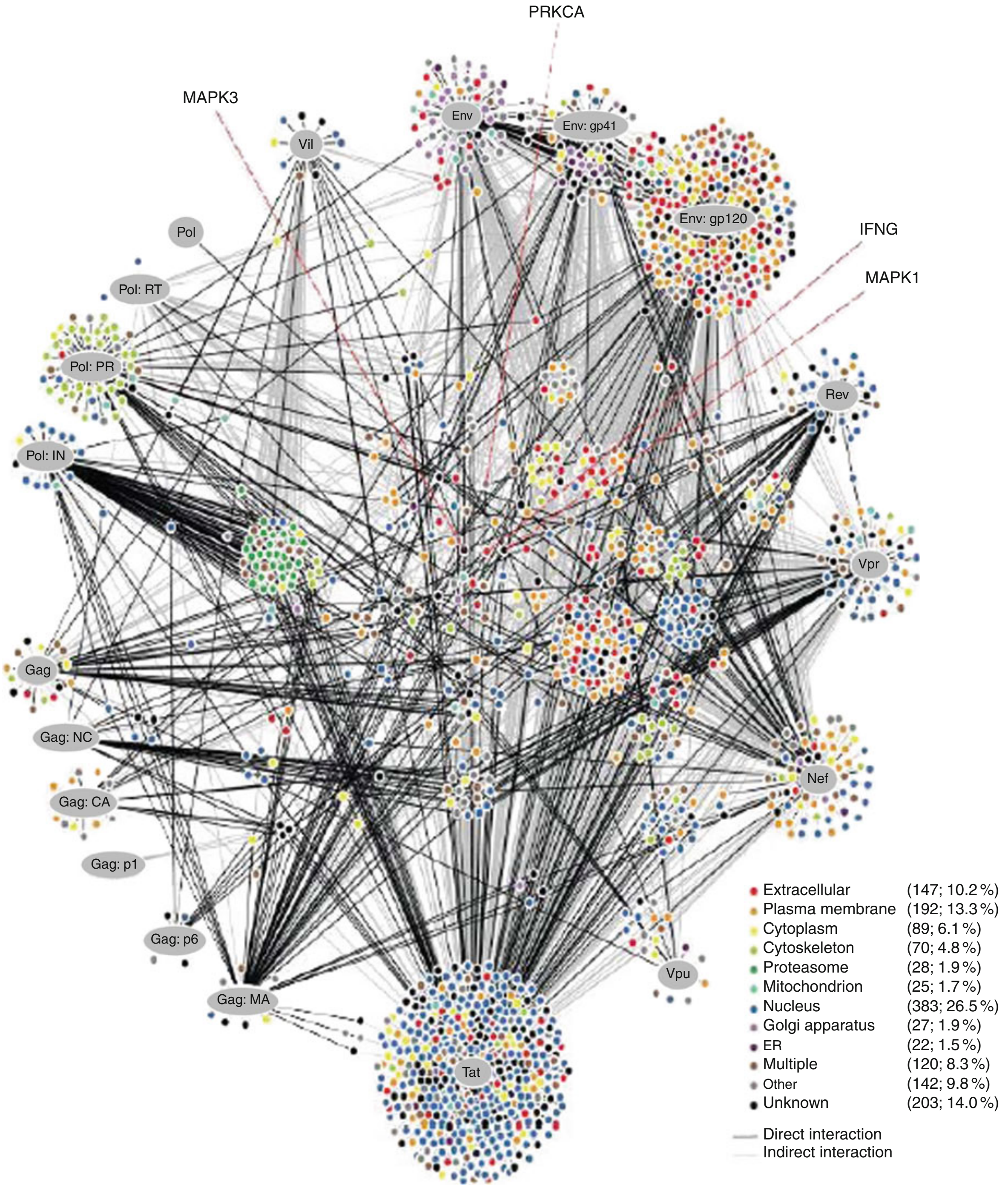
Advances in immunological sciences have mainly been achieved through analysis of individual biological components at any given time. There is certainly great value to this approach, for instance we now have rich and detailed data about the function for all the 17 proteins that are expressed and processed by HIV-1 and understand many of their interactions with host cell proteins. However, this approach may promote an oversimplified view with a focus on single factors without explicit regard for the network of elements within which each individual component carries out its function. Recently, however, as new data types and technologies have become available through high-throughput techniques such as gene expression microarrays, deep sequencing, and mass spectrometry to characterize genes, transcripts, and proteins, the concept of a system-level analysis, which aims to relate all of these individually determined patterns to each other has become feasible. A conceptual representation of a system immunology approach, as shown in Fig. 1, can be examined through iterative cycles of perturbing model systems at different levels of complexity, measuring the response with high-throughput techniques, analyzing the results with multidimensional data mining, and (re)developing a predictive model, with evaluation and refinements


Generate mechanistic Hypothesis
Perturb the model with virus
Iterative processing
Experimental validation on the differentially expressed sub-networks
Cell line and / or animal model
Sample preparation and high throughput data collection at different time point and / or condition
Application to more complex system such as human
High througput data
Other omics data
Genomics
Proteomics
Data integration
Samples
s1 s2 s3 s4 s5 s6 s7 s8
g1
g2
g3
g4
g5
g6
Genes / proteins
Gene / protein expression matrix
High throughput data analysis and clustering of gene with similar expression
Protein-protein interaction network

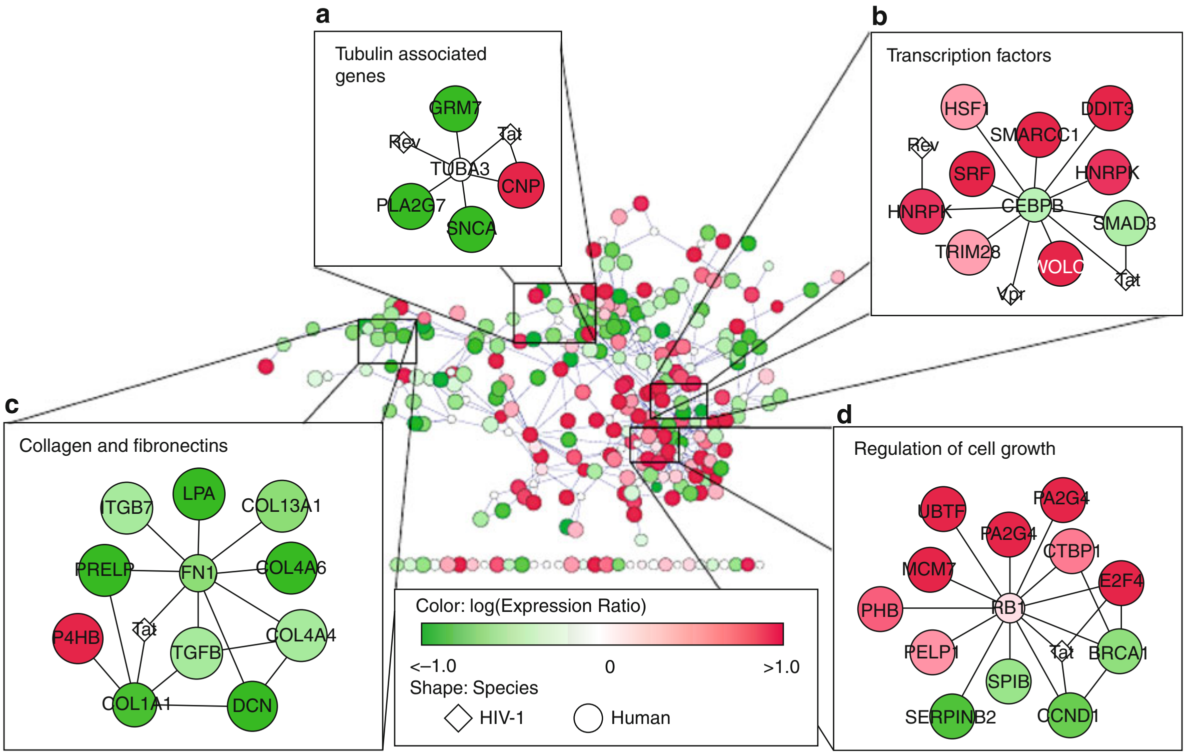
Systems Immunology, Adaptive Immune Response to HIV Infection, Fig. 1 Schematic overview of high-throughput data comprising a systems model of host response to viral infection form the view of protein-interaction networks as a means to understand disease biology. This figure is adapted from (Tan et al. 2007) with some modifications

to the model. It is apparent that each data type from any given high-throughput technology may reveal only one layer of the system. However, as these data sets have global scope, when combined and critically evaluated they provide new perspectives on the organization, complexity, functionality, and dynamics of biological entities.

The framework of Fig. 1 illustrates a system-wide analysis of quantitative and/or qualitative changes in biomolecules including transcripts, proteins, lipids, and metabolites followed by clustering and establishing networks linking the biomolecules to perturbations related to HIV infection and its pathogenesis. For instance a number of microarray and/or proteomics studies of CD4 T-cells following HIV-1 infection or exposure to HIV-1 accessory proteins have illustrated a very broad perturbation in host gene and/or protein expression profiles. Clustering the gene, or protein expression, according to similarity of its expression profile is often followed by annotation using functional classification tools such as The Database for Annotation, Visualization, and Integrated Discovery (DAVID), expression analysis systematic Explorer (EASE), or Ingenuity pathways analysis (IPA) to identify overrepresented gene families and pathways that are involved in viral replication and/or CD4 T cell apoptosis. With this high-throughput approach, it has been demonstrated that HIV-1 infection up-regulates transcriptional factors that are shown to support HIV-1 long terminal repeat (LTR) transcription while at the same time down-regulating expression of negative regulators of



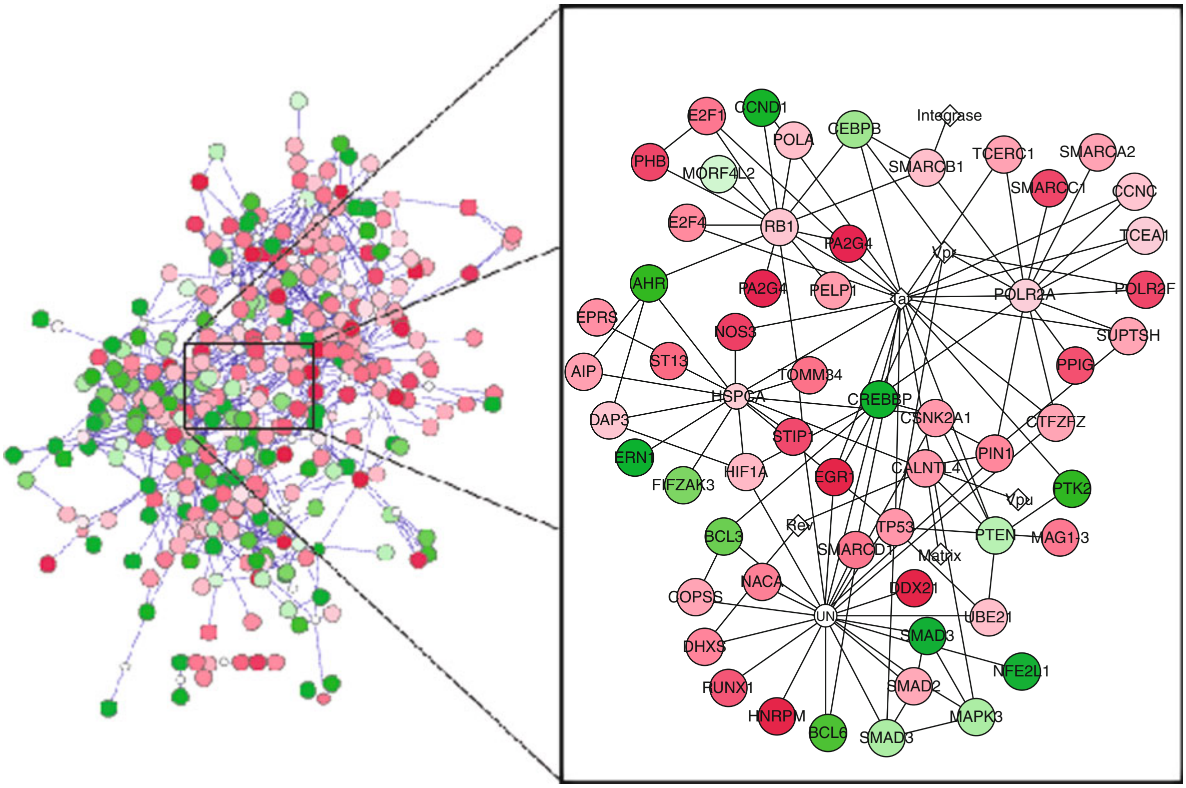
Systems Immunology, Adaptive Immune Response to HIV Infection, Fig. 2 A visual representation of HIV-1 human interaction network. On this network the nodes and edges represent HIV-1 (gray oval) or human (colored circles) proteins, and HIV-1–human protein interaction, respectively. Colors correspond to human protein categories based on cellular component gene ontology (GO) term. The number of proteins in each category and percentage of the total 1,448 human proteins interacting with HIV-1 are indicated in parentheses. Black edges for direct and gray for indirect interaction. This figure is adapted (from Ptak et al. 2008) with permission



Systems Immunology, Adaptive Immune Response to HIV Infection, Fig. 3 Active network of gene expression during HIV latency. This active network is further broken down into top five subnetworks. The color of each node corresponds to the mean change in gene expression (red corresponds to up-regulation, green down-regulation, white with no expression value) and the shape corresponds to the species (HIV-1 diamond and human circle). This figure is adapted (from Bandyopadhyay et al. 2006) with permission

HIV transcription to tip the balance in the T-cell activation in favor of viral replication. Parallel to modulating the T-cell activation, HIV-1 infection has a direct impact on the cell cycle and apoptosis of the host cells by up-regulating the genes that are involved in death receptor pathways while at the same time down-regulating DNA repair genes. HIV-1 has also been shown to induce genes that are involved in cholesterol biosynthesis to enhance viral infectivity and replication (Giri et al. 2006). These approaches have provided a global view of host gene modulation by HIV-1, have provided answers to many global biological questions, and have suggested novel hypothesis related to immune dysregulation, susceptibility to apoptosis, virus replication, and viral persistence following in vitro or in vivo infection by HIV-1. On the other hand, a limitation to this approach is that many human gene functions have not been (correctly) annotated to defined pathways.

Recent systems approaches, in addition to clustering of the gene and/or protein expression data with predefined pathways, aim to integrate protein–protein interaction network information to identify interesting groups of genes that have not been pre-identified in pathways in an ontological framework. This approach combines measures derived from expression data and metrics on biological networks into single coherent framework. To this end, expansion of protein–protein interaction (PPI) databases mainly either through high-throughput experimental techniques (such as two-hybrid screens and affinity purification followed by mass spectrometry (AP/MS)) or literature curation (assembled from publicly available datasets), coupled to advances in software technology, can be used to integrate data within novel interaction networks to identify disease-specific, functional subnetworks (Galitski et al. 2004; Sharan et al. 2007). One of the approaches aims to integrate the molecular networks with the gene or protein expression data with the goal



Systems Immunology, Adaptive Immune Response to HIV Infection, Fig. 4 Active network of gene expression during the early stages of HIV reactivation. The zoomed-in image shows the top five proteins that are highly connected with proteins with differential expression. This figure is adapted (from Bandyopadhyay et al 2006) with permission

to extract relevant network modules (expression activated modules) based on coherent expression patterns of their genes. Ideker and colleagues pioneered this approach (Ideker et al. 2002) that has been later extended and improved by several groups. The idea behind this approach is that by interrogating a protein interaction network with high-throughput data such as gene/protein expression it is possible to extract subnetworks whose protein states are perturbed by the condition of interest and functionally connected.

