N

Named Entity

Jörg Hakenberg
Department of Computer Science and Department of Biomedical Informatics, Arizona State University, Tempe, AZ, USA

Definition

A named entity is an atomic element in a text that is assigned to a predefined class such as person, location, date and time, disease, or protein.

Characteristics

In ▶ named entity recognition in systems biology, named entities pertain to categories relevant to biology, such as individual genes or gene families, cell lines, tissues, cell compartments, species, chemical compounds, and kinetic constants. In many cases, membership to a particular class of entities is indicated by certain characteristics of the terms in that class: for example, the names of persons which start with an initial uppercase letter, the names of many enzymes which end with the suffix "-ase," names of cellular compartments which are of Latin origin and use syllables uncommon in English (as in "nucleus," "cytoplasm," etc.). Named entities often have synonyms (such as "human" and "Homo sapiens" or "MAPK1"and "ERK-2") and often, homonyms exist as well (such as "p40" designating IL9, RPSA, or MAPK1, among others). To distinguish whether an occurrence of "p40" in a text refers to a protein, a Propranolol dosage of 40 mg, or "page 40" is the task for ▶ word sense disambiguation.

Named Entity Recognition

Jörg Hakenberg
Department of Computer Science and Department of Biomedical Informatics, Arizona State University, Tempe, AZ, USA

Definition

Named entity recognition (NER) is a subtask to ▶ information extraction and ▶ text mining, concerned with spotting and classifying (▶ Classification) atomic elements in a text, named entities (▶ Named Entity), such as persons, locations, genes, proteins, or ▶ gene ontology terms.

Characteristics

The goal of named entity recognition is to find all occurrences pertaining to a given class of entities, such as genes, enzymes, and drugs, in a piece of text. Knowing about the entities contained in a text facilitates indexing and search (▶ Information Retrieval) and summarization of documents and passages. It is also a key step for subsequent ▶ information extraction that focuses on single entities as well as their associations with others of the same or different classes, such as protein-protein interactions or gene-disease associations.

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

NER consists of two basic steps: recognizing that a word or phrase refers to a concept of interest (▶ Automated Term Recognition) and assigning the proper entity class. Most of the current NER methods focus on a single class of entities (to spot just drug names, for example), and these two steps often are handled at the same time. Related to NER is ▶ word sense disambiguation, which deals with assigning the proper semantic category (class) to a term in case of ambiguities (see examples below).

The challenges for NER arise from homonyms, synonyms, and acronyms: one and the same term might refer to multiple entities ("insulin" as a protein or a drug; "white" as a color or a *Drosophila* gene), an entity can have multiple names ("Death Receptor 5" and "TNF-related apoptosis-inducing ligand receptor 2" as synonyms for the "Tumor necrosis factor receptor superfamily member 10B" gene), and acronyms/abbreviations often overlap with abbreviations referring to other entities (see example on "ACE" below). Tamames and Valencia (2006) discuss nomenclature guidelines and usage of gene names over time in more detail.

Methods

The basic idea behind most methods for named entity recognition is to spot terms based on characteristics that are particular to names from a given entity class, and then expand "seed" terms to larger phrases when possible. Such characteristics can be inherent to a single word: Examples are the suffix "-ase" hinting on an enzyme; the word "white" being marked as a noun rather than an adjective in one particular sentence (see "▶ Part-of-Speech Tagging"); or capitalization which usually indicates a proper noun as in "Death Receptor 5." In addition, the surrounding words in a sentence or a larger passage often contain similar hints: If a sentence mentions "peptidase activity," the acronym "ACE" is more likely to refer to "Angiotensin-converting enzyme" than to "affinity capillary electrophoresis," thus NER can assign the class "protein."

The most successful NER methods that yield high recall and precision employ supervised machine learning (▶ Learning, Supervised) techniques that draw all such characteristic properties (also called features) from a pre-annotated set of examples, usually a set of sentences in which all occurrences of one or more entity class are marked as such, so that the learner encounters positive and negative examples for a class. Recently, efforts are under way to create such datasets (called corpora) by automatic means instead of manual curation, also see "▶ Automated Corpus Generation (CALBC)".

There exist several community challenges in the spirit of CASP (Moult et al. 1995) that include tasks on NER, the most prominent in the biomedical text mining domain being BioCreative (Smith et al. 2008). Individual systems tested on BioCreative NER tasks to recognize gene names achieved an f-score of above 87% and a theoretical joint system could reach more than 90%.

Current, openly available systems include BANNER (Leaman and Gonzalez 2008) and ABNER (Settles 2004). Tools that recognize gene names and map them to identifiers (▶ Named Entity Recognition and Normalization of Species, LINNAEUS) are GNAT (Hakenberg et al. 2011) and GeneTUKit (Huang et al. 2011), for example.

Cross-References

- ▶ Automated Corpus Generation (CALBC)
- ▶ Automated Term Recognition
- ▶ Classification
- ▶ Gene Ontology
- ▶ Information Extraction
- ▶ Information Retrieval
- ▶ Learning, Supervised
- ▶ Named Entity
- ▶ Named Entity Recognition and Normalization of Species, LINNAEUS
- ▶ Part-of-Speech Tagging
- ▶ Text Mining
- ▶ Word Sense Disambiguation

References

Hakenberg J et al (2011) The GNAT library for local and remote gene mention normalization. Bioinformatics 27(19):2769–2771

Huang M et al (2011) GeneTUKit: a software for document-level gene normalization. Bioinformatics 27(7):1032–3

Leaman R, Gonzalez G (2008) BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput 13:652–663

Moult J et al (1995) A large-scale experiment to assess protein structure prediction methods. Proteins 23(3):ii–iv

Settles B (2004) Biomedical named entity recognition using conditional random fields and rich feature sets. In: Proceedings of workshop on natural language processing in biomedicine and its applications (NLPBA), Geneva

Smith L et al (2008) Overview of BioCreative II gene mention recognition. Genome Biol 9(Suppl 2):S2

Tamames J, Valencia A (2006) The success (or not) of HUGO nomenclature. Genome Biol 7(5):402

Named Entity Recognition and Normalization of Species, LINNAEUS

Martin Gerner[1] and Goran Nenadic[2]
[1]Faculty of Life Sciences, University of Manchester, Manchester, UK
[2]Manchester Interdisciplinary Biocentre, University of Manchester, Manchester, UK

Definition

Species named entity recognition and normalization (here referred to as species NER) is the problem of finding mentions of species in unstructured text (► Named Entity Recognition) and linking (normalizing) these mentions to appropriate database identifiers. LINNAEUS (Gerner et al. 2010) is an application that has been developed to perform species NER, and serves as an example of how the problems involved with species NER can be solved.

Characteristics

Introduction

Knowledge of what species are mentioned in a document is useful in a range of situations. It can aid in ► information retrieval, that is, in the discovery of relevant research articles by providing document search systems with a method for filtering article search results based on the species discussed in them. Species NER can also help in a range of other ► text mining applications: for example, to accurately identify gene or anatomical location names, identities of which cannot be fully determined without knowing what species they belong to (for an example of a typical text-mining problem where species NER functionality is crucial, see ► Gene Normalization with GNAT).

Although species names are well-defined, species NER is not trivial: many documents mention several species, some use various non-standardized synonyms, whereas some assume that – given the context – it is obvious what species is referred to. For example, many documents use "*C. elegans*" to refer to *Caenorhabditis elegans* only, although this acronym could be used for 41 different species. The aim of a species named entity recognition and normalization system is to identify every mention of species in a given document and link it to a database entry that describes the correct species. Various species taxonomies and databases can be used for normalization, for example, the NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/). Any software tool aiming to perform species NER with high accuracy need to overcome several challenges including

- Ambiguous species names, acronyms, and abbreviations (e.g., "yeast" and "*C. elegans*")
- Misspellings (complex and less frequent scientific names can often be misspelled)
- Different lexical variations of a species name that authors may use (e.g., "fruit fly," "Fruit flies," "*D. melanogaster*")

There are various approaches that can be used to identify species mentions in text. They typically follow three steps: build an extensive dictionary of species names of interest with the corresponding links to the taxonomies; apply the dictionary to text to identify candidate species mentions; and finally determine correct mappings through disambiguation. There are also computational challenges in regard to achieving efficient runtimes when applying such software, given the number of potential species (the NCBI taxonomy, for example, contains names and synonyms for half a million species) to very large document sets such as MEDLINE (containing about ten million scientific abstracts, with a total of about 27 million species mentions).

Generating a Dictionary of Species Names

In order to locate species names and link them to database identifiers, existing taxonomies and ontologies (► Ontology; ► Ontology Lookup Service for Controlled Vocabularies and Data Annotation) are required. These resources contain names and synonyms for various species. The resource that is most commonly used for species identifiers in the biomedical domain is the NCBI taxonomy (also used by LINNAEUS). Other alternatives exist that may be more suitable for other domains, such as *uBio* (www.ubio.org; which may be more suitable for bio-diversity-focused species NER).

Existing taxonomies and ontologies typically list the scientific name (e.g., "Homo sapiens") and common synonyms (e.g., human, man) for species. Still, authors will often use other variations of these terms, such as "H. sapiens," "Human," or "humans." In order

to detect these variations, the species terms need to be expanded to cover most common lexical variations of the original terms. This is typically done using a manually constructed set of rules automatically applied to the whole taxonomy. In addition, manual modifications may still be necessary after generating the dictionary by expanding the terms in order to keep the number of false negatives and positives (▶ False Positive Rate) to a minimum. To reduce the number of false negatives, some terms may need to be added to the dictionary. For example, although "patient" strictly is not the name of a species, for many biomedical text-mining applications, linking the term "patient" to human would make a large difference. Another example, this time where it may be difficult to generate the plural version of a term, is "mice" ("mouse" is often included as a term in ontologies, while "mice" typically is not). Depending on the taxonomy used, it may also be necessary to manually remove some terms in order to avoid false positives. For the NCBI taxonomy, two extreme examples are the terms "name" and "spot" that are listed as synonyms for *Dioscorea trifida* and *Leiostomus xanthurus*, respectively.

Applying the Dictionary to Research Articles

A typical second step is the identification of strings in text that match dictionary entries (candidate mentions). When applying the dictionary of expanded terms to a large number of research documents, the choice of algorithm for matching will have a large impact on the amount of time and computer memory required. For LINNAEUS and a dictionary extracted from the NCBI taxonomy, this was solved by first generating dictionary regular expressions, which in turn are used to generate a list of all possible variations of the regular expressions. This list is matched against the text using a custom search algorithm, but alternatives that use regular expressions in other ways are also available (Møller 2008).

Determining the Correct Species Identifier for Species Mentions

By the nature of using a dictionary for finding the species mentions, all located mentions will also be associated with the species identifiers associated to the recognized term. However, a significant portion of mentions can be ambiguously matched to several identifiers (11% of all mentions in MEDLINE) and need to be disambiguated (▶ Word Sense Disambiguation). These include terms such as "C. elegans" (matching 41 different species) or "CMV" (matching both cytomegalovirus and cucumber mosaic virus). The disambiguation methods used by LINNAEUS include searching for explicit mentions of candidate species elsewhere in the document (e.g., an explicit mention of "*Caenorhabditis elegans*" earlier or later in the document) and – in the case of overlapping mentions – only retaining the mentions of longest length. Also used are "background frequencies" of explicit mentions of species in MEDLINE. More precisely, for each species name we can calculate how many times it non-ambiguously appears in MEDLINE, and then use these to filter out extremely rare interpretations of particularly ambiguous species acronyms or synonyms. Using these background frequencies, mentions such as for example "*C. elegans*" can be disambiguated, with high accuracy, to *Caenorhabditis elegans* (since this species is mentioned almost 250 times more often than the second most frequently mentioned species alternative, *Cunninghamella elegans*).

Evaluating the Accuracy of Species NER Systems

A key part in the development of text-mining software in general is to perform an evaluation of its accuracy in order to estimate the quality of the data generated by the tool. This is typically performed by applying the tool to a set of documents, called corpus (▶ Named Entity Recognition; ▶ Text mining), for which one or more human annotators have determined what the correct output should be. The output generated by the tool can then be compared against the manual annotations, enabling the computation of precision and recall accuracy levels.

The only currently freely available corpus manually annotated for species mentions is a corpus constructed as part of the LINNAEUS project. This corpus consists of 100 open access full-text documents and annotations for all mentioned species, linked to NCBI taxonomy identifiers. In total, the corpus contains 4,259 references to 233 different species.

LINNAEUS

LINNAEUS is a software package that has been developed in order to enable accurate and fast species NER of biomedical articles. It can process documents in a number of input formats (MEDLINE XML, PubMed Central XML, BioMed Central XML, or plain-text files) and provides the input text annotated with

disambiguated species mentions as output (for an example, see Fig. 1).

The general processing workflow of LINNAEUS is given in Fig. 2. The NCBI taxonomy and manual additions are combined to construct a dictionary, which is applied to a set of documents. Disambiguation is performed using the rules mentioned above. Additional rule-based algorithms are used to detect author-declared acronyms, remove common false positives, and assign identity probabilities (using background mention frequencies) to mentions that still are ambiguous.

Compared against the manually annotated corpus described in the previous section, LINNAEUS achieves an accuracy of 94% recall and 97% precision. It is capable of processing text documents at a fast rate: consuming about 2 GB memory and running on a 2.66 GHz CPU, LINNAEUS can process documents at a rate of about 1,100 MEDLINE abstracts or about 70 PubMed Central full-text documents per second. Additionally, the user can instruct LINNAEUS to utilize multiple threads, which would lead to faster processing on multicore CPUs.

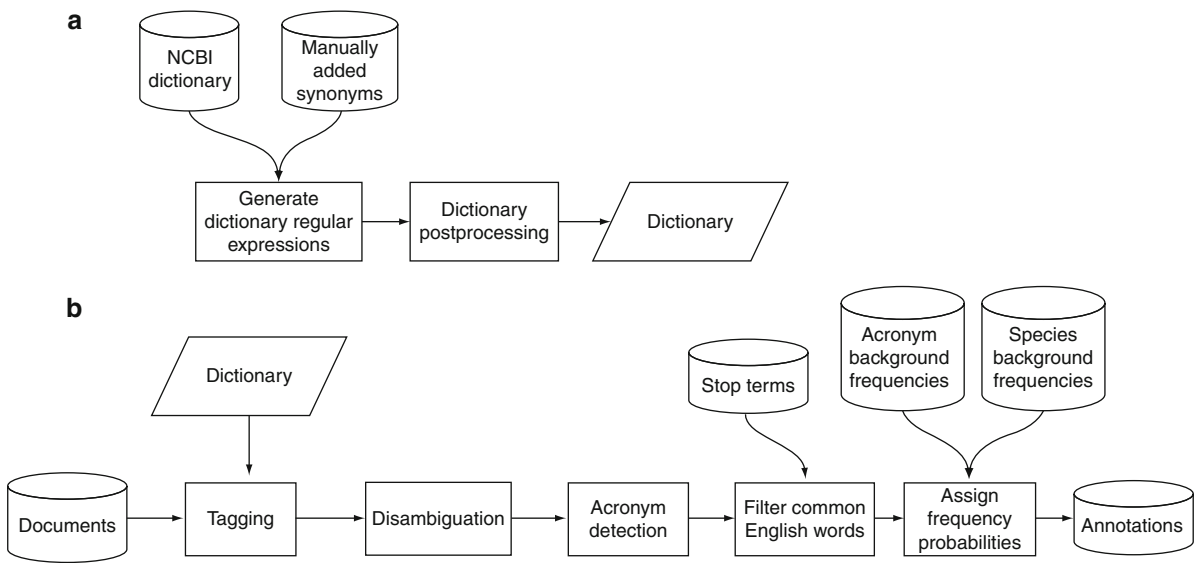
LINNAEUS is available for download as a stand-alone application and as a web service at http://linnaeus.sourceforge.net.

BACKGROUND: Mutations in the gene encoding the E3 ubiquitin ligase parkin (PARK2) are responsible for the majority of autosomal recessive parkinsonism. Similarly to other knockout mouse models of PD-associated genes, parkin knockout mice do not show a substantial neuropathological or behavioral phenotype, while loss of parkin in Drosophila melanogaster leads to a severe phenotype, including reduced lifespan, apoptotic flight muscle degeneration and male sterility. In order to study the function of parkin in more detail and to address possible differences in its role in different species, we chose Danio rerio as a different vertebrate model system.

