D

D Gene

► Diversity (D) Gene

DAH

► Differential Adhesion Hypothesis

Damage-associated Molecular Patterns

Harpreet Singh
Computational Biology Service Unit, Cornell University, Ithaca, NY, USA
Biomedical Informatics Centre, Indian Council of Medical Research, New Delhi, India

Synonyms

Alarmins; DAMP

Definition

Damage associated molecular patterns (DAMPs) are molecules released by stressed cells. Interactions between PRRs and DAMP initiate and perpetuate immune response similar to pathogen-associated molecular pattern molecules (PAMPs) that drive initiation and perpetuation of the inflammatory response (Janeway 1989). Many DAMPs are nuclear or cytosolic proteins with defined intracellular function that, when released outside the cell following tissue injury, move from a reducing to an oxidizing milieu resulting in their functional denaturation (Rubartelli and Lotze 2007). Also, following necrosis (a kind of cell death), tumor DNA is released into the extranuclear space/extracellular microenvironment and functions as a DAMP (Farkas et al. 2007).

Protein DAMPs include intracellular proteins, such as heat-shock proteins or HMGB1 (high-mobility group box 1), and proteins derived from the extracellular matrix that are generated following tissue injury, such as hyaluronan fragments. Examples of nonprotein DAMPs include ATP, uric acid, heparin sulfate, and DNA.

Cross-References

► Microbe-associated Molecular Pattern
► Pattern Recognition Receptors

References

Farkas AM, Kilgore TM, Lotze MT (2007) Detecting DNA: getting and begetting cancer. Curr Opin Investig Drugs 8(12):981–986
Janeway C (1989) Immunogenicity signals 1,2,3 ... and 0. Immunol Today 10(9):283–286
Rubartelli A, Lotze MT (2007) Inside, outside, upside down: damage-associated molecular-pattern molecules (DAMPs) and redox. Trends Immunol 28(10):429–436

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

DAMP

▶ Damage-associated Molecular Patterns

Data and Model Management Platform, SEEK

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

Definition

The ▶ SysMO-DB project in the ▶ SysMO consortium is developing a general platform to enable the sharing and exchange of research data/models/processes in systems biology consortia. This platform is known as the SEEK (1) and emphasis is placed on conserving and increasing the reusability of research outputs. The SEEK provides an index of consortium resources and acts as a gateway to other tools and services. Examples include integration with the ▶ JWS Online model repository to enable model simulations, and the PubMed plugin that allows publications to be linked to supporting data and author profiles. This transforms the SEEK from a static repository to an active, dynamic resource.

Characteristics

The SEEK is accessible via a web-based client and built on an open source philosophy. It has been adopted by projects outside the ▶ SysMO consortium, including The Virtual Liver project (http://seek.virtuelle-leber.de/), EraSysBio + (http://www.erasysbio.net), and UniCellSys (http://www.unicellsys.eu/) by the time of writing. Development of the SEEK follows a rapid, incremental cycle that is strongly user oriented by virtue of frequent interactions, in site visits and workshops with a focus group of users, the ▶ SysMO-PALS (Project Area Liasons), who are representatives from each ▶ SysMO project and are a mixture of experimentalists, modelers, and informaticians. The SEEK is also actively codeveloped by one of its first adopters, the Virtual Liver project, and is available as a free-standing platform (http://www.sysmo-db.org/) for academic research groups.

The SEEK encourages the use of community standards by providing tools to assist with data management and annotation, as data exchange and reuse rely on sufficient annotation, consistent metadata descriptions, and the use of standard exchange formats for models, data, and the experiments that they are derived from. The SEEK makes use of its own set of minimum information models for each data type, known as JERMs (Just Enough Results Model). The JERMs are derived from the minimum information models established by MIBBI and are available as spreadsheet templates. As entries in such templates often include items from controlled vocabularies, such as ontologies, the ▶ SysMO-DB team also developed an open source tool called RightField, which makes dynamic browsing of ontologies and embedding of their elements possible within JERM templates.

The SEEK's recommended model format is SBML (▶ Systems Biology Markup Language (SBML)), which is used by the JWS Online model repository and simulator, that has been a part of the SEEK since its inception. The aim of this integration is to provide researchers and modelers with easy access to modeling standards such as ▶ Systems Biology Markup Language (SBML), ▶ SBGN, and ▶ MIRIAM compliant model annotation.

Cross-References

▶ JWS Online
▶ MIRIAM
▶ SBGN
▶ SysMO
▶ Systems Biology Markup Language (SBML)

References

Wolstencroft K, Owen S, du Preez FB, Krebs O, Mueller W, Goble C, Snoep JL (2011) The SEEK: a platform for sharing data and models in systems biology. Methods Enzymol 500:629–655, Elsevier, Amsterdam

Data Collection

► Data Integration

Data Collection (Integration) from Distributed Sources

► Distributed Data Access

Data Deluge

► Data-intensive Research

Data Integration

Roberta Alfieri and Luciano Milanesi
Institute for Biomedical Technologies – CNR (Consiglio Nazionale delle Ricerche), Segrate, Milan, Italy

Synonyms

Data collection; Data warehouse

Definition

In systems biology, data integration is an important approach to better understand the main features of a biological process, because it represents a way to combine interesting information related to the reaction involved in specific network. One possible method to develop an integration system is the data warehousing approach (Stein 2003), which allows the integration of information stored in different biological databases. The necessity for data integration is widely approved in the bioinformatics and systems biology community since bioinformatics data are currently spread across different databases and they are stored in a wide variety of formats. Moreover, the achievement of interesting results in most bioinformatics and systems biology–related activities, from functional characterization of genomic and proteomic data to the development of mathematical models of biological processes, requires an integrated view of all relevant data useful to accomplish those tasks.

Cross-References

► Cell Cycle Database

References

Stein LD (2003) Integrating biological databases. Nat Rev Genet 4(5):337–345

Data Integration and Visualization

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

Synonyms

Interoperability; Semantic integration

Definition

Data integration is the process of bringing together information from multiple, diverse sources such that it can be interrogated as a whole to provide holistic knowledge that is greater than the sum of its parts. In particular, data integration aims to seamlessly expose information inherent in the relationships between concepts. There are numerous technological challenges relating to the scalability of data-integration systems, as well as complex issues concerning both the nature of the data itself and the means by which the data may be understood by humans. Visualization is the process of making data human intelligible, enabling human intuition and expert knowledge to be applied in areas where algorithmic interrogation is unrealistic.

Systems biology attempts to take a "universal" view of complex biological phenomena, treating them as an integrated whole rather than as individual, independently functioning components: it is as much about understanding the interactions *between* biological entities as it is about the entities themselves. This approach necessitates studies at a variety of levels, from individual small molecules and macromolecular complexes to their interactions in a variety of interrelated contexts: e.g., the specific biochemical pathways in which they participate, the molecular networks those pathways may form, and, ultimately, the complete organisms to whose evolution and functioning those networks and pathways contribute. Supporting systems perspectives with information technology is challenging in terms of both data integration and visualization: managing the individual components in isolation is hard enough; integrating this information to provide "system-wide" views is even more daunting.

Characteristics

Issues with Data-Integration Technologies

Traditionally, data integration has involved either data warehousing or database federation. In data warehousing, data are extracted from various sources (databases, "flat files," and other information-management systems), transformed into a common schema (and also possibly audited for quality), and finally loaded into what is logically a single (usually relational) data-store for querying by end-users. This process is often referred to as "ETL" (extract, transform, load). Federating databases, however, retains the original data sources and schemas at their original locations, instead providing a distributed query mechanism that interrogates the sources in unison, as if they were logically a single resource. In essence, federated systems perform a similar process to ETL but "on the fly." Warehousing is a predominantly centralized approach, and federation is inherently distributed; thus, the same issues of balancing space complexity (storage) and time complexity (computational load) apply as with other distributed architectures. Similarly, the usual pros and cons associated with scalability, run-time performance, and data integrity must be taken into consideration when designing data-integration platforms.

Many data-integration issues arise from the need to reconcile disparate source database schemas; the more diverse the schemas, the harder the problem. Traditional (relational) databases explicitly encode schemas into the database architecture; this requires the schemas to be "decomposed" when loaded into a warehouse, or at run-time in a federated system. Schemas can be very simple, with all information encoded in a single monolithic table of records – this makes some queries trivial, but is inflexible, making re-purposing the data difficult. Alternatively, they may encode records with finer levels of granularity – this provides greater flexibility, but imposes on users the need to formulate significantly more complex queries. Contemporary approaches attempt to obviate the need for schemas, relying on the use of ontologies to give meaning and structure to database contents. Here, "triple stores" and "schemaless" databases provide a generic underlying technology, and shared ontologies for reconstituting those data are managed by the community.

Various community-driven initiatives are evolving data-integration standards (e.g., BioSharing, BioDBcore, BioPax) in close collaboration with publishers, journals, database curators, software developers, and so on (Field et al. 2009; Gaudet et al. 2011; Demir et al. 2010). These are intended to help users locate and access information dispersed within databases; to help shape the data-preservation/management/sharing policies implemented by journal editors and funders; and to encourage software developers to embrace and extend community-endorsed standards, like the systems biology markup language (SBML, Hucka et al. 2004) and CellML (Garny et al. 2008). Alongside ontologies and schemas, numerous "minimum information" standards have evolved as a means of establishing a baseline for the information deposited in data repositories (Taylor et al. 2008). These checklists and guidelines aim to improve the quality and integrity of recorded data, leading to improved opportunities for data integration.

Issues with the Data

Irrespective of data-integration technologies, there are also issues at the data level, especially with identity and nomenclature. Getting humans to agree on the meaning of words is hard; getting them to understand when they are using the same name to mean different

Data Integration and Visualization, Table 1 The variety of names for 'the same' gene and its protein product in three different species, including the many different protein alternative names and gene synonyms

Species	Protein	Alternative name	Gene	Synonym	UniProtKB:ID
Saccharomyces cerevisiae	DNA replication licensing factor MCM4	cell division control protein 54	MCM4	CDC54, HCD21	MCM4_YEAST, P30665
Drosophila melanogaster	DNA replication licensing factor MCM4	protein disc proliferation abnormal	dpa		MCM4_DROME, Q26454
Mus musculus	DNA replication licensing factor MCM4	CDC21 homolog, P1-CDC21	Mcm4	Cdc21, Mcmd4	MCM4_MOUSE, P49717

things is harder; getting a machine to understand the difference is harder still. Precision is key; but unfortunately, adherence to standard nomenclatures has been limited in the life sciences.

Consider the protein shown in Table 1: this has five protein and six gene names in three different species. Humans can unravel such complexity relatively easily; but once we involve computers in the loop of comprehension, the task becomes more complicated, and the level of precision required to specify a particular protein/gene becomes more difficult to achieve: e.g., a text-mining algorithm exploring the literature for gene dpa or protein "disc proliferation abnormal" would miss other mentions of the same gene (Mcm4, HCD21, CDC54, etc.) or protein (cell division control protein 54, CDC21 homolog, P1-CDC21) unless it had been programmed to use comprehensive synonym dictionaries.

Consistent use of identifiers is also an issue. Consider ovine rhodopsin. This protein entered the PIR database with identifier (ID) OOSH, accession number (AC#) A03155. Successive changes to its sequence and annotation then led to changes of its AC# to A93264, A90319 and then A30407. Later, the protein also appeared in Swiss-Prot with ID OPSD $SHEEP, AC# P02700. Its ID then changed to OPSD_SHEEP; finally, the PIR and Swiss-Prot entries acceded to UniProtKB, where changes and refinements continued to be made. Today, UniProtKB archives 82 versions of the entry for ovine rhodopsin, including these three different IDs and five different AC#s; the sequence itself is also stored in UniParc, with AC#s UPIUPI000059C30D and UPI0000130E18, the latter being the currently active sequence of record. Initiatives such as the MIRIAM Registry (Laibe and Le Novère 2007) and its associated naming scheme will be crucial in untangling such "identity crises" in future.

Visualization

Humans are intuitive pattern matchers. Computers, by contrast, are incapable of the leaps of intuition that are the essence of human thought processes. Machines must be programmed to find specific patterns; this requires programmers to characterize those patterns in terms that are machine comprehensible. Visualization bridges the gap between human intuition and machine pattern-matching, and constitutes the set of tools and paradigms that allow computers to aid humans in knowledge discovery from large, complex data-sets.

Aside from standard histograms, scatter graphs, etc., the life sciences also tend to use network-, structure-, and sequence-visualization approaches.

Network visualization is important because systems biology tries to understand relationships *between* biological entities. Computationally, these relationships can be represented as "graphs" in the mathematical sense, i.e., collections of objects (represented by "nodes"), some of which are linked in some way (with relationships represented by "edges"). Numerous exchange formats (e.g., GraphML) have been developed for encapsulating the properties of generic mathematical graphs, with more specific formats (like SBML, CellML, and BioPax) for capturing networks and pathways in machine-readable form. The systems biology graphical notation (SBGN) initiative defines a mechanism for making such networks human readable, providing both a graphical notation for displaying and a schema for encoding network diagrams (Le Novère et al. 2009).

Structure visualization encompasses techniques and formats for representing small molecules (consisting of perhaps a few tens of atoms and bonds), through to macromolecular structures (involving hundreds or thousands of atoms and bonds). For small molecules, compact representations such as the

IUPAC international chemical identifier (InChi™) are often sufficient to define a molecule's topological structure (but the use of InChi as a chemical identifier is hotly contested, with many arguing that "semantic-free" identifiers offer a more reliable foundation for data integration). A plethora of other file formats explicitly capture atomic coordinates for small molecules; for large molecules like proteins, where the atomic structure is typically determined experimentally using techniques like X-ray crystallography, formats such as that defined by the protein data bank (PDB, Berman et al. 2000) are often used for data exchange. As well as encoding the empirically determined coordinate data, this format supports basic annotation of the underlying sequence.

Sequence visualization techniques usually involve displaying ordered lists of amino acids (in a protein) or nucleotides (in a chromosome or genome). Here, the main challenge is to provide mechanisms for mapping regions of interest (topological domains, coding/non-coding regions, etc.) onto a sequence, and to allow comparison between multiple sequences (e.g., to determine their similarity). The data formats for sequences are relatively simple compared to those for structure and network visualization: e.g., the FastA and PIR formats primarily comprise strings of characters from the relevant sequence alphabet; the "general feature format" (GFF) and the schema of the distributed annotation system (DAS) also encode regions of interest that may be mapped to the sequence. Typically, sequences are depicted visually as horizontal rows of color-coded blocks, representing residues or nucleotide bases. Such visualizations allow users to scroll back and forth, zoom in and out, reorder the sequences or their features, render them in 3D, etc. Until recently, users would tend to align tens or hundreds of sequences; however, developments in next-generation sequencing (NGS) are likely to drive this into the thousands, pushing the limits of sequence visualization into new dimensions.

Conclusions

Capturing life science data for the purposes of data integration and visualization – whether for sequence, structure, or network analysis – is challenging. As techniques like NGS create more data than ever before, the problems become more complex. In attempting to paint holistic pictures, systems biology must take into account the increasing convolutions of integrating disparate data drawn from a field that is itself growing in complexity. It has become clear that the ad hoc mechanisms and file formats that have served the field for decades are no longer sufficient, and greater use of ontologies, standards, and identification schemas will be required to help extract knowledge from our growing data collections. With increasing complexity, however, our ability to integrate data meaningfully using ontologies and schemas is being pushed to its limits. Ultimately, if we are to be able to grasp the subtleties of communication, we will need to be able to more effectively capture the relationships between data and the biomedical literature (i.e., how data are described in scientific articles). Tools for managing this relationship will become crucial in future.

Cross-References

- ▶ Alignment, Protein Interaction Networks
- ▶ Applied Text Mining
- ▶ Biological Network Model
- ▶ Biological System Model
- ▶ Bio-Ontologies
- ▶ Controlled Vocabulary
- ▶ Data Integration
- ▶ Data Mining
- ▶ Directed Acyclic Graph
- ▶ Distributed Data Access
- ▶ Gene Ontology
- ▶ Graph
- ▶ Graph Algorithms in Network Analysis
- ▶ Graph Alignment, Protein Interaction Networks
- ▶ Graphical Model
- ▶ Interoperability
- ▶ Knowledge
- ▶ MIRIAM
- ▶ MIRIAM URI
- ▶ Ontology
- ▶ Ontology Structure
- ▶ SBML
- ▶ Text Mining

References

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

Demir E, Cary MP, Paley S et al (2010) The BioPAX community standard for pathway data sharing. Nat Biotechnol 28: 935–942

Field D, Sansone SA, Collis A et al (2009) Omics data sharing. Science 326:234–246
Garny A, Nickerson DP, Cooper J et al (2008) CellML and associated tools and techniques. Philos Transact A Math Phys Eng Sci 366:3017–3043
Gaudet P, Bairoch A, Field D et al (2011) Towards BioDBcore: a community-defined information specification for biological databases. Nucleic Acids Res 39:D7–D10
Hucka M, Finney A, Bornstein BJ et al (2004) Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. Syst Biol 1:41–53
Laibe C, Le Novère N (2007) MIRIAM resources: tools to generate and resolve robust cross-references in systems biology. BMC Syst Biol 1:58
Le Novère N, Hucka M, Mi H et al (2009) The systems biology graphical notation. Nat Biotechnol 27:735–741
Taylor CF, Field D, Sansone SA et al (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 26: 889–896

Data Integration, Breast Cancer Database

Ettore Mosca, Ivan Merelli and Luciano Milanesi
Institute for Biomedical Technologies – CNR (Consiglio Nazionale delle Ricerche), Segrate, Milan, Italy

Synonyms

Genes-to-systems breast cancer database

Definition

The Genes-to-Systems Breast Cancer Database (shortly, G2SBC Database) is a freely available Web resource that collects data about genes, transcripts, and proteins reported in the scientific literature as altered in *breast cancer* cells; alterations encompass different types of mutations and expression variations of transcript and proteins. These data are integrated in a multilevel database (from genes, transcripts, and proteins to *molecular networks*, cell populations, and tissues), which is coupled with a series of analysis tools concerning cellular biochemical pathways, protein–protein physical interactions, protein structures, and mathematical models of cell behavior.

Characteristics

Motivation

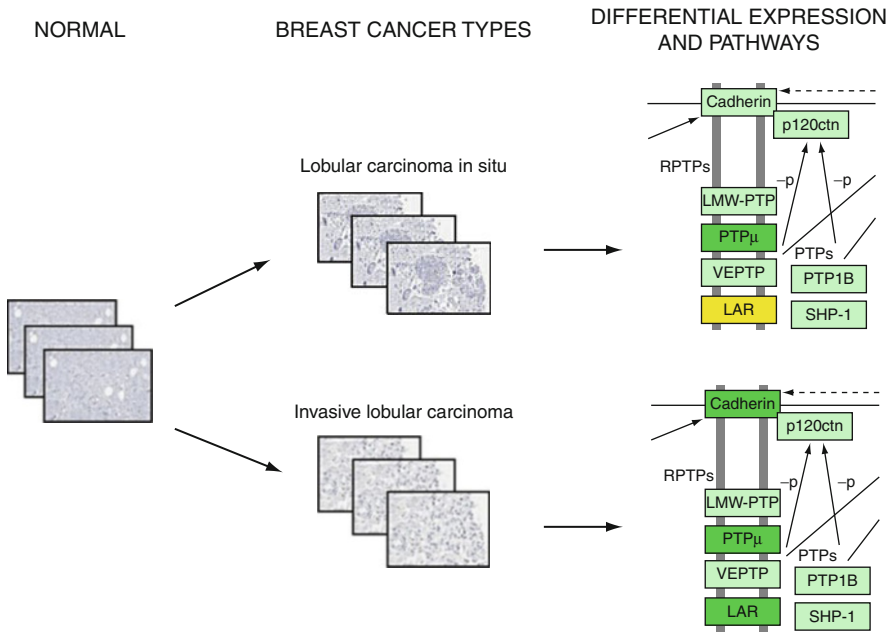
Breast cancer is one of the most common cancer types: Approximately, it affects 1 out of 10 women and represents the 25% of all tumors that hit women. From a scientific point of view, it is increasingly believed that using a systems biology perspective it is possible to develop better strategies for cancer treatment; this consideration is due to the fact that the complex behavior of living systems can be hard to predict from the properties of individual parts (such as genes, proteins, and cells). In this context, a multilevel integration of the available knowledge regarding both biological components and pathways is a crucial task in order to promote the system perspective.

Functionality

The G2SBC Database contains several types of molecular alterations associated with breast cancer. These alterations encompass the genome (mutations and SNPs), the transcriptome (RNA expression level and splicing variations), and the proteome (protein expression level, sequence, structure, and localization). Molecular alteration data are integrated with information concerning the molecular network layer (pathways and interactions, retrieved from databases as *KEGG Pathway* and *BioGRID*, respectively) and the cellular layer (breast cancer types and tissue images from the *Human Protein Atlas*, mathematical models of carcinogenesis, tumor growth, and tumor response from literature).

The Web site (implemented employing PHP and JavaScript) by which the database can be accessed is mainly divided into three sections. The first concerns the query system, which allows to retrieve data from the three levels of biological entities: molecular components, the molecular systems, and the cellular layer.

The second section concerns the analysis tools available by means of the Web interface. In particular there are two tools that rely on the application of *graph theory* to *biological networks* for highlighting interactions, protein complexes, and hubs. Concerning the *gene-annotation enrichment analysis*, the G2SBC Database provides a tool that enables the functional annotation of gene lists provided by the user or retrieved exploring the data from the database.



Data Integration, Breast Cancer Database, Fig. 1 Integration between protein expression in different breast cancer types and biochemical pathways data. A use case showing the differential expression of some proteins belonging to the "adherens junction" KEGG map in lobular carcinoma in situ and invasive lobular carcinoma tissue images collected from the Human Protein Atlas database. *Green*: downregulation; *yellow*: similar expression

Lastly, the G2SBC Database maintains a model-oriented section, which involves two aspects. The first one concerns the interaction among cell cycle regulation and breast cancer: due to this connection it is possible to retrieve the breast cancer genes involved in cell cycle control and simulate the associated mathematical models. The second regards the mathematical models related to carcinogenesis, tumor growth, and response to treatments.

The data integration approach employed to develop this resource enables the collection of a large amount of records and the Web tools provided allow to infer nontrivial knowledge: one example is the integration of protein expression data available for different types of breast cancer with the cellular pathways, Fig. 1.

Accessibility

The G2SBC Database is freely accessible at http://www.itb.cnr.it/breastcancer. An extensive help section is available and contains some use cases covering the different sections of the G2SBC Database.

Cross-References

- ▶ Comparative Analysis of Molecular Networks
- ▶ Data Integration, Breast Cancer Database
- ▶ Functional Enrichment Analysis
- ▶ Gene Set and Protein Set Expression Analysis
- ▶ Interaction Networks
- ▶ KEGG Pathway Database

References

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(Database issue):D355–D360

Mosca E, Alfieri R, Merelli I, Viti F, Calabria A, Milanesi L (2010) A multilevel data integration resource for breast cancer study. BMC Syst Biol 4:76

Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M (2011) The BioGRID interaction database: 2011 update. Nucleic Acids Res 39(Database issue):D698–D704

Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Björling L, Ponten F (2010) Towards a knowledge-based human protein atlas. Nat Biotechnol 28(12):1248–1250

Data Management

- ▶ Knowledge Management

Data Management: Data Processing

▶ Distributed Data Management

Data Mining

Aleksi Kallio[1] and Jarno Tuimala[2]
[1]CSC, IT Center for Science Ltd, Espoo, Finland
[2]Finnish Red Cross Blood Service, Helsinki, Finland

Definition

Data mining is the process of applying computational methods to large amounts of data in order to reveal new non-trivial and relevant information. Data mining is not only used for finding interesting patterns from the data but also for exploring large data sets, for building models that describe the relevant properties of data, and for making predictions based on the data (Hand et al. 2001). Due to rapidly evolving biological measurement instruments such as sequencers and array scanners, data mining problems have become central in modern biosciences (Baldi and Brunak 2001).

Characteristics

There is some confusion on the meaning of the term data mining and its relation to overlapping terms such as machine learning and statistics. As a field of science, data mining sits between information technology and statistics. Data mining originated from knowledge discovery research in information science, and it has traditionally had very little interaction with statistics. This history also explains much of the critique toward data mining, to a large extent from the statistical community. Early data mining methods have been criticized for so-called "data fishing": searching for complex patterns without controlling the probability of them being just random artifacts in the data (Hand et al. 2001).

However, the difference between statistics and data mining is vague: statistics is especially interested in inference and prediction, typically based on some a priori hypotheses and supporting experimental designs. In contrast, data mining is more interested in making inferences from data without much background information or typically without control over data producing process, for example, experimental design. John Chambers (1993) coined the term "greater statistics" as everything related to learning from data, and data mining falls within that definition. Thus, data mining can also be thought of as a field of statistics making heavy use of visualizations and computationally intensive methods with an emphasis on algorithmic efficiency.

Machine learning is a related field that shares many of its methods with data mining. The distinction between the two is largely historical; machine learning evolved from artificial intelligence research, whereas data mining evolved from information science. Artificial intelligence problems are often conceptually well defined such as making a correct medical diagnosis based on data from a patient, whereas the data mining problem could be, for example, to mine the database of medical records to find new and surprising associations. Partly due to the somewhat different goals, these two communities of computational data analysis typically rely on different statistical frameworks (Hand et al. 2001; Baldi and Brunak 2001). Machine learning research has embraced the flexible Bayesian framework (▶ Bayesian inference) for probabilistic modeling, while data mining often uses computationally lightweight frequentist statistics to process large data sets.

Data Mining Methods

Data mining is to a large extent defined by the methods that are used in the field. The methods and their goals are varied, but can be roughly divided into two main categories: they either try to model the data (learn the global structure of the data) or try to find patterns of interest (learn local structures from the data). From a computational point of view, data mining problems are often NP-hard, meaning that they do not have an optimal solution algorithm and need approximated solutions. This inherent complexity is also one of the reasons why data mining is seen as an interesting area for computer science research.

The most common data mining tasks can be divided into clustering (unsupervised learning), classification (supervised learning), regression (▶ Regression Analysis), association mining, and text mining (▶ Text Mining) (Hand et al. 2001; Hastie et al. 2009).

Clustering (▶ Clustering) tries to find groups of similar or similarly behaving entities, such as genes, proteins, or metabolites, from the data. In general, clustering is blind to the known structures of the data in a sense that it does not use any knowledge outside the data to assign the entities to the clusters. Clustering methods are at the core of data mining, although arguably no longer a central point of research. A large variety of clustering methods exists, but among the most well-known ones are *k*-means clustering (▶ Clustering, k-Means) and hierarchical clustering (▶ Clustering, Hierarchical) (Fig. 1). One typical application area for clustering is gene expression analysis, where clustering can be used for, for example, grouping cancer samples to discover novel subtypes.

Classification (▶ Classification) falls into two categories. First, if knowledge of the grouping of the entities, such as the sex of the providers of the biological samples, is available, then classification can be applied to find the themes that make the groups different. Second, once such themes are identified, classification can predict the grouping of new entities. Perusing the previous example on cancer gene expression data, classification methods can be used to learn the expression patterns of known cancer subtypes and then to classify new cancer samples. Common classification techniques are discriminant analysis, decision trees, nearest-neighbor methods, neural networks, and support vector machines.

Regression (▶ Regression) comprises a very large family of methods. The general idea of regression is to find a function of predictor variables that fits the observed response with the least error. Hence, regression can be used for, for example, studying an organism's response to the dosage of a drug. The range of methods varies from the classical linear regression (▶ Linear Regression) through generalized linear models to generalized additive mixed models and mixture models.

Association methods (▶ Association Rule) search for relationships between variables. Association rule mining has found only a few applications in bioinformatics, since most implementations of it need categorical data, and typical measurements in biology are made on a continuous scale. However, especially in systems biology, mining associations in the form of networks has been studied extensively. Example applications include gene regulatory networks (▶ Data Mining–based Transcriptional Regulatory Network Construction) and protein interaction networks.

Text mining (▶ Text Mining) methods locate patterns and trends from textual data sets. The two major application areas in systems biology are analyzing genomic and proteomic sequences for patterns of interest and mining literature databases for associations between, for example, genes and proteins.

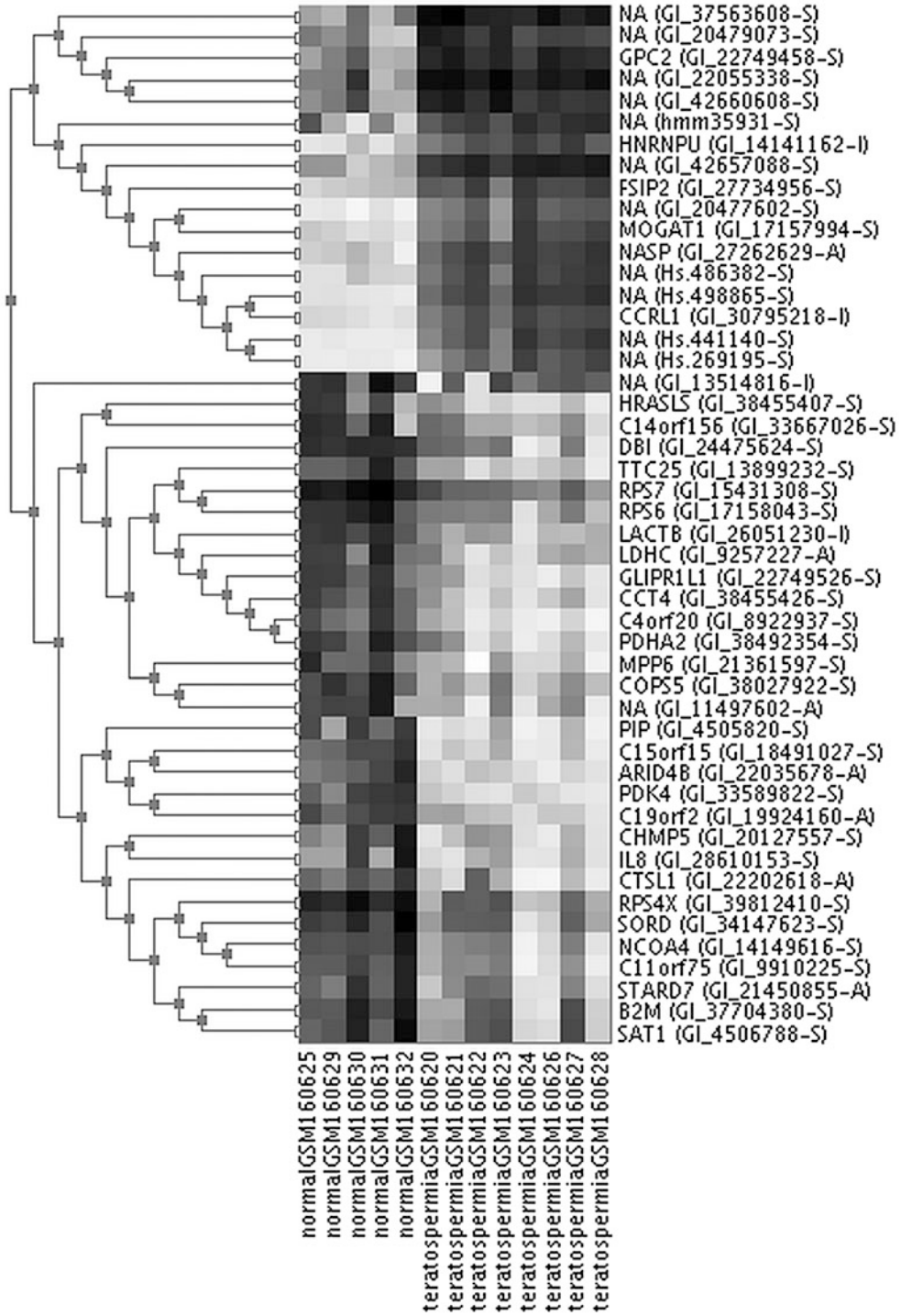
Visualization

Humans are skilled at quickly locating patterns from visual representations, which can be harnessed for efficient data exploration. In statistics, data is often summarized with variables such as mean and variance. Visualizations can be thought of as an extension to simple variables, allowing the overview of more complex structures (e.g., contour plots) and multiple variables simultaneously (e.g., scatter plots). Visualizations can also be coupled with data mining techniques. For example, by using dimensionality reduction techniques to bring the number of dimensions down to two, and then by using scatter plots to show the result, a high-dimensional data set can often be shown in an understandable form.

Validation

Not all discovered patterns are necessarily robust or valid. Data mining methods often find patterns that are unique to the analyzed data, and cannot be detected from independent data sources. The same basic phenomenon exists in different forms in different data mining tasks: in modeling, the problem is known as "overfitting" (▶ Overfitting), whereas in pattern searching and statistical testing, it is known as "false positives." In all cases, one is led to draw overly optimistic conclusions.

There are different approaches to guarantee valid and robust results. One approach is to split the data into a training set (▶ Model Training, Machine Learning) and a test set (▶ Model Testing, Machine Learning). The selected data mining method is applied on the training data only, and once the pattern has been discovered, its validity is assessed using the test data (▶ Model Validation, Machine Learning). Some data types allow randomization methods to be used, where the whole data is used in training, but the trained method is also applied to a large number of randomized versions



Data Mining, Fig. 1 Visualization of a hierarchical clustering of gene expression samples. Genes are clustered so that they have similar expression levels in the sampled conditions. Visualization was produced with Chipster software

of the original data. Only those original results that are unlikely in randomized data are selected.

Validation is hard to do well, and it often takes considerable effort to program a validation approach that is suitable for the particular data at hand. Ready-made solutions in most software programs often do not perform satisfactorily.

Tools and Data Management

It is common to store structurally simple data sets as flat files, such as comma separated tables. More complex data sets are typically stored in SQL databases, in XML files or in custom data storage systems. Data warehouse technologies commonly used in the industry are less often seen in bioscientific data mining.

For preprocessing and analysis, most commonly used tools are either generic scripting languages (e.g., Perl and Python), interactive data analysis environments (e.g., R (▶ R, Programming Language) and Matlab), and graphical tools (e.g., Weka). Knowledge discovery professionals in the industry often use specialized data mining packages, but in systems biology in academia, a set of open source command line tools is favored. As the need for data mining methods is growing due to developments in measurement technology, commercial companies have started to provide specialized graphical tools, which have been later followed by open source developments. These tools aim to provide advanced data analysis capabilities to users without computational background.

Analysis workflows differ, but they all contain some variants of the following steps: preprocessing, data exploration, and main analysis. In preprocessing, the data is formatted correctly, filtered for unnecessary or dubious content, and often normalized. Exploration ranges from outputting simple statistics to elaborate visualizations. The main analysis is often constructed in the form of a statistical test, although sometimes a less strict approach is used. Especially for systems biology work, an additional step is taken after the main analysis: result integration and explanation. In the integration and explanation step, the results are often validated (Model Cross-validation, Machine Learning) against independent data sources, such as the many biological databases that are publicly available. The intent is to find support for the results and also to find sound explanations that help to understand the results as part of the larger "systems view".

Cross-References

- ▶ Bayesian Inference
- ▶ Classification
- ▶ Clustering, Hierarchical
- ▶ Clustering, k-Means
- ▶ Clustering, Model-Based
- ▶ Data Mining–based Transcriptional Regulatory Network Construction
- ▶ Linear Regression
- ▶ Model Testing, Machine Learning
- ▶ Model Training, Machine Learning
- ▶ Model Validation, Machine Learning
- ▶ Overfitting
- ▶ R, Programming Language
- ▶ Text Mining

References

Baldi P, Brunak S (2001) Bioinformatics: the machine learning approach, 2nd edn. MIT Press, Massachusetts

Chambers J (1993) Greater or lesser statistics: a choice for future research. Stat Comput 3:182–184

Hand DJ, Mannila H, Smyth P (2001) Principles of data mining. MIT Press, Massachusetts

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

Data Mining–based Transcriptional Regulatory Network Construction

Xing-Ming Zhao
Institute of System Biology, Shanghai University, Shanghai, China

Synonyms

Classification; Data mining; Regression

Definition

Data mining is a new field that aims to analyze large datasets and extract knowledge from data, and it is an interdisciplinary field involving statistics, pattern

recognition, information theory, and so on. Generally, data mining techniques can extract informative patterns from data and construct decision-making systems. The most popular techniques in data mining include clustering, classification, regression, etc. Data mining is being widely used in various fields, including image processing, marketing relationship, and medicine (Fayyad et al. 1996).

Recently, data mining has been playing an important role in bioinformatics, based on which various tools have been developed to construct transcriptional regulatory networks. Instead of determining individual protein-gene regulations with traditional biological experiments, the integration of emerging high-throughput data and data mining is able to predict transcriptional regulations in an automatic way. In general, the regulation relationship can be predicted as a classification or regression problem, which aims to extract regulation patterns from the experimental data.

Characteristics

Data Representation and Preprocess

In data mining, the data are usually formulated as vectors, where each sample is represented as one vector. Therefore, the similarity between samples can be estimated. In constructing transcriptional regulatory network, one of the most commonly available data source is microarray data that describes the expression profile of genes under different conditions. The gene expression data can be easily used for data mining. Other kinds of data, for example, ChIP-on-chip (also known as ChIP-chip) data, can also be represented as numeric vectors.

However, it is not a trivial task to extract informative patterns due to the noise inherited in the experimental data. Some data mining techniques are therefore used to reduce noise and sometimes also help to reduce the dimensionality of the data, such as principal component analysis (PCA). Since the gene expression data can be measured as time series data, some signal processing methods, for example, Fourier transform, are helpful to reduce noise and transform the data into another description space so that the signal in the data can be easily detected. Another important issue in data mining is feature selection, which aims to select important patterns so that the performance of the classifier can be improved. Popular feature selection techniques include maximum relevance/minimum redundancy (MRMR), Entropy, and Mutual information, etc. The selected features can help to interpret the model and important patterns.

Data Mining Algorithms

Various data mining methods can be grouped into three groups, that is, supervised methods, unsupervised methods, and semi-supervised methods, among which the supervised and unsupervised methods are widely used in transcriptional regulatory network construction. In supervised methods, there are some regulations that are known in advance and are therefore used as training set to infer the patterns for prediction of new regulations. The supervised methods widely used to construct transcriptional regulation networks include artificial neural network, decision tree, genetic algorithm, and support vector machine. On the other hand, if there are not any known regulations available, the unsupervised methods will be helpful, among which the most popular one is clustering.

Artificial Neural Networks

Artificial Neural Network (ANN) model is designed to mimic the real biological neural network that can process information and make decision. In general, the ANN model consists of three parts, including input layer, output layer, and hidden layer. There are some nodes in each layer that work like the neurons, and the nodes are interconnected to simulate the information flow between neurons. The structure of the ANN model reflects the mapping from input variables to output variables. The ANN model can handle large dataset and is robust against noisy data. However, the ANN model works as black box and it is difficult to interpret the results obtained in some cases.

Veiga et al. (2008) designed a feed-forward (FF) and bi-fan (BF) regulatory motif prediction model for *Escherichia coli* based on multilayer perception artificial neural networks (ANNs). The regulatory motifs predicted consist of transcription factors and regulated genes, and are highly enriched. Hart et al. (2006) found that single-layer feed-forward ANN models can effectively discover gene network structure by integrating global in vivo protein-DNA interaction data (ChIP/Array) with genome-wide microarray data. The ANN models were successfully applied to construct the yeast cell cycle transcriptional regulatory network, which is composed of hundreds of genes.

Genetic Algorithms

The genetic algorithm (GA) is a heuristic optimization method that works like the evolution of biological processes and aims to find out the global optimization in the solution space. There are some chromosomes in GA, which denote possible solutions, and combination and mutation are used to change the structure of the chromosomes so that the evolution proceeds to better solution. The objective function of GA is called fitness function and is specially designed for the algorithm's goal based on expert knowledge. GA performs very well in searching for optimal solutions although it is time consuming in some cases.

In the construction of transcriptional regulatory network, it is important to determine the topological structure of the regulatory network. Seema and Ramanatha (2010) successfully applied GA to infer the structure of transcriptional regulatory network. Kikuchi et al. (2003) presented a modified version of GA that not only predicts the structure of the regulatory network but also predicts the dynamics of the regulations. In particular, a new fitness function was designed to improve the performance.

Decision Tree

Decision tree is a decision-making system that utilizes a tree-like data structure. Decision tree has been widely used as a decision-making system for a long time due to its simple structure and interpretability. The popular decision tree learning methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). Decision tree is robust against noisy data and can handle large dataset in a short time. In addition, the decision tree can perform feature selection automatically, which can help to interpret the model.

Ruan et al. (2009) built an ensemble classifier of decision trees to construct transcriptional regulatory network, where each decision tree is learned based on one specific gene expression dataset. The yeast transcriptional regulatory network was constructed by this classifier that associates the gene expression level with binding affinity between transcription factors and DNAs detected by chromatin immunoprecipitation and microarray (ChIP-chip) experiment.

From the decision tree ensembles, the logical rules can be extracted to explain how a set of transcription factors act in concert to regulate the expression of their targets.

Support Vector Machine

Support vector machine (SVM) is a newly developed supervised classifier that is especially useful for dataset with high dimensionality and small samples, and it can work on both classification and regression problems. The success of SVM is attributed to its two new features. Firstly, SVM uses kernel technique to describe the relationship between samples, which enables SVM to work efficiently without considering the dimensionality of the data. Secondly, SVM only uses the samples near the classification hyperplane, that is, support vectors, to build the classifier so that it can work robustly against noise.

Qian et al. (2003) used support vector machines (SVMs) to predict the targets of a transcription factor based on the association relationships between their expression profiles. In particular, SVMs successfully predicted the targets of 36 transcription factors for *Saccharomyces cerevisiae* based on the microarray data obtained under different physiological conditions. Kumar et al. (2007) presented another framework for predicting protein-DNA interactions based on sequence information and SVMs, and gave promising results.

Clustering

Clustering is one of the most popular unsupervised learning methods in data mining. Clustering generally groups samples into different clusters so that the samples in the same cluster are similar based on the patterns in the data while dissimilar between clusters. The most important and difficult things in clustering are how to describe the similarity between samples and what criteria should be used to make a group for a set of samples.

In biology, the genes that are always co-expressed under various conditions are assumed to be co-regulated by same regulators. Therefore, the genes can be clustered into groups based on their expression profiles. If a set of genes are clustered into one group, these genes are regulated by either regulators outside of the group or regulators within the group. Based on the above assumptions, various clustering techniques have been applied to construct

transcriptional regulatory network. For example, de Hoon et al. (2003) first clustered the genes into different groups and then constructed a transcriptional regulatory network based on the clustering results for *Bacillus subtilis*.

Cross-References

- ▶ Chromatin Immunoprecipitation
- ▶ Entropy
- ▶ Feature Selection
- ▶ Maximum Relevance/Minimum Redundancy (MRMR)
- ▶ Mutual Information

References

de Hoon MJL, Imoto S, Kobayashi K, Ogasawara N, Miyano S (2003) Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. Pac Symp Biocomput 8:17–28

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Mag 17:37–54

Hart CE, Mjolsness E, Wold BJ (2006) Connectivity in the yeast cell cycle transcription network: inferences from neural networks. PLoS Comput Biol 2(12):e169

Kikuchi S, Tominaga D, Arita M, Takahashi K, Tomita M (2003) Dynamic modeling of genetic networks using genetic algorithm and s-system. Bioinformatics 19(5):643–650

Kumar M, Gromiha M, Raghava G (2007) Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics 8:463

Qian J, Lin J, Luscombe NM, Yu H, Gerstein M (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 19(15):1917–1926

Ruan J, Deng Y, Perkins EJ, Zhang W (2009) An ensemble learning approach to reverse-engineering transcriptional regulatory networks from time-series gene expression data. BMC Genomics 10(Suppl 1):S8

Seema S, Ramanatha KS (2010) Inference of gene regulatory network using modified genetic algorithm. In: Proceedings of the international symposium on biocomputing, Calicut, Kerala, pp. 1–8

Veiga DF, Vicente FF, Nicolás MF, Vasconcelos AT (2008) Predicting transcriptional regulatory interactions with artificial neural networks applied to *E. coli* multidrug resistance efflux pumps. BMC Microbiol 8:101

Data Parallel

- ▶ General-Purpose Computation, Graphics Processing Units
- ▶ Grid Computing, Parallelization Techniques

Data Sampling

Haiying Wang and Huiru Zheng
School of Computing and Mathematics, Computer Science Research Institute, University of Ulster, Jordanstown, UK

Synonyms

Sampling; Statistical sampling

Definition

In machine learning (▶ Model Validation, Machine Learning), data sampling (Cochran 1977) is a widely accepted process concerned with the selection of an unbiased subset of data, which are representative of a larger population, for the purposes of constructing predictive models with machine learning algorithms. The size of the sample is the number of data items in a sample, typically denoted as an integer number N.

The major benefit of applying data sampling in the context of machine learning is it can effectively speed up the modeling process, which allows the analyst to build a model and make prediction with relatively little cost and effort. To achieve this, a sample is required to contain the essence and reflect the characteristics of the entire dataset. The key to meet this fundamental requirement is randomness, i.e. allowing each data item in the database to have the same probability of being selected.

Types of data sampling commonly used in machine learning include the following:

- Simple Random Sampling
  Each data item has the same chance of being selected in the data set.

- Stratified Random Sampling
  The entire dataset is first divided into k disjoint groups with each group having the size of N_1, N_2, . . ., N_k. These subgroups are non-overlapping and together they are comprised of the whole population, i.e., $N_1 + N_2 + \ldots + N_k = N$. Then for each subgroup, a simple random sample is taken in a number proportional to its size when compared to the whole population. The collection of these subsets constitutes a stratified sample. The subgroups are referred as strata and the whole procedure is called stratified random sampling.
- Cluster Sampling
  The entire dataset is divided into groups of items, i.e., clusters, and each cluster becomes a sample unit. This is an example of two stage sampling. In the first stage, analysis is carried out on a population of clusters and each cluster has the same chance of being included in the sample. After this process, a random number of items within these clusters is selected.

Cross-References

▶ Model Validation, Machine Learning

References

Cochran WG (1977) Sampling Techniques. Wiley, New York

Data Storing and Querying

▶ Distributed Data Management

Data Warehouse

▶ Data Integration

Database AC

▶ Database Accession Number

Database Accession Number

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

Synonyms

AC#; Database AC

Definition

A database accession number, rather like a database identifier, is a short code used to uniquely identify a particular entry or record within a particular database. The code normally contains alphanumeric characters, and is usually designed to be machine readable (they are seldom, if ever, human readable). For example, P02700 is the accession number that identifies the entry for ovine rhodopsin in the UniProtKB:Swiss-Prot (UniProt consortium 2011) protein sequence database: Unlike its largely human-readable database identifier (OPSD_SHEEP), this accession number is neither informative nor particularly memorable to humans.

As already mentioned, accession numbers are database specific, and different databases adopt different numbering conventions. Hence, for example, in the PIR protein sequence database, ovine rhodopsin has the accession number A03155.

Information pertinent to ovine rhodopsin, which belongs to a superfamily of G protein–coupled receptors (GPCRs), may also be found in protein family databases like InterPro (Hunter et al. 2009), PROSITE, PRINTS, Pfam, and so on: Examples of accession numbers for the GPCR superfamily entries in these databases are IPR00026, PS00237, PR00237, and PF0001, respectively. By contrast with protein sequence database numbering schemes, it is broadly possible to decipher which is the parent database from protein family database accession numbers: For example, IPR denotes InterPro; PS denotes PROSITE; PR, PRINTS; and PF, Pfam. The number itself, however,

gives nothing away about the protein family entry with which it is associated.

Database accession numbers are intended to provide a stable means of tracking down particular database entries. Consider, for example, the DNA replication licensing factor MCM4. Although its Swiss-Prot identifier has oscillated since March 1993, from CD21_YEAST, to C21H_YEAST, CC54_YEAST, CDC54_YEAST, and finally to MCM4_YEAST, its accession number remained the same throughout – that is, P30665. Nevertheless, accession numbers can and do change between database releases. Thus, for example, the accession number for the same protein in the PIR protein sequence database began as S25527, then became S26641, and finally S56050. Sometimes, different database entries in the same database are merged or replaced with or by others (e.g., when a computer-annotated TrEMBL sequence is discovered to be redundant with an existing manually annotated Swiss-Prot sequence). In such cases, the accession number of the deprecated sequence is retained so that it too can be tracked. For example, in November 2010, the TrEMBL entry D6W429 (ID, D6W429_YEAST) was replaced by UniProtKB: Swiss-Prot entry P30665 (ID at that time, CDC54_YEAST): Effectively, the entries were merged and D6W429 was retired. At that point, the accession number record of the revised entry was updated to read, "AC P30665; D6W429." In this case, P30665 is denoted the primary accession number and D6W429 the secondary accession number. Retaining all the accession numbers in this way makes tracking the history of particular database entities much easier.

Cross-References

▶ Data Integration and Visualization

References

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A et al (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37(Database issue):D211–D215

UniProt Consortium (2011) Ongoing and future developments at the universal protein resource. Nucleic Acids Res 39(Database issue):D214–D219

Database Code

▶ Database Identifier

Database ID

▶ Database Identifier

Database Identifier

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

Synonyms

Database code; Database ID; DBID

Definition

A database identifier is a short code or name that is used to uniquely and reliably identify a particular entry or record within a particular database. The code normally contains alphanumeric characters (but may also sometimes contain other symbols), and, by contrast with database accession numbers, is usually designed to be human readable or, at least, human decipherable. For example, OPSD_SHEEP is the code that identifies the entry for ovine rhodopsin in the UniProtKB:Swiss-Prot (UniProt consortium 2011) protein sequence database: Here, OPSD is a shorthand that identifies the protein "rhodopsin"; SHEEP is, self-evidently, a label that identifies the species in which this particular rhodopsin is found.

As already mentioned, identifiers are database specific, and different databases adopt different

naming conventions. Hence, for example, in the PIR protein sequence database, ovine rhodopsin has identifier OOSH.

Information pertinent to ovine rhodopsin, which belongs to a superfamily of G protein–coupled receptors (GPCRs), may also be found in protein family databases like InterPro (Hunter et al. 2009), PROSITE, PRINTS, Pfam, and so on: Examples of identifiers for the GPCR superfamily entries in these databases are 7TM_GPCR_Rhodpsn, G_PROTEIN_RECEP_F1_1, GPCRRHODOPSN, and 7TM_1, respectively. Although some of the identifiers here are more cryptic than others, in each case, the code is broadly readable or decipherable, each in some way pointing to 7TM proteins, to GPCRs and/or to rhodopsins.

Although database identifiers were intended to provide a stable means of tracking down particular database entries, in practice, they can and do change between database releases. Thus, for example, prior to March 1993, the Swiss-Prot identifier for ovine rhodopsin, OPSD_SHEEP, was OPSD$SHEEP. Less subtle and more frequent changes also occur; hence, for example, in March 1993, the Swiss-Prot identifier for DNA replication licensing factor MCM4 was CD21_YEAST: This changed to C21H_YEAST, CC54_YEAST, CDC54_YEAST, and finally to MCM4_YEAST in 1994, 1995, 2005, and 2011, respectively. This kind of identifier volatility can make tracking particular database entities problematic, and is partly why additional forms of identification, in the form of accession numbers, are also vital.