There exists a rich literature on methods for integrating protein–protein interaction networks and pathway databases with protein/gene expression data with the goal of identifying expression-activated subnetworks or modules that are better diagnostic and prognostic markers for cancer (for review, see Ideker and Sharan 2008; Nibbe and Chance 2009). Of interest, recently it is demonstrated that subnetwork markers extracted based on protein-network approach are more reproducible classifiers of breast cancer metastasis than individual gene markers selected from conventional expression-alone analysis (Chuang et al. 2007). This approach has also recently been improved to combine both protein and gene expression data (Nibbe et al. 2010).

HIV-1–human interaction databases based on high-throughput protein interaction measurements are under development, and there is a publicly available HIV-1 and human protein interaction database curated from multiple studies accessible at http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/. This database represents a unique and continuously updated scientific resources and a graphical visual representation of HIV-1–human protein interactions (Ptak et al. 2008) and is shown in Fig. 2. It is important to understand this complex interaction between HIV-1 proteins and a vast array of host protein as demonstrated in Fig. 2 to understand the underlying mechanisms in viral replication and pathogenesis.

Even though there is not an established secondary level literature review on annotating network information for viral pathogenesis, some progress has been made recently. For example, with the availability of HIV-1–human interaction database, Human Protein Reference Database (HPRD), and high-throughput time series gene expression data for HIV-1 reactivation in human T-cell lines, it has been illustrated that a network-level analysis (integrating the expression clustering with the protein–protein interaction networks) can identify significant "activated networks" and/or subnetworks that are unique to the latent and early stages of viral replication (Bandyopadhyay et al. 2006). According to the authors, with this integrated approach they were able to extract highly significant active networks ($P < 0.05$) for both latent (un-induced) and in the early stage of HIV-1 activation. As shown in Figs. 3 and 4, the active networks for the latent and early-stage HIV-1 activation, respectively, are enriched with Tat-interacting proteins. One of the interesting outcomes from this analysis is to see how the active networks topology changes from the latent to the early stages in HIV-1 activation. In addition, the approach offers additional information such as identifying the genes (e.g., Tuba3, Fig. 1) that are not annotated as indicting changes in expression but strongly interact with genes that have significant differential expression. Thus activated networks and or subnetworks are novel diagnostic and/or prognostic markers and they also provide multiple layers of information related to the causes of disease or viral pathogenesis that can be tested.

Acknowledgments This work is supported in part by NCRR Grant UL1-RR-024989, NIDA Grant P20-DA-026133, CFAR Grant P30-AI-036219, and NIDCR Grant P01-DE-019759.

Cross-References

- ▶ Gene Ontology
- ▶ Systems Immunology

References

Bandyopadhyay S, Kelley R, Ideker T (2006) Discovering regulated networks during HIV-1 latency and reactivation. Pac Symp Biocomput 11:354–366

Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140–149

Galitski T (2004) Molecular networks in model systems. Annu Rev Genomics Hum Genet 5:177–187

Giri MS, Nebozhyn M, Showe L, Montaner LJ (2006) Microarray data on gene modulation by HIV-1 in immune cells: 2000–2006. J Leukoc Biol 80:1031–1043

Ideker T, Sharan R (2008) Protein networks in disease. Genome Res 18:644–652

Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18(Suppl 1):S233–S240

Kindt TJ, Goldsby RA, Osborne BA (2007) Immunology, 6th edn. W. H. Freeman, New York, pp 1–53

Nathanson N, Rafi A et al (2007) Viral pathogenesis and immunity, 2nd edn. Academic Press, Amsterdam/Boston, pp 185–200

Nibbe RK, Chance MR (2009) Approaches to biomarkers in human colorectal cancer: looking back, to go forward. Biomark Med 3:385–396

Nibbe RK, Koyutürk M, Chance MR (2010) An integrative -omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput Biol 6:1–15

Ptak RG, Fu W, Sanders-Beer BE, Dickerson JE, Pinney JW, Robertson DL, Rozanov MN, Katz KS, Maglott DR, Pruitt KD, Dieffenbach CW (2008) Cataloguing the HIV type 1 human protein interaction network. AIDS Res Hum Retroviruses 24:1497–1502

Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 88:1–13

Tan SL, Ganji G, Paeper B, Proll S, Katze MG (2007) Systems biology and the host response to viral infection. Nat Biotechnol 25:1383–1389

Systems Immunology, Data Modeling and Scripting in R

Ramachandran Srinivasan, Rupanjali Chaudhuri, Rajni Verma, Ab Rauf Shah, Rituparna Sen and Chaitali Paul
G.N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Delhi, India

Synonyms

Antigenic determinant; Epitope

Definition

Immunological data refers to data corresponding to the molecules and activities of the immune system.

The discipline of immunoinformatics deals with applying bioinformatics principles and tools to the molecular activities of the immune system. Immunoinformatics provides databases and predictive tools, which are used in discovering novel vaccines and this approach is referred to as computer-aided vaccine design. The focus of immunoinformatics has been to enable identification of antigens or epitopes capable of eliciting immune response.

B-cells and T-cells are important cells of the immune system. Subsequent to recognition of epitopes, these cells are activated. An epitope, also known as "antigenic determinant" is a surface localized part of antigen capable of eliciting an immune response. A B-cell epitope is the region of the antigen recognized by soluble or membrane-bound antibodies. B-cell epitopes are classified as either linear or discontinuous epitopes. Linear epitope constitutes of a single continuous stretch of amino acids within a protein sequence, whereas epitopes whose residues are distantly placed in the sequence brought together by physicochemical folding constitute discontinuous epitopes. T-cell epitope is a short region presented on the surface of an antigen-presenting cell, where they are bound to MHC molecules.

An important goal in the process of discovering new vaccines is to identify protein sequences capable of generating a potent protective immune response while minimizing the possibility of developing cross-reactions with host system components. Data using immunoinformatics can help in this endeavor to a significant extent. To this end, in order to help users mine the data through a set of criteria meeting the requirements, user-friendly software platforms are built for query analysis using scripts.

Characteristics

Tools of Immunoinformatics

Epitope prediction tools are hallmark of immunoinformatics. The main goal of these tools is to aid in reliable epitope identification. Computational T-cell epitope prediction methods have been developed such as algorithms based on artificial neural networks and weight matrices – NetMHC, predictive IC(50) values IEDB-ARB method (Zhang et al. 2008; Bui et al. 2005), predicted half-time of dissociation – Bimas, and quantitative matrices – Propred. Reliable and accurate B-cell epitope prediction is still in development although we have some tools such as ABCpred and Bcepred.

It is desirable that a candidate vaccine is nonallergic. In this direction various tools of immunoinformatics have been developed with aim to predict allergenic proteins. AlgPred allows prediction of allergens through single or combination of support vector machine, motif-based method, and searching the database of known IgE epitopes. Allermatch performs BLAST search against allergen representative peptides using a sliding window approach. References for all tools are available from the Journal article MalVac (Chaudhuri et al. 2008).

Process Initiation

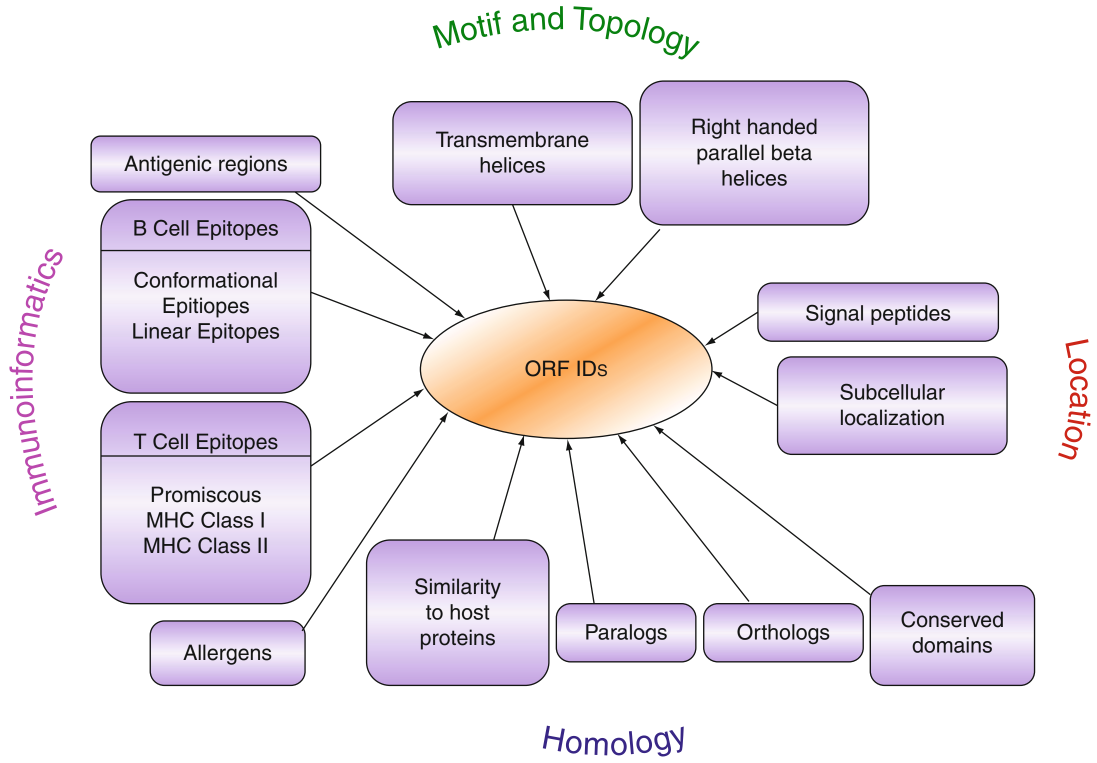
Starting of an analysis involves identification of the relevant subproblems under the major problem undertaken. The fundamentals of the main problem need to be understood elaborately. This should be followed by subsequent data mining using literature search and relevant bioinformatics and immunoinformatics tools. As a case in point, where the problem undertaken is identification of potential adhesin vaccine candidate, the fundamental importance of the problem lies as under:

- Induced antibody response at cell surface against adhesins can prevent attachment of pathogen to cell surface and thus abrogate colonization at the very first stage of infection.
- Adhesins also show a high degree of antigenic conservation (bind to invariant host receptors)

The subproblems involve identification of probable adhesin proteins from a pathogen's proteome to identify probable adhesins (Sachdeva et al. 2005) or surface located proteins using subcellular localization tools. Other bioinformatics tools of relevance are orthologs, paralogs, transmembrane topologies, beta helix supersecondary structural motifs, signal peptides, similarity against human proteins, and conserved domains (Fig. 1). Each prediction comes with associated confidence level. References for all tools are available from the Journal article MalVac (Chaudhuri et al. 2008). This set of analytical data constitutes the first layer for datamining. The immunoinformatics data constitute the second layer.

Creation of Datasets in R Platform

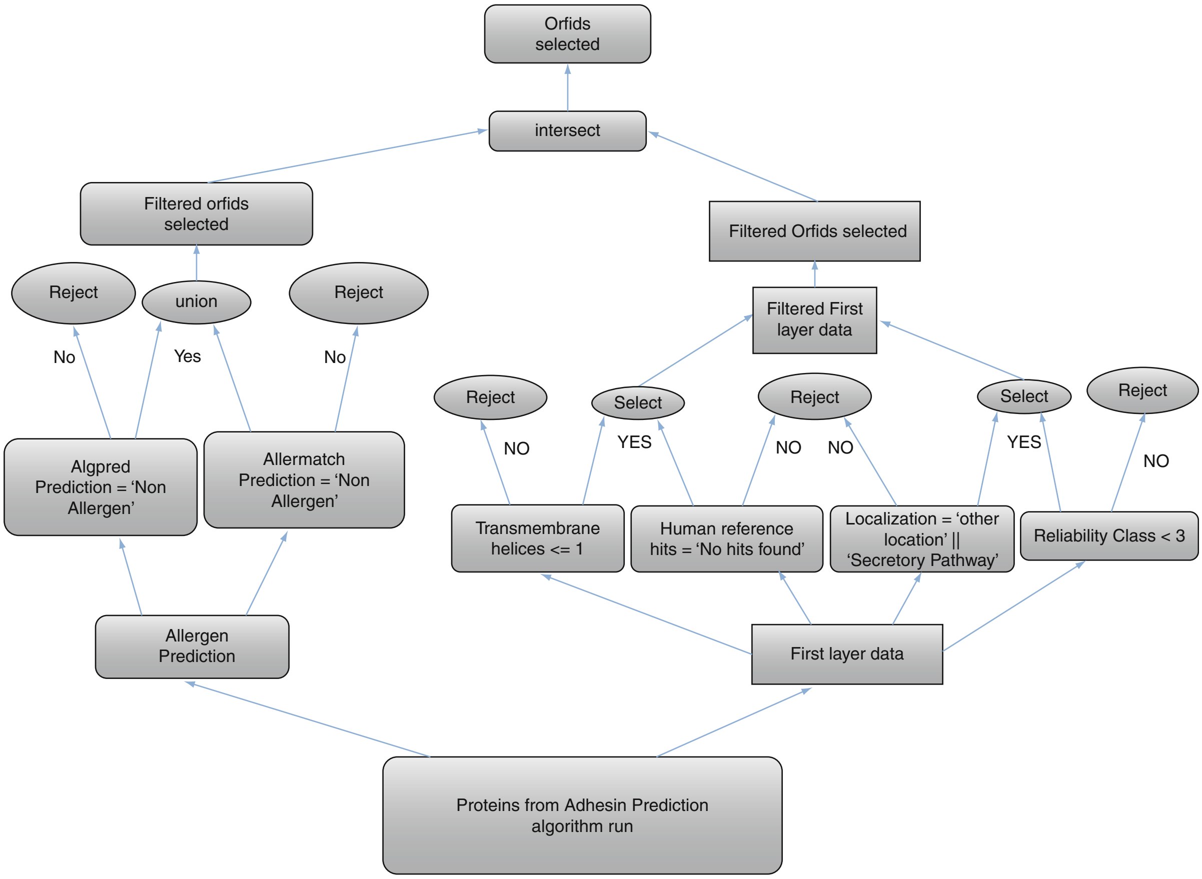
R is a high-level interpreted language suitable for developing new computational methods



Systems Immunology, Data Modeling and Scripting in R, Fig. 1 Data layout of bioinformatics and immunoinformatics for both first layer and second layer

(R Development Core Team, 2010). Several computational biology packages have been developed in R language. Developing computational packages in R provides advantage as to carry out the analysis locally and also build further tools and scripts. This facilitates development of both new applications and extension of existing applications. R thus facilitates accomplishment of complex tasks using simple scripts. Another major advantage of preparing datasets and computational biology tools in R is that a large set of statistical and mathematical tools can be applied on the datasets for analysis. R being an open source controlled by GNU General Public License, allows future developments and customizations more widely. The responsibility for the maintenance of R is taken upon by a core group thereby ensuring its availability for long life.

Data for R platform can be prepared as ".RData". For the problem addressed in the above section, data were prepared as CSV files from various pathogens *Plasmodium falciparum, Plasmodium vivax, Plasmodium yoelii* (malaria causing species), *Aspergillus fumigatus, Candida albicans, Candida glabrata, Blastomyces dermatitidis, Histoplasma capsulatum, Coccidioides immitis, Coccidioides posadasii, Paracoccidioides brasiliensis* (fungal pathogens), *Mycobacterium tuberculosis* (H37Rv and H37Ra strains) and *Chikungunya Virus*. The orthologs, paralogs, transmembrane topologies, beta helix supersecondary structural motifs, subcellular localization, similarity against human proteins, antigenic regions, conserved domains informations were placed into one CSV file as first layer data and read into a single R object. The epitope and allergen information collected using various immunoinformatics prediction tools were read into separate R objects, each corresponding to the specific result. These R objects were saved together as R image data files *.RData. These can be sourced from the link http://sourceforge.net/projects/sysbior/. All the data objects can be accessed instantly by loading the R image data files using the load command.