Named Entity Recognition and Normalization of Species, LINNAEUS, Fig. 1 An example of a portion of an abstract marked up for species names (that are hyperlinked to the NCBI taxonomy). Output is also generated in a table format for further software processing

Related Software

There are several tools available for species NER. Whatizit organisms (Rebholz-Schuhmann et al. 2007) is a species NER and normalization web service hosted by the European Bioinformatics Institute (EBI). Similar to LINNAEUS, Whatizit organisms is also based on the NCBI taxonomy. Taxongrab (Koning et al. 2006) is able to recognize a wide variety of scientific species names using a set of rules, and is not limited to any



Named Entity Recognition and Normalization of Species, LINNAEUS, Fig. 2 LINNAEUS processing workflow. (**a**) Generating a species dictionary. (**b**) Locating and identifying species names in text using the constructed dictionary

specific taxonomy. It, however, does not provide normalization of species mentions (linking to database identifiers) or recognition of common species names.

Cross-References

- ▶ Named Entity
- ▶ Named Entity Recognition
- ▶ Ontology Lookup Service for Controlled Vocabularies and Data Annotation
- ▶ Ontology
- ▶ Text Mining
- ▶ Word Sense Disambiguation

References

Gerner M, Nenadic G, Bergman CM (2010) LINNAEUS: a species name identification system for biomedical literature. BMC Bioinform 11:85

Koning D, Sarkar IN, Moritz T (2006) TaxonGrab: extracting taxonomic names from text. Biodivers Inform 2:79–82

Møller A (2008) dk.brics.automaton. http://www.brics.dk/automaton/

Rebholz-Schuhmann D, Arregui M, Gaudan M, Kirsch H, Jimeno A (2007) Text processing through Web services: calling Whatizit. Bioinformatics 23(2):e237–e244

National Cancer Institute Thesaurus

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

Definition

The NCI Thesaurus is a comprehensive collection of terms relating to cancer biology, clinical oncology, and cancer epidemiology. The NCI Thesaurus was conceived in the 1990s as a compendium of everything that the NCI cared about in its research portfolio.

Cross-References

- ▶ Protégé Ontology Editor

National Center for Biomedical Ontology

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

Definition

One of the eight National Centers for Biomedical Computing established under the original NIH Roadmap in the 2000s, the NCBO is centered at Stanford University with collaborators at the Mayo Clinic (Rochester, MN), the University at Buffalo, and the University of Victoria (Canada), among other institutions. The NCBO maintains a comprehensive repository of biomedical terminologies, ontologies, and models; develops software tools to assist biomedical investigators in the use of ontologies; and collaborates with a wide range of investigators on the use of semantic technology in biomedicine. See http://bioontology.org.

Cross-References

- ▶ Protégé Ontology Editor

Natural Killer Cells, Mycobacterial Infection

Rohan Dhiman
Center for Pulmonary and Infectious Disease Control, University of Texas Health Science Center at Tyler, Tyler, TX, USA

Synonyms

Killer cells; M. tb infection; NK cells, M. tb infection

Definition

Natural killer cells are lymphocytes that comprise an important arm of the ▶ innate immunity. They are

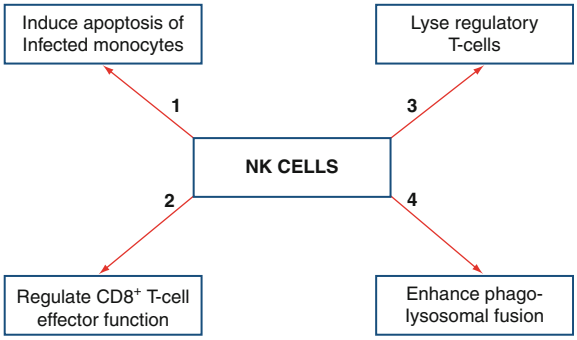
named natural killer cells because of their ability to lyse certain tumor cells without prior sensitization. This ability makes them different from B and T lymphocytes. Natural killer cells mediate protection against viruses, bacteria, and parasites by lysing infected cells and by secreting ▶ cytokines that augment the adaptive immune response.

Characteristics

Natural killer cells are the third main lymphocyte population. They share same progenitors which generate B and T lymphocytes but they do not express any of the T or B lymphocyte markers. NK cells are characterized phenotypically by the expression of CD56 and lack of expression of CD3.

Monoclonal antibodies specific for NK-cell markers led to the discovery of two distinct populations of human NK cells based upon the cell-surface expression of CD56, $CD56^{bright}$, or $CD56^{dim}$ (Robertson and Ritz 1990; Lanier et al. 1986). CD56 is an isoform of the human neural-cell adhesion molecule and a glycoprotein expressed by various cells like neurons, glia, skeletal muscle, and natural killer cells. CD56 plays an important role in cell–cell adhesion, synaptic plasticity, and memory. $CD56^{dim}$ human NK cells constitute a major population (around 90%) of NK cells and express high levels of Fcγ receptor III (FcγRIII, CD16), whereas less than 10% of NK cells are $CD56^{bright}CD16^{dim}$ or $CD56^{bright}CD16^{-}$ (Cooper et al. 2001a). Both these subsets uniformly express NK group 2, member D (NKG2D), CD161, NK-cell protein 46 (NKp46), and CD122 but differ in the expression of various other NK receptors (Freud et al. 2006). $CD56^{bright}$ NK cells express no or very low expression of KIRs (killer-cell immunoglobulin-like receptors) and ILT-2 (an inhibitory receptor) but high-level expression of CD94 and NKG2A inhibitory receptors compared to $CD56^{dim}$ NK cells (Andre et al. 2000; Voss et al. 1998; Colonna et al. 1997). These two subsets also differ in the expression of various cytokines and chemokine receptors like c-kit, IL-1R1, CCR7, CXCR1, and CX_3CR1 (Matos et al. 1993; Cooper et al. 2001b; Campbell et al. 2001), and adhesion molecules like CD2, CD62L, CD44, and CD49 (Lanier et al. 1986; Frey et al. 1998; Sedlmayr et al. 1996).

NK-cell subsets also differ in their functional responses. $CD56^{dim}$ NK cells have been found to be more cytotoxic than $CD56^{bright}$ NK cells (Nagler et al. 1989). On the other hand, $CD56^{bright}$ NK cells produce high levels of immunoregulatory cytokines like interferon- γ (IFN-γ), tumor necrosis factor β (TNF-β), IL-10, IL-13 and granulocyte–macrophage colony-stimulating factor (GM-CSF) compared to $CD56^{dim}$ NK cells (Cooper et al. 2001c).



Natural Killer Cells, Mycobacterial Infection, Fig.1 Differential functional facets of NK cell in mycobacterial infection

Role in Mycobacterial Infection

Tuberculosis is a leading cause of death from infectious diseases worldwide, claiming an estimated 1.3 million lives annually. Multidrug-resistant tuberculosis continues to spread in many parts of the world, requiring therapy with potentially toxic agents for a long time, compared to that for drug-susceptible tuberculosis. Development of various strategies which augments innate immunity against ▶ *Mycobacterium tuberculosis* constitutes an important component to fight against both drug-resistant and drug-susceptible tuberculosis (Dhiman et al. 2009).

Natural killer cells, an important part of innate immune defense, have been found to play an essential role in immune defenses against cancer and infectious diseases in various experimental settings. NK cells kill autologous infected cells without prior sensitization through perforin or Fas/Fas ligand pathway and by secretion of various cytokines, thus playing a central role in innate immunity against microbial pathogens (Vankayalapati and Barnes 2009).

It has been shown that NK cells exert antimycobacterial activity using various mechanisms (Fig. 1). First, they kill mycobacteria in vitro by inducing apoptosis in infected monocytes mediated by NKp46 recognition of vimentin and NKG2D recognition of its ligand ULBP-1 (Vankayalapati et al. 2002;

Vankayalapati et al. 2005; Garg et al. 2006). Mycobacterial ▶ infection leads to increase in vimentin and ULBP-1 expression on infected macrophages and NK cells lyse these cells via ligation of NKp46 with vimentin and NKG2D with ULBP-1. Secondly, they shape the adaptive immune response by regulating $CD8^+$ T-cell effector function against mononuclear phagocytes infected with *M. tuberculosis* (Vankayalapati et al. 2004). NK cell–depleted peripheral blood mononuclear cells of healthy tuberculin reactors show reduction in frequency of *M. tuberculosis*–responsive $CD8^+IFN\text{-}\gamma^+$ cells. These $CD8^+$ cells also show decreased capacity to lyse infected monocytes. Thirdly, they lyse regulatory T cells (Roy et al. 2008) which are expanded T cells that express a regulatory phenotype (CD25 + FoxP3+). It has been shown recently that regulatory T cells prevent efficient clearance of ▶ infection in infected mice by proliferating and accumulating at sites of infection (Kursar et al. 2007; Scott-Browne et al. 2007). T-regs have also been shown to inhibit IFN-γ production by BCG stimulated CD4 + CD25- cells, thus clearly showing that they inhibit an effective immune response (Li et al. 2007; Garg et al. 2008). Fourthly, they also secrete IL-22, member of a group of cytokines called the IL-10 superfamily. IL-22 has been shown to restrict mycobacterial growth by enhancing phagolysosomal fusion (Dhiman et al. 2009).

Cross-References

- ▶ CD8+ Cytotoxic T Lymphocytes (CTL)
- ▶ Cytokines
- ▶ Infection
- ▶ Innate Immunity
- ▶ Mycobacterium Tuberculosis

References

Andre P, Spertini O, Guia S, Rihet P, Dignat-George F, Brailly H, Sampol J, Anderson PJ, Vivier E (2000) Modification of P-selectin glycoprotein ligand-1 with a natural killer-cell restricted sulphated lactosamine creates an alternate ligand for L-selectin. Proc Natl Acad Sci USA 97(7):3400–3405

Campbell JJ, Qin S, Unutmaz D, Soler D, Murphy KE, Hodge MR, Wu L, Butcher EC (2001) Unique subpopulation of CD56+ NK and NK-T peripheral blood lymphocytes identified by chemokine receptor expression repertoire. J Immunol 166(11):6477–6482

Colonna M, Navarro F, Bellon T, Llano M, Garcia P, Samaridis J, Angman L, Cella M, Lopez-Botet M (1997) A common inhibitory receptor for major histocompatability complex class I molecules on human lymphoid and myelomonocytic cells. J Exp Med 186(11):1809–1818

Cooper MA, Fehniger TA, Caligiuri MA (2001a) The biology of human natural killer-cell subsets. Trends Immunol 22(11):633–640

Cooper MA, Fehniger TA, Ponnappan A, Mehta V, Wewers MD, Caligiuri MA (2001b) Interleukin-1beta costimulates interferon-gamma production by human natural killer cells. Eur J Immunol 31(3):792–801

Cooper MA, Fehniger TA, Turner SC, Chen KS, Ghaheri BA, Ghayur T, Carson WE, Caligiuri MA (2001c) Human natural killer cells: a unique innate immunoregulatory role for the CD56(bright) subset. Blood 97(10):3146–3151

Dhiman R, Indramohan M, Barnes PF, Nayak RC, Paidipally P, Rao LV, Vankayalapati R (2009) IL-22 produced by human NK cells inhibits growth of *Mycobacterium tuberculosis* by enhancing phagolysosomal fusion. J Immunol 183(10): 6639–6645

Freud AG, Yokohama A, Becknell B, Lee MT, Mao HC, Ferketich AK, Caligiuri MA (2006) Evidence for discrete stages of human natural killer cell differentiation in vivo. J Exp Med 203(4):1033–1043

Frey M, Packianathan NB, Fehniger TA, Ross ME, Wang WC, Stewart CC, Caligiuri MA, Evans SS (1998) Differential expression and function of L-selectin on CD56bright and CD56dim natural killer cell subsets. J Immunol 161(1):400–408

Garg A, Barnes PF, Porgador A, Roy S, Wu S, Nanda JS, Griffith DE, Girard WM, Rawal N, Shetty S, Vankayalapati R (2006) Vimentin expressed on *Mycobacterium tuberculosis*-infected human monocytes is involved in binding to the NKp46 receptor. J Immunol 177(9):6192–6198

Garg A, Barnes PF, Roy S, Quiroga MF, Wu S, García VE, Krutzik SR, Weis SE, Vankayalapati R (2008) Mannose-capped lipoarabinomannan- and prostaglandin E2-dependent expansion of regulatory T cells in human *Mycobacterium tuberculosis* infection. Eur J Immunol 38(2):459–469

Kursar M, Koch M, Mittrücker HW, Nouailles G, Bonhagen K, Kamradt T, Kaufmann SH (2007) Cutting Edge: Regulatory T cells prevent efficient clearance of *Mycobacterium tuberculosis*. J Immunol 178(5):2661–2665

Lanier LL, Le AM, Civin CI, Loken MR, Phillips JH (1986) The relationship of CD16 (Leu-11) and Leu-19 (NKH-1) antigen expression on human peripheral blood NK cells and cytotoxic T lymphocytes. J Immunol 136(12):4480–4486

Li L, Lao SH, Wu CY (2007) Increased frequency of CD4(+) CD25(high) Treg cells inhibit BCG-specific induction of IFN-gamma by CD4(+) T cells from TB patients. Tuberculosis (Edinb) 87(6):526–534

Matos ME, Schnier GS, Beecher MS, Ashman LK, William DE, Caligiuri MA (1993) Expression of a functional c-kit receptor on a subset of natural killer cells. J Exp Med 178(3):1079–1084

Nagler A, Lanier LL, Cwirla S, Phillips JH (1989) Comparative studies of human FcRIII-positive and negative natural killer cells. J Immunol 143(10):3183–3191

Robertson MJ, Ritz J (1990) Biology and clinical relevance of human natural killer cells. Blood 76(12):2421–2438

Roy S, Barnes PF, Garg A, Wu S, Cosman D, Vankayalapati R (2008) NK cells lyse T regulatory cells that expand in response to an intracellular pathogen. J Immunol 180(3):1729–1736

Scott-Browne JP, Shafiani S, Tucker-Heard G, Ishida-Tsubota K, Fontenot JD, Rudensky AY, Bevan MJ, Urdahl KB (2007) Expansion and function of Foxp3-expressing T regulatory cells during tuberculosis. J Exp Med 204(9):2159–2169

Sedlmayr P, Schallhammer L, Hammer A, Wilders-Truschnig M, Wintersteiger R, Dohr G (1996) Differential phenotypic properties of human peripheral blood CD56dim + and CD56bright + natural killer cell subpopulations. Int Arch Allergy Immunol 110(4):308–313

Vankayalapati R, Barnes PF (2009) Innate and adaptive immune responses to human *Mycobacterium tuberculosis* infection. Tuberculosis (Edinb) 89(Suppl 1):S77–S80

Vankayalapati R, Wizel B, Weis SE, Safi H, Lakey DL, Mandelboim O, Samten B, Porgador A, Barnes PF (2002) The NKp46 receptor contributes to NK cell lysis of mononuclear phagocytes infected with an intracellular bacterium. J Immunol 168(7):3451–3457

Vankayalapati R, Klucar P, Wizel B, Weis SE, Samten B, Safi H, Shams H, Barnes PF (2004) NK cells regulate CD8+ T cell effector function in response to an intracellular pathogen. J Immunol 172(1):130–137

Vankayalapati R, Garg A, Porgador A, Griffith DE, Klucar P, Safi H, Girard WM, Cosman D, Spies T, Barnes PF (2005) Role of NK cell-activating receptors and their ligands in the lysis of mononuclear phagocytes infected with an intracellular bacterium. J Immunol 175(7):4611–4617

Voss SD, Daley J, Ritz J, Robertson MJ (1998) Participation of the CD94 receptor complex in costimulation of human natural killer cells. J Immunol 160(4):1618–1626

Natural Language Processing

Karin Verspoor[1,2] and Kevin Bretonnel Cohen[2]
[1]Victoria Research Laboratory, National ICT Australia, University of Melbourne, Melbourne, VIC, Australia
[2]Center for Computational Pharmacology, University of Colorado, Aurora, CO, USA

Synonyms

Biomedical natural language processing (BioNLP); Computational linguistics; Information extraction; Natural language understanding; Text mining; Text processing

Definition

Natural language processing is the analysis of linguistic data, most commonly in the form of textual data such as documents or publications, using computational methods. The goal of natural language processing is generally to build a representation of the text that adds structure to the unstructured natural language, by taking advantage of insights from linguistics. This structure can be *syntactic* in nature, capturing the grammatical relationships among constituents of the text, or more *semantic*, capturing the meaning conveyed by the text.