Cross-References

► Data Integration and Visualization

References

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A et al (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37(Database Issue): D211–D215

UniProt Consortium (2011) Ongoing and future developments at the universal protein resource. Nucleic Acids Res 39(Database issue):D214–D219

Database of Quantitative Cellular Signaling (DOQCS)

G. V. Harsha Rani and Upinder S. Bhalla
Neurobiology, National Center for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India

Synonyms

DOQCS: Database of quantitative cellular signaling; GENESIS: General neural simulation system; MATLAB: Matrix laboratory; MIRIAM: Minimal information required in the annotation of models; MySQL: My structured query language; ODE: Ordinary differential equation; PHP: Hypertext preprocessor; URL: Uniform resource locator

Definition

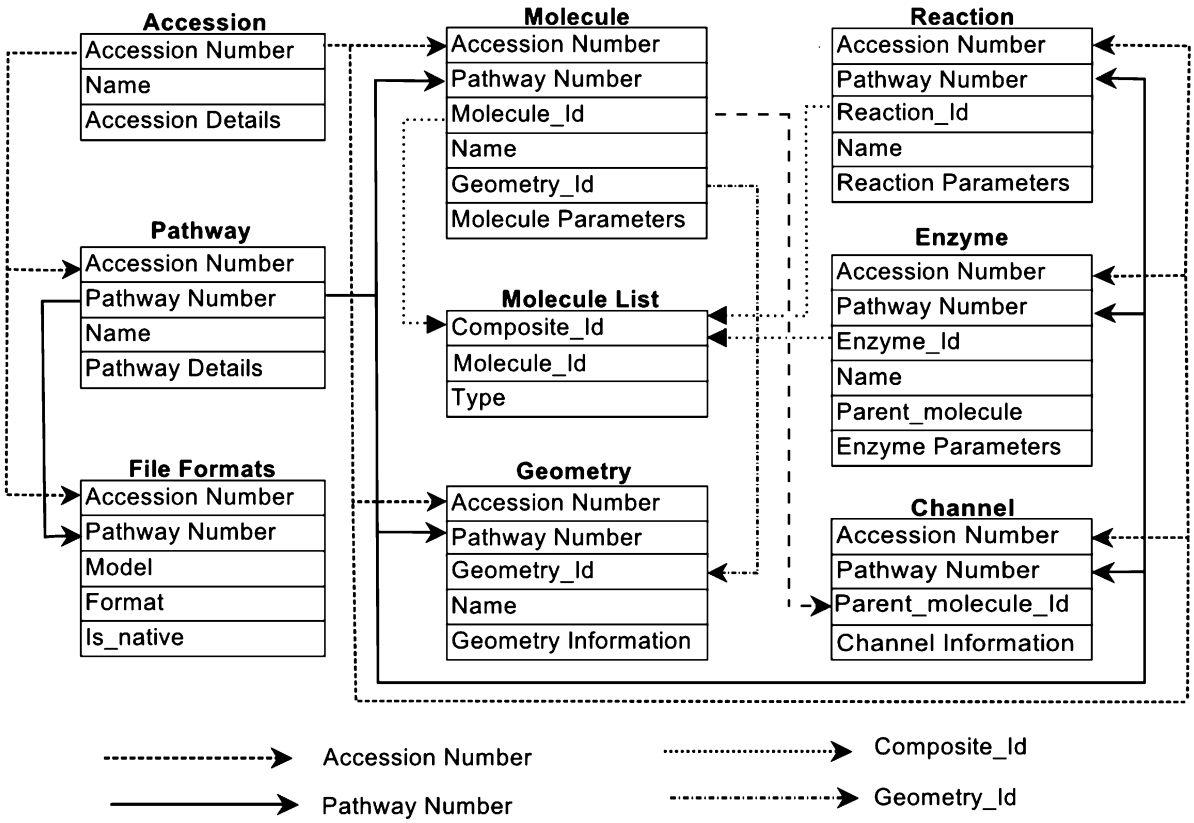
The Database of Quantitative Cellular Signaling (DOQCS) is a repository of models of signaling pathways available at http://doqcs.ncbs.res.in. This is a curated database in which all models have been implemented and tested. The database is free for content access and download. Models are presented in various data formats, and each entry includes reaction diagram, annotation, and links to other models and literature.

Characteristics

Role

Technical advances in biological information capture have yielded intimidating quantities of data pertaining to many areas of biology, including biochemical signaling. In parallel, continuing developments in software and simulators have helped researchers to develop and explore increasingly detailed biological models of complex signaling pathways. Model databases are an emerging category of bioinformatics tools that serve the intersection of these trends (Ghosh et al. 2011).

DOQCS was designed to integrate several modeling requirements to provide a resource for model



Database of Quantitative Cellular Signaling (DOQCS), Fig. 1 Entity Relationship diagram for the DOQCS database. The accession table is used as an entry point into the model. The accession number is the unique identifier for each model. Most other tables in the database refer to this (*broken line*). Similarly, each pathway has a unique pathway number used by other tables (*the black line*). Model parameters are stored in the molecule, enzyme, and reaction tables. The identity of reactants (substrates and products of reactions and enzymes) and hence the connectivity of the model is stored in the "Molecule List" table. The channel and geometry information are stored in respective tables. Original and converted model files are stored as large text objects in the "File Format" table

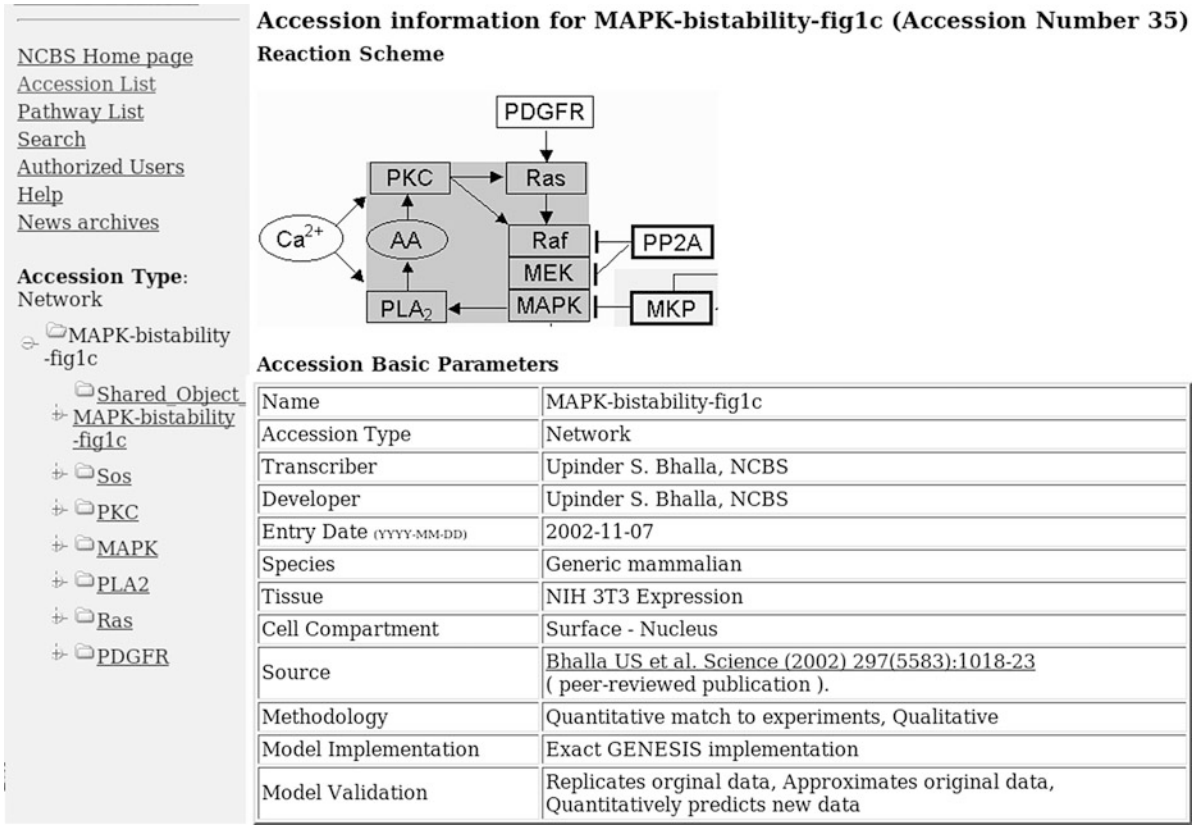
development. It includes a collection of models of signaling pathways, but has a special emphasis toward organizing the data both in terms of schematic descriptions as well as searchable model data. It provides reaction schemes and associated rate constants, enzyme parameters, concentration terms, as well as contextual data including data sources and tissue type.

Scope

DOQCS includes compartmental chemical kinetic models solved using systems of ▶ ordinary differential equations (ODE), and also using stochastic methods such as the ▶ Gillespie Stochastic Simulation (Gillespie 2007). It also includes a few one-dimensional spatial models which have been expanded out into large ODE systems by treating inter-compartment diffusion as a reaction term. DOQCS is particularly rich in models on synaptic plasticity and MAPK signaling.

Database Structure

DOQCS is implemented using ▶ relational database; it is structured as a set of accessions, each of which represents a complete model. Accessions may consist of one or more signaling pathways, and each pathway is specified in terms of several molecules, enzymatic reactions, and binding reactions. This conceptual structure has been previously described (Sivakumaran et al. 2003) and is briefly recapitulated here. The underlying table structure has been revised based on additional navigation and data requirements considered in this paper (Fig. 1).



Name	MAPK-bistability-fig1c
Accession Type	Network
Transcriber	Upinder S. Bhalla, NCBS
Developer	Upinder S. Bhalla, NCBS
Entry Date (YYYY-MM-DD)	2002-11-07
Species	Generic mammalian
Tissue	NIH 3T3 Expression
Cell Compartment	Surface - Nucleus
Source	Bhalla US et al. Science (2002) 297(5583):1018-23 (peer-reviewed publication).
Methodology	Quantitative match to experiments, Qualitative
Model Implementation	Exact GENESIS implementation
Model Validation	Replicates orginal data, Approximates original data, Quantitatively predicts new data

Database of Quantitative Cellular Signaling (DOQCS), Fig. 2 Screenshot of overview page of DOQCS accession, showing directory-tree-like hierarchy structure, block diagram and basic annotations

As before, DOQCS is implemented using Linux/Apache/MySQL/PHP.

The accession table specifies the primary model entry into the database and holds substantial contextual data as well as a summary diagram for each model. The contextual data include ▶ MIRIAM compliant annotations. Further fields including tissue, cell type data in addition to fields on curation level and model type annotation are added.

The Pathway table provides additional overview information about the models, including reaction diagrams and annotations.

The remaining tables provide low-level model details: molecular identity, rate constants, and substrate/product lists.

Compartmental details are stored in Geometry table.

Original and converted model files are stored as Large Text Objects in the File Formats table.

Database Interface

The web interface to DOQCS facilitates navigation either through links (URL-based navigation) or through searches (form-based navigation) (Fig. 2).

The URL-based navigation matches the conceptual organization of the database into accessions, pathways, and biochemical entities. These levels are explicitly represented using a directory-tree-like hierarchy in the web interface. The root of the tree is the accession, folders include different pathways within the accession, and the molecules, reactions, and enzymes are represented as entries within the pathways. A click on any level of the tree brings up the details pertaining to that level of the database, as mentioned above.

Overview data is provided on the accession and pathway pages, and detailed biochemical parameters are found on the reaction and molecule pages.

Each of these pages includes richly hyperlinked results from the navigation. For example, each pathway page has links to all related pathways in the database. Further, each biochemical entity page has links to all the reactions or other molecules that interact with it.

The second major navigation option is through searches. This uses a query-based form present at the top of each page of DOQCS. In additional to simple queries as previously described (Sivakumaran et al 2003), additional several complex searches are possible for textual pattern matching within various subfields, as it is common on most databases. More sophisticated searches can also be carried out based on connectivity or functional criteria. Typical connectivity criteria specify that a given molecule is a substrate of another. A functional criterion might be that a given molecule acts as an enzyme.

Data Sources and Curation

Except for few models that were explicitly developed for the database by the curation team, all other models are published models which have been tested before putting online and in the accession table citation information is included as a link to PubMed. This procedure frequently reveals ambiguities in model representation, and in some complex models, the original implementation is difficult to replicate. In all cases, the extent of the fit between published and internally tested representations is reported as an important entry in the validation field in the table.

File Formats

DOQCS supports three file formats: Kinetikit, SBML, and MATLAB. All models in the database have been implemented using GENESIS/Kinetikit simulator (Bhalla 2002; Vayttaden and Bhalla 2004) and are available in its internal model file format which includes annotation information. In cases where the original publication includes model files, these are also presented unchanged for download. Most models have also been converted to MATLAB and SBML. All converted models have been tested for equivalent function to the GENESIS/Kinetikit form.

Cross-References

- ▶ Gillespie Stochastic Simulation
- ▶ MIRIAM
- ▶ Ordinary Differential Equation (ODE)
- ▶ Relational Database
- ▶ Systems Biology Markup Language (SBML)

References

Bhalla US (2002) Use of Kinetikit and GENESIS for modeling signaling pathways. Methods Enzymol 345:3–23

Ghosh S, Matsuoka Y, Asai Y, Hsin KY, Kitano H (2011) Software for systems biology: from tools to integrated platforms. Nat Rev Genet 12(12):821–832

Gillespie DT (2007) Stochastic simulation of chemical kinetics. Annu Rev Phys Chem 58:35–55

Sivakumaran S, Hariharaputran S, Mishra J, Bhalla US (2003) The database of quantitative cellular signaling: management and analysis of chemical kinetic models of signaling networks. Bioinformatics 19(3):408–415

Vayttaden SJ, Bhalla US (2004) Developing complex signaling models using GENESIS/Kinetikit. Sci STKE 219:l4

Database Search

- ▶ Protein Identification Analysis

Databases for Kinetic Models

Jacky L. Snoep
Department of Biochemistry, Stellenbosch University, Matieland, South Africa
Manchester Institute for Biotechnology, University of Manchester, Manchester, UK
Molecular Cell Physiology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Definition

Over the last decade, systems biology (SB) has been one of the fastest growing fields in the life sciences. There are many interpretations and definitions for the

SB field (Westerhoff and Alberghina 2005), but most scientists agree that an SB study combines experimental and theoretical (including modeling) approaches. As such the increase in number of SB studies correlates with the increase in mathematical models for biological system, which is evident from a recent inventory (Huebner et al. 2011). Not only did the number of models increase, also the size of the models has increased. The latter effect is due to the larger size of the systems being studied but more importantly due to a more complete representation of the biological system in the models.

The aim of many SB studies is to relate systems behavior to characteristics of the components of the system. This is a subtle but important difference from the classic theoretical biology approaches, where core models were constructed to describe biological behavior using mathematical equations that were as simple as possible, not necessarily reflecting biological mechanisms. Specifically in so-called bottom-up SB models, a strong mechanistic link is maintained between model components and biological entities. Such models can be used to critically test our biochemical knowledge of a given system (e.g., Teusink et al. 2000).

The increase in number and size of kinetic models and the more realistic representation of biological components in these models have led to the development of a number of model database initiatives. In this section, descriptions of several of these initiatives are given and the advantages of storing models in such databases and the consequences for model reuse, annotation, and dissemination are discussed.

Characteristics

Why Model Databases

The increased size of kinetic models in SB studies necessitates the storage of the models in a publically accessible form. Whereas it is easy to code a two-variable model from a publication, the effort and chances of making errors becomes much larger with increased model sizes. Clearly, if a model were available in a repository from which it can be downloaded and used without recoding the model, this would save a lot of work and would also eliminate coding errors (if the model in the database were properly curated).

Models have increased in size for two reasons: firstly because larger systems are being studied and secondly because the models are more detailed. In some cases, the modelers are actually attempting to build replicas of the real system. In the latter case, where the model variables have a well-defined mechanistic interpretation, it is important to annotate the model. The information in such annotated models is ideally suited for storing in relational databases.

Strong searching capabilities, easy access and dissemination via standard formats and protocols, linking to other databases, and possibilities of incorporation of automated workflows are just a few of the added advantages of storing models in databases.

Minimal Requirements for Model Databases

A model repository or database must fulfill three minimal criteria to make it useful. Firstly, the models' descriptions must be correct, i.e., they must be identical to the model description in the publication where the model was first presented (see ▶ CellML Model Curation). Secondly, the models must be available for download from the repository in a standard model description format, such as ▶ Systems Biology Markup Language (SBML) or ▶ CellML. Thirdly, the models must be annotated (model annotation). A minimal annotation should link the model to the reference where it was published and it should be clear how the model variables and parameters link to the species and constants in that publication.

A nice-to-have functionality for kinetic model repositories is a simulation tool, such that the models in the repository can be directly inspected and simulated, for instance in a web browser. The first initiatives that stored kinetic models of biological systems started out as model repositories with a simulation engine (see ▶ JWS Online and the ▶ Virtual Cell (VCell) Modeling and Analysis Platform).

A large number of model databases have been initiated in the last decade. Some of these initiatives focus on a specific subset of models, such as models for signal transduction pathways (▶ DOQCS: Database of Quantitative Cellular Signaling), for the cell cycle (▶ Cell Cycle Database), for neuronal models (modelDB), while other databases are more general repositories (▶ BioModels Database: a repository of mathematical models of biological processes, ▶ JWS Online, ▶ CellML, ▶ Virtual Cell, ▶ WebCell).

Model Annotation

Whereas the focus in theoretical biology studies used to be on the mathematical aspects of a model, i.e., on the analytical and numerical analyses, in systems biology studies the direct link to experimental data is much stronger and it therefore becomes more important to relate model components to biological entities, i.e., to annotate models.

In addition to a minimal reference annotation to link models in the database to the scientific publication in which the model is described, a further annotation of model components to biological entities using controlled vocabularies (e.g., SBO terms) greatly enhances the application strength of mathematical models. The annotation of models makes it much easier to search for specific components in models, or to compare or even link models for overlapping systems.

The Minimal Information Required in the Annotation of Models (▶ MIRIAM) is an example of guidelines for annotating models, and it links model variables to ontology terms (e.g., to a systems biology ontology, SBO term) and thereby to unique identifiers. Thus, where modelers can use names as G6P, Glu6P, Glc6p, or *X*2 to denote a variable such as glucose 6-phosphate, if such variable names are linked to an identifier such as a ChEBI number they can all be related to glucose 6-phosphate. This annotation relates model constituents unambiguously to a known entity, where the references information is given in a {"data type," "identifier," "qualifier"} triplet. The "data type" is given as a Unique Resource Identifier and can be an Uniform Resource Locator (url) or a Uniform Resource Name (urn), for instance for a model variable Ca2+ which is a reactant in a reaction the data type could be a url: "http://www.ebi.ac.uk/chebi/" with identifier: "CHEBI:29108" and qualifier: "is." MIRIAM extends beyond the annotation of model components; it also requires a reference correspondence, standard model format, and the possibility of instantiation of a model simulation.

Biomodels (▶ BioModels Database: a repository of mathematical models of biological processes) was one of the first model database initiatives to strongly promote model annotation and has played an important role in development in many model description, annotation, and simulation standards.

What Information Should Be Stored in Model Databases?

For each model entry in the database, its description file in a standard format and an annotation file would be the minimal information to be stored (in SBML, the two can be combined in a single xml file). In this section, the focus is on models described as deterministic ordinary differential equations, which is the class in which the majority of the kinetic models in SB studies are described. For a more general treatment, see the section on kinetic models (▶ Kinetic Modeling and Simulation).

A typical model description would include the following: (1) description of a set of reactions, (2) rate equations for the reactions, (3) parameter values, and (4) initial conditions. In addition, models can contain events, functions, and algebraic equations, but the above four components are generic. The reaction description and the rate equations can be combined to give a set of differential equations, and some models are defined in differential equations only, without specifying the reactions and their rate equations. The standard model description formats allow for these different ways to describe kinetic models. Whether a model is described in terms of reactions or as ODEs does not influence the numerical integration (since these are performed as differential equations), but the model description can affect the functionality of the model.

The Importance of Specifying Reactions

If a model is described in ODEs without explicit reference to the reactions in a system, it is not always possible to dissect the contributions of the individual reactions to an ODE.

The famous Lotka-Volterra model for predator-prey interaction might serve as an example to illustrate the point. This model is usually presented in the form of ODEs:

$$d(x[t])/dt = x[t] * (a - b * y[t])$$

$$d(y[t])/dt = y[t] * (c * x[t] - d)$$

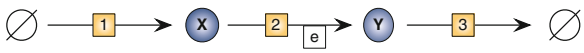
with x[t] and y[t] the densities of prey and predator, respectively, and a, b, c, and d positive parameters. From these ODEs, it is not immediately clear that the system consists of three reactions: (1) a birth rate of prey (a*x[t]), (2) a death rate of predator (d*y[t]), and

(3) a rate for prey consumption by predator leading to a decrease of prey ($-b^*y[t]^*x[t]$) and an increase of predator ($c^*y[t]^*x[t]$). From the ODEs, it cannot be automatically derived that (3) is a single reaction with a different stoichiometry for prey consumption and predator growth. If the ODEs would be given in terms of reactions this uncertainty would not exist, for instance:

$$d(x[t])/dt = v1 - v2$$

$$d(y[t])/dt = e^*v2 - v3$$

with $v1 = a^*x[t]$, $v2 = b * y[t] * x[t]$, and e = c/b, and v3 = d * y[t]. When the ODEs are given in terms of reactions, a reaction network can automatically be drawn:



The Lotka-Volterra model is not a mechanistic model and the translation to explicit reactions is not so important. In addition, it is not difficult to make the translation into reactions with some background information. However on the basis of the ODEs only, a computer cannot make the translation to the network automatically. For larger models, such a translation becomes much more difficult, even for humans. See Conradie et al. 2010, for an example of a study where considerable effort was made to translate a model defined in ODEs back to the original reactions for which it was defined.

If a model is defined in terms of reactions, it is easier to identify the biological system it refers to and to search for components or compare (parts of) models. A reaction is defined in terms of substrates and products and a unique identifier can be given for a reaction (e.g., KEGG, see entry ▶ KEGG pathway database; www.genome.jp/kegg/). If the process is enzyme catalyzed, as is the case for most reactions in the cell, a corresponding E.C. number exists. Such identifiers are important as they can give information on thermodynamic constants (e.g., Keq for a reaction), and they can give indications for enzyme kinetic constants as stored in enzyme kinetic databases. Furthermore, by defining a mathematical model in terms of reactions it is possible to describe the modeled system in a standard format. BioPAX (Biological Pathway Exchange; www.biopax.org/) is a standard language to represent biological pathways and is widely supported by database initiatives.

Thus, although the traditional way of defining models in terms of ODEs without explicit reference to reactions has no consequence for the time integration of the model, it does affect the functionality of the model both in terms of comparing to other resources and in terms of analyses that can be made with the model. For instance, automated drawing of reaction network graphs or ▶ metabolic control analysis (MCA) can only be performed for models defined in terms of reaction steps.

Rate Equations

For some model organisms, a fairly complete set of reactions occurring in a cell has been constructed (at least for metabolism). Such structural models, which define a reaction network, have been defined for much larger system than have been used for kinetic models. The reason for this is that the information needed for building a kinetic model is more extensive and not as simple and condition independent as for structural models.

For well-studied enzymes, a kinetic mechanism might be known and then a rate equation can be derived from the mechanism. However, for many enzymes no mechanism is known and then often a generic rate equation is used based on a simplified mechanism such as a random order rapid equilibrium binding mechanism. If the enzyme can be studied in isolation, kinetic parameters can be estimated by fitting the rate equation on the experimental data set for the isolated enzyme. If the enzyme cannot be studied in isolation, the kinetic parameters are often fitted on behavior of the complete system. In the latter case, it is much more difficult to make specific perturbations to the enzyme and even simpler rate equations must be used since the parameters are often not identifiable.

The above paragraph indicates some of the difficulties of obtaining suitable rate equations for kinetic models, which are treated in more detail in the sections on model construction and ▶ model validation. For this section, it is sufficient to know that for many systems we can define the reaction network with good confidence and can store this information well in a database. But for the equations that can be used to describe the rate of the reactions, the situation is not so simple.

Although the number of possible rate equations that can be used in mathematical models might appear limitless, the situation is not quite that bad. Due to the limited number of substrates and products for most reactions, the possible combinations, even when taking several mechanisms into account, is not that big. Thus, libraries for rate equations have been constructed, and when storing models in databases a reference to such a library can be made if a standard mechanism was implemented for a reaction. In practice many models will also include nonstandard rate equations that are derived specifically for a reaction, and these must be included in the model storage as well. Some model simulators such as COPASI (www.copasi.org/) refer to an internal rate equation library. When building a model, the user can select a rate equation from the library or use a user-defined reaction (that is then added to the library). When exporting the model from the simulator, each equation is given explicit in the SBML file.

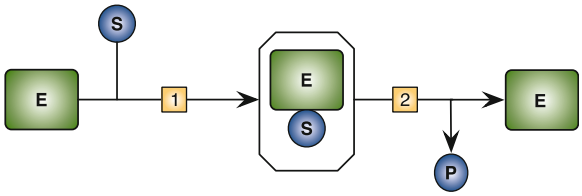
Kinetic Parameters

One of the advantages of using rate equation libraries is that it is possible to give identifiers to each of the model parameters. In much the same way that modelers pick names for variable species, they also use names for parameters that cannot always be uniquely related to known constants. If parameters in rate equations were annotated, it would be possible to relate them to existing kinetic constants and make links to database resources.

Unique identifiers for small chemical species and for enzymes are quite well accepted and used, but for rate equations this is not so common. This makes it harder to store kinetic information in a database and limits the strength of database searches and functional comparisons with other rate equations or parameter values. The situation is worse for reactions for which we do not know the kinetic mechanism as now different rate equations can be used to describe the same reaction. The functional behavior of different rate equations might be similar (this is to be expected if they were constructed from the same data set), and would therefore not lead to very different model simulation results. However, the different equations make it hard to compare (or use) kinetic parameters obtained for different rate equations.

This touches on an important aspect of one of the advantages of using databases for model storage: interactivity and exchange between different databases. Whereas molecules, reactions, and enzyme identifiers are largely species and condition independent, this does not hold for rate equations or parameter values. Thus, the strong point of connecting databases to search for kinetic parameters for rate equations should be used with caution and might be difficult to automate.

A simple example might make the problem clearer. Let us consider the well-known Michaelis-Menten equation for the enzyme-catalyzed conversion of S to P



$$V = Vm^{*}S/(Km + S)$$

For the original derivations, Michaelis and Menten assumed equilibrium binding of enzyme and substrate and then the Km value has a mechanistic interpretation as the dissociation constant (Kd) of the enzyme-substrate complex (Km = Kd = kd/ka; with kd the rate constant for the dissociation reaction and ka the rate constant for the association reaction). In a later derivation, Briggs and Haldane relaxed the equilibrium binding assumption to a quasi-steady-state approximation for the enzyme-substrate complex, which gives a different mechanistic interpretation for the Km value (Km = (kd + kcat)/ka). Although the assumptions for the derivation and the resulting interpretation of the Km value are different, the equation for the description of the enzyme activity is identical.

Interestingly, in practice the Km value is determined as the substrate concentration giving half-maximal reaction rate, irrespective of the mechanism or assumptions. The enzyme characteristics and the experimental conditions determine whether the equilibrium binding assumption or the quasi-steady-state approximation assumption holds (if any). However, the mechanistic interpretation of the Km value is independent of the experimental determination, and many researchers have become a bit careless in using and reporting Km values. If the Km value is operationally defined as the substrate concentration giving half-maximal enzyme

activity, this can still work fine. However, for multiple substrate/product reactions, the rate equation becomes dependent on the kinetic mechanism used, and for ordered binding mechanisms the definition of the inhibition constants is related to the mechanism. For such reactions, it is important to publish the kinetic rate equation together with the Ki values.

Irrespective of the mechanistic interpretation of kinetic parameters, their values are usually dependent on the assay conditions under which they were determined. The kinetic parameter values should be determined under conditions that closely reflect the conditions that are simulated in the model. Where a model simulates the intracellular cytosol, it is hard to imitate those conditions ex vivo. Often kinetic parameters are determined in vitro, and some guidelines for mimicking cytosolic conditions have been formulated (van Eunen et al. 2010).

The contents of databases are dependent on the way the data are represented in the scientific literature. For the representation of enzyme kinetic data, standards have been formulated by the ▶ STRENDA commission (Standards for Reporting Enzymology Data, http://www.beilstein-institut.de/en/projects/strenda/).

A number of databases exist where enzyme kinetic data are collected. This might seem in conflict with the above advice to use published kinetic parameters with caution, but these database initiatives give information on the following: (1) the experimental conditions under which the parameters were determined, (2) the associated rate equation, and (3) the experimental data used for the parameter determination (e.g., SABIO-RK). Some kinetic parameters such as Keq and kcat are relatively constant (e.g., not sensitive to a particular kinetic mechanism assumed in its estimation), when they are measured under the correct physiological conditions. Other kinetic parameters, such as Vmax values can vary largely as they are not only dependent on the assay conditions but also on expression level of the enzyme, i.e., on the growth conditions of the organism.

The relevance of this long section on enzyme kinetic parameters for model databases is that databases that store the kinetic parameters will facilitate the reuse of the parameter values. There is a significant risk involved in reusing parameter values, if the user does not check carefully how the parameter values were determined. A close link of the parameter values and rate equations to experimental data seems necessary.

Linking Models to Data Sets

One can roughly distinguish two types of data sets that are connected to mathematical models: data for model construction and data for model validation. In principle these should be independent data sets, and they can be very different. For instance in a typical bottom-up modeling approach, it is possible to have kinetic data for each of the individual enzymes that is used for parameterization of the enzyme kinetic rate equation as part of the model construction. In such a study, data for the complete system could be used for model validation. Of course, this is just a scenario and other types of data sets are possible. In many studies it is not possible to separate the model components functionally from the system, and then one cannot characterize the individual components in isolation.

Because the bottom-up approach nicely separates the model construction data sets from the model validation data sets, I will expand a little further on how such data sets can be linked to model storage in databases. The first question to ask is whether experimental data should be stored in a model database. On the one hand, one can argue that the data are not part of the model description, and are not necessary for model simulation or analysis and could therefore be kept separate from the model. On the other hand, the data for model construction are closely linked to the model and can enhance the model functionality significantly. Firstly, if all data for model construction are made available, the model construction becomes a completely transparent process and could be reproduced by other scientists, which is an important aspect of scientific research. Secondly, it enables researchers to make changes to the model, for instance using the same experimental data sets but applying different rate equations. A last advantage to explicit linking of the data sets to the model is that it makes it possible to completely separate the construction and validation data sets.

The above given advantages should be sufficient motivation to link experimental data sets to the experimental data that was used for the model construction, but it does not provide with a strong argument to store the data in the same database as where the models are stored. For instance it would be possible to store all the experimental data in a specialized database for enzyme kinetics, such as SABIO-RK (sabio.villa-bosch.de/) or Brenda (http://www.brenda-enzymes.info/), and then link the model to the external database. Whereas such

a link to external databases is easily made, it is sometimes more convenient to also store the experimental data in the model database. For instance, several of the model databases (e.g., JWS Online and Biomodels) are part of data management systems of large research projects, and then it can be advantageous to keep the data in a secure database and have direct access to the data without external links. For instance, JWS Online is used in active model development where for each reaction there is a data set available that is directly available via the network schema. These data sets are sometimes updated and linked to versioned models. Only the final model is released to the public and then the data is made available.

Clearly, different solutions can be used to store models, kinetic parameters, rate equations, and the associated experimental data. It is important that a link between the model and the experimental data is made explicit and that it can be followed, how the link is made is not so important. In some of the above-mentioned systems biology projects, in addition to experimental data there is a lot of additional information stored and often a much greater functionality is provided in a complete data management structure.

Data and Model Management Structures

The multidisciplinary character of systems biology projects and the scale of many of these projects necessitate the collaboration between sometimes-large numbers of research groups. Although each research project could in principle come up with its own unique solution for data and model management, it would make more sense to develop more generic tools that can be used by many research groups. An example of such a data management system that is being developed centrally is the data management group for the ▶ SysMO projects (www.sysmo.net/). The SEEK (▶ Data and Model Management Platform, SEEK) (Wolstencroft et al. 2011) is the centrally developed software package that gives a wide functionality to each of the individual research groups, one of which is the storage and curation of mathematical models, but it also makes it possible to make explicit links between models and experimental data.

The SEEK has been a large success and is now also implemented in several other SB projects. SEEK can be downloaded, installed, and adapted to anyone's needs, and this has helped strongly in the package being incorporated in, e.g., the UniCellSys project (www.unicellsys.eu/), EviMalaR (www.evimalar.org/), the Virtual Liver (www.virtual-liver.de/), and many more research projects.

Web-Services and Workflows

One of the biggest advantages of using a database for storing data is the possibility to structure the data according to its expected use. One can store the data such that expected queries run most efficiently. It is a relatively simple step to not only structure the data storage but also the way in which queries are made and output is given. Most databases allow access to their data via so-called web-services (see ▶ Web Service), which are structured queries. There are different types of web-services, which are treated in more detail in another section. For the model databases, such web-services can be a simple search query to select all models for a certain organism, or pathway, but it can also involve running simulations of models that have been selected before. Such workflows that run a set of web-services in succession can be defined in software tools specifically designed for this task, such as Taverna, or in more generic programs such as Mathematica. Workflows are very powerful tools that automate tedious and repetitious jobs.

An important aspect of web-service construction is the formulation of the specific format in which the query must be made and the answer will be given. Recent initiatives have focused on the formulation of standards for model simulation descriptions. MIASE (Minimal Information About a Simulation Experiment) proposes a minimal set of information needed to reproduce simulation experiments, and SED-ML (Simulation Experiment Description Markup Language) encodes this information.

The Silicon Cell Initiative

Model databases greatly enhance the accessibility of published kinetic models. The storage of curated models is important to prevent losing the models, which are often custom coded, and to make the models publicly available in standard formats. The reproducibility of (model simulation) results is an important aspect of the scientific process and this is ensured in the curation process of the model databases.

Model accessibility is also important for future model reuse. Whether a specific model will be reused is largely dependent on how it was constructed. If a model was made to address a specific research question, the model structure and kinetic parameters

might be very dependent on the specific conditions that were applied. Then model reuse might be limited to these conditions. However, if a model was constructed using the above-discussed bottom-up approach, using experimental data for each of the individual reaction steps, the model (or parts of it) can be reused. If the kinetics of the individual reaction steps were measured under physiological conditions, they are to a large extent model independent. This holds for the kinetic mechanism and the kinetic parameters but to a lesser extent to the Vmax values (which are dependent on the enzyme expression levels). Since the enzyme expression levels will be dependent on the growth conditions, these could very well be model dependent.

The ▶ silicon cell initiative (e.g., Snoep and Westerhoff 2004) advocates the use of rate equations based on kinetic parameters that were experimentally determined for the individual reactions. Construction of kinetic models with such rate equations followed by an independent model validation for the system would lead to kinetic models that can be reused in a modular approach to building models of larger systems. Whereas it is unlikely that a detailed kinetic model for a whole cell can be constructed in a single step with a bottom-up approach, the Silicon Cell approach would be to validate models for parts of the cell and then merge them to simulate a larger part of the system. After merging, the resulting model can again be validated. In such a modular approach, a gradual increase in model size will prevent the accumulation of experimental error in the individual model parameters. Model validation is crucial in this approach, and the definition of modules should largely be determined by whether they can be validated, i.e., define a module such that it can be validated.

In this section, several aspects of model databases were introduced and discussed. A number of these database initiatives will be highlighted in assays and several additional definitions will be formulated.

Cross-References

- ▶ BioModels Database: A Repository of Mathematical Models of Biological Processes
- ▶ Cell Cycle Database
- ▶ CellML
- ▶ CellML Model Curation
- ▶ Data and Model Management Platform, SEEK
- ▶ DOQCS: Database of Quantitative Cellular Signaling
- ▶ JWS Online
- ▶ KEGG Pathway Database
- ▶ Kinetic Modeling and Simulation
- ▶ Metabolic Control Analysis
- ▶ Model Validation
- ▶ Silicon Cell
- ▶ STRENDA
- ▶ SysMO
- ▶ Systems Biology Markup Language (SBML)
- ▶ Virtual Cell (VCell) Modeling and Analysis Platform
- ▶ WebCell
- ▶ Web Service

References

Conradie R, Bruggeman FJ, Ciliberto A, Csikász-Nagy A, Novák B, Westerhoff HV, Snoep JL (2010) Restriction point control of the mammalian cell cycle via the cyclin E/Cdk2:p27 complex. FEBS J 277:357–367

Huebner K, Sahle S, Kummer U (2011) Applications and trends in systems biology in biochemistry. FEBS J 278:2767–2857

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

Teusink B, Passarge J, Reijenga CA, Esgalhado E, Van der Weijden CC, Schepper M, Walsh MC, Bakker BM, Van Dam K, Westerhoff HV, Snoep JL (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. Eur J Biochem 267:5313–5329

van Eunen K, Bouwman J, Daran-Lapujade P, Postmus J, Canelas AB et al (2010) Measuring enzyme activities under standardized in vivo-like conditions for systems biology. FEBS J 277:749–760

Westerhoff HV, Alberghina L (eds) (2005) Systems biology, definitions and perspectives, vol 13, Topics in current genetics. Springer, Berlin, pp 13–30

Wolstencroft K, Owen S, du Preez F, Krebs O, Mueller W, Goble C, Snoep JL (2011) The SEEK: a platform for sharing data and models in systems biology. Methods Enzymol 500:629–655, Chapter 29

Data-Driven Science

- ▶ Data-intensive Research

Data-Intensive Research

Sabina Leonelli
ESRC Centre for Genomics in Society, University of Exeter, Exeter, Devon, UK

Synonyms

Computational data analysis; Data deluge; Data-driven science; Data-intensive science

Definition

Data-intensive research can be characterized as the attempt to extract biological knowledge from the huge amounts of data produced through experiments and high-throughput technologies (e.g., new generation ▶ DNA sequencing) and disseminated through cyberinfrastructures (e.g., community databases and ▶ Bio-Ontologies). Data-intensive research encompasses a wide variety of scientific methods, whose common feature is to rely on the accumulation and sharing of evidence on a large scale and across research contexts as a starting point for the research process. Also central to data-intensiveness is the idea of automated data analysis, defined as the extraction of biologically significant patterns from data through computational means, with as little human intervention as possible (see also ▶ Automated Reasoning). Computational tools for ▶ data mining are expected to facilitate the generation of new hypotheses and thus to help identify fruitful research directions, which can be explored further through in vivo experimentation. At the same time, both champions and critics of data-intensive research recognize that data analysis cannot be understood as purely inductive and that the human input and skills involved, such as the ability to interpret data, construct models, and formulate hypotheses, cannot be fully automated. Data-intensive research is thus not one single, inductive, computer-driven method for discovery. Rather, it encompasses a variety of methods of data analysis, all of which rely on iterative feedback between ▶ experimentation in vivo and the consultation of data available in silico; and between inductive, deductive, and explorative reasoning (Kell and Oliver 2004; O'Malley et al. 2009).

Characteristics

What does it mean for research to be based on empirical evidence? This question, one of the oldest within the philosophy and history of science, is being reformulated and reconsidered within contemporary biological and biomedical science. In these areas, and particularly within system biology with its emphasis on data sharing and interdisciplinary integration, technological innovation and shifting ideas about what counts as evidence have transformed practices of data collection, dissemination and analysis, with profound methodological consequences for experimental and modeling practices. This entry sketches some characteristics of this broad trend.

The Data Deluge

The activities of data gathering and data use appear to have acquired relative independence from other scientific activities such as hypothesis-testing, modeling, and explanation. Up to the second half of the twentieth century, biological data were largely produced as evidence to support a specific experimental hypothesis. This is still the case in several research areas, but not within molecular and system biology, where high-throughput technologies such as sequencing and micro-array experiments have changed the way in which data are produced. In these fields, the activity of data gathering has become increasingly automated and technology-driven, resulting in the production of billions of data-points in need of a biological interpretation (Hey et al. 2009). Consequently, massive research efforts are being devoted to the dissemination of data online, in the hope that free and widespread access to large datasets will enable scientists to use them to understand biological phenomena, thus generating new paths toward discovery.

Thanks to the variety of computational tools developed to collect, store, and distribute them, data are now available to researchers on an unprecedented scale. This partnership between biology and computer science constitutes both the strength and the weakness of data-intensive research. Several commentators have argued that the extraction of knowledge from such large, cross-disciplinary datasets constitutes a new scientific method, often depicted as "data-driven." The underlying idea is that data already available online constitute formidable sources of insight, which can be used to generate research programs without

necessarily starting from a specific hypothesis to be tested and without necessarily possessing the same expertise as the original data producers (Kell and Oliver 2004). At the same time, it has become clear that the sheer scale and diversity of data to be analyzed requires the creation of sophisticated tools for data mining, which in turn needs to be informed by relevant expertise in the theory and practice of all the relevant domains within biology (Blake and Bult 2005, Buetow 2005).

Three Data-Intensive Research Methods

The following three cases provide good examples of the advantages and limitations of data-intensive research methods:

- Discovery through triangulation of existing evidence
  The opportunity to access data through a web of interlinked repositories and databases is enabling scientists to retrieve more and more data of possible relevance to their research interests. It is now possible to gather and integrate data obtained on a wide variety of organisms by laboratories across the globe, no matter the specific expertises and interests guiding the production of data at each location. This is particularly true in the case of research on ▶ model organisms, where researchers can retrieve large portions of the data available on the same set of phenomena through community databases. This unrivaled level of data sharing is fueling the discovery of new regulatory roles of specific genes or pathways. The triangulation of existing evidence thus furthers and transforms existing understandings of phenomena. This research strategy is hardly new, yet the use of digital technologies makes it tremendously more efficient. It is crucial that the data in question are in a digital format, which makes it possible to disseminate them widely and retrieve them instantly.
- Discovery through data mining
  Online databases can also be searched for emerging patterns or correlations that could not have been predicted otherwise. A striking instance of this method is the idea of "random walks" through data, where software is used to mine datasets to spot *statistically* significant patterns (e.g., gene expression). It is not obvious that these patterns also have *biological* significance, and what they teach us about the biology of organisms needs to be investigated through further experimentation. Yet, computational analysis here offers a shortcut toward discovery by pointing to patterns that have at least the potential to enhance existing understandings of biological phenomena.
- Discovery through spotting gaps
  Data mining can lead to discovery by pointing researchers toward areas of investigation that were not previously charted. A good example is the discovery of "ultra-conserved regions" in vertebrate DNA and their regulatory role in development (Blake and Bult 2005); or, more generally, the discovery of correlations between the presence of specific alleles in an individual's genotype and the occurrence of a phenotypic trait, which might open the way to the investigation and discovery of mechanisms responsible for the development of that trait (this is the approach underlying genome-wide association studies). In these ways, data-intensive research helps to identify gaps in the existing knowledge about a given entity, thus opening up new areas for investigation.

Beyond Induction

As exemplified by the above cases, at the core of data-intensive research is an emphasis on the epistemic value of data beyond the experimental context in which they are originally produced. Supporters of data-intensive research stress that the way in which data can be used as evidence varies depending on the scientific context in which they are considered. This flexibility as "raw materials" of science is what makes the dissemination and integration of data across biological domains into an effective route toward discovery (Leonelli 2009). This does not necessarily mean that the accumulation and sharing of data constitutes the best starting point for inquiry, yet critics of data-intensive approaches have stressed the risk of "induction beckoning again" (Allen 2001) and emphasized that proponents of data-driven science tend to portray data as the primary source of scientific knowledge and to stress the value of inductive procedures over and above other research methods. This impression is heightened by the frequent juxtaposition of data-intensive methods to

"hypothesis-driven," deductive research, which suggests that data mining can lead to the formulation of testable claims without recourse to preconceived hypotheses (Evans and Rzhesky 2010). These interpretations of data-intensive research as purely inductive and independent from existing theoretical expectations are very problematic and untenable in the face of actual scientific practice. Reusing data for the purposes of discovery involves a complex ensemble of skills and methodologies, which go well beyond an inductive approach. Extracting biologically meaningful inferences from high-throughput genomic data involves, for instance, reliance on theories about gene expression and regulation, specific models of the biological processes being regulated, common standards for the formatting and visualization of genomic data, and familiarity with the instruments and organisms from which data were originally obtained. This methodological complexity is what makes it difficult to pinpoint the epistemic characteristics of data-driven research as a unique, emerging mode of inquiry. At the same time, methodological complexity aligns this form of research with the goals and methods favored in system biology: the pursuit of complex, interdisciplinary interactions; and the use of a variety of methods to achieve an integrated understanding of living organisms (Philippi 2006).

The Limits of Automation

Another characteristic common to the three examples of data-driven methods given above is reliance on automated data analysis: "smart" software is assigned a prominent role in facilitating the extraction of patterns from data, either through statistical analysis or through search mechanisms in databases. While automated techniques for data analysis and hypothesis generation are proliferating and becoming increasingly sophisticated, there are two good reasons to believe that biological research cannot and should not be *fully* automated:

1. Research within the field of bioinformatics, and particularly database curation, has shown that data mining processes are only reliable when resources and expertise are invested in selecting high-quality data for insertion in databases; and in building databases that make use of efficient search engines, visualization systems for data, and interoperable ▶ ontologies (Renear and Palmer 2009). Indeed, it has been claimed that data-intensive science cannot advance without substantial investment in a reliable cyberinfrastructure which can be easily updated by users, particularly when this approach is put to the service of systems biology (Philippi 2006).
2. Successful examples of data-intensive science illustrate the need for these methods to be embedded in a wider spectrum of scientific practices, ranging from theoretical analysis to ▶ experiments and field observations. Data-intensive research thrives when exposed to iterative feedback with in vivo research (O'Malley et al. 2009).

The Heuristic Value of Data-Intensive Science

The relation between data-intensive science and other forms of research is difficult to describe in simple and general terms, yet this complexity is precisely what makes data-intensive methods interesting as a new approach to scientific inquiry. Recent advances in bioinformatics and successful applications of data-driven methods have shown that the analysis of data in silico cannot be fully automated, nor should it be disjoint from experimental research in vivo. However, computational tools have the power to substantially transform how research is performed and the ways in which ▶ experiments are set up, carried out, and verified. Data-intensive science has great heuristic value, since findings emerging from the analysis of large datasets can be used to challenge and re-direct the means and targets of experimental research. This does not mean that data-driven methods should always be conceived as the starting point for experimental inquiry. They constitute resources that can potentially complement all stages of research and can thus be fruitfully applied to any project, even if in each case their function is likely to be different depending on the specific goals, context, and resources available.

Cross-References

▶ Automated Reasoning
▶ Bio-Ontologies
▶ Community Database
▶ Data Mining

- ▶ DNA Sequencing
- ▶ Experiment
- ▶ Ontology Lookup Service for Controlled Vocabularies and Data Annotation

References

Allen JF (2001) Bioinformatics and discovery: induction beckons again. Bioessays 23(1):104–107
Blake JA, Bult CJ (2005) Beyond the data deluge: data integration and bio-ontologies. J Biomed Inform 39(3):314–320
Buetow KH (2005) Cyberinfrastructure: empowering a 'third way' in biomedical research. Science 308:821–824
Evans J, Rzhesky A (2010) Machine science. Science 329(5990):399–400
Hey T, Tansley S, Tolle K (eds) (2009) The fourth paradigm. Data-intensive scientific discovery, Microsoft Research, Redmond. http://research.microsoft.com/en-us/collaboration/fourthparadigm. Accessed 31 August 2010
Kell DB, Oliver SG (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. Bioessays 26(1):99–105
Leonelli S (2009) On the locality of data and claims about phenomena. Philos Sci 76(5):737–749
O'Malley MA, Elliott KC, Haufe C, Burian RM (2009) Philosophies of funding. Cell 138:611–615
Philippi S (2006) Addressing the problems with life-science databases for traditional uses and systems biology. Nature 7:769–773
Renear AH, Palmer CL (2009) Strategic reading, ontologies, and the future of scientific publishing. Science 325(5942):828–32

Data-Intensive Science

- ▶ Data-Intensive Research

Date Hub

- ▶ Hub

DBID

- ▶ Database Identifier

dbSNP

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

Synonyms

Single-nucleotide polymorphism database

Definition

A database containing information about genetic variations. It was established by the National Center for Biotechnology Information as a central repository of data for both single-base nucleotide substitutions and short deletion and insertion polymorphisms (Sherry et al. 2001).

Cross-References

- ▶ SNPedia

References

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

DDBJ Genome Resources

Akos Dobay[1] and Maria Pamela Dobay[2]
[1]Institute of Evolutionary Biology and Environmental Studies (IEU), University of Zurich, Zurich, Switzerland
[2]Department of Physics, Ludwig-Maximilians University, Munich, Germany

Definition

The DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp) is a collection of nucleotide

sequence data. Although DDBJ collects data worldwide, most of the direct submissions are from Japanese researchers. DDBJ continuously exchanges the collected data with the European Bioinformatics Institute (▶ EBI Genome Resources) and the National Center for Biotechnology Information Database (NCBI; http://www.ncbi.nlm.nih.gov). The three databases are part of the International Nucleotide Sequence Database (INSD; http://www.insdc.org) consortium, whose function is to ensure the integrity of the shared information. Apart from the genome resources, DDBJ also has resources for protein sequences and protein structures.

Characteristics

DDBJ was an initiative of Japanese molecular biologists and biophysicists (Tateno and Gojobori 1997) that began its activities in 1986, in collaboration with the European Molecular Biology Laboratory (EMBL; http://www.embl.de) and the genetic sequence database (Genbank; http://www.ncbi.nlm.nih.gov/genbank) of the National Institutes of Health (NIH; http://www.nih.gov). The maintenance and the development of DDBJ are organized by the Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ; http://www.cib.nig.ac.jp/) of the National Institute of Genetics (NIG; http://www.nig.ac.jp/english/index.html). Ninety-nine percent of the data coming from Japanese researchers and available from INSD are submitted through DDBJ (Kaminuma et al. 2010, 2011). Researchers from China, Korea, and Taiwan are also mostly submitting their data through DDBJ. When the submission does not include a large number of sequences, as the case is from whole-genome shotgun (WGS) data or mass sequence data for genome annotation (MGA) (Sugawara et al. 2009), DDBJ offers a user interface called SAKURA. Up to now, DDBJ has released several databases. As the case is in the ▶ NCBI BioProject genome resource, researchers have the option to submit their projects prior to full completion. DDBJ also provides genomic information as well as resources for protein sequences and protein structures. Table 1 summarizes the available databases in DDBJ.