Orfids selected
intersect
Filtered orfids selected
Filtered Orfids selected
Reject
union
Reject
Filtered First layer data
No
Yes
No
Reject
Select
Reject
Select
Reject
NO
YES
NO
NO
YES
NO
Algpred Prediction = 'Non Allergen'
Allermatch Prediction = 'Non Allergen'
Transmembrane helices <= 1
Human reference hits = 'No hits found'
Localization = 'other location' || 'Secretory Pathway'
Reliability Class < 3
Allergen Prediction
First layer data
Proteins from Adhesin Prediction algorithm run

Systems Immunology, Data Modeling and Scripting in R, Fig. 2 Decision tree for selection of non allergenic proteins fulfilling the conditions on first layer data

Similarly multiple R image data files can be loaded as desired within seconds.

Samples and Scripting

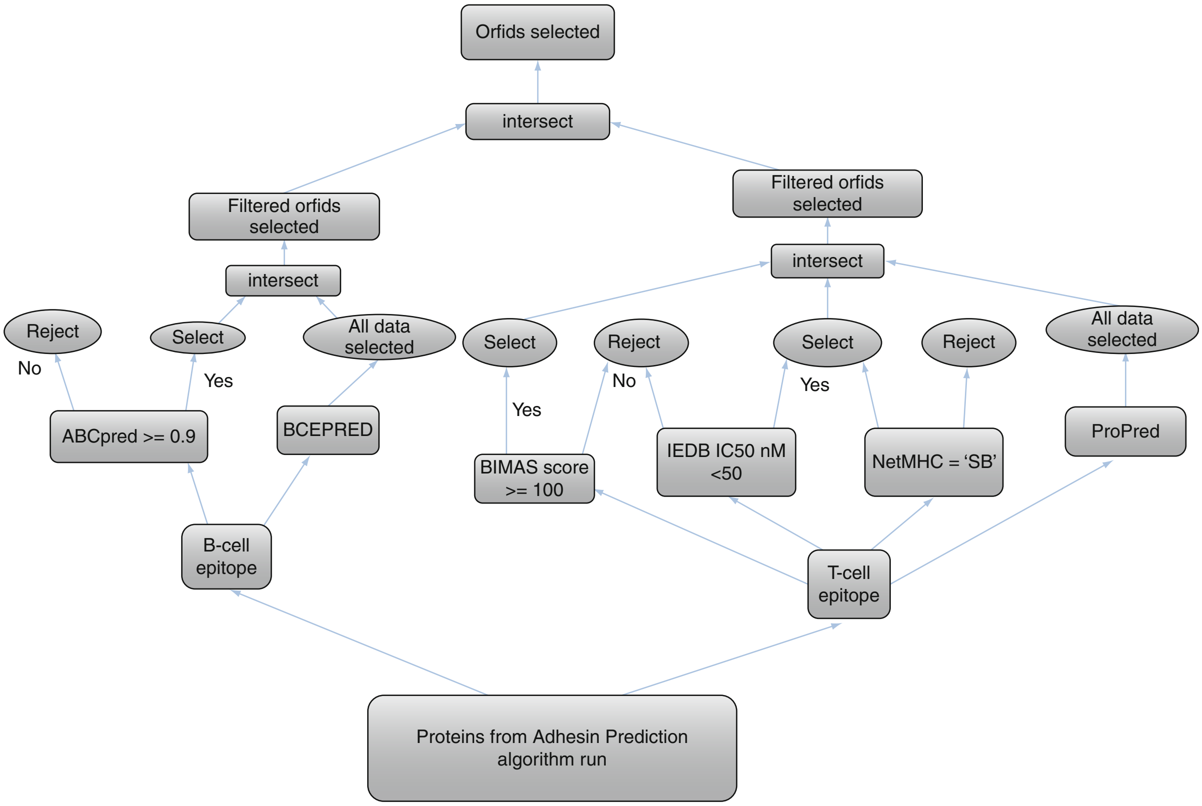
1. Knowing the contents: After loading the R image data files, ls() command allows listing of the loaded objects. The command names(dataobjectname) displays characteristics of the data table corresponding to a given R object. The dim(dataobjectname) command displays its size. Further work depends on the question asked and knowing the data types present in the R object. This information will help formulate scripts to carry out searches for complex queries in order to meet conditional criteria.

   We represent here a decision process taken using immunoinformatics data. The process can be implemented using scripts for conditional searches and operations of set theory

2. Sample Scripts: To address the example problem undertaken as under

(a) From the first layer data certain subproblems to target a potential adhesin vaccine candidate can be stated as follows: the protein should be an adhesin, the protein should not be intracellularly located, it should not have similarity to human reference proteins, and it should not have more than one tansmembrane helices facilitating proper cloning and expression. The following script retrieves the resulting proteins meeting these criteria:

To get ORFids of the filtered proteins fulfilling all conditions made directly into an R object firstlayer_filteredorfidsfirstlayer_filteredorfids<-NULL; for (i in 1:322) {if(((as.character(firstlayer_malaria[i,10])=="Other_Location") || (as.character(firstlayer_malaria[i,10])=="Secretory_Pathway")) &&



Systems Immunology, Data Modeling and Scripting in R, Fig. 3 Decision tree for selection of proteins having high scoring B-cell and T-cell epitopes

(firstlayer_malaria[i,11]<3) && (firstlayer_malaria [i,14]<=1) && (as.character(firstlayer_malaria [i,17])=="No hits found")) firstlayer_filteredorfids <-c(firstlayer_filteredorfids, as.character(firstlayer_malaria[i,1]))} here the 14th column refers to the number of tansmembrane helices, the 10th column has data for localization of the protein, the 11th column has reliability class (RC) of localization prediction (this value ranges from 1–5, and lower the RC, greater the confidence of prediction), and 17th column has data for similarity of the protein to human reference proteins. Post execution of this script z5 stores the row numbers of the proteins fulfilling above-mentioned criteria.

(b) To get the ORFids of high scoring B-cell and T-cell epitopes containing proteins, scripts can be written using conditional operators. If, for example, to select high scoring B-cell from ABCpred abcpredfilteredorfids <- NULL; for (i in 1: 32299) {if (as. numeric(as.character(abcpred_malaria[i,6]))>=0.9) abcpredfilteredorfids <-c(abcpredfilteredorfids, as.character(abcpred_malaria[i,1]))} here the 1st column has ORF id of the protein and 6th column has scores against individual epitopes predicted from ABCPred server for the corresponding protein.

In this way, scripts can be run on First layer, B-Cell, T-cell, and Allergen data provided to select the filtered ORF Ids fulfilling all criteria which come out to be the list of potential vaccine candidates (Figs. 2 and 3).

References

Bui HH, Sidney J, Peters B, Sathiamurthy M, Sinichi A, Purton KA, Mothé BR, Chisari FV, Watkins DI, Sette A (2005) Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications. Immunogenetics 57(7):304–314

Chaudhuri R, Ahmed S, Ansari FA, Singh HV, Ramachandran S (2008) MalVac: database of malarial vaccine candidates. Malar J 7:184

R Development Core Team (2010) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Sachdeva G, Kumar K, Jain P, Ramachandran S (2005) SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. Bioinformatics 21(4):483–491

Zhang Q, Wang P, Kim Y, Haste-Andersen P, Beaver J, Bourne PE, Bui HH, Buus S, Frankild S, Greenbaum J, Lund O, Lundegaard C, Nielsen M, Ponomarenko J, Sette A, Zhu Z, Peters B (2008) Immune epitope database analysis resource (IEDB-AR). Nucleic Acids Res 36:W513–W518, Web Server issue

Systems Immunology, Novel Evaluation of Vaccine

Bertrand Bellier, Adrien Six, Véronique Thomas-Vaslin and David Klatzmann
Immunology-Immunopathology-Immunotherapy, UPMC University Paris 06, UMR7211, CNRS, UMR7211, INSERM, U959, Paris, France

Synonyms

Systems vaccinology; Vaccine molecular signatures; Vaccinomics

Definition

Systems biology offers a new approach to vaccine design by understanding the molecular networks mobilized by vaccination. Systems vaccinology approaches investigate global correlates of successful vaccination, beyond antigen-specific immune responses, providing new methods for measuring early vaccine efficacy and generating hypotheses for understanding the mechanisms that underlie successful immunogenicity. Using functional genomics, vaccine-specific molecular signatures can be identified and used as predictors of efficient immune responses.

Characteristics

Vaccine

Vaccines are the most effective tools to prevent infectious diseases in humans or animals. A vaccine typically contains an agent resembling a disease-causing microorganism. The agent (antigens) stimulates the body's immune system to recognize the agent as foreign, inducing specific immune responses and memory for long-term protection. Vaccines can be prophylactic (e.g., to prevent or ameliorate the effects of a future infection), or therapeutic (e.g., vaccines against cancer).

Immune Responses Induced by Vaccine

Following vaccination, both innate and adaptive immune system components synergize to elicit an immune response. Antigen-presenting cells – notably dendritic cells – take up antigens and traffic to the draining lymph nodes where they present processed antigens to naïve CD4+ and CD8+ T lymphocytes. Naïve T cells are stimulated to proliferate and differentiate into effector and memory T cells. Activated, effector, and memory CD4+ T cells provide help to B cells to mount antibody responses, and to naïve CD8+ T cells to enhance their clonal expansion and differentiation into cytotoxic CD8+ T lymphocytes (CTL). The quality of the vaccine-induced immune response depends on several factors, e.g., antigen nature, route of administration, antigen presentation, vaccine preparation adjuvants, and timing between challenges.

Immune Responses Required for Protection

To be effective a vaccine should be capable of eliciting:

- Activation of antigen-presenting cells to initiate antigen processing and presentation to T cells
- Activation of T and B cells:
  - Production of antibodies that bind and neutralize antigens and/or target invading pathogens for destruction by complement- or antibody-dependent cellular cytotoxicity
  - Generation of memory B cells
  - Generation of memory CTL
  - Generation of memory CD4+ T cells

Evaluation of Vaccine Efficacy

Classically, the effectiveness of vaccination is ascertained until vaccinated individuals exposed to infection are protected. A central goal of vaccine research is to identify whether an early vaccine-induced immune response is predictive of later protection. An immune correlate can be used for guiding vaccine development, for predicting vaccine efficacy in different settings, and for guiding vaccination policies and regulatory decisions.

Most current successful vaccines were developed with little or no understanding of cellular immune responses. Prior to the 1990s, most vaccination programs have been evaluated based on the efficacy to induce high antibody titres without assessing T-cell responses. Moreover, ▶ immunomonitoring methods to evaluate vaccine immunogenicity are not suitable to predict vaccine efficacy; e.g., antibody assays are mainly based on antigen-binding parameters that do not reveal the antibody function. Therefore, immune correlates of protection are poorly characterized.

New methods are now emerging to assess a growing number of vaccine-associated immune parameters. In particular, T-cell functions and interactions with other cells (e.g., antigen-presenting cells) are evaluated: tetramer binding to TCR, epitope immunoreactivity, cytokine production, cell phenotype, etc.

The expanding knowledge on the molecular mechanisms of immune responses and the development of genomic and proteomic analysis now offers new approaches for modeling vaccine-induced immune responses and opens the possibility to establish predictive signatures of effective responses.

Still, these methods assess vaccine efficacy at the individual cell, when investigations should be done at the various scales of the organism. Indeed, induction of antigen-specific memory immune responses does not imply that antibodies, memory cells, or cytokines represent surrogates or correlates of vaccine efficacy. This highlights the requirement for an integrated evaluation of vaccine efficiency including numerous multiparametric variables:

- Potential host efficiency (age-, disease-, or treatment-related immunodeficiency)
- Antigenicity of the vaccine preparation
- Cellular and humoral immunogenicity

The goal is to establish correlates between protection and cellular and molecular responses: mucosal response, local antibody production, timely B- and T-cell responses, and appropriate effector or regulatory biological pathways. Systems immunology provides new tools to investigate the immune system dynamics following immunization, and derive models of efficient vaccine-induced immune responses.

Current State-of-the-Art of Systems Vaccinology

Historically, progress in vaccine development has come in waves produced by technological revolutions (for a review, see Germain 2010). Current developments translate vaccinology as a combinatorial science which studies the diversity of pathogens and the complexity of the immune system, throughout screening or immunoinformatic tools. The future advances for vaccine development will be based on taking a systems biology approach to the immune system, leading to the creation of a virtual or *in silico* immune system capable of complex simulations. This systematic approach aims to predict vaccine immunogenicity allowing its advancement into clinic without the uncertainties of the current vaccine development processes.

▶ *Reverse vaccinology* involves the *in silico* screening of the entire genome of a pathogen to find genes that encode proteins with the attributes of good vaccine targets, using either the genome of a single pathogenic isolate or the pan-genome of a pathogenic species.

Computational vaccinology models antigen processing and presentation in order to support T-cell epitope mapping (▶ T Cell Epitope, Prediction with Peptide Libraries). Web-accessible computational methods have been developed for each of the different antigen processing steps including proteasome cleavage, transport by the transporter associated with antigen processing, peptide binding to MHC molecules, and cell surface presentation (Flower 2007; Brusic et al. 2004).

A new era of genomic vaccinology and computational prediction methods comes out, enabling systematic screening of multiple complete genomes of pathogens together with analysis of the variability of pathogens and/or MHC complex. The field of ▶ vaccinomics investigates the host genetic heterogeneity, with the aim of predicting and minimizing vaccine failure or adverse events.

Systems Vaccinology

Systems biology is bringing more robust approaches to vaccine design based upon understanding of the molecular network and relationships among the various immune system components. While genomics has successfully identified new vaccine antigens, it is also promising for evaluating vaccine-induced immune response–specific signature and assessing their predictive value.

Proof of concept was recently brought by Pulendran and colleagues who used a systems biology approach to build a predictive algorithm of yellow fever vaccine immunogenicity (Querec et al. 2009). Their method involves immunology, genomics, and bioinformatics. The investigators identified gene expression signatures in the blood a few days after vaccination that could predict, with up to 90% accuracy, the strength of the immune response to the yellow fever vaccine. The consistency of these predictive signatures across two trials for CD8+ T cell and antibody responses raises the possibility that these rules have broad applicability for different types of immunogens and vaccines.

Therefore, systems biology approaches permit the observation of a global picture of vaccine-induced immune responses at an early time point after vaccination. These gene expression signatures of early innate immune activation predict the ensuing adaptive immune responses. Thus, in addition to providing a potential tool for the forward assessment of vaccine efficacy, the findings from this systems approach provide a starting point for the development of new hypotheses aimed at elucidating the parameters that control memory T cell and antibody production.

Rational Development of Novel Genetic Vaccines

By bringing together high-throughput experimental methods and information technology, our ability to decipher complex interactions that occur in the immune system has significantly improved. The current developments in computational vaccinology, including systems modeling of vaccine responses, aim to establish immune correlates and thus to accelerate the development of effective vaccines.

In this line a number of research initiatives have been supported, such as follows:

- VIOLIN (Vaccine Investigation and Online Information Network, www.violinet.org) integrates in a dedicated database of curated vaccine experimental data, a vaccine target prediction algorithm, and a vaccine ontology.
- CompuVac (Rational Design and Standardized Evaluation of Novel Genetic Vaccines, compuvac.cs.put.poznan.pl), a European FP6 integrated project devoted to (1) rational development of a novel platform of genetic vaccines and (2) standardization of vaccine evaluation, assembled a platform of viral vectors and virus-like particles and developed standardized protocols and database to comparatively evaluate vector platform efficacy.
- ImmSim is a cellular automata-based simulator of immune responses used to compare the behavior of 64 virtual viruses with various speeds of growth, infectivity level, and lethal load. Protection against infection conferred by different vaccine strategies could be tested and showed how different viruses are more susceptible to either antibody or T-cell-mediated responses (Kohler et al. 2000).