Natural language processing is used in systems biology to develop applications that integrate information extracted from the literature with other sources of biological data (see ▶ Applied Text Mining).

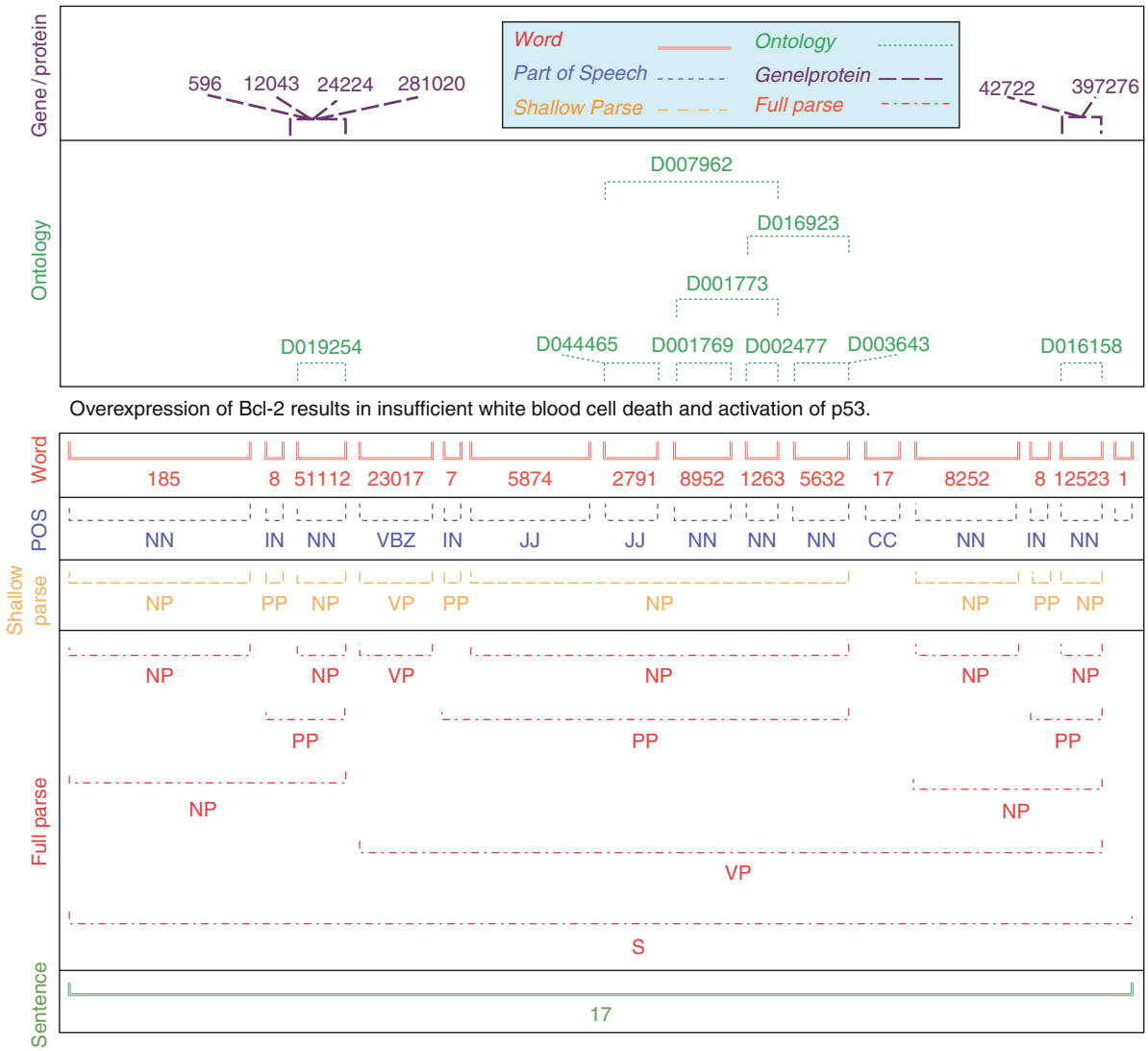
Characteristics

The typical natural language processing system consists of a pipeline of components that manipulate an input text in increasingly sophisticated ways. Generally, the aim of each component is to add structure to the text that can be used to facilitate downstream processing. The components early on in the pipeline handle tasks that are close to the surface strings of the text, while later components aim to analyze concepts and relationships. Various methods may be used to accomplish component tasks, ranging from rule-based methods, such as regular expressions and finite state automata, to statistical and machine learning models.

In Fig. 1, we can see an example of the processing of a single sentence from a biomedical text. Each level will be discussed in more detail below.

Tokenization and Sentence Demarcation

Natural language processing is strongly word based in that words are generally considered to carry the meaning of a text. It is therefore important as a preprocessing step to any further analysis to delimit the individual *word tokens* that make up a text. This is seen at the "Word" level in Fig. 1. This process is referred to as *tokenization*. While a simple approach is to split the text on any whitespace or punctuation, some care must be taken in biomedical texts to appropriately handle punctuation that has special meaning in certain contexts, such as a single quote in the representation of a DNA strand (*5′-GCRTGNCCAT-3′*), the characters in some chemical names (*tricyclo (3.3.1.13,7)decanone*), hyphens which can indicate charge (*Cl-*), constitute part of a gene or cell name



Natural Language Processing, Fig. 1 The levels of analysis for a sentence processed by a typical NLP pipeline (From Hunter and Bretonnel Cohen [2006], adapted from Nakov et al. [2005])

(*hsp-60, t-cell*), or a knocked out gene (*lush- flies*), etc. Thus, tokenization tools sensitive to the biomedical context are required.

As a precursor to syntactic analysis, it is also important to delimit individual sentences within a text. This is because the sentence is generally the grammatical unit of a text. Similarly to tokenization, *sentence splitting* generally involves taking advantage of a basic heuristic: look for normal sentence-final punctuation (period or question mark) followed by a capital letter. However, again complications can be introduced in the context of names containing initials (e.g., *H.G. Wells* or *Dr. Bronner's soap*) where the heuristic would incorrectly split a sentence into multiple pieces. Similarly, domain-specific conventions that sometimes require a sentence-initial, lowercase letter can cause problems for sentence demarcation, such as in the following example:

> The process of activation involves [. . .] phosphorylation of tyrosine **kinases**. **p21(ras)**, a guanine nucleotide binding factor, mediates T-cell signal transduction . . . (from PMID 8887687, with thanks to Bob Carpenter for finding it)

Syntactic and Morphological Analysis

Syntactic information about the text can be important to assist in resolving ambiguities and in establishing the appropriate relations among the words in a text. At the most basic level, determining whether a word is a noun or a verb (or some other part of speech) can be useful. This is accomplished through tools that perform ▶ part-of-speech tagging. Then, identification of phrases in the text can be important, such as recognizing that a sequence of words forms a single conceptual unit (e.g., breast cancer (▶ Data Integration, Breast Cancer Database) and *NF kappa beta inhibitor*). A commonly used strategy for this is *shallow* ▶ parsing, which involves identifying coarse phrasal structures, such as noun phrases, without identifying the specific grammatical relationships among them. In contrast, *deep* ▶ parsing determines the full set of grammatical relations among words in a sentence, producing a complete *parse tree* to represent these relations.

The surface forms of words will vary depending on their syntactic usage in a sentence, for instance, a noun appearing in plural form or a verb appearing in various tenses (*regulated, regulating, regulates*). Often, it is desirable to normalize such variation to a base form of the word in order to appropriately associate different occurrences of the same term. This is called *morphological normalization* and is often accomplished in practical NLP applications through *stemming* tools which strip off inflected word endings. The *Porter* algorithm, based on suffix stripping, is a popularly used strategy for stemming (Porter 1980).

Information Extraction

Information extraction in general refers to the extraction of specific types of information from text and normally formalized in a structured representation, such as an event template or a concept from an externally defined ontology. It can refer to the association of particular strings of a text to a category of interest, for instance, identifying protein names in a publication.

Named Entity Recognition

In the upper levels of Fig. 1, we see annotations of ontology terms and gene/protein terms. Many of such terms correspond to ▶ *named entities*, i.e., to objects that are generally referred to by name. This is in contrast to terms that correspond to processes or events, which normally require identification of higher-order relations. Examples of named entities in the biological domain that are often targeted for extraction are genes, diseases, chemicals, or experimental methods.

Various methods exist for performing named entity recognition. The most basic approach is to compile a dictionary of the relevant names for a specific category of entities, and to perform a string match into the dictionary. Empirical methods based on supervised machine learning will often use a dictionary match as one feature of a model that also considers surrounding words, syntax, and other textual evidence to identify likely instances of terms from a particular category.

Relation and Event Extraction

Beyond extraction of entities, many applications require extraction of *relations* among those entities. One popular example, addressed in several shared tasks such as BioCreative (Hirschman et al. 2005; Krallinger et al. 2008; Leitner et al. 2010) and BioNLP′09 (Kim et al. 2009), is identification of protein–protein interactions from text. This first requires the recognition of the proteins as entities and then identification of an interaction relation among at least two of the recognized proteins. In the sentence in Fig. 1, for instance, we can identify an activation relationship between Bcl-2 and p53, i.e., one of the key pieces of information in the sentence can be summarized as *Bcl-2 activates p53*. Strategies for relation extraction again vary from high-precision linguistic-based methods (Cohen et al. 2011) to high-recall supervised learning methods (Dai et al. 2010), and hybrid methods that achieve more balanced performance (Hakenberg et al. 2010).

Co-reference Resolution

Co-reference resolution refers to identifying multiple occurrences in the text of the same entity or event. It includes resolving pronouns such as "it" to their references, as well as other kinds of references such as definite noun phrases (a noun phrase that starts with "the," e.g., "the protein"). Note that these references can include references to events previously mentioned, e.g., "the process" or "this interaction."

Implementation Aspects

Natural language processing systems are implemented in the form of software. Such systems tend to have modular architectures where components such as those outlined above are run serially in a "pipeline."

Document Format Issues

Before any more sophisticated linguistic processing can be performed, documents must be converted into a format that is easy for computational tools to work with. Since source documents can be available in various formats, including HTML, XML, Microsoft Word, and PDF, in addition to plain text, NLP systems must clearly specify the kinds of input documents they can handle. In general, documents must be converted to a simpler plain text representation without the structure and formatting information available in other formats. There are tools available to assist with these conversions, but they can vary in quality and effectiveness.

In addition, NLP systems must be sensitive to the character encoding of a given document. Documents can be encoded in numerous formats, including UTF-8 and ISO-8859-1. Some characters, in particular special two-byte UNICODE characters such as Greek letters, will not be correctly interpreted if the correct encoding is not utilized when loading the document. Since such characters can be meaningful in biomedical texts, (e.g., in the name of the TGF-β gene), this is an issue that cannot be overlooked.

For most applications, it is preferable to retain as much of the original document structure as possible. Certain formatting information can have semantic import. For instance, italics are sometimes used to highlight a gene name in a document. In addition, sensitivity to the sections of a document can provide a system with an advantage in solving certain problems, such as for detecting *new* experimental protein interactions described in the text – one would not expect these to be mentioned in a background or methods section. Document sections are generally most reliably identified by taking advantage of the previously demarcated structure of the document, but sophisticated algorithms to perform document zoning might need to be employed if such demarcations are unavailable.

Unstructured Information Management Architecture

The Unstructured Information Management Architecture, or UIMA, is a commonly used architecture for computational systems that aim to perform Natural Language Processing (Ferrucci et al. 2009). It provides a common representation for a document and its meta-data, which can be shared across components. It is the foundation of several repositories of tools supporting biomedical text mining, such as bionlp.org and u-compare.org.

Cross-References

► Applied Text Mining

References

Cohen KB, Verspoor K, Johnson H, Roeder C, Ogren P, Baumgartner W Jr., White E, Tipney H, Hunter L (2011) High-precision biological event extraction: Effects of system and data. Comput Intell 27(4)

Dai H-J, Lai P-T, Tsai RT-H (2010) Multistage gene normalization and svm-based ranking for protein interactor extraction in full-text articles. IEEE/ACM Trans Comput Biol Bioinformatics 7(3):412–420

Ferrucci D, Lally A, Verspoor K (eds) (2009) Unstructured information management architecture (UIMA) Version 1.0. OASIS Standard, 2 Mar 2009

Hakenberg J, Leaman R, Ha Vo N, Jonnalagadda S, Sullivan R, Miller C, Tari L, Baral C, Gonzalez G (2010) Efficient extraction of protein–protein interactions from full-text articles. IEEE/ACM Trans Comput Biol Bioinformatics 7(3):481–494

Hirschman L, Yeh A, Blaschke C, Valencia A (2005) Overview of biocreative: critical assessment of information extraction for biology. BMC Bioinformatics 6(Suppl 1):S1

Hunter L, Bretonnel Cohen K (2006) Biomedical language processing: what's beyond PubMed? Mol Cell 21:589–594

Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii J (2009) Overview of BioNLP'09 shared task on event extraction. In: Proceedings of the Workshop on BioNLP: Shared Task, Association for Computational Linguistics, Boulder, Colorado, pp 1–9

Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L, Wilbur J, Hirschman L, Valencia A (2008) Evaluation of text-mining systems for biology: overview of the second BioCreative community challenge. Genome Biol 9(Suppl 2):S1

Leitner F, Chatr-aryamontri A, Mardis SA, Ceol A, Krallinger M, Licata L, Hirschman L, Cesareni G, Valencia A (2010) The FEBS letters/BioCreative II.5 experiment: making biological information accessible. Nat Biotechnol 28:897–899

Nakov P, Schwartz A, Wolf B, Hearst M (2005) Supporting annotation layers for natural language processing. In: ACL 2005 Poster/Demo Track, Ann Arbor

Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137

Natural Language Understanding

► Natural Language Processing

Natural Product Databases

► Natural Product Resources

Natural Product Resources

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

Synonyms

Natural product databases

Definition

Natural product resources are databases which store information on natural products. These are information tools which facilitate the screening of natural compounds during the drug discovery process.

Characteristics

Content of Natural Product Resources

Containing thousands of natural compounds, natural product databases contain the following types of information for each compound:

1. *Descriptive data* which includes systematic and common names, synonyms, chemical structure, compound type, molecular formula, and CAS registry numbers.
2. *Physicochemical data* which includes, among others, a compound's boiling point, melting point, molecular weight, and optical rotation.
3. *Spectroscopic data* which includes measurements in the following spectra: infrared (IR), mass, nuclear magnetic resonance (NMR), and ultraviolet (UV).
4. *Origin data* which consists of taxonomic information on source organisms.
5. *Biological data* which consists of information on biological activity and toxicity.
6. *Bibliographic data* which consists of citations of primary or secondary literature from which the compound information was abstracted.

The inclusion of information on biological activities is a feature of natural product databases that makes them different from general chemical compound databases. Typically, biological activity information includes details such as the drug targets against which the compound was reportedly active, and the measured activity.

Each database is supported by an interface which allows its users to search for compounds by either entering any or a combination of the details above, or by drawing a chemical structure or substructure.

Significance of Natural Product Resources

Due to the large number of compounds which have already been published, natural product chemists face a challenge when screening compounds for novel, pharmaceutically relevant chemical structures. Structure elucidation, the process of determining the structure of chemical substances such as natural products, becomes more efficient when already known compounds are rapidly characterized or dereplicated (Corley and Durley 1994).

Dereplication involves the comparison of one's preliminary findings on a compound against published information to accomplish any of two tasks: to determine if the compound in question has already been reported, or to use a partial structure to arrive at a complete chemical structure (Dinan 2005). Natural product resources facilitate these tasks by enabling chemists to access, search, and analyze published and curated information in a systematic manner.

Available Natural Product Resources

Natural product resources can be categorized into two according to their availability: public and commercial.

Public databases are accessible to anyone who has access to the Internet. Most of them contain small molecules in general and are not limited to natural compounds only. Under this category are the following databases:

1. *ChemBank*. Created by the National Cancer Institute's Initiative for Chemical Genetics (ICG), *ChemBank* contains information on small molecules and biological assays. It is dedicated to the storage, organization, analysis, and visualization of raw screening data (Seiler et al. 2008). As of version 2.0, *ChemBank* houses data on more than 1.2 million unique small molecules and 2,500 biological assays.
2. *ChemBL Database*. Provided by the European Bioinformatics Institute's European Molecular Biology Laboratory (EMBL-EBI), the *ChemBL Database* (*ChemBLdb*) stores information on bioactive drug-like small molecules including biological activities and assay data, all abstracted from the primary literature (Warr 2009). As of version 0.9, *ChemBLdb* contains more than 650,000 unique compounds with more than three million biological activity records.