DDBJ Genome Resources, Table 1 List of the databases available at DDBJ

Service name	Description
INSD-core	The INSD-core data contain all the traditional sequences of complete genomes, but exclude whole-genome shotgun (WGA) sequences, mass sequence for genome annotation (MGA), and third party annotation (TPA) (Kaminuma et al. 2010; Kaminuma et al. 2011)
WGS	The whole-genome shotgun contains large sets of overlapping and finished sequences without annotation from ongoing genome projects (Kaminuma et al. 2010; Kaminuma et al. 2011)
MGA	The mass sequence for genome annotation is comprised of sequences that are produced in a large quantity for the purpose of genome annotation (Kaminuma et al. 2010)
TPA	Third party annotation data are assembled using primary entries of publicized nucleotide sequence data collections, with additional features determined by experimental or inferential methods from a pool of experts (Cochrane et al. 2010)
DTA	DDBJ trace archive is a permanent repository of DNA sequence chromatograms
DRA	DDBJ sequence read archive is a repository for sequencing data from next-generation sequencing technologies. A web-based metadata creation tool called MetaDefine has been released in March 2010 to facilitate the submission process
DAD	The DDBJ amino acid database contains amino acid sequences extracted from the nucleotide flat files present in the DDBJ periodical release and TPA dataset
GTPS	The Gene trek in prokaryotic space is a re-annotated database that uses motif scans
GIB	The genome information broker is a comprehensive data repository of complete microbial genomes
GIB-V	The genome information broker for viruses is a repository for complete virus genomes. For other virus genome resources, refer to the ▶ NCBI viral genomes resources
CIBEX	The Center for Information Biology Gene Expression database (http://cibex.nig.ac.jp) is a public database for mircoarray data (▶ Microarray data, parallel and distributed preprocessing)
DOR	The DDBJ omics archive stores quantitative data from both microarrays (▶ Microarray data, parallel and distributed preprocessing) and new high-throughput sequencing platforms. DOR also integrates CIBEX and exports the data to ArrayExpress database (▶ EBI genome resources; Kodama et al. 2010)
GTOP	The genomes TO protein structures and function database consists of data analyses of proteins by application of various computational tools to the amino acid sequences of genome projects sequenced to its entirety (Kawabata et al. 2002; Fukuchi et al. 2009)

Other Resources

The DDBJ also includes patent data transferred from the Japan Patent Office (JPO; http://www.jpo.go.jp), the Korean Intellectual Property Office (KIPO; http://www.kipo.go.kr), as well as from the United States Patent and Trademark Office (USPTO; http://www.uspto.gov) and the European Patent Office (EPO; http://www.epo.org).

Cross-References

- ▶ EBI Genome Resources
- ▶ Genome Annotation
- ▶ Microarray Data, Parallel and Distributed Preprocessing
- ▶ NCBI Bioproject Genome Resources
- ▶ NCBI Viral Genomes Resources

References

Cochrane G, Bates K, Apweiler R, Tateno Y, Mashima J, Kosuge T, Mizrachi IK, Schafer S, Fetchko M (2010) Evidence standards in experimental and inferential INSDC third party annotation data. OMICS J Int Biol 10 (2):105–113

Fukuchi S, Homma K, Sakamoto S, Sugawara H, Tateno Y, Gojobori T, Nishikawa K (2009) The GTOP database in 2009: updated content and novel features to expand and deepen insights into protein structures and functions. Nucleic Acids Res 37(suppl 1):D333–D337

Kaminuma E, Mashima J, Kodama Y, Gojobori T, Ogasawara O, Okubo K, Takagi T, Nakamura Y (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. Nucleic Acids Res 38:D33–D38

Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y (2011) DDBJ progress report. Nucleic Acids Res 39:D22–D27

Kawabata T, Fukuchi S, Homma K, Ota M, Araki J, Ito T, Ichiyoshi N, Nishikawa K (2002) GTOP: a database of protein structures predicted from genome sequences. Nucleic Acids Res 30:294–298

Kodama Y, Kaminuma E, Saruhashi S, Ikeo K, SugawaraH TY, Nakamura Y (2010) Biological databases at DNA data bank of Japan in the era of next-generation sequencing technologies. Adv Exp Med Biol 680:125–135

Sugawara H, Ikeo K, Fukuchi S, Gojobori T, Tateno Y (2009) DDBJ dealing with mass data produced by the second generation sequencer. Nucleic Acids Res 37:D16–D18

Tateno Y, Gojobori T (1997) DNA data bank of Japan in the age of information biology. Nucleic Acids Res 25:14–17

De Novo Computational Discovery of Motifs

Jianhua Ruan
Department of Computer Science, University of Texas at San Antonio, San Antonio, TX, USA

Definition

A class of computational methods for finding transcriptional binding motifs within a set of promoter sequences. This is different from the scenario where one needs to search promoter sequences for the binding sites of a known transcription factor binding motif (e.g., a consensus or a PSWM).

Death Rate

- ▶ Life Span, Turnover, Residence Time
- ▶ Lymphocyte Population Kinetics

Decentralized Version Control

- ▶ Distributed Version Control System (DVCS)

Decision Rule

- ▶ Prediction Rule

Decision Theory

- ▶ Bayesian Decision Analysis

Decision Tree

Daniel Berrar[1] and Werner Dubitzky[2]
[1]Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Midori-ku, Yokohama, Japan
[2]Biomedical Sciences Research Institute, University of Ulster, Coleraine, UK

Synonyms

Classification tree; Regression tree

Definition

A decision tree refers to both a concrete decision model used to support decision making and a method to construct such models automatically from data. As a model, a decision tree refers to a concrete information or knowledge structure to support decision making, such as classification (▶ Model Testing, Machine Learning) and regression (▶ Regression Analysis) tasks, processes or analyses. As a method, a decision tree comprises various techniques to construct decision tree models in a highly automated fashion using specific algorithms and measures from the field of statistics, machine learning (▶ Model Validation, Machine Learning) and artificial intelligence.

Characteristics

Decision-Tree Structure

A decision tree is composed of *nodes* and *branches* that connect the nodes (Quinlan 1993; Hastie et al. 2001; Duda et al. 2001). Two basic node types are distinguished: *leaf nodes* and *non-leaf nodes* (a special non-leaf node is the *root node*). Each non-leaf node is labeled with an attribute or a question. The branches emanating from a non-leaf node correspond to the possible values of the attribute or the answers to the question. The leaf nodes of a decision tree are labeled with a class or category.

Figure 1 illustrates decision tree structures based on two simple examples from everyday life. To emphasize the tree analogy, the decision trees depicted in Fig. 1 have their root node at the bottom of the diagram and the leaf nodes at the top. In practice, decision trees are usually visualized with the root node at the top and the leaf nodes at the bottom.

Consider Fig. 1. In the Name Title example, there are three leaf nodes labeled *Mr*, *Mrs*, and *Miss*, and two non-leaf nodes labeled with the attributes *Gender* (root node) and *Marital Status*. The branches are labeled with the corresponding attribute values: *male* and *female* (for attribute *Gender*) and *married* and *not married* (for attribute *Marital Status*). In the Jogging example, there are seven leaf nodes (labeled *Yes* and *No*, respectively), three non-leaf nodes (labeled with the attributes *Outlook*, *Humidity*, and *Temperature*), and ten branches labeled with the corresponding attribute values. Figure 1 also illustrates the corresponding decision rules (▶ Prediction Rule).

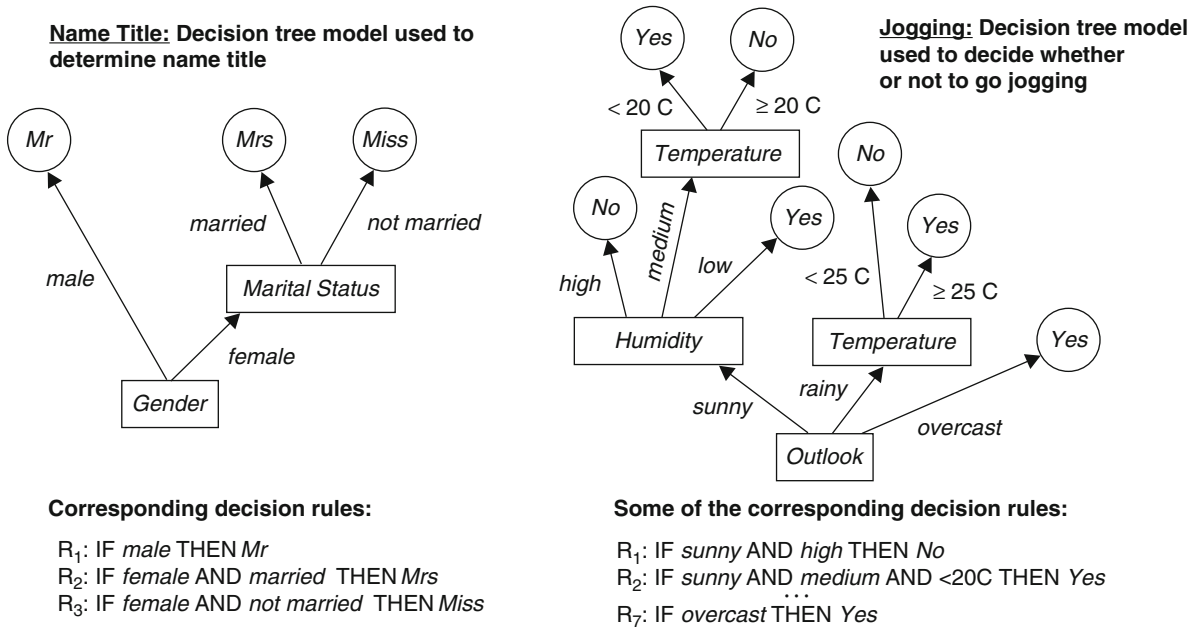
A decision tree whose leaf nodes are labeled with discrete class labels is referred to as classification tree (▶ Model Testing, Machine Learning). A decision tree that uses continuous values or value ranges is referred to as regression tree (▶ Regression Analysis) (Breiman et al. 1984). A decision-tree structure represents a ▶ directed acyclic graph which satisfies the following properties:

1. There is exactly one *node*, called the *root*, into which no *edges* (branches) enter.
2. Each *node* other than the *root* has exactly one entering *edge* (branch).
3. There is a unique path from the *root* to each non-root *node*.

Each path from a decision tree's root node to a leaf node can be interpreted as a decision rule (▶ Decision Rule, ▶ Machine Learning) which has a *condition* and *conclusion* part. This may be expressed using the IF-THEN notation or the symbol for logical implication as follows:

$$\text{IF } \textit{condition} \text{ THEN } \textit{conclusion}$$
$$\textit{condition} \Rightarrow \textit{conclusion}$$

If the input information meets *all* the conditions described in the condition part, then the conclusion



Decision Tree, Fig. 1 Simple decision trees facilitating classification tasks in everyday life: determining the English (name) title or honorific (*left diagram*), and deciding whether or not to go jogging (*right diagram*). *Circles* depict leaf nodes, *boxes* depict non-leaf nodes, and *arrows* represent branches. The rule sets below the decision trees describe the corresponding decision rules

stated in the conclusion part is asserted. Thus, the decision structure of a decision tree can be formulated as a set of decision rules (this is illustrated in Fig. 1). The decision rules derived from a decision tree may be associated with quantities expressing confidence and support as illustrated in Fig. 2.

Decision-Tree Construction

Once a decision tree is constructed, it can be used to aid decisions of a decision maker, humans or machines. There are two basic approaches for creating decision trees. Firstly, decision trees may be generated manually by knowledge engineers working with human experts or textbooks. This approach is effective in well-understood domains but involves a considerable human effort and time. Secondly, decision trees may be derived automatically from available examples (data) by suitable machine learning (▶ Induction) algorithms. The two basic steps involved in the machine learning approach are:

1. Obtain facts (data) about the decision making problem or studied phenomenon.
2. Derive the decision tree (or equivalent set of decision rules) by means of an inductive process that generalizes from the available facts (data).

In not so well-understood domains (like biological knowledge discovery), the machine learning approach has the added advantage that by automatically generating a decision tree from available data it may be possible to reveal relationships in the data that may not be obvious to the investigator or decision maker. A decision-tree learning algorithm determines, for instance, which attribute should be placed at the root node given the input data, thus providing an insight as to what attribute has the strongest influence in the partitioning of the underlying instances. The knowledge-discovery aspect as well as the symbolic encoding of the knowledge they represent make decision trees a useful method for exploring biological data. Decision trees have been widely used for exploratory data analysis, classification, and regression tasks in biology (Zhang et al. 2001; Kingsford and Salzberg 2008).

Training Set (Learning Instances)

Instance #	Attributes X	Attributes Y	Class Label
#1	1	0	A
#2	1	0	A
#3	1	1	A
#4	1	1	A
#5	1	0	A
#6	0	1	A
#7	0	0	B
#8	1	0	B
#9	1	0	B
#10	0	1	B

entropy (*training set*) = 0.97

Partition of Learning Set Based on Attribute *X*

Element of partition whose instances have $X = 1$

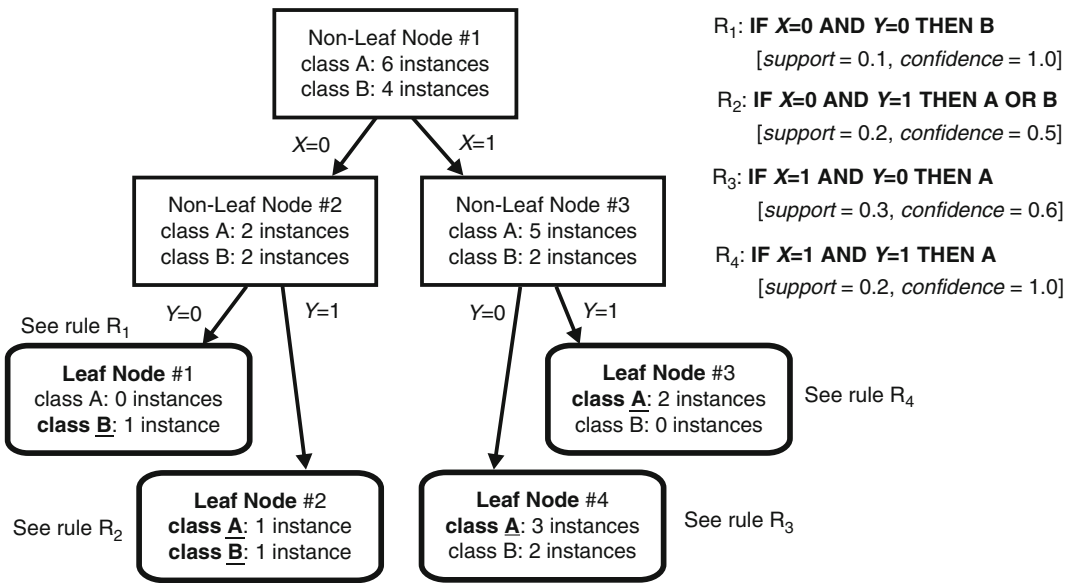
Instance #	Attributes X	Attributes Y	Class Label
#1	1	0	A
#2	1	0	A
#3	1	1	A
#4	1	1	A
#5	1	0	A
#8	1	0	B
#9	1	0	B

entropy (X=1) = 0.86

Element of partition whose instances have $X = 0$

Instance #	Attributes X	Attributes Y	Class Label
#6	0	1	A
#7	0	0	B
#10	0	1	B

entropy (X=0) = 0 .92



Decision Tree, Fig. 2 Illustration of decision-tree learning based on a training set with ten instances and two attributes. *Top*: Training set and partition determined for root node. *Bottom*: resulting final decision tree (*left*) and decision rules (*right*)

Decision-tree learning generalizes from observations by a process of induction. This process takes as input a set of specific observations or instances and generates general rules that cover them. To facilitate automated decision-tree learning, the learning observations or instances (referred to as training set [▶ Model Training, Machine Learning]) are usually expressed as a table where a row represents an instance and a column represents either an attribute and its values or the class labels.

Using the training instances as input, a decision-tree learning algorithm generates a decision-tree model by recursively partitioning the instances via the following main steps:

1. *Initialize*. Provide the full set of observations (training set) as input to the decision-tree learning algorithm.

2. *Analyze attributes*: Determine the attribute that produces the most uniform grouping or partitioning of instances based on the class label of the instances.
3. *Create node*: If the predefined uniformity threshold is reached, create a leaf node and label it with the corresponding class label. Otherwise, create a non-leaf node that tests the attribute, assign to it the instances of the corresponding partition, then go to Step 2.

The eventual result of the decision-tree learning procedure outlined above is a decision-tree model in which all or the majority of instances at a leaf node show the same class label.

A challenge in decision-tree learning is to maximize uniformity or purity of the instances assigned to the leaf nodes of a decision-tree model while ensuring that the model generalizes well to unseen instances, i.e., instances that do not feature in the training set. The latter is also referred to as generalization ability, or bias-variance trade-off.

Measures of uniformity, purity, or homogeneity commonly employed by classification tree learning algorithms include the information theory measures of entropy (▶ Entropy), the Gini index, and the Kolmogorov–Smirnov distance. Regression-tree learning algorithms make use of variance minimization techniques for the same purpose.

The following example illustrates the concept of decision tree learning for a binary classification task (▶ Model Testing, Machine Learning), i.e., a task involving exactly two class labels. The training set (▶ Model Training, Machine Learning) in this example comprises ten instances, each described by two numeric attribute values and one class label. The attributes are called X and Y and assume values from the set {0,1}, and the class labels are drawn from the set {A,B}. Six instances in the training set carry the class label A, and four the label B. Figure 2 depicts the training set as well as a partition (composed of two groups of instances) obtained from applying the information-theoretic entropy uniformity measure. According to this measure, higher uniformity corresponds to more information, which is corresponds to lower entropy.

Because the ▶ information gain for the partition derived from the values of attribute X is higher than that for attribute Y, attribute X is chosen to split the data at the root node. This is illustrated by the decision tree depicted in Fig. 2. The same learning process is applied to the remaining attributes (in this case only attribute Y) at the next level. This leads to the decision-tree model and decision rule (▶ Prediction Rule) depicted in Fig. 2. The model consists of three non-leaf nodes (including the root node) and four leaf nodes. Whereas three leaf nodes are unambiguously associated with the class label A and B, respectively, Leaf Node #2 cannot be assigned to a unique class label.

The class label frequencies represented by the leaf nodes are used to express likelihood estimates for the classification of unseen instances. Unseen instances are those that comply with the structure of the learning instances but have not been used in the decision learning process. To illustrate this idea, consider the decision tree model in Fig. 2. An unseen instance that is characterized by the attribute values $X = 1$ and $Y = 0$ would be labeled (classified) with the class label A. Because three out of five instances in the corresponding Leaf Node #4 carry the class label A, the conditional probability for this classification would be given as 3/5. The conditional probability associated with the predicted outcome of decision tree and similar models is also referred to as confidence. Another measure that is frequently used in the context of decision trees is called support. The support of a decision rule (corresponding to a path from root to a leaf node in a decision tree model) is determined as the proportion of instances in the learning set that satisfy the rule. For example, the decision rule R3 corresponding to Leaf Node #4 in the decision tree model in Fig. 2 has a support of 3/10 because 3 of 10 items (#1, #2, and #5) satisfy the rule. The aim in decision tree learning is to construct a decision tree model with a high confidence and support.

Strengths and Weaknesses of Decision Trees

Strengths

- Decision-tree models capture knowledge in an easy-to-interpret knowledge structure, either as hierarchical trees or sets of decision rules.
- Decision-tree learning offers a powerful approach to automatically discover decision-tree models from large data sets. Decision-tree learning can be used to reveal important relationships in data.
- The computational costs associated with decision-tree learning are low to moderate.
- Decision trees can process both discrete and continuous variables and have an intrinsic ability to handle missing values.

Weaknesses

- Decision-tree learning is highly sensitive to changes in the training set. Small changes in the learning set may lead to considerably different decision-tree models. This is also referred to as the stability problem in machine learning.
- For a given data set or decision problem, the orthogonal decision regions generated by axis-parallel decision-tree learning methods may not sufficiently approximate the true underlying decision regions.

Decision-Tree Algorithms and Tools

There is a large variety of decision-tree algorithms and tools; open source tools widely used in computational biology include Weka (Weka, Machine Learning Tool) and R (▶ R, Programming Language).

Cross-References

- ▶ Directed Acyclic Graph
- ▶ Entropy
- ▶ Induction
- ▶ Information Gain
- ▶ Model Training, Machine Learning
- ▶ Model Validation, Machine Learning
- ▶ Overfitting
- ▶ Prediction Rule
- ▶ R, Programming Language
- ▶ Regression Analysis

References

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman and Hall/CRC, Boca Raton

Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley-Interscience, New York

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning, Springer Series in Statistics. Springer, New York

Kingsford C, Salzberg S (2008) What are decision trees? Nat Biotech 26(9):1011–1013

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Mateo

Zhang H, Yu CY, Singer B, Xiong M (2001) Recursive partitioning for tumor classification with gene expression microarray data. Proc Natl Acad Sci USA 98(12):6730–6735

Declarative Language

Amanda Clare[1] and Martin Swain[2]
[1]Department of Computer Science, Aberystwyth University, Aberystwyth, UK
[2]Institute of Biological, Environmental, and Rural Sciences, Aberystwyth University, Aberystwyth, Ceredigion, UK

Synonyms

Declarative programming

Definition

A declarative programming language is a programming language where the programmer specifies the goal or *what* should be achieved, rather than *how* a goal should be achieved. Logic programming languages, such as Prolog, are declarative languages, as are many database query languages. They are often contrasted with imperative languages such as C/C++, Java, FORTRAN, Python, and Perl. Imperative languages are concerned with procedures: the programmer specifies a series of instructions or statements that must be executed one after another in a specific order to achieve the output or goal of the program. Declarative languages are concerned with the relations between statements. In declarative languages based on logic such as Prolog, computations are based on the evaluation of logical statements using mechanisms such as ▶ deduction, ▶ induction, or ▶ abduction: there is an important separation between the logical statements and the mechanisms of reasoning with those statements.

Cross-References

- ▶ Abduction
- ▶ Deduction
- ▶ Induction
- ▶ PROLOG

Declarative Programming

▶ Declarative Language

Decrease

▶ Reduction

Deduction

C. Maria Keet
KRDB Research Centre, Free University of Bozen-Bolzano, Bolzano, Italy

Definition

Deduction is a way to ascertain if some theory T, which consists of one or more axioms expressed in a suitable logic language, entails a conclusion, which is an axiom α that is not explicitly asserted in T; this is written as $T \models \alpha$. That is, α can be *derived* from the premises using a set of *deduction rules*.

Usage

Deduction is used widely in ▶ knowledge representation and the semantic web, including checking consistency of ▶ bio-Ontologies, using automated reasoners (▶ Automated Reasoning).

Characteristics

There are various ways how to ascertain $T \models \alpha$, be it manually or automatically. One can construct a step-by-step proof (▶ Proof, Logic) "forward" from the premises by applying the deduction rules or prove it indirectly such that $T \cup \{\neg\alpha\}$ must lead to a contradiction. The former approach is called *natural deduction*, whereas the latter is based on techniques such as resolution, matrix connection methods, and sequent deduction (which includes *tableaux*).

Concerning deduction rules for tableaux and first order predicate logic formulae, we have, as with the example in proof, the two deduction rules that if a model satisfies a conjunction, then it also satisfies each of the conjuncts,

$$\frac{\phi \wedge \varphi}{\begin{array}{c}\phi\\ \varphi\end{array}}$$

and if a model satisfies a disjunction, then it also satisfies one of the disjuncts,

$$\frac{\phi \quad \vee \quad \varphi}{\phi \quad | \quad \varphi}$$

In addition, there are two rules for the quantified formulas. First, if a model satisfies a universally quantified formula ($\forall$), then it also satisfies the formula where the quantified variable has been substituted with some term (and the prescription is to use all the terms which appear in the tableaux),

$$\frac{\forall x.\phi}{\begin{array}{c}\phi\{X/t\}\\ \forall x.\phi\end{array}}$$

and, second, for an existentially quantified formula, if a model satisfies it, then it also satisfies the formula where the quantified variable has been substituted with a new Skolem constant,

$$\frac{\exists x.\phi}{\phi\{X/a\}}$$

Example

Let us take some arbitrary theory T that contains two axioms stating that relation R is reflexive ($\forall x.R(x,x)$, a thing relates to itself) and asymmetric ($\forall x,y.\ R(x,y) \rightarrow \neg R(y,x)$; if a thing a relates to b by relation R, then b does not relate back to a). We then can deduce, among others, that $T \cup \{\neg\forall x,y.R(x,y)\}$ is satisfiable. We do this by demonstrating that the *negation* of the axiom is *unsatisfiable*.

To enter the tableau, we first rewrite the asymmetry into a disjunction using equivalences, that is, $\forall x,y.\ R(x,y) \rightarrow \neg R(y,x)$ is equivalent to $\forall x,y.\ \neg R(x,y) \vee \neg R(y,x)$, and add a negation to $\{\neg\forall x,y.R(x,y)\}$, which

Deduction, Table 1 Tableau example

Number	Tableau	Explanation
1	$\forall x.R(x,x)$	Reflexivity axiom in the original theory T
2	$\forall x,y.\ \neg R(x,y) \vee \neg R(y,x)$	Asymmetry axiom in the original theory T
3	$\forall x,y.R(x,y)$	The negated axiom added to theory T
4	|	Substitute x for term a in 1,2,3
5	$R(a,a)$	
6	$\forall y.\ \neg R(a,y) \vee \neg R(y,a)$	
7	$\forall y.R(a,y)$	
8	|	Substitute y for term a in 2 and 3
9	$R(a,a)$	
10	$\neg R(a,a) \vee \neg R(a,a)$	
11	$R(a,a)$	
12	/ \	Split the disjunction of 10
13	$\neg R(a,a)$ $\neg R(a,a)$	Which each generate a clash with 9 and 11, hence, $\neg\forall x,y.R(x,y)$ is entailed by T

thus becomes $\forall x,y.R(x,y)$. Then, to start the tableau, we have three axioms (Table 1).

Cross-References

▶ Automated Reasoning

References

Hedman S (2004) A first course in logic – an introduction to model theory, proof theory, computability, and complexity. Oxford University Press, Oxford, UK

Portoraro F (2010) Automated reasoning. In: Zalta E (ed) Stanford encyclopedia of philosophy. Stable URL, http://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=reasoning-automated

Deductive Reasoning

Angelika Kimmig
Departement Computerwetenschappen, Katholieke Universiteit Leuven, Heverlee, Belgium

Definition

Deduction is the process of inferring whether a given statement is entailed, that is, whether it logically follows from a theory.

Characteristics

Deductive reasoning allows one to infer statements that are logical consequences of a set of given statements (Genesereth and Nilsson 1987). For instance, given a theory stating that all swans are white ($\forall x.swan(x) \rightarrow white(x)$) and that Odette is a swan ($swan(Odette)$), it can be deduced that Odette is white ($white(Odette)$). The example uses one of the general inference rules of first order logic, which can be written as $\frac{\forall x.p(x) \rightarrow q(x) \text{ and } p(a)}{q(a)}$. The top part denotes the statements present in the given theory (read as "whenever a property p holds for some object x, another property q also holds for x" and "p holds for the given object a"), and the bottom part the deduced statement ("q holds for a").

In contrast to ▶ induction, deduction does not hypothesize new knowledge, but makes information implicitly contained in a logical theory explicit. It thus is *truth-preserving*: whenever the premises of deductive inference hold, the conclusions must hold as well. Deduction either generates additional statements that follow logically from the theory, or verifies a given statement by constructing a proof (▶ Proof, Logic). Deductive reasoning is the basis for theorem proving and logic programming (Flach 1994).

Cross-References

- ▶ Induction
- ▶ PROLOG
- ▶ Proof, Logic

References

Flach P (1994) Simply logical – intelligent reasoning by example. Wiley, The Netherlands
Genesereth M, Nilsson N (1987) Logical foundations of artificial intelligence. Morgan Kaufmann, San Francisco

Deductive-nomological (DN) Analysis

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

Definition

The deductive-nomological (DN) analysis, explicitly formulated by Hempel and Oppenheim (1948/1965), has had enormous influence as an account of both causation and scientific explanation, in particular in the tradition of logical empiricism. Instead of considering that explanation by laws *replaces* causal explanation, the DN analysis suggests that causation can be analyzed in terms of lawful, or "nomological," explanation. Carnap has given a classical statement of this view: "What is meant when it is said that event *B* is caused by event *A*? It is that there are certain laws in nature from which event *B* can be logically deduced when they are combined with the full description of event *A*." (Carnap 1966, p. 194).

Cross-References

- ▶ Causality

References

Carnap R (1966) Philosophical foundations of physics. Basic Books, New York
Hempel CG, Paul O (1948/1965) Studies in the logic of explanation. In: Hempel CG (ed) Aspects of scientific explanation. Free Press, New York, pp 245–295

Degree Centrality

Deepak Sharma[1] and Avadhesha Surolia[2]
[1]Translational Health Science and Technology Institute, Gurgaon, India
[2]Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India

Definition

Degree centrality is defined as the number of links incident upon a node (i.e., the number of ties that a node has). If the network is directed (meaning that ties have direction), then two separate measures of degree centrality are defined, namely, indegree and outdegree. Indegree is a count of the number of ties directed to the node (head endpoints) and outdegree is the number of ties that the node directs to others (tail endpoints). In such cases, the degree is the sum of indegree and outdegree.

Cross-References

- ▶ Pathway Targeting, Antimycobacterial Drug Design

References

Bang-Jensen J, Gutin G (2007) Digraphs: theory, algorithms and applications, 1st edn. Springer, London
Diestel R (2010) Graph theory, 4th edn. Springer, Heidelberg
Freeman LC (1978/1979) Centrality in social networks: conceptual clarification. Soc Netw 1:215–239
Sabidussi G (1966) The centrality index of a graph. Psychometrika 31:581–603

Delocalized

▶ Function, Distributed

Deltaretroviridae

Christian Schönbach
Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka, Japan

Synonyms

Deltaretrovirus

Definition

Retroviridae is a family of retroviruses whose replication involves one reverse transcription step. Older publications often refer to three retrovirus subfamilies Oncovirinae, Lentivirinae, and Spumavirina. Nowadays, the classification is obsolete and retrovirus is grouped into two subfamilies Orthoretrovirinae and Spumaretrovirinae. The genera Betaretrovirus, Gammaretrovirus, Alpharetrovirus, Deltaretrovirus, and Lentivirus belong to the subfamily of Orthoretrovirinae (Index of Viruses – Retroviridae 2006). Deltaretroviridae are complex retroviruses, whose genomes contains besides LTR-gag-pol-env-LTR a number of accessory genes. Deltaretrovirus infects vertebrates. Bovine leukemia virus (BLV), Human T-lymphotropic virus 1, 2, 3, and 4 (HTLV-1, -2, -3, and −4) and Simian T-lymphotropic virus 1, 2, and 3 (STLV-1, -3, -4) were isolated from cow, human, and various nonhuman primates. HTLV-1 and its subtypes are thought to originate from various independent interspecies transmissions from simians to humans starting around 50,000 years ago. The most recent HTLV-1 subtype f probably emerged some 3,000 years ago (Van Dooren et al. 2001).

HTLV-1 and HTLV-2 are human-pathogenic viruses that are transmitted by sexual contact, breastfeeding, and needle sharing during intravenous drug abuse. Worldwide an estimated 10–20 million people are infected with HLTV-1, the etiological agent of adult T-cell leukemia/lymphoma (ATL), and HTLV-1 associated myelopathy/tropical spastic paraparesis (HAM/TSP). Endemic areas include Southwest Japan, Caribbean islands, Central Africa, South America, and Melanesia. HTLV-2 infections are endemic in Central Africa and native populations of North, Central, and South America and associated with HAM/TSP-like illness. HTLV-3 and HTLV-4 are largely uncharacterized in terms of pathogenicity and epidemiology (Wolfe et al. 2005).

Cross-References

▶ HTLV, Cellular Transcription

References

Index of Viruses – Retroviridae (2006) In: Büchen-Osmond C (ed) ICTVdB – The Universal Virus Database. Columbia University, New York (Version 4)

Van Dooren S, Salemi M, Vandamme AM (2001) Dating the origin of the African human T-cell lymphotropic virus type-i (HTLV-I) subtypes. Mol Biol Evol 18(4):661–671

Wolfe ND, Heneine W, Carr JK, Garcia AD, Shanmugam V, Tamoufe U, Torimiro JN, Prosser AT, Lebreton M, Mpoudi-Ngole E, McCutchan FE, Birx DL, Folks TM, Burke DS, Switzer WM (2005) Emergence of unique primate T-lymphotropic viruses among central African bushmeat hunters. Proc Natl Acad Sci USA 102(22):7994–7999

Deltaretrovirus

▶ Deltaretroviridae

Dense Overlapping Regulon Motif

▶ Dense Overlapping Regulons

Dense Overlapping Regulons

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

Synonyms

Dense overlapping regulon motif; DOR

Definition

Dense overlapping regulons (DORs) were found in the *Escherichia coli* transcriptional regulation network. A set of operons, $Z_1, \ldots, Z_m$, are each regulated by a combination of a set of input transcription factors, $X_1, \ldots, X_m$.

References

Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet. 2002;31:64–8

Density of a Subgraph

► Clustering Coefficient

Deoxyribonucleic Acid Sequencing

► DNA Sequencing

Depletion Rate

► Life Span, Turnover, Residence Time
► Lymphocyte Population Kinetics

Design Explanation

► Explanation, Functional

Design of Experiments

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

Synonyms

DOE; Optimal experiment design

Definition

The aim is to design experiments in order to maximize the information content of the measurements in the context of their utilization for estimating the model parameters. This is equivalent to minimizing the variances of the parameters to be estimated. The variances are a measure for the uncertainty of the parameters, also represented by individual confidence interval approximations. Design of experiments (DoE) aims at minimizing the variances of the parameters to be estimated. Experiment design for parameter precision aims at determining optimal experimental settings and measurement times in order to maximize the information content from the measured data generated by these experiments. This is equivalent to minimizing the confidence ellipsoid of the parameters to be estimated.

Characteristics

DoE aims to address the following questions:

- What should be the initial conditions for the experiment?
- How long should we run the experiment?

- How should we vary the controls (e.g., the time profiles of feed flowrates)?
- When should we take the measurement samples?

The overall aim is to generate the maximum amount of information for a subsequent estimation of the model parameters, while trying to maintain the process within the required operating envelope. In mathematical terms, we want to minimize some measure ψ of the variance-covariance matrix, v_θ, of the parameters (θ) to be estimated:

$$\min_\xi \Psi(V_\vartheta). \tag{1}$$

The variance-covariance matrix is given by:

$$V_\vartheta = \left(H^*_\vartheta\right)^{-1}, \tag{2}$$

where H^*_ϑ is the information matrix which is a n_θ x n_θ matrix (n_θ is the number of parameters (θ) to be estimated) and is given by:

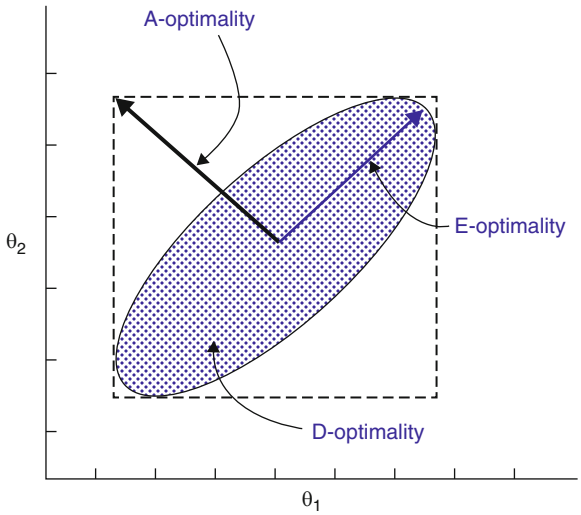
$$H^*_\vartheta = \sum_{l=1}^{N\exp} \sum_{i\in SV} \sum_{m=1}^{Nm,il} \left(\frac{\left(\frac{\partial}{\partial\theta_\mu} z_{il}(\rho_{iml})\right)\left(\frac{\partial}{\partial\theta_\nu} z_{il}(\rho_{iml})\right)}{\sigma^2_{il} z_{il}(\rho_{iml}, \beta_{il})} \right).$$
$$\mu, \nu = 1, ..., n_\theta \tag{3}$$

where ξ is the set of experiment decision variables in all experiments, N_{EXP} is the number of experiments, SV_i is the set of measured state variables in experiment l, N_{il} the number of sampling points for measured variable i in experiment l, ρ_{iml} the m-th measurement time for variable i in experiment l, $z_{il}(\rho_{iml})$ the model-predicted value of variable i at time point ρ_{iml} in experiment l and $\sigma^2_{il} z_{il}(\rho_{iml}, \beta_{il})$ The variance of the measurement error of variable i at time point ρ_{iml} in experiment l.

In order to compare the magnitude of different variance-covariance matrices, various real-valued functions have been suggested as a measure of optimality. The three most commonly used criteria are:

A-optimality: minimize the trace of the variance-covariance matrix:

$$\Psi_A(V_\vartheta) = \frac{1}{N_\theta} \sum_{\mu=1}^{N\theta} (V_\vartheta)_{\mu,\mu}. \tag{4}$$



Design of Experiments, Fig. 1 The different design criteria for a two-dimensional confidence ellipsoid (Adapted from PSE, Ltd)

This minimizes the sum of the variances of the individual parameter estimates. It corresponds to minimizing the dimensions of the smallest hyper rectangle within which the confidence ellipsoid can be inscribed.

D-optimality: minimize the determinant of the variance-covariance matrix:

$$\Psi_D(V_\vartheta) = \det\left(V_\vartheta\right)^{\frac{1}{N\theta}}. \tag{5}$$

This is also known as the minimum volume criterion since it minimizes the volume of the confidence ellipsoid.

E-optimality: minimize the largest Eigen value of the variance-covariance matrix:

$$\Psi_E(V_\vartheta) = \lambda_{MAX}(V_\vartheta). \tag{6}$$

The Eigen values of the variance-covariance matrix correspond to the lengths of the minor and major axes of the confidence ellipsoid. By minimizing the largest Eigen value, the design renders the confidence ellipsoid as spherical as possible.

Figure 1 shows a graphical interpretation of the different design criteria for a two-dimensional confidence ellipsoid.

Cross-References

- ▶ Designing Experiments for Sound Statistical Inference
- ▶ Global Sensitivity Analysis
- ▶ Model-based Experimental Design, Global Sensitivity Analysis
- ▶ Optimal Experiment Design, Ill-Posed Problems

References

Process Systems Enterprise (1997–2011) gPROMS. www.psenterprise.com/gproms

Designing Experiments for Sound Statistical Inference

Melissa Key[1] and Olga Vitek[2]
[1]Center for Computational Diagnostics, Indianapolis, IN, USA
[2]Department of Statistics, Department of Computer Science, Purdue University, West Lafayette, IN, USA

Synonyms

Design of experiments; Experimental planning; Statistical experimental design

Definition

An ▶ *experimental design* is a protocol that defines all aspects of a planned experiment. This includes a definition of the populations of biological organisms and the conditions of interest, procedures for selecting the individuals from the populations or allocating them to treatments, and the organization of experimental material from the selected individuals in the experimental framework. *Statistical experimental design* (Kreutz and Timmer 2009; Montgomery 2001; Oehlert 2000) is conducted in conjunction with *statistical inference*, i.e., in situations where measurements on the selected individuals are used to make biologically relevant conclusions regarding the unknowns. Statistical experimental design avoids ▶ *bias*, i.e., it avoids systematic errors in the conclusions, and ensures that the experiment is *efficient*, (▶ Efficiency) i.e., it minimizes the uncertainty in the conclusions for a given amount of cost.

Characteristics

As an example, consider an experiment which compares the expression of genes in subjects with type II diabetes to healthy controls using a whole transcriptome shotgun sequencing (RNA-seq) (Wang et al. 2009) technology. Statistical experimental design is characterized by the following steps.

Step I: Define the Problem

Define Who or What Is Being Studied

The first step in experiment planning is to clearly define the *populations* and the *conditions* to be represented by the subjects in the study. In systems biology, it may be practical to initially focus the study on smaller populations. In the example of type II diabetes, if literature suggests that the effect of the disease is different in men and women, the experiment may limit the scope of the study to a single gender initially, then plan a follow-up experiment to verify that the conclusions apply to the other gender. Subject availability may impose additional constraints. For example, if the diabetes study can only access subject samples from a single health-care institution, the population is limited to that institution only, and a larger follow-up experiment is necessary to broaden the conclusions.

Define What Is Being Measured

It is also important to define the ▶ *response*, i.e., a measurable aspect of the biological samples for which the variation between conditions or treatments is of interest. In systems biology, many responses are quantitative measurements at the molecular level, such as gene expression. It is common to simultaneously measure a large number of responses. For example, the diabetes experiment simultaneously quantifies the expression of tens of thousands of genes on each biological sample. The definition of the response can be nontrivial. In the example RNA-seq experiment, gene expression can be quantified separately for each isoform, or as the sum of the expression of all of its isoforms, and the latter can be calculated over the unions or the intersections of the exons (Garber et al. 2011). The definitions can have

important implications for interpreting the data and for the biological conclusions.

Translate the Research Question into a Statistical Hypothesis

Many systems biology experiments aim at finding relative changes in the response between conditions. In the diabetes example, of interest are (log)-fold changes in the transcription rate between the populations of healthy subjects and the subjects with the disease, separately for each gene. In statistical terminology, this translates into testing the *null hypothesis* of "no change" against the *alternative hypothesis*, e.g., log-fold change different from zero. Not all nonzero values of (log)-fold change are of practical importance, and the range of biologically relevant values needs to be specified in advance.

Step II: Sample Selection and Resource Allocation

Define the Experimental Unit

► *Experimental units* are the basic currency of sample selection and resource allocation. An experimental unit is the subject, sample, or object that carries the condition or the treatment. For example, in the diabetes experiment, the experimental unit is a person (i.e., a patient). If multiple treatments are assigned to the same subject in different areas, (e.g., a different topical treatment is applied to different patches of the skin), then the experimental unit is an area (e.g., skin patch). If a treatment is assigned to pools of biological material from multiple subjects, then the experimental unit is a pool.

Replication

Quantities such as transcript abundance vary naturally in the populations and introduce *biological variation* in the response. In addition, sample processing and handling can interfere with its composition, and measurement technologies can be somewhat imprecise, introducing *technical variation*. As the result, the observed difference in the response can be either the systematic effect of the condition or treatment, or a random artifact of these sources of variation. The goals of replication are to (1) evaluate the extent of biological and technical variation, (2) assess whether the observed change in the response is likely to arise from this variation by random chance, and (3) increase the precision of the conclusions, since increasing the number of replicates increases our confidence that the difference is indeed systematic.

The two sources of variation lead to two possible replication types. Biological replicates are multiple experimental units with the same condition or treatment. For example, in the diabetes experiment, biological replicates are multiple subjects from a disease group. Technical replicates are multiple measurements on the same experimental units. They address the precision of the measurement protocol but not the biological variation. Even sensitive technologies such as RNA-seq do not eliminate the presence of biological variation. Therefore, experiments that focus on the populations always require biological replicates as part of the design.

Randomization

Randomization guards the experiment against ► *bias*. Bias occurs when the experimental units with different conditions are selected or handled in systematically different ways, not intended by the purpose of the study. The two sources of variation (biological and technical variation) lead to two sources of bias. The first occurs when subjects selected from the groups differ in known or unknown biological characteristics (such as age, gender, or ethnicity) that affect the response. The second occurs due to systematic differences in the technical aspects of the experiment between conditions, such as in protocols of specimen collection or time of data acquisition.

When these sources are unaccounted for, they become *confounding factors*, i.e., they affect the response in addition to the condition or treatment, and bias the results. Confounding cannot be removed by increasing the sample size or by demonstrating the reproducibility of the results in a repeated instance of the same workflow. Instead, confounding can be removed by randomization. A randomization of the experimental units (e.g., a random selection of samples from the underlying population) and the randomization of the order of sample processing and data acquisition distribute the confounding factors roughly equally across groups and eliminate the bias.

Blocking

A completely randomized design has two drawbacks. First, although randomization averages the allocation of confounding factors between conditions, it can yield unequal allocations in experiments with a small number of replicates. Second, in randomized experiments, the variability of the response within each group is the

combination of the biological and technical variation, and it may be difficult to detect the systematic differences between conditions.

Blocking improves upon these two aspects when confounding factors are known in advance, and is performed by imposing restrictions on randomization. The two sources of variation (biological and technical) lead to two types of blocking. In the context of sample selection, blocking is sometimes referred to as *matching*. In the diabetes example, a block is a combination of known confounding factors (e.g., age, gender, ethnicity). The design enforces an equal number of subjects with diabetes and controls in a block, and randomly selects subjects with these characteristics from the population. When multiple measurements are possible on a same subject, e.g., two treatments can be applied to two skin patches, the subject forms a block and receives both treatments, and the pairs of the patches are matched.

For data acquisition, blocking is sometimes referred to as *multiplexing*. If a particular experimental step is noisy but can accommodate multiple samples, it can be viewed as a block. Blocking enforces the constraint that the step processed an equal number of samples from each condition. An example of block is a two-color cDNA mocroarray, and a variety of strategies for allocating samples to arrays have been proposed (Dobbin and Simon 2002). In an RNA-seq experiment, blocking can utilize the capacity to label the samples with sample-specific sequences ("bar codes") that allow multiple samples to be sequenced in a same lane of a flow-cell, making within-lane differences more consistent (Auer and Doerge 2010). Blocking strategies for mass spectrometry-based proteomic experiments have also been discussed (Oberg and Vitek 2009). A randomization of samples within blocks is always required to account for the unknown confounding effects.

Blocking enforces a strict balance of the confounding factors between conditions and prevents bias. In addition, if the downstream statistical analysis distinguishes the variation of the response between the blocks from the within-block variation between the conditions, it increases the sensitivity of tests.

Step III: Statistical Modeling

Describe the Anticipated Properties of Data in a Statistical Model

Unfortunately experimentalists often consider statistical modeling as a separate task, which can be deferred to a statistician once the data are collected. In practice, experimental design and statistical modeling are tightly interconnected. The understanding of the downstream statistical analysis helps us determine the optimal design and maximize our ability to detect quantitative changes in the response for a given amount of replication and cost.

Statistical inference requires a probability model that describes three aspects of the experiment. First, the model describes the sources of systematic variation in the response (such as conditions, subjects, and blocks) in the experimental design. Second, the model describes the scope of interpretation that we associate with the biological replicates, i.e., whether we restrict our conclusions to the subjects in the study or generalize the conclusions to the entire underlying populations. In systems biology, many experiments aim at generating initial hypotheses for a subsequent follow-up, and the number of biological replicates is too small to adequately represent the underlying populations. In these cases, it is useful to restrict the scope of the conclusions to the selected samples only and treat them as fixed units. On the other hand, experiments at the validation stage aim at generalizing the conclusions to the underlying populations, and in this case, the biological replicates are best viewed as random selections from the populations and are represented in a mixed or multilevel model.

Finally, the probability model describes the nature of the nonsystematic variation in the response, based on the information from a pilot study or from the literature. For example, in gene expression microarrays, the variation of log-response can be assumed normally distributed and leads to models such as analysis of variance (ANOVA) (► Analysis of Variance (ANOVA) Tables). In the RNA-seq example, the response is the count of reads and can be described using a Poisson or a negative binomial distribution. Frequentist modeling treats the unknown parameters of these distributions as fixed, while Bayesian models specify the prior distributions of all the parameters, and Empirical Bayes models estimate the parameters of the prior distributions from the data.

Describe the Model-Based Testing

A consequence of each probability model is the procedure for testing the scientific hypothesis of interest.

For example, in analysis of variance, testing the null hypothesis of the same expected gene expression in two conditions leads to the student's *t*-test. Count response leads to the Fisher's exact test or tests based on Poisson regression. In experiments with multiple responses, such as the expression of genes, a separate hypothesis is tested for each gene and the testing requires controlling a multivariate error rate, such as the false discovery rate.

Derive Model-Based Requirements of Sample Size

The combination of the protocol of replication, randomization, and blocking and of the probability model allows us to calculate the *sample size* (▶ Power and Sample Size) (Lenth 2001; Wittes 2002), i.e., the desirable number of biological and technical replicates, with two goals. First, sample size calculations are used to evaluate the expected operational characteristics of the experiment. The number of replicates should be large enough to enable the detection of biologically significant changes in the response, but not too large in order to optimize the cost. Second, sample size calculations allow us to compare various strategies of sample selection and resource allocation, e.g., various strategies of allocation samples to blocks. The optimal strategy will maximize our ability to detect a change in the response for a fixed number of biological and technical replicates.

In frequentist modeling, each test is characterized by five properties: (1) the probability of type I error (i.e., the probability of rejecting the null hypothesis when it is true), (2) the probability of type II error (i.e., the probability of not rejecting the null hypothesis when the alternative is true), (3) the extent of known sources of variation, (4) the biologically significant change in the response, and (5) the number of biological and technical replicates in each condition. A modification is introduced for experiments with multiple responses, such as RNA-seq. The probability of type I error in (1) is replaced with the false discovery rate, and an additional quantity (6), the expected proportion of the true null hypotheses in all tests, is specified. The quantities (l)–(6) are interconnected, and a prior specification of all of them but one allows us to solve for the remainder. In particular, the specification of (l)–(4) and (6) allows us to solve for the sample size.

Several general conclusions can be made from the calculations of sample size. Detection of smaller changes in the response between conditions requires more replication. Experiments with larger variation also require more replication. An increase in the number of biological replicates is always more efficient than an increase in the number of technical replicates, and experiments without biological replication cannot generalize their conclusions to the underlying populations. Experiments with more responses, and also experiments with a smaller proportion of expected changes in multiple responses, have more opportunity for false discoveries and therefore require more replication.

The implementation of the steps of the experimental designs described in this entry depends substantially on the characteristics of the biological system and on the technology at hand. The rapid technological advances introduce constant changes into how the experiments are performed and into the properties of the resulting datasets. At the same time, they motivate new developments in statistical methodology. As the result, experimental planning becomes increasingly complex and cannot always be achieved in a fully automated fashion. A close collaboration between experimentalists and statisticians at all steps of the experiment, starting from the earliest stages of experiment planning, is highly recommended.