Vaccine design and evaluation should also gain from mathematical and computer models of host/pathogen interactions and immune responses (see for reviews Cohn and Mata 2007; Flower and Timmis 2007 and ▶ Vaccine Antigen Databases). For example, models of influenza viral epitope spread over years, their spatial dissemination and antisera responses will certainly guide the design of more adapted vaccines (Park et al. 2009).

Conclusion

Despite their great success, mechanisms describing how effective vaccines stimulate protective immune responses are poorly known. A major challenge in vaccinology is to prospectively determine vaccine efficacy. High-throughput technologies, such as gene expression profiling, multiplex cytokine analysis, multiparametric flow cytometry, and imaging, combined with computational modeling, offer new perspectives. Elucidation of clusters of signatures which correlate with vaccine immunogenicity should facilitate the rapid screening of vaccines but also propose new hypotheses on how vaccines mediate protective immune responses, and identify early predictive signatures of vaccine efficacy. Systems vaccinology offers great promise for future translation of basic immunology research advances into successful vaccines.

Cross-References

▶ Immunomonitoring
▶ Lymphocyte Dynamics and Repertoires, Modeling
▶ Reverse Vaccinology
▶ T Cell Epitope, Prediction with Peptide Libraries
▶ Vaccine Antigen Databases
▶ Vaccinomics

References

Brusic V, Bajic VB, Petrovsky N (2004) Computational methods for prediction of T-cell epitopes – a framework for modelling, testing, and applications. Methods 34:436–443

Cohn M, Mata J (2007) Quantitative modeling of immune response. Immunol Rev 216:1–236

Flower DR (2007) Immunoinformatics and the in silico prediction of immunogenicity. An introduction. Meth Mol Biol 409:1–15

Flower DR, Timmis J (eds) (2007) In silico immunology. Springer, New York

Germain R (2010) Vaccines and the future of human immunology. Immunity 33:441–450

Kirschner DE, Chang ST et al (2007) Toward a multiscale model of antigen presentation in immunity. Immunol Rev 216:93–118

Kohler B, Puzone R et al (2000) A systematic approach to vaccine complexity using an automaton model of the cellular and humoral immune system. I. Viral characteristics and polarized responses. Vaccine 19(7–8):862–876

Park AW, Daly JM et al (2009) Quantifying the impact of immune escape on transmission dynamics of influenza. Science 326(5953):726–728

Pulendran B, Li S, Nakaya HI (2010) Systems vaccinology. Immunity 33:516–529

Querec TD, Akondy RS et al (2009) Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. Nat Immunol 10:116–125

Systems Immunology, Vaccine Adjuvant

Nikolai Petrovsky
Department of Endocrinology, Flinders Medical Centre/Flinders University, Adelaide, Australia

Synonyms

Immune enhancer; Immuno-stimulant

Definition

Adjuvants are compounds that when added to vaccines enhance the specific immune response against co-inoculated vaccine antigens. The word adjuvant comes from the Latin word *adjuvare*, which means to help or to enhance.

Characteristics

The concept of vaccine adjuvants arose 90 years ago from observations that the inclusion with vaccine antigens of inflammatory substances such as mineral oil or live or killed microorganisms boosted the levels of antibody generated against the co-injected antigen (Petrovsky and Aguilar 2004).

These days vaccine adjuvants are used to

- Enhance the humoral and/or cellular immune response against of purified or recombinant antigens
- Reduce the amount of antigen needed for protective immunity (antigen sparing)
- Reduce the number of immunizations needed for protective immunity (dose reduction)
- Obtain faster vaccine protection
- Improve vaccine effectiveness in individuals with impaired immunity (newborns, the elderly, transplant patients, or those with chronic disease)
- Improve the uptake of antigens by immune cells (antigen delivery)
- Drive the immune response in a particular desired direction (immune deviation)

Altogether, several hundred natural and synthetic compounds have been identified to have adjuvant activity, including bacterial, fungal, and viral compounds, such as RNA, DNA, proteins, lipopeptides, lipopolysaccharides, and glycolipids (Vogel et al.; Hackett and Harn 2006). Adjuvants principally work via activation of specific innate immune receptors which sense tissue damage or invasion and provide the immune system with an early warning or “danger signal” (Matzinger 2007). This results in production of inflammatory molecules including cytokines and chemokines such as tumor necrosis factor alpha, interleukin 1, and type 1 interferons, which together activate and attract immune cells to the site of origin of the danger signal, in this case the site of immunization, where they phagocytose the foreign antigen(s) and then migrate to the draining lymph node where they present the digested antigen to resident T and B cells. In this way antigen-specific memory T and B cells are expanded which in turn are able to elicit an adaptive immune memory response when reexposed to the relevant antigen(s).

The specific innate immune receptors which sense tissue damage and danger signals have recently been identified to include multiple members including the toll-like receptor (TLR) family, components of the inflammasome, and other cytoplasmic, lysosomal, or membrane receptors that recognize pathogen-associated molecular patterns (PAMPs) or endogenous danger signals.

Adjuvants can be classified according to their source, mechanism of action, or physicochemical properties, and can be subdivided into (1) carriers, being immunogenic proteins that provide T cell help; (2) vehicle adjuvants, being oil emulsions, liposomes, or particles that bind antigens; and (3) immunostimulants, being substances that increase the immune response to the antigen by activation of the innate immune system. Adjuvants can also be classified according to their site of delivery; that is, parenteral adjuvants are injected, mucosal adjuvants are applied to mucosal surfaces, and transdermal adjuvants are applied directly to the skin.

As new methods of vaccination are developed, for example, DNA vaccines, a major challenge has been to find new adjuvant forms compatible with these new vaccine technologies. For example, DNA vaccines can be adjuvanted by inserting the coding sequence for inflammatory cytokines such as interleukin 12 or GM-CSF into the actual DNA vaccine vector. Similarly, mucosal vaccines require their own unique adjuvants, with the most potent mucosal adjuvant being cholera toxin.

Unfortunately due to their propensity to induce severe inflammatory responses, the most potent adjuvants are often the most toxic (Petrovsky 2008). Local adjuvant reactions include pain, local inflammation, swelling, injection site necrosis, lymphadenopathy, granulomas, ulcers, and the generation of sterile abscess. Systemic adjuvant reactions include nausea, fever, adjuvant arthritis, uveitis, eosinophilia, allergy, anaphylaxis, organ specific toxicity, and autoimmune disease. Aluminum hydroxide, while not a potent adjuvant, is relatively well tolerated, helping explain its effective monopoly over human vaccine use for the last 90 years. The monopoly that aluminum adjuvants have held over human vaccines is slowly being forced back with development of new adjuvants.

One of the biggest breakthroughs in the adjuvant field was the identification and characterization over the last 20 years of the innate immune receptors including the TLRs through which many longstanding adjuvants were found to be working. For example, alum was recently found to work via activation of NALP3 leading to inflammasome activation and production of inflammatory cytokines including IL-1 and IL-18 that in turn stimulate adaptive immune responses. Similarly, bacterial and viral components, such as RNA, DNA, protein, lipopolysaccharides, and lipopeptides, bind and activate specific innate immune receptors including the TLRs, leading to activation of NFkB and an inflammatory response, thereby explaining their adjuvant activity.

The major challenge for adjuvant development remains the question of whether adjuvant reactogenicity and potency can ever be separated. Historically, compounds inducing the greatest inflammation and thereby the greatest immune danger signal have been the strongest adjuvants, with Freund’s complete adjuvant being a case in point. Unfortunately the inflammation that is critical to the potency of such adjuvants is also responsible for toxicity including local injection site pain and swelling to fevers, muscle aches, and autoimmunity. The solution to this conundrum would be to develop adjuvants that specifically enhance adaptive immunity but do not induce innate immune activation responsible for toxicity. Recent adjuvant research indicates that this objective can be met and adjuvant potency and reactogenicity potentially separated. A good example of this is a newly developed range of adjuvants based on the polysaccharide delta inulin, which show potent adjuvant activity on humoral and cellular immunity but are not reactogenic.

While vaccine adjuvants have traditionally been identified by trial and error experimentation, the identification of specific innate immune receptors, including the TLR receptors and the inflammasome and the inflammatory pathways through which adjuvants are working, has for the first time allowed a more systematic approach to adjuvant identification (Singh 2007). It is now possible to express individual TLR receptors on cell lines and then use these cell lines to screen candidate compounds for their ability to bind and activate these receptors to identify novel adjuvants. Similarly, it should be possible to use gene array signatures generated by a known adjuvant in human immune cells to screen for other compounds which generate a similar signature, and may thereby share similar adjuvant properties. On the adjuvant toxicity side, a systems approach now provides the opportunity to use similar tools to screen candidate adjuvants for potential toxicity, this time using gene array signatures in cell lines from toxin-sensitive tissues such as the liver and kidneys to deselect potentially toxic candidates without the time and cost of needing to undertake extensive animal experimentation.

Cross-References

▶ Systems Immunology, Novel Evaluation of Vaccine
▶ Vaccine Antigen Databases

References

Hackett CJ, Harn DA (eds) (2006) Vaccine adjuvants: immunological and clinical principles. Humana Press, New Jersey

Matzinger P (2007) Friendly and dangerous signals: is the tissue in control? Nat Immunol 8:11–13

Petrovsky N, Aguilar JC (2004) Vaccine adjuvants: current state and future trends. Immunol Cell Biol 82:488–496

Petrovsky N (2008) Freeing vaccine adjuvants from dangerous immunological dogma. Expert Rev Vaccin 7(1):7–10

Singh M (ed) (2007) Vaccine adjuvants and delivery systems. Wiley, New York

Vogel FR, Powell MF, Alving CR (eds) A compendium of vaccine adjuvants and excipients (2nd edn)

Systems Medicine

Gilles Clermont
Center for Inflammation and Regenerative Modeling,
University of Pittsburgh, Pittsburgh, PA, USA

Definition

The availability of high-throughput techniques opened a new epistemological era in biology. New disciplines such as genomics, functional genomics, proteomics, metabolomics, and structural biology leveraged these technological advances to deepen our understanding of cellular physiology and regulation. The need to extract knowledge from these large data streams stimulated the development of computational techniques to store, organize, mine, and model the data, facilitated by the increasing availability of cheaper computer hardware and improved performance. These developments led to the emergence of the umbrella discipline of computational systems biology (Kitano 2002). The initial intent and motivation underlying the large resources and funding devoted to this major effort was to enhance human health. The human genome project (HGP), where the bulk of the work was conducted between 1990 and 2003, and which successfully led to sequencing the entire human genome, is arguably the archetypal example of the necessary combination of high-throughput techniques and computation contributing to a monumental accomplishment. The relatively modest practical output from the HGP, such as advances in diagnostic testing for cancer, hematological and liver disease, and deeper understanding of comparative biology and evolution, has demonstrated that the road to translating these new disciplines into tools that would indeed contribute to diagnosing and treating human disease in a personalized fashion will be a challenging one.

Characteristics

One can define systems medicine as the application of systems biology to the diagnosis, prevention, pathophysiologic understanding, and treatment of developmental disorders, disease, and recovery processes in humans (Clermont et al. 2009). The concept of system as an assembly is fundamental: organisms are comprised of a large number of parts, or sub-systems, creating a whole which accomplishes biological functions beneficial to the integrity of the system, and admits observables that are relevant at the system level, not merely its constituent parts. In this sense, there are strong parallels between the discipline of systems engineering and systems medicine (Parker and Clermont 2010). Systems engineering is an interdisciplinary science that combines the expertise of industrial engineering, control engineering, management science toward the design, logistical execution, and maintenance of large complex projects such as submarine and airplane design, the international space station, and other large-scale projects such as the Internet. Importantly, systems engineering conceives of a project over its entire life cycle. Similarly, as systems medicine strives to pursue its goals, it draws not only from rich multi-scalar data streams, but also from several quantitative fields that are not otherwise naturally aligned, such as statistics and control engineering.

Many consider that the system is not bound to individual living organisms, but also extend to communities of such organisms and how these interact. At a biological level, it is increasingly recognized that human interactions with their physical and microbiological environments have a key effect on health. An argument can be made that the goal of systems medicine is to be more global in scope and to extend beyond health of individuals. This perspective has driven large-scale research efforts, such as the human microbiome project (Turnbaugh et al. 2007), which seeks to understand the relation between human health and the microbial flora inhabiting gut, oral cavity, upper respiratory system, and skin. At the population scale and from a societal viewpoint, the interaction between health-care delivery models and population health indicators is further pushing the envelope as to what could still be considered systems medicine. Clearly, there is an extensive need for modeling how different factors, including financial constraints and corporate and public policies, impact health at this level.

Some researchers have suggested the concept of translational systems biology (An et al. 2007) to describe several of the efforts, goals, and promises described above, and we clearly see a conceptual distinction between systems medicine and translational systems biology.

Systems Medicine and Personalized Therapies

A fundamental goal of systems medicine is to open an evidence-based path toward personalized therapies. The concept of personalized therapy is pervasive in clinical medicine in that clinicians will implement a broad concept of treatment for specific diseases, while constantly adapting and refining treatments to the perceived circumstances of their patients. The lack of rigor in the implementation of this approach combined with a fundamental lack of scientific evidence lead to the recognition that the initial ad-hoc attempt at personalized medicine was well-intended, but prone to error, often misguided or frankly harmful (Kohn et al. 2000). Standardizing and organizing the delivery of health care and building of the evidence, mostly from randomized clinical trials has started to address these concerns (Leape and Berwick 2005), but also to the realization that predictably effective personalized medicine remains a challenging long-term goal. The human genome project's promise was to present a full description of the human genome and to leverage this knowledge toward the goal of personalized medicine. The discipline of genome medicine offers a genome-centric approach to understanding the association between the human genome, human disease, and pharmacological approaches to treatment. Genome medicine falls under the umbrella of systems medicine.

Systems Medicine and Models

The role of mathematical models as an integrative, formal framework that represents how constituent parts of a system are dynamically linked is more central to systems medicine than it is in more traditional systems biology, or medicine. At the very least, formalizing knowledge into a model will identify critical knowledge gaps and guide experimental efforts, including experimental design. Beyond, models represent a new vehicle through which interdisciplinary teams of quantitative, biological, and clinical investigators can focus discussions and evaluate the merit of competing hypotheses prior to experimental evaluation. Models could represent the preferred method to integrate and interpret data that opens the way to personalized medicine and, thus, redraws the playing field for systems medicine.

Future Developments

Advances in basic science and mathematics will be required for systems medicine to deliver on its promise beyond the initial results of association studies. Less than 2% of the human genome codes for proteins and the majority of proteins have regulatory roles which are not primarily involved in core cellular functions. Rather, they promote system robustness. The scientific community is getting early glimpses into the role and significance of the non-protein world and the fundamental role of epigenetics in disease. Systems medicine will likely provide a strong motivation for the development of methods and applications that will integrate this evolving knowledge in more comprehensive theories of health and disease.