3. *PubChem*. Hosted by the US National Institutes of Health (NIH), *PubChem* consists of three component databases: *PubChem Substance*, *PubChem Compound* and *PubChem BioAssay*. *PubChem Substance* contains information on chemical samples or substances submitted by contributors. *PubChem Compound* contains chemical structures derived from the substances while *PubChem BioAssay* stores the results of biological activity testing on them. *PubChem* currently contains 81 million substance records, 32 million unique chemical structures, and more than 500,000 biological assay records. It is an open repository where any organization can become a contributor and deposit data (Wang et al. 2010).
4. *SuperNatural*. Developed at the Berlin Center of Genome-Based Bioinformatics, *SuperNatural* contains 3D structures, conformers and supplier information on natural compounds, as well as their analogues and derivatives. It also contains data on the biological activity of compounds against several tumor cell lines. As of version 3.0.2, *SuperNatural* contains data for around 50,000 natural compounds (Dunkel et al. 2006).

Commercial databases are only accessible upon payment of a fee. They usually come in the form of electronic media (e.g., CD or DVD) although some are web-based. A few of the commercially available databases are:

1. *AntiBase*. Developed at the University of Göttingen, *AntiBase* contains information on natural compounds, abstracted from primary and secondary literature. A unique feature of AntiBase is the inclusion of predicted Carbon-13 NMR (13 C-NMR) spectra, calculated by Wiley's spectrum prediction system *SpecInfo*. 13 C-NMR spectra data helps chemists determine the structure of unknown organic molecules. Available in CD form, the 2011 version of *AntiBase* contains information on more than 38,000 natural compounds (Laatsch 2011).
2. *Dictionary of Natural Products*. Chapman and Hall's *Dictionary of Natural Products* (*DNP*) is a database produced through the compilation of natural product information from the well-known Dictionary of Organic Compounds. The compounds are organized such that the users can easily view under one entry the compounds which are biosynthetically and structurally related. Also, each compound is indexed using a controlled vocabulary of more than 1,000 headings to allow faster searching. Available in DVD form, the current version of *DNP* contains more than 230,000 compounds (Buckingham 2010).
3. *MarinLit*. Developed at the University of Canterbury, *MarinLit* is a database of natural products focusing on compounds from the marine environment. Aside from comprehensive bibliographic information, it includes detailed taxonomic data, allowing the user to explore relationships among source organisms at various taxonomic levels. Available as a stand-alone application, *MarinLit* currently contains information on more than 22,000 marine natural products (Blunt 2011).
4. *NAPRALERT*. Developed at the University of Illinois at Chicago, *NAPRALERT* is a database of natural products, formed by abstracting primary and secondary literature (Graham and Farnsworth 2010). It includes information on a compound's uses in traditional medicine, aside from the biological activities established in assays. It currently contains data from more than 200,000 scientific papers and reviews. However, due to financial constraints, only 15% of the literature has been included from 2004 to present. A user is required to pay a fee in order to retrieve results from the database using the online interface.

Cross-References

- Biological Activity
- Biological Assay
- Drug Discovery
- Drug Target
- Small Molecule

References

Blunt J (2011) MarinLit database. University of Canterbury, Department of Chemistry, New Zealand

Buckingham J (2010) Dictionary of natural products on DVD. Chapman & Hall/CRC, London

Corley DG, Durley RC (1994) Strategies for database dereplication of natural products. J Nat Prod 57(11):1484–1490

Dinan L (2005) Dereplication and partial identification of natural products. In: Sarker SD, Latif Z, Gray AI (eds) Methods in biotechnology, vol 20, 2nd edn, Natural products isolation. Humana Press, Totowa, pp 297–321

Dunkel M, Fullbeck M, Neumann S, Preissner R (2006) SuperNatural: a searchable database of available natural compounds. Nucleic Acids Res 34(suppl 1): D678–D683

Graham JG, Farnsworth NR (2010) The NAPRALERT database as an aid for discovery of novel bioactive compounds.

In: Mander L, Liu H (eds) Comprehensive natural products II. Elsevier, Oxford, pp 81–94
Laatsch H (2011) AntiBase 2011: the natural compound identifier. Wiley-VCH, Weinheim
Seiler KP, George GA, Happ MP, Bodycombe NE, Carrinsky HA, Norton S, Brudz S, Sullivan JP, Muhlich J, Serrano M, Ferraiolo P, Tolliday NJ, Schreiber SL, Clemons P (2008) ChemBank: a small-molecule screening and cheminformatics resource database. Nucleic Acids Res 36(suppl 1):D351–D359
Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker BA, Suzek TO, Wang J, Xiao J, Zhang J, Bryant SH (2010) An overview of the PubChem BioAssay resource. Nucleic Acids Res 38(suppl 1):D255–D266
Warr WA (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). J Comput Aided Mol Des 23(4): 195–198

NC1

► Cofactors

NC2

► Cofactors

NCBI BioProject 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

Synonyms

Entrez genome project; Genome project; NCBI genome project resource

Definition

The NCBI BioProject database provides a set of complete and in-progress large-scale sequencing, assembly, annotation, and mapping projects originating from a single organization or a consortium. The BioProject database is designed for complex collaborations looking at various aspects of cellular organisms and generating different types of records, including genome sequences and assemblies, metagenomes (► Metagenomics), transcriptome sequences and expression, and epigenetic records.

Characteristics

The NCBI BioProject operates as a web portal for large-scale genome sequencing and other biomedical projects across several taxonomies (► TaxonRank) as well as for projects focusing on a particular locus (Sayers et al. 2011; Pruitt et al. 2011). BioProject offers a synoptic table containing an overview of all the organisms covered by the projects and their completion status. The status can be either complete or in progress. Data relative to an organism can be retrieved from the main page of the website using a direct query. Each project has a dedicated webpage and the user can cross-reference the information with other databases at NCBI such as Refseq or Genbank.

BioProject Page

A BioProject has a dedicated webpage, which contains information about the genome lineage, the project status, external resources, and the genomic records linked to other databases in NCBI. These records contain information about the nucleotide sequences, the number of genes, the number of proteins and their structures, if available. It also contains links to related literature. The data can be viewed using the standard tools available from NCBI. A summary of the experimental methods used in the project is also indicated under the attributes list.

Umbrella Project

It is possible to group related projects that belong to a single collaborative effort, but which are different in terms of the methodology, sample material, or result type, into umbrella projects. Umbrella projects of a BioProject are always indicated in the BioProject page.

Organism Overview

Organism overview is a special type of umbrella project. Unlike the conventional umbrella projects,

which are established on the basis of an organizational link, organism overviews group projects that are derived from the same organism.

Cross-References

- ▶ Metagenomics
- ▶ TaxonRank

References

Pruitt K, Clark K, Tatusova T, Mizrachi I (2011) NCBI bioproject help document. (http://www.ncbi.nlm.nih.gov/books/NBK54015/)

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J (2011) Database resources of the National Center for Biotechnology Information. Nucl Acids Res 39:D38–D51

NCBI Genome Project Resource

- ▶ NCBI BioProject Genome Resources

NCBI Viral Genomes Resources

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

Synonyms

Virus reference genomes

Definition

The NCBI viral genomes resource (http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239) is a comprehensive online portal that provides access to more than 3,800 sequences for more than 2,600 curated viral genomes. The resource also contains the genomes of subviral agents known as viroids. Individual genomes can be accessed using Entrez, the text-based search and retrieval system of NCBI; or from a list of viruses, viral groups, or host organisms. Virus species are represented by a reference genome sequence. In the event that a species has more than one strain, variant or isolate, the reference sequence is chosen on the basis of how well characterized it is, as well as its practical importance (Bao et al. 2004). All information in the viral genomes resource is fully integrated with other NCBI databases.

The resource is primarily used in the taxonomic classification of viruses, and as a source for sequences used as standards for developing annotation tools (Brister et al. 2010). More recently, it has been used in constructing virus-host interaction networks and in the development of tools for this purpose (Kozhenkov et al. 2011; Huang et al. 2009), mapping taxonomy standards between the NCBI and the International Committee on the Taxonomy of Viruses (ICTV) (Valdivia-Granda and Larson 2009), and as a source of reference sequences in viral metagenomics studies (▶ Metagenomics, Andrews-Pfannkoch et al. 2010; Marhaver et al. 2008; Dinsdale et al. 2008).

Characteristics

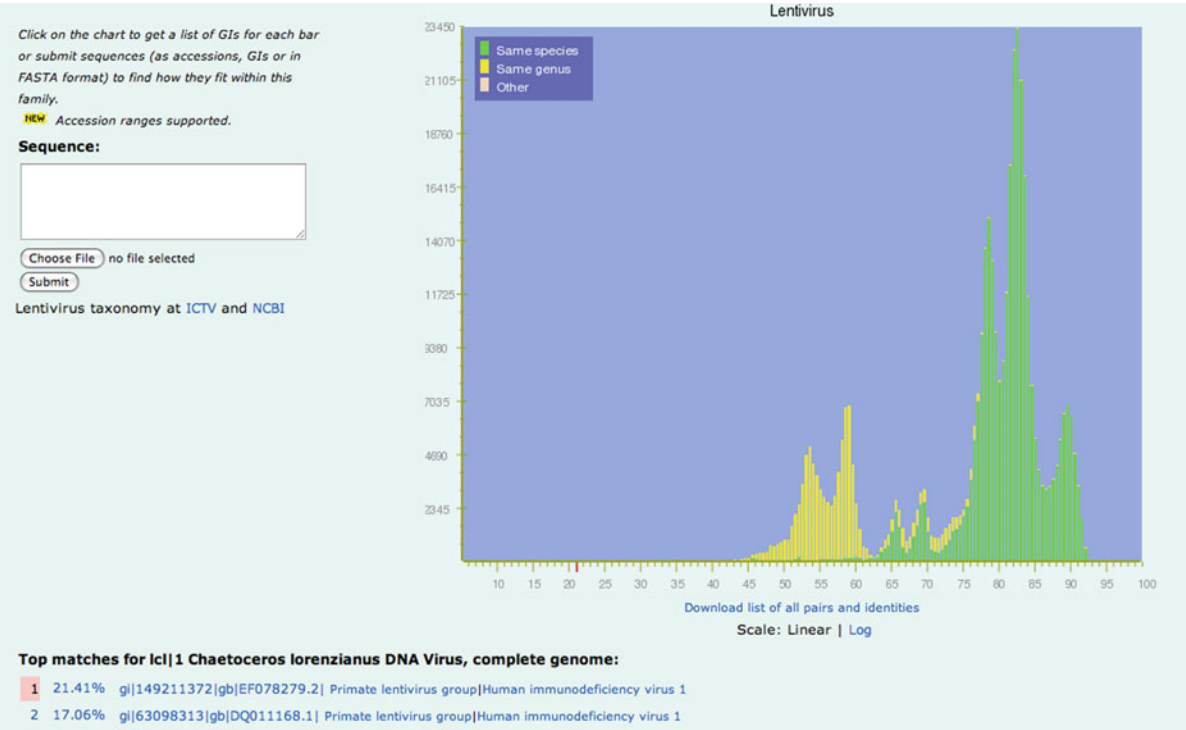
Content

Candidate viral genome sequences are selected automatically from GenBank based on sequence topology, user description, and sequence length with respect to known sequences in a virus genus. Candidate sequences are verified by NCBI curators and external scientific advisors. Sequences are taxonomically classified based on user-specified information, which are generally standardized to conform with ICTV reports.

Tools

Pairwise Sequence Comparison (PASC)

The PASC interface primarily permits the comparison of an external sequence, such as a new viral genome, with genomes in a selected virus family, for which the percentages of identity have been generated from the pairwise global alignments of complete genome



NCBI Viral Genomes Resources, Fig. 1 Frequency distribution of pairwise identities from 917 lentivirus sequences. Note the clustering of the number of virus pairs per percentage identity. The distribution reflects demarcations at different taxonomic levels

sequences (Bao et al. 2004). PASC yields the closest matches, whose positions can be visualized with respect to the identity distribution chart. Apart from placing newly sequenced viruses in a taxonomy group, the identity distributions in PASC can be used for defining taxonomic demarcations and identifying questionable classifications (Fig. 1).

Viral COG: Clusters of Related Viral Proteins (VOG)

The VOG tool was developed for the phylogenetic and functional classification of proteins associated with viral genomes. VOG clusters were generated based on the pairwise alignment of proteins from complete reference viral genome sequences. The VOG link is currently available in the viral genomes resources page (as "Protein Clusters") but updates have been discontinued in 2005. The contents of VOG have been assimilated under Entrez Protein Clusters (http://www.ncbi.nlm.nih.gov/sites/entrez). For each protein cluster, it is possible to view the detailed alignment of sequences, a phylogenetic tree of the proteins, and the cluster patterns, which show the functions associated with the proteins in a cluster.

Related NCBI Resources

Virus-Specific Resources

- The NCBI viral genome database contains links to virus-specific resources. Currently, resources for influenza, retrovirus genomes, and SARS Coronavirus are available. These sites also automate searches for virus-specific information, including literature, across all NCBI databases. Disease-related links are also provided.
- The virus variation resources (VVR) is another set of virus-specific extended resources (http://www.ncbi.nlm.nih.gov/genomes/VirusVariation/index.html), which was extended from the influenza virus sequence database. This resource serves as an alternative interface for retrieving virus sequences based on fields not available in the virus genome resources page, such as genotype, disease severity, collection year, and region of acquisition. The virus variation site also hosts pre-calculated alignment and phylogenetic tools, which permit a more efficient processing of queries (Resch et al. 2009). This resource is currently available for influenza and dengue, and will also be available for the West Nile virus.

External Resource Links

The NCBI viral genomes resource has links to external resources including independently maintained databases for dsRNA viruses, HIV, and plant viruses, among others. Selected viral genomes are included in the NCBI BioProject resources (► NCBI BioProject Genome Resources).

Cross-References

► Metagenomics
► NCBI BioProject Genome Resources

References

Andrews-Pfannkoch C, Fadrosh DW, Thorpe J, Williamson SJ (2010) Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. Appl Environ Microbiol 76:5039–5045

Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S, Rozanov M, Tatusov R, Tatusova T (2004) National Center for Biotechnology Information viral genomes project. J Virol 78:7291–7298

Bao Y, Kapusting Y, Tatusova T (2008) Virus classification by pairwise sequence comparison (PASC). In: Mahy BWJ, Van Regenmortel MHV (eds) Encyclopedia of virology. Elsevier, Oxford, pp 342–348

Brister JR, Bao Y, Kuiken C, Lefkowitz EJ, Le Mercier P, Leplae R, Madupu R, Scheuermann RH, Schobel S, Seto D, Shrivastava S, Sterk P, Zeng Q, Klimke W, Tatusova T (2010) Towards viral genome annotation standards, report from the 2010 NCBI annotation workshop. Viruses 2:2258–2268

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, Mcdaniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Huang FR (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–632

Huang T, Cui WR, He Z, Hu L, Liu F, Wen TQ, Li Y, Cai Y (2009) Functional association between influenza A (H1N1) virus and human. Biochem Biophys Res Commun 390:1111–1113

Kozhenkov S, Sedova M, Dubinina Y, Gupta A, Ray A, Ponomarenko J, Baitaluk M (2011) BiologicalNetworks – tools enabling the integration of multi-scale data for the host-pathogen studies. BMC Syst Biol 5:7–21

Marhaver KL, Edwards RA, Rohwer F (2008) Viral communities associated with healthy and bleaching corals. Environ Microbiol 10:2277–2286

Resch W, Zaslavsky L, Kiryutin B, Rozanov M, Bao Y, Tatusova TA (2009) Virus variation resources at the National Center for Biotechnology Information: dengue virus. BMC Microbiol 9:65–72

Valdivia-Granda W, Larson F (2009) ORION-VIRCAT: a tool for mapping ICTV and NCBI taxonomies. Database 2009: bap014

NcRNA

► microRNA, Disease and Therapy

NcRNA Databases

► Non-coding RNA Databases

Negative Autoregulation

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

Definition

Negative autoregulation (NAR) occurs when a transcription represses the transcription of its own gene. This network motif occurs in about half of the repressors in *Escherichia coli* (Shen-Orr et al. 2002), and in many eukaryotic repressors.