Cross-References

- ▶ Analysis of Variance
- ▶ Design of Experiments
- ▶ Experimental Design, Variability
- ▶ False Discovery Rate (FDR)
- ▶ Fisher's Test
- ▶ Frequentist Inference
- ▶ Hypothesis Testing
- ▶ Hypothesis Testing, Bayesian vs Frequentist
- ▶ Hypothesis Testing, Parametric vs Nonparametric
- ▶ Mixed and Multi-Level Models
- ▶ Multiple Hypothesis Testing
- ▶ Poisson Regression
- ▶ Power and Sample Size
- ▶ Quantitative Experiment Design
- ▶ RNA-Seq
- ▶ Sample Variability, Inter-groups
- ▶ Sample Variability, Intra-groups
- ▶ Statistical Methods in Systems Biology
- ▶ Student's t-Test

References

Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. Genetics 185(2):405–416
Dobbin K, Simon RM (2002) Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18(11):1438–1445
Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8(6):469–477
Kreutz C, Timmer J (2009) Systems biology: experimental design. FEBS J 276:923–942 From Melissa
Lenth RV (2001) Some practical guidelines for effective sample size determination. The Am Statist 55(3):187–193
Montgomery DC (2001) Design and analysis of experiments, 5th edn. Wiley, New York
Oberg AL, Vitek O (2009) Statistical design of quantitative mass spectrometry-based proteomic experiments. J Proteome Res 8(5):2144–2156 Exp design
Oehlert GW (2000) A first course in design and analysis of experiments, 1st edn. W. H. Freeman, New York
Wang Z, Gerstein MB, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev 10:57–63
Wittes J (2002) Sample size calculations for randomized controlled trials. Epidemiol Rev 24:39–53, From Stephane

Deuterated Glucose (^{2}H-Glucose)

► Lymphocyte Labeling, Cell Division Investigation

Deuterated Water (^{2}H2O)

► Lymphocyte Labeling, Cell Division Investigation

Developmental Biology, Classical Sea Urchin Experiments

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

Definition

Manipulations of sea urchin embryos have been important benchmarks in the development of embryology, after the framework set by the *Entwicklungsmechanik* (e.g., Roux, His, etc.) in the 1890s: perturbing at various stages the development in order to unravel the mechanisms at each stage of development. Sea urchins were easy to handle model organisms, even if until Hans Spemann the embryological techniques were quiet crude.

Developmental Biology, Classical Sea Urchin Experiments, Fig. 1 Spemann Mangold experiment – When a small region of the embryo, the dorsal lip, is grafted to the opposite (ventral) side of a host gastrula embryo (on the left), the resulting Xenopus laevis tadpole develops a Siamese twin 3 days later (right)

The first crucial experiment was done by Hans Driesch (1894): a sea urchin, when divided after the first cell stage (or before the fourth), develops into two sea urchins. The later the stage, the smaller the embryos. This experiment was, first by Driesch himself, seen as an argument for vitalism (the vital power being efficient in both embryo; cutting it off should have simply broken a purely material disposal).

Hans Spemann did two important experiments: in the first, one half of a blastomere of a sea urchin develops into a complete one. In the second one, done with Hilde Mangold (1924), a set of sea urchin cells from one embryo induces Siamese twins when grafted in another urchin embryo (although not any cut of the urchin leads to this result) (Fig. 1). This second kind of experiments has been understood as an argument against preformism (see ► Preformation and Epigenesis). Specifically, it gave an evidence for embryonic induction. Spemann and Mangold saw as an "organizer" the substance which induces the second sea urchin embryo. More precisely, because the fate of the transplanted cells could therefore be traced during development, Spemann and Mangold were able to demonstrate that the graft became a notochord, yet induced neighboring cells to change fates. These neighboring cells adopted differentiation pathways that were more dorsal, and

produced tissues such as the central nervous system, somites, and kidneys. Afterward, against Spemann's vitalism, it has been proven that even killed cells from the organizer substance were able to induce the development, which prevented vitalistic interpretations of the organizers.

Cross-References

- ▶ Explanation, Developmental
- ▶ Preformation and Epigenesis

Developmental Cancer Networks

- ▶ Cancer Networks

Developmental Control 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

Cancer networks; Cenes; Developmental networks; Stem cell networks

Definition

A developmental control network or cene is a network that controls the development of multicellular organisms by controlling cell states. The nodes in the network are cell control states. Edges in the network denote cell actions including jumps to new cell states. Branches in the network denote cell division where each daughter cell enters a possibly new control state.

Characteristics

Developmental control networks or cenes can be linked together to form larger cenes. The global developmental control network in a genome is called the cenome. The topology of the cene determines its ideal dynamic phenotype. Developmental control networks can be deterministic or stochastic. They can be conditionally activated by satisfaction of some condition Φ. They can involve cell-cell signaling. Cenes are abstract but executable networks that guide the development of multicellular organisms. Changes in developmental control networks can result in major evolutionary transitions in the morphology and function of organisms.

Examples of cenes include ▶ stem cell networks, ▶ cancer networks and terminal, and progenitor cell Developmental Control Networks (for details see Werner 2011a, b).

Developmental Networks are Executable Networks

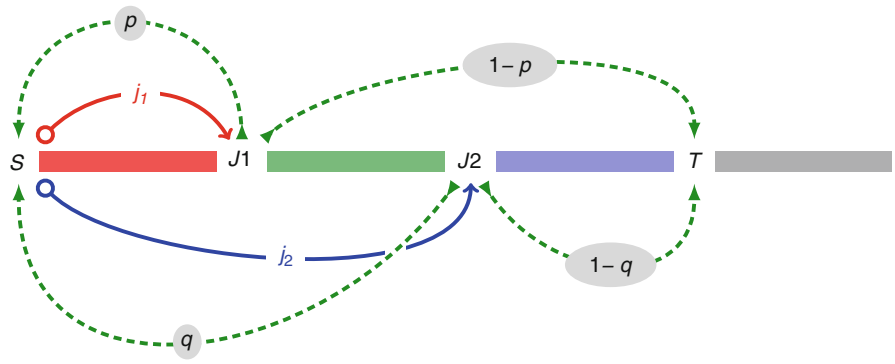
Developmental control networks or cenes are executable networks. The cell has an interpretive executive system, the IES, that interprets and executes the directives in developmental control networks. This system co-evolved with the developmental control networks (Werner 2011a).

Subnetworks within Developmental Control Networks

Any network can link subnetwork of another developmental network. This means the link has further developmental progeny generated by that subnetwork. In this way more and more complex cenes are built up.

Developmental Networks Subsume Gene Networks

Developmental control networks subsume and control lower level reactive network in the cell. The cell is a living organism that needs to react to local information to survive. Thus, the cell is locally reactive, but is globally controlled by its developmental networks. Hence, there are levels of network control in the cell. In generating the embryo, the global developmental network makes use of a cell's ability to move, to communicate, and to react to its environment. Hence, while the global developmental network does not uniquely determine the outcome of the ontogeny of an organism, it constrains to reach its ultimate form and functions.



Developmental Control Networks, Fig. 1 *Flexible exponential-linear stochastic stem cell network*. One stem cell (S) divides to produce two cells (J1 and J2) that each stochastically activates either the stem cell itself or a terminal cell (Werner 2011b)

Stochastic Developmental Networks

A stochastic developmental control network contains links to nodes that have an associated probability p and only jump to that node with probability p. Consider an example of a stochastic, flexible developmental network whose characteristics and topology are determined by the probabilities of its links:

The network in Fig. 1 is very flexible. Depending on the probability distribution, the network can range between being exponential, linear, or terminal, as well as every mixture in between. Furthermore, the network can be deterministic, stochastic, or mixture of both. This flexibility is partly the result of separating out the probability distributions for the behaviors of the two daughter cells. It shows that the architecture of the network imposes constraints on what kinds of developmental dynamics are in principle possible. The probability distribution presupposes a network architecture of possible developmental paths.

The cell S is only a stem cell stochastically and not intrinsically when $p < 1$ and $q < 1$. When $p = q = 1$ the network is deterministic exponential. When $p = 1$, $q = 0$ or $p = 0$, $q = 1$ the network is deterministic linear, that is, a deterministic 1st order geometric stem cell network. If $p = q = 0$ the network is terminal. When $p = 1$ and $q < 1$, or when $q = 1$ and $p < 1$, then the network is mixed deterministic linear with stochastic exponential tendencies.

If probabilities $p = q$ then as p and q approach 1 the network approaches the behavior of a deterministic exponential network. However, if p and q are different then this network can simulate a linear stochastic network as well when, for example, p approaches 1 and q approaches 0, or vice versa. As the probabilities p and q decrease, the more frequently the cancer stem cell results in a terminal tumor that does not develop further because it consists only of cells of terminal type T. This shows that stochastic cancer Stem Cell Networks (▶ Cancer Networks) can in some cases go into spontaneous remission. While this network can also exhibit exponential growth even in a stochastically linear probability distribution, because of the two backward loops, there is a diminishing probability that it remains exponential. Thus, whether this network results in linear or exponential proliferation depends on the probability distribution.

For this stochastic network, if the probabilities $p = 1-q$ then the higher the probability of p the more the network approximates a deterministic linear developmental network. Since, in this case the distribution is antisymmetric, the cell population partition of cell types consists of an equal number of exponential stem cells and terminal cells, with the majority of cells being linear stem cells. This corresponds to the observed distribution in epidermal basal stem cells. If, on the other hand, we have a symmetric distribution where $p = q$ then the higher the probability of p the more the network approximates a deterministic exponential network. Thus, the type of cancer network we have depends on the probability distributions over the connecting stochastic links.

The Network Evolution of Multicellular Organisms

The evolution of multicellular organisms is directly linked to the increasing complexity of their developmental control networks. These networks are an autonomous layer on top of normal gene networks (Werner 2011a, b).

Developmental control networks provide a powerful framework for explaining diverse developmental

phenomena beyond stem cells and cancer. These include bilateral symmetry and the evolution of the internal skeleton in bilateral symmetric animals (Werner 2012a, b).

Cross-References

- ▶ Cancer Networks
- ▶ Cene
- ▶ Cenome
- ▶ Stem Cell 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

Werner E (2012a) The origin, evolution and development of bilateral symmetry in multicellular Organisms. arXiv:12073289v1 [q-bioTO]. http://arxiv.org/abs/1207.3289

Werner E (2012b) How to grow an organism inside-out: evolution of an internal skeleton from an external skeleton in bilateral organisms. arXiv:12073624v1 [q-bioTO]. http://arxiv.org/abs/1207.3624

Developmental Module

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

Definition

A developmental module is a set of cells, or genes, which is more intrinsically connected than connected to its surroundings, which is constant across some clades, and which plays a specific causal role in development. For example endoderm is a developmental module, but also the Gene Regulatory Network of the skeletogenic micromere cell lineage in sea urchins (Oliveri et al. 2008). Developmental modules are important units of analysis because they are constant across many species, and developmental theory (unlike the evolutionary viewpoint) is mostly interested in commonalities across phyla (e.g., in regular constant developmental mechanisms such as apoptosis) rather than in differences. Developmental modules exist at many levels: genetic (GRNs), cellular (morphogenetic fields), and tissues (germ layers: ectoderm, etc.), and therefore may have some overlap.

Although quasi-independence defines modules in general, developmental modules are not necessarily the same modules as the ones identified by physiology or morphology (Winther 2001). For instance, the mesoderm – a developmental module – gives rise to the heart (a physiological module), but is also involved in the production of the vertebrate eye (another physiological module).

Modularity seems tied to evolvability (Wagner and Altenberg 1996) because it entails that no variation in a subpart is likely to change the functioning of the whole; therefore, mosaic evolution can be possible. Developmental modularity fulfills the same evolutionary requirements. It raises the question of its evolutionary origins: is it given with the first elementary eukaryote and generally the most basic of cell mechanisms? Or has it been selected for some advantages, or evolved as a by-product of selection for some developmental mechanisms? No consensus is yet attained.

Cross-References

- ▶ Explanation, Developmental

References

Oliveri P, Tu Q, Davidson E (2008) Global regulatory logic for specification of an embryonic cell lineage. PNAS 105(16):5955–5962

Wagner G, Altenberg L (1996) Complex adaptations and the evolution of evolvability. Evolution 50(3):967–976

Winther R (2001) Varieties of modules: kinds, levels, origins, and behaviors. J Exp Zool 291:116–129

Developmental Networks

- ▶ Cene
- ▶ Developmental Control Networks

Developmental Systems Theory

Niall Palfreyman
Biotechnology and Bioinformatics, Weihenstephan-Triesdorf University of Applied Science, Freising, Germany

Definition

Developmental Systems Theory (DST) regards evolution as change not in gene frequencies, but in the spatial and temporal structure of developmental processes or systems. DST arose out of Oyama's (2000) concerns about the role of ▶ *information* in the Modern Synthesis (MS) of evolutionary biology, which holds that genes, whether individually or in combination, encode information about the developmental construction of phenotypes. The MS thus makes an inherently dualistic distinction between *material* genes and bodies, and the ▶ *information* encoded in them, which "passes through bodies and affects them, but it is not affected by them on its way through" (Dawkins 1995).

A central difficulty of this account is that it views the genome as a *representation* of development, whereas the genome is in reality just one among many dynamical components in a developmental system which includes ribosomes, methylation, RNA splicing, transcription factors, intercellular signaling, and environmental and cultural resources. If we wish to hold to the metaphor of information, we are forced to regard that information as distributed throughout the entire developmental system. Yet as Oyama (2000) points out, this "means that information is not 'out there,' that it is not in the nucleus or anyplace else, that it is a way of talking about certain interactions rather than their cause or a prescription for them."

Oyama's point is that there certainly exist continuities and correlations between organisms and their environments which we may refer to as information; however, information in this sense is not a commodity which can be carried, stored, or transmitted in genes or in any other way. Nijhout (in Oyama et al. 2001), for instance, reports on a computer simulation of gene frequencies in a population subject to phenotypic selection. While the resulting phenotype exhibits gradual change over time, the correlation between genes and phenotype varies wildly, making it infeasible to talk about the informational content of any particular determinant in isolation from the entire genetic and environmental matrix.

Rather, information is an intrinsic aspect of the developmental process which is reconstructed in each individual organism, and is therefore strongly dependent on time and context. Consequently, in DST, the concept of a genetically programmed organism is replaced by the idea of a *life cycle* process which reconstructs itself out of an entire *developmental system* of genetic, environmental, and other resources available to the life cycle.

Like Oyama, various authors have objected to the use of the information metaphor. A recent analysis concluded that, instead of expanding it as in DST, *information* should be dropped from the biological lexicon because it is inimical to biological thought and thus is hindering the development of a theory of organisms (Longo et al. 2012).

The rather abstract formulation of DST has on the one hand excited criticism that it is far removed from the practicalities of evolution in a slowly changing environment. On the other hand it has also facilitated dialogue between a growing community of authors who emphasize the need to integrate developmental, evolutionary, and ecological processes into a single coherent theory. The link between DST and evolutionary-developmental biology is clear, and the manifestly hierarchical definition of the life cycle meshes well with multilevel theories of selection. Also, from the DST perspective, ontogeny, phylogeny, and niche-construction are respectively the enaction of life cycle, evolutionary and cultural continuity in the internal structural relationships of a developmental system.

References

Dawkins R (1995) River out of eden: a Darwinian view of Life. Basic Books, New York
Longo G, Miquel PA, Sonnenschein C, Soto AM (2012) Is information a proper observable for biological organization? Prog Biophys Mol Biol 109:108–114
Oyama S (2000) The ontogeny of information: developmental systems and evolution. Duke University Press, Durham
Oyama S, Griffiths PE, Gray RD (eds) (2001) Cycles of contingency: developmental systems and evolution. MIT Press, Cambridge, MA

Differential Adhesion Hypothesis

Anja Voss-Böhme
Center for Information Services and High Performance Computing (ZIH), Technical University Dresden, Dresden, Germany

Synonyms

DAH

Definition

The Differential Adhesion Hypothesis (DAH) is a theory advanced by Steinberg (1962) to explain the mechanisms of *cell sorting*. The latter are in vitro observations, where mixed heterotypic cell aggregates sort out into homotypic clusters. The sorting proceeds via the coalescence of small clusters into larger ones until a complete de-mixing of cell types is achieved. The DAH postulates that, analogous to the de-mixing of immiscible fluids, *differences in the cell-type-specific strengths* of intercellular adhesion cause measurable tissue surface tensions which drive the sorting process to minimize these tensions. It predicts that round cell aggregates emerge where either the cell type with the highest homotypic intercellular adhesion is in the center of the aggregate and is surrounded by cells with lower homotypic intercellular adhesion or a serial arrangement of homotypic clusters arises. The DAH has been challenged both by experimental and theoretical works; for a review see Green (2008). By now, it is fairly generally excepted that differential adhesion causes cell sorting, although there is a recent debate on whether additional intercellular interactions could contribute to cell sorting and affect the final sorted pattern as well (Green 2008; Krieg et al. 2008; Voss-Boehme and Deutsch 2010).

References

Green JB (2008) Sophistications of cell sorting. Nat Cell Biol 10(4):375–377
Krieg M, Arboleda-Estudillo Y, Puech PH, Käfer J, Graner F, Müller DJ, Heisenberg CP (2008) Tensile forces govern germ-layer organization in zebrafish. Nat Cell Biol 10(4):429–436
Steinberg MS (1962) On the mechanism of tissue reconstruction by dissociated cells. I. Population kinetics, differential adhesiveness, and the absence of directed migration. PNAS 48(9):1577–1582
Voss-Boehme A, Deutsch A (2010) On the cellular basis of cell sorting kinetics. J Theor Biol 263(4):419–436

Differential Equations with Deviating Arguments

▶ Dynamical Systems Theory, Delay Differential Equations

Differential Evolution

Zhong-Yuan Zhang
School of Statistics, Central University of Finance and Economics, Beijing, China

Definition

Differential evolution (DE) (Xu et al. 2007) tries to find the optimal solution of an objective function $f(x) : \mathbb{R}^n \to \mathbb{R}$ by collective-intelligence-based-search strategy. DE is initialized with a swarm of particles $\{x_i \in \mathbb{R}^n : i = 1, 2, \cdots, N\}$ that serves as the candidate solutions, and the particles are updated by turns. For example, when updating the particle x_i at step $t + 1$, one randomly selects three other distinct particles a, b, and c firstly, then every dimension j of x_i is mutated with the predefined probability Pr as follows:

$$x_{ij}^{(t+1)} = a_j + \gamma(b_j - c_j),$$

where the parameter $\gamma \in [0, 2]$ is predefined and called the differential weight.

Otherwise:

$$x_{ij}^{(t+1)} = x_{ij}^{(t)}.$$

Of $x_i^{(t)}$ and $x_i^{(t+1)}$, the one that has higher fitness with respect to the objective function $f(x)$ is passed on to the next generation.

The iteration process is terminated when some stop criterion is satisfied.

Cross-References

► Identification of Gene Regulatory Networks, Neural Networks

References

Xu R, Venayagamoorthy GK, Wunsch DC II (2007) Modeling of gene regulatory networks with hybrid differential evolution and particle swarm optimization. Neural Networks 20(8):917–927

Differential Expression Analysis

► Gene Expression Biomarkers, Ranking
► Relative Expression Analysis

Differential Expression of Homologous Genes

► Genomic Imprinting

Differential-Difference Equations

► Dynamical Systems Theory, Delay Differential Equations

Differentiation Potency

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

Definition

Differentiation potency is the property of a particular cell, such as a stem cell, to give rise to multiple distinct cell types. Three major categories of potency are totipotency, pluripotency, and multipotency. Totipotent cells are capable of differentiating into every cell type of a particular organism in addition to the extraembryonic tissues. An example of totipotent cells is those produced upon the fusion of a sperm and an egg up to the stage of a morula. Pluripotency describes the potential of a stem cell to differentiate into cells comprising any of the three germ layers: ectoderm (gut, lung, etc.), mesoderm (blood, bone, muscle, etc.), and endoderm (skin, nervous system, etc.). Examples of pluripotent stem cells include embryonic stem cells and induced pluripotent stem (iPS) cells. Multipotent cells are those that can differentiate into multiple cell lineages, but only to a restricted family of closely related cell types. Hematopoietic stem cells are an example of multipotent stem cells as they can give rise to all blood cells but not other tissue types such as neurons, muscle, or epithelium.

Cross-References

► Single Cell Assay, Mesenchymal Stem Cells

Diffuse Division

► Modeling, Cell Division and Proliferation

Diffusion Approximation to Chemical Master Equation

► Stochastic Processes, Fokker-Planck Equation

Diffusion Driven Lattice-Gas Model for Translation

► Stochastic Modeling of Translation Elongation and Termination

Diffusion Processes

▶ Stochastic Processes, Fokker-Planck Equation

Digital Metrics

Virginio Cantoni, Riccardo Gatti and Luca Lombardi
Department of Computer Engineering and Systems Science, University of Pavia, Pavia, Italy

Definition

Given two pixels p and q in a bi-dimensional space, with coordinates (x, y) and (s, t), the following distances are defined as follows:

- City block distance (4-connectivity) (Table 1)

$$D_4(p,q) = |x-s| + |y-t| \quad (1)$$

Digital Metrics, Table 1 City block distance: pixels classification in a 5 × 5 neighborhood

4	3	2	3	4
3	2	1	2	3
2	1	0	1	2
3	2	1	2	3
4	3	2	3	4

Digital Metrics, Table 2 Euclidean distance: pixels classification in a 5 × 5 neighborhood

$2\sqrt{2}$	$\sqrt{5}$	2	$\sqrt{5}$	$2\sqrt{2}$
$\sqrt{5}$	$\sqrt{2}$	1	$\sqrt{2}$	$\sqrt{5}$
2	1	0	1	2
$\sqrt{5}$	$\sqrt{2}$	1	$\sqrt{2}$	$\sqrt{5}$
$2\sqrt{2}$	$\sqrt{5}$	2	$\sqrt{5}$	$2\sqrt{2}$

Digital Metrics, Table 3 Chessboard distance: pixels classification in a 5 × 5 neighborhood

2	2	2	2	2
2	1	1	1	2
2	1	0	1	2
2	1	1	1	2
2	2	2	2	2

- Euclidean distance (Table 2)

$$D_E(p,q) = \sqrt{(x-s)^2 + (y-t)^2} \quad (2)$$

- Chessboard distance (8-connectivity) (Table 3)

$$D_8(p,q) = \max(|x-s|, |y-t|) \quad (3)$$

These definitions can be easily extended to the 3D space.

Cross-References

▶ Distance Transform and Travel Depth

Digital Organism

Christoph Adami
Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA

Synonyms

Avidian

Definition

A digital organism is a self-replicating computer program, usually within the Tierra or Avida evolution platforms.

Cross-References

▶ Artificial Evolution

Dimer

Xiaoping Liu
Institute of Systems Biology, Shanghai University, Shanghai, China

Definition

A dimer is a molecule consisting of two subunits called monomers. In biochemistry and molecular biology, dimers of proteins or nucleic acids are often observed. If the two subunits constituting a dimer are the same monomers, the dimer is called a homodimer. If the two subunits constituting a dimer are different monomers, the dimer is called a heterodimer.

Directed Acyclic Graph

Lin Wang
School of Computer Science and Information Engineering, Tianjin University of Science and Technology, Tianjin, China

Synonyms

Acyclic digraph; Directed acyclic network

Definition

A directed graph, or digraph, is a graph with directions assigned to its edges. A digraph is usually denoted by $D = (V, A)$ where $V=\{v_1, \cdots, v_n\}$ is a finite set of nodes and A is a set of ordered pairs of nodes called arcs. In a digraph, a cycle $C=\{c_1, \cdots, c_k\}$ of D is a closed and non-repetitive sequence of nodes in V such that $(c_j, c_{j+1}) \in A,\ j = 1, \cdots, k-1,\ c_1 = c_k$, and $c_i \neq c_j, i, j = 1, \cdots, k-1$. A directed acyclic graph is a digraph without any cycles.

References

Zhang XS (2000) Neural networks in optimization. Kluwer, Dordrecht

Directed Acyclic Network

▶ Directed Acyclic Graph

Directory

▶ Workspace

Disassembly of the Pre-initiation Complex

▶ PIC Disassembly

Discontinuous Epitope

Ramachandran Srinivasan
G.N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Delhi, India

Synonyms

Conformational epitope

Definition

Epitopes whose residues are distantly placed in the sequence brought together by physicochemical folding constitute discontinuous epitopes. The epitope structure is defined by protein folding process when the residues forming a discontinuous epitope are juxtaposed, enabling the antibody to recognize its three-dimensional structure.

Discrete Model

▶ Logical Model

Disease Classification or Discrimination

James J. Chen
U.S. Food and Drug Administration, National Center for Toxicological Research (HFT-20), Jefferson, AR, USA

Synonyms

Disease identification; Disease taxonomy

Definition

Disease classification is to systematically group diseases into classes in a hierarchical structure according to characteristics of disease: etiology, pathology, physiology, prognosis, and combinations of these. Every disease is grouped into one and only one class. Disease discrimination is used to identify a set of features that classy diseases into classes. Disease identification is to examine and classify diseases into specific classes. Disease taxonomy is the practice and science of classification of diseases. Disease taxonomy is a completing list of all disease types in the classification system.

International Statistical Classification of Diseases and Related Health Problems is a system of codes for classifying diseases and health problems published by the World Health Organization. The Medical Dictionary for Regulatory Activities is a clinically validated international medical terminology used by the biopharmaceutical industry and is used as the adverse event classification dictionary endorsed by the ICH.

Molecular classification of diseases based on genomic and proteomic profiling is used within the field of systems biology. Molecular classification uses statistical and machine learning methods to identify molecular markers of a specific disease or to develop prediction models to classify disease types (Baek et al. 2009). Classification models select a set of molecular features to discriminate between different types of disease or between disease and normal groups (Ramaswamy et al. 2001). The selected molecular features, individually or as a set, may be further developed to be probable biomarkers or valid biomarkers for disease classification or disease discrimination.

References

Baek S, Tsai C-A, Chen JJ (2009) Development of biomarker classifiers from high-dimensional data. Brief Bioinform 10:537–546

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci 98:15149–15154

Disease Databases

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

Definition

Disease databases are resources containing information on diseases, syndromes, and other medical conditions. They may provide specific information on the signs and symptoms, risk factors, treatment regimen, and/or results of studies done on known diseases and medical conditions. Diseases databases are available both in printed and online publications. Nowadays, online resources have already supplanted many printed publications. Selected online disease databases used in systems biology researches are presented below.

Characteristics

Classification of Disease Databases

A disease database can be categorized as general or specialized database depending on its scope. General disease databases contain a wider scope of diseases and medical conditions, while specialized databases are limited to certain types of diseases such as cancer, tropical diseases, or genetic diseases. Many online disease databases such as the "MedlinePlus," "Diseases Database," and the "Online Mendelian Inheritance in Man (OMIM)" are publicly accessible for free. Some databases are maintained by universities and government institutions, while others are maintained by private organizations or interest groups.

General Disease Databases

1. MedlinePlus (http://www.nlm.nih.gov/medlineplus/) is an online publication of the United States National Library of Medicine (NLM) that provides information on diseases, health conditions, treatments, and wellness issues. This database provides reliable and up-to-date health information for free. The health topics on the Website are categorized by body location/system, disorders or conditions, diagnosis and therapy, demographic groups, and health and wellness issues. One can find meaning of medical terms, search information on a disease, or view medical illustrations and videos on the Website. It is linked to the ▶ MEDLINE/▶ PubMed database for further information on researches on a certain disease and other related topics of interest such as systems biology (National Library of Medicine 2011).
2. Diseases Database (http://www.diseasesdatabase.com) is one of the free Websites that provides a medical textbook-like index and search options about human diseases, medical disorders, signs and symptoms, medications, and other information. It has dictionary-type definitions of terms which are linked to the National Library of Medicine's Unified Medical Language System. It also has subject-specific links to other Web resources. This cross-referenced database is funded by the Medical Object Oriented Software Enterprises Ltd and has been designed for physicians and other health workers and students. Researchers will find general knowledge about their disease of interest in the database and numerous links to other related databases and references. The Website follows the Health on the Net (HON) code of conduct for medical and health Websites and is edited by Dr. Malcolm H. Duncan who is also the Director of the database's sponsoring company (Duncan 2011).
3. OpenMED@NIC (http://openmed.nic.in/) is an open-access database which contains archives of peer-reviewed scientific articles and technical documents in the field of Medical and Allied Sciences. It includes a big collection of published articles about diseases which is indexed by year and subject. The site also includes conference proceedings and journal articles on biological phenomena, cells, enzymes, and genetic processes among others. No registration is required for searching the archive and downloading documents. However, a one-time registration is required for authors or owners who wish to submit documents for sharing to enable them to upload their files to the Website. All submitted documents are first reviewed by the OpenMED editor before it is posted online for free public access. Non-English contributed documents are also accepted provided that the abstract and keywords are in English. OpenMED is hosted by the Bibliographic Informatics Division of National Informatics Centre in India (OpenMED@NIC 2011).

There are other online databases that may be useful in doing research on diseases such as the MediLexicon (http://www.medilexicon.com/), the NCBI Biosystems Database (http://www.ncbi.nlm.nih.gov/biosystems/), Free Medical Journals (http://www.freemedicaljournals.com/), GPnotebook (http://www.gpnotebook.co.uk), Jeghers Medical Index (http://www.jeghers.com/), SciVerse Scopus (http://www.info.sciverse.com/scopus/), and also country-specific databases like the Indian Biomedical Journals Database (http://medind.nic.in/imvw/).

Specialized Disease Databases

1. Prion Disease Database or PDDB (http://prion.systemsbiology.net) is a public database for systems biology research on Prion disease, a fatal neurodegenerative disorder found in humans and animals. The database is a product of the joint effort between the Institute for Systems Biology in Seattle, Washington and the McLaughlin Research Institute in Great Falls, Montana, USA. It contains genetic data, microarray data, and other datasets that are useful for genetics and systems biology studies related to Prion disease. The information available in the PDDB are collected from public sources and collaborating laboratories. PDDB users can also share, store, and also analyze their own data through the Website. Analysis software tools are provided on the Website. PDDB is powered by the GDxBase software which is also used in different disease Websites (Gehlenborg et al. 2009).
2. The Human microRNA Disease Database (HMDD, http://202.38.126.151/hmdd/mirna/md/) is a database containing information on microRNAs (▶ Cell Cycle Regulation, microRNAs) and their disease associations. It contains microRNA names, diseases, dysfunction evidences, literature citations,

and tissue expression pictures in some cases. The database is not only useful for studying the association of microRNAs with diseases but also for investigating their roles in biological processes such as tissue differentiation, embryonic development, cell growth, proliferation, and ▶ Apoptosis (Lu et al. 2008).

3. Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/omim/) is a public database primarily designed for physicians, health professionals, students, and researchers concerned with genetic disorders. OMIM is the online version of the database of ▶ Mendelian traits and disorders which was initiated by Dr. Victor A. McKusick in the early 1960s. The printed versions entitled "Mendelian Inheritance in Man (MIM)" were published between 1966 and 1998. It was made available online starting 1987 through collaborative efforts of the National Library of Medicine and the Willian H. Welch Medical Library at Johns Hopkins University and was further developed in 1995 by the National Center for Biotechnology and Information (NCBI). OMIM contains information on Mendelian traits and disorders and more than 12,000 genes. It also contains full citation information, pictures of disorders (where appropriate), and links to other genetic resources. Its content is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine (OMIM 2011). Baxevanis (2002) published an article about the details of OMIM's layout of records and specific data entries as a guide for searching information for genes involved in human diseases.

There are numerous databases available online for specific diseases. It would be a big challenge to count and index these continuously growing resources. What was mentioned above are just the commonly accessed databases in systems biology research. To name additional ones, there is the Pathogenic Pathway Database for Periodontitis (http://bio-omix.tmd.ac.jp/disease/perio/), the Integrated Clinical Omics Database (http://omics.tmd.ac.jp/icod_en/portal/top.do), the Kyoto Encyclopedia of Genes and Genome (KEGG, http://www.genome.jp/kegg/), the Rare Metabolic Diseases Database (RAMEDIS, http://www.ramedis.de), BIOBASE Biological Databases (http://www.biobase-international.com/), and many more.

Cross-References

- ▶ Apoptosis
- ▶ Cell Cycle Regulation, microRNAs
- ▶ MEDLINE
- ▶ Mendelian Traits
- ▶ MEDLINE and PubMed

References

Baxevanis AD (2002) Searching Online Mendelian Inheritance in Man (OMIM) for information for genetic loci involved in human disease. Curr Protoc Bioinform 1.2.1–1.2.15 doi:10.1002/0471250953.bi0102s00/

Bibliographic Informatics Division of National Informatics Centre, India (2011) OpenMED@NIC. World Wide Web URL: http://openmed.nic.in/. Accessed 18 May 2011

Duncan MH (ed) (2011) Diseases database. World Wide Web URL: http://www.diseasesdatabase.com. Accessed 18 May 2011

Gehlenborg N, Hwang D, Lee IY, Yoo H, Baxter D, Petritis B, Pitstick R, Marzolf B, Dearmond SJ, Carlson GA, Hood L (2009) The Prion disease database: a comprehensive transcriptome resource for systems biology research in prion diseases. Database (Oxford): bap011. Epub 2009 Sep 17. doi:10.1093/database/bap011

Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q (2008) An analysis of human microRNA and disease associations. PLoS ONE 3(10):e3420. doi:10.1371/journal.pone.0003420

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD) (2011). Online Mendelian Inheritance in Man, OMIM (TM). World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/. Accessed 18 May 2011

National Library of Medicine (2011) About MedlinePlus. World Wide Web URL: http://www.nlm.nih.gov/medlineplus/aboutmedlineplus.html. Accessed 18 May 2011

Disease Identification

- ▶ Disease Classification or Discrimination

Disease Marker Identification

- ▶ Host–Pathogen Systems, Target Discovery

Disease Mechanism Networks

▶ Ontology Analysis of Biological Networks
▶ Functional/Signature Network Module for Target Pathway/Gene Discovery

Disease Ontology

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications, US National Library of Medicine, Bethesda, MD, USA

Definition

Background

Disease ontologies can be traced back to the seventeenth century, when health authorities in London used a standard list of about 200 causes of death to compile accurate health statistics known as the Bills of Mortality (Bodenreider 2008). This list was later integrated into the International Classification of Diseases, the 11th revision of which is currently under active development. Among the several hundredths of biomedical ontologies currently available, a few dozen provide coverage of diseases. A selection of these ontologies is presented in this brief review.

Biomedical Ontologies

Biomedical ontologies represent the properties of biomedical entities and their relations to other biomedical entities. As such, biomedical ontologies are artifacts used to represent and share knowledge about the biomedical domain (Bodenreider and Stevens 2006). More specifically, biomedical ontologies tend to focus on definitional knowledge (i.e., what is always true of biomedical entities), as opposed to assertional knowledge, usually found in knowledge bases. Ontologies also differ from terminologies, whose focus is purely naming, and from thesauri, in which knowledge is usually organized for a specific purpose (e.g., information retrieval). Despite these differences, the term "ontology" is often used loosely, as an umbrella name for these various kinds of artifacts.

Disease Ontologies

Disease ontologies are biomedical ontologies providing coverage for the domain of diseases, disorders, illness, etc. The degree to which these words are synonymous is subject to debate (Ceusters and Smith 2010), but disease ontologies generally cover conditions understood as or suspected of being a deviation from a healthy status and, less frequently, diagnostic criteria for these conditions. Some disease ontologies also cover the manifestations of these conditions, that is, the signs and symptoms associated with them. However, the relations between conditions and their manifestations are usually not recorded in ontologies, as such relations are not definitional for most diseases. In fact, except for the so-called pathognomonic manifestations, there is a probabilistic, not systematic relation between a manifestation and a condition. Phenotypes are the observable characteristics of organisms resulting from the genetic makeup of a particular organism. There is partial overlap between phenotypes and disease manifestations, and phenotypes may be covered by disease ontologies.

The principal use of disease ontologies is to support the annotation of diseases in biomedical datasets, including the curation of knowledge bases, the clinical documentation of electronic health records and the indexing of the biomedical literature. Disease ontologies are also used for aggregation purposes (e.g., for grouping myocardial infarction and mitral stenosis under cardiovascular diseases), as well as for clinical decision support (e.g., drugs contra-indicated with asthma should not be prescribed to patients diagnosed with specific forms of asthma, such as seasonal asthma and occupational asthma).

Characteristics

Existing Disease Ontologies and their Characteristics

In this section, we review 17 ontologies, which roughly qualify as disease ontologies according to the definition above. The list of ontologies is shown in Table 1, along with a URL from which more information can be obtained. A list of salient characteristics for disease ontologies is presented in Table 2. Finally, for each ontology, we provide a brief description and a list of

Disease Ontology, Table 1 List of disease ontologies

DO	*Disease Ontology* – http://diseaseontology.sourceforge.net/
DSM	*Diagnostic and Statistical Manual of Mental Disorders* – http://www.psych.org/
HPO	*Human Phenotype Ontology* – http://www.human-phenotype-ontology.org/
ICD	*International Classification of Diseases* – http://www.who.int/classifications/icd/en/
ICPC	*International Classification of Primary Care* – http://www.globalfamilydoctor.com/wicc/
IDO	*Infectious Disease Ontology* – http://www.infectiousdiseaseontology.org/
LOINC	*Logical Observation Identifiers Names and Codes* – http://loinc.org
MEDCIN	*MEDCIN* – http://www.medicomp.com/
MedDRA	*Medical Dictionary for Regulatory Activities* – http://www.meddramsso.com/
MeSH	*Medical Subject Headings* – http://www.nlm.nih.gov/mesh/
MPATH	*Mouse Pathology Ontology* – http://www.pathbase.net/
MPO	*Mammalian Phenotype Ontology* – http://www.informatics.jax.org/searches/MP_form.shtml
NCI Thes.	*NCI Thesaurus* – http://ncit.nci.nih.gov/
NDF-RT	*National Drug File-Reference Terminology* – http://evs.nci.nih.gov/ftp1/NDF-RT/
OMIM	*Online Mendelian Inheritance in Man* – http://www.ncbi.nlm.nih.gov/omim/
PATO	*Phenotypic Quality Ontology* – http://obofoundry.org/wiki/index.php/PATO:Main_Page
SNOMED CT	*SNOMED CT* – http://www.ihtsdo.org/

Disease Ontology, Table 2 List of salient characteristics for disease ontologies

Component	The disease ontology is a component of a broader ontology
Specialized	The disease ontology only covers a specific group of diseases
Human	The disease ontology mainly covers human diseases
OBO	The disease ontology is part of the Open Biomedical Ontologies (OBO) family of ontologies
Clinical	The disease ontology is mainly used in clinical practice
Definitions	The disease ontology includes definitions (textual or logical)
Translations	The disease ontology is available in other languages than English
Publicly available	The disease ontology is publicly available
X-ref	The disease ontology has cross-references to other disease ontologies (natively or through the UMLS)

characteristics summarized in Table 3, with additional notes in Table 4.

- *Disease Ontology* (DO): Controlled terminology originally created for annotation purposes as part of the NuGene project at Northwestern University. Still under development.
- *Diagnostic and Statistical Manual of Mental Disorders* (DSM): Standard classification of mental disorders in the United States, developed by the American Psychiatry Association and used by a wide range of mental health professionals across clinical settings.
- *Human Phenotype Ontology* (HPO): Controlled vocabulary for the phenotypic features encountered in human hereditary and other diseases, used for the annotation of the genetic diseases listed in OMIM. Developed by a consortium including Charité Hospital (Berlin) and the University of Cambridge (UK).
- *International Classification of Diseases* (ICD): Classification from the World Health Organization (WHO) family of health classifications, with many local adaptations. ICD9-CM, developed by the Center for Medicare & Medicaid Services (CMS) for use in the US, includes clinical modifications. Broad coverage of diseases and health problems.
- *International Classification of Primary Care* (ICPC): Classification of reasons for encounter, diagnoses or problems, and process of care. Developed by the World Organization of Family

Disease Ontology, Table 3 Some characteristics of 17 disease ontologies (see Table 2 for the definitions of the characteristics)

	Component	Specialized	Human	OBO	Clinical	Definitions	Translations	Publicly available	X-ref
DO	No	No	Yes	Yes	No	Textual	No	Yes	Native
DSM	No	Yes	Yes	No	Yes	None*	No	No	UMLS
HPO	No	Yes	Yes	Yes	No*	Textual	No	Yes	Native
ICD	No	No	Yes	No	Yes	None	Yes	No	UMLS
ICPC	No	Yes*	Yes	No	Yes	None	Yes	Yes	Native to ICD, UMLS
IDO	No	Yes	Yes	Yes	No	Textual	No	Yes	None
LOINC	No	Yes	Yes	No	Yes	Logical*	Yes	Yes	UMLS
MEDCIN	Yes	No	Yes	No	Yes	None	No	No	UMLS
MedDRA	No	Yes*	Yes	No	Yes	None	Yes	No	UMLS
MeSH	Yes	No	Yes*	No	No	Textual	Yes	Yes	UMLS
MPATH	No	No	No	Yes	No	Textual	No	Yes	None
MPO	No	No	No	Yes	No	Textual	No	Yes	None
NCI Thes.	Yes	Yes*	Yes	No	Yes*	Logical, textual	No	Yes	Native, UMLS
NDF-RT	Yes	No	Yes	No	Yes	Logical*	No	Yes	UMLS
OMIM	No	Yes	Yes	No	Yes	Textual*	No	Yes	UMLS
PATO	No	No	Yes*	Yes	No	Logical, textual	No	Yes	None
SNOMED CT	Yes	No	Yes	No	Yes	Logical	Yes	Yes*	UMLS

*Refer to additional notes in Table 4

Disease Ontology, Table 4 Additional notes on Table 3 items

DSM	Provides diagnostic criteria for many mental disorders
HPO	PhenExplorer is a clinical diagnostic tool based on HPO annotations of OMIM diseases
ICPC	Primary care can be considered a specialty
IDO	IDO borrows concepts from other OBO ontologies
LOINC	Although LOINC does not use description logics (DL), its organization is close to DL representation
MedDRA	MedDRA is not restricted to any medical specialties, but focuses on adverse events
MeSH	MeSH covers, but is not limited to human diseases
NCI Thes.	NCIt essentially covers cancers and cancer-related diseases; used in clinical research
NDF-RT	Weak logical definitions (primitive classes)
OMIM	OMIM contains extensive narrative descriptions more than definitions
PATO	PATO represents all kinds of phenotypes, including in humans
SNOMED CT	Freely available for use in the IHTSDO member countries

Doctors (Wonca). Coverage of diseases and health problems at the level of detail required for primary care.

- *Infectious Disease Ontology* (IDO): Set of ontologies for specific infectious diseases, including malaria, influenza, and tuberculosis, sharing a core ontology. Covers entities relevant to both biomedical and clinical aspects of most infectious diseases. Developed by the Infectious Disease Ontology Consortium.
- *Logical Observation Identifiers Names and Codes* (LOINC): Set of names and codes for laboratory and other clinical observations (elements of clinical phenotypes). Developed at the Regenstrief Institute. Coverage restricted to clinical observations.
- *MEDCIN*: Developed by Medicomp Systems, MEDCIN is a vocabulary for clinical documentation and a knowledge base for clinical decision support. It provides coverage for elements

including symptoms, medical history, physical examination, tests, and diagnoses.

- *MedDRA*: Created by a pharmaceutical industry trade group, the Medical Dictionary for Regulatory Activities (MedDRA) is a medical terminology used to classify adverse event information associated with the use of medications, vaccines, and medical devices, especially for reporting to regulatory agencies.
- *Medical Subject Headings* (MeSH): Controlled vocabulary developed by the US National Library of Medicine for the indexing and retrieval of the biomedical literature, especially in the MEDLINE bibliographic database. Broad coverage including diseases.
- *Pathbase pathology ontology* (MPATH): Ontology of mutant and transgenic mouse pathology phenotypes used for the annotation of Pathbase, a repository of histopathology images. Developed by the Pathbase European Consortium.
- *Mammalian Phenotype Ontology* (MPO): Controlled vocabulary for the annotation of mammalian phenotypes, currently used for the annotation of phenotypic data in mouse and rat databases. Developed at the Jackson Laboratory. Coverage restricted to phenotypes.
- *NCI Thesaurus* (NCIt): Controlled vocabulary developed by the National Cancer Institute to support the integration of information related to cancer research. Broad coverage including diseases.
- *National Drug File-Reference Terminology* (NDF-RT): Reference terminology for medications, providing information including pharmacologic class, therapeutic intent, mechanism of action, and physiologic effect. Produced by the US Department of Veterans Affairs, Veterans Health Administration (VHA). Coverage of diseases through their relations to drugs (therapeutic intent).
- *Online Mendelian Inheritance in Man* (OMIM): Knowledge base on human genetic diseases developed at Johns Hopkins University and available through the NCBI Entrez system. Coverage restricted to genetic diseases.
- *Phenotypic Quality Ontology* (PATO): Ontology of phenotypic qualities, intended for use in a number of applications, primarily defining composite phenotypes and phenotype annotation. Coverage restricted to phenotypes.
- *SNOMED CT*: The largest clinical terminology, maintained by the International Health Terminology Standard Development Organization (IHTSDO) for use in electronic health records and adopted by seventeen countries to date. Broad coverage including diseases.

Disease Ontology, Table 5 Ontology repositories providing coverage of disease entities

UMLS	*Unified Medical Language System* – https://uts.nlm.nih.gov
BioPortal	*NCBO BioPortal* – http://bioportal.bioontology.org/

Ontology Repositories

Many of the disease ontologies listed above are present in ontology repositories (Table 5), which offer a convenient way of integrating disease resources annotated to different ontologies. The *Unified Medical Language System* (UMLS) is a terminology integration system developed by the US National Library of Medicine. The UMLS establishes a correspondence among terms from different terminologies for a given biomedical entity. It integrates a number of the terminologies presented above, as well as many other biomedical terminologies. Developed by the National Center for Biomedical Ontology (NCBO), The *BioPortal* is another such repository, which offers mapping among terms from different ontologies. The BioPortal provides systematic coverage of the ontologies from the Open Biomedical Ontologies (OBO) family, as well as many other ontologies. The NCBO also indexes resources, such as clinical trials and gene expression databases, in reference to ontology entities from the BioPortal.

Acknowledgments This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). This manuscript is loosely based on a study of desiderata for disease ontologies, coauthored with Dr. Anita Burgun and presented to the First International Conference on Biomedical Ontology in 2009.

References

Bodenreider O (2008) Biomedical ontologies in action: role in knowledge management, data integration and decision support. Methods Inf Med 47(suppl 1):67–79

Bodenreider O, Stevens R (2006) Bio-ontologies: current trends and future directions. Brief Bioinform 7(3):256–274

Ceusters W, Smith B (2010) Foundations for a realist ontology of mental disease. J Biomed Semantics 1(1):10

Disease Progression

▶ Disease Progression Modeling

Disease Progression Modeling

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

Disease system analysis; Disease progression

Definition

Disease progression describes the change of disease status over time as function of disease process and treatment effects. The status of a subject, such as a patient, may be represented by a numerical quantity S (Chan and Holford 2001). In practice, ↗ *biomarkers* are frequently used as a proxy to monitor disease status. In a healthy subject, mechanisms of homeostasis ensure that the status S is relatively constant and remains within a normal range. The change in disease status shows minimal variation over time t, which mathematically is expressed as $dS/dt \approx 0$

A disease process is characterized by a change of S, taking it outside of the normal range. As disease progresses, the status S continues to move further away from the normal range, or in terms of change over time, $dS/dt \neq 0$.

The change of status S over time can be described by mathematical expressions that model biological processes. As an example from Dayneka et al. (1993), change can be explained in terms of a constant rate of synthesis k_{in} and a first-order process of decay or elimination k_{out}:

$$dS/dt = k_{in} - k_{out}S. \quad (1)$$

If $k_{in} = k_{out}S$, synthesis and elimination cancel each other out and the status is in homeostasis. A disease process (dp) may affect the rate of synthesis or elimination. This may be modeled by introducing temporal change in the rate of synthesis

$$dk_{in}/dt = f_{dp,\ synth}(k_{in},\ t) \quad (2)$$

or in the rate of elimination

$$dk_{out}/dt = f_{dp,\ elim}(k_{out},\ t). \quad (3)$$

Models of disease progression may provide insight into the biology behind the evolution of a target disease, and help identify the best treatment that either control or stop disease progression. The selection of the best course of action is a function of the effect of a given treatment on disease status over time. It is therefore paramount that disease progression models include biological aspects (i.e., genetic, transcription, and cross talks) that are involved in the evolution of the disease.

References

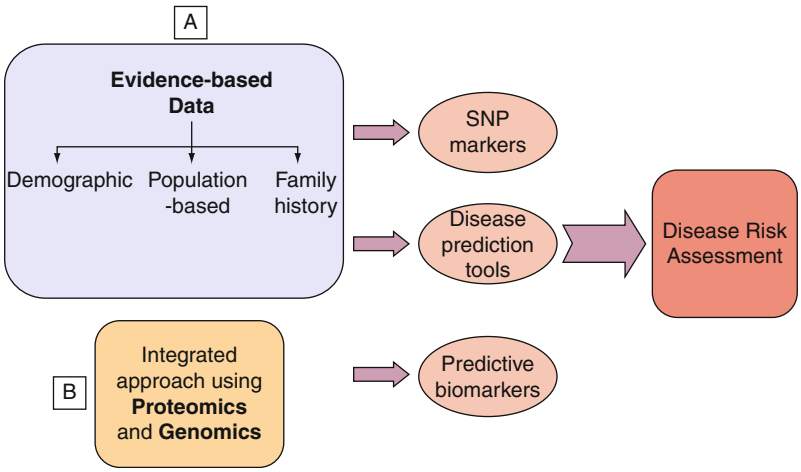
Chan PLS, Holford NHG (2001) Drug treatment effects on disease progression. Annu Rev Pharmacol Toxicol 41(1):625–659
Dayneka NL, Garg V, Jusko WJ (1993) Comparison of four basic models of indirect pharmacodynamic responses. J Pharmacokinet Pharmacodynamics 21:457–478. doi:10.1007/BF01061691

Disease Risk Assessment

Sanjeev Kumar[1] and Shipra Agrawal[2]
[1]BioCOS Life Sciences Private Limited, Bangalore, Karnataka, India
[2]BioCOS Life Sciences Pvt. Limited, Institute of Bioinformatics and Applied Biotechnology, Bangalore, Karnataka, India

Definition

Disease risk assessment could be defined as the systematic evaluation and identification of ▶ *risk factors* responsible for a disease, *estimation of risk levels* and finding possible *ways to counter the onset and progression* of a disease within the population.



Disease Risk Assessment, Fig. 1 *Factors involved in disease risk assessments*: The information from evidence-based data (demographic, population-based data, and family history) is processed to get the SNP markers. The SNP marker data is used to predict the disease risk. An integrated approach from proteomics and genomics provide clues to identify predictive biomarkers (expression), which also help in disease risk assessment

The following facts are very important and considered for disease risk assessment (Fig. 1):

- *Estimation of risk factors* involves the evaluation of variables and factors which can be indicative of the likelihood development of a particular disease. For example, the risk factors such as age, smoking, elevated levels of cholesterol and LDL, low levels of HDL, family history of premature coronary heart disease, etc. are indications of incurring cardiovascular disease.
- As the current focus has shifted toward preventive intervention than the disease cure, the risk assessment involves genotyping of tissues to identify "SNP markers" associated with a genetic disease. The SNP/genetic markers could be used to estimate the disease susceptibility in a new born/unborn child. This becomes useful in administrating the preventive treatment to the child.
- Integrated approaches in *proteomics* and *genomics* can create a rich resource of "predictive biomarkers" that is, signature molecules, which are biochemically measurable. Such biomarkers can help in *secondary disease prediction* (i.e., whether a diabetic person has a susceptibility of developing cardiovascular disease).
- *Evidence-based data* (▶ Evidence-based Medicine) such as family history, demographic, and clinical data can be used to develop *disease prediction tools*. Such predictive models having variables (factors which may influence the disease) gathered from *cohort-based studies* can be used in these studies. *Cohort-based studies* which are analytical investigations, are conducted on a group of people to gather evidences on a probable cause for a disease.