The field of pharmacokinetics has pioneered the development of model-based individualized predictions of pharmacokinetic data. Similar predictions will be much more difficult to achieve for more complex system for which experimental or clinical data are considerably sparser and for which model representations are less well known or subject to ongoing controversy. General methods that extend mixed effect modeling in standard statistical theory to nonlinear dynamical systems are under development, as are model selection algorithms that allow comparing the relative merit of competing models. In several fields, such as weather prediction, the concepts of model ensembles and consensus models have emerged and early approaches using similar methods to express incomplete information and other source of variations

Systems Medicine, Systems Medicine, Table 1 A roadmap for systems medicine. The success of systems medicine as a discipline is contingent upon the participation of a diverse group of promoting entities, contributing enabling initiatives according to their domains of expertise and influence. The color scheme reflects the anticipated importance of their relative contribution to each initiative (Adapted from Clermont et al. 2009)

Inititiave	Rationale	Promoting entities			
		Academia	Government	Biotechnology industry	Pharmaceutical industry
Organization, funding and regulation					
Education and training	Institution-wide initiatives promoting cross-disciplinary training and research between different faculties, where clinical researchers are included. Medical school curricula minimize training in the basic and quantitative sciences that would promote interest in interdisciplinary research.	▓			
Centers of Excellence	Grow a critical mass of researchers from the clinical, basic and quantitative sciences with adequate funding, logistic support and sufficient access to patients.	▓	▓	░	░
Validation studies	A fundamental obstacle to acceptance of model-based methods in the clinical arena is the lack of animal and human validation studies of the predictive abilities of translationally relevant models.	▓	▓		▓
Investigator initiated research	Favoring of multidisciplinary research, preferable following a multiple principal investigator format will lead to better study design and faster progress from cross-fertilization.	▒	▒		
Regulation requirements	Requirements by regulatory bodies of the inclusion of model-aided study design in new drug portfolios would constitute a strong incentive for collaborative effort.		▓		▓
Technologies					
Model-aided clinical study design	Irrespective of regulatory requirements, model-aided study desing will lead to improved chances of therapeutic success, improved selection of biomarkers of therapeutic efficacy, and selection of better outcome measures, through more accurate population targetting (enrichment), drug dosing, or disease activity assessment.	▒	░		▒
Point of care technologies	Despite much progress, acceptance of such technologies has been slow because their clinical useful has not been proven.	▒		▓	▒
Knowledge acquisition methods					
Ontology development	Mechanistic insight into pathophysiology is obscured by an incomplete, possibily incorrect, understnading of biologically related functions and their interaction in the creation of clinical phenotypes.	░		░	
Computational algorithms	Extraction of knowledge from data reamins a key obstacle to progress. This is particularly true in systems medicine and clinical medicine. Theoretical challenges faced by statisticians, mathematicians and computational scientists remain monumental. Increasing computational power in the laboratory and at the bedside is a distinct, yet complementary challenge.	▓			
Disease models	The development of biologically sound computational models is of high importance and key to successful validation studies and accurate predictions.	▓			▓
Knowledge dissemination					
Access to mainstream clinical literature	Publications that have gone beyond a simple mapping of genotypes to phenotypes are extremely rare in the clinical literature. Familiarity with model-aided experimental design and insight into mechanistic pathophysiology and clinically relevant prediction in severely lacking on most editorial boards. Metrics to judge the quality of this emerging literature have not been defined and therefore conservatism prevails.	▓			
Access to scientific fora of clinical relevance	Very few meetings with a major clinical following will have sessions devoted to systems biology, modeling, or personalized medicine. Expertise of scientific committees is lacking.	▓			

* The intensity of the boxes reflect the relative urgency of the different roadmap initiatives listed, as well as the relative contribution of academia, government, biotechnology companies and the pharmaceutical industry to the initiative.

in a system in terms of parametric and structural model uncertainty. It would stand to reason that a proximal goal of this research would be to describe "similar patients" in terms of such a model ensemble, on the way to a fully probabilistic description of individual patients.

How accurate would inferences made about individual patients based on such model ensembles need to be? Scientists do not require full understanding of a system before useful applications can be realized. Gravity and electromagnetism come to mind as canonical examples of physical phenomena for which a full theory is not yet established, but applications based on incomplete knowledge have, nevertheless, shown extreme usefulness. How much knowledge is necessary before practical predictions can be extracted from existing models of a system is a difficult question. This applies directly to systems medicine. It is likely that current knowledge is sufficient to predict the effect of certain interventions. Several biosimulation companies have developed prediction engines for clinical trials based on in silico models of disease. Whether such contributions have actually impacted drug design on delivery is unclear at this stage, but some degree of success is almost certain within the next few years.

A Roadmap for Systems Medicine

The construction of a roadmap for systems medicine is imperative for its development as a successful science and will require the involvement of several stakeholders, including academic faculty and trainees, the scientific dissemination industry, institutions of higher knowledge, government and other regulatory and funding entities, the biotechnology industry, and pharmaceutical companies (Table 1). Each of these promoting entities is an essential stakeholder in systems medicine.

References

An G, Hunt CA, Clermont G, Neugebauer E, Vodovotz Y (2007) Challenges and rewards on the road to translational systems biology in acute illness: four case reports from interdisciplinary teams. J Crit Care 22:169–175

Clermont G, Auffray C, Moreau Y et al (2009) Bridging the gap between systems biology and medicine. Genome Med 1:88

Kitano H (2002) Computational systems biology. Nature 420:206–210

Kohn LT, Corrigan J, Donaldson MS (eds) (2000) To err is human: building a safer health system. National Academy Press, Washington, DC

Leape LL, Berwick DM (2005) Five years after to err is human: what have we learned? JAMA 293:2384–2390

Parker RS, Clermont G (2010) Systems engineering medicine: engineering the inflammation response to infectious and traumatic challenges. J R Soc Interface 7(48):989–1013

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI (2007) The human microbiome project. Nature 449:804–810

Systems Microbiology

▶ Metagenomics

Systems Network in HIV

▶ Systems Immunology, Adaptive Immune Response to HIV Infection

Systems Pathology

Dana Faratian[1] and David Harrison[2]
[1]Centre for Molecular Pathology, Institute of Cancer Research, Surrey, UK
[2]School of Medicine, University of St Andrews, St Andrews, Scotland, UK

Synonyms

Clinical systems pathology; Predictive pathology; Systems histopathology

Definition

Systems pathology is the study of disease through the integration of clinical, morphological, quantitative, and molecular parameters using mathematical analytical frameworks. The aim is to create coherent models which enable the understanding of pathophysiological processes in their entirety and generate hypotheses that can be tested experimentally. In practice, systems pathology aims to personalized therapy and predictive outcomes for patients (▶ Personalized Medicine; ▶ Predictive Medicine).

Characteristics

Pathology is the study of mechanisms and diagnosis of disease and is the basis of all components of Laboratory Medicine including hematology, immunology, biochemistry, microbiology, genetics, and anatomic/histopathology. The word pathology comes from the Ancient Greek πάθος, *pathos*, "feeling, suffering"; and -λογία, *-logia*, "the study of." Pathology addresses four main components of disease, namely, etiology (cause), mechanism (pathogenesis), cell and tissue structure (morphology), and the clinical manifestations or consequences of disease. The main branch of pathology is anatomical pathology, which is concerned with the diagnosis of disease based on examination of the gross, microscopic, immunological, or molecular examination of tissues, organs, and whole bodies. These examinations often demand qualitative rather than quantitative analysis, such as human interpretation of thin slices of tissue (histopathology) by pathologists to make a diagnosis of cancer, or the presence or absence of a protein labeled with an antibody (immunohistochemistry) in order to determine whether a particular target for therapy is present and, therefore, whether the agent should be given. However, anatomical pathology remains the mainstay of analysis of tissues, organs, and bodies, and frequently determines the need and mode of therapy for patients with both neoplastic and nonneoplastic diseases, often very accurately.

Four main developments have challenged the use of this traditional framework for anatomical pathology:

1. An awareness that within disease groups there is considerable heterogeneity both between patients, and within each individual patient, which may influence the natural progression of disease and response to therapy.
2. The development of high-throughput analytical techniques, such as ▶ DNA-microarrays and next generation sequencing (see ▶ DNA Sequencing), which give unprecedented detail on the underlying molecular abnormalities of disease, that is pathogenesis.
3. An understanding that the quantitative, as well and qualitative differences in biological parameters influence cellular outcomes and therefore disease pathogenesis (e.g., sustained versus transient MAPK signaling can result in differentiation or cell division, respectively).
4. Diseases change phenotypic characteristics as they progress, and therefore both spatial and temporal parameters must be taken into account (e.g., differences between primary and metastatic disease in ▶ Cancer).

Therefore it is implicit that in systems pathology:

1. Data generation should be quantitative rather than qualitative.
2. Data generation should aim to quantitate both temporal (disease progression, pharmacokinetic/pharmacodynamics) and spatial (site in organism, morphology) characteristics of disease.
3. Disparate, multiscale data sources should be integrated (e.g., clinical and radiological parameters, histopathology, molecular) in order to characterize how the disease process impacts on the organism as a whole.
4. There should be a mathematical and computational framework which can handle quantitative, qualitative, and dynamic data types.

Algorithms and Tools (Experimental)

The tools for experimental data generation for systems pathology can broadly be divided into molecular, spatial, and temporal quantification.

Molecular data generation is increasingly high-throughput, especially with the advent of next generation sequencing, which permits quantitative analysis of DNA and RNA sequence at base-pair resolution. Likewise, ▶ mass spectrometry has advanced protein analysis to the level of post-translational modifications.

Spatial resolution usually employs the use of imaging, either isolation or with molecular labeling (such as ▶ Fluorescence Microscopy with in situ hybridization or protein labeling with antibodies), in order to quantify the relationship of structures on the subcellular, tissue, or organism level. Examples of imaging include digital imaging of histological sections using light or ▶ fluorescence microscopy, or more recently imaging of molecules from tissue sections using MALDI-TOF ▶ mass spectrometry.

Temporal quantification for systems pathology makes use of suitable experimental models (in vitro or in vivo) in order to quantitate biological phenomena over time, or relevant clinical models in order to quantify changes in disease process over time. An example might be to measure changes in gene expression in cancer tissue samples before and after therapy, as has frequently been performed in breast cancer clinical trials.

Ideally, the above methods are combined in order to gain a systems level understanding of pathology, such as ultra-deep sequencing in order to resolve spatial complexity and molecular heterogeneity, or in vivo models with molecular imaging.

Algorithms and Tools (Mathematical)

No single mathematical framework is universally appropriate for analyzing systems pathology data. However, methodologies may broadly be divided into hypothesis-driven and data-driven approaches. Hypothesis-driven approaches that are relevant to the study of the dynamics of cell networks, which are important in cancer, include those where the network is prescribed a priori, such as mechanistic, deterministic ordinary differential, equation-based mathematical models (e.g., those describing ▶ Receptor Tyrosine Kinase signaling). Important examples of data-driven approaches are general ▶ power-law functions (a set of mathematical tools for the approximation, modeling, numerical simulation and analysis of nonlinear systems) including ▶ biochemical systems theory (BST) and ▶ metabolic control analysis. The alternative data-driven approach can be used when qualitative and/or imprecise measures are present to harness ▶ fuzzy logic, a mathematical term used to describe decisions or biological readouts that have a continuous range between 0 and 1, that is, they are imprecise but within boundaries, in contrast to binary logic, when the decision is discrete, either 0 or 1. Fuzzy logic is a type of mathematical approach which might be useful where qualitative data are available, such as from gene expression array or clinical data.

Examples

Systems pathology has been used with good effect to predict outcome in prostate cancer, through the integration of clinical, molecular, and morphometric data using support vector machine methods (Saidi et al. 2007). Likewise, in breast cancer, systems pathology was used in order to predict sensitivity to targeted therapy and aid the ▶ biomarker discovery process in breast cancer using an ▶ ordinary differential equation based approach (Faratian et al. 2009).

Cross-References

- ▶ Biochemical Systems Theory (BST)
- ▶ Biomarker Discovery, Typical Process
- ▶ Biomarkers, Solid Tissue
- ▶ Cancer
- ▶ Cancer Pathology
- ▶ DNA Microarrays
- ▶ DNA Sequencing
- ▶ Fluorescence Microscopy
- ▶ Fuzzy Logic
- ▶ Mass Spectrometry, Proteomics, and Metabolomics
- ▶ Metabolic Control Analysis
- ▶ Ordinary Differential Equation (ODE)
- ▶ Personalized Medicine
- ▶ Power-Law Functions
- ▶ Predictive Medicine
- ▶ Receptor Tyrosine Kinase

References

Faratian D, Goltsov A, Lebedeva G, Sorokin A, Moodie S, Mullen P, Kay C, Um IH, Langdon S, Goryanin I, Harrison DJ (2009) Systems biology reveals new strategies for personalizing cancer medicine and confirms the role of PTEN in resistance to trastuzumab. Cancer Res 69:6713–6720

Saidi O, Cordon-Cardo C, Costa J (2007) Technology Insight: will systems pathology replace the pathologist? Nat Rev Urol 4:39–45

Systems Pharmacology

Shan Zhao and Ravi Iyengar
Department of Pharmacology and Systems
Therapeutics, Mount Sinai School of Medicine,
New York, NY, USA

Synonyms

Systems biomedicine

Definition

System pharmacology describes the application of systems biology to answer questions relating to the action of drugs, which include pharmcokinetics, pharmacodynamics, toxicology, and drug discovery.

Characteristics

The number of drugs the average individual is being prescribed has increased in the past decade, and in the first 8 months of 2010, another 69 first-time generic drugs have been approved. In this setting of increasing pharmacotherapy, it is important to know how drug combinations may interact to produce both beneficial and ▶ adverse events. We need an integrated understanding of how drugs are affected by both genetic and environmental factors. The field of systems pharmacology helps with this integration, by considering data at various levels (Wist et al. 2009). Systems pharmacology helps determine how the organization of cellular networks, within which drug targets reside, affects tissue-and organ-level functions to produce both the desired beneficial effects and ▶ adverse events. This knowledge can be used to tailor individualized therapies, which is at the core of ▶ personalized medicine.

Sequencing genomes, profiling mRNAs and proteomic studies allow scientists to examine the biology of diseases at a higher resolution. It has become apparent that disease profiles are complex and most are phenomena emerging from the interactions of multiple gene products with environmental factors. Even known single gene mutation disease, such as Marfan’s (El-Hamamsy and Yacoub 2009), has been shown to involve many downstream interactions, which influence the phenotypes of the disease. Systems medicine and pharmacology offer a framework for integrating the hundreds of thousands of interactions and identify important interactors and interactions (▶ Edge Betweenness Centrality) involved.

Goals of Systems Pharmacology

The overall goals of drug therapy are to determine what drugs should be prescribed to which patients, the dosage for maximum beneficial effect, and what side effects may be expected. This knowledge is required for the practice of personalized and ▶ evidence-based medicine. However, predicting the balance between clinical benefit and adverse side effect is a complex because a patient’s response to a drug is affected by various factors, including diet, co-administered medications, genetic background, and behavioral responses. Systems pharmacology can help us predict the balance between clinical benefit and risk of ▶ adverse events by analyzing the drugs in conjunction with the human ▶ interactome, and integrating clinical, behavioral, environmental, and genomic background (Fig. 1).

An important obstacle in pharmacology concerns drug design and therapy optimization. Even if a drug can be designed to not to bind to ▶ off-targets, the multiplicity of downstream effects of the ▶ primary-target may lead to ▶ adverse events (Fig. 2a). As a result, pharmacologists have started to consider combination therapies that each on their own have minimal effect on the phenotype, but together interact to achieve an optimal effect (Fig. 2b). Systems pharmacology offers a method to aid in the design of a combination profile, by predicting how various targets may participate together to give a beneficial effect profile.