NAR has been shown to display two important functions (Alon 2006, 2007). The first function is response acceleration. NAR was shown to speed up the response to signals. The second function is to increase the stability of the autoregulated gene product concentration and the robustness against stochastic noise; thus it can reduce the cell-to-cell variations in protein levels.

References

Alon U (2006) Introduction to systems biology: design principles of biological circuits. CRC, Boca Raton

Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8:450–461

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

Negative Feedback

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

Definition

Negative feedback is the diminution or counteraction of an effect by its own influence on the process giving rise to it. It occurs when the output of a system acts to oppose changes to the input to the system.

Many biological systems exhibit negative feedback. For example, in hormone secretion, a high level of a particular hormone in the blood may inhibit further secretion of that hormone to maintain the stability of hormone concentration (Chiras 2008).

In gene regulatory networks, negative autoregulation is the simplest motif of negative feedback.

References

Chiras D (2008) Human Biology, Jones & Bartlett Publishers. 7th edn.

Negative Predictive Value

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

Definition

In machine learning, the negative predictive value is defined as the proportion of predicted negatives which are real negatives. It reflects the probability that a predicted negative is a true negative.

Let TP be true positives (samples correctly classified as positive), FN be false negatives (samples incorrectly classified as negative), FP be false positives (samples incorrectly classified as positive), and TN be true negatives (samples correctly classified as negative). The relationship between these prediction outcomes can then be summarized using a confusion matrix (Kohavi and Provost 1998) as illustrated in Table 1.

Negative Predictive Value, Table 1 Sample confusion matrix for two possible outcomes positive and negative

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

The negative predictive value can then be computed using the following equation (Eq. 1)

$$negative\ predictive\ value = \frac{TN}{(TN + FN)} \quad (1)$$

In medical research, the negative predictive value can be used to assess the usefulness of a diagnostic test. However, the negative predictive values depend on the prevalence of disease that is being examined. The significance of its values should be interpreted with caution (Altman and Bland 1994).

As an illustration, suppose that a total of 200 people are tested for a disease. If the prevalence is 15%, 30 people actually have the disease. For a given diagnostic test with sensitivity equal to 67% and specificity 53%, the values of TP, FP, TN, and FN are 20, 80, 90, and 10 respectively. Thus, the negative predictive value is 90%. However, if the prevalence is set to 30%, the negative predictive value will decrease to 74% for the same test with the same sensitivity and specificity values.

Cross-References

▶ Model Cross-Validation
▶ Model Validation, Machine Learning

References

Altman DG, Bland JM (1994) Diagnostic tests 2: predictive values. Brit Med J 309(6947):102
Kohavi R, Provost F (1998) Glossary of terms. Mach Learn 30:271–274

Neoplasia

Barbara Davis
Section of Pathology, Tufts Cummings School of Veterinary Medicine Biomedical Sciences, North Grafton, MA, USA

An excessive, disorderly dysregulated abnormal "new growth" that exceeds and is uncoordinated with the mass of normal tissues, and persists in an excessive manner after the cessation of any stimuli that evoked the change.

Cross-References

► Cancer Pathology

Neoplasms

Carlos Sonnenschein and Ana M. Soto
Department of Anatomy and Cellular Biology, Tufts University School of Medicine, Boston, MA, USA

Synonyms

Cancer; Tumor

Definition

"A neoplasm is an abnormal mass of tissue, the growth of which exceeds and is uncoordinated with that of the normal tissues, and persists in the same excessive manner after cessation of the stimulus which evoked the change" (Willis 1967). The hallmark of neoplasms is altered tissue organization and excessive accrual of cells.

Characteristics

Much has been written about how to define cancer. Entire single- and multiple-author books and reviews have been dedicated to this complex subject (1, 2). The difficulty stems from the lack of uniformity in discriminating whether cancer is a cell-based disease or a tissue-based disease. An additional difficulty is provided by the fact that all comprehensive definitions incorporate features that also manifest in normalcy, for example, excessive proliferation (observed in organ regeneration, ► hyperplasia) or local invasion (embryonic implantation, mammary gland development). Triolo (1965) summarized two tendencies prevailing during the nineteenth century: There were the concept of tissue structure based on the cell and the concept of disease based on the lesion. This dual appreciation has dominated the discussion and generated misunderstandings up to the present. Of note, as technology developed along the last and current centuries, questions about the nature of cancer became more precise.

During the last decades, two tendencies have been noticed: (a) an organizational one that incorporates elements of embryonic fields, epidemiology, immunology, differentiation, wound repair, and organismal organization and (b) a mechanistic tendency based on biochemistry, genetics, and finally molecular biology. This tendency, which is based on hypotheses about molecular mechanisms, is subject to constant change as new technologies enter the experimental laboratory. Regardless of impressive technological advances in molecular biology, light microscopy originally incorporated during the nineteenth century for the diagnosis of cancers has remained as the gold standard. Today, as in the nineteenth century, tumors are diagnosed by pathologists using light microscopes; they ultimately identify tumors by describing their tissue architectural characteristics which provide the oncologist not only a rather precise idea about the origin of the tumor but also a prognosis for the patient.

Core definitions about cancer usually fall short of what a neoplasm is. For example, the classical definition by Willis has been criticized by many, and the corrections introduced by others have also been criticized in turn. Additional concepts are necessary to provide a sense of the dynamic, hierarchical, and interactive properties of biological organization that may be affected by defective controls (Rowlatt 1995). Until the scientific community reaches a consensus definition, we suggest the following definition: The hallmark of neoplasms is altered tissue organization and excessive accrual of cells.

Causes and Explanations of Cancer

A common misperception has been to link the causes of cancers with explanations of the emergence of cancers in humans. By now, the overwhelming majority of the causes of cancers are known. For instance, *viral infections* (hepatitis, HPV, herpes, etc.) have been linked with the eventual appearance of cancers in different tissues or organs where those infections have taken place. *Radiation* (X-, α-, γ-rays) is also known as a cause of cancers. It is equally acknowledged that exposure to *environmental pollutants* (tobacco, DDT, BPA, BP, asbestos, hormones, etc.) and *inflammations* by a diversity of agents (Leishmania, Schistosoma, *Helicobacter pylori*, EBV, etc.) results in the appearance of tumors in exposed populations. However, this wide variety of causes does not explain how they eventually end up forming tumors in affected individuals.

Explanations are proposed by theories of carcinogenesis and metastases. For almost a century, the *cell-centered* somatic mutation theory (SMT) has been the prevalent theory. The tissue organization field theory (TOFT) postulates, instead, the alternative view that neoplasms are a result of faulty histogenesis. Reports about the plausibility of explaining carcinogenesis by experimentally testing the SMT and the TOFT have been published (Maffini et al. 2004, 2005; Bizzarri et al. 2008; Soto and Sonnenschein 2011; Vaux 2011).

Cross-References

- ▶ Cancer
- ▶ Cancer and Environmental Influences
- ▶ Cancer Pathology
- ▶ Classification of Cancer Genesis
- ▶ Oncogene
- ▶ Somatic Mutations
- ▶ Stroma

References

Bizzarri M, Cucina A, Conti F, D'Anselmi F (2008) Acta Biotheor 56:173–1961
Maffini MV, Soto AM, Calabro JM, Ucci AA, Sonnenschein C (2004) J Cell Sci 117:1495–1502
Maffini MV, Calabro JM, Soto AM, Sonnenschein C (2005) Am J Pathol 167:1405–1410
Rowlatt C (1994) In: Iversen OH (ed) New frontiers in cancer causation. Taylor & Francis, Washington, DC, pp 45–58.4
Soto AM, Sonnenschein C (2011) Bioessays 33:332–340
Triolo VA (1965) Cancer Res 25:76–98
Vaux DL (2011) Bioessays 33:341–343
Willis RA (1967) Pathology of tumors. Butterworths, London

Neovascularization

Marsha A. Moses and Di Jia
Department of Surgery/Harvard Medical School, Vascular Biology Program/Children's Hospital Boston, Boston, MA, USA

Definition

Neovascularization is the process of new blood vessel formation. It is essential during normal processes such as embryonic development and reproduction as well as during pathological processes such as tumor growth and progression, tissue repair during wound healing, and certain ocular diseases.

Characteristics

Structure of the Vasculature

In a healthy adult, every part of the body is nourished by blood vessels which carry oxygen, nutrients, growth factors, and cells, and provide normal gas exchange. The vessels are lined by a monolayer of endothelial cells. These cells are in tight association with mural cells (pericytes or smooth muscle cells), together with which they generate a layer of ▶ basement membrane which provides a stable and functional unit. The adult vasculature is quiescent and the endothelial cells rarely divide. This quiescent state is achieved by a fine balance between levels of proangiogenic factors and angiogenesis inhibitors. Under such physiological insults as metabolic stress, low oxygen or pH, mechanical stress, immune and inflammatory challenges, and genetic mutations, the quiescent vasculature can be activated and can initiate the process of neovascularization. In contrast to the normal vasculature, tumor capillaries are structurally and functionally abnormal. They are often leaky with uneven diameters and excessive branching. Endothelial cells lining the capillaries can be irregular in shape and in weak

association with pericytes. As a result, blood flow in the tumor can be chaotic, which can contribute to a hypoxic and acidic environment within the tumor.

Modes of Neovascularization

The most well-studied mode of neovascularization is sprouting from preexisting vessels, a process also known as angiogenesis (Folkman 1971). When challenged by angiogenic mitogens such as vascular endothelial growth factor (VEGF) and fibroblast growth factors (FGFs), the basement membrane of the parent vessel is degraded by matrix metalloproteinases (MMPs), a multi-gene family of metal-dependent enzymes whose activity is the rate-limiting step in ▶ extracellular matrix degradation. This process is regulated by the activity of their cognate inhibitors, the tissue inhibitors of metalloproteinases (TIMPs). Following degradation of the basement membrane, cell junctions of activated endothelial cells are loosened to facilitate the subsequent steps. These cells are equipped with tyrosine kinase receptors such as VEGFRs, which are able to initiate a migratory or proliferative phenotype upon binding of their corresponding ligands. Within the endothelial cell monolayer, cells that lead the way in the branching vessel, called tip cells, are responsible for sensing the angiogenic signals in the microenvironment with their filopodia and directing the sprout. Cells adjacent to tip cells assume a stalk cell fate, elongating the sprout by proliferation. After fusing with neighboring sprouts, junctions between endothelial cells are restored and basement membrane is redeposited by the cells. A pericyte coat soon covers the nascent vessel, which is perfused and becomes functional. In addition to MMPs and TIMPs, this multistep process is tightly regulated by VEGF and Notch signaling pathways and numerous angiogenic regulators including FGFs, platelet-derived growth factors (PDGFs), and angiopoietins (ANGs) (Harper and Moses 2006; Roy et al. 2006).

In addition to sprouting angiogenesis, several other modes of neovascularization have been identified. Vasculogenesis refers to de novo formation of blood vessels. In this process, bone marrow–derived endothelial progenitor cells are recruited to angiogenic sites, incorporate into nascent vessels, and differentiate into mature endothelial cells. Vasculogenesis was first thought to occur only during embryonic development but was later found to play an important role during tumor vascularization as well. Intussusception, another mode of neovascularization, is a process that is faster and more energy efficient than angiogenesis and vasculogenesis in forming new blood vessels. During this process, endothelial cells are remodeled to become larger and thinner, and subsequently form an interstitial pillar and cause the splitting of a mother vessel into two daughter vessels (Carmeliet and Jain 2011).

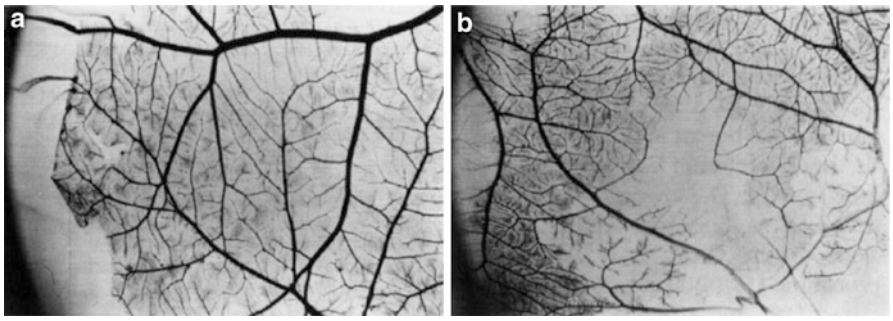
Neovascularization in Health and Disease

Neovascularization is an indispensable process during embryonic development. It is required for the establishment of the circulation system and for organ formation. In adults, neovascularization is restricted temporally to certain physiological situations such as specific stages of the female reproductive cycle (▶ corpus luteum formation and endometrial remodeling during the menstrual cycle) and placenta formation during pregnancy. Insufficient or dysregulated neovascularization is associated with, or can directly lead to, many pathological conditions, including cerebral ischemia, diabetic retinopathy, preeclampsia, moyamoya, wound healing, and cancer.

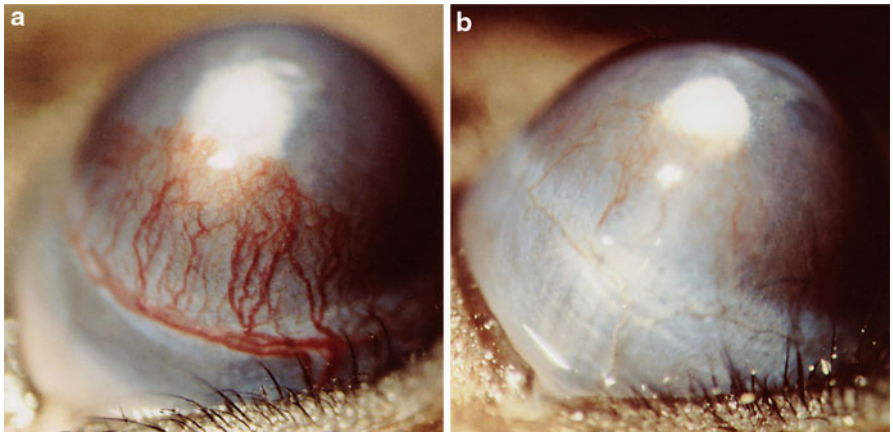
Due to the limitation of oxygen diffusion in a tissue, most solid tumors need to recruit their own vasculature in order to reach a size larger than 1–2 mm in diameter (Folkman 1971). Once vascularized, the blood vessels within the tumor carry oxygen and nutrient supply and remove metabolic waste. In addition, they also function as a conduit through which tumor cells enter the circulation, travel to other sites of the body, extravasate, and form distant metastases. In addition to facilitating tumor growth, the tumor vasculature is also necessary for the successful delivery of therapeutic reagents (Carmeliet and Jain 2000; Nyberg et al. 2008).

Models for Studying Neovascularization

Several models to study neovascularization have been developed over the last few decades. In vitro assays became available after the first successful culture of primary endothelial cells by Folkman and colleagues (Folkman et al. 1979). These assays are based on the activities of endothelial cells during neovascularization: their ability to proliferate, migrate, invade, and form tube-like structures upon their stimulation by angiogenic mitogens. Proliferation assays measure the growth of endothelial cells in culture over a certain period of time. Migration assays quantify the endothelial cell



Neovascularization, Fig. 1 CAM assays showing the inhibition of angiogenesis by cartilage-derived inhibitor (CDI) (Moses et al. 1990). (**a**) Normal vasculature of control CAMs implanted with empty methylcellulose disks. (**b**) Large avascular zones of CAMs implanted with CDI-containing disks



Neovascularization, Fig. 2 Cornea pocket assays showing the inhibition of bFGF-induced angiogenesis by systemic administration of the anti-angiogenic factor Troponin I (Moses et al. 1999). (**a**) Angiogenesis induced by bFGF in control corneas. (**b**) Significant inhibition of bFGF-induced angiogenesis in Troponin I-treated corneas

movement in response to proangiogenic factors or angiogenic inhibitors. Invasion assays measure the ability of endothelial cells to invade through biologically relevant matrices. In tube formation assays, endothelial cells are cultured in a three-dimensional matrix such as Matrigel and the tube-like structures are evaluated by microscopy.