Disease Risk Assessment is a Challenge for Complex Diseases

Complex diseases result from the combined effects of multiple genetic and environmental factors and each such factor alone has only marginal contribution to the disease. Type 2 diabetes, coronary heart disease, and myocardial infarctions are examples of such diseases where genetic profiling has led to limited predictive values. Complete knowledge of causal mechanisms, which involves identification of all the possible combinations of the causal factors (gene–gene interactions, gene–environmental interactions), is required.

Monogenic diseases like Huntington disease, PKU, and hereditary cancers where identification of causal mechanisms and risk prediction in these diseases are comparatively straightforward are caused by DNA variations in a single gene. (Teramoto et al. 2008; Wei et al. 2009).

References

Teramoto T, Ohashi Y, Nakaya N, Yokoyama S, Mizuno K, Nakamura H, MEGA Study Group (2008) Practical risk prediction tools for coronary heart disease in mild to moderate hypercholesterolemia in Japan: originated from the MEGA study data. Circ J 72(10):1569–75

Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SF, Polychronakos C and Hakonarson H (2009) From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genetics 5(10): e1000678

Disease System Analysis

► Disease Progression Modeling

Disease System, Malaria

Pragyan Acharya, Manish Grover and Utpal Tatu
Department of Biochemistry, Indian Institute of Science, Bangalore, Karnataka, India

Synonyms

Bioinformatics; Genomics; Metabolomics; Parasitology; Proteomics; Transcriptomics

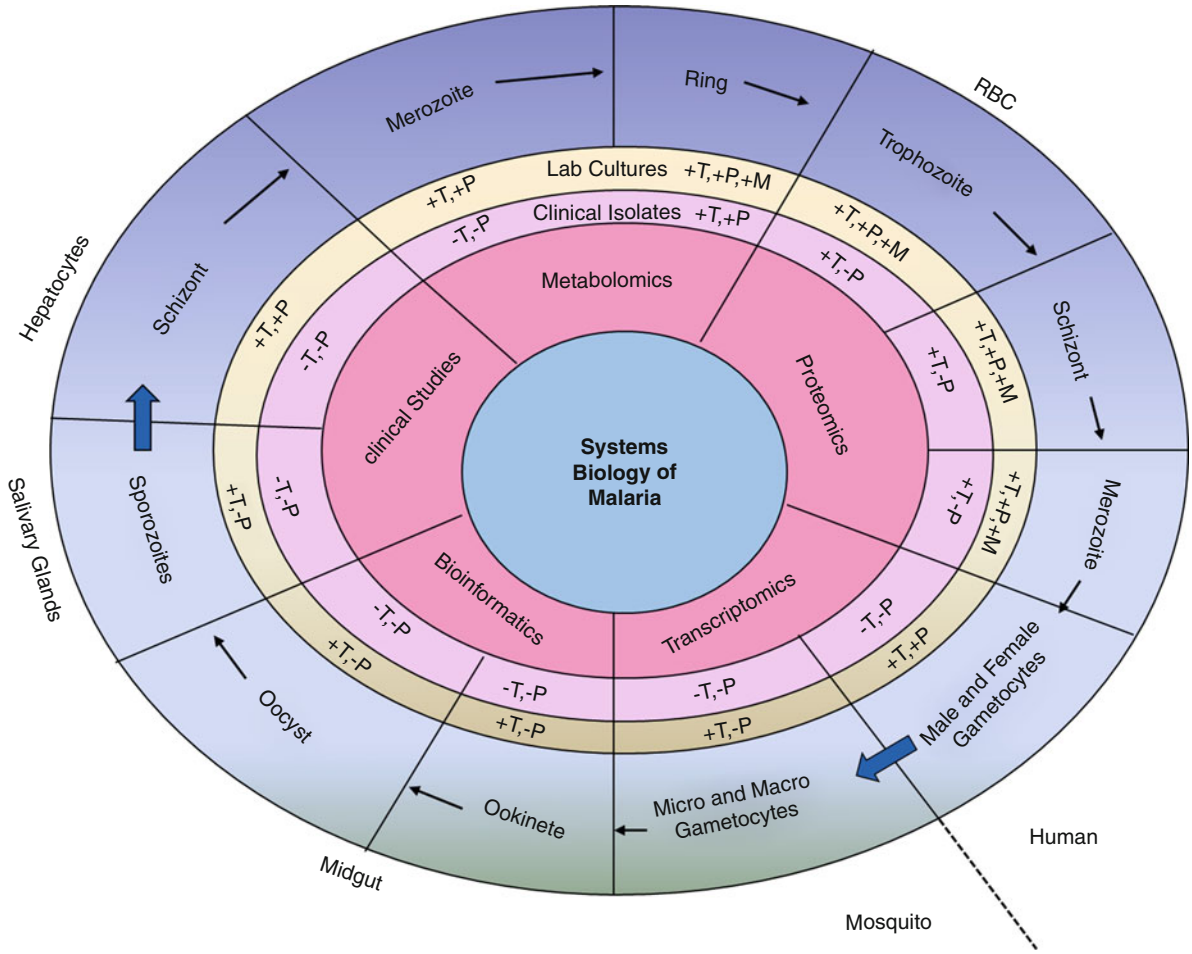
Definition

Malaria continues to be an enormous threat to the society, and the emergence of drug-resistant parasite strains has severely challenged the methods of disease prevention and control. In order to combat this situation, better understanding of the disease is required to identify new drug targets and vaccine candidates. Hence, we need to expand our research horizon and move beyond the regular reductionist approach of studying a single gene/protein or a biochemical pathway. With the help of high-throughput technologies and interdisciplinary tools, a holistic approach toward the understanding of the disease has been developed (Fig. 1). These initiatives commonly referred to as "systems biology" are further classified into genomics, proteomics, and metabolomics based on the component of the biological system they deal with. Genomics is concerned with the study of genomes of organisms and majorly involves DNA sequencing. It provides insights into the mechanism of gene expression and regulation and also establishes phylogenetic relationships between different organisms. An upcoming extension of genomic studies is transcriptomics which involves identification of actively expressing genes at any given point of time. Microarray technology is generally used to quantify the expression levels of various mRNAs and RNA-Seq provides details about the nucleotide sequence of transcripts being expressed. Proteomics deals with the comprehensive analysis of the entire complement of proteins present in a given system. Mass spectrometry is the main tool used for proteomic studies and it has revolutionized this field due to its extremely high sensitivity and diverse qualitative and quantitative applications. Metabolomics involves the study of the metabolites (metabolic intermediates, hormones, signaling molecules, etc.) present in a given system at a particular time and provides an instantaneous glimpse into the physiology of any system. Bioinformatics makes use of computational and statistical approaches to develop algorithms and databases which help to integrate and analyze the information obtained from genomic, proteomic, and metabolomic studies.

Characteristics

Malaria

Malaria is a debilitating disease affecting almost 200–300 million people worldwide annually. It is caused by the protozoan parasite belonging to the genus *Plasmodium*, which is carried by the female *Anopheles* mosquito and transmitted by mosquito bites in humans. Malaria inflicts mostly children and is centered around the tropical and subtropical regions of the world, being most widespread in Africa, some parts of South America, and Asian countries including India. In humans, malaria is caused by infection of the red blood cell with any of the five species of the parasite – *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale*, and *Plasmodium knowlesi*. Of these, maximum mortality and morbidity is caused by *P. falciparum* followed by *P. vivax*. All types of malaria are associated with febrile episodes with periodic paroxysms and chills in addition to nausea, headache, and general weakness. Population groups such as children, pregnant women, HIV/AIDS-infected individuals and travelers to



Disease System, Malaria, Fig. 1 Life cycle of *P. falciparum* and the systems biology approaches used for understanding the disease. "T," "P," and "M," respectively, represent the availability of transcriptomic, proteomic, and metabolomic evidence for the particular stage in the parasite's life cycle (Adapted and modified from Das A et al., Systems Biology of Malaria: An Indian Perspective. Biobytes Vol. 5, 2009)

malaria endemic areas are at a higher risk for malaria due to lower immunity (reviewed by Gracia 2010).

Malaria is a disease that has been a major research focus for several years. Yet, vaccines against malaria are elusive. The best bet in the development of malaria vaccines has been the RTS,S vaccine (Mosquirix) which has been recently demonstrated by Joe Cohen, a GlaxoSmithKline research scientist, to provide about 50% protection in infants (Olotu 2011). Still, the most popular and effective methods of controlling malaria are control of its mosquito vector and the use of bed nets. However, malaria is not completely preventable and continues to be an enormous threat to the society. This problem is further compounded by the emergence of drug-resistant strains of the parasite. Both *P. falciparum* and *P. vivax* have become resistant to common antimalarials and recently new *P. falciparum* strains resistant to artemisinin, one of the most effective antimalarial available, have also emerged (reviewed by Gracia 2010). Traditionally, most studies on malaria have focused on the biochemical, cell biological, and genetics of single gene or protein. In recent times however, the focus in malaria research has shifted to the analyses of gene and protein expression at the global level, revealing several interesting features of this parasite that may be useful in the discovery of novel antimalarial drug targets.

Systems Biology of Malaria

The recent availability of genome sequences of many pathogens has enabled systems analyses of infectious diseases. The advantage of using systems biology approaches is that they allow visualization of a system as a whole in response to its environment. In case of the malaria parasite, functional genomics data, global transcriptome data, and proteome data from both laboratory strains and clinically isolated parasites as well as metabolome data from laboratory strains are now available and have made striking revelations about the physiology of the malaria parasite.

Genomics

The deadliest form of malaria is caused by *P. falciparum*, which alone is responsible for about a million deaths annually. The complete life cycle of *P. falciparum* consists of three stages – the mosquito stage, the liver stage, and the human blood stages. The blood stages of the malaria parasite are responsible for most of the pathophysiology associated with malaria. As a result, the blood stages of *P. falciparum* have received maximum research focus. The release of the complete genome sequence of *P. falciparum* in 2002 incited systems biology analyses of the parasite. The 22.8 Mb genome of *P. falciparum* was shown to contain 14 linear chromosomes, a circular plastid-like genome, and a linear mitochondrial genome (Gardner et al. 2002). About 3.9% of the *P. falciparum* genome was shown to encode for different families of antigenic determinants essential for parasite virulence. Subsequent transcriptome analysis showed that about 60% of the 5,409 predicted open reading frames of the parasite were transcriptionally active in the blood stages of the parasite (Bozdech et al. 2003). This study utilized 7,462 individual 70 mer oligonucleotides representing 4,488 of the 5,409 ORFs manually annotated by the malaria genome sequencing consortium. The transcriptome analysis of the blood stages revealed for the first time that induction of any parasite gene occurred once per cycle and only at a time when it is required. More recently, transcriptome analyses of parasites isolated from malaria patients have been carried out, which have revealed the existence of distinct parasite physiologies in the wild (Daily et al. 2007). This study analyzed the transcriptome of malaria parasites isolated from venous blood samples from 43 malaria patients from Senegal with a diverse age range, parasitemia and hematocrit (reflecting severity of disease). The parasite transcriptome profiles thus obtained were then statistically clustered to obtain gene expression patterns for different parasite groups, if any. The genes were also mapped onto their *S. cerevisiae* orthologs in order to identify the possible pathways. This study revealed the presence of three distinct physiological clusters of *P. falciparum* as found within the malaria-infected individuals. These corresponded to first, active growth based on glycolytic metabolism, second, a starvation response accompanied by metabolism of alternative carbon sources, and third, an environmental stress response. The glycolytic state closely resembled the ring stages of the parasite in culture; however, the other two states were novel and specific to the parasites in vivo, indicating that gene expression profiles in the wild may differ significantly from those in cultures.

Proteomics

Although interesting, transcriptomes may not reveal the true physiological states of clinical malaria parasites. There is increasing awareness about greater reliability and accuracy of proteomics over transcriptome analyses. It has been demonstrated that in many cases transcriptome profiles are not reflective of the protein complement of a cell, as temporal and spatial differences arise between the two. In fact, studies that correlate the transcriptome and proteome data from laboratory cultures of *P. falciparum* indicate that there is a significant time delay between the abundance of transcripts and corresponding proteins in the asexual stages of the parasite (Le Roch et al. 2003). Most recently, the clinical proteome of *P. falciparum* and *P. vivax* has been reported for the first time (Acharya et al. 2009; 2011). Although proteomic analyses of the parasite-infected RBC had been carried out earlier, the proteomics analysis of clinical parasites isolated directly from patients is a challenge due to several factors such as low parasite density in the venous blood of patients and the presence of abundant host proteins that mask identification of low abundant parasite proteins. This study had identified about 100 proteins from the major malaria parasite *P. falciparum*. The highlights of the study were the identification of several well-known and putative drug targets in *P. falciparum* and the detection of several proteins

in clinical parasites that were not expressed by the laboratory culture of the parasite. This study is the first of its kind since it utilized parasites directly isolated from the peripheral blood without culture adaptation or influence by external factors.

Metabolomics

More recently, metabolome analysis of the parasite has been carried out using laboratory cultures. Metabolomic analyses in models of infectious diseases have been especially useful in elucidating metabolic modulation of the host by the parasite and may aid in the detection of biomarkers. The most striking revelation of the metabolome analysis of *P. falciparum* has been the presence of citrate, aconitate, and α–ketoglutarate, which are metabolic intermediates of the tricarboxylic acid (TCA) cycle, in the parasite asexual stages. The study not only provided the evidence for a functional TCA cycle but also suggested it to be organized in a branched pathway rather than the canonical cyclic pathway, with amino acids glutamate and glutamine as the major carbon sources (Olszewski et al. 2010). This study further highlighted the importance of metabolomic technologies in elucidating the architecture of metabolic networks and identification of novel plausible drug targets. Metabolomic analysis of urine and plasma samples from *P. berghei*–infected mice (rodent model of malaria) revealed the presence of a unique biomarker, pipecolic acid, in the urine of malaria-infected mice which was completely absent in urine samples collected from normal mice (Li et al. 2008). This suggests the use of metabolic profiling as a diagnostic tool in malaria infections.

Bioinformatics

In addition to the above experimental approaches, bioinformatics analysis of *Plasmodium* genes at a systems level has revealed several interesting features of the parasite. A systems level interactome analysis of chaperones in *Plasmodium falciparum* has been reported based on the presence of orthologs of parasite proteins and established yeast-two-hybrid data (Pavithra et al. 2007). This study predicted the possible roles of several hypothetical proteins based on their interactions with *P. falciparum*–encoded chaperones and uncovers parasite-specific chaperone-dependent pathways. In addition, the study proposed a putative mechanism by which Geldanamycin, an Hsp90 inhibitor shown to abrogate parasite growth in cultures, may work reiterating the usefulness of such a systems level approach in generating testable hypotheses (Banumathy et al. 2003). Similarly, another bioinformatic study addressed the question of PfEMP1 diversity, which is the major antigenic protein present on the infected erythrocyte membrane and a potential vaccine candidate for malaria. The study was conducted in seven genomes using sequence alignment and distance tree analysis (Rask et al. 2010). It described multiple novel features about PfEMP1 domain organization thereby providing a platform for the understanding of PfEMP1 expression and function.

Systems Analysis of Malaria Vector

Malaria eradication programs center around vector control methods in a large way and thus understanding of the biology of the *Anopheles* mosquito is equally important as that of the parasite biology inside the human host. Malaria transmission in the wild is largely determined by vectorial competence i.e., ability of the particular mosquito species or strain to replicate and transmit the parasite to human host. Systems biology approaches have unraveled some aspects of this host-pathogen interaction.

Whole genome sequencing of *Anopheles gambiae* and other insects has enabled systematic analysis of divergent protein families which have evolved to facilitate specific interactions with the parasite. The virulence caused by the sporogenic development of the parasite in the mosquito imposes a fitness cost on the vector which can be expressed as reduction of mosquito survival or decrease in parasite fecundity. Vector longevity will definitely have a better impact on malaria transmission and this is also emphasized by the expanded immune repertoire present in the mosquito. Large-scale functional genetic screen of mosquito genes pointed out PRRs (putative pattern recognition receptors), TEP (thioester-containing proteins), CTL (C-type lectins) and LRIM (leucine rich-repeat immune proteins) to be especially important in this regard (reviewed by Bongfen et al. 2009). This conjecture was further supported by a global proteomic study on the saliva of *A. gambiae* which revealed overrepresentation of proteins involved in signaling and immune response functions (Choumet et al. 2007). A recent large-scale genomic study addressed

the susceptibility of different *Anopheles* populations to infection by human malaria parasites and found that the exophilic subgroup is more susceptible than the endophilic subgroup (Reighle et al. 2011). Such studies will play an important role in determining the epidemiology of malaria and develop better control and prevention strategies.

Conclusion

From the above discussion, it is clear that systems level approaches have been extensively initiated to understand the biology of the malaria parasite and have indeed been fruitful in uncovering several novel aspects of parasite biology that would not have been possible by classical studies involving one gene or one protein. There are emerging evidences which highlight that physiology of the parasite is different under in vitro and in vivo conditions (reviewed by LeRoux et al. 2009), and, thus, efforts should be made to understand the biology of the parasite directly isolated from malaria patients. This will give insight into the real disease scenario and enable better understanding of disease pathogenesis, transmission, and therapeutic targets. In the future, these studies may reveal several features of malaria infection and will thereby contribute to the discovery of novel antimalarial drug targets as well as vaccine candidates. Such studies have provided an opportunity for scientific groups with different expertise to pursue research in a coherent manner, and this forms the key feature for the success and impact of systems biology studies.

Cross-References

- Biological Disease Mechanism Networks
- Biomarkers
- Biomarkers, Clinical Relevance
- Graph Alignment, Protein Interaction Networks
- Gene Expression Biomarkers
- Host–Pathogen Interactions
- Interactome
- Mass Spectrometer
- Proteomics
- Systems Pharmacology, Drug-Target Networks
- Vaccinomics

References

Acharya P, Pallavi R, Chandran S, Chakravarti H et al (2009) A glimpse into the clinical proteome of human malaria parasites *Plasmodium falciparum* and *Plasmodium vivax*. Proteomics Clin Appl 3:1314–1325

Acharya P, Pallavi R, Chandran S, Dandavate V et al (2011) Clinical proteomics of the neglected human malarial parasite *Plasmodium vivax*. PLoS One 6(10): e26623

Banumathy G, Singh V, Pavithra SR, Tatu U (2003) Heat shock protein 90 function is essential for *Plasmodium falciparum* growth in human erythrocytes. J Biol Chem 278(20): 18336–18345

Bongfen SE, Laroque A, Berghout J, Gros P (2009) Genetic and genomic analyses of host-pathogen interactions in malaria. Trends Parasitol 25(9):417–422

Bozdech Z, Llinas M, Pulliam BL, Wong ED, Zhu J, DeRisi JL (2003) The transcriptome of the intraerythrocytic cycle of *Plasmodium falciparum*. PLoS Biol 1(1):E5

Choumet V, Carmi-Leroy A, Laurent C, Lenormand P et al (2007) The salivary glands and saliva of *Anopheles gambiae* as an essential step in *Plasmodium* life cycle: a global proteomic study. Proteomics 7(18):3384–3394

Daily JP, Scanfeld D, Pochet N, Le Roch K et al (2007) Distinct physiological states of *Plasmodium falciparum* in malaria infected patients. Nature 450:1091–1095

Gardner MJ, Hall N, Fung E, White O et al (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419(6906):498–511

Gracia LS (2010) Malaria. Clin Lab Med 30(1):93–129

Le Roch KG, Zhou Y, Blair PL, Grainger M et al (2003) Discovery of gene function by expression profiling of the malaria parasite life cycle. Science 301(5639):1503–1508

LeRoux M, Lakshmanan V, Daily JP (2009) *Plasmodium falciparum* biology: analysis of in vitro versus in vivo growth conditions. Trends Parasitol 25(10):474–481

Li JV, Wang Y, Saric J, Nicholson JK et al (2008) Global metabolic responses of NMRI mice to an experimental *Plasmodium berghei* infection. J Proteome Res 7(9):3948–3956

Olotu A, Moris P, Mwacharo J, Vekemans J et al (2011) Circumsporozoite-Specific T Cell Responses in Children Vaccinated with RTS,S/AS01(E) and Protection against *P falciparum* Clinical Malaria. PLoS One 6(10): e25786

Olszewski KL, Mathew MW, Morrisey JM, Garcia BA, Vaidya AB, Rabinowitz JD, Llinas M (2010) Branched tricarboxylic acid metabolism in *Plasmodium falciparum*. Nature 466(7307):774–778

Pavithra SR, Kumar R, Tatu U (2007) Systems analysis of chaperone networks in the malarial parasite *Plasmodium falciparum*. PLoS Comput Biol 3(9):1701–1715

Rask TS, Hansen DA, Theander TG, Gorm Pedersen A, Lavstsen T (2010) *Plasmodium falciparum* erythrocyte membrane protein 1 diversity in seven genomes – divide and conquer. PLoS Comput Biol 6(9):e1000933

Reighle MM, Guelbeogo WM, Gneme A, Eiglmeier K et al (2011) A cryptic subgroup of *Anopheles gambiae* is highly susceptible to human malaria parasites. Science 331(6017):596–598

Disease System, Parkinson's Disease

Rajeswara Babu Mythri[1], Shireen Vali[2] and M. M. Srinivas Bharath[1]
[1]Department of Neurochemistry, National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore, Karnataka, India
[2]Cell Works Group Inc., Bangalore, India

Synonyms

Movement disorders; Neurodegenerative diseases

Definition

Parkinson's disease (PD) is a movement disorder and an age-associated neurodegenerative disease. Motor impairment in PD is caused by degeneration of dopaminergic neurons in the substantia nigra (SN) of the midbrain which leads to depletion of the neurotransmitter dopamine in the striatum. Neurodegeneration in PD is a culmination of several interdependent processes including oxidative stress, mitochondrial damage, protein aggregation, proteasome inhibition, etc. However, the dynamics and interdependence of the disease pathways, their temporal order, synergy, and regulation cannot be understood completely by isolated experiments. Systems biology based dynamic modeling and predictive analyses supported by experimental data can address this issue and provide a superior understanding of PD pathology at the molecular level aimed at improved diagnosis and therapy.

Characteristics

Parkinson's disease (PD) is an age-associated neurodegenerative disease clinically defined as a movement disorder. The clinical symptoms of PD include akinesia (impaired body movement), rigidity, resting tremor, and postural instability. The patients also exhibit nonmotor symptoms including autonomic dysfunction, cognitive, neurobehavioral, sensory, and sleep dysfunctions and dementia. Most PD cases are sporadic arising spontaneously with unknown origin. PD is common among subjects aged >60 years with the prevalence and severity increasing with age. Interestingly, men are more affected than women (Hoehn and Yahr 1967; Chaudhuri et al. 2006). Diagnosis of PD even today is symptomatic and depends on neurological recognition of clinical symptoms.

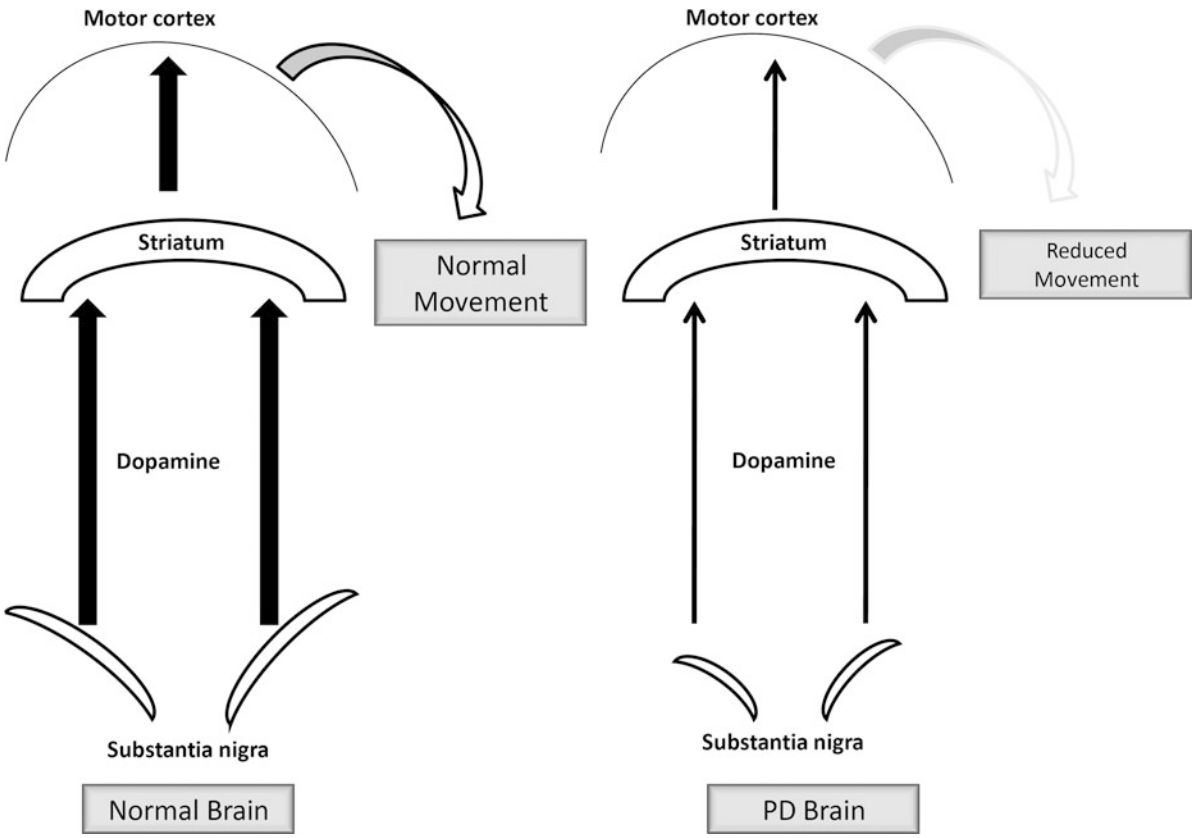
Pathology of PD

Voluntary movement is controlled in the brain by the nigrostriatal pathway involving dopaminergic neurons originating from the substantia nigra (SN) region of the ventral mid brain and projecting into the striatum (Fig. 1). These neurons synthesize and supply the neurotransmitter ▶ dopamine (DA) to the striatum where it participates in a complicated neurochemical network that ultimately controls body movement. Biochemical, pathological, and imaging data from PD patients have indicated that a gradual loss of these dopaminergic neurons causing decreased DA supply result in motor impairment and PD symptoms (Burke 1998).

Although there are several approaches for PD therapy, a permanent cure is not available. Most pharmacological drugs strive at replenishing the lost DA; but many patients develop motor complications with chronic treatment (Diaz and Waters 2009). Further, most drugs do not exhibit significant neuroprotection. The failure to obtain an effective PD drug is attributed partly to the lack of complete understanding of the pathology at the molecular level. Therefore, there is a need to obtain a comprehensive network of interacting molecules and pathways involved in neuronal death in PD.

Molecular Mechanisms in PD

The etiology of PD is contributed by a combination of physiological ▶ aging, environmental factors, and genetic mutations. Neurodegeneration in PD involves interdependent mechanisms such as ▶ oxidative stress, mitochondrial damage (▶ Mitochondrial Dysfunction, Parkinson's Disease), proteasome inhibition (▶ Proteasome Inhibition, Parkinson's Disease), protein ▶ aggregation, neuroinflammation, etc. (Betarbet et al. 2002). The brain is particularly vulnerable to oxidative stress because, it (1) consumes relatively



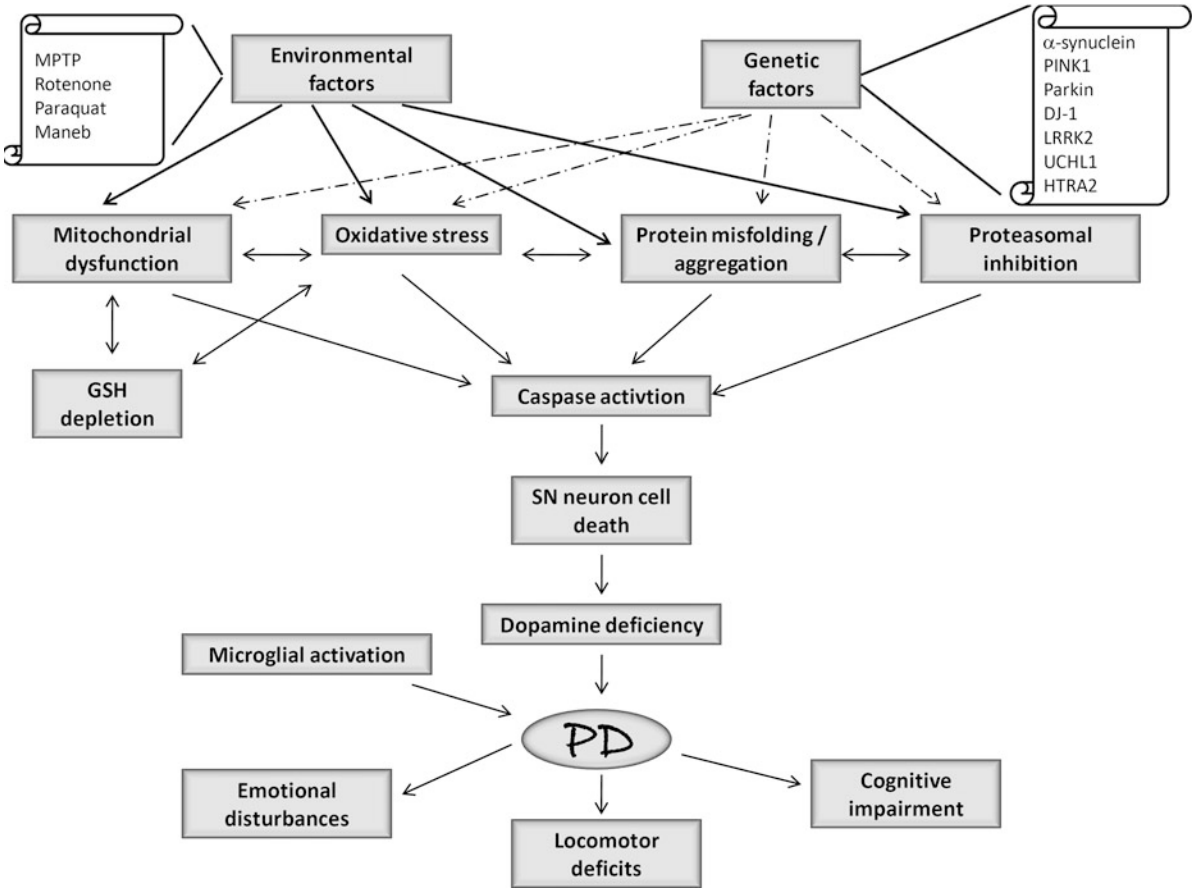
Disease System, Parkinson's Disease, Fig. 1 Nigrostriatal pathway in normal and PD brains

higher amounts of oxygen (2) accumulates lipids and iron that promote oxidative damage (3) has lower antioxidant defenses. Among the brain cells, neurons in general and dopaminergic neurons in particular are highly susceptible to oxidative damage because they (1) generate oxidative stress during dopamine metabolism (2) have lower levels of the antioxidant ▶ glutathione (GSH) and (3) increased iron content. Mitochondrial damage in SN neurons is caused by oxidative stress mediated selective inhibition of mitochondrial complex I (CI) (Bharath et al. 2002; Abou-Sleiman et al. 2006).

Inhibition of protein processing is linked to aggregation of cellular proteins in the SN dopaminergic neurons as intraneuronal protein deposits called "Lewy bodies (LBs)" (▶ Lewy Bodies, Role of Alpha-syn). LBs represent a pathological hallmark of PD with α-synuclein (α-syn) protein as the major component (Shults 2006). Neuroinflammatory pathways also contribute to the degenerative process in PD (Glass et al. 2010). Although PD involves a complex network of events, the precise relationship, synergy, and temporal order among these pathways are not clear (Fig. 2).

Systems Biology Applications in Neurodegeneration

It could be surmised that designing experiments to analyze the comprehensive dynamics of all the events in PD could be difficult. However, systems biology based in silico predictive technology can address this issue by recapitulating an accurate and simultaneous view of individual and interlinked pathways at the molecular level. Such a virtual platform should include all the relevant proteins and their genes and transcripts with their relationship quantitatively represented. The platform should integrate these species in intra and intercellular pathways providing a comprehensive view at the cellular and tissue level. The platform should be certified against predefined in vitro and in vivo studies reported in scientific literature with flexibility to include new data. Such a platform could assay all intermediate and endpoint biomarkers and



Disease System, Parkinson's Disease, Fig. 2 Interplay among different disease pathways in PD

manipulate different triggers, inhibitors, and activators. This also includes percentage, knockout, dose–response, overexpression, and mutational analysis. Different customized studies can be defined and executed through an interactive graphical user interface mode or a high-throughput approach.

Accordingly, a typical in silico PD platform should include an exhaustive list of ▶ molecular markers representing important pathways in normal physiology and neurodegeneration as follows:

1. Markers that represent physiological aging to distinguish aging and neurodegeneration. Aging state is depicted by markers of oxidative stress, mitochondrial activity, and quantitative changes in insulin growth factor, melatonin, homocysteine, etc.
2. Pathways representing PD state including neurodegenerative and neuroinflammatory events and related endpoint biomarkers of mitochondrial dysfunction, oxidative stress, endoplasmic reticulum stress, proteasomal dysfunction, protein aggregation, neurotrophin/growth factor signaling, etc.
3. Amalgamation of signals from all cell types involved, including dopaminergic neurons, microglia, and astrocytes and relevant pathways covering all disease stages.
4. Incorporation of different triggering factors including environmental toxins (rotenone, 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine or MPTP (▶ Parkinson's Disease, MPTP) and genetic factors (familial mutations) which induce neurodegeneration in the aging neurons.

We recently generated such a platform to carry out simulations linking disease pathways in PD with experimental validation (Vali and Bharath 2009). The methods in model construction are as follows and are applicable to similar biological phenomena: A bottom–up approach was adopted to build the platform. Initially, the various phenomena related to

neuronal dynamics were built as base modules and progressively integrated. These modules included:

(a) Comprehensive mitochondrial bioenergetics
(b) Generation of reactive oxygen and nitrogen species (ROS/RNS) including dopamine metabolism and iron
(c) GSH and Ca^{2+}dynamics
(d) Proteasome inhibition and protein aggregation

Modeling involved a detailed study of the individual pathways and components. Interactions among these components and their regulation were arranged to obtain a basic interaction map. The static map was made dynamic by incorporating (1) physiological concentrations of individual molecules and enzymes/proteins (2) catalytic rates and reaction mechanisms in flux equations. Most of these data were obtained from published scientific literature involving experimental methods. After individual modules were built and validated against published data, the cross-talk among these phenomena was integrated such that the output from a module either created the input to another module or provided regulatory control. These interactions cross-linked and integrated the base modules thus generating a complete integrated system. The performance of such a complex network would be different from isolated experimental data obtained from these phenomena. The modeling of kinetic phenomena such as time-dependent potential differences in trans-mitochondrial membrane potential, proton motive force, ion fluxes, and other metabolic reactions were performed utilizing modified "Ordinary Differential Equations" and "Mass Action Kinetics" and the integration of the pathways were solved by the Radau and Euler methods. Such modeling experiments reconcile numerous formerly unrelated features of PD in a chronological manner thus elucidating disease progression based on the combined molecular actions of different mechanisms.

Using this platform, we carried out few studies exploring the mechanistic and therapeutic aspects of PD (Vali and Bharath 2009). Firstly, we integrated GSH metabolism and mitochondrial dysfunction associated with PD. This study inferred that the mitochondrial damage affected the cellular GSH synthesis thereby enhancing the oxidative damage and exacerbating neurodegeneration. Secondly, we have also traced the neuroprotective function of curcumin (a polyphenol from turmeric) and its bioconjugates. We found that curcumin and its conjugates induced GSH production, protected against oxidative stress and mitochondrial damage with therapeutic potential in PD. In a third study, modeling envisaged that the A53T mutant of α-syn can accumulate and disrupt mitochondrial function in dopaminergic neurons. In the presence of proteasome inhibition, mitochondrial turnover is further reduced, resulting in decreased ATP synthesis and in turn decreasing GSH synthesis.

Cross-References

- ▶ Aging
- ▶ Dopamine
- ▶ Genetic Factor in Parkinson's Disease
- ▶ Glutathione
- ▶ Lewy Bodies, Role of Alpha-syn
- ▶ Mitochondrial Dysfunction, Parkinson's Disease
- ▶ Molecular Markers, In Silico Parkinson's Disease Platform
- ▶ Oxidative Stress, Protein Damage
- ▶ Parkinson's Disease, MPTP
- ▶ Proteasome Inhibition, Parkinson's Disease

References

Abou-Sleiman PM, Muqit MM, Wood NW (2006) Expanding insights of mitochondrial dysfunction in Parkinson's disease. Nat Rev Neurosci 7:207–219

Betarbet R, Sherer TB, Di Monte DA, Greenamyre JT (2002) Mechanistic approaches to Parkinson's disease pathogenesis. Brain Pathol 12:499–510

Bharath S, Hsu M, Kaur D, Rajagopalan S, Andersen J (2002) Glutathione, iron and Parkinson's disease. Biochem Pharmacol 64:1037–1048

Burke RE (1998) Parkinson's disease. In: Koliatsos VE, Ratan RR (eds) Cell death and disease of the nervous system. Humana, Totowa, pp 459–475

Chaudhuri KR, Healy DG, Schapira AH (2006) Non-motor symptoms of Parkinson's disease: diagnosis and management. Lancet Neurol 5:235–245

Diaz NL, Waters CH (2009) Current strategies in the treatment of Parkinson's disease and a personalized approach to management. Expert Rev Neurother 9:1781–1789

Glass CK, Saijo K, Winner B, Marchetto MC, Gage FH (2010) Mechanisms underlying inflammation in neurodegeneration. Cell 140:918–934

Hoehn MM, Yahr MD (1967) Parkinsonism: onset, progression and mortality. Neurology 17:427–442

Shults CW (2006) Lewy bodies. Proc Natl Acad Sci USA 103:1661–1668

Vali S, Bharath MMS (2009) Dynamic virtual prototype of Parkinson's disease at the cellular and molecular abstraction level: focus on neurodegeneration physiology. Biobytes 5:40–46

Disease Taxonomy

▶ Disease Classification or Discrimination

Disease-oriented Causal Networks

Sanjeev Kumar[1] and Shipra Agrawal[2]
[1]BioCOS Life Sciences Private Limited, Bangalore, Karnataka, India
[2]BioCOS Life Sciences Pvt. Limited, Institute of Bioinformatics and Applied Biotechnology, Bangalore, Karnataka, India

Definition

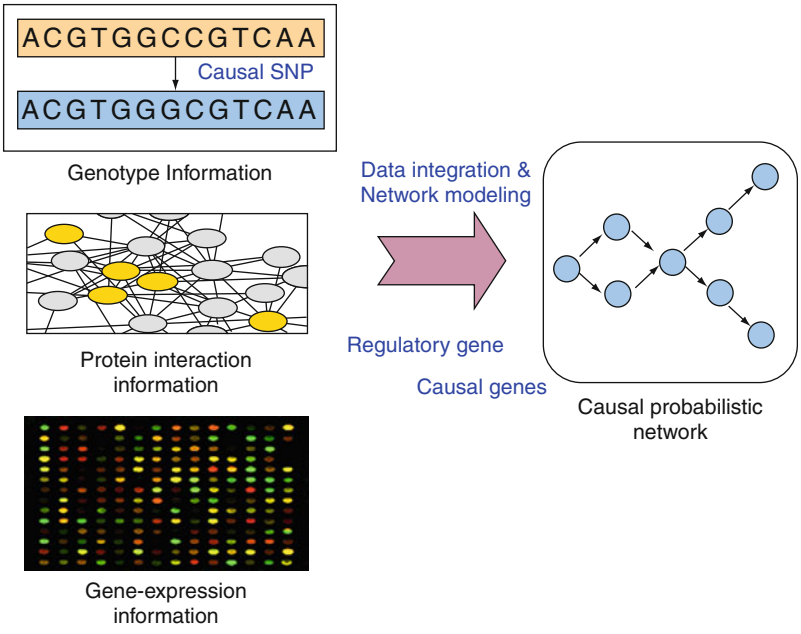
DOCN represents the relationship across disease, candidate genes, regulatory genes and their functions to define the causal relationship through a gene or protein network. The connectivity across the network is established through directed graphs where the nodes or genes are variables, which are connected through edges. The graph indicates interaction between and across the genes (nodes) to ultimately describe the causal mechanism of a disease. In the DOCN, the graph shows how the change of the state of one variable affects the certainty of the state of another variable hence these causal networks are graphical representations of causal relationships between the variables of the graph. For example, obesity is one of the factors for developing insulin resistance, which in turn is the cause for type 2 diabetes.

Characteristics

Usage of the Disease-Oriented Causal Networks

Disease mechanisms can be explained using causal networks. These networks are used to identify genes, underlying pathways or interactions that can be causal driver of the disease. For example, causal genes involved in the pathogenesis of a disease can be revealed by studying gene regulatory networks, which are directed graphs and depict causal interactions and functional correlation between the genes.

The intricate genetic interactions and gene-to-phenotype correlation of a complex disease could be better understood by integrating data from multiple sources (genotype information, gene-expression, PPI information, etc.) to construct the causal probabilistic networks. Such networks are based on probabilistic approaches where it is believed that one node in the network affects the state of other nodes in the network (Fig. 1).



Disease-oriented Causal Networks, Fig. 1 Data integration and network modeling to create a causal probabilistic disease network

Data Integration to Create Causal Disease Networks

The advantage of integrating the multiple molecular data that is, RNA profiling, genotyping/SNP typing, protein expression and/or protein–protein interaction is to enhance the identification of genes in genomic loci or disease loci. Further, the data integration methods also help in knowing whether given loci are jointly associated with the disease. This association may be due to the co-alteration in transcript/protein levels, or two closely linked loci are altered independently to affect the RNA/protein levels. The statistical procedures are used examine the joint probabilities of genotype association, RNA/protein expression, and clinical disease data. The entire data could be modeled to know whether they are related in a causal or reactive relationship.

References

Barabási A-L, Gulbahce N, Loscalzo J (2012) Network medicine: a network-based approach to human disease. Nat Rev 8(4):286–295

Schadt EE, Friend SH, Shaywitz DA (2009) A network view of disease and compound screening. Nat Rev Drug Disc 8(4):286–295

Diseasome

► Biological Disease Mechanism Networks

Disposition

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

Synonyms

Capacity; Potentiality; Power; Tendency

Definition

Dispositions are properties that need enabling conditions for being manifest.

Characteristics

Dispositional Versus Categorical Properties

Properties of objects or systems are usually distinguished into dispositional properties and categorical properties. Everyday paradigm cases of dispositional properties are, for instance, courage and fragility; paradigm cases of categorical properties are shape and structure of an object or a system. Whether it is possible to explicitly define individual dispositions (and to provide a general definitional scheme for "disposition" as a generic term) is a major dispute in the debate about dispositions (see section "Conditional Analysis").

In the case of dispositional properties, it is important to distinguish between an object or system having a property on the one hand and manifesting (► manifestation) the property under ► enabling conditions on the other hand. A person may be a courageous person all his life, but has only few occasions to show courageous behavior. Similarly, a glass may be fragile but this disposition will become manifest (i.e., the glass will break) only if given certain enabling or stimulus conditions obtain (e.g., striking of the glass). On the contrary, if objects or systems possess categorical properties (e.g., the roundness of a billiard ball) they will be manifest unconditionally. Thus, concepts for categorical properties do not entail the distinction of having the property and manifesting it.

In the biological sciences many examples of dispositional properties can be found. These examples include: the capacity of amino acid chains to fold into a specific three-dimensional structure, the capacity of genes to become activated, the ability of muscle fibers to contract, the pluripotency and totipotency of cells, the fitness (capacity to reproduce successfully and survive) of organisms, the ► evolvability/adaptability of populations, and the sustainability of ecosystems.

The Importance of Dispositions in Science

Dispositions have been controversial since early modern times because they were conceived of as

hidden causes (▶ causality) that bring about effects (i.e., their manifestations). Moliére in his *Le Malade imaginaire* ridicules explanations (▶ Explanation in Biology) in terms of dispositions by pointing out that one might explain why opium puts people to sleep by appealing to its "dormitive virtue." However, it is explanatory empty to refer to hidden causes that are epistemically accessible only via a single effect.

Since the 1930s (cf. Carnap 1936) it became apparent that dispositional concepts do play an important role in science and furthermore interest in dispositions as an analytical tool for characterizing science has resurged considerably in the last two decades (e.g., Mellor 2000; Choi and Fara 2012). The concept of a disposition is an important tool in the analysis of science because it points to the fact that the properties/behavior of systems may only be manifest given certain enabling or stimulus conditions, for example, contextual factors. This is particularly true in the biological sciences. We attribute many properties to biological systems (e.g., the ability of muscle fibers to contract) that become manifest only given the presence of specific enabling conditions (e.g., the presence of ATP and an appropriate stimulus).

Conditional Analysis

Certain aspects of dispositions have been debated (see Mumford 1998 for a comprehensive overview). We will discuss some of these issues in order to clarify what is implied by the attribution of a disposition to an object or a system.

First, what are the conditions under which we can legitimately attribute a disposition D to a system s? To give a precise answer to this question requires specifying the relation between having a disposition and manifesting it. One major issue in the debate about dispositions is whether this connection can be made more precise – whether particular dispositions can be defined explicitly in terms of their manifestations and enabling conditions.

The starting point for such attempts is the so-called ▶ simple conditional analysis (SCA). 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, the necessary and sufficient conditions for s having D can be symbolized as follows:

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

which is to be read as: s has Disposition D if and only if: if s were confronted with E, then s would necessarily manifest M. Thus, given SCA and given the knowledge regarding how to test the counterfactual claim "Es → Ms," we know under which conditions we can legitimately attribute D to s.

One problem with the SCA is that manifestations cannot easily be specified. What exactly are the manifestations of being courageous or of fragility (cf. Prior 1985, 6–10)? Likewise, it is difficult to spell out the exact enabling conditions for a disposition (e.g., breaking, hitting, and throwing in particular ways). This is even truer for biological dispositions because the way in which the context is involved in the manifestation is diverse and complicated, and the enabling conditions are very complex.

Another significant problem for the simple conditional analysis is a family of counterexamples that shows that the right hand side of SCA (Es → Ms) is neither necessary nor sufficient for the left hand side (Ds). There are various such counterexamples discussed under the headings of "antidotes," "finks," "masks," etc. For example, if we understand "fatally poisonous" as "disposed to kill if ingested," someone might take the poison but, nevertheless, survive because of some antidote that has been ingested as well (Bird 2007, 27). In such a case, the substance is fatally poisonous, but the manifestation does not take place even though the enabling conditions (ingestion) did occur. *A fortiori* the right hand side of SCA is not a necessary condition for the left hand side. There are possible interferences, which invalidate SCA. Thus, the manifestation of a disposition requires not only enabling conditions but also the absence of interfering factors. Only if all of these conditions can be listed explicitly, the SCA would provide an explicit definition of a dispositional concept. It is, however, a controversial issue whether it is even in principle possible to list all relevant factors. Take the example of the differentiability of cells. The process of manifestation, that is, the differentiation of a cell into a specific cell type is a very complex and temporally

extended process, which requires that many genes are correctly activated or repressed, that plenty of proteins are properly synthesized and interact in the right way with each other. According to the ▶ complexity of the differentiation process, numerous factors could disturb this process and prevent the manifestation. It is hardly imaginable that one could (even in principle) prepare a complete list of all possible interfering factors.

Intrinsicality

A second instructive debate concerns the ▶ intrinsicality of dispositions. Roughly speaking, a property is intrinsic if a system possesses the property independently of what is going on in its context. Shape is an intrinsic property, whereas being smaller than everybody else in the room is an extrinsic (relational) property.

The rationale for attributing a disposition to a particular system seems to imply that dispositions are intrinsic. The rationale is as follows: The phenomenon of sugar dissolving in water is, strictly speaking, a property of a combined system – sugar plus water. If we describe the phenomenon in terms of a disposition being manifest rather than in terms of a property of a compound system, we usually introduce a distinction between a system (e.g., sugar), which is endowed with a disposition, and external, for example, contextual conditions. If we ascribe solubility to sugar, then we focus on those conditions for obtaining of the phenomena that are due to sugar only. The disposition (solubility) comprises exactly those conditions of the phenomenon that the system (sugar) possesses independently of what is going on in the context. Thus, even though the manifestation of dispositions (e.g., the dissolving in the case of solubility) depends on extrinsic factors, it is usually held that the *disposition itself* (e.g., the solubility of salt) is intrinsic. But intrinsicality may not be a necessary feature of dispositions (cf. McKitrick 2003; Choi and Fara 2012). The challenge is particularly clear in the case of some biological systems: The importance of the context undermines the claim that all dispositional properties are intrinsic. As Alan Love (2003) has pointed out for the example of the ▶ evolvability of populations (▶ adaptation), in many cases external factors are not only the enabling conditions for biological dispositions. Rather, they determine jointly with intrinsic factors the very nature of the disposition as well as its causal efficacy. For example, whether a population is evolvable or not is not independent of contextual factors like migratory abilities and landscape topography. Hence, the intrinsic character of the biological disposition "evolvability" is called into question.

Single-Track Versus Multi-Track Dispositions

Courage, it seems, is a disposition that will be manifested in different situations by different behaviors. It is a multi-track disposition, that is, one disposition with multiple possible manifestations. However, the SCA-tradition has often assumed that dispositions are individuated in terms of one set of enabling conditions and one manifestation (single-track dispositions). The drawback is a proliferation of dispositions, for example, different courage-dispositions – one for each kind of courageous behavior, for example, courage in the face of death and courage in the face of financial stress.

In biology, there are many possible candidates for multi-track dispositions: the manifestation of evolvability for a population can result in different changes of gene frequency of a population; the pluripotency of stem cells can become manifest in muscle cells, bone cells, etc. But on closer inspection it becomes apparent that the characterization of these dispositions as "multi-track" depends on a fine grained analysis of the manifestation states. If we raise the graininess of the analysis, just one and not multiple possible manifestation states can be identified. For example, the evolvability of a population will be manifest if its gene frequency has changed independent of the kind of gene whose frequency has changed and independent of the exact dimension of the change.