Data Sources

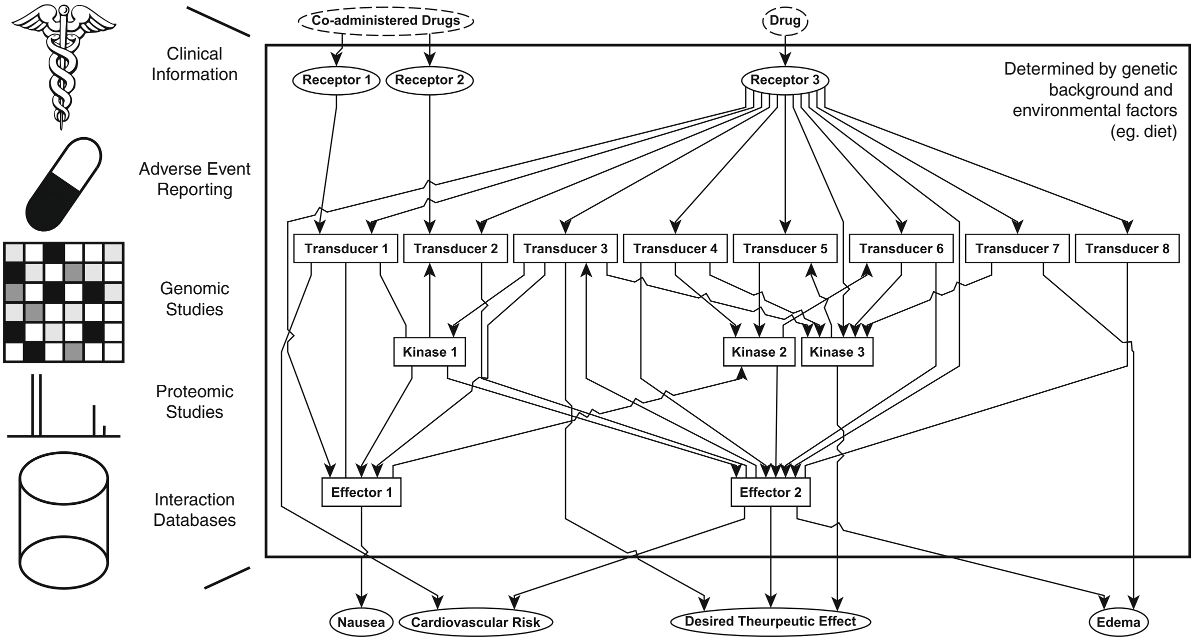
An essential feature of systems pharmacology is the integration of data sources. Some of the important types of data sources include Drug structure and target databases, database of interactions of human gene products, genome wide association studies, and ▶ adverse events databases.

Drug Databases

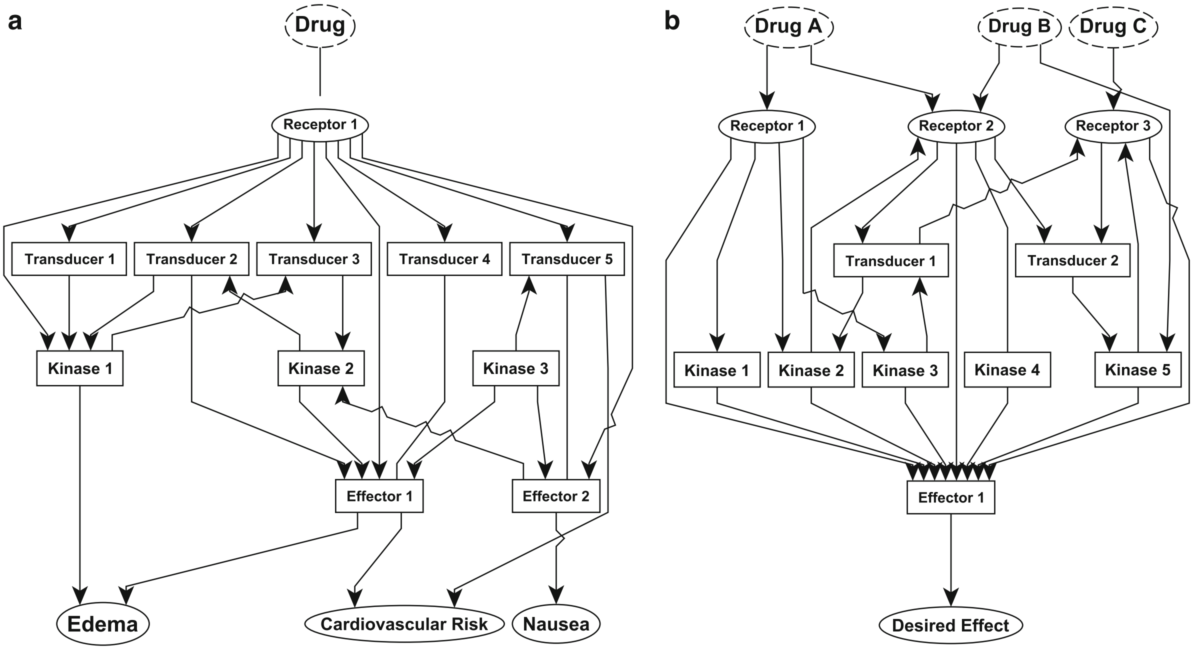
In order to study drugs using a systems pharmacology approach, one first needs to link the drug and their targets together. This task is often done utilizing databases, such as Drugbank (http://drugbank.ca/) or PharmGKB (http://www.pharmgkb.org/). These databases offer information on drug properties, including gene targets as well as metabolic information, which can then be analyzed using systems biology methods.

Databases of Gene Product Interactions

Pharmacological treatments perturb cellular systems interacting through the underlying ▶ interactome, that make up the regulatory pathways and cellular machines. Systems pharmacology uses interaction



Systems Pharmacology, Fig. 1 A schematic representation where ▶ systems pharmacology integrates various types of information (Clinical Phenotypes, Adverse Events, Genomic and Proteomic Studies, and Interaction databases) into an ▶ interaction network (represented by *black squares*), performs analysis adverse event reporting to predict the effects profile



Systems Pharmacology, Fig. 2 (**a**) A representation of the traditional paradigm of drug design, where a drug is optimized for a particular target, but despite the targeting being optimized, the effect profile downstream may still vary. (**b**) The use of ▶ systems pharmacology to combine multiple drugs together to give a targeted effect, also minimizing for side effects databases, to construct the ▶ interactome. Together, these databases help to highlight how different drugs may influence each other and how these interactions may lead to unexpected beneficial or harmful events. Examples of these databases may include large networks, such as Biogrid (http://thebiogrid.org/) and HPRD (http://www.hprd.org/), as well as more detailed pathway networks, such as NCIPathway (http://pid.nci.nih.gov/) and KEGG (http://www.genome.jp/kegg/).

Genome Wide Association Studies (GWAS)

Systems pharmacology often includes genomic information that can help to capture the individual variations between patients. This information has become increasingly available. Many large post-approval drug safety clinical trials have started to conduct genomic wide association studies. Such studies can be found at various websites, including http://www.genome.gov/gwastudies/ and http://gwas.lifesciencedb.jp/cgi-bin/gwasdb/gwas_top.cgi. One tool that is often used during the analysis is plink, available at http://pngu.mgh.harvard.edu/~purcell/plink/. For a more detailed review on how to interpret GWAS studies, see review by Pearson and Manolio (2008).

Adverse Events Databases

An adverse event database can help in studying side effects at a systems level. Using such a database, one can identify common side effects to various classes of drugs, and determine how drug combinations can be applied to reduce a particular side effect. An example of this kind of database is the adverse event reporting system database, available through the United States Food and Drugs Administration (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm).

Methods

Systems pharmacology uses both experimental and computational techniques. These methods act to complement each other in answering pharmacological questions in an integrated way.

Experimental Methods

An important part of the experimental methods is the collection of clinical data, which range from family medical history to medication histories. Such information allows the clinician and investigator to integrate them into the ▶ interactome and ▶ genome-wide association studies, which can then be used to give predictions on the potential effects of a drug.

Genomic analyses provide genetic signatures of disease. In particular, a genomic screen can help identify genotypes, which leads to a resistance to drug therapy. An example of this is the screening for estrogen receptor positive breast cancer, leading to the prescription of tamoxifen.

Proteomic analysis such as stable isotope labeling with amino acids (SILAC) (Pimienta et al. 2009) and drug screens provide a data-rich background for ▶ systems pharmacology approaches (Zhu and Cuozzo 2009). Additionally, model organisms such as fruit fly, worms, and zebra fish have offered methods of screening disease gene phenotypes. Furthermore, robotics and high-throughput microscopy imaging are also becoming increasingly available. Nevertheless, new biological components, such as microRNA, are emerging and new experimental techniques are required to identify how they may influence our existing knowledge base.

Beyond studying a disease of interest, experimental and clinical methods can also be used to monitor patient responses to therapy, providing feedback to how system pharmacological models need to be modified and refined for discovering rare unanticipated mechanisms.

Computational Methods

Systems pharmacology largely uses computational techniques that revolve around the interactome. Statistical methods are used to analyze network features. This type of analysis is called graph theory. Centralities, maximum flow, and eigenvalue analysis may be applied to deduce information about the network (West 2000). Specialized algorithms have been constructed to identify various network architectures. These algorithms can be found in tools which are available on various websites: genes2networks (http://actin.pharm.mssm.edu/genes2networks/) for identifying significant gene interactions, and cytoscape for visualization and analysis (http://www.cytoscape.org/). Commercial products are also available through ingenuity (http://www.ingenuity.com/) or metacore (http://www.genego.com/metacore.php). Network analysis–based computational tools can be used to analyze ▶ adverse events for correlations with influencing factors and can propose cellular and molecular mechanisms based on these correlations of ▶ adverse events. Such integrate analysis can potentially enable physicians to prescribe medicine in an evidence-driven manner.

Dynamical modeling is done through computations wherein the interactions are treated as a series of differential equations with rate constants (Hoppensteadt and Peskin 2010). These models can allow us to explore drug action as the drug is used for varying treatment periods. Tools for this are also available through various websites: virtual cell for dynamical stimulation of chemical and biological species, and stochsim (http://www.sys-bio.org/sbwWiki/sysbio/stochsim) for stochastic simulations of biochemical networks. If a custom algorithm is desired, one can find frameworks for developing such as the systems biology workbench (http://www.sys-bio.org/research/sbwIntro.htm). These dynamical modeling allows for integration of network models with the classical pharmacokinetic data of drug action.

Conclusion

▶ Systems pharmacology is a new field that is a branch of both systems biology and pharmacology. The applications of ▶ systems pharmacology offer methods for exploring various pharmacological areas, including pharmacokinetics, pharmacodynamics toxicology, drug design and discovery in an integrated manner. As a result, one of the goals of ▶ systems pharmacology is to bring evidence-based medicine to drug therapy. With the advent of electronic medical record systems and personalized genomics, ▶ systems pharmacology can be developed to help physicians treat patient optimally on an individual basis.

Cross-References

▶ Adverse Events
▶ Drug Target, Off-Target
▶ Evidence-based Medicine
▶ Genome-wide Association Study
▶ Interaction Networks
▶ Interactome
▶ Personalized Medicine
▶ Primary Target

References

El-Hamamsy I, Yacoub MH (2009) Cellular and molecular mechanisms of thoracic aortic aneurysms. Nat Rev Cardiol 6(12):771–786. doi:10.1038/nrcardio.2009.191, Nature Publishing Group

Hoppensteadt FC, Peskin CS (2010) Modeling and simulation in medicine and the life sciences. Texts in applied mathematics. Springer, New York, p 376. Retrieved from http://www.amazon.com/Modeling-Simulation-Medicine-Sciences-Mathematics/dp/1441928715

Pearson TA, Manolio TA (2008) How to interpret a genome-wide association study. J Am Med Assoc 299(11):1335–1344. doi:10.1001/jama.299.11.1335

Pimienta G, Chaerkady R, Pandey A (2009) Phospho-proteomics. In: de Graauw M (ed) Methods 527: 107–116. Humana Press, Totowa. doi:10.1007/978-1-60327-834-8

West DB (2000) Introduction to graph theory, 2nd edn. Prentice Hall, Upper Saddle River, p 470. Retrieved from http://www.amazon.com/Introduction-Graph-Theory-Douglas-West/dp/0130144002

Wist AD, Berger SI, Iyengar R (2009) Systems pharmacology and genome medicine: a future perspective. Genome Medicine 1(1):11. doi:10.1186/gm11

Zhu Z, Cuozzo J (2009) Review article: high-throughput affinity-based technologies for small-molecule drug discovery. Journal of Biomolecular Screening 14(10):1157–1164. doi:10.1177/1087057109350114

Systems Pharmacology, Drug Disease Interactions

Jean-Marc Schwartz[1] and Jose C. Nacher[2]
[1]Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, Manchester, UK
[2]Department of Information Science, Faculty of Science, Toho University, Funabashi, Chiba, Japan

Synonyms

Drug disease networks

Definition

Relationships between drugs and pathologies can be characterized in a comprehensive manner by interaction networks, giving a global view of drug disease interactions.

Characteristics

The relationships between drugs and diseases are complex. In the classical view, a particular drug used to be associated to a particular molecular target and to a particular treatment. However, cellular systems are highly interconnected by nature and this view is therefore incomplete. To obtain a comprehensive description of interactions between drugs, targets, and diseases, integrated approaches are needed where the relationships between all these components can be analyzed globally. Network approaches enable such analyses and have the potential to better explain the influence of drugs on diseases, as well as to understand interdependencies between drugs themselves (Spiro et al. 2008).

Network representations of drug–disease interactions necessitate classifications of drugs on the one side and of diseases on the other. Several classifications are available and have been used in different contexts.

Disease Classifications

The Mendelian Inheritance in Man (MIM) and its online version, the Online Mendelian Inheritance in Man (OMIM), is a comprehensive catalog of human genes and their associated phenotypes (Amberger et al. 2009). It details Mendelian traits associated to genetic disorders and includes links to literature references, sequences, and chromosomal localizations. The nomenclature of OMIM is mainly text based, making it complex for use in computational analyses. A subset of the database, the Morbid Map, therefore provides a more condensed view of the relationships between genes and diseases.

The World Health Organization (WHO) developed the International Statistical Classification of Diseases and Related Health Problems (ICD-10). It uses a hierarchical structure, where the first level classifies diseases into 22 main categories.

The Unified Medical Language System (UMLS) provides a series of controlled vocabularies for many areas of biomedicine, including the description of diseases. It is aimed at enabling the development and interoperability of computer systems to handle health-related information.

The UMLS system was used to develop the Disease Ontology (DO), where diseases are classified following a hierarchical structure (Osborne et al. 2009). This structure makes it possible to carry out enrichment analyses in a similar manner to the Gene Ontology (GO). For example, a set of overexpressed genes detected in high-throughput experiments can be queried against DO to reveal diseases that are significantly overrepresented in the data.

The Medical Subject Headings (MeSH) provide another controlled vocabulary for medical terms, which are organized in a hierarchical structure. They include disease terms and have been used to map genes to diseases.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) contains a database of diseases, KEGG DISEASE (Kanehisa et al. 2010). They developed their own classification system based on knowledge of genetic and environmental perturbations, but also reference diseases by their ICD-10 Disease Classification.

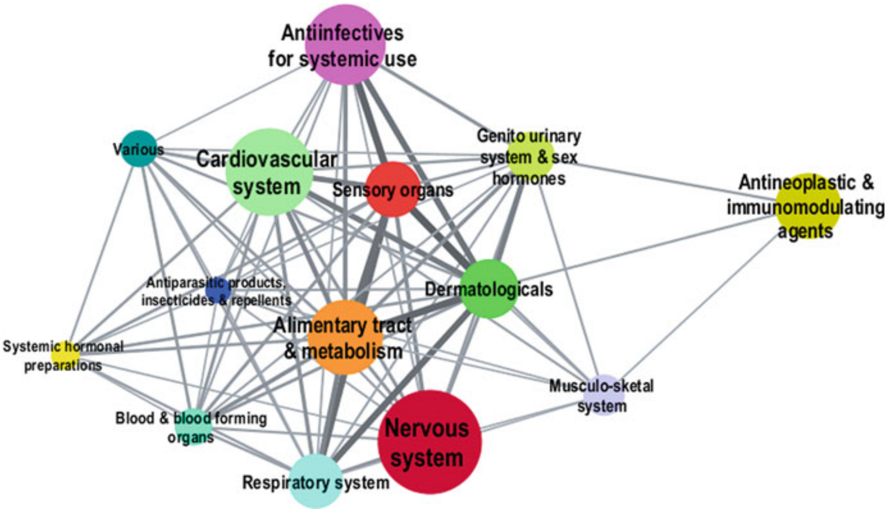
Drug Classifications

The Anatomical Therapeutic Chemical (ATC) classification system classes drugs according to their therapeutic properties or the organ on which they act. The ATC system is developed by the World Health Organization (WHO 2010). ATC codes have a hierarchical structure composed of five levels, representing increasingly detailed levels of anatomical and therapeutic properties. The first level comprises 14 main anatomical groups; the second level represents a therapeutic subgroup; the third level represents a pharmacological subgroup; the fourth level distinguishes between chemical subgroups, and the fifth level identifies the chemical substance itself. A drug can be assigned multiple ATC codes if it possesses multiple therapeutic applications or routes of administration.