In contrast, the in vivo assays of neovascularization assess the angiogenic or anti-angiogenic potential of cells, proteins, or reagents in a living system. The chicken ▶ chorioallantoic membrane (CAM) assays (Fig. 1) and ▶ corneal pocket assays (Fig. 2) measure the effect of test materials on the vasculature of fertilized chicken embryos and on the avascular animal corneas, respectively. In Matrigel plug assays, cells, tissues, or other test materials are combined with liquid Matrigel which is composed of basement membrane proteins and are injected subcutaneously into animals. At the endpoint of the assay, vascularization of the Matrigel plug can be visualized by immunohistochemistry following removal of the plug, fixation, and sectioning.

Recently, additional in vivo models such as the zebrafish system have been utilized to study neovascularization. Due to the optical clarity of zebrafish embryos, the availability of transgenic lines that help in visualizing the vasculature, and the time- and cost-efficient screening of mutant fish, this model represents a very promising new in vivo model to help us better understand the process of neovascularization (Figg and Folkman 2008).

Cross-References

- ▶ Basement Membrane
- ▶ Chorioallantoic Membrane (CAM) Assay
- ▶ Corneal Pocket Assay
- ▶ Corpus Luteum
- ▶ Extracellular Matrix

References

Carmeliet P, Jain RK (2000) Angiogenesis in cancer and other diseases. Nature 407:249–257

Carmeliet P, Jain RK (2011) Molecular mechanisms and clinical applications of angiogenesis. Nature 473:298–307

Figg WD, Folkman J (2008) Angiogenesis: an integrative approach from science to medicine. Springer, New York

Folkman J (1971) Tumor angiogenesis: therapeutic implications. N Engl J Med 285:1182–1186
Folkman J, Haudenschildf CC, Zetter BR (1979) Long-term culture of capillary endothelial cells. Proc Natl Acad Sci USA 76:5217–5221
Harper J, Moses MA (2006) Molecular regulation of tumor angiogenesis: mechanisms and therapeutic implications. In: Cancer: cell structures, carcinogens and genomic instability. Birkhauser, Basel, pp 223–268
Moses MA, Sudhalter J, Langer R (1990) Identification of an inhibitor of neovascularization from cartilage. Science 248:1408–1410
Moses MA, Wiederschain D, Wu I, Fernandez CA, Ghazizadeh V, Lane WS, Flynn E, Sytkowski A, Tao T, Langer R (1999) Troponin I is present in human cartilage and inhibits angiogenesis. Proc Natl Acad Sci USA 96:2645–2650
Nyberg P, Salo T, Kalluri R (2008) Tumor microenvironment and angiogenesis. Front Biosci 13:6537–6553
Roy R, Zhang B, Moses MA (2006) Making the cut: protease-mediated regulation of angiogenesis. Exp Cell Res 10:608–622

Nested Models

▶ Mixed and Multi-Level Models

Network Alignment

Shihua Zhang[1] and Zhenping Li[2]
[1]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[2]School of Information, Beijing Wuzi University, Beijing, China

Synonyms

Global network alignment; Local network alignment; Network comparison; Network querying

Definition

Network alignment problem is to find a common or approximative subgraph (i.e., a set of conserved edges) across the input networks (Sharan and Ideker 2006). Corresponding to these conserved edges, there exists a mapping between the nodes of the input networks with or without gaps.

The network alignment problem can be formulated in various ways, depending on the kind of input (pairwise vs. multiple alignments) and the scope of node mapping desired (local and global) (Berg and Laig 2004; Berg and Lässig 2006; Flannick et al. 2006; Kelley et al. 2003; Koyutürk et al. 2005; Li et al. 2007; Sharan et al. 2005; Singh et al. 2008; Zaslavskiy et al. 2009).

Cross-References

▶ Comparative Analysis of Molecular Networks
▶ Global Network Alignment
▶ Link Score, Graph Alignment
▶ Local Network Alignment
▶ Multiple Network Alignment
▶ Network Querying
▶ Networks Comparison
▶ Node Score, Graph Alignment
▶ Parameter Estimation, Graph Alignment

References

Berg J, Laig M (2004) Local graph alignment and motif search in biological networks. Proc Natl Acad Sci USA 101:14689–14694
Berg J, Lässig M (2006) Cross-species analysis of biological networks by Bayesian alignment. Proc Natl Acad Sci USA 103:10967–10972
Flannick J et al (2006) Graemlin: general and robust alignment of multiple large interaction networks. Genome Res 16(9):1169–1181
Kelley BP et al (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci USA 100:11394–11399
Koyutürk M, Grama A, Szpankowski W (2005) Pairwise local alignment of protein interaction network guided by models of evolution. RECOM LNBI 3500:48–65
Li Z, Zhang S, Wang Y, Zhang X, Chen L (2007) Alignment of molecular networks by integer quadratic programming. Bioinformatics 24(4):594–596
Sharan R, Ideker T (2006) Modeling cellular machinery through biological network comparision. Nat Biotechnol 24:427–433
Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005) Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci USA 102:1974–1979
Singh R, Xu J, Berger B (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc Natl Acad Sci USA 105(35):12763–12768
Zaslavskiy M, Bach F, Vert JP (2009) Global alignment of protein-protein interaction networks by graph matching methods. Bioinformatics 25(12):i259–i267

Network Alignment, Protein Interaction Networks

▶ Graph Alignment, Protein Interaction Networks

Network Analysis

▶ Graph Mining

Network Building Blocks

▶ Canonical Network Motifs

Network Clustering

Long Jason Lu[1] and Minlu Zhang[2]
[1]Division of Biomedical Informatics, Cincinnati Children's Hospital Research Foundation, Cincinnati, OH, USA
[2]Department of Computer Science, University of Cincinnati, Cincinnati, OH, USA

Synonyms

Graph clustering; Graph partitioning; Network partitioning

Definition

Network clustering (or graph clustering) refers to both a computational problem to extract densely connected but relatively isolated subnetworks from a network and a set of algorithms and methods to solve this problem.

Due to its application-oriented nature in molecular network analysis, the definition and requirement of the output clusters may vary for specific problems and applications. Some methods seek to find only the densely connected subnetworks based on a density criterion while ignoring the rest of the network that are relatively loosely connected, and the output clusters or communities may consist of only a portion of the original network. Other methods output a partition of the whole network, that is, every node in the network must exist in one cluster, possibly by removing a number of edges in the network based on certain measures. In addition, some algorithms allow the output clusters to overlap with each other, that is, a node may be included in two or more clusters, while others do not. The quality of the output clusters can often be measured by quantitative scores, such as the modularity score (Newman 2004; Zhang et al. 2010). The ultimate assessment of the quality of the output clusters is whether the grouping of the nodes in the clusters are biologically meaningful, for example, whether nodes in a cluster from a protein-protein interaction network correspond to a known (or partially known) protein complex or genes/proteins in a signaling pathway.

Cross-References

▶ Biological Applications of Network Modules
▶ Modules in Networks, Algorithms and Methods
▶ Modules, Identification Methods and Biological Function

References

Newman MEJ (2004) Detecting community structure in networks. Eur Phys J B 38(2):321–330
Zhang M, Deng J, Fang C, Zhang X, Lu LJ (2010) Biomolecular network analysis and applications. In: Alterovitz G, Ramoni M (eds) Knowledge-based bioinformatics: from analysis to interpretation. Wiley, Chichester, pp 253–288

Network Comparison

▶ Network Alignment

Network Component Analysis

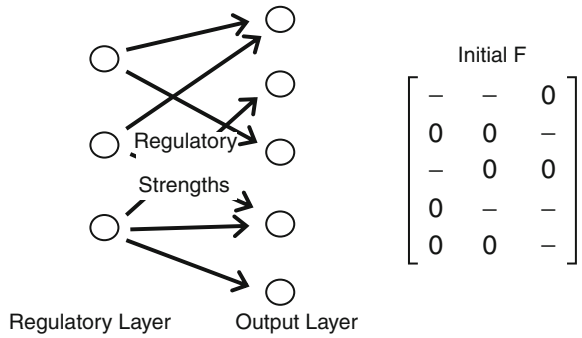
Zhong-Yuan Zhang
School of Statistics, Central University of Finance and Economics, Beijing, China

Definition

Network component analysis (NCA) tries to model gene regulatory network as a bipartite graph whose vertices can be divided into two parts: the regulatory factors and the genes regulated by the factors (indeed, genes interact with each other indirectly through their products such as RNA or proteins).

The expression data of the genes are available by experiments and can be formulated as a matrix X of size $n \times m$ in which the ith row represents the ith gene's expression level across the m time points (or different conditions, or samples). To find the intensities of the regulatory factors and the regulatory strengths, NCA factorizes X into two low-rank matrices F and G such that $X \approx FG^T$, where F is of size $n \times r$ and G is of size $m \times r$, and $r \ll n.m$. The jth column of G is the intensities of the jth regulatory factor, and the ith row of F denotes the sensitivities of the ith gene to the regulatory factors. For example, $F_{ij} = 0$ means that the jth TF does not regulate the ith gene. Different from the traditional matrix factorization models such as principal component analysis (PCA) and independent component analysis (ICA), NCA does not have any statistical constraints on F and G which may not necessarily find biological meaningful components. It takes advantage of partial connectivity information as prior knowledge which is available by experiments and can result in a unique decomposition without considering the scale problem (Liao et al. 2003) if the following three assumptions are satisfied:

- F should have full column rank, which means the regulatory mode of each factor cannot be formulated as a linear combination of the other regulatory modes.
- Each regulatory factor can regulate at most $n - (r - 1)$ genes, which means the graph should be sufficiently sparse.
- G should have full column rank, which means the intensities of each regulatory factor cannot be formulated as a linear combination of the other intensities.



Network Component Analysis, Fig. 1 An example of completely identifiable network in network component analysis. It is a bipartite network. The expression levels of the genes in output layer and a partial knowledge of the regulatory strengths are available, and NCA seeks to identify the intensities of the regulatory factors and reconstruct the underlying network topology. The initial F satisfies the assumptions of NCA. "–" can be arbitrarily replaced by random nonzero values and finally be identified by NCA

In other words, once the above three assumptions are satisfied, the network is identifiable. Figure 1 gives an example of completely identifiable network.

Cross-References

▶ Identification of Gene Regulatory Networks, Machine Learning

References

Liao JC, Boscolo R, Yang Y-L, Tran LM, Sabatti C, Roychowdhury VP (2003) Network component analysis: reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci 100(26):15522–15527

Network Construction, NetSynthesis

Luis Tari
Pharma Early Development Informatics, Hoffmann-La Roche Inc., Nutley, NJ, USA

Definition

A biological network refers to the representation of relationships among various types of biological

entities. Each node in the network refers to a biological entity and a pair of related entities is connected with an edge in the network. An example of a biological network is a protein–protein interaction network, in which the nodes are the proteins and a pair of proteins connected with an edge corresponds to an interaction between the protein pair. NetSynthesis is a method to build such kind of biological networks by means of text-mining queries over Medline abstracts.

Characteristics

While interaction data from manually curated databases are highly useful as a concise resource for biologists, the level of detail about the interactions is a priori defined by the databases. The interactions are often restricted to specific kinds of information so that information one might be interested, such as the structure or strength of the interactions, might not be captured in the databases. Biologists who use these interactions have to be aware of the limitations of the data, which can be unclear if the biologists are not familiar with the curation protocol for the particular database. In other words, biologists can only use the interaction data in a passive manner as they are not engaged in the curation process of the interactions. Biologists can perform filtering or visualization on the interactions provided by the databases as users, but not how the interactions are collected. Such passive use of interactions limits the applicability of the interaction data into research. This presents a single view of the knowledge to the biologists, and it may not be suitable to researchers' specific needs.

NetSynthesis (Tari et al. 2009) is a method that enables users to create biological networks to issue their own keyword-like queries over Medline abstracts. The core idea behind NetSynthesis is through the querying of a specialized database of Medline abstracts. This specialized database differs from traditional indices for document retrieval in the sense that both syntactic and semantic information of sentences and words. The search mechanism of NetSynthesis utilizes both syntactic structures of sentences and semantic information such as biological entities that include proteins, drugs, and diseases. With this approach, users can specify precisely what kind of information, such as entity types, they want in the resulting networks. By using simple-to-use queries to the specialized database of Medline abstracts, these networks convey the information needed by the users, such as strength of the interactions, and such information might be missing in the networks that are synthesized from curated data. In addition, users do not have to depend on the time-consuming curation process and synthesize biological networks from curated data that do not include the latest findings. Such approach is capable of synthesizing biological networks with high precision and even finds relations that have yet to be curated in public databases.

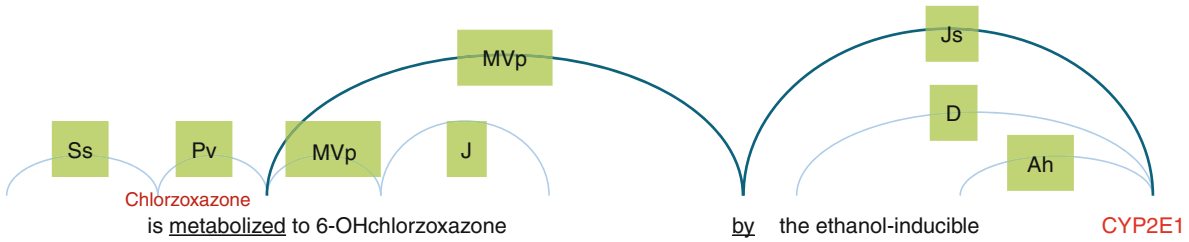
Suppose a user is interested in constructing a network of gene–drug relations, in which the drugs are metabolized by enzymes. The following query can be used:

```
<DRUG> _ metabolized by<GENE>
```

The symbols <DRUG> and <GENE> infer that the sequences of words have to be a drug name and a gene/protein name in the matching sentences. Unlike keyword queries, ordering of the keywords is taken into consideration in our *PTQL^{LITE} query language*. In the rest of the section, we describe the parse tree database as well as the PTQL^{LITE} query language.

Parse Tree Database

The essential component of NetSynthesis is *parse trees* of Medline abstracts; parse trees are syntactic structures that represent the grammatical structures of sentences. Parse trees include *constituent trees* and *linkages*, in which constituent trees are hierarchical syntactic structures of sentences and linkages are composed of *links* that represent syntactic dependencies between pairs of words. Fig. 1 shows an example of a linkage of a sentence that is produced by the Link Grammar parser (Grinberg et al. 1995). These parse trees are generated automatically by the Link Grammar parser. Such parse trees are ideal to be used for expressing linguistic patterns, which are commonly utilized in automated extraction systems. To store the parse trees, a database is needed to capture the hierarchical representation of abstracts, which include the sections of the abstracts such as title or body of the abstracts, parse trees, and the semantic information of words. Semantic information includes the entity type of a sequence of words, such as whether it is a gene/protein name, a drug name, or a disease name. (Gene, protein, and enzyme names are indistinguishable by current automated entity recognizers, and sometimes even by human readers. From here on, we use "gene" to refer to genes/proteins/enzymes.) To cope with the



Network Construction, NetSynthesis, Fig. 1 Linkage of the sentence "Chlorzoxazone is metabolized to 6-OHchlorzoxazone by the ethanol-inducible CYP2E1"

high variation of gene names, an entity recognition system based on a statistical machine learning technique named BANNER (Leaman and Gonzalez 2008) is utilized to identify gene names in text. Lists of drug and disease names from Medical Subject Headings (MeSH) (http://www.nlm.nih.gov/mesh/), DrugBank (http://www.drugbank.ca/), and PharmGKB are employed to recognize drug and disease names. We called the database as the *parse tree database*, and the database is implemented using a relational SQL database. Since standard SQL queries are not ideal for expressing queries that involve linguistic patterns, we develop a query language called *parse tree query language* (*PTQL*) that are used to express linguistic patterns and query parse trees. The details of the PTQL query language and its implementation can be found in Tari et al. (2010). Similar to standard database query languages such as SQL, PTQL is designed to be used by developers and people who are familiar with linguistics.