Reduction

A further frequently disputed question concerns the issue of ▶ reduction. Dispositions, such as fragility, are necessary conditions for the obtaining of the manifestation (provided the simple conditional analysis or something akin is correct). This is often analyzed as: Fragility is causally efficacious (▶ causality) in bringing about the manifestation. An ensuing question that has been widely discussed is whether a disposition can be considered causally efficacious on its own or whether it is causally efficacious in virtue of an underlying causal basis, such as molecular structure.

It is important to distinguish two issues in this debate. First, fragility and other every-day dispositions

are *macroscopic* properties. We tend to assume that macroscopic properties of systems can be reduced to their molecular structure. A glass, for instance, is fragile in virtue of its molecular structure. This, however, is true for dispositional and categorical properties alike. The glass has its shape (a categorical property) in virtue of its molecular structure (and/or arrangement) as well. So this is not a special issue for dispositions.

A second, different, issue is whether there can be bare dispositions or whether every dispositional property needs to be reduced to categorical properties, such as the microstructural configuration. The question is whether there might be irreducible dispositional properties that cannot be identified with a set of categorical (e.g., microstructural) properties. The physical property "charge" or other fundamental dispositions might be candidates for bare dispositions because there are no microstructural properties that they might be identified with.

Cross-References

- ▶ Adaptation
- ▶ Causality
- ▶ Complex System
- ▶ Complexity
- ▶ Enabling Conditions
- ▶ Evolvability
- ▶ Explanation in Biology
- ▶ Intrinsicality
- ▶ Manifestation
- ▶ Reduction
- ▶ Simple Conditional Analysis (SCA)

References

Armstrong DM, Martin CB, Place UT (1996) Dispositions – a debate. Routledge, London
Bird A (2007) Nature's metaphysics: laws and properties. Oxford University Press, Oxford
Carnap R (1936) Testability and meaning. Phil Sci 3:420–471
Choi S, Fara M (2012) Dispositions. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (Spring 2012 Edition), http://plato.stanford.edu/archives/spr2012/entries/dispositions/
Love A (2003) Evolvability, dispositions, and intrinsicality. Phil Sci 70:1015–1027
McKitrick J (2003) A case for extrinsic dispositions. Aust J Phil 81:155–174
Mellor DH (2000) The semantics and ontology of dispositions. Mind 109:757–780
Mumford S (1998) Dispositions. Oxford University Press, Oxford
Prior E (1985) Dispositions. Aberdeen University Press, Aberdeen

Distance Field

- ▶ Distance Transform
- ▶ Distance Transform and Travel Depth

Distance Map

- ▶ Distance Transform
- ▶ Distance Transform and Travel Depth

Distance Transform

Virginio Cantoni[1,2], Riccardo Gatti[1] and Luca Lombardi[1]
[1]Department of Computer Engineering and Systems Science, University of Pavia, Pavia, Italy
[2]Computational Biology, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Distance field; Distance map

Definition

The Distance Transform (DT) is an operator usually applied into the domain of 2D binary image (where each point is classified as foreground or background) but it can be extended to 3D domains too. The result of the transform is a new image whose foreground pixels are labeled with a value that represents the minimum distance from the background.

There are many different types of DT which differ mainly on the type of metric used to evaluate the

Distance Transform,

0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	2	2	1	1	0	0
0	0	1	2	3	3	2	1	0	0
0	1	1	2	3	3	3	2	1	0
0	0	0	1	2	2	2	2	1	0
0	0	0	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	2	1	1	1	0	0
0	0	1	1	2	2	2	1	0	0
0	1	1	1	2	3	2	1	1	0
0	0	0	1	2	2	2	1	1	0
0	0	0	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	0	0

Fig. 1 Distance transform examples with city block and chessboard metrics

distance. Most commonly used ▶ digital metrics are "chessboard" metric, "Euclidean" metric, or "city block" metric (Fig. 1).

The simplest method to obtain a DT is to apply a sequence of erosion operations from mathematical morphology with a proper structuring element defined through the chosen metric. This sequence of operations must be performed until all foreground pixels are covered.

The DT is applied in skeletonization, shape description, and symmetry evaluation processes.

Cross-References

▶ Distance Transform and Travel Depth

Distance Transform and Travel Depth

Virginio Cantoni[1,2], Riccardo Gatti[1] and Luca Lombardi[1]
[1]Department of Computer Engineering and Systems Science, University of Pavia, Pavia, Italy
[2]Computational Biology, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Distance field; Distance map

Definition

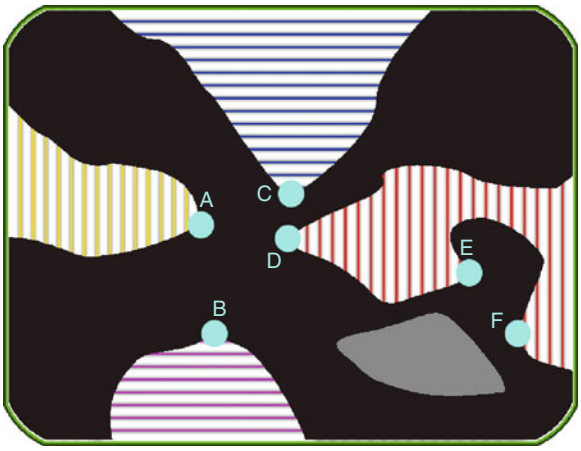
For the pockets analysis in proteomics, the minimum distance of a point from a reference surface, that is the ▶ convex hull (or a related surface), is called travel depth. For the evaluation of this feature, an effective tool is given by the ▶ distance transform.

Characteristics

In biology, the travel depth parameter has been introduced by Coleman and Sharp (Coleman and Sharp 2006) (Coleman and Sharp 2010). The travel depth is the value of the distance associated to each point of a pocket surface from the ▶ convex hull (a 2D example is shown in Fig. 1). The shape and the properties of the molecular surface determine what interactions are possible with ligands or other macromolecules. In particular, the active sites are generally described as shallow and deep spots on the molecular volume. Experimentations have shown that active sites usually correspond with areas on the surface having a high travel depth value and thus they coincide with the bottom of protein's pockets.

How to Obtain Travel Depth for a Given Protein 3D Structure

The first step is to define the reference protein's volume inside a cubic grid of voxels (Cantoni et al. 2010). This reference volume could be, for example, the van der Walls volume, the solvent excluded surface (SES), or the solvent accessible surface (SAS). The second step is to build the convex hull of the molecule's volume that will define the boundaries in the 3D space in which to apply the distance transform algorithm. The convex hull of a molecule is the smallest convex polyhedron that contains the molecule voxels. In R^3 the convex hull is constituted by a set of facets, usually triangles and a set of ridges (boundary elements) that are edges. Each triangle that belongs to the convex hull must then be inside the 3D grid and



Distance Transform and Travel Depth, Fig. 1 A 2D example. A, B, C, and D are points with high travel depth value and possible active sites candidates

also all the voxels belonging to the volume of the convex hull must be labeled inside the grid. The region on which the distance transform is applied is called concavity volume and is obtained by:

$$R = CH \cap \overline{RV} \tag{1}$$

where R is the concavity volume, CH the convex hull, and RV the reference molecular volume (e.g., the SES). Within the region R the following propagation is applied:

$$D_i = \left\{ \begin{array}{l} 1 \ if \ i \in B_{CH} \\ 0 \ otherwise \end{array} \right\}$$

$A = B_{CH}$;
$N = (A \oplus K) \cap R$;
$E = N - A$;
while $E \neq \emptyset$ *do*
  $\forall e \in E : d_e = \min_{n \in n_e} (d_n + w_n)$;
  $A = N$;
  $N = (A \oplus K) \cap R$;
  $E = N - A$;
done

Where:

- A represents the increasing set of voxels contained in R; E corresponds to the recruited set of near neighbors of A contained in R (i.e., the voxels reached by the last propagation step).
- B_{CH} represents the surface of the convex hull.
- $min_{n \in ne}(d_n + w_n)$ represents the minimum value among the distances d_e in the near neighbors belonging to D already defined, incremented by the displacement w_j between the locations (e, n): that is, if e and n have a common face $w_n = 1$; if e and n have a common edge $w_n = \sqrt{2}$; if e and n have a common vertex $w_n = \sqrt{3}$. At each iteration, new voxels, inside R, are reached by the propagation process and the value they take is determined by the neighbor distance (from the convex hull) and the voxels distance from the neighbor involved; this in order to simulate an isotropic propagation process and the proper distance evaluation.
- $E = \emptyset$ corresponds to the regime condition: no other changes are given and the connected component of R, adjacent to the border B_{CH}, is completely covered.

The travel depth represents the distance of each voxel of A from B_{CH}.

Cross-References

▶ Convex Hull
▶ Distance Transform

References

Cantoni V, Gatti R, Lombardi L (2010) Segmentation of SES for protein structure analysis. In: Proceedings of bioinformatics 2010, Valencia, INSTICC Press: Setubal, PT, pp 83–89
Coleman RG, Sharp KA (2006) Travel depth, a new shape descriptor for macromolecules: application to ligand binding. J Molecular Biology 362:441–458
Coleman RG, Sharp KA (2010) Protein pockets: inventory, shape and comparison. J Chem Inf Model 50:589–603

Distributed Data Access

Eugenio Cesario
ICAR-CNR (Istituto di Calcolo e Reti ad Alte Prestazioni), Rende, CS, Italy

Synonyms

Data collection (integration) from distributed sources

Definition

Distributed Data Access refers to the search and exploration of data stored in distributed data repositories, as well as the gathering of data from various sources to find interesting and useful information and answer a specific scientific question.

Characteristics

Biological Data Issues

Biological data (protein structure and function, DNA sequences, and metabolic pathways) can concern many life science domains, such as genetics, structural biology, microarrays, pharmacology, etc. They are characterized by two main features: *heterogeneity* and *huge volume*.

Heterogeneity

Biological repositories are highly diverse in both variety and granularity. For example, biological data types can vary from images and drawings to graph structures, from unstructured text to data tables, from sequence to three-dimensional protein structures, etc. Moreover, each researcher can focus on different levels of biological problems. This makes the repository store data with different granularities, such as, from genomic and protein sequences to protein activities, from cell structure to two-dimensional or three-dimensional structured data of huge molecules (Tao 2006). Table 1 shows a list of the biological data types and a related schematic description of their content (BioInfoBank Library 2010).

Huge Volume

Due to the advances of the research activity in the life science, huge amounts of data are generated by researchers all over the world. For example, an organization that generates data, that is, a sequencing laboratory, starts by processing raw data (collected by sample tracking), followed by analytical processing to translate the signal to measurements, and finally to obtain biological data (such as sequence tags or abundance of gene or proteins). Daily production rates are in the order of tens of gigabytes (Topaloglou et al. 2004). For example, from 1996 to 2010 the GenBank (the NIH genetic sequence database populated by an annotated collection of all publicly available DNA sequences) increased the number of entries with an exponential rate, that is, from 1 million sequences to 49 million (GenBank 2011).

Data Access and Integration

Biological information is highly interconnected and often context-dependent. In practice, genomic data and its associated information generated by experimental or computational methods is stored in hundreds of independent, overlapping, and heterogeneous data resources geographically distributed. They are stored in a variety of formats, ranging from unstructured data (i.e., textual data) to strongly structured database data, depending on its content (Baralis and Fiori 2008). Moreover, there are millions of articles composing the scientific research literature, most of them accessible on the Web.

Biological scientists often require the execution of "cross-queries," that is, the discovery of information from different repository locations and the merging of results retrieved by various datasets (Haider et al. 2009). For such a reason, a typical data-integration problem is the gathering of data from various sources to find relevant information and answer a specific scientific question. To do that, the simple solution of moving all these data into a central location for integrated querying with other resources is unfeasible, due

Distributed Data Access, Table 1 Types of biological data

Experimental data	Data revealed from direct laboratory experiments (observations, digital images, notes, etc.)
Raw data	Data which have never been a subject of manipulation or processing
Sequence data	Data containing protein sequences or obtained from a DNA sequencing process
Structure data	Data modeling three-dimensional protein structures, DNA, RNA, or small molecules
Phylogenetic data	Data about evolutionary relations among various groups of organisms (information is revealed through molecular sequencing data and morphological data matrices)
Metabolic data	Data containing metabolic pathways (enzymatic reactions in living organisms) and systems biology information

to physical transfer challenges (too long transfer time) and privacy-preserving issues. On the contrary, the most common approach currently exploited consists in maintaining the information stored in geographically distributed databases whereby individual data providers are responsible for updates and release cycles (Smedley et al. 2008). The query response is obtained by collecting and merging the information obtained from various data sources and return a final result to the user. Lastly, the results to be returned should be in standard formats and where possible, semantically annotated to ensure interoperability with other databases and tools (Haider et al. 2009). In the following we describe two systems that have been developed for biological data integration.

The Distributed Annotation System (DAS) (Dowell et al. 2001) is a widely adopted protocol for dynamically integrating a wide range of biological data from geographically diverse sources. Its applicability is growing and evolving in response to new challenges facing integrative bioinformatics. An extended version of the DAS specification (version 1.53E) incorporates several recent developments, including its extension to serve new data types and an ontology for protein features (Jenkinson et al. 2008). Data distribution, performed by DAS servers, is separated from visualization, which is done by DAS clients that integrate information from multiple servers. It allows a single machine to gather up sequence annotation information from multiple distant Web sites, collate the information, and display it to the user in a single view. The DAS specification has several client implementations (Jenkinson et al. 2008): the Ensembl genome browser (to display data from a wide variety of genomic, gene, and protein sequence coordinate systems), SPICE (to combine protein sequence and structural annotations), the DASMIweb portal (to integrate protein-protein and domain-domain interaction datasets), and iPfam (to compare the interaction topologies of different sources by overlaying them in a node graph).

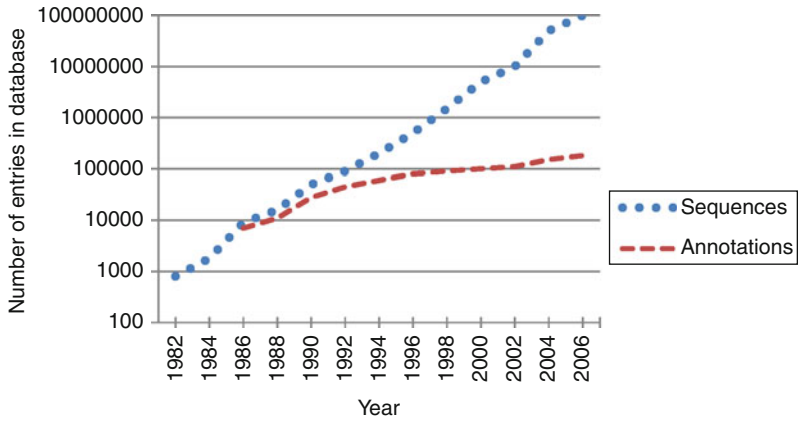
Recently the *BioMart Central Portal*, a system offering access to a wide set of biological datasets, has been implemented (Haider et al. 2009). It is a Web server interface of BioMart software (Smedley et al. 2009) and provides a unified view over disparate data sources that enable bioscientists to retrieve data from one or multiple sources in a simple and efficient way. It provides access to a variety of datasets that can be queried independently or in a federated way enabling users to ask complex questions over data sources that may be located at different geographical locations. It is used to access many of the large biological datasets in the public domain, such as dbSNP Ensembl genomic, Uniprot protein, Reactome pathway, HGNC gene name, Wormbase genomic, and PRIDE proteomic data (a complete list is available at http://www.biomart.org/biomart/martview/). As of March 2009, BioMart Central Portal brings together an extensive range of databases serving more than 100 datasets with an average monthly usage of over one million server hits (Haider et al. 2009).

Data Analysis

The number of available complete genomic sequences is doubling almost every 12 months (Meyer 2006), whereas according to Moore's law, available compute cycles (i.e., computational power) double every 18 months. That is, biological data to be processed are growing more quickly than computational and technological instruments. Additionally, since the analysis of genomic sequences requires binary comparisons of the genes involved in it, the computational overhead is very high. The impact of such issues is plotted in Fig. 1 (Meyer 2006), which contrasts the number of genetic sequences obtained with the number of annotations generated. The figure shows that the knowledge (annotations, models, patterns) has a sublinear rate with respect to the available data sequences which they are extracted from.

To handle this abundance in data availability (whose rate of production often far outstrips the capability of the scientists to analyze it), automatic data analysis techniques are used. In particular, the exploitation of Data Mining algorithms (Fayyad et al. 2006; Grossman et al. 2001) in science helps scientists in hypothesis formation and gives them a support on their scientific practices and solving environments, getting the benefits coming from knowledge that can be extracted from large data sources. Moreover, since data is large and is maintained over geographically distributed sites, the computational power of distributed systems is often exploited for knowledge discovery in scientific data. Distributed Data Mining algorithms are very suitable to such a purpose.

The Grid (Foster et al. 2003) is a privileged computing infrastructure to develop applications over geographically distributed sites. The Grid involves the integrated and collaborative use of remote computing



Distributed Data Access, Fig. 1 Using a logarithmic scale, the growth of sequence databases and annotations

power, storage, software, and data managed and shared by different organizations. This technology has shown to be very reliable in solving large-scale bioinformatics-related problems and improving the efficiency and effectiveness of the computation on biological data (Cesario and Talia 2010). In the last years, various international scientific projects have been developed (and are currently under development) on this field. The most important are Euro BioGrid (Eurogrid 2001), Asia Pacific BioGrid (APBiogrid 2001), UK BioGrid (UKBiogrid 2001), and North Carolina BioGrid (NCBiogrid 2001) showing that the Grid is a reliable and useful infrastructure for the management and analysis of distributed biological data.

Cross-References

- ▶ General-Purpose Computation, Graphics Processing Units
- ▶ Grid Computing, Parallelization Techniques

References

APBiogrid (2001) http://compaq.apbionet.org/grid/
Baralis E, Fiori A (2008) Exploring heterogeneous biological data sources. In: 19th international workshop on database and expert systems applications, Turin, Italy, pp 647–651
BioInfoBank Library (2010). http://lib.bioinfo.pl/courses/view/160
Cesario E, Talia D (2010) Using grids for exploiting the abundance of data in science. Scalable Comput Pract Exp 11(3):251–262
Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L (2001) The distributed annotation system. BMC Bioinformatics 2:7
Eurogrid (2001) http://www.eurogrid.org/
Fayyad U, Haussler D, Stolorz P (2006) Mining scientific data. Commun ACM 39(11):51–57
Foster I, Kescbrselman C, Nick J, Tuecke S (2003) The physiology of the grid. In: Berman F, Fox G, Hey A (eds) Grid computing: making the global infrastructure a reality. Wiley, New York, pp 217–249
GenBank (2011) http://www.ncbi.nlm.nih.gov/genbank/
Grossman RL, Kamath C, Kegelmeyer P, Kumar V, Namburu RR (2001) Data mining for scientific and engineering applications. Kluwer Academic, Dordrecht/Boston
Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A (2009) BioMart Central Portal – unified access to biological data. Nucleic Acids Res 37:23–27
Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, Hermjakob H, Hubbard TJP, Jimenez RC, Jones P, Kahari A, Kulesha E, Macías JR, Reeves GA, Prlic A (2008) Integrating biological data – the distributed annotation system. BMC Bioinformatics 9:S3
Meyer F (2006) Genome sequencing vs. Moore's law: cyber challenges for the next decade. Trends and tools in bioinformatics and computational biology. CTWatch Quart 2(3) http://www.ctwatch.org/quarterly/articles/2006/08/genome-sequencing-vs-moores-law/
NCBiogrid (2001) http://www.ncbiogrid.org/
Smedley D, Swertz MA, Wolstencroft K, Proctor G, Zouberakis M, Bard J, Hancock JM, Schofield P (2008) Solutions for data integration in functional genomics: a critical assessment and case study. Brief Bioinform 9(6):532–544
Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A (2009) BioMart – biological queries made easy. BMC Genomics 10:22
Tao C (2006) Toward making online biological data machine understandable. In: Proceedings of the 5th international semantic web conference (ISWC 2006) doctoral consortium, Athens, GA, November 2006, pp 992–993

Topaloglou T, Davidson SB, Jagadish HV, Markowitz VM, Steeg EW, Tyers M (2004) Biological data management: research, practice and opportunities. In: Proceedings of the thirtieth international conference on very large data bases, Toronto, pp 1233–1236
UKBiogrid (2001) http://www.mygrid.org.uk/

Distributed Data Management

Pietro Hiram Guzzi[1], Giuseppe Tradigo[2] and Pierangelo Veltri[2]
[1]Department of Experimental Medicine and Clinic, University Magna Gratia of Catanzaro, Catanzaro, Italy
[2]Department of Medical and Surgical Sciences, University Magna Græcia of Catanzaro, Catanzaro, Italy

Synonyms

Data Management: Data processing; Data storing and querying

Definition

Database

A database is a set of information semantically organized and correlated such that it can be used to guide processes or opportunely combined to form knowledge. Databases can be reconducted to database management systems that *include* all procedures necessary to organize, store, index, and query data and to keep data always consistent, that is, always valid.

Distributed Processing

The distributed processing means solving an input problem by means of subprocesses evaluated on different processing unit. Processing unit may be located on single computational machine or can be dislocated on different machine wired and located in the same building or geographically distributed.

Data Management

Organizing information in data containers for storing, saving, and maintaining them allowing querying and information retrieval.

Distributed Data

Information organized in a semantically related way, also divided in pieces of information with rules to be able in reconstructing them. Pieces of information may represent data duplication (minimizing data loss risks) or partition (saving space).

Characteristics

Availability of data in digital way with an increasing data precision and quality is increasing the need of both support and spaces for data storing as well as procedures and structures for data exchanging. In such a scenario, the term *Distributed Data Management* refers to a set of methodologies, architectures, and tools enabling the efficient management of data stored in geographically distributed databases aiming both to reduce access time and to allow efficient knowledge extraction. In research contexts, distributed data management refers to a set of technologies enabling the realization of a distributed laboratory in which different research unit collaborate performing different experiments on the same project.

Distribution may help in saving spaces and improve data availability and replication. It is the case of biological data management where data produced are huge and information extraction often is a hard task. For example, data *produced* by *many* experimental platforms used in the biological field have been largely used in many studies to identify molecules that may be related to human diseases (Cannataro et al. 2010). Computational intensive applications for data manipulation require distributed processing, as in biological applications, where, thanks to always more accurate techniques to manipulate or to simulate information obtained from molecules are available (Cannataro et al. 2010), a typical study involves large number of samples and huge amount of data. Data distribution allows retrieving information in similar way as in centralized data management structure, allowing scalability in terms of data and users, where parallel data manipulation from different users allows to improve knowledge of the database.

Main requirements of distributed data storing with several nodes each with part of database and part of local data are as follows:

- The introduction of a commonly shared data model able to capture both raw data of the experiment and

related metadata; currently existing approaches are often based on XML-based languages for the representation of data and metadata, for example, all the languages developed by the HUPO-PSI initiative (www.hupo-psi.org).
- The definition of a uniform and widely accepted access and manipulation strategy for such large datasets enabling the sharing of information coming from single experiment on a specified laboratory has to be shared and validated in a distributed laboratory environment.
- The definition of a high performance data transfer strategy.
- The definition of a set of rules and prescription guaranteeing data privacy.

The advantages of distribution for data management has been considered in previous studies (see for instance Valduriez (1993)), and it has been becoming always more and more required for solving problems where computational time requires several nodes and input data is very large (it is the case for instance of proteins interactions simulation or protein structure prediction Branden and Tooze (1999)). As an application example, consider a distributed laboratory as a framework environment composed of several nodes, each one associated to a laboratory running its local database. The framework efficiently supports distributed storage and manipulation of experimental data. Each node contains an application programming interface mounted on a data storage system hosting data produced by the laboratory. Scientists may cooperate working each in his/her own laboratory each one running tens of experiments on different available biological data sets (as for instance for mass spectrometers as in Veltri (2008)). Such a configuration can potentially lead to terabytes of data produced each week or even each day. Pushing these considerations to their limit data processing (e.g., reduction of noise and allowing data comparable in terms of instrument accuracy) can easily scale-up minimal computational requirements. Algorithms for data manipulation and information extraction, that often use access to external databases, in a such scaled-up context become heavy time-consuming tasks, whereas in a distributed data management environment allow the cooperation and problems tractable.

Cross-References

- ▶ General-Purpose Computation, Graphics Processing Units
- ▶ Grid Computing, Parallelization Techniques

References

Branden C, Tooze J (1999) Introduction to protein structure, 2nd edn. Garland Science, New York. ISBN 0815323050

Cannataro M, Guzzi PH, Veltri P (2010) Protein-to-protein interactions: technologies, databases, and algorithms. ACM Comput Surv 43(1):1

Valduriez P (1993) Parallel database systems: open problems and new issues. Distrib Parallel Databases 1(2):137–165. doi:10.1007/BF01264049, ISSN 0926–8782

Veltri P (2008) Algorithms and tools for analysis and management of mass spectrometry data. Brief Bioinform 9(2): 144–155

Distributed Query Optimization

▶ Distributed Query Processing

Distributed Query Processing

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

Synonyms

Distributed Query Optimization; Distributed Querying

Definition

A distributed query is a kind of database query that interrogates multiple databases. These can be colocated (typically for performance), or distributed

across geographically separate locations (typically used for data integration). Each component database usually holds only a part of the integrated whole. The source query is translated into several individual queries, which are executed on the separate databases. The results of each query are then reassembled to give the required result.

Cross-References

► Data Integration and Visualization

Distributed Querying

► Distributed Query Processing

Distributed Revision Control

► Distributed Version Control System (DVCS)

Distributed Version Control System (DVCS)

Catherine M. Lloyd
Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand

Synonyms

Decentralized Version Control; Distributed revision control; Revision control; Source control; Version control

Definition

Version control is the management of changes to documents, programs, and other information stored as computer files. It is most commonly used in projects where a team of people may be working on the same files concurrently. In contrast to a centralized version control system, in a distributed version control system there is no single central repository. In the case of the CellML model repository, this allows modelers to be able to work independent of the online CellML model repository, and share their changes directly with each other until they decide the model is ready to be uploaded into the repository.

The version control software employed by the CellML model repository is Mecurial (http://mercurial.selenic.com/).

Cross-References

► CellML Model Repository

References

Mecurial http://mercurial.selenic.com/

Distribution-Free Tests

► Hypothesis Testing, Parametric vs Nonparametric

Disturbance

► Perturbation

Diversity

► Diversity (D) Gene

Diversity (D) Gene

Marie-Paule Lefranc
Laboratoire d'ImmunoGénétique Moléculaire, Institut de Génétique Humaine UPR 1142, Université Montpellier 2, Montpellier, France

Synonyms

D gene; Diversity; Diversity gene

Definition

The diversity (D) gene, or "*diversity*" is a ▶ leafconcept of the "▶ *GeneType*" concept of identification (generated from the ▶ IDENTIFICATION Axiom) of ▶ IMGT-ONTOLOGY, the global reference in ▶ immunogenetics and ▶ immunoinformatics (Giudicelli and Lefranc 1999; Lefranc et al. 2004, 2005, 2008; Duroux et al. 2008), built by IMGT®, the international ImMunoGeneTics information system® (http://www.imgt.org) (▶ IMGT® Information System). "*Diversity*" identifies a gene that rearranges at the DNA level and codes the diversity region of the variable domain of an immunoglobulin (IG) or antibody or of a T cell receptor (TR) chain (▶ Chain Type).

It is one of the four leafconcepts that are characteristics of the IG and TR loci (Lefranc and Lefranc 2001a, b), the other three being "variable" (V), "joining" (J), and "constant" (C) (▶ Variable (V) Gene, ▶ Joining (J) Gene, ▶ Constant (C) Gene).

The diversity (D) genes are observed in three loci: IG heavy (IGH), TR beta (TRB), and TR delta (TRD), where they participate to V-D-J rearrangements (Lefranc and Lefranc 2001a, b) (▶ Immunoglobulin Synthesis). An IG or TR diversity (D) gene has three possible and exclusive configurations (▶ Configuration Type): a germline configuration (before DNA rearrangement), a partially rearranged configuration (after D-J DNA rearrangement or, less frequently, V-D or D-D rearrangements), and a rearranged configuration (after a complete V-D-J DNA rearrangement, that eventually may involve several D). In the germline configuration, a diversity (D) gene possesses in 5′ (upstream) and in 3′ (downstream) a recombination signal (5′D-RS and 3′D-RS, respectively) (▶ Recombination Signal (RS)). These 5′D-RS and 3′D-RS are specifically recognized by the enzyme recombinase that, in the most usual chronology, allows first a D gene to be rearranged to a J gene, and then a V gene to be rearranged to the previously rearranged D-J gene. The sequence resulting from the V-D-J rearrangement encodes the V-DOMAIN (▶ Variable (V) Domain) of the IGH, TRB and TRD chains (▶ Chain Type) (Lefranc and Lefranc 2001a, b).

Cross-References

▶ Chain Type
▶ Configuration Type
▶ Constant (C) Gene
▶ Conventional Gene
▶ Gene Type
▶ IMGT® Information System
▶ IMGT-ONTOLOGY
▶ IMGT-ONTOLOGY, IDENTIFICATION Axiom
▶ IMGT-ONTOLOGY, Leafconcept
▶ Immunogenetics
▶ Immunoglobulin Synthesis
▶ Immunoinformatics
▶ Joining (J) Gene
▶ Recombination Signal (RS)
▶ Variable (V) Domain
▶ Variable (V) Gene

References

Duroux P, Kaas Q, Brochet X, Lane J, Ginestoux C, Lefranc M-P, Giudicelli V (2008) IMGT-Kaleidoscope, the Formal IMGT-ONTOLOGY paradigm. Biochimie 90:570–583

Giudicelli V, Lefranc M-P (1999) Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 12:1047–1054

Lefranc M-P, Lefranc G (2001a) The immunoglobulin FactsBook. Academic Press, London, pp 1–458

Lefranc M-P, Lefranc G (2001b) The T cell receptor FactsBook. Academic Press, London, pp 1–398

Lefranc M-P, Giudicelli V, Ginestoux C, Bose N, Folch G, Guiraudou D, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, Combres K, Girod D, Jeanjean S, Protat C, Yousfi Monod M, Duprat E, Kaas Q, Pommié C, Chaume D, Lefranc G (2004) IMGT-ONTOLOGY for immunogenetics and immunoinformatics. In Silico Biol 4:17–29

Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, Combres K, Ginestoux C, Giudicelli V, Chaume D,

Lefranc G (2005) IMGT-Choreography for immunogenetics and immunoinformatics. In Silico Biol 5:45–60
Lefranc M-P, Giudicelli V, Regnier L, Duroux P (2008) IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. Brief Bioinform 9:263–275

Diversity Gene

▶ Diversity (D) Gene

Dividing Cells Depletion

▶ Quantifying Lymphocyte Division, Methods

DNA Chip

▶ DNA Microarrays

DNA Content

▶ Lymphocyte Labeling, Cell Division Investigation
▶ Quantifying Lymphocyte Division, Methods

DNA Damage

Paolo Plevani
Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, Milan, Italy

Definition

DNA is modified by numerous chemico-physical agents causing a variety of lesions on the DNA molecule.

Cross-References

▶ Cell Cycle Checkpoints

DNA Damage Checkpoint

▶ Cell Cycle Checkpoints

DNA Damage Response

▶ Cell Cycle Arrest After DNA Damage

DNA Labeling

▶ Modeling, Cell Division and Proliferation

DNA Methylation

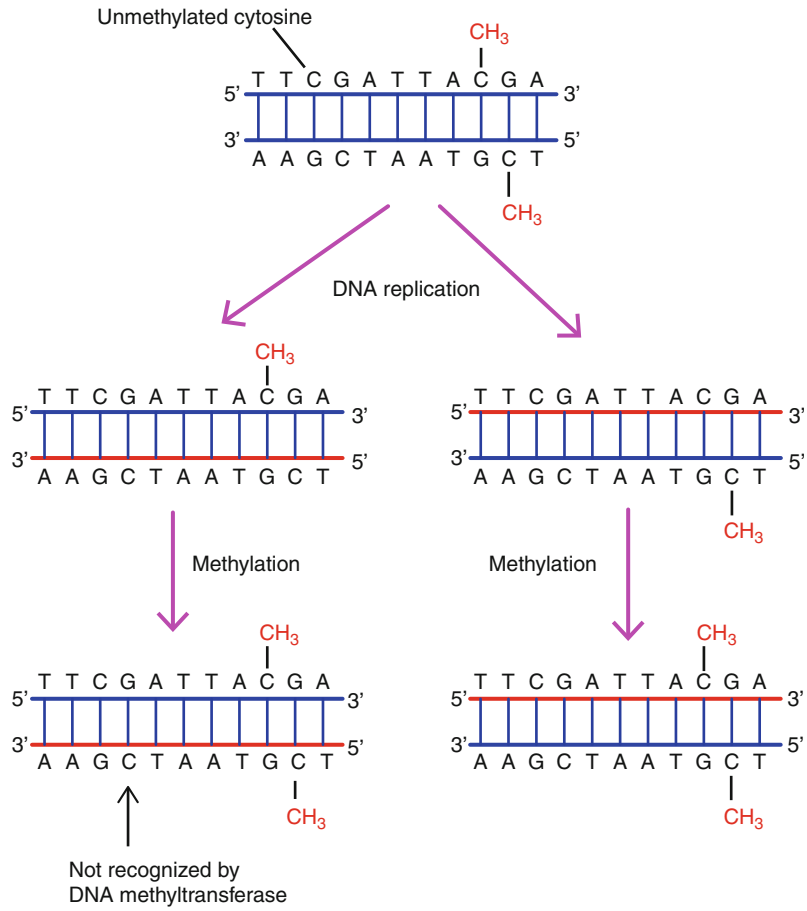
Yan Zhang
Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Definition

DNA methylation refers to the addition of a methyl group to the 5 position of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring in DNA strand. This modification can be inherited through cell division. The attachment of a methyl group to these nucleotides can serve many important biological purposes and is a crucial part of normal development and cellular differentiation in higher organisms.

Characteristics

The DNA in many different types of organisms can undergo DNA methylation, though it does not always necessarily serve the same function. In plants, for example, scientists believe that methylation occurs to deactivate genes that could otherwise cause harmful mutations. In fungi, DNA methylation is used to moderate and control the expression of certain genes based on the particular conditions affecting the fungus.



DNA Methylation, Fig. 1 Inheritance of the DNA methylation pattern. The DNA methyltransferase can methylate only the CG sequence paired with methylated CG. The CG sequence not paired with methylated CG will not be methylated

Methylation in mammals similarly moderates and inhibits the expression of certain genes; additionally, it is involved in the production of chromatin, a protein-DNA complex that makes up the structure of chromosomes.

DNA methylation stably alters the gene expression pattern in cells. DNA methylation is typically removed during zygote formation and reestablished through successive cell divisions during development. However, the latest research shows that hydroxylation of methyl group occurs rather than complete removal of methyl groups in zygote (Iqbal et al. 2011). Some methylation modifications that regulate gene expression are inheritable and are referred to as epigenetic regulation (Fig. 1).

In addition, DNA methylation suppresses the expression of viral genes and other deleterious elements that have been incorporated into the genome of the host over time. DNA methylation also forms the basis of chromatin structure, which enables cells to form the myriad characteristics necessary for multicellular life from a single immutable sequence of DNA. DNA methylation also plays a crucial role in the development of nearly all types of cancer (Jaenisch and Bird 2003).

DNA methylation at the 5 position of cytosine has the specific effect of reducing gene expression and has been found in every vertebrate examined. In adult somatic tissues, DNA methylation typically occurs in a CpG dinucleotide context; non-CpG methylation is prevalent in embryonic stem cells (Lister et al. 2009).

References

Iqbal K, Jin SG, Pfeifer GP, Szabó PE (2011) Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proc Natl Acad Sci USA 108(9):3642–3647

Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33(suppl (3s)):245–254

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462(7271):315–322

DNA Microarrays

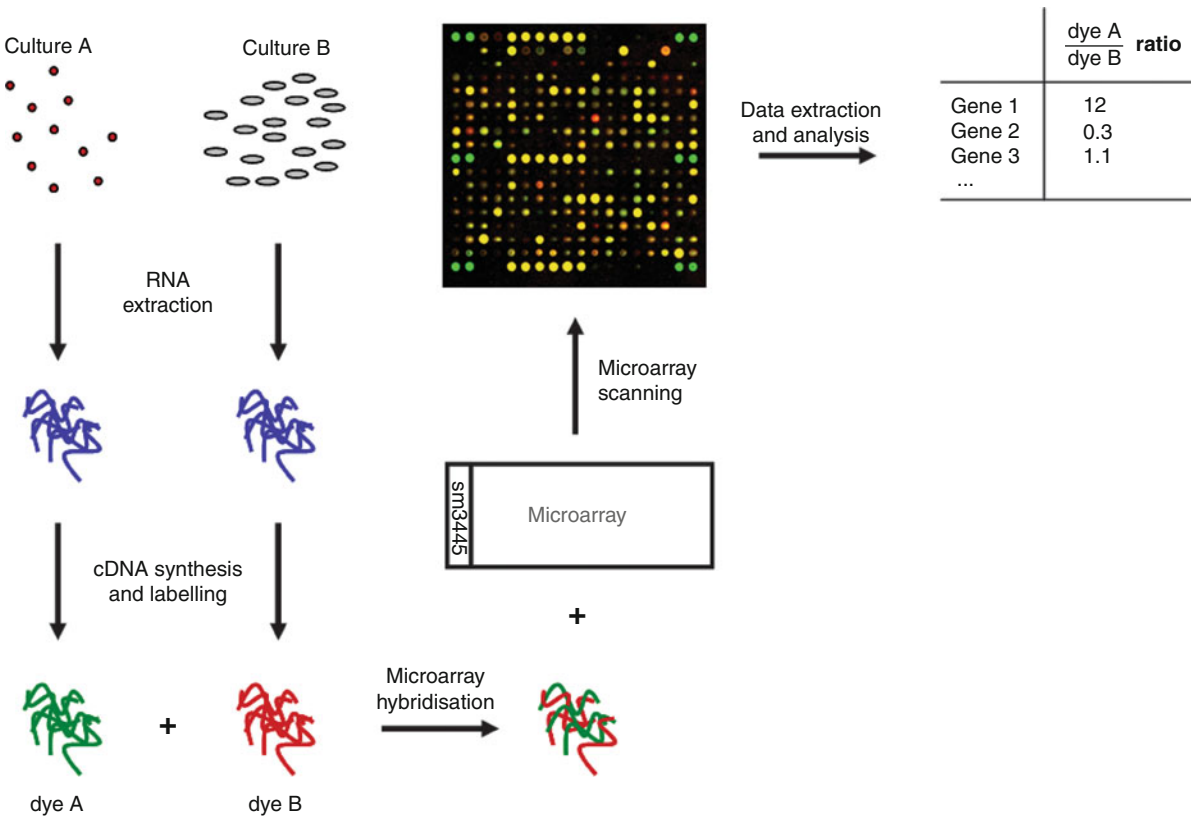
Jürg Bähler and Samuel Marguerat
Department of Genetics, Evolution & Environment and UCL Cancer Institute, University College London, London, UK

Synonyms

Arrays; cDNA microarrays; DNA chip; Microarrays; Oligo microarrays

Definition

This term refers to hybridization-based analytic platforms composed of hundreds to millions of DNA probes with defined sequences arrayed on a solid surface to measure, in parallel, nucleic acids present in a complex sample (Wheelan et al. 2008). The sample is first labeled with a fluorescent dye and then hybridized to the DNA microarray. After hybridization and washing, the fluorescence signal intensities from each probe are recorded using a laser scanner, providing a semiquantitative measure for the amount of nucleic acid molecules whose sequence is complementary to any given probe. One or two samples can be hybridized on a single microarray. When two samples are analyzed together, they are labeled with different dyes and the relative intensities of the two dyes are analyzed for each probe (two-color array; Fig. 1). When only one sample is analyzed, the absolute signals of the probes are used instead. Microarray probes are either PCR



DNA Microarrays, Fig. 1 Schematic representation of a two-color microarray experiment

products or shorter DNA primers (typically 25–60 nucleotides). Probes are either printed on a solid surface using a robot, or synthesized directly on the surface.

DNA microarrays provide a versatile platform to analyze RNA transcript levels and structures as well as genome structure (array CGH). When applied to the study of ► transcriptomes, DNA microarrays typically consist of probes directed against annotated gene features. However, a specialized type of microarray, called "tiling array," consists of probes with sequences tiled systematically across a genome. Tiling arrays permit the analysis of the transcriptional landscape of cells or tissues without being restricted by existing gene annotations.

Cross-References

► Cell Cycle Analysis, Expression Profiling

References

Wheelan SJ, Martínez Murillo F, Boeke JD (2008) The incredible shrinking world of DNA microarrays. Mol Biosyst 4:726–732

DNA Modification

► Post-Replication Modification

DNA Polymerases

Zoi Lygerou
School of Medicine, Laboratory of General Biology, University of Patras, Patras, Greece

Definition

DNA polymerases are enzymes that synthesize new DNA by moving along the template strand and synthesizing a new strand of complementary DNA sequence by nucleotide polymerization in the 5′ to 3′ direction.

Cross-References

► DNA Replication

DNA Repair

Paolo Plevani
Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, Milan, Italy

Definition

The various types of molecular processes repairing specific classes of lesions in the DNA.

Cross-References

► Cell Cycle Checkpoints

DNA Replication

Zoi Lygerou[1], K. K. Koutroumpas[2] and John Lygeros[2]
[1]School of Medicine, Laboratory of General Biology, University of Patras, Patras, Greece
[2]Automatic Control Laboratory, ETH Zurich, Zurich, Switzerland

Synonyms

DNA synthesis

Definition

DNA replication is the process of making an identical copy of the genetic material within each cell (Alberts et al. 2007; DePamphilis 2006). In eukaryotes, DNA replication takes place during a defined period of the ► cell cycle, called S (for synthesis) phase. DNA replication must be carried out with great precision every time the cell divides, so that genetic

information is preserved. Control mechanisms ensure that every base of the genome is replicated once and only once per cell cycle, thereby safeguarding genomic integrity.

Characteristics

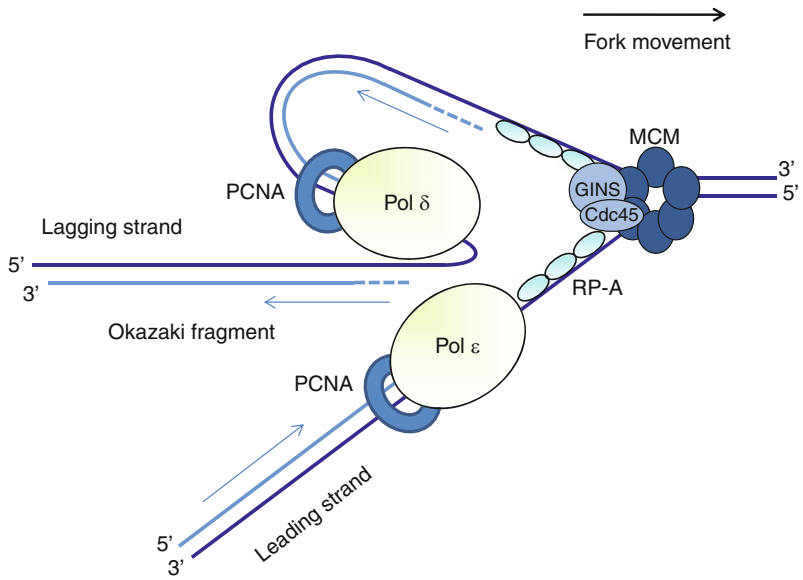
We present key characteristics of DNA replication in eukaryotic cells.

Replication Forks Move Continuously Along the Genome as Replisomes Catalyze DNA Synthesis

Each cell must accurately copy millions of bases of DNA (six billion base pairs in a human cell) before every cell division. DNA replication initiates from thousands of sites along eukaryotic chromosomes, called ▶ replication origins. Unwinding of the double-stranded DNA at replication origins (origin firing) creates a replication bubble consisting of two Y-shaped DNA structures called ▶ replication forks where DNA synthesis takes place (Fig. 1). Replication forks move outward in both directions from the origin as the DNA is replicated, eventually merging with a replication fork moving in the opposite direction (fork conversion). Since the two strands of the template DNA have opposite orientations (antiparallel) and DNA synthesis can only take place in the 5′ to 3′ direction, one of the DNA strands in a replication fork is replicated continuously (leading strand) while the other is replicated in a backstitch fashion in pieces of around 100–200 base pairs, called Okazaki fragments (lagging strand). Multisubunit protein complexes comprising over 100 proteins (▶ Replisome) carry out the steps required for DNA synthesis (DePamphilis 2006). A DNA helicase, made up from the hetero-hexameric MCM complex assisted by the tetrameric GINS complex and Cdc45 (CGM complex), unwinds the double-stranded template ahead of the replication fork, while replication protein A (RP-A) binds and stabilizes the single-stranded DNA exposed by the helicase. ▶ DNA Polymerase α-primase lays down RNA-DNA primers for replication and is then replaced (polymerase switching) by polymerase ε on the leading strand and polymerase δ on the lagging strand. The replication clamp PCNA (proliferating cell nuclear antigen), a homo-trimer loaded by the clamp loader RF-C (replication factor C) encircles the DNA, holds the polymerases in place, and choreographs the multiple transitions that take place at the replication fork. Okazaki fragments on the lagging strand are stitched together by the action of the endonuclease Fen1 and DNA ligase. A fork protection complex (consisting of timeless, tipin, claspin, and And1/ctf4) safeguards integrity of the fork when polymerases are forced to stall. Topoisomerases ensure that topological tension introduced by replication is relieved. Newly synthesized DNA is repackaged into chromatin by histone deposition complexes such as CAF-1, which is recruited to the replication fork by interactions with PCNA. The replisome copies the leading and lagging strand at the same time and replication forks move continuously along the genome, producing identical copies of the cell's genetic material.

DNA Replication in Eukaryotes Is Complex and Uncertain

Origin selection and activation is a crucial part of replication and various organisms have evolved different ways to define origins of replication (Gilbert 2004). In bacteria, there is a single, sequence-specific origin of replication and origin activation is deterministic: the origin fires in every cell cycle with high fidelity. At the other extreme, in early fly and frog embryos origin selection is a stochastic process. In *Xenopus* preblastula embryos, where replication must be completed fast, replication initiates apparently at random and at short intervals (8–15 kb) without discernible sequence specificity. Most eukaryotic cells seem to follow an intermediate route between a fully deterministic and a fully random origin selection mechanism. Replication initiates from relatively specific regions along the genome. In each cell cycle a fraction of these regions are activated, giving rise to a different distribution of initiation events along the genome at every S-phase. Moreover, the timing of firing of each origin of replication is not fully determined: though some origins tend to fire on average early and others late, generating a reproducible timing program in a population of cells, the time at which a given origin will fire may differ from cell to cell, giving rise to uncertainty also in the time domain. The process of DNA replication thus follows a unique pattern in each cell in a population. Every cell must therefore remember, at every point in time, which parts of its genome have been replicated, and should not be replicated a second time and which parts remain unreplicated. Such molecular memory is brought about by origin-bound multisubunit protein complexes.



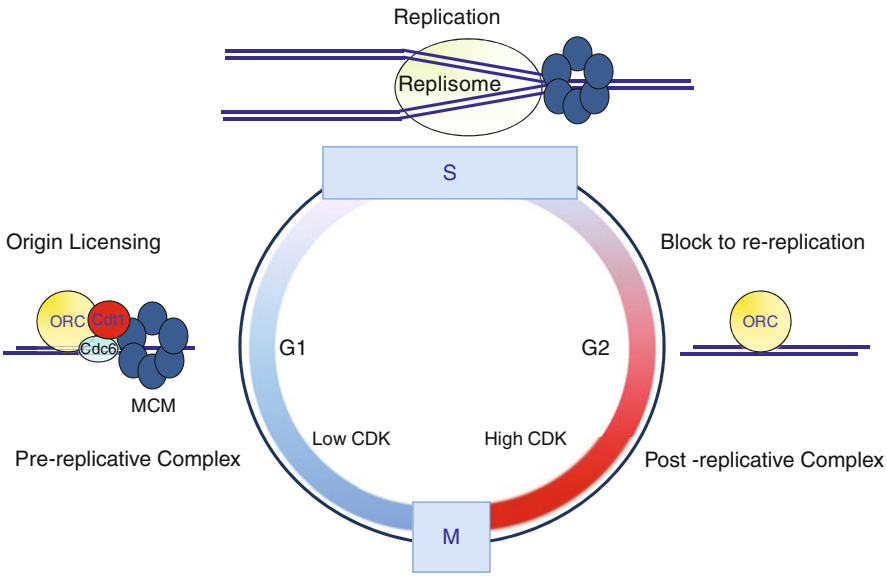
DNA Replication,
Fig. 1 The eukaryotic DNA replication fork

Dynamic Origin-Bound Complexes Safeguard Once per Cell Cycle Replication

Every part of the genome must be replicated once and only once per cell cycle. This is ensured by ordered transitions in protein complexes bound at origins of replication (Blow and Dutta 2005, Fig. 2). When mitosis is completed and cells move into a new G1 phase, all putative origins become licensed (▶ DNA Replication Licensing) for a round of replication by the assembly of ▶ prereplicative complexes onto origins, consisting of the six subunit origin recognition complex (ORC), the loading factors Cdc6/Cdc18 and Cdt1, and the six subunit MCM complex, which will later act as the replicative helicase. As cells move into S-phase, specific origins are activated by modifications and recruitment of additional factors (Cdc45, the GINS complex, Dpb11/TopBP1, Sld2/RecQL4, Sld3/Treslin, Mcm10, and others) which turn the pre-replicative complex into a pre-initiation complex, lead to activation of the helicase, recruitment of polymerases, and origin firing. Complexes which remain at origins after firing (or passive replication from a fork coming in from a nearby origin) are at the post-replicative state, and cannot support replication initiation again until mitosis has been completed and a new round of licensing has taken place. It is therefore the composition of complexes along the DNA which dictate when and where replication will initiate.

The Cell Cycle Control System Orchestrates Transitions

DNA replication must be accurately controlled in space and time and must be coordinated with cell cycle progression. How are the ordered transitions of origin-bound complexes brought about at the correct point in time during the cell cycle? Cyclin dependent kinases (CDKs) (▶ Cyclins and Cyclin-dependent Kinases), the master regulators of the cell cycle, cross-talk with origin-bound complexes to ensure once per cell cycle replication (Blow and Dutta 2005, Fig. 2). When mitosis is completed, CDK activity levels are low. Only then can pre-replicative complexes assemble at origins of replication (window of opportunity). Increase in CDK activity levels at the G1/S transition signals conversion of pre-replicative complexes to pre-initiation complexes and entry into S-phase. CDK activity levels over the G1/S transition threshold however inhibit further assembly of pre-replicative complexes, ensuring that licensing will only take place again after mitosis has been completed and CDK activity levels have dropped (▶ Cell Cycle Transitions, Mitotic Exit). The CDK cycle therefore restricts licensing and replication to different windows of the cell cycle, guarding against re-replication. To ensure that mitosis only occurs after DNA replication has been completed, replication complexes signal to the cell cycle control system to inhibit CDK activity



DNA Replication, Fig. 2 DNA replication licensing

from reaching the levels required for mitotic entry (▶ G2/M Checkpoint) until every part of the genome has been replicated. Defects in the cross-talk between CDKs and origin-bound complexes can lead to over-replication of the genome or a catastrophic entry into mitosis with unreplicated DNA.