The DrugBank database is a comprehensive repository of approved and experimental drugs. It contains detailed information on the chemical, pharmaceutical, and pharmacological properties of drugs, including their associated ATC codes, as well as drug target data (Wishart et al. 2008).

The Approved Drug Products with Therapeutic Equivalence Evaluations, commonly known as the Orange Book, lists drugs approved by the US Food and Drug Administration (FDA) with their therapeutic equivalence evaluations.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) contains the KEGG DRUG database,


Antiinfectives for systemic use
Various
Cardiovascular system
Sensory organs
Genito urinary system & sex hormones
Antineoplastic & immunomodulating agents
Antiparasitic products, insecticides & repellents
Dermatologicals
Alimentary tract & metabolism
Systemic hormonal preparations
Musculo-sketal system
Blood & blood forming organs
Nervous system
Respiratory system

Systems Pharmacology, Drug Disease Interactions, Fig. 1 First ATC level of the therapy projection of a drug–therapy bipartite network. All nodes in the network represent therapeutic classes as defined in the ATC classification; the name associated to each node is the first-level ATC heading. The node size is proportional to the number of drugs in each therapy class; the edge thickness is proportional to the number of common drugs involved in each pair of therapies

a unified database of drugs approved in Japan, USA, and Europe (Kanehisa et al. 2010). Drugs are represented with details of their chemical structure and classified according to ATC codes as well as to the Therapeutic Category of Drugs in Japan, a system derived from the Japan Standard Commodity Classification.

Integrated Representations

Network representations are efficient tools to combine heterogeneous information and reveal patterns of connections between different types of data. Paolini et al. (2006) integrated several pharmacological resources to construct a global mapping of drugs and human targets (▶ Systems Pharmacology; ▶ Drug Target), including drug indications indexed by a disease code. This approach reveals a high level of promiscuity in the pharmacological space.

A bipartite graph of interactions between human genes and associated disorders was constructed. In this representation, one set of nodes represents genetic disorders and the other represents disease-related genes (Goh et al. 2007). Two projections can be obtained from this bipartite graph; in the disease projection, two diseases are connected if there is a common gene involved in both; in the gene projection, two genes are connected if they are associated to a common disease. Several types of cancers, including colon cancer and breast cancer, appear as hubs in the disease projection which are genetically connected to a large number of other disorders. However a majority of disease genes are found to be nonessential, showing no tendency to encode hub proteins and occupying more peripheral positions.

Interactions between drugs and associated therapies are revealed by constructing a bipartite graph whose nodes are either drugs (D) or therapies (T), such that each edge connects a node in D and a node in T (Nacher and Schwartz 2008). This bipartite graph can be decomposed into two projections; in the drug projection, two nodes from D are connected if a common therapy is involved in both of them; in the therapy projection, two nodes from T are connected if a drug is implicated in both of them. When the hierarchical ATC classification of therapies is used, these networks can furthermore be represented at different anatomical and therapeutic levels. The first ATC level shows that the therapeutic space is fully connected, revealing unexpected links between several areas of therapeutic applications (Fig. 1).

The importance of particular drugs in the drug–therapy network can be assessed by computing measures of network centrality (▶ Network Metrics). Drugs with a high betweenness centrality constitute key bottleneck connections between distinct classes of therapies; these include scopolamine, morphine, tretinoin, tolbutamide, among others.

Cross-References

▶ Drug Target
▶ Systems Pharmacology

References

Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's online Mendelian inheritance in man (OMIM). Nucleic Acids Res 37:D793–D796

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL (2007) The human disease network. Proc Natl Acad Sci USA 104:8685–8690

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38:D355–D360

Nacher JC, Schwartz JM (2008) A global view of drug-therapy interactions. BMC Pharmacol 8:5

Osborne J, Flatow J, Holko M, Lin S, Kibbe W, Zhu L, Danila MI, Feng G, Chisholm RL (2009) Annotating the human genome with Disease Ontology. BMC Genomics 10:S6

Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nat Biotechnol 24:805–815

Spiro Z, Kovacs IA, Csermely P (2008) Drug-therapy networks and the prediction of novel drug targets. J Biol 7:20

WHO Collaborating Centre for Drug Statistics Methodology (2010) ATC classification index with DDDs, Oslo, Norway

Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) DrugBank: a knowledgebase of drugs, drug actions and drug targets. Nucleic Acids Res 36:D901–D906

Systems Pharmacology, Drug-Target Networks

Jose C. Nacher[1] and Jean-Marc Schwartz[2]

[1]Department of Information Science, Faculty of Science, Toho University, Funabashi, Chiba, Japan

[2]Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, Manchester, UK

Definition

Relationships between drugs and their molecular targets can be characterized in a comprehensive manner using interaction networks.

Characteristics

For many years, the dominant approach of drug design has focused on the search for highly selective ligands targeting a specific disease-causing agent. However, this reductionist approach is now being questioned as many drugs are failing in late clinical development stages. It is estimated that as much as 30% of newly developed drugs fail due to a lack of efficacy and a similar rate due to harmful side effects (Hopkins 2008).

Systems biology has challenged this view by raising awareness that no component or process is isolated in biological systems. This principle applies to diseases as well, since many genes are related to more than one disease, and the pharmacology of drug-disease interactions reveals intricate connections between heterogeneous classes of therapies (▶ Systems Pharmacology, Drug Disease Interactions).

A new approach to drug discovery is therefore emerging, which has been termed "polypharmacology." It requires a comprehensive description of the relationships between diseases and molecular biological components, and aims to identify the best combinations of targets to achieve a desired therapeutic effect. In this context, network pharmacology is expected to provide valuable information by relating the topological properties of potential targets to their biological functions and assisting in the development of new therapeutic strategies.

Drug-Target Databases

Recently, thanks to an increasing number of available databases, many systems have been described in terms of networks where individual nodes are connected by specific relationships or interactions. Interestingly, the drug-target system can be investigated along a similar approach and interaction data are available in several databases.

The DrugBank database is a bioinformatics-chemoinformatics resource that combines drug data with comprehensive drug-target information with over 4,900 drug entries. However, it often occurs that to find specific information about drugs, ligands, therapies, and related disease categories, it is necessary to navigate through different databases. A knowledge base of human and genetic disorders can be found at the Online Mendelian Inheritance in Man (OMIM). The KEGG database has also devoted a large part of its storage categories to drugs and diseases. Associated therapeutic properties of each drug are classified using the Anatomic Therapeutic Chemical (ATC) classification.

Drug-Target Networks

A rich variety of complex networks can be constructed in systems pharmacology (Spiro et al. 2008; Berger and Iyengar 2009). At different complexity levels, from top to bottom, patient records can be used to define individual patient-drug target networks. This information can be closely related to features of phenotype variability of complex disorders. Similarly, data related to patient symptoms can lead to symptom-drug interactions. These networks can be entirely represented using bipartite graphs, but still remain largely unexplored.

The nodes in a bipartite graph can be divided into two disjoint sets such that each edge links nodes from different sets. The bipartite graph representation has been the framework used in most of the network analyses carried out to study drug-target networks. In network pharmacology, drugs and targets define the two disjoint sets used to define bipartite graphs structures.

By going into higher levels of complexity, connections between therapies and specific drugs also define bipartite networks. Key drugs that connect distinct classes of therapies in a few steps could be identified (Nacher and Schwartz 2008). Moreover, each therapy is linked to specific drug targets. Complex networks defined by drug targets and drug interactions represent a higher level of detail to study polypharmacology phenomena.

A bipartite network connecting drug targets and drugs made it possible to analyze targets in the context of a global protein-protein interaction network (Ma'ayan et al. 2007) or in its projections (Yıldırım et al. 2007). Network projection is a general technique in graph theory that allows us to transform a bipartite graph into two simple graphs. In the drug network projection, each node represents a drug and two drugs are linked to each other if they share one or more drug targets. On the other hand, in the drug-target network projection, nodes represent proteins or gene targets. Two targets are linked if they share at least one common drug. Both projections revealed a power-law decay for the node degree, highlighting the heterogeneous character of these networks. This topology is governed by the statistics of hubs. Although the probability of finding a hub in a scale-free network is very low, each hub tends to have many links. Projections of drug and drug-target networks follow the same principles, containing a few drugs (proteins) that share many drug targets (drugs).

Moreover, drug targets can be linked to disease-gene products and this information can be mapped back to the protein-protein interaction networks (PPI). The human disease network was collected from OMIM-based disorder-disease gene associations (Goh et al. 2007). Etiological and palliative properties can then be investigated using network metrics (▶ Network Metrics). The shortest path then estimated the number of molecular steps that separate a drug target from the corresponding disease cause (Yıldırım et al. 2007). Enhancement of the distribution was observed for short distances when compared to random groups of proteins. This result suggested a dominance of palliative drugs. Network analysis also highlighted the recent trend toward more rational drug design thanks to the increased knowledge on complex disorders. A similar shortest path analysis was conducted between networks composed of drugs approved in the last 10 years and before 1996. The results indicated a higher frequency in the distribution of shorter paths indicating a more rational drug design. They also showed the importance of developing new multi-target drugs that shorten the routes in the drug-target network. The study of drug targets from a systemic point of view has not been limited to bipartite network approaches. For example, a global mapping of pharmacological space enabled the identification of human targets for which chemical tools and drugs have been discovered to date.

On the other hand, the distribution of targets associated to approved drugs has been reported to follow a power law. It shows that a large proportion of drugs can act on only one target but a few of them can act on multiple targets. This fact can be used to develop multi-target drugs which already showed promising results. Complementary approaches may consist in developing drugs which combine the action on multiple targets with topological features in the drug-target network space, such as, for example, high centrality values (Nacher and Schwartz 2008).

Recent studies have started working on networked interactions of drugs with single complex diseases or subtypes of the same diseases like cancer (Dalkic et al. 2010). Network analyses of drug and mutation targets led to new insights on how drugs are shared between cancer types and vice versa, since some cancer variants share drug targets but not mutation target. This finding suggests that new drug targets or mutation targets could be discovered for these cancers.

Beyond purely network-based analysis, data integration based on phenotypic and chemical indexes in pharmacological space and protein interaction networks has also been suggested as a way to identify new drug targets as well as to find new applications for existing drugs (Shiwen and Li 2010). First, drug therapeutic similarity among drugs that belong to ATC categories was defined using a probabilistic model. The chemical similarity was derived using the Tanimoto coefficient. On the other hand drug-protein closeness and its genomic relatedness were computed as an exponential functional that decays with the square of the shortest distance between a given protein p and the p_k targets of the protein in the protein-protein interaction networks. By computing the three possible linear correlations among these three metrics, concordance scores between a given protein and drug could be obtained. The results showed that unexpected drug-drug interactions emerged in more than 500 cases, highlighting new possible applications. This approach perfectly illustrates the network pharmacology concept (Hopkins 2008) where the network framework is naturally embedded in the analysis. Drug-protein closeness is computed thanks to the shortest path in the protein interaction network and the knowledge of drug-protein interactions.

Drug Scopes

A complementary approach to investigate the relationships between drugs and metabolic systems is offered by the metabolic drug scope (▶ Drug Scope, Metabolic). The drug scope represents the largest possible set of compounds that a drug can influence in a metabolic system, when only metabolic network connections are taken into account, irrespective of kinetic or thermodynamic laws (Schwartz and Nacher 2009). A systematic investigation of the scopes of drugs from the DrugBank database revealed that drugs can be classified into different categories, where some of them have small scopes corresponding to localized action, while others have large scopes corresponding to potential widespread systemic action. These different categories are furthermore associated to distinct classes of therapies. This approach opens new possibilities to determine appropriate sets of targets aimed at achieving a desired therapeutic effect.

Cross-References

- ▶ Drug Scope, Metabolic
- ▶ Network Metrics
- ▶ Systems Pharmacology, Drug Disease Interactions

References

Berger SI, Iyengar R (2009) Network analyses in systems pharmacology. Bioinformatics 25:2466–2472

Dalkic E, Wang X, Wright W, Chan C (2010) Cancer-drugs associations: a complex system. PLoS One 5:e10031

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL (2007) The human disease network. Proc Natl Acad Sci USA 104:8685–8690

Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4:682–690

Ma'ayan A, Jenkins SL, Goldfarb J, Iyengar R (2007) Network analysis of FDA approved drugs and their targets. Mt Sinai J Med 74:27–32

Nacher JC, Schwartz JM (2008) A global view of drug-therapy interactions. BMC Pharmacol 8:5

Schwartz JM, Nacher JC (2009) Local and global modes of drug action in biochemical networks. BMC Chem Biol 9:4

Shiwen Z, Li S (2010) Network-based relating pharmacological and genomic spaces for drug target identification. PLoS One 5:e11764

Spiro Z, Kovacs IA, Csermely P (2008) Drug therapy networks and the prediction of novel drug targets. J Biol 7:20

Yıldırım MA, Goh KI, Cusick ME, Barabási AL, Vidal M (2007) Drug-target network. Nat Biotechnol 25:1119–1126

Systems Vaccinology

- ▶ Systems Immunology, Novel Evaluation of Vaccine
- ▶ Vaccinomics

Systems Virology

Narsis Aftab Kiani[1] and Lars Kaderali[2]
[1]ViroQuant Research Group Modeling, BioQuant, University of Heidelberg, Heidelberg, Germany
[2]Institute for Medical Informatics and Biometry Technical University, Dresden, Germany

Synonyms

Systems biology of viral pathogens; Systems biology of virus–host interactions

Definition

Systems Virology can be defined as the application of systems biology approaches and methods to the field of virology. The aim of Systems Virology is to develop a system-level understanding of viral infection, by focusing on the dynamic interplay between virus and host. A hallmark of Systems Virology approaches is the use of quantitative and dynamic mathematical models to simulate and predict viral interactions with the host, to develop a systems-level understanding of these interactions, and ultimately to identify new potential targets for antiviral drugs using model analysis. Systems Virology thus aims at the development and analysis of models of the influence of virus infection on cellular signaling pathways, of the role of specific viral and host genes in all steps of the viral life cycle, and of the host response to viral infection.

There is no clear boundary between systems biology and systems virology, the latter could be characterized as applied systems biology in the field of virology. While the central aim of systems biology is the development of a fundamental understanding of biological networks in general, systems virology ultimately aims at predictive and therapeutic applications (Clermont et al. 2009).

Characteristics

Viruses are ideal systems to argue for the necessity of systems approaches, since they depend on host processes for almost every single step in their life cycle. Surprisingly, in spite of the fact that viruses encode a relatively small number of genes, virus–host interactions lead to extremely complex molecular interaction networks, encompassing both viral and host processes, and which are far from being well understood. For example, hundreds of different host factors have been implicated in the HIV life cycle using RNA interference, with most of them only poorly annotated and no obvious direct role in the infection process. Correspondingly, antiviral drug design remains an extremely challenging goal and faces serious obstacles, as witnessed by the tendency of viruses to quickly develop resistance against many antiviral drugs, failure to develop efficient vaccines for many viruses, and only slow development of novel antiviral drugs, guided often by trial and error instead of a systematic drug design.