PTQL^{LITE} Queries

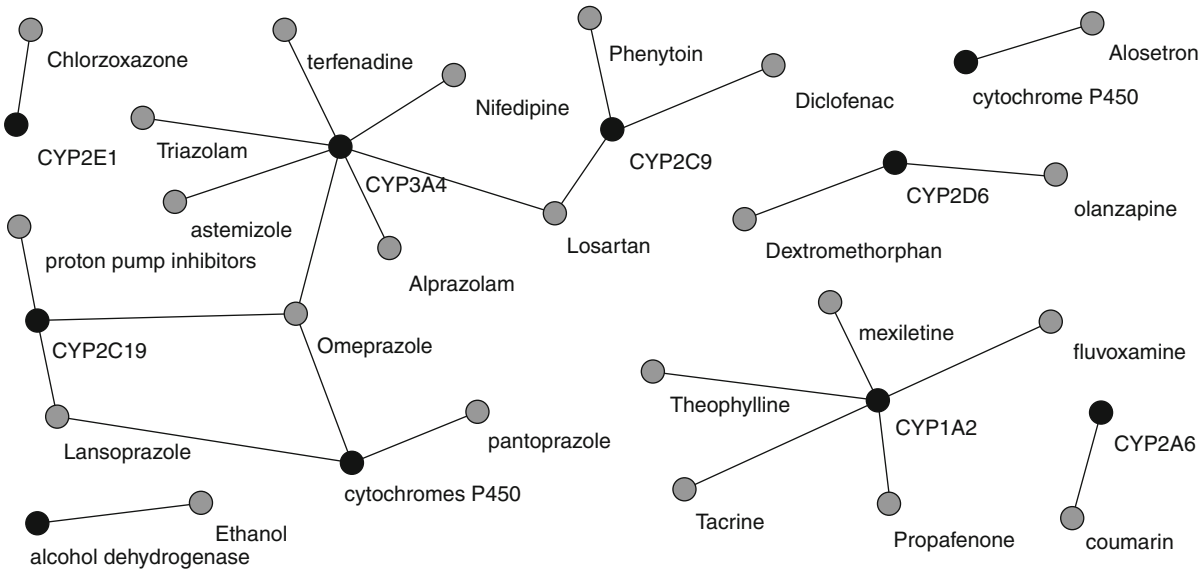
To facilitate the synthesis of biomolecular networks through querying of parse trees of sentences in Medline abstracts by biologists, a simpler query language called *PTQL^{LITE}* is used as the input of NetSynthesis. While *PTQL^{LITE}* queries are not as expressive as PTQL queries, the syntax is close to keyword-based queries used in search engines so that they are easy to use. The sample queries shown in the beginning of this section are PTQL^{LITE} queries. The following query is used to illustrate PTQL^{LITE} queries:

```
<DRUG> _ metabolized by <GENE>
```

The symbols <DRUG> and <GENE> infer that the sequences of words have to be a drug name and a gene/protein name in the matching sentences. The order of the tokens in the query matters, so that the above query specifies that the grammatical structures of the matching sentences include a syntactic dependency between the words "*metabolized*" and "*by*." Similarly, "*by*" has to be syntactically dependent on <GENE>. The operator _ is a wildcard operator that <DRUG> and "*metabolized*" may not have any syntactic dependency between them in the matching sentences. This query can retrieve support evidences such as "Diclofenac is widely used in the treatment of rheumatic diseases and is mainly *metabolized* in the liver *by* CYP2C9"(PMID: 8793607). The grammatical structure of the sentence reveals that there are syntactic dependencies between "*metabolized*" and "*by*," as well as "*by*" and "*CYP2C9*." Table 1 shows sample sentences that are retrieved by the above query, and the resulting network of 33 nodes (10 genes and 23 drugs) with 27 edges is generated from a collection of 13015 Medline abstracts, as shown in Fig. 2. Such kind of drug–enzyme metabolic networks can be used to study how drug metabolism influences the effects of drug chemicals, and genetic variations can affect the effectiveness of drug metabolism. This also allows the discovery of potential relations to draw new hypotheses. For instance, the drugs omeprazole are metabolized by CYP3A4 and CYP2C19, and

Network Construction, NetSynthesis, Table 1 Sample sentences of drug–enzyme relations that are extracted by our approach using the PTQL^{LITE} query "`[DRUG] _ substrates of [GENE] | [DRUG] _ metabolized by [GENE]`"

Gene/drug	Support evidence
CYP2C9/Lovastatin; CYP2C9/Simvastatin; CYP2C9/Atorvastatin	Lovastatin, simvastatin, and atorvastatin are substrates of CYP3A4, whereas *fluvastatin is* metabolized by CYP2C9. (PMID:11029845)
CYP1A2/Propafenone	Propafenone is mainly *metabolized by* CYP2D6 (PMID:10917404)



Network Construction, NetSynthesis, Fig. 2 A gene–drug network in which each edge represents a drug metabolized by an enzyme. Each edge is supported by at least two support evidences

users might want to study a potential relation between CYP3A4 and CYP2C19.

By allowing users to perform their own queries, users can specify their own criteria in their target interactions. One way of specifying the strength of the interaction is to include the word *extensively* in the query as follows:

```
<DRUG> _ extensively metabolized by
<GENE>
```

Here we are interested in drug–enzyme metabolic relations in which the strength of the interactions is described as "extensive." The support evidence "Tacrine is extensively metabolized by CYP1A2." (PMID:9209244) is an example retrieved by the query. There are cases when negative relations are reported in the literature. Our current system simply disregards sentences with words that indicate negation, such as "not," "no," so that sentences such as "Hesperetin was not metabolized by human CYP1A2" (PMID:10781868) are not retrieved as support evidences.

References

Grinberg D, Lafferty J, Sleator D (1995) A Robust Parsing Algorithm For Link Grammars. CMU-CS-TR-95-125

Leaman R, Gonzalez G (2008) BANNER: An executable survery of advances in biomedical named entity recognition. Pac Symp Biocomput 13:652–663

Tari L, Hakenberg J, Gonzalez G, Baral C (2009) Querying parse tree database of Medline text to synthesize user-specific biomolecular networks. Pac Symp Biocomput 14:87–98

Tari L, Tu PH, Hakenberg J, Chen Y, Son TC, Gonzalez G, Baral C (2010) Incremental Information Extraction Using Relational Databases. IEEE Trans Knowl Data Eng 99

Network Decomposition

▶ Top-down Decomposition of Biological Networks

Network Interacting Pattern

▶ Network Topology Motif

Network Measures

▶ Network Metrics

Network Metrics

Jose C. Nacher
Department of Information Science, Faculty of Science, Toho University, Funabashi, Chiba, Japan

Synonyms

Network measures

Definition

Real systems, from metabolic pathways to airline routes, can be represented by graphs composed of a set of nodes and a set of edges that represent interactions between nodes. The structure of a graph can be represented using an adjacency matrix a_{ij} whose entries are 1 if vertices i and j are connected, and 0 otherwise. The information contained in the adjacency matrix is used to analyze the network. Network metrics refer to a variety of mathematical measures that use the adjacency matrix to capture specific properties of the network topology. There are many measures that involve both local and global properties of the network (Albert and Barabasi 2002; Newman 2010). Many concepts behind these metrics were first introduced in earlier studies of social network analysis (Wasserman and Faust 1994).

Characteristics

Node Degree, Distance, and Clustering Degree

The simplest metric to quantify the importance of a node in a network is the *node degree* k that indicates the number of edges connected to it. Highly connected nodes are called ▶ hub nodes. The analysis of the probability to find hubs in a network led to the concept of *scale-free networks,* whose structural features significantly deviate from *random networks.* The *distance* l_{ij} between two nodes i and j is defined as the shortest number of edges to travel from node i to node j. It is often useful to define the *mean path length* which represents the average over the shortest paths between all node pairs. The *diameter* of a network is the largest distance between any two nodes. The *clustering degree* C_i, in contrast, measures the proportion of the number of edges between the neighbors of node i and the maximum number of edges that could exist between the neighbors of node i by using the following expression:

$$C_i = \frac{2n_i}{k_i(k_i - 1)}$$

where n_i indicates the number of edges connecting the k_i neighbors of node i (or equivalently the number of triangles that pass through node i). The total number of triangles that can be constructed through node i is described by $k_i(k_i - 1)/2$.

Networks that combine highly clustering values with a small average shortest path are known as *small-world networks*.

Centrality Measures

While the clustering degree is a property that indicates local order and involves only the neighbors of a given node, centrality measures are computed using the information from the complete network.

The *closeness centrality* CC_i computes the inverse of the distance from a given node i to all other nodes. This distance is defined as the shortest distance l_{ij} between a pair of nodes i and j. Although several normalizations have been proposed, we can write the closeness centrality as follows:

$$CC_i = \frac{n}{\sum_j l_{ij}}$$

A node is classified as a highly central node if its closeness centrality is high; information from a highly central node can quickly reach distant nodes in the network.

The *betweenness centrality* BC_i measures the number of times a node is contained on all possible shortest paths between other pairs of nodes. Nodes with a high betweenness centrality lie on the geodesic paths between many distant pairs of nodes and are able to bridge them together. A targeted removal of nodes with high betweenness leads to a fast network fragmentation. The expression for the betweenness centrality reads as follows:

$$BC_i = \frac{1}{n(n-1)} \sum_{s \neq t \neq i} \frac{l_{st}(i)}{l_{st}}$$

where $l_{st}(i)$ indicates the number of shortest paths that lie on node I from s to t.

Cross-References

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

References

Albert A, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
Newman MEJ (2010) Networks: an introduction. Oxford University Press, Oxford
Wasserman S, Faust K (1994) Social network analysis. Cambridge University Press, Cambridge

Network Modeling of Biochemical Transport Phenomena

Andrés Fernando González Barrios[1], Nubia Velasco[2] and Jorge Mario Gomez Ramirez[1]
[1]Department of Chemical Engineering, Universidad de los Andes, Bogotá, Colombia
[2]Department of Industrial Engineering, Universidad de los Andes, Bogotá, Colombia

Synonyms

Flux balance analysis; Modeling in compartment models

Definition

Network modeling of biochemical transport phenomena is referred to approaches intended to model biochemical systems taking into account the metabolic flux from reaction nature and its variation in concentration based on constitutive equations, mass, and energy balances. Reaction rates are usually determined utilizing optimization algorithms restricted by mass and energy balances in an algorithm coined flux balance analysis (FBA). Nevertheless, regular FBA deals with homogeneous systems; as the metabolites concentration is not a function of position, it is necessary to utilize constitutive equations such as Fourier's and Fick's law, capable of establishing a relation between energy and mass gradients with temperature and concentrations, respectively. These equations are merged with energy and mass balances to determine the temperature and the concentration in any part of the control volume.

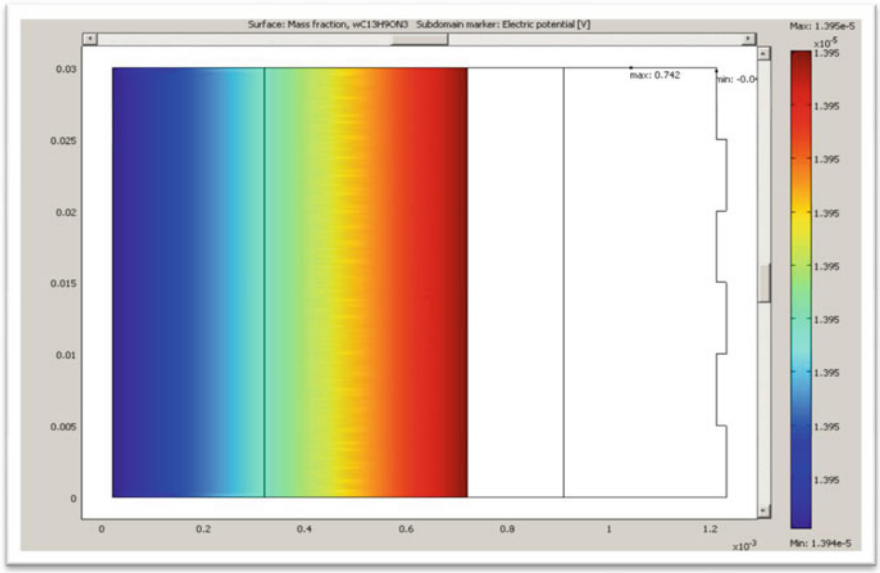
Characteristics

The big outburst of genomic information during the last decade has led to the development of techniques, approaches, and tools that facilitate or aim to interpret the gene function in nature (Stephanopoulos et al. 1998) Nowadays, systems biology which conceives the interaction of biochemical entities (DNA, RNA, and proteins) in a more holistic tactic has fostered the understanding regarding the underpinnings of those interactions and functions of these entities. Mass and energy conservation are the center of the standard methods for modeling biochemical networks which are mostly classified in mass action, stochastic, and optimization based. Nevertheless, the appearing of numerous kinetic constants impedes the development of a model capable of describing the evolution of metabolic flows in a deterministic manner as the number of degrees of freedom is excessive. Then an optimization approach could be utilized to quantify metabolic flows in a biological system. Once the objective function is defined, the equality and inequality restrictions are derived from mass balance and substrate availability, respectively, so we end up having a linear programming problem named flux balance analysis (FBA) (Kim et al. 2008). The optimization problem with m components and n stoichiometric equation can be formulated as follows:

$$\text{Maximize } \upsilon_{cellular\ objective} \tag{1}$$

$$\text{Subject to}: \sum_{j=1}^{m} S_{ij}\upsilon_j = 0 \qquad \forall\, i \in \{1 \ldots n\} \qquad \alpha \le \upsilon_j \le \beta\ \forall\ \in \{1 \ldots m\} \tag{2}$$

where α and β are upper and lower bounds, υ_j represents the fluxes, and S_{ij} the stoichiometric values.



Network Modeling of Biochemical Transport Phenomena,

Fig. 1 Phenazine mass fraction distribution in an air cathode microbial fuel cell model for *Pseudomonas aeruginosa* based on a flux balance analysis and a mass transport 2D model in Comsol® Multiphysics

The fundamental Eq. 2 is easily derived assuming that the cell behaves like a homogeneous reactor regarding the concentration, as it does not consider space variation of the metabolites concentration. Nevertheless, biochemical systems, either biological based such as cells or engineering based such as bioreactors, require the establishment of mathematical equation that relates the concentration gradient with the metabolite flow that in near equilibrium is named Fick's law and can be described for binary systems using the following equation:

$$J_A = -C * D_{AB} * \frac{dX_a}{dz} \tag{3}$$

where J_A is the metabolic flux, D_{AB} is the diffusivity coefficient, and $\frac{dX_a}{dz}$ is the molar fraction gradient with respect to z. Nevertheless, multicomponent diffusion, as is the general case in biochemical systems, demands the use of the Maxwell-Stefan's equation:

$$\nabla x_a = -\sum_{\beta=1}^{N} \frac{x_\alpha x_\beta}{D_{a\beta}} \left(v_\alpha - v_\beta\right) \tag{4}$$

where v_α describes the velocity for component α.

Microbial fuel cells are examples where it is necessary to incorporate the constitutive equations for mass transport in addition to FBA. Figure 1 displays the concentration distribution of phenazine, the electron shuttle in *Pseudomonas aeruginosa* in an air cathode microbial fuel cell model developed by our groups (Mejía et al. 2012). Concentration profiles allow determining if there exist axial dispersion, influence of diffusivity coefficients, and influence of metabolites flow in the distribution of the compound in charge of transporting electrons.

References

Kim HU, Kim TY, Lee SY (2008) Metabolic flux analysis and metabolic engineering of microorganisms. Mol Biosyst 4:113–120

Mejía JD, Rojas CS, Avellaneda F, Urbina D, Velasco Rodríguez N, Vives-Flórez MJ, González Barrios AF (2012) Multi-objective optimization approach of a microbial air-cathode single chamber fuel cell based on metabolic flux analysis and a fuel cell model. Bio Biosys Eng (Submitted)

Stephanopoulos G, Aristodou A, Nielsen J (1998) Metabolic engineering: principles and methodologies. Elsevier, San Diego

Network Modularity

▶ Modularity-based Network Decomposition

Network Motif

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

Definition

Network motifs are patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks.

The present of network motifs in gene regulatory networks was first discovered in transcriptional regulation networks of the bacteria *Escherichia coli* (Shen-Orr et al. 2002), and then in a large set of natural networks (Milo et al. 2002).

Characteristics

Network Motifs in Gene Regulatory Networks

In ▶ gene regulatory networks, network motifs are patterns of genes regulating each other's transcription rate. When analyzing transcription networks, it is seen that the same network motifs appear frequently in diverse organisms from bacteria to human. In *Escherichia coli*, for example, much of the transcriptional interactions are composed of repeated appearances of three highly significant motifs: ▶ feed-foward loop, single-input module (SIM), and dense overlapping regulons (DOR) (Shen-Orr et al. 2002). Each network motif has a specific function in determining gene expression, such as generating temporal expression programs and governing the responses to fluctuating external signals.