Mathematical Models of DNA Replication

As eukaryotic DNA replication is characterized by a high degree of uncertainty, both in the location and in the time of activation of origins along the genome, mathematical models have been employed to capture how DNA replication may be progressing in each cell in a population and to allow accurate interpretation of experimental data (reviewed in Hyrien and Goldar 2010). One of the first models to be developed was a stochastic model for DNA replication based on the KJMA model of phase transition kinetics, originally used to analyze single-molecule data of DNA replication in cell-free extracts of *Xenopus laevis* embryos and further exploited for analyses of DNA replication dynamics (Herrick et al. 2002; Hyrien and Goldar 2010). A stochastic hybrid model of eukaryotic DNA replication incorporating exact locations and firing propensities of origins along a complete genome was proposed by Lygeros et al. 2008. The model was instantiated using experimental data for *Schizosaccharomyces pombe* and Monte Carlo simulations were used to reproduce full genome replication at the single-cell level and statistically analyze the properties of the process at the population level. Spiesser et al. (2009) modeled replication in *Saccharomyces* deterministically using data for location and firing times of a fraction of replication origins. De Moura et al. 2010 developed a stochastic model of DNA replication which was employed for quantitative analysis of the dynamics of replication of chromosome VI of *S. cerevisiae*, while Yang et al. 2010 presented an analytical model of DNA replication with which replication timing across the *S. cerevisiae* genome was analyzed. A model to analyze replication fork failure in metazoan cells was developed by Blow and Ge 2009. The model assumes stochastic origin activation in a cluster of 5–100 potential origins on a circular 250 kb DNA molecule, modeling replication in a series of discrete time steps.

Uncertainty and Robustness in DNA Replication

Model predictions and single cell experiments indicate that random selection and activation of origins of replication results in an exponential distribution of distances between active origins. Such a distribution would produce infrequent large inter-origin gaps, which would need a long time to be replicated. This complication of random origin selection has been named the *random completion* or *random gap* problem. Several hypotheses have been proposed to resolve this paradox (reviewed in Legouras et al. 2006;

Hyrien and Goldar 2010), including redistribution of a limiting factor, defined spacing of active origins, or origin redundancy. It has also been suggested that S-phase may in fact last longer than previously assumed, occupying much of what is currently thought of as the G2 phase of the cell cycle (Lygeros et al. 2008).

Given the uncertainty inherent in stochastic origin selection, why have eukaryotes not opted for a deterministic mode of origin selection? It is likely that stochastic origin selection offers robustness because of redundancy (Legouras et al. 2006): there are many more origins ready to fire than those actually required to complete S-phase which can be used if need arises. One example of this is DNA damage: when cells experience DNA damage and active forks arrest, the presence of dormant origins becomes essential for the completion of DNA replication (Blow and Ge 2009). Differences in transcriptional programs in different cell types offers another example: transcription cross-talks with origin selection and origin usage changes under different metabolic conditions or differentiation. The excess in putative origins may therefore ensure timely completion of replication under various, often adverse, conditions.

Higher-Order Organization of DNA Replication Within the Cell Nucleus

We have thus far considered DNA replication as a linear process along the DNA. DNA is however tightly packaged within the cell nucleus and DNA replication is topologically organized, providing an additional level of regulation. Replication origins appear to fire in clusters (of 6–12 active origins within an approximately 1 Mb region) which are co-regulated and are visible within the cell nucleus as replication foci (or factories). Such a topological organization may offer a number of advantages: sequestration of replication proteins within factories may help increase their local concentration facilitating the kinetics of DNA replication; co-replication of origins within a similar chromatin context may facilitate the inheritance of epigenetic modifications at a given locus; local organization in clusters provides the possibility for differential regulation within a cluster (e.g., local firing of dormant origins in the vicinity of DNA damage) and outside the cluster (e.g., global inhibition of replication when DNA damage is present elsewhere in the genome). Future work will hopefully elucidate how the process of DNA replication is organized in time and space and how it cross-talks with other cellular processes, such as transcription and the inheritance of epigenetic states.

Cross-References

- ▶ Cell Cycle
- ▶ Cyclins and Cyclin-dependent Kinases
- ▶ DNA Polymerases
- ▶ DNA Replication Licensing
- ▶ Prereplicative Complex
- ▶ Replication Fork
- ▶ Replication Origin
- ▶ Replisome

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2007) Molecular biology of the cell, 5th edn. Garland Science, New York

Blow JJ, Dutta A (2005) Preventing re-replication of chromosomal DNA. Nat Rev Mol Cell Biol 6:476–486

Blow JJ, Ge XQ (2009) A model for DNA replication showing how dormant origins safeguard against replication fork failure. EMBO Rep 10:406–412

de Moura APS, Retkute R, Hawkins M, Nieduszynski CA (2010) Mathematical modelling of whole chromosome replication. Nucleic Acids Res 38:5623–5633

DePamphilis ML (2006) DNA replication and human disease. CSHL Press, Cold Spring Harbor

Gilbert DM (2004) In search of the holy replicator. Nat Rev Mol Cell Biol 5:848–855

Herrick J, Jun S, Bechhoefer J, Bensimon A (2002) Kinetic model of DNA replication in eukaryotic organisms. J Mol Biol 320:741–750

Hyrien O, Goldar A (2010) Mathematical modelling of eukaryotic DNA replication. Chromosome Res 18:147–161

Legouras I, Xouri G, Dimopoulos S, Lygeros J, Lygerou Z (2006) DNA replication in the fission yeast: robustness in the face of uncertainty. Yeast 23:951–962

Lygeros J, Koutroumpas K, Dimopoulos S, Legouras I, Kouretas P, Heichinger C, Nurse P, Lygerou Z (2008) Stochastic hybrid modeling of DNA replication across a complete genome. Proc Natl Acad Sci USA 105:12295–300

Spiesser TW, Klipp E, Barberis M (2009) A model for the spatiotemporal organization of DNA replication in *Saccharomyces cerevisiae*. Mol Genet Genomics 282:25–35

Yang SC, Rhind N, Bechhoefer J (2010) Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. Mol Syst Biol 6:404

DNA Replication Licensing

Zoi Lygerou
School of Medicine, Laboratory of General Biology, University of Patras, Patras, Greece

Definition

DNA replication licensing is the process of marking origins of replication as competent for a new round of replication. It takes place in late mitosis and G1 and consists of the loading of the hexameric MCM complex, which will later act as the replicative helicase, onto origins of replication. It requires the origin recognition complex (a six subunit complex which binds to origin DNA) and the MCM loading factors Cdc6/Cdc18 and Cdt1.

Cross-References

▶ DNA Replication

DNA Sequencing

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

Synonyms

Deoxyribonucleic acid sequencing

Definition

DNA sequencing is the procedure of determining the order of nucleotide bases, i.e., adenine, guanine, cytosine, and thymine, in the DNA molecule (Maxam and Gilbert 1977).

References

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Biochemistry 74(2):560–564

DNA Synthesis

▶ DNA Replication

Document Classification/Categorization

▶ Text Classification

Document Retrieval

▶ Information Retrieval

DOE

▶ Design of Experiments

Domain Ontology

▶ Cell Cycle Ontology (CCO)

Domain Type

▶ IMGT-ONTOLOGY, DomainType

Doob–Gillespie Algorithm

▶ Stochastic Simulation Algorithm

Dopamine

Rajeswara Babu Mythri[1], Shireen Vali[2] and M. M. Srinivas Bharath[1]
[1]Department of Neurochemistry, National Institute of Mental Health and Neurosciences (NIMHANS), Bangalore, Karnataka, India
[2]Cell Works Group Inc., Bangalore, India

Definition

DA is a catecholamine that occurs in a wide variety of animals, including both vertebrates and invertebrates. In the brain, DA functions both as a neurotransmitter and a neurohormone. DA functions by activating the five types of DA receptors termed D1–D5. DA is produced in several areas of the brain, including the SN and the ventral tegmental area. DA is released by the hypothalamus as a neurohormone which inhibits the release of prolactin from the anterior lobe of the pituitary. Swedish scientist Arvid Carlsson, the Nobel laureate performed fundamental biochemical experiments demonstrating the neurotransmitter role of DA which later paved the way for the first clinical therapy of PD.

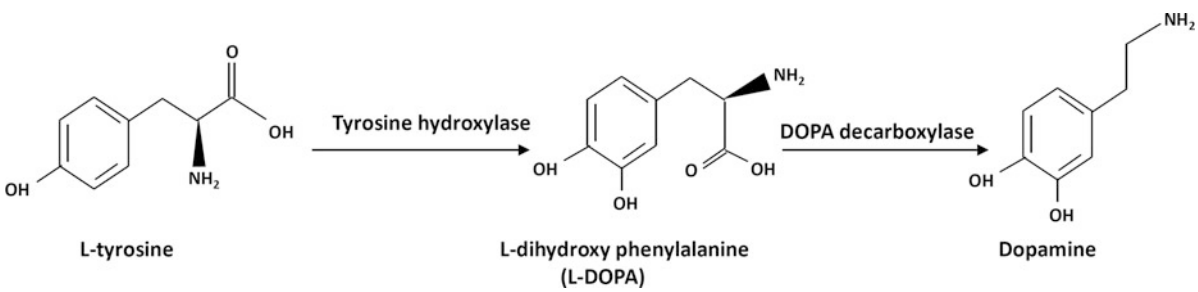
DA is also a precursor for other catecholamine neurotransmitters, epinephrine and norepinephrine. Biosynthesis of DA in the body occurs from the amino acid precursor, L-tyrosine, by a two-step reaction (Fig. 1).

DA synthesized at the presynaptic terminal following suitable stimulus, is released into the synaptic cleft and is taken up by DA receptors on the postsynaptic terminal. D1 and D5 receptors belong to D1-like family of receptors which are stimulatory in function and their activation results in increased production of the second messenger cyclic adenosine-5′-monophosphate (cAMP), while D2, D3, and D4 receptors belong to the D2-like family of receptors which are primarily inhibitory and inhibit the production of cAMP. Whatever be the type of receptor, on completion of its function, DA is degraded by either monoamine oxidase A/B (MAO A/B) or catechol-*O*-methyl transferase (COMT) at the synaptic cleft or following reuptake into the presynaptic neuron. Alternatively, the DA undergoing reuptake into the presynaptic neuron is repackaged into vesicles for the next cycle of synaptic activity.

DA synthesizing (dopaminergic) neurons project axons to larger brain regions which control behavior and cognition, voluntary movements, motivation, punishment and reward, inhibition of prolactin production, sleep, mood, attention, working memory, and learning. DA is commonly associated with the pleasure system of the brain, providing feelings of enjoyment and reinforcement to motivate a person proactively to perform certain activities. DA is released by naturally rewarding experiences such as food, sex, use of certain drugs, and stimuli that are associated with them.

The nigrostriatal pathway that involves dopaminergic neurons from the SN into the striatal region of the brain is the region of focus in PD pathology. This pathway is directly involved in controlling voluntary movement. Gradual loss of these neurons in PD patients causes drastic DA deficiency in the striatum consequently reducing the ability to perform smooth and controlled movements.

It could therefore be surmised that DA can be administered as a therapeutic molecule for controlling the motor symptoms of PD. However, DA supplied as



Dopamine, Fig. 1 Synthesis of DA from L-tyrosine

a drug acts on the sympathetic nervous system, producing effects such as increased heart rate and blood pressure. Since DA cannot cross the blood-brain barrier, it does not directly affect the central nervous system. Therefore, the therapeutic approach for PD involves the administration of DA precursor L-dihydroxy phenyl alanine (L-DOPA or levodopa), which can cross the blood-brain barrier and induce DA synthesis.

Cross-References

▶ Disease System, Parkinson's Disease

DOQCS: Database of Quantitative Cellular Signaling

▶ Database of Quantitative Cellular Signaling (DOQCS)

DOR

▶ Dense Overlapping Regulons

Dosage Compensation

▶ X Chromosome Inactivation

Doubling Time

▶ Modeling, Cell Division and Proliferation

Downward Looking

▶ Reduction

DRIP

▶ Mediator

Drug Discovery

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

Synonyms

Pharmaceutical discovery

Definition

Drug discovery is "the process of identifying chemical entities that have the potential to become therapeutic agents" (Decker and Sausville 2007).

There are two different approaches to drug discovery: empirical and rational. Empirical drug discovery involves finding a compound that produces a desired therapeutic effect in vitro. Initially, there is no understanding of the candidate drug's mechanism of action. In rational drug discovery, on the other hand, the target is known from the beginning; scientists then attempt to find or design compounds which would interact with the target of interest (Decker and Sausville 2007).

Cross-References

▶ Biological Activity
▶ Drug Target
▶ Natural Product Resources

References

Decker S, Sausville E (2007) Drug discovery. In: Atkinson AJ, Abernethy DR, Daniels CE, Dedrick RL, Markey SP (eds) Principles of clinical pharmacology, 2nd edn. Academic, Burlington, pp 439–447

Drug Disease Networks

▶ Systems Pharmacology, Drug Disease Interactions

Drug Metabolism

▶ Epigenetics, Drug Discovery
▶ Pharmacokinetics and Pharmacodynamics

Drug Scope, Metabolic

Jean-Marc Schwartz
Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, Manchester, UK

Definition

The metabolic drug scope is an extension of the concept of scope in metabolic pathways. The scope of a metabolic compound is the set of all compounds that can be generated in principle by transformations of the seed compound, irrespective of kinetic and thermodynamic laws that determine the rate at which these transformations might actually take place (Handorf et al. 2005). When a set of several seed compounds is considered, the resulting scope is the set of all compounds that can be generated by transformations and combinations of the seeds.

The scope of metabolic compounds is generated through an expansion process. The principle used is that for any reaction to take place, all necessary substrates must be present. Starting from the seed compounds, products generated by metabolic reactions using the seeds are iteratively added, until no further reaction is possible.

By extension, the metabolic drug scope is the scope generated by the enzymatic targets of a drug. The metabolic drug scope is constructed by the expansion of a set of seeds containing the substrates and products of all metabolic reactions targeted by the drug. Essentially, the metabolic drug scope represents the largest possible network that a drug might influence in a metabolic system (Schwartz and Nacher 2009).

Cross-References

▶ Systems Pharmacology, Drug Disease Interactions
▶ Systems Pharmacology, Drug-Target Networks

References

Handorf T, Ebenhöh O, Heinrich R (2005) Expanding metabolic networks: scopes of compounds, robustness, and evolution. J Mol Evol 61:498–512
Schwartz JM, Nacher JC (2009) Local and global modes of drug action in biochemical networks. BMC Chem Biol 9:4

Drug Target

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

Synonyms

Molecular drug target; Target

Definition

A drug target is a macromolecule that undergoes a specific interaction (e.g., binding, inhibition) with a drug. A target is linked to a specific disease; the interaction between a drug and a target is expected to elicit an effect on the course of the disease.

There are four main types of drug targets: proteins, polysaccharides, lipids, and nucleic acids. Among them, proteins are considered the best source of drug targets as most drugs have been shown to interact with them. Proteins can be further divided into seven families: G-protein-coupled receptors, ion channels, protein kinases, zinc metalloproteases, serine proteases,

nuclear hormone receptors, and phosphodiesterases (Giegel et al. 2007).

Some targets belong to or are associated with pathogenic organisms (e.g., bacteria, viruses, and fungi), which cause infectious diseases (Gies and Landry 2008).

Cross-References

▶ Natural Product Resources

References

Giegel DA, Lewis AJ, Worland P (2007) Diversity versus focus in choosing targets and therapeutic areas. In: Taylor JB, Triggle DJ (eds) Comprehensive medicinal chemistry II. Elsevier, Oxford, pp 753–770

Gies J, Landry Y (2008) Molecular drug targets. In: Wermuth CG (ed) The practice of medicinal chemistry, 3rd edn. Elsevier Science and Technology, Amsterdam, pp 85–105

Drug Target Strategies

▶ Pathway Targeting, Antimycobacterial Drug Design

Drug Target, Off-Target

Ravi Iyengar
Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, NY, USA

Definition

Target of a drug that is not the *primary target*, which may give rise to undesirable pathophysiology leading to *adverse events*.

Cross-References

▶ Adverse Events
▶ Primary Target
▶ Systems Pharmacology

Drug Targets

▶ Epigenetics, Drug Discovery
▶ Host–Pathogen Systems, Target Discovery
▶ Pharmacogenomics

Drugs

▶ Xenobiotics

Duration Analysis

▶ Survival Analysis
▶ Survival Analysis, Fundamental Statistical Techniques

Dye Dilution

▶ Quantifying Lymphocyte Division, Methods

Dynamic Bayesian Networks

Lin Wang
School of Computer Science and Information Engineering, Tianjin University of Science and Technology, Tianjin, China

Synonyms

Dynamic conditional independent graphs

Definition

Dynamic Bayesian network is a representation of stochastic evolution of a set of random variables $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ over discretized time. It consists of a directed graph representing conditional independences and a family of conditional distributions

$\boldsymbol{P(x_i(t)|Pa[x_i(t\text{-}1)])}$, where $\mathbf{Pa[x_i(t\text{-}1)]}$ represents the set of parent nodes of the node $\mathbf{x_i(t)}$. The temporal process of a dynamic Bayesian network is assumed to be a time-homogeneous Markov process:

$$P[X(t)|X(0), X(1), \cdots, X(t-1)] = P[X(t)|X(t-1)], \tag{1}$$

The joint distribution over all the possible trajectories of the process is decomposed into the following product form:

$$P[X(0), X(1), \cdots, X(T)] = P(0) \prod_{t=1}^{T} P[X(t)|X(t-1)], \tag{2}$$

Therefore, given an initial state of random variables, their evolution is given by:

$$\begin{aligned} P[X(1), \cdots, X(T)|X(0)] &= \prod_{t=1}^{T} P[X(t)|X(t-1)] \\ &= \prod_{t=1}^{T} \prod_{i=1}^{n} P(x_i(t)|\mathrm{Pa}[x_i(t-1)]), \end{aligned} \tag{3}$$

References

Chen L, Wang RS, Zhang XS (2009) Biomolecular networks. Wiley, New York

Dynamic Conditional Independent Graphs

▶ Dynamic Bayesian Networks

Dynamic Mechanistic Explanation

▶ Mechanism, Dynamic

Dynamic Metabolic Flux Analysis

Yun Lee, I-Chun Chou, Melissa L. Kemp and Eberhard O. Voit
The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Definition

It is easier to deduce the static structure of a biological system from experimental data than to account for its full dynamic repertoire. Nonetheless, knowledge of the static structure reveals strict constraints that can form the basis for constructing dynamic models. While this conversion step from static to dynamic models is anything but solved, this essay discusses some generic strategies toward accomplishing the task.

Characteristics

Experimental and Computational Strategies of Model Construction

Metabolic pathway systems are characterized by three classes of components: metabolites and their concentrations, enzymes and modulators, and fluxes that describe how much material flows through the systems. Sole knowledge of metabolites at the normal state of a pathway system is not sufficient to deduce fluxes, and sole knowledge of fluxes does not permit inferences on the metabolite concentrations. Three trends in metabolic analysis are in the process of converging and have the potential to offer new insights:

1. Systematic profiling of metabolite concentrations using NMR spectroscopy and mass spectrometry
2. Experimental metabolic flux analysis
3. Computational methods

Many platforms are now available to provide targeted analysis of hundreds of metabolites in a single biological sample. The *metabolic profiles* thus obtained have been used to distinguish molecular alterations between two samples, whether derived from benign and metastatic cancer, or from a mutant and the corresponding control. Current profiling methods are mainly based on two analytical techniques: nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) coupled to a pre-separation technique such as chromatography or electrophoresis.

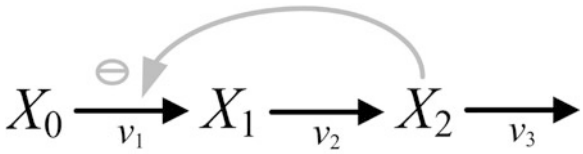
Of the two platforms, MS-related techniques are considered more sensitive and have been used to obtain one-time snapshots of thousands of metabolites. NMR spectroscopy has its own advantages, because it is noninvasive and of a high-throughput nature and can, therefore, be used to execute dense in vivo time-series measurements of moderately large collections of metabolites. With considerable advances being made on both analytical fronts, it is to be expected that time-resolved profiles of thousands of metabolites will become the norm rather than the exception in the foreseeable future.

In contrast to the concentrations of enzymes and metabolites that define cellular metabolism, metabolic fluxes cannot be measured directly but rather need to be inferred from measurable quantities. A method developed for this purpose, *metabolic flux analysis* (MFA), dates back to the 1970s, where one of its early applications was to infer the rates of intracellular reactions from the measurements of secreted metabolites and from coarse knowledge of the pathway structure (Aiba and Matsuoka 1979). Since then, the use of stable isotopes (mostly by feeding ^{13}C-labeled substrates to cultured cells) was shown to refine the flux estimates considerably, especially for ratios of fluxes at branch points (Stephanopoulos 1999). These local flux ratios, when combined with a stoichiometric model of the metabolic pathway system, furthermore permit the determination of absolute rates for all fluxes through an iterative fitting algorithm (Wiechert 2001). Evidently, this process poses a challenge for undertaking MFA in higher organisms because their pathway structure is more complex and, in many cases, ill-defined. To address this challenge, more advanced experimental design and techniques like dynamic labeling are required, but so far their application has been limited (Voit et al. 2004; Schaub et al. 2008).

In addition to experimental approaches, computational methods have been developed and successfully used to study metabolic pathway systems, thanks in part to advances in computers and the increasingly easy access to them. Most existing methods fall into two categories: (1) stoichiometric and flux-based models; and (2) dynamic, kinetics-based models. Models from the former class are based on the assumption that the metabolic transients are much faster than both cellular growth and environmental fluctuations. Consequently, the metabolic fluxes are assumed to be in a quasi-steady state where, for any metabolite pool, the fluxes governing its synthesis and degradation are equal, thus leading to the name "flux balance analysis" (FBA; Palsson 2006). FBA extends stoichiometric network analysis by also accounting for thermodynamic and other physicochemical constraints that limit the reactions within the network. Furthermore, and in contrast to MFA, FBA identifies a particular flux distribution under the assumption that the cell strives to meet a specific objective such as maximizing growth.

A significant feature – but also its major weakness – of FBA is that it contains no information about metabolite concentrations. This may pose a problem if the purpose of an investigation is, for example, to investigate the effect of a certain mutation on the intracellular level of a metabolite serving as a biomarker. In such a case, dynamic, kinetics-based models that center on metabolite concentrations appear to be a better fit. Traditionally, the formulation of dynamic models of metabolic pathway systems starts with finding a functional representation for each reaction that best describes its kinetics in vitro. Given the explicit representations of individual reactions, the next step is to integrate them into a system of ordinary differential equations (ODEs) where each equation describes the temporal change in one metabolite as a difference between the sums of rates (fluxes) of its synthesis and degradation. Lastly, having determined the initial concentrations, one solves the ODE model to obtain the metabolic concentrations at different time points, which are not necessarily at steady state, and compute fluxes if needed. Overall, the design of dynamic models requires many kinetic details, but these models eventually offer the ability to predict all metabolite concentrations and fluxes under non-steady-state conditions.

Interestingly, there is little overlap between the two different kinds of metabolic models. The only common characteristic is that they both yield flux estimates at the steady state, although with distinct tactics: one uses a top-down approach by directly predicting the fluxes under a quasi-steady-state assumption, whereas the other takes a bottom-up approach through solving the integrated ODE model toward the steady-state. For a metabolic pathway system, however, the kinetic data obtained individually with purified enzymes may not reflect the true kinetic behavior in vivo. Therefore, a sensible approach would be first to determine the flux



Dynamic Metabolic Flux Analysis, Fig. 1 A generic linear pathway with feedback inhibition by the product (X_2)

distribution with methods of MFA or FBA and then to infer the kinetic parameters based on concentration measurements and functional descriptions of individual reactions. Such a merger of two otherwise distinct modeling techniques has the following attraction: not only is the resulting model more accurate at the level of fluxes, but one also gains information on metabolite concentrations. This type of merger is greatly facilitated by using canonical models within the framework of *Biochemical Systems Theory* (BST) (Savageau 1976; Voit 2000) because mechanistic assumptions regarding the enzymatic processes can be minimized.

Let us illustrate the idea with a linear pathway (Fig. 1). Suppose a fixed amount of substrate X_0 is converted into product X_2 in two steps, with X_2 inhibiting the synthesis of the intermediate X_1 through feedback.

In FBA, the assumption of a metabolic quasi-steady state leads to the following flux balance equations:

$$\begin{aligned} v_1 - v_2 &= 0 \\ v_2 - v_3 &= 0 \end{aligned} \tag{1}$$

Equation 1 is underdetermined since the number of fluxes exceeds the number of metabolites (equations) where the steady-state constraints are imposed (i.e., X_1 and X_2). Therefore, infinitely many solutions exist. In this case, a particular solution may be obtained if one of the three fluxes can be directly measured, as it might be the case for substrate uptake (v_1) or product formation (v_3). If none of the three fluxes is measurable, one might be able to make reasonable assumptions regarding the biomass composition of the system, such as:

$$a_1 X_1 + a_2 X_2 \xrightarrow{v_4} Biomass, \tag{2}$$

where a_1 and a_2 represent stoichiometric coefficients for the two species. Linear optimization can be used to identify a flux distribution that achieves maximal growth by solving the task in Eq. 3:

$$\begin{aligned} &\text{maximize } v_4 \\ &\text{subject to } v_1 - v_2 - a_1 v_4 = 0 \\ &\qquad\qquad v_2 - v_3 - a_2 v_4 = 0 \end{aligned} \tag{3}$$

Notably, no explicit accounts of metabolite concentrations are involved in FBA, and the result exclusively addresses fluxes.

As an alternative, one can construct an ODE model using traditional enzyme kinetics. For instance, assuming that all reactions follow the classic Michaelis–Menten type kinetics and that the inhibition by X_2 is competitive, the corresponding ODE model is formulated as:

$$\begin{aligned} X_0 &= \text{constant} \\ \dot{X}_1 &= v_1 - v_2 = \frac{V_1 X_0}{X_0 + K_1(1 + X_2/K_4)} - \frac{V_2 X_1}{X_1 + K_2} \\ \dot{X}_2 &= v_2 - v_3 = \frac{V_2 X_1}{X_1 + K_2} - \frac{V_3 X_2}{X_2 + K_3} \end{aligned} \tag{4}$$

Solving Eq. 4 requires knowledge of the initial metabolite concentrations and of all kinetic parameters (V_i and K_i). In practice, these are often determined using purified enzymes and thus prone to uncertainties due to the fact that in vivo systems are quite different from in vitro experiments. A novel means of parameter estimation, thanks to the advent of metabolic time-series profiles, is to substitute the differentials on the left-hand side of Eq. 4 with estimated slopes at discrete time points, transform the coupled system of differential equations into several sets of decoupled algebraic equations, and identify the optimized values of parameters via regression (Voit and Almeida 2004). For some metabolic systems, we can even derive the dynamic flux profiles from the slope data in the first place and then estimate the parameters on a flux basis (Goel et al. 2008).

In cases where the time-series metabolic profiles are not accessible, the steady-state flux distributions, as determined by MFA or FBA, can be valuable for parameter estimation. The idea is that instead of using the estimated fluxes at multiple time points within one experiment, one may slightly perturb the system many times and record or predict the steady-state responses. With flux and concentration data from multiple perturbation experiments, the task is again

reduced to fitting parameters individually for each flux. To illustrate the point, it is convenient to model the linear pathway in Fig. 1 with power-law functions, as proposed in BST:

$$\begin{aligned} X_0 &= \text{constant} \\ \dot{X}_1 &= v_1 - v_2 = \gamma_1 X_0^{f_{1,0}} X_2^{f_{1,2}} - \gamma_2 X_1^{f_{2,1}} \\ \dot{X}_2 &= v_2 - v_3 = \gamma_2 X_1^{f_{2,1}} - \gamma_3 X_2^{f_{3,2}} \end{aligned} \quad (5)$$

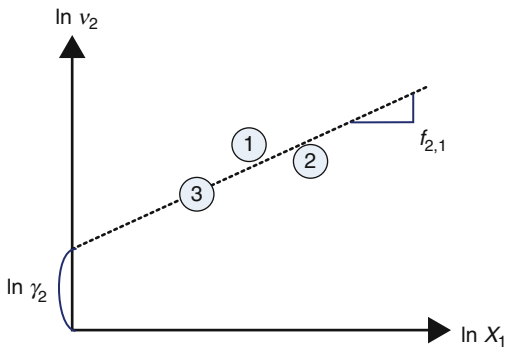
In Eq. 5, each flux has two or three unknown parameters (rate constants γ_i and kinetic orders $f_{i,j}$). Thus, we need flux and concentration data at three or more different steady states for parameter estimation. This argument is valid, because a flux, say v_2, reduces to a linear equation, when we take logarithms of the power-law representation:

$$\ln v_2 = \ln \gamma_2 + f_{2,1} \ln X_1. \quad (6)$$

We can see that the kinetic order $f_{2,1}$ is the slope in the plot of $\ln v_2$ versus $\ln X_1$, while the logarithm of the rate constant $\ln \gamma_2$ is the y-intercept (Fig. 2). If experimentally feasible, it is important that enough data are obtained to allow a statistical regression analysis.

Many experimental strategies are available for artificially driving the pathway of interest to different metabolic steady states. For instance, one may slightly perturb the system, which is initially at a nominal steady state, by changing the amount of substrate fed into the pathway. Another approach is to genetically modify the activity of pathway enzymes, for instance, through a gene knockdown experiment. In the latter approach, one must also measure the relative enzyme activities compared to those in the control to adjust for the bias introduced to the rate constants.

Overall, the transformation of steady-state, flux-based models into dynamic, kinetics-based models requires large amounts of concentration measurements, which despite the substantial improvements that have recently been made in the field of metabolomics, remains a challenge. The issue is especially significant in plant biology: not only do plants synthesize a far more diverse array of metabolites than do animals and microorganisms, but we also lack the ability to identify the majority of signals from metabolic profile data (Saito and Matsuda 2010). Nevertheless, even in this complicated case of plants, it was shown that even without a comprehensive set of concentration data, the conversion can still be accomplished through a combination of model reduction and optimization methods (Lee and Voit 2010).



Dynamic Metabolic Flux Analysis, Fig. 2 Estimation of $f_{2,1}$ and γ_2 using data from three observed steady states. The colored circles refer to the log-transformed values ($\ln X_1$, $\ln v_2$) taken either from the nominal steady state (1) or from new steady states where, for instance, the input substrate X_0 is increased (2) or the activity of enzyme catalyzing v_3 is decreased (3)

Cross-References

- ^{13}C Metabolic Flux Analysis
- Flux Balance Analysis
- Metabolic Flux Analysis
- Metabolic Networks, Structure and Dynamics
- Metabolic Pathway Analysis
- Ordinary Differential Equation (ODE)
- Optimization Algorithms for Metabolites Production
- Parameter Estimation, Metabolic Network Modeling
- Pathway Modeling, Metabolic

References

Aiba S, Matsuoka M (1979) Identification of metabolic model: citrate production from glucose by *Candida lipolytica*. Biotechnol Bioeng 21:1373–1386

Goel G, Chou I-C, Voit EO (2008) System estimation from metabolic time-series data. Bioinformatics 24:2505–2511

Lee Y, Voit EO (2010) Mathematical modeling of monolignol biosynthesis in *Populus* xylem. Math Biosci 228:78–89

Palsson BØ (2006) Systems biology: properties of reconstructed networks. Cambridge University Press, Cambridge
Saito K, Matsuda F (2010) Metabolomics for functional genomics, systems biology, and biotechnology. Annu Rev Plant Biol 61:463–489
Savageau MA (1976) Biochemical systems analysis: a study of function and design in molecular biology. Addison-Wesley, Reading
Schaub J, Mauch K, Reuss M (2008) Metabolic flux analysis in *Escherichia coli* by integrating isotopic dynamic and isotopic stationary ^{13}C labeling data. Biotechnol Bioeng 99:1170–1185
Stephanopoulos G (1999) Metabolic fluxes and metabolic engineering. Metab Eng 1:1–11
Voit EO (2000) Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists. Cambridge University Press, Cambridge
Voit EO, Almeida J (2004) Decoupling dynamical systems for pathway identification from metabolic profiles. Bioinformatics 20:1670–1681
Voit EO, Alvarez-Vasquez F, Sims KJ (2004) Analysis of dynamic labeling data. Math Biosci 191:83–99
Wiechert W (2001) ^{13}C metabolic flux analysis. Metab Eng 3:195–206

Dynamic Metabolic Networks, k-Cone

Isaac F. López-Moyado and
Osbaldo Resendis-Antonio
Center for Genomic Sciences-UNAM, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

Synonyms

Feasible parameter space; k-cone space; Modal analysis

Definition

Genome-scale metabolic reconstruction constitutes a paradigm in systems biology that currently extends its escope to eukaryotes, prokaryotes, and archaea (Oberhardt et al. 2009). Among a variety of biological questions that can be surveyed with these metabolic reconstructions, the dynamical description of metabolism is a noteworthy issue for exploring how the metabolic phenotype changes under external perturbations or internal gene deletions. In the assumption that linear perturbations occur around a metabolic steady state, one can expect that even though the perturbation induces changes in the metabolic concentrations of the network, the system recovers its initial metabolic profile. The dynamical description of this relaxation process is described by:

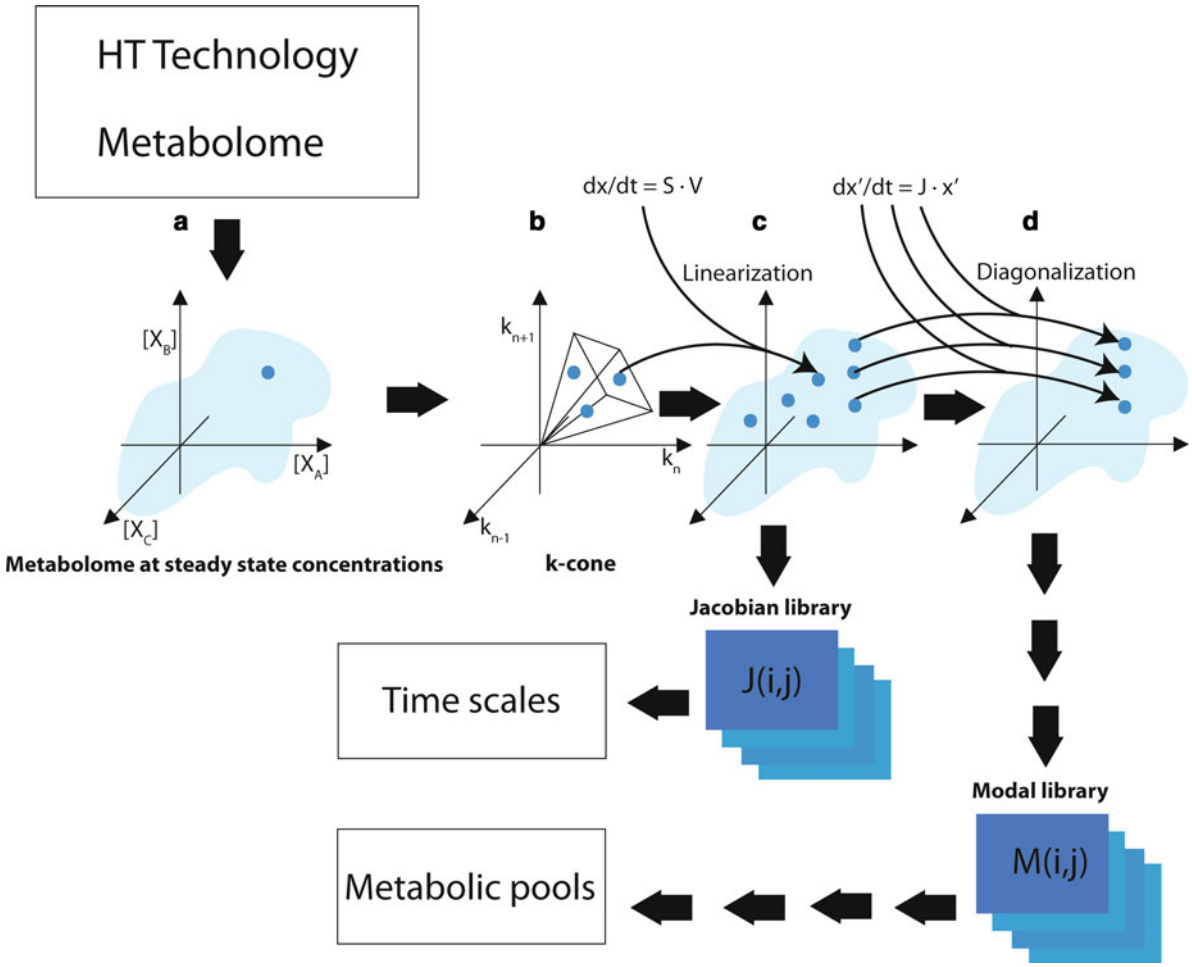
$$\frac{dm}{dt} = J' \cdot m \tag{1}$$

where J' is a diagonal matrix whose entries are the eigenvalues of the Jacobian matrix while m is a vector that specifies the modes of the system, i.e., the set of metabolites whose concentrations dynamically correlate at each timescale. Timescale distribution is defined by the negative inverse of J' eigenvalues.

Even though this theoretical framework can be straightforwardly applied to genome-scale metabolic reconstructions, a major limitation in dynamic modeling is the current lack of kinetic information. With the purpose to overcome this limitation, the *k-cone* has been suggested as an appealing scheme for exploring dynamic behavior of genome-scale reconstructions. This formalism refers to the feasible space of kinetic parameters that ensures a steady-state behavior in metabolic networks. From a mathematical point of view, this space is defined through the next equation:

$$S \cdot diag(C) \cdot \vec{k} = 0 \tag{2}$$

where S is the stoichiometric matrix and $diag(C)$ is a diagonal matrix whose entries are determined by a function of metabolic concentrations at steady state ($C = \Pi x_i^{|S_{ij}^R|}$, where S_{ij}^R refers to the reactant stoichiometric coefficients). In addition, $\vec{k}$ represents a vector whose dimensionality is determined by the number of metabolic fluxes in the network. The combination of modal theory and *k-cone* space supply with a statistical pipeline for exploring and surveying the dynamical behavior of genome-scale metabolic reconstructions, this framework being independent of a complete knowledge of the kinetics underlying the metabolic network and strongly dependent on metabolome data, see Fig. 1.



Dynamic Metabolic Networks, k-Cone, Fig. 1 General overview. (**a**) The metabolic state of a system can be obtained from metabolome data. (**b**) From this information and assuming that law mass action govern all metabolic reactions in the network, the feasible space of kinetic parameters, the k-cone, can be calculated. (**c**) Once defined the feasible set of kinetic parameters ensuring a steady state in the metabolic systems a Jacobian library can be calculated. (**d**) Finally, the modal library is recovered from the diagonalization of the Jacobian library. The Jacobian and the modal libraries contain information about the timescales and the metabolic pools formed during the relaxation process, respectively

Characteristics

Modal Analysis

Cells are continuously exposed to environmental perturbations whose biological effects are controlled by genetic, protein, and metabolic circuits for ensuring a proper functional state. In the particular situation where the perturbation is small, one expects that initially the metabolite concentrations change inside the cell; however, with time, these concentrations recover their original values characterizing a functional state in the cell. Thus, with the purpose of studying the dynamic properties of metabolism, it is convenient to find a way to study how a metabolic network relaxes to its steady state after an environmental perturbation has occurred (Kauffman et al. 2002). Modal analysis is a conceptual framework which is useful for this purpose by assuming that the perturbations occurred very close to a metabolic steady state. In this context, the temporal evolution of the systems is obtained by linearizing the dynamic mass balance equations around a reference point, i.e., the steady state.

In general, the temporal behavior of a metabolite concentration is calculated as a balance among the metabolic flux of those reactions that contribute to increase and decrease the production and use of metabolites. This principle is conserved at genome scale and the temporal behavior of a metabolic network integrated by n reactions and m metabolites is given by:

$$\frac{d\vec{x}}{dt} = S \cdot \vec{V} \tag{3}$$

where S denotes the ▶ stoichiometric matrix which is formed from the stoichiometric coefficients of the reactions that comprise a metabolic reconstruction. This matrix is organized in such a way that every column corresponds to a reaction and every row corresponds to a metabolic compound (as exemplified by Fig. 2 panel D). In addition, $\vec{V}$ refers to a vector that contains the fluxes of the reactions included in the reconstruction. Here $\frac{d\vec{x}}{dt}$ is a vector with time derivatives of the metabolite concentrations and it is equal to zero when the system has reached a steady state.

Here we limit our description to analyze the dynamical metabolic profile when small perturbations are applied to the metabolic reconstruction. Thus, in order to model the deviation from the steady-state concentration, we selected $\vec{x}'=\vec{x}-\vec{x}_{st}$ as our central variable where $\vec{x}_{st}$ and $\vec{x}$ indicate the concentrations of the metabolites at the steady and perturbed state, respectively. Under this assumption, a Taylor expansion around $\vec{x}$ lets us to obtain:

$$\frac{d\vec{x}'}{dt} = J \cdot \vec{x}' \tag{4}$$

with J being the Jacobian matrix, which in turn is obtained through:

$$J = S \cdot G \tag{5}$$

where G is the gradient matrix formed with the partial derivatives between the flux of the *i-esime* reactions of $\vec{V}$, V_i, and the concentration x_j:

$$G = \frac{\partial V_i}{\partial x_j} \tag{6}$$

In this contextual scheme, the Jacobian matrix contains information about the topology of the metabolic network and the thermodynamics of the reactions (Jamshidi and Palsson 2008).

As we mentioned earlier, the objective of using the theory of modal analysis is to recover dynamical information of the system, in particular (1) how the metabolism rearranges its constituents while it relaxes toward a steady state and (2) which are the timescales during this process. Even though these questions can be explored from Eq. 5, modal analysis is a proper formalism to explore these issues in a more direct way. Thus, by applying a diagonalization on Eq. 5 we obtain:

$$J = M \cdot J' \cdot M^{-1} \tag{7}$$

Notably, J' defined in Eq. 7 is a diagonal matrix with the eigenvalues of J ordered descendingly and the term M is a matrix whose columns are integrated by the eigenvectors of J. By substituting Eq. 7 into Eq. 4 we obtain:

$$\frac{d\vec{x}'}{dt} = J \cdot \vec{x}' = M \cdot J' \cdot M^{-1} \cdot \vec{x}' \tag{8}$$

$$M^{-1} \cdot \frac{d\vec{x}'}{dt} = M^{-1} \cdot M \cdot J' \cdot M^{-1} \cdot \vec{x}' \tag{9}$$

By defining the metabolic modes as:

$$m = M^{-1} \cdot \vec{x}' \tag{10}$$

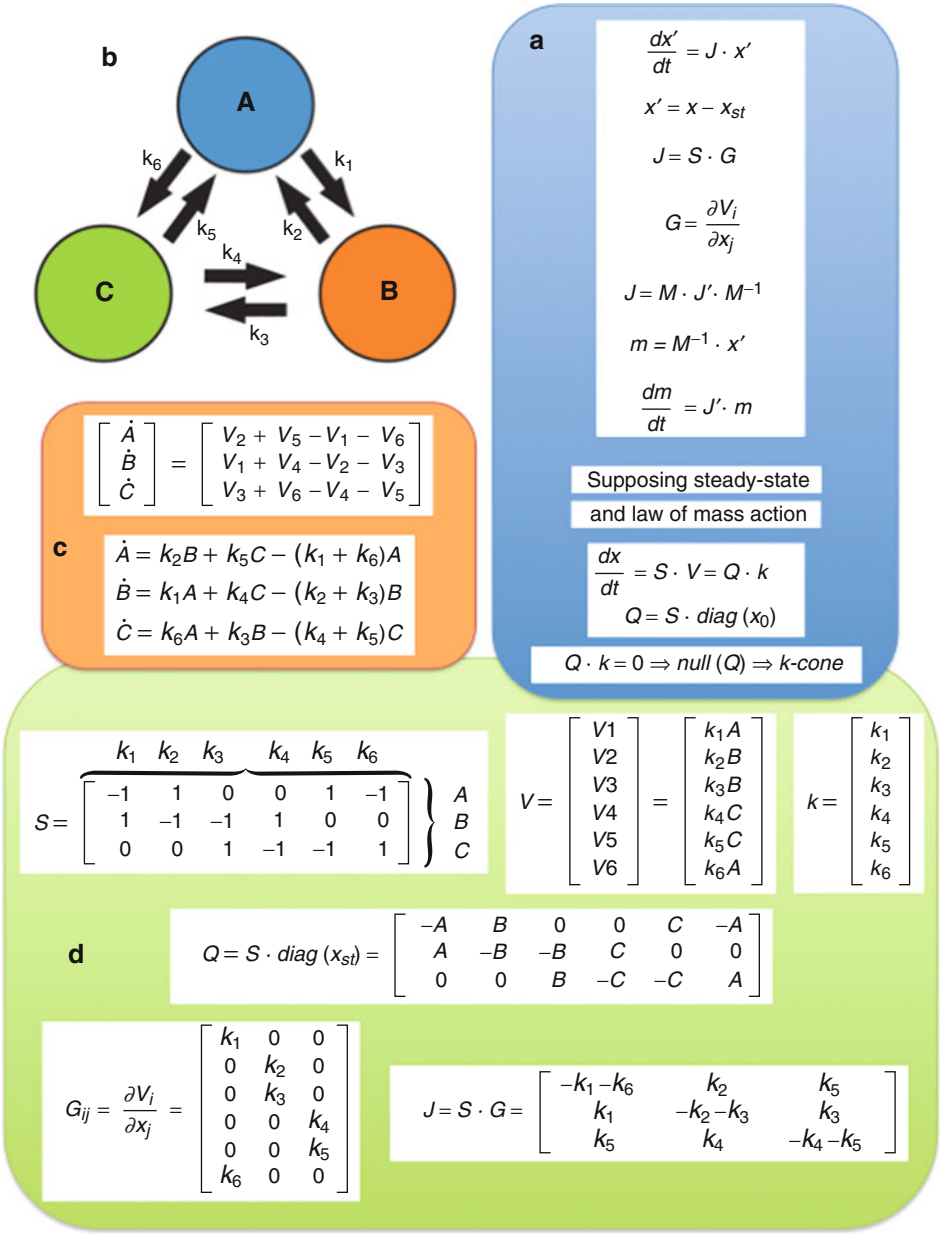
Equation 9 can be written in compact form as:

$$\frac{dm}{dt} = J' \cdot m \tag{11}$$

Thus, meanwhile m contains information of how the network coordinates and organizes its metabolites to reach the steady-state condition, and the diagonal matrix J specifies the timescales in which this process happens. This latter issue is given by the negative inverse of the eigenvalues of J (Kauffman et al. 2002; Resendis-Antonio 2009).

k-Cone Space

As we mentioned in the last section, the dynamical description of metabolism constitutes a cornerstone



Dynamic Metabolic Networks, k-Cone, Fig. 2 Example of the application of the theory of modal analysis and the *k-cone* formalism to a hypothetical metabolic network. (**a**) The *box* shown in this figure recapitulates the mathematical equations described in this essay. (**b**) A hypothetical metabolic network composed of three metabolites which interconvert to each other with reaction constants k_i. (**c**) *Box* showing the balance equations in terms of the fluxes V_i or the metabolite concentrations and kinetic constants. (**d**) *Box* exemplifying the matrices mentioned in the text constructed according to the network from (**b**)

in systems biology for surveying the mechanism by which cells respond under external perturbations. However, given that a variety of kinetic information is unknown in most of the biological systems, this formalism has been applied in few cases at genome scales. For instance, in the example depicted in Fig 2, it can be seen that the knowledge of the kinetic constants is needed to obtain the Jacobian matrix J (panel D). With the purpose of overcoming this issue, some databases are emerging (Rojas et al. 2007;

Scheer et al. 2011) that store the numerical values of some parameters required for analyzing certain biological circuits at well-defined physiological conditions. Although, these databases represent important contributions to enrich the model and experimentally assess its outcomes predictions, there is evidence that the numerical values of the kinetic parameters can vary depending on the physiological context. In fact, there is evidence that measurement in vitro and in vivo can differ by some orders of magnitude (Famili et al. 2005). In this contextual scheme, the elaboration of alternative procedures that allows us to estimate the range of each parameter participating in the reactions conforming a metabolic network would be of great value.

In general terms, the temporal evolution of the metabolite concentrations can be founded by solving Eq. 3, but this requires a complete knowledge of the values of S, $\vec{V}$, and the initial conditions, a fact that is not always fulfilled. However, if we suppose that all the reactions obey law of mass action, Eq. 3 can be written as:

$$\frac{d\vec{x}}{dt} = S \cdot diag(C) \cdot \vec{k} \tag{12}$$

where $diag(C)$ is a diagonal matrix whose entries are determined by a function of metabolic concentrations at steady state ($C = \Pi x_i^{|S_{ij}^R|}$, such that S_{ij}^R refers uniquely to the reactant stoichiometric coefficients) and $\vec{k}$ is a vector with the unknown kinetic constants and whose dimensionality is determined by the number of metabolic fluxes included in the metabolic reconstruction. Given that the numerical values of the kinetic constants remain along time, the numerical value of $\vec{k}$ can be straightforward identified at the steady-state regime. In such a situation, Eq. 12 is written as:

$$Q \cdot \vec{k} = 0 \tag{13}$$

where, we have stated that $S \cdot diag(C) = Q$ for simplicity in notation. Thus, we concluded that the numerical range of the kinetic parameters can be estimated through the right null space of Q. The feasible space of kinetic parameters that ensure the presence of a steady-state behavior in metabolic networks is called the *k-cone* (Famili et al. 2005). This multidimensional space let us identify and explore the potential response of the metabolic phenotype for a microorganism and allows us to survey how the metabolism coordinately acts to reach a steady state after one perturbation occurs. Remarkable, by selecting an ensemble of kinetic parameters belonging to the *k-cone*, one can build a library of feasible dynamic behavior and explore the average or most frequent behavior. Furthermore, this dynamic library can contribute to classify those parameters that have a high from those with a low numerical variability for recovering the steady state.

In order to apply this framework at the genome scale, an important issue is to have a well-defined metabolic profile at a steady state. As described in Fig. 1, the variety of technologies used in metabolome high-throughput data can fulfill this latter requirement. As we have seen throughout this essay, the combined effect of theory of modal analysis and the *k-cone* formalism supply with a pipeline to explore and survey the dynamical behavior of genome-scale metabolic reconstructions. This method is strongly dependent on quantitative metabolic profile of the organism in study and independent of a complete or partial knowledge of the kinetics underlying the metabolic network (Resendis-Antonio 2009).