Interestingly, most of the currently available antiviral drugs target virus-encoded enzymes that are essential for successful viral replication. Examples include transcriptase or protease inhibitors, which are used, for example, in combination therapy against HIV. Such drugs targeting viral enzymes are often subject to the quick development of viral resistance, due to high mutation rates in particular of RNA viruses. To circumvent development of resistance, an alternative treatment strategy is to target host processes, for example, by stimulating or manipulating the host viral response, or by targeting host factors required by the virus in its life cycle (Tan et al. 2007). Examples of compounds targeting host factors include Maraviroc, which inhibits a cofactor required by HIV to enter cells, or pegylated interferon and ribavirin, which are nonspecific antiviral drugs used in treatment of hepatitis C. Such a strategy may be more successful for several reasons: (1) Resistance might be less of a problem since the virus would have to replace the entire host process and thus evolve a far more complex strategy to evade the drug. (2) It may become possible to develop antiviral drugs with broad applicability against many different viruses, if the same host process is exploited by different viruses. (3) Many more potential drug targets become available, when targeting host factors instead of one of maybe only a few dozen viral proteins. On the downside, side effects are already a problem for drugs targeting viral enzymes, and this problem may clearly be aggravated if host processes are being targeted. On the other hand, a large number of small molecules exist for the treatment of many diseases with manageable side effects, and a systems-level understanding of virus–host interactions will be of considerable utility in the development of safe and efficient antiviral drugs targeting host processes instead of viral enzymes (Tan et al. 2007).

The development of such drugs will require a fundamental understanding of viral infection, and therefore, demands a study of the dynamic interplay between the virus and its host. Systems virology attempts to fill this gap by focusing not just on the pathogen alone, but by integrating models of the virus and viral processes with processes in the host cell, thus achieving a systems view. Systems Virology thus aims at the development and analysis of models of the influence of virus infection on cellular signaling pathways, of the role of specific viral and host genes in all steps of the viral life cycle, and of the host response to viral infection. The aim of systems virology is the identification and characterization of key network components or connections, and their interplay in the virus–host interaction network as a whole. Using model analysis, systems virology aims to identify load- and choke points of viral infection and replication processes, which can be used as potential new targets for antiviral drug design. Ultimately, the promise of systems virology is to provide profound knowledge about the complex virus–host system, and to translate this knowledge into predictive, preventive, and personalized medicine to combat viral infection.

To achieve these objectives at a systems level, large-scale experimental data sets are required. Systems virology, therefore, benefits greatly from major advances in molecular virology and from the development of high-throughput experimental techniques and associated data processing and analysis methods in the recent years. These include microarray-based functional genomics, high-throughput and high-content siRNA screening, live cell imaging, high-throughput protein interaction measurements using Yeast-2-Hybrid screens, automated mass spectrometry and protein arrays, and next generation sequencing (Peng et al. 2009). These technological developments are paralleled by novel developments in data processing, data integration, and data analysis techniques in the fields of statistical data analysis, bioinformatics, data mining, and machine learning, which are employed to reconstruct virus–host interaction networks and develop a basis for more detailed, quantitative, and dynamic models of virus–host interactions.

Systems virology typically proceeds in an iterative cycle, consisting of systematic and large-scale perturbation of individual entities in the virus–host system, measuring the outcome using high-throughput technologies, and then trying to relate the change at the molecular level to global properties of the system during the infection, using modeling and simulation, followed by the design of further experiments to fill the knowledge gap highlighted by the difference between the model simulation and the real system (Kitano et al. 2002). As an example strategy, large-scale siRNA screens to identify new host factors involved in viral replication are followed by live cell imaging and more detailed biochemical characterization of identified host processes to develop quantitative, dynamic models of HIV and HCV infection at Heidelberg University, laying the basis for computational modeling and model analysis of virus–host systems. Modeling and data analysis are then carried out using a combination of machine learning approaches for the data-driven reconstruction of virus–host networks, bioinformatics annotation and database queries, and forward modeling using knowledge-based approaches and based on differential equations. Ultimately, all these approaches are mapped onto one, virus and host cell type–specific, integrated model of virus–host interactions. Such models integrating viral and host processes can then be used to identify critical points in the infection cycle, to design new drugs with optimal efficiency and minimizing side effects, and to gain a better understanding of host immune response and thus vaccines development.

Cross-References

- ▶ Systems Immunology

References

Clermont G, Auffray C, Moreau Y, Rocke DM, Dalevi D, Dubhashi D, Marshall DR, Raasch P, Dehne F, Provero P, Tegner J, Aronow BJ, Langston MA, Benson M (2009) Bridging the gap between systems biology and medicine. Genome Med 1:88.1–88.3

Kitano H (2002) Systems biology: a brief overview. Science 295:1662–1664

Peng X, Chan EY, Li Y, Diamond DL, Korth MJ, Katze MG (2009) Virus-host interactions: from systems biology to translational research. Curr Opin Microbiol 12:432–438

Tan S, Ganji G, Paeper B, Proll S, Katze MG (2007) Systems biology and the host response to viral infection. Nat Biotechnol 25:1383–1389

Systems, Autopoietic

Leonardo Bich and Arantza Etxeberria
Department of Logic and Philosophy of Science,
IAS-Research Center for Life, Mind, and Society,
University of the Basque Country, UPV/EHU,
Donostia-San Sebastián, Spain

Definition

The authors' definition of the autopoietic system has evolved through the years. One of them states that *an autopoietic system is organized (defined as a unity) as a network of processes of production (transformation and destruction) of components that produces the components which: (1) through their interactions and transformations regenerate and realize the network of processes (relations) that produced them; and (2) constitute it (the machine) as a concrete unity in the space in which they exist by specifying the topological domain of its realization as such a network* (Varela 1979, p. 13). Nearly the same formula was earlier used to define an *autopoietic machine* (Maturana and Varela 1973/1980, 1984/1987, p. 135).

Characteristics

The Chilean biologists H. Maturana and F. Varela proposed the term *autopoiesis* in the early 1970s to account for the organization of individual living beings, characterized as a process by which they produce their own identity in a mechanistic way. The autopoietic approach to life is very different from that of the Theory of Evolution and Molecular Biology: On the one hand, instead of reproduction or evolution, the theory focuses on autonomy and identity to naturalize them as marks of life; on the other hand, it considers that all system components have the same status to explain the self-referent dynamics by which they produce a unity; that is to say, living phenomenology is not explained in terms of some components being information carriers.

Autopoietic systems, also called initially *autopoietic machines*, explore the general relational scheme common to all living systems as the configuration of transformative processes whose result is the configuration itself, so that identity and activity, producer and product coincide. Unlike *Turing machines*, set by external programmers (thus being *heteropoietic*) to compute problems referring to issues other than the system itself (thus being *allopoietic*), *autopoietic machines* realize a self-defined identity in a space of interactions. Already in 1974 (Varela et al. 1974), the authors presented their account of living organization with a computational model in cellular automata which was later rehearsed by Barry McMullin (in Di Paolo 2004).

Some of these distinctions, for example, between autopoietic and heteropoietic, already appear in Canguilhem's *La connaisance de la vie*. In fact, the autopoietic approach belongs to a systemic tradition focused on the problem of the relational unity of the living, associated to Kant's understanding of organisms in the *Critique of Judgment*, Claude Bernard's concept of *milieu intérieur*, and the organicist tradition that considers life as organization (G. Canguilhem, H. Jonas, J. Piaget among others, see Weber and Varela 2002), and opposed to the mainstream of the time, such as some of the views of Jacob´s *La logique du vivant*. Other clear associations are with the cybernetic movement, especially with second-order cybernetics. The influence of the autopoietic approach has been significant in theoretical Biology (especially on work on the definition of life and origins and organization of minimal living systems), Artificial Life, and Cognitive Science. In contrast, it has had no comparable effect on mainstream biology (e.g., Molecular and Evolutionary Biology), although it appears to be more present in Systems Biology, whose approach is less centered on master molecules and information.

The Main Conceptual Development

Autopoietic systems aim to grasp what makes an organism be a unity of a specific kind, that is to say, how a system appears out of a continuous flux of transformations at the level of its components.

The system is characterized by its *organizational closure*, a notion that provides a reinterpretation of the cybernetic notion of *circular self-stabilization*, which instead of considering single regulatory processes in isolation and then coupling them together (as homeostatic machines, acting on internal variables, behave) refers to the whole living system: The autopoietic system is organized in such a way that it does not only maintain the interval of stability of some variables, but also the global organization is kept invariant.

Some of the main concepts of the theory refer to distinctions, such as the following:

- *Organization and structure*: This emphasizes that an organism is not characterized by its material or physicochemical processes, but by how the interactions are related to produce and maintain the integrated biological unity they belong to. The *structure* refers to the variant aspect of a living system: to its physical realization, whereas the notion of *organization* aims to grasp the invariant one: the topology of the relations that constitute it. Thus, the authors embrace a particular form of *multiple* *realizability* between organization and structure, as the autopoietic organization is proposed as a main invariant underlying the diverse biological phenomenology, that is conserved through the ontogenetic and phylogenetic changes.

- *Openness and closure*: Whereas living systems are open to the exchange of matter and energy at the level of structure, the network of processes that constitutes their organization is closed in the form of a global cyclical process that determines and regenerates itself. Rosen developed a similar view independently and expressed it mathematically in the notions of the system being *open to material causation* and *closed to efficient causation* (Letelier et al. 2006). The distinction between open structure and closed organization can be also found in Piaget's *Biologie et connaissance*, complemented by an internal mechanism of adaptation to perturbations in terms of Waddington's *assimilation* and *accommodation*.

Another characteristic feature of the theory is its internalism, present through the notion of *structural determinism*. In each time step, the system interacts and changes in a way totally determined by its structure, which specifies the set of all possible changes to effective perturbations. The latter do not define, but only trigger structural changes. Thus, environmental perturbations do not have intrinsic meaning, their effect depends on the structure of the receiver: Unlike in input-output relations, the same stimulus can cause different alterations. F. Varela (in Varela et al. 1991) showed this peculiarity through a cellular automata model called Bitorio. Similar to this is the idea that in the communication between two systems, there is no transmission of information but a *structural coupling*.

In this framework, evolution is reinterpreted in neutralist terms as a natural drift. The idea of adaptation as optimization of the organism's traits by natural selection is replaced by one of conservation of adaptation, as the maintenance of a specific form of coupling between the living system and its environment (Maturana and Varela 1984).

Further Developments

Developments of the autopoietic theory have particularly been connected with the definition of life and autonomy and with agency and cognition. Finally, some have tried, without success so far, to extend the notion of autopoiesis from the cellular level to that of multicellular organisms and social systems.

- *Definition of life as autonomy*. The main influence of autopoietic systems has been in fields related to the definition of life and its organization, such as *Artificial Life*, *Synthetic Biology*, *Astrobiology* or, in general, *Systems Biology*. The main impact of the autopoietic theory in these areas has been through the notion of autonomy as an ingredient of the definition of life.

  The goals of the initial approach to *Artificial Life* were congenial to the theory of autopoietic systems in the significance of form above matter, but very different in what concerns the nature of life, which was there thought to be connected to reproduction and evolution by the mainstream, not to autonomy or organization as the autopoietic theory maintains. Nevertheless, for some authors, it is problematic to consider the operations of the living only at a formal abstract level, without considering the complexities of material and historical realizations of life as we know it. For example, the formal account of autonomy fails to meet the thermodynamic criteria required to realistically maintain the state of activity of any candidate system in its environment, and this has been one of the main developments of the original theory by researchers who, accepting the relevance of autonomy, would not want to explore it only in formal models but related to material constraints.

  Similarly, in the *Origins of Life* field, the theory of autopoiesis has been particularly influential among those pursuing the cellular origins of life (as opposed to molecular origins) in the generation of self-maintaining and self-reproducing systems (Luisi 2006).

  In *Systems Biology*, autopoietic theory has revealed itself promising as a theoretical guideline in developing a notion of system as a integrated unity, in modeling the cellular metabolism as a closed and intertwined network of processes, in reinterpreting the role of the genomes in the cell in a more ecological fashion, and in pointing out the relevance of self-regulation at different hierarchical levels (Boogerd et al. 2007).

- *Agency and cognition.* From the autopoietic perspective, cognition is the system's capability to provide meaning to the world, a property connected, if not coincident, with life. An increasingly relevant issue raised in the investigation of cognition is the one concerning how to characterize the specific mechanism of self-maintenance instantiated by biological metabolism in its basic form, being the notion of self-production insufficient to account for agency as the ability to act in the environment. There have been proposals to expand the definition of self-production through the introduction of active mechanisms of self-regulation.

With respect to the impact on the study of cognition in the conventional sense, autopoietic theory has provided an analysis of the biological roots of knowledge by considering human observers as structurally determined systems. In doing so it has pointed out the limits of the notions of representation and objectivity and contributed to the development of an epistemological perspective known as "radical constructivism" according to which the natural world emerges as coherences in the coupling between the observer and its medium. In cognitive sciences, the autopoietic theory has pointed to the need to develop embodied and situated accounts to characterize autonomous agents, by inaugurating the so-called *enactive approach* (Varela et al. 1991).

- *Other levels of organization*. As autopoietic systems define life at the cellular level, multicellular living systems and social ones – respectively defined as autopoietic systems of the second and third order – are considered as derivative, even if not trivially, with respect to the properties of cellular ones. But satisfactory criteria for this operation of expansion of the theory have not been provided in the original formulations.

In spite of these acknowledged difficulties, the notion of autopoiesis brings forth a relevant scenario for inquiry about the nature of life, providing an intuitive idea of what it means to be alive, autonomy, which is lacking in other approaches.

Acknowledgments LB has a *Juan de la Cierva Research Contract ((MICINN/UPV-EHU). AE's research was funded by the MICINN (FFI2008-06348-C02-01/FISO & FFI2008-06348-C02). AE's and LB's research was funded by the Basque Government grant (IT 505-10).*

Cross-References

- ▶ Autonomy
- ▶ Closure, Causal
- ▶ Organization
- ▶ Self-Organization
- ▶ Synthetic Biology, Predictability and Reliability

References

Boogerd FC, Bruggerman FJ, Hofmeyr J-HS, Westerhoff HV (eds) (2007) Systems biology. Philosophical foundations. Elsevier, Amsterdam

Di Paolo EA (2004) Special issue on: unbinding biological autonomy: Francisco Varela's contributions to artificial life. Artificial Life 10:231–360

Letelier J-C, Soto-Andrade J, Guinez-Abarzua F, Cardenas M-L, Cornish-Bowden A (2006) Organizational invariance and metabolic closure: analysis in terms of (M, R) systems. Journal of Theoretical Biology 238:949–961

Luisi PL (2006) The emergence of life. From chemical origins to synthetic biology. Cambridge University Press, New York

Maturana M, Varela F (1973/1980) De máquinas y seres vivos. Autopoiesis: La organización de lo vivo. Editorial Universitaria, Santiago. (English translation: Autopoiesis and cognition. The realization of the living. Reidel, Boston)

Maturana H, Varela F (1984/1987) El árbol del conocimiento. Editorial Universitaria, Santiago de Chile. (English translation: The tree of knowledge. Shambhala, Boston)

Varela F (1979) Principles of biological autonomy. Elsevier North Holland, New York

Varela F, Maturana H, Uribe R (1974) Autopoiesis: the organization of living systems, its characterization and a model. Biosystems 5:187–196

Varela F, Thompson E, Rosch E (1991) The embodied mind. Cognitive science and human experience. MIT Press, Cambridge, MA

Weber A, Varela FJ (2002) Life after kant: natural purposes and the autopoietic foundations of biological individuality. Phenomenology and the Cognitive Sciences 1(2):97–125