Cross-References

- Jacobian Matrix
- Law of Mass Action
- Stoichiometric Matrix

References

Famili I, Mahadevan R, Palsson BO (2005) k-Cone analysis: determinig all candidate values for kinetic parameters on a network scale. Biophys J 88(3):1616–1625

Jamshidi N, Palsson BO (2008) Formulating genome-scale kinetic models in the post genome era. Mole Syst Biol 4:171

Kauffman KJ, Pajerowski JD, Jasmshidi N, Palsson BO, Edwards JS (2002) Description and analysis of metabolic connectivity and dynamics in the human red blood cell. Biophys J 83(2):646–662

Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Molec Syst Biol 5:320

Resendis-Antonio O (2009) Filling kinetic gaps: dynamic modeling of metabolism where detailed kinetic information is lacking. PLoS ONE 4(3):e4967

Rojas I, Golebieswki M, Kania R, Krebs O, Mir S, Weidemann A, Wittig U (2007) SABIO-RK: a database for biochemical reactions and their kinetics. BMC Syst Biol 1(Suppl 1):S6
Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, Sohngen C, Stelzer M, Thiele J, Schomburg D (2011) BRENDA, the enzyme information system in 2011. Nucleic Acids Res 39:670–676

Dynamic Modeling and Simulation

▶ Kinetic Modeling and Simulation

Dynamic Modularity

Junhua Zhang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
Key Laboratory of Random Complex Structures and Data Science, Chinese Academy of Sciences, Beijing, China
National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing, China

Synonyms

Community; Modular; Modularity; Module network

Definition

The main idea of dynamic modularity comes from the need to systematically explain the influence of a simple genetic change or environment perturbation on the behavior of an organism, which is also the ultimate goal of studying biological networks. However, research on dynamics of molecular networks remains very challenging even though a huge number of high-throughput data are available nowadays because the state space of networks grows exponentially with the number of network components (Alexander et al. 2009). So, methods to reduce the complexity of the analysis are of great interest. Thus, dynamic modularity appears, which can be used to explore molecular network dynamics to some extent by two steps. First, the network is decomposed into smaller building blocks that can be analyzed more easily, among which usually network modules (▶ Modularity; ▶ Module network) are of an imaginable form for decomposition because of network ▶ modularity. And second, the dynamical connectivity between different modules (▶ Modularity; ▶ Module network) is quantified under various conditions. The typical method for this purpose is modular response analysis (MRA) proposed by Kholodenko et al. (2002).

Cross-References

▶ Modular Organization of Gene Regulatory Networks

References

Alexander RP, Kim PM, Emonet T, Gerstein MB (2009) Understanding modularity in molecular networks requires dynamics. Sci Signal 2(81):pe44
Kholodenko BN, Kiyatkin A, Bruggeman FJ, Sontag E, Westerhoff HV et al (2002) Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. Proc Natl Acad Sci USA 99:12841–12846

Dynamical System Model Invalidation

▶ Model Invalidation

Dynamical Systems Theory, Asymptotics and Singular Perturbations

John R. King
School of Mathematical Sciences, University of Nottingham, Nottingham, UK

Synonyms

Matched asymptotic expansions; Multiscale and homogenization methods; Regular and singular perturbation methods

Definition

Asymptotic methods comprise a class of systematic mathematical procedures that allow the dominant

processes operating in a given model to be identified, for the model to be reliably simplified by neglecting those effects that are identified as negligible and for the error arising from the simplification to be rather precisely characterized. The techniques are widely applied in the study of differential-equation models but are equally applicable to discrete systems, differential-delay equations, and stochastic models, for example. Typical applications in systems biology include in reducing the complexity of network models (possibly expressing them in modular form) and in integrating across disparate spatial and/or temporal scales in multiscale formulations. The techniques permit the extraction of parameter dependencies from, and more extensive analytical treatment of, the governing models, as well as complementing their numerical study.

Characteristics

The numerous general texts available on asymptotic methods include Murray (1984); Hinch (1991); Holmes (1995); and Kevorkian and Cole (1996). An ordinary differential equation (ODE) model that exemplifies many of the key aspects of singular-perturbation methods as applied in systems biology is one that is widely adopted as a deterministic description of the enzymatic reactions

$$S + E \underset{k_1}{\overset{k_{-1}}{\rightleftharpoons}} C \xrightarrow{k_2} P + E$$

in which a substrate S reacts with an enzyme E to form a complex C that decomposes irreversibly into the enzyme and a product P. Using the same notation for the concentrations of each of these species then gives the ▶ Ordinary Differential Equation (ODE)

$$\begin{aligned} \frac{dS}{dt} &= -k_1 ES + k_{-1} C, \\ \frac{dE}{dt} &= -k_1 ES + (k_{-1} + k_2) C, \\ \frac{dC}{dt} &= k_1 ES - (k_{-1} + k_2) C, \end{aligned} \tag{1}$$

typically subject to the initial conditions

$$S = S_0,\ E = E_0,\ C = 0 \text{ at } t = 0 \tag{2}$$

for prescribed constants S_0 and E_0. An essential first step in the application of asymptotic techniques is to non-dimensionalize the model, thereby reducing the number of parameters present, identifying the relevant parameter groupings and, most importantly in the current context, characterizing the relative importance of these dimensionless groups and hence of the processes that each embodies. The choice of non-dimensionalization is somewhat arbitrary, but it is often convenient to scale out the initial data, i.e., in the case of Eqs. 1 and 2 to set, having noted that $E = E_0 - C$,

$$S = S_0 \hat{S}, \quad C = E_0 \hat{C}, \quad t = \hat{t}/k_1 S_0. \tag{3}$$

This yields

$$\begin{aligned} \frac{d\hat{S}}{dt} &= r\left(-\left(1 - \hat{C}\right)\hat{S} + \kappa_{-1}\hat{C}\right), \\ \frac{d\hat{C}}{d\hat{t}} &= \left(\left(1 - \hat{C}\right)\hat{S} - (\kappa_{-1} + \kappa_2)\right)\hat{C}, \end{aligned} \tag{4}$$

with

$$\hat{S} = 1,\ \hat{C} = 0 \quad \text{at } \hat{t} = 0. \tag{5}$$

Here the number of parameters has been reduced from five in Eqs. 1 and 2 to three dimensionless groupings

$$r = E_0/S_0, \quad \kappa_{-1} = k_{-1}/k_1 S_0, \quad \kappa_2 = k_2/k_1 S_0, \tag{6}$$

the first of which is a concentration ratio, while the other two are ratios of possible timescales. The extent to which the number of parameters can be reduced depends on the problem; a direct application of the Buckingham π theorem gives a lower bound on how many fewer there are in the dimensionless formulation (and the actual value, two, in the above example), while the number of variables to be scaled, e.g., three in the case of Eq. 3, is typically an upper bound.

If reasonable (at least order of magnitude) estimates are available for each of the parameters, the next step is to determine which of the dimensionless groupings is small (or large) and hence available for exploitation in an asymptotic analysis (subtleties can arise in identifying the combination of dimensionless parameters that can be most effectively exploited in this way – see Segel and Slemrod (1989), for example – and

distinguished limits, described below, have a role to play in this regard). In the case of *regular perturbation* problems, a uniformly valid approximation is obtained simply by discarding all terms multiplied by the small parameter. *Singular perturbation* problems are more delicate, since different approximations hold on different scales, thereby capturing the ▶ slow-fast dynamics: this can be illustrated by considering the limit of Eqs. 4 and 5 in which r is small. Then for $\hat{t} = \mathrm{O}(1)$ (the shortest relevant timescale, corresponding to the *inner region* or *boundary layer*) the appropriate approximation to Eq. 4 is, at leading order in r,

$$\frac{d\hat{S}_0}{d\hat{t}} = 0, \quad \frac{d\hat{C}_0}{d\hat{t}} = (1 - \hat{C}_0)\hat{S}_0 - (\kappa_{-1} + \kappa_2)\hat{C}_0, \quad (7)$$

so that $\hat{S}_0 = 1$ and

$$\hat{C}_0 = \left(1 - e^{-(1+\kappa_{-1}+\kappa_2)\hat{t}}\right)/(1 + \kappa_{-1} + \kappa_2). \quad (8)$$

The *outer region* sets $\hat{t} = \bar{t}/r\kappa_{-2}$ (to obtain a balance in the first of Eq. 4), $\hat{S} = \bar{S}$ and $\hat{C} = \bar{C}$ and a different approximation then holds as $r \to 0$, namely,

$$\kappa_2 \frac{d\bar{S}_0}{d\bar{t}} = -(1 - \bar{C}_0)\bar{S}_0 + \kappa_{-1}\bar{C}_0,$$
$$0 = (1 - \bar{C}_0)\bar{S}_0 - (\kappa_{-1} + \kappa_2)\bar{C}_0,$$

whereby the second equation in Eq. 4 is replaced by its quasi-steady approximation (▶ Flux Balance Analysis, for example) leading to a Michaelis-Menten expression (a special case of the ▶ Hill Equation):

$$\frac{d\bar{S}_0}{d\bar{t}} = -\frac{\bar{S}_0}{\kappa_{-1} + \kappa_2 + \bar{S}_0}, \quad \bar{C}_0 = \frac{\bar{S}_0}{\kappa_{-1} + \kappa_2 + \bar{S}_0}. \quad (9)$$

The initial data on Eq. 8 are provided by the *matching condition*

$$\lim_{\bar{t}\to 0} \bar{S}_0(\bar{t}) = \lim_{\hat{t}\to+\infty} \hat{S}_0(\hat{t}) = 1$$

(more generally, however, the matching conditions need not correspond to the original initial conditions). The approximations valid on each of these timescales can be combined into a *uniformly valid* or *composite approximation*.

The *reduced problems* obtained in the above fashion are more amenable to analytical solution (as in Eq. 8) than is the full problem, are numerically better conditioned (their stiffness having been removed), and contain fewer parameters (e.g., Eqs. 7 and 9 each involve only the single grouping $\kappa_{-1} + \kappa_2$), so can more readily be fitted to experimental data. It is noteworthy in particular that for such purposes it may not be necessary to have any quantitative information about the small parameter, beyond that it is indeed small. In the above, only *leading order* approximations have been given – more accurate expressions can be obtained by expanding the solutions in each region in powers of the small parameter and balancing the terms that involve the same power (in more complicated examples, terms depending on the logarithm of the small parameter may also need to be introduced in order to match successfully etc.).

The above assumes the problem to contain a single small (or large) parameter. In many applications (and almost invariably in the case of complex systems, such as are common in the modeling of ▶ gene regulatory networks and of ▶ metabolic and signaling networks) multiple such parameters are present; typically one of these would be identified as that in which the expansions are to be performed and the relative sizes of the others are then characterized by expressing them in terms of powers of this small parameter; there can, however, be ambiguity in how this is accomplished. *Distinguished limits*, in which such characterizations are identified in part on mathematical grounds as giving the fullest set of relevant balances within the equations, can then be of particular value, as they can also be (because of their broad validity) when limited information is available about the sizes of some of the parameters.

The discussion above pertains to circumstances in which the method of matched *asymptotic expansions* applies. Another broad set of techniques (*multiple scales*, having *two-timing* as a special case) was originally developed largely in the analysis of nonlinear oscillations (whereby rapid oscillations are subject to a slow rate of decay, for example: since the oscillations persist, the fast and slow scales must be captured for all times, rather than the former being relevant only in the boundary layer); here *secularity conditions*, instead of matching, play a central role. This class of techniques is of particular importance in integrating between scales (*homogenization* – see Mei and Vernescu (2010), for

instance), having the potential within systems biology to embed cell-scale behavior systematically within tissue-level models, say (cf. ▶ mixed and multi-level models).

Since the above example is an ODE system, it should be emphasized that the techniques are equally effective in the analysis of, *inter alia*, ▶ partial differential equation (PDE) models, differential-delay equations, discrete systems, and stochastic models (cf. van Kampen (2007), for example, reduction of the ▶ master equation to a ▶ Fokker-Planck equation being a common upshot) and may be of value in underpinning the development of a ▶ mean-field approximation. Given the disparate timescales that almost invariably arise in such applications, they can be of particular effectiveness in simplifying a complex ▶ biological network model containing numerous distinct pathways and may then provide systematic approaches to ▶ modularity-based network decomposition.

The methodologies in question remain areas of active research and issues that go beyond those described above include the following.

(a) Asymptotics *beyond all orders*. In certain applications, the calculation of algebraic terms (those involving powers of the small parameter) does not suffice in the sense that important phenomena manifest themselves only in terms that are exponential small: these terms are "hidden" beyond a (divergent) algebraic series and hence lead to additional complexities; a particular application is to the failure of signal propagation in spatially discrete systems – see King and Chapman (2001) and references therein.

(b) *Intermediate asymptotics* (Barenblatt 1996). An independent variable, rather than one of the dimensionless constants, can be taken to be the small or large quantity, the analysis of large-time behavior (such as traveling-wave propagation in Fisher's equation) being a common such application.

Cross-References

▶ Biological Network Model
▶ Flux Balance Analysis
▶ Fokker–Planck Equation
▶ Gene Regulatory Networks
▶ Hill Equation
▶ Master Equation
▶ Mean-Field Approximation
▶ Metabolic and Signaling Networks
▶ Mixed and Multi-level Models
▶ Modularity-Based Network Decomposition
▶ ODE: Ordinary Differential Equation
▶ Partial Differential Equation (PDE), Models
▶ Slow-Fast Dynamics

References

Barenblatt GI (1996) Scaling, self-similarity and intermediate asymptotics: dimensional analysis and intermediate asymptotics. Cambridge University Press, Cambridge
Hinch EJ (1991) Perturbation methods. Cambridge University Press, Cambridge
Holmes MH (1995) Introduction to perturbation methods. Springer, New York
Kevorkian JK, Cole JD (1996) Multiple scale and singular perturbation methods. Springer, New York
King JR, Chapman SJ (2001) Asymptotics beyond all orders and stokes lines in nonlinear differential-difference equations. Eur J Appl Math 12:433–463
Mei CC, Vernescu B (2010) Homogenization methods for multiscale mechanics. World Scientific, Singapore
Murray JD (1984) Asymptotic analysis. Springer, New York
Segel LA, Slemrod M (1989) The quasi-steady-state assumption: a case study in perturbation. SIAM Rev 31:446–447
van Kampen NG (2007) Stochastic processes in physics and chemistry. Elsevier, Amsterdam

Dynamical Systems Theory, Bifurcation Analysis

Alan Champneys and Krasimira Tsaneva-Atanasova
Department of Engineering Mathematics, University of Bristol, Bristol, UK

Synonyms

Numerical continuation; Parameter studies; Path-following; Qualitative analysis

Definition

Bifurcation theory refers to the study of qualitative changes to the state of a system as a parameter is varied.

It can be applied to ▶ steady state systems, or to dynamical systems and can be understood best at the level of a mathematical model, although recent techniques allow the method to be applied to experiments with feedback control. Typically the theory is applied to a ▶ continuous model, but can also be used in ▶ discrete models and mathematics, difference equations. There are dedicated numerical implementations of bifurcation theory using path-following, or numerical continuation. There is a distinction between a local ▶ bifurcation, which can be understood in terms of a change to the number or stability of simple steady states, and a global bifurcation, which cannot. Often global bifurcations cause catastrophic changes to the ▶ attractor of the system. Typical local examples are the ▶ Hopf bifurcation which leads to the onset of oscillation and the ▶ saddle-node bifurcation where a stable steady state is created or destroyed, often leading to bistability.

Characteristics

Once a model of a biological system has been constructed and, consequently, the number of parameters in the model carefully defined and evaluated – perhaps using ▶ optimization and parameter estimation – one may wonder how the long-term behavior of a dynamical system changes when these parameters are changed. This question is the basis of bifurcation theory and it underlies a qualitative understanding of many biological processes and transitions, such as the onset of oscillation, switching, morphogenesis, multi-stability, emergence, and localization.

Bifurcation theory can be applied to a wide variety of deterministic models and processes, including ▶ partial differential equation (PDE) models and ▶ dynamical systems theory, delay differential equations but for simplicity this entry shall consider only the context of ▶ dynamical systems theory, ordinary differential equations. Specifically, consider such a parametrized ▶ ODE model written in state-space form:

$$x = f(x, \mathrm{A}), \tag{1}$$

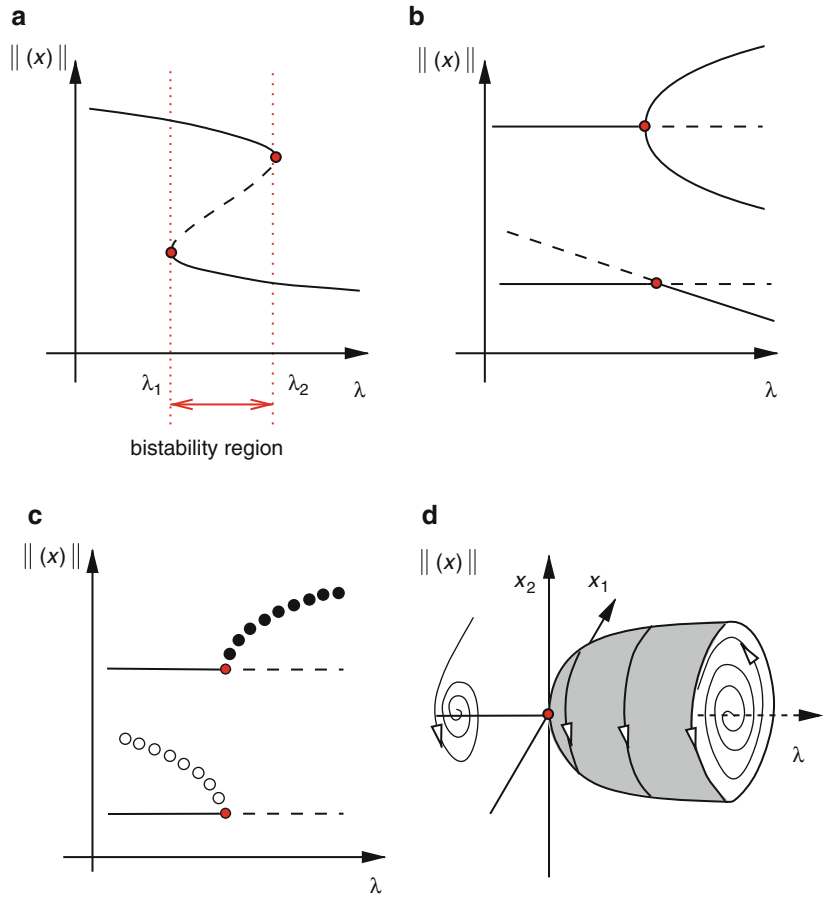
where $x \in \mathrm{R}^n$ is the set of states of the system, $\lambda \in \mathrm{R}^p$ is a parameter set, and a dot denotes differentiation with respect to time.

In contrast with many equations arising in physics, engineering, and economics, a key feature of most biological models of the form (Danino et al. 2010) is that they are nonlinear, essentially because of large deformations, excitability, thresholding and interaction processes governed by the ▶ law of mass action. Nonlinear systems can have multiple ▶ stable steady states, or ▶ attractors, that can have many different stable regions, or *basins of attraction* in state space. They can also often support ▶ limit cycle periodic oscillations. Moreover, as a parameter is varied, a new attractor can emerge out of "thin air," typically when an unstable state gains local stability. For this reason, usual numerical methods and computer simulation for ordinary differential equations (▶ Partial Differntial Equations, Numerical Methods and Simulations; ▶ Ordinary Differential Equation (ODE), Model) can be highly unreliable in understanding the true dynamics of the system, if the method is based on simulation from fixed initial conditions. Bifurcation theory can be especially useful in this context as it enables the tracing of paths of unstable states as parameters vary and hence determine precisely the transitions (or "bifurcation points") at which qualitatively distinct stable behavior emerges.

The *codimension* of a bifurcation is defined as the number of parameters required to observe a bifurcation in a structurally stable way. So a codimension-one bifurcation can be observed at an isolated value of a single parameter, whereas a codimension-two bifurcation would typically only be seen at an isolated point in a two-parameter diagram. There is a distinction drawn between *local bifurcations* that can be understood in terms of loss of ▶ stability of a simple state such as an equilibrium or a limit cycle and *global bifurcations* that cannot. Often global bifurcations involve rearrangements of *stable and unstable manifolds* of other simple states, such as in *homoclinic bifurcations*.

A codimension-one bifurcation can often be represented in a *bifurcation diagram* that depicts a measure, or norm, of a system state against a single parameter; see Fig. 1 for examples. Codimension-one bifurcations can also be used to divide regions in a parameter plain in which qualitatively distinct bifurcations can occur.

A comprehensive treatment of local bifurcations of codimension-one and two, and many examples of global bifurcations can be found in Kuznetsov (2004). That book also contains analytical and



Dynamical Systems Theory, Bifurcation Analysis, Fig. 1 Schematic bifurcation diagrams depicting codimension-one local bifurcations. (**a**) An *s*-shaped fold featuring two saddle-node bifurcations (at parameter values λ_1 and λ_2) and a consequent parameter interval between these two values in which bistability is observed. In this and subsequent panels, *solid lines* represent paths of stable steady states and dashed lines unstable steady states. Also $\|x\|$ represents a characteristic norm or scalar measure of the vector state x. (**b**) A supercritical pitchfork bifurcation (*upper plot*) and a transcritical bifurcation (*lower plot*). (**c**) A supercritical Hopf bifurcation (*upper plot*) and a subcritical Hopf bifurcation (*lower plot*). Here, a curve composed of *solid circles* represents a path of stable limit cycle oscillations, whereas a curve of *open circles* represents a path of unstable limit cycles. (**d**) A representation of a supercritical Hopf bifurcation in state and parameter space

numerical techniques for analyzing bifurcations in practical examples, principally using the theory of center manifolds and normal forms. That theory can be seen as a counterpart to more traditional ▶ dynamical systems theory, asymptotics, and singular perturbations. The qualitative geometry or *topology* of bifurcations is also stressed by Shilnikov et al. (2001). Many examples in biological systems, especially in ▶ reaction-diffusion-advection equations are found in Murray (2007), and applications to cell biology in Fall et al. (2002). A more elementary introduction to bifurcation theory and nonlinear dynamical systems theory, stability analysis in general can be found in Strogatz (1994). For more on numerical techniques for performing parameter continuation and bifurcation analysis, see Krauskopf et al. (2007).

Rather than reiterate this theory, this entry shall try to give a qualitative flavor to how bifurcation theory can underlie several key phenomena in biological systems. Specifically treated are threshold behavior, oscillation, bi- and multi-stability, synchronization and emergence of collective behavior, before some final remarks on bifurcation theory applied directly to experiments.

Threshold

It is common in biological systems for a threshold concentration of some chemical signal to be there in order for certain behavior to be triggered. Thresholds can often be understood in terms of the phenomenon of *excitability*, which in itself is a property of a dynamical system with two timescales, where a transient pushes the system beyond the region where a large excitation occurs. In systems that reach steady state though, the variation of a parameter can cause a global bifurcation that makes the large excitation occur.

Examples:

- Control of calcium oscillations, see for example (Sneed and Keener 2008)
- ▶ Cell cycle model analysis, bifurcation theory, which is a canonical example of the practical applicability of bifurcation theory in explaining how biological decisions occur as emergent bifurcation events upon integrating the various environmental and internal parameters

Oscillation

Periodic oscillations are important in biology, appearing in diverse areas such as ▶ Circadian rhythms, metabolic networks and their evolution, heart beats, cell signaling, etc. Oscillations can be described by simple harmonic motion, governed by second-order linear ODEs, but such descriptions suffer from several serious limitations. First, they are not robust, as a small amount of damping destroys the oscillation. Second, ▶ oscillation amplitude and phase depends on the initial condition. Finally, the frequency is typically not adaptable. In contrast, stable limit cycles of nonlinear models are robust, because following a small ▶ perturbation away from the cycle, the system will return to the cycle by itself. Also, if the dynamics changes a little, a limit cycle will still exist, close to the original one.

The canonical way to generate oscillation from a system that is otherwise at rest is via a ▶ Hopf bifurcation. These oscillatory instabilities can occur in two ways (see Fig. 1c), either as a ▶ supercritical bifurcation in which a stable limit cycle is created at small amplitude, or as a ▶ subcritical bifurcation, often accompanied by bistability which would cause the jump to a fully formed large-amplitude attractor.

Examples:

- The ▶ Goodwin oscillator is a simplified model of circadian rhythms based on a negative ▶ feedback loop.
- The ▶ repressilator and oscillating network is a simple ▶ network motif constructed by ▶ modeling and simulation: synthetic models and methods. It models a ▶ gene regulatory network composed of three genes which mutually repress each other in sequence according to a ring structure.
- Spatiotemporal regulation of extracellular signal-regulated kinase has been shown to result in rapid and sustained nuclear-cytoplasmic oscillations (Shankaran et al. 2009).

Bistability and Multi-stability

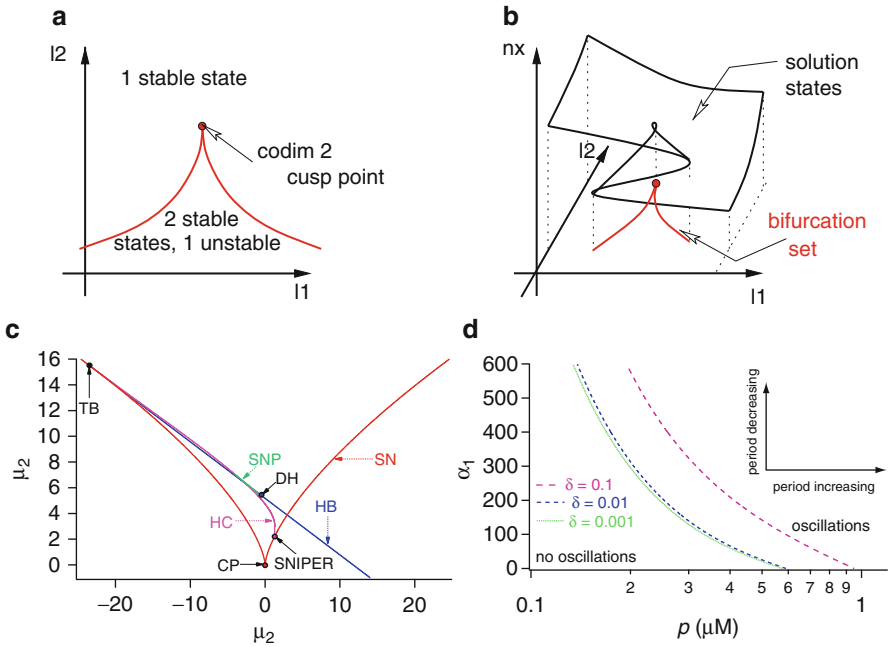
Bistability refers to the existence of two distinct attractors to which the system may evolve given different initial states or perturbations. A typical bifurcation scenario in which bistability occurs is via a pair of ▶ saddle-node bifurcations that are connected in an S-shaped fold, see Fig. 1a. Such folded structures are often seen as part of the the unfolding of a codimension-two *cusp bifurcation* point (see Fig. 2a, b) which is one of the elementary *catastrophes* of singularity theory.

Bistability also often accompanies subcritical bifurcations, where an extra fold causes the unstable bifurcating branch to turn around, become stable, and coexist with the primary branch, again see Fig. 1.

Multi-stability refers the situation when there are more than two competing attractors, each with distinct basins of attraction, which can occur via a sequence of bifurcations, or directly via a global bifurcation, such as that caused by a *Shilnikov-type homoclinic bifurcation*.

Examples:

- The *Toggle switch* (Gardner et al. 2000) is a synthetic biology construct that matches what is believed to be a common ▶ network motif in systems biology. It is composed of two genes that mutually repress each other, which causes bistability between steady states in which either one gene or the other is expressed at high levels.
- Multi-stability is also seen in unusual cell-division phenotypes in ▶ cell cycle model analysis, bifurcation theory and in recent systems biology models of tumors as competing attracting states in cancer dynamics (Huang et al. 2009).



Dynamical Systems Theory, Bifurcation Analysis, Fig. 2 Examples of two-parameter bifurcation diagrams indicating parameter regions in which qualitatively distinct behavior is observed. (**a**) A cusp bifurcation point linking two curves of saddle-node bifurcations. (**b**) A representation of the cusp in 3D showing a norm of the solution state on the vertical axis. (**c**) Codimension-two bifurcations in a simple model for plateau bursting in excitable systems that arises in the unfolding of a certain degenerate codimension-three bifurcation point; see (Golubitsky et al. 2001). Here CP represents a cusp bifurcation point, TB a Takens-Bogdanov point (where Hopf and saddle-node combine), DH a degenarate Hopf bifurcation (a transition between super and subcritical cases), SNIPER represents a Saddle Node of Infinite PERiod bifurcation point (a kind of global bifurcation), SN a saddle node of equilibria, SNP a saddle node of periodic orbits, HB a Hopf bifurcation, and HC a homoclinic bifurcation. (**d**) Two-parameter bifurcation diagram of an open-cell model for calcium oscillations (Sneed and Keener 2008). The lines (for three different values of a parameter δ) represent Hopf bifurcation curves in the parameter plane and separate regions in which oscillations do and do not exist

Synchronization and Emergence of Behavior

Many biological systems composed of near identical cells, components, or organisms are known to undergo common collective dynamics. The simplest such state is that of synchronization, where each state oscillates, typically periodically perfectly in time with the others. This is an example of a symmetric state of a dynamical system and the onset or loss of synchronicity can be understood as a form of symmetry breaking bifurcation. Other collective states include spatially localized patterns of either steady state or dynamic behavior. For more on the bifurcation theory of pattern formation systems, see (Hoyle 2006).

Example:

- Collective oscillations in proliferating bacterial population (Danino et al. 2010).

Experimental Bifurcation Theory

Recently, the possibility of performing bifurcation analysis directly in ▶ feedback controlled experiments has raised (Sieber et al. 2008). This poses the possibility of direct intervention in vitro or in vivo in order to analyze, or indeed to influence and control bifurcations to desirable or undesirable states as an external or internal parameter is varied. Such technology is likely to have a significant impact on synthetic biology and personalized medicine.

Cross-References

- ▶ Attractor
- ▶ Bifurcation
- ▶ Bifurcation, Supercritical and Subcritical

- Cell Cycle Model Analysis, Bifurcation Theory
- Circadian Rhythm
- Continuous Model
- Differential-Difference Equations
- Dynamical Systems Theory, Asymptotics and Singular Perturbations
- Dynamical Systems Theory, Delay Differential Equations
- Feedback Regulation
- Gene Regulatory Networks
- Goodwin Oscillator
- Hopf Bifurcation
- Law of Mass Action
- Limit Cycle
- Metabolic Networks, Evolution
- Network Motif
- Partial Differntial Equations, Numerical Methods and Simulations
- Optimization and Parameter Estimation, Genetic Algorithms
- Ordinary Differential Equation (ODE)
- Oscillation Amplitude
- Partial Differential Equation (PDE), Models
- Periodic Oscillation
- Perturbation
- Reaction-Diffusion-Advection Equation
- Repressilator and Oscillating Network
- Saddle-Node Bifurcation
- Stability
- Stability, States and Regions
- Steady State

References

Danino T, Mondragn-Palomino O, Tsimring L, Hasty J (2010) A synchronized quorum of genetic clocks. Nature 463(7279):326–330

Fall CP, Marland ES, Wagner JM, Tyson JJ (2002) Computational cell biology. Springer, New York, In memory of Joel Keizer

Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in *Escherichia coli*. Nature 403(6767):339–342

Golubitsky M, Josic K, Kaper TJ (2001) An unfolding theory approach to bursting in fast-slow systems. Global Analysis of Dynamical Systems. Festschrift dedicated to Floris Takens for his 60th birthday, pp 277–308

Hoyle R (2006) Pattern formation: an introduction to methods. Cambridge University Press, Cambridge

Huang S, Ernberg I, Kauffman S (2009) Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. Semin Cell Dev Biol 20(7):869–876

Krauskopf B, Osinga HM, Galn-Vioque J (2007) Numerical continuation methods for dynamical systems: path following and boundary value problems. Springer, New York

Kuznetsov YA (2004) Elements of applied bifurcation theory, 3rd edn. Springer, New York

Murray JD (2007) Mathematical biology. Springer, New York, third (in two parts) edition

Shankaran H, Ippolito DL, Chrisler WB, Resat H, Bollinger N, Opresko LK, Wiley HS (2009) Rapid and sustained nuclear-cytoplasmic erk oscillations induced by epidermal growth factor. Mol Syst Biol 5:332

Shilnikov LP, Shilnikov AL, Turaev DV, Chua LO (2001) Methods of qualitative theory in nonlinear dynamics. World Scientific, Singapore

Sieber J, Gonzalez-Buelga A, Neild SA, Wagg DJ, Krauskopf B (2008) Experimental continuation of periodic orbits through a fold. Phys Rev Lett 100:244101

Sneed J, Keener J (2008) Mathematical physiology. Springer, New York, second, in two parts edition

Strogatz SH (1994) Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering. Westview, Boulder

Dynamical Systems Theory, Delay Differential Equations

Patrick Nelson
CCMB 2017 Palmer Commons,
University of Michigan, Ann Arbor, MI, USA

Synonyms

Differential equations with deviating arguments; Differential-difference equations; Functional differential equations; Time delay

Definition

Delay differential equations (DDE) are equations whose solution depends on not just a single initial condition at time, $t = t_0$, but also on the past history of the system. DDEs can be classified as retarded or

Dynamical Systems Theory, Delay Differential Equations, Table 1 Summary of Research on DDEs

Topic	Problem formulation	Approach	Applications
Nonlinear	$\dot{y}(t) = f(y(t), y(t-\tau), u(t))$	Perturbation method with Lambert function	HIV, chatter
Multiple delays	$\dot{y}(t) + \sum_{i=1}^{N} A_i y(t-\tau_i) + By(t) = u(t)$	Superposition or modified Lambert function	HIV, multiple regenerative effect in chatter
Time-varying	$\dot{y}(t) + A(t)y(t-\tau) + B(t)y(t) = u(t)$	Floquet theory, Wronskian matrix with Lambert function	Milling chatter

neutral and continuous or discrete. In general, a discrete delay differential equation can be written as

$$\frac{d^n y}{dt^n} = f(t, a_0(t)y(t), a_1(t)y'(t), \ldots a(t)_{n-1}y^{n-1}(t), \\ b_0(t)y(t-\tau)\ldots b_{n-1}(t)y^{n-1}(t-\tau)) \quad (1)$$

where τ represents a point of time in the past history of the equation. Note, when $b_i = 0$ for all i that the system is just an ordinary differential equation. However, when $b_i = 0$ for $i > 0$ then the system is defined as a retarded differential equation. If $b_i \neq 0$ for any $i > 0$ the system is defined to be of neutral type. Equation 1 is an example of a DDE with only one time delay, $t = \tau$ but there can be multiple time delays, $\tau_1, \tau_2, \ldots, \tau_n$, in the system. Most DDEs in physics, engineering, and biology are of the discrete type with one time delay. Another representation for a DDE is defined as a continuous delay differential equation.

$$\frac{dy}{dt} = f\left(t, y(t), \int_0^\infty y(t-\tau)g(\tau)d\tau\right) \quad (2)$$

Note that when one considers the distribution function $g(\tau)$ to be a gamma distribution (more precisely an Erlang distribution), the system can reduce to the discrete delay type.

Characteristics

Delays are inherent in many physical, biological, economic, and engineering systems. The first appearance of DDEs was in a paper by Kondorse in 1777 but their use did not become popular until the early 1940s and 1950s when Nyquist, Chebotarev, and Pontryagin pioneered work investigating the stability of DDEs (Pontryagin 1955; Krall 1965). More recently, the advances into the theory, solution, and application of DDEs has been led by Bellman, Cooke, Hale, Kuang, and Stepan (Bellman and Cooke 1963; Stepan 1989; Hale 1977; Corless et al. 1996). Time delays can be used to represent self-oscillating systems, economic futures, and in engineering, pure delays are often used to ideally represent the effects of transmission, transportation, and inertial phenomena. They are also quite popular in biology, where they can be used to model gestation, maturation, transcription, and numerous cell-cycle phenomena. Delay differential equations constitute basic mathematical models for such real phenomena. However, there is a principal difficulty in studying DDEs hidden in their special transcendental character which leads to an infinite spectrum of frequencies. In other words, all DDEs will result in an infinite number of eigenvalues. Hence, they are often solved using numerical methods, asymptotic solutions, approximations (e.g., Padé) and graphical approaches or through the study of bifurcation of their characteristic equations to determine their stability.

Characteristic Equation (zeros of the transcendental equation) The characteristic equation for a discrete time delay equation at steady state, with a single delay ((Eq. 1), with $b_i = 0$ for $i > 1$) can be written as

$$P_1(\lambda) + P_2(\lambda)e^{-\lambda\tau} = 0 \quad (3)$$

where P_1 and P_2 are functions of the eigenvalues, λ of the equation. $P_1 + P_2 = 0$ represent the characteristic equation for the system when $\tau = 0$. For the equation given by Eq. 2, the characteristic equation looks like

$$P_1(\lambda) + P_2(\lambda)F(\lambda) = 0 \quad (4)$$

where the P_i are the same as in Eq. 3 but instead of the explicit exponential term, $e^{\lambda\tau}$, we find, $F(\lambda)$ to be the Laplace transform of the delay kernel, defined as $F(\lambda) = \int_0^\infty g(\tau)e^{-\lambda\tau}$

The basic idea behind studying the characteristic equation is to determine conditions on the parameters and the time delay, τ, that causes a bifurcation in stability of the system. One starts by looking at the conditions for stability when $\tau = 0$ (all roots of characteristic equation lie in the left half plane) and then consider $\tau > 0$ to find when the roots of the characteristic equation cross the imaginary axis (El'sgol'ts and Norkin 1973). As τ varies, these roots change. Of interest is any critical values of τ at which a root of this equation transitions from having negative to having positive real parts. If this is to occur, there must be a boundary case, a critical value of τ, such that the characteristic equation has a purely imaginary root.

Early stability methods developed and presented in the classic papers of Pontryagin (1955) and Nyquist (Krall 1965) have been used for many years to study bifurcations in transcendental equations. However, these methods rely heavily on the principal of the argument for determining where the poles of the transcendental equations are located. In other words, they use geometric principles to determine the number of roots of these equations. The monograph by Chebotarev and Meiman (1949) shows how to extend the Routh-Hurwitz criteria for polynomials to quasi-polynomials. However, it has been noted that the application of the Chebotarev criterion as an analytical tool is not effective practically. Recent results using Sturm sequences (Forde and Nelson 2000) relax the need for the application of the argument principle and provides an analytical criterion that is practical to use. The Sturm sequence provides an algorithm for determining stability of low degree, i.e., less than degree 4, polynomials and is explained by the following.

Given a system of differential equations $\frac{dy}{dt} = f(y(t), y(t-\tau))$ with a discrete delay τ, and a stable steady state, y_s, for $\tau = 0$, will lead to

$$\sum_{i=1}^{N} a_i \lambda^i + e^{-\lambda\tau} \sum_{i=1}^{M} b_i \lambda^i = 0$$

as the characteristic equation of the system about y_s. Then there exists a $\tau^* > 0$ for which y_s undergoes a nondegenerate change of stability if and only if the equation

1. $S(\mu) = 0$ (defined by substituting $\lambda = \mu + iv$ into the characteristic equation, separating the real and imaginary parts and squaring both sides. Then allowing $\mu = v^2$ and defining the resulting equation as $S(\mu)$) has a positive real root $\mu^* = (v^*)^2$, such that
2. $S'(\mu^*) \neq 0$

Example

$$\lambda^2 + a\lambda + b + (c\lambda + d)e^{-\lambda\tau} = 0. \tag{5}$$

A steady state with this characteristic is stable for $\tau = 0$ if all of the roots of

$$\lambda^2 + (a + c)\lambda + (b + d) = 0$$

have negative real part. By the Routh-Hurwitz conditions, this occurs if and only if $a + c > 0$ and $b + d > 0$. Letting $\lambda = iv$ we arrive at the following form of equation

$$S(\mu) = \mu^2 + (a^2 - c^2 - 2b)\mu + (b^2 - d^2) = 0. \tag{6}$$

Let $A \equiv a^2 - c^2 - 2b$ and $B \equiv b^2 - d^2$. Equation 6 has a positive real root in two circumstances. Clearly, since the lead coefficient is positive, if $B < 0$ then there is a positive real root. If $B > 0$, the roots of Eq. 6 are

$$\frac{-A \pm \sqrt{A^2 - 4B}}{2},$$

and there is a simple positive root if and only if $A < 0$. Thus one can conclude

Proposition 1. *A steady state with characteristic* (Eq. 5) *is stable in the absence of delay, and becomes unstable with increasing delay if and only if*

1. $a + c > 0$ *and* $b + d > 0$
2. *Either* $b^2 < d^2$, *or* $b^2 > d^2$ *and* $a^2 < c^2 + 2b$.

Analytical Solutions

Finding analytical solutions for DDEs is difficult and most of the theory to date focuses on computation methods for solutions or methods for determining stability via computation or analytics. However, a group at the University of Michigan (Asl and Ulsoy 2003; Yi et al. 2007) recently developed an analytic approach, based on the matrix Lambert function, for the complete solution of a system of linear constant coefficient DDEs. This method can be applied to study eigenvalue

assignment, pole placement, controllability and observability, and time-varying coefficients. The method has been validated in engineering problems where delay is significant, e.g., regenerative chatter in a machining operation on a lathe and a biological problem, e.g., control of drug therapies. The matrix Lambert function-based solution approach for DDEs is analogous to the use of the matrix exponential for the free and forced solution of linear constant coefficient ordinary differential equations. Systems with multiple time delays and nonlinearities arise quite naturally in engineering and biology and yet little attention has been paid to their analyses. To explain, look at a second-order linear system of DDEs with state space given by $\vec{x}$.

$$\mathbf{x}(t) + \mathbf{A}\mathbf{x}(t) + \mathbf{A_d}\mathbf{x}(t-\tau) = 0. \tag{7}$$

$\mathbf{A}$ and $\mathbf{A_d}$ are the linearized coefficient matrices and are functions of the dynamical system. The analytical method to solve scalar DDEs, and systems of DDEs using the matrix Lambert W function was introduced by Asl and Ulsoy (2003) and extended by Yi (2007) to obtain the solution of general systems of DDEs in matrix-vector form. First assume a solution form for Eq. 7 as

$$\mathbf{x}(t) = e^{\mathbf{S}t}\mathbf{x}_0, \tag{8}$$

where $\mathbf{S}$ is $n \times n$ matrix. In the usual case, the characteristic equation for Eq. 7 is obtained from the equation by looking for nontrivial solution of the form $e^{st}\mathbf{C}$ where s is a scalar variable and $\mathbf{C}$ is constant (Hale 1977). However, such an approach can neither lead to any interesting result nor help in deriving a solution to systems of DDEs in Eq. 7. Alternatively, one could assume the form of Eq. 8 to derive the solution to systems of DDEs in Eq. 7 using the matrix Lambert W function. Substituting into Eq. 7 yields

$$\mathbf{S}e^{\mathbf{S}t}\mathbf{x}_0 + \mathbf{A}e^{\mathbf{S}t}\mathbf{x}_0 + \mathbf{A_d}e^{\mathbf{S}(t-T)}\mathbf{x}_0 = \mathbf{0}, \tag{9}$$

and using the property of the exponential

$$e^{\mathbf{S}(t-T)} = e^{\mathbf{S}(-T+t)} = e^{\mathbf{S}(-T)}e^{\mathbf{S}t} \tag{10}$$

one can rewrite as

$$\mathbf{S}e^{\mathbf{S}t}\mathbf{x}_0 + \mathbf{A}e^{\mathbf{S}t}\mathbf{x}_0 + \mathbf{A_d}e^{-\mathbf{S}T}e^{\mathbf{S}t}\mathbf{x}_0 = (\mathbf{S} + \mathbf{A} + \mathbf{A_d}e^{-\mathbf{S}T})e^{\mathbf{S}t}\mathbf{x}_0 = \mathbf{0}. \tag{11}$$

Because the matrix $\mathbf{S}$ is an inherent characteristic of a system and independent of initial condition, we can conclude that for Eq. 11 to be satisfied for any arbitrary initial condition, $\mathbf{x}_0$, and every time, t, we must have

$$\mathbf{S} + \mathbf{A} + \mathbf{A_d}e^{-\mathbf{S}T} = \mathbf{0}. \tag{12}$$

In the special case that $\mathbf{A_d} = 0$, the delay term in Eq. 7 disappears, Eq. 7 becomes ODE, and Eq. 12 is

$$\mathbf{S} + \mathbf{A} = \mathbf{0} \Leftrightarrow \mathbf{S} = -\mathbf{A}. \tag{13}$$

Then, substitution into Eq. 8 yields

$$\mathbf{x}(t) = e^{-\mathbf{A}t}\mathbf{x}_0 \tag{14}$$

This is the typical solution to ODE in terms of the matrix exponential. Multiply $Te^{\mathbf{S}T}\,e^{\mathbf{A}T}$ on both sides of Eq. 12 and rearrange to obtain

$$T(\mathbf{S} + \mathbf{A})e^{\mathbf{S}T}e^{\mathbf{A}T} = -\mathbf{A_d}Te^{\mathbf{A}T}. \tag{15}$$

In the general case, when the matrices $\mathbf{A}$ and $\mathbf{A_d}$ do not commute, neither do $\mathbf{S}$ and $\mathbf{A}$; thus

$$T(\mathbf{S} + \mathbf{A})e^{\mathbf{S}T}e^{\mathbf{A}T} \neq T(\mathbf{S} + \mathbf{A})e^{(\mathbf{S}+\mathbf{A})T}. \tag{16}$$

Consequently, to adjust the inequality in Eq. 16 and to take advantage of the property of the matrix Lambert W function defined by

$$\mathbf{W}(\mathbf{H})e^{\mathbf{W}(\mathbf{H})} = \mathbf{H}, \tag{17}$$

we introduce an unknown matrix $\mathbf{Q}$ so that satisfies,

$$T(\mathbf{S} + \mathbf{A})e^{(\mathbf{S}+\mathbf{A})T} = -\mathbf{A_d}T\mathbf{Q}. \tag{18}$$

Comparing Eqs. 17 and 18 we note that

$$(\mathbf{S} + \mathbf{A})T = \mathbf{W}(-\mathbf{A_d}T\mathbf{Q}). \tag{19}$$

Then from Eq. 19, solving for $\mathbf{S}$ gives

$$\mathbf{S} = \frac{1}{T}\mathbf{W}(-\mathbf{A_d}T\mathbf{Q}) - \mathbf{A}. \tag{20}$$

Substituting Eq. 20 into Eq. 15 yields the following condition, which can be used to solve for the unknown matrix $\mathbf{Q}$:

$$\mathbf{W}(-\mathbf{A}_d T\mathbf{Q})e^{\mathbf{W}(-\mathbf{A_d}T\mathbf{Q})-\mathbf{A}T} = -\mathbf{A_d}T. \tag{21}$$

To date, many examples have been studied and Eq. 21 always has a unique solution $\mathbf{Q}_k$ for each branch, k. However, a proof of this result is needed. The solution is obtained numerically, for a variety of initial conditions, using the "fsolve" function in Matlab. The matrix Lambert W function defined in Eq. 17 contains an infinite number of branches (Corless et al. 1996). Corresponding to each branch, k ($= -\infty, \ldots, -1, 0, 1, \ldots, \infty$), of the Lambert W function, for $\mathbf{H}_k = -\mathbf{A_d}T\mathbf{Q}_k$, we compute the eigenvalues $\hat{\lambda}_{ki}$, $i = 1, 2$, of $\mathbf{H}_k$ and the corresponding eigenvector matrix $\mathbf{V}_k$. Hence, the matrix Lambert W function is

$$\mathbf{W}_k(\mathbf{H}_k) = \mathbf{V}_k \begin{bmatrix} W_k(\hat{\lambda}_{k1}) & 0 \\ 0 & W_k(\hat{\lambda}_{k2}) \end{bmatrix} \mathbf{V}_k^{-1}. \tag{22}$$

Finally, $\mathbf{S}_k$ is computed corresponding to $\mathbf{W}_k$ from Eq. 20 and summated to be the solution to the systems of DDEs Eq. 7 as

$$\mathbf{x}(t) = \sum_{k=-\infty}^{\infty} e^{\mathbf{S}_k t}\mathbf{C}_k \tag{23}$$

where the $\mathbf{C}_k$ is a 2×1 coefficient matrix computed from a given preshape function $\mathbf{x}(t) = \mathbf{g}(t)$, which is initial state of DDEs Eq. 7, for $t \in [-T, 0]$ (Yi et al. 2007).

Each branch of the Lambert W function can be computed analytically as shown in Corless et al. 1996), and one of the merits of the matrix Lambert W function approach is that one can compute all of the branches of the function using commands already embedded in the various commercial software packages, such as Matlab, Maple, and Mathematica.

The following table provides an overview of where the current state of studying DDEs is.

Cross-References

- ▶ Dynamical Systems Theory, Asymptotics and Singular Perturbations
- ▶ Mathematical Model, Model Theory
- ▶ Systems Network in HIV

References

Asl FM, Ulsoy AG (2003) Analysis of a system of linear differential equations. ASME J Dyn Syst Meas Control 125:215–223

Bellman R, Cooke K (1963) Delay differential equations. Academic, New York/London

Chebotarev NG, Meiman NN (1949) The Routh-Hurwitz problem for polynomials and entire functions. Trudy Mat Inst Steklov 26

Corless RM et al (1996) On Lambert's W function. Adv Comput Math 5:329–359

El'sgol'ts LE, Norkin SB (1973) An introduction to the theory and application of differential equations with deviating arguments. Academic, New York

Forde J, Nelson P (2000) Applications of Sturm sequences to bifurcation analysis of delay differential equation models. J Math Anal Appl 300:273–284

Hale JK (1977) Theory of functional differential equations. Spring, New York

Krall AM (1965) Stability criteria for feedback systems with a time lag. SIAM J Control 2:160–170

Kuang Y (1993) Delay differential equations with applications to population biology. Academic, Boston

Pontryagin LS (1955) On the zeros of some elementary transcendental functions. Am Math Soc Transl 2:95–110

Stepan G (1989) Retarded dynamical systems: stability and characteristic functions. Longman Scientific and Technical, Burnt Mill

Yi S, Ulsoy AG, Nelson PW (2007) Delay differential equations via the matrix Lambert W function and bifurcation analysis: application to machine tool chatter. Math Biosci Eng 4:1–12

Dynamics Modeling

- ▶ Lymphocyte Dynamics and Repertoires, Modeling

Dysplasia

Barbara J. Davis
Section of Pathology, Tufts Cummings School of Veterinary Medicine Biomedical Sciences, North Grafton, MA, USA

Definition

Dysplasia is an excessive, disorderly grown tissue in which cells "abnormally increase in number," lose uniformity and polarity, and may either stop proliferating after cessation of the stimulus that evoked the growth or progress to neoplasia.

Cross-References

► Cancer Pathology