The leading hypothesis for the repeated appearances of motifs is that the network motifs were independently selected by the evolution of gene regulation in a converging manner (Babu et al. 2004; Conant and Wagner 2003). Furthermore, both experiments and computational studies on the dynamics generated by network motifs indicate that they have characteristic dynamical functions (Dekel and Alon 2005; Alon 2007; Ma et al. 2009). This suggests that network motifs serve as building blocks in gene regulatory networks that are beneficial to the organism.

There are two types of transcription network: sensory networks that respond to signals such as stresses and nutrients, and developmental networks that guide differentiation events. Network motifs of sensory networks are common to both types of networks, while some motifs are specific for developmental networks.

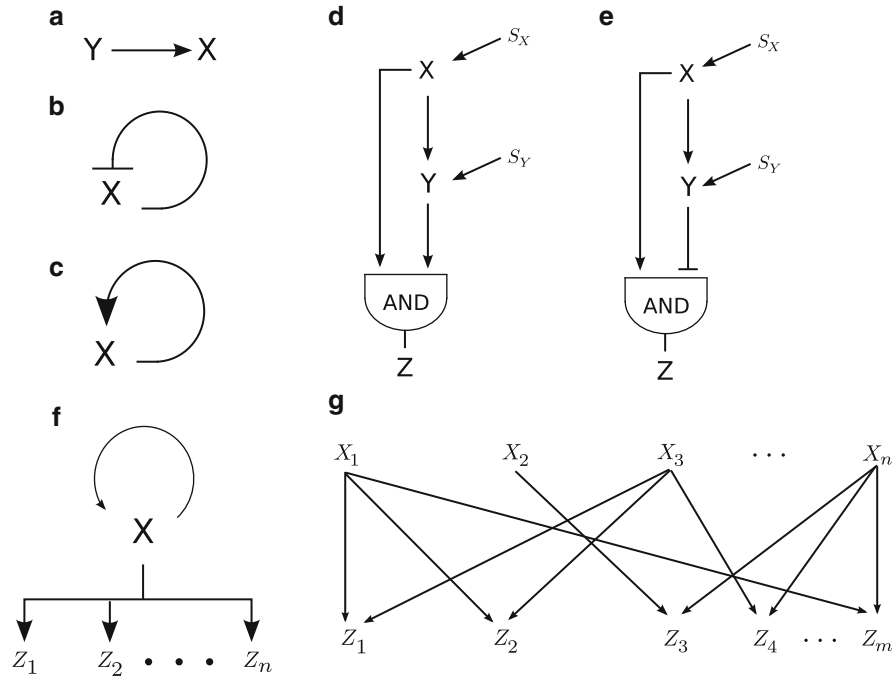
Network Motifs in Sensory Networks

Network motifs in sensory networks include simple regulation and auto-regulation, ▶ feed-forward loop, single-input modules (SIM), and dense overlapping regulons (DOR) (Alon 2006, 2007). The four motif families seem to cover most of the known interactions in the transcription networks of *Escherichia coli* and yeast.

Simple regulation is a basic transcription interaction in which transcription factor Y regulates gene X with no additional interactions (Fig. 1a). The transcription factor Y is usually activated by a signal. The signal can be an inducer molecule that directly binds Y, or a modification of Y by a signal-transduction cascade, and so on. When transcription begins, the concentration of gene product X rises and converges to a steady-state level. When production stops, the concentration of the gene product decays exponentially. In both cases, the response time is equal to half-life of the gene product. The faster the degradation rate, the shorter the response time (Alon 2007).

There are two types of auto-regulations, ▶ negative autoregulation (NAR) in which a transcription factor represses the transcription of its own gene (Fig. 1b), and ▶ positive autoregulation (PAR) in which a transcription factor enhances its own rate of production (Fig. 1c). Usually, NAR accelerates the response time relative to a simple regulation system that has the same steady-state expression level, while PAR slows down the response time. In additional to speeding up response, NAR can reduce cell-cell variation in protein levels, while PAR tends to increase the cell-cell variability (Alon 2007).

▶ Feed-forward loop (FFL) is a family of network motifs. The motif consists of three genes: a regulator X, which regulates Y, and Z, which is regulated by both X and Y. Because each of the three regulatory interactions in the FFL can be either activation or repression, there are eight possible structural types. Two of them are far more common than others in transcription networks (Alon 2006, 2007).



Network Motif,
Fig. 1 Network motifs in gene regulatory networks. (**a**) Simple regulation. (**b**) Negative auto-regulation. (**c**) Positive auto-regulation. (**d**) Coherent type-1 FFL with an AND input function at the Z promoter. (**e**) The incoherent type-1 FFL with an AND input function at the Z promoter. (**f**) The single-input module (SIM) network motifs. (**g**) The dense overlapping regulon (DOR) network motif

The most common form, called ▶ coherent type-1 FFL (C1-FFL) (Fig. 1d), is a sign-sensitive delay element that can protect against unwanted responses to fluctuating inputs. Thus, it can function as a persistence detector, filtering away brief fluctuations from the input signal (Alon 2006).

The second common FFL type, the ▶ incoherent type-1 FFL (I1-FFL) (Fig. 1e), can act as a pulse generator and a response accelerator. This acceleration can be used in conjunction with the other mechanisms of acceleration, such as increased degradation and negative auto-regulation (Alon 2006).

The FFLs in transcription networks tend to combine to form multi-output FFLs, in which X and Y regulate multiple output genes Z_i, $(i = 1, 2, \cdots, n)$. In these configurations, each of the output genes benefits from the dynamical functions that are described above. In addition, the multi-output FFL can generate temporal orders of gene activation and inactivation by means of a hierarchy of regulation thresholds for the different promoters (Alon 2007).

The single-input module (SIM) network motif is a simple pattern in which one regulator regulates a group of target genes (Fig. 1f). The SIM has an interesting dynamical function: It can generate temporal programs of expression, in which genes are turned on one by one in a definite order. This kind of strategy can prevent protein production before it is needed (Alon 2006, 2007).

The DOR network motif is a dense array of regulators that combinatorially control output genes (Fig. 1g). The DORs can carry out decision-making calculations, based on the input functions of each gene (Alon 2006, 2007).

Network Motifs in Developmental Networks

Developmental transcription networks transduce signals into cell-fate decisions. These networks have different constraints: They usually function on the timescale of one of several cell generations, and often need to make reversible decisions that last even after the input signal has vanished (Alon 2007).

Developmental transcription networks use all the network motifs as in sensory transcriptional network. In addition, as a result of their specific requirements, developmental networks use other network motifs that are not commonly found in sensory networks.

Developmental transcription networks often use ▶ positive feedback loops making up of two transcription factors that regulate each other. There are two kinds of positive feedback loops, a double-positive loop and a double-negative loop. The positive feedback loop can display two steady states: In double-positive loop, either both activators are ON or

are OFF; in double-negative loop, one of the repressor is ON, and the other is OFF. In this sense, this network motif can provide memory of an input signal even after the signal is gone. The double-negative feedback loop is often used as a ▶ toggle switch between two different fates (Alon 2007).

In additional to feedback loops, developmental transcription networks tend to have much longer cascades than sensory transcription networks. These cascades pass information on a slow timescale that can be on the order of one cell generation at each cascade step, an appropriate pace for many developmental processes. Development often uses repressor cascades, the timing properties of which can often be more robust to noise in protein production rates than those of activator cascades (Alon 2007).

Network Motifs in Other Biological Networks

In addition to transcription networks, there are composite network motifs that include different types of interactions. One of the most common composite motifs is a negative feedback loop between two proteins, in which one arm is a transcriptional interaction, and the other aim is a protein-protein interaction. The separation of timescales between the slow transcription arm and the faster protein-protein interaction arm might help to stabilize the dynamics of composite loop. Networks of protein modification and synaptic connection between neurons also seem to exhibit network motifs including FFLs connections (Alon 2007).

Detection of Network Motifs

To detect network motifs, one can start with real networks where the interactions between nodes are represented by directed edges. Each network is scanned for all possible *n*-node subgraphs, and the number of occurrences of each subgraph is recorded. The occurrence numbers are compared with those in random networks with the same size and connectivity properties. Network motifs are patterns that occur more often in the real networks than in random networks (Milo et al. 2002).

Open-source software that can detect network motifs from an input network is available (Kashtan et al. 2004). The software accepts network data in the form of a list that details the interactions occurring between different nodes, and outputs the recurring network motifs and depicts these motifs within the network.

Cross-References

▶ Motif
▶ Network Topology Motif

References

Alon U (2006) Introduction to systems biology: design principles of biological circuits. CRC, Boca Raton
Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8:450–461
Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann S (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14:283–291
Conant GC, Wagner A (2003) Convergent evolution of gene circuits. Nat Genet 34:264–266
Dekel E, Alon U (2005) Optimality and evolutionary tuning of the expression level of a protein. Nature 436:588–592
Kashtan N, Itzkovitz S, Mllo R, Alon U (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics 20:1746–1758
Ma W, Trusina A, El-Samad H, Lim WA, Tang C (2009) Defining network topologies that can achieve biochemical adaptation. Cell 138:760–773
Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827
Shen-Orr S, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31:64–68

Network Motifs of Gene Regulatory Networks

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

Synonyms

Building blocks of gene regulatory network

Definition

In the gene regulatory network of *Escherichia coli* and yeast, network motifs refer to those regulatory interacting patterns that occur in original networks at

numbers that are significantly higher than those in randomized networks (Milo et al. 2002; Shen-Orr et al. 2002; Alon 2007). These network motifs are responsible for diverse functions and behaviors of *E. coli* and yeast. They comprise ▶ feed forward loop (FFL), ▶ single-input module motif, SIM, and ▶ dense overlapping regulons (DOR).

Characteristics

Network Motif Structure

A network motif of a regulatory network is composed of *nodes* and *regulations* that connect the nodes. It defines the interacting patterns that are preferred by the gene regulatory network. Many genes in the regulatory network are organized by the preferred regulatory patterns. For example, a network motif may be $A \rightarrow B \rightarrow C$, and the interacting pattern can be employed by the genes in the regulatory network, $crp \rightarrow araC \rightarrow araBAD$.

Detection of Network Motifs

Generation of randomized networks is a key step to detect network motifs.

1. For regulatory networks, the generated randomized networks should have the same incoming and outgoing degree per node as the original network. Two different algorithms with identical results were given to generate randomized networks.

Algorithm A. A Markov-chain algorithm, based on starting with the real network and repeatedly swapping randomly chose pairs of connections ($X1 \rightarrow Y1, X2 \rightarrow Y2$ is replaced by $X1 \rightarrow Y2, X2 \rightarrow Y1$), is employed until the network is well randomized. Switching is prohibited if either of the connections $X1 \rightarrow Y2, X2 \rightarrow Y1$already exists.

Algorithm B. Connectivity matrix was used to a direct construction algorithm. Each network was presented as a connectivity matrix M, such that $M_{ij} = 1$ if there is a connection directed from node i to node j, and 0 otherwise. The goal is to create a randomized connectivity matrix M_{rand}, which has the same number of nonzero elements in each row and column as the corresponding row and column of the real connectivity matrix: $R_j = \sum_j M_{rand,ij}$, $C_i = \sum_i M_{rand,ij} = \sum_i M_{rand,ij}$. To generate the randomized networks, they start with an empty matrix M_{rand}. They then repeatedly choose a row n according to the weights $p_i = R_i / \sum R_i$ and a column m according to the weights $q_j = R_j / \sum R_j$ in a randomized manner. If $M_{rand,mn} = 0$, they set $M_{rand,mn} = 1$. They then set $R_m = R_m - 1$ and $C_n = C_n - 1$. If the entry (m, n). This process is repeated until all $R_i = 0$ and $C_j = 0$.

2. Each of the randomized networks has the same $(n-1)$-node subgraph count as the real network, as a null hypothesis for detecting n-node motifs. This is done to avoid assigning high significance to a structure only because of the fact that it includes a highly significant substructure. To ensure the null hypothesis as a basis for detecting three-node motifs, algorithm A in (1) can preserve the numbers of the in- and outdoing edges for each node, as well as the number of edges and single edges separately. For a random null hypothesis network for assigning significance to the four-node subgraphs, they generate randomized networks that have the same three-node subgraph counts as the real network.

Another step is to count all connected n-node subgraphs in a connectivity matrix **M**. The algorithm loops through all rows I. For each nonzero element (I, j), it loops through all connected elements $M_{ik} = 1, M_{ki} = 1, M_{jk} = 1,$ and $M_{kj} = 1$. This is recursively repeated with elements $(i,k), (k,i), (j,k),$ and (k,j) until n-node subgraph is obtained. This process is repeated for each of the randomized networks. The number of appearances of each type of subgraph in the random ensemble is recorded, to assess its statistical significance.

Last step is to find the network motifs that meet the following criteria:

1. The probability that it appears in a randomized network an equal or greater number of times than in the real network is smaller than $P = 0.01$. In the present study, P was estimated (or bounded) by using 1000 randomized networks.
2. The number of times it appears in the real network with distinct sets of nodes is at least $U = 4$.
3. The number of appearances in the real network is significantly larger than in the randomized networks: $N_{real} - N_{rand} > 0.1N_{rand}$. This is done to avoid detecting as motifs some common subgraphs that have only a slight difference between N_{rand} and N_{real} but have a narrow distribution in the randomized networks.

Software for Network Motif Detection

Mfinder: http://www.weizmann.ac.il/mcb/UriAlon/
FANMOD: http://theinf1.informatik.uni-jena.de/~wernicke/motifs/index.html (Fast)
MAVisto: http://mavisto.ipk-gatersleben.de/

Cross-References

- ▶ Dense Overlapping Regulons
- ▶ Feed Forward Loop
- ▶ Single-Input Module

References

Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8:450–461
Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827
Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31:64–68

Network Partitioning

- ▶ Network Clustering

Network Querying

Shihua Zhang
National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing, China

Synonyms

Network alignment

Definition

Network querying is a special type of network alignment problem (Sharan and Ideker 2006; Zhang et al. 2008). It is to map molecules (such as proteins or genes) of one network of interest (e.g., a complex, a pathway, a functional module, or a general molecular network) to another network or network database for uncovering conserved (sub) networks (Ferro et al. 2007; Pinter et al. 2005; Shlomi et al. 2006).

Cross-References

- ▶ Comparative Analysis of Molecular Networks
- ▶ Multiple Network Alignment
- ▶ Network Alignment
- ▶ Networks Comparison

References

Ferro A, Giugno R, Pigola G, Pulvirenti A, Skripin D, Bader GD, Shasha D (2007) NetMatch: a Cytoscape plugin for searching biological networks. Bioinformatics 23:910–912
Pinter RY, Rokhlenko O, Yeger-Lotem E, Ziv-Ukelson M (2005) Alignment of metabolic pathways. Bioinformatics 21:3401–3408
Sharan R, Ideker T (2006) Modeling cellular machinery through biological network comparision. Nat Biotechnol 24:427–433
Shlomi T, Segal D, Ruppin E, Sharan R (2006) QPath: a method for querying pathways in a protein-protein interaction network. BMC Bioinformatics 7:199
Zhang S, Zhang XS, Chen L (2008) Biomolecular network querying: a promising approach in systems biology. BMC Syst Biol 2:5

Network Targeting

- ▶ Pathway Targeting, Antimycobacterial Drug Design

Network Topology Motif

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

Synonyms

Network interacting pattern; Network motif

Definition

In complex network, network motifs are those patterns of interconnections occurring in original networks at numbers that are significantly higher than those in randomized networks (Alon 2007 and Milo et al. 2002). The interacting patterns are helpful to understand how the nodes in the network interact and how the network is constructed by the subgraphs of the interacting nodes. The subgraphs of network motifs are, thereby, considered as the simple building blocks of complex network. They are widely identified in transcription networks, neuron synaptic connection networks, ecological food webs, electronic circuits, and the World Wide Web.

References

Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8:450–461

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827

Network Visualization and Exchange

► SBGN

Network-based Biomarkers

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

The complex phenotypes observed during the development of a disease are rarely due to single proteins. Hence, recently, it has been shown that protein networks are a source for identifying powerful biomarkers. These biomarker networks in many cases are more useful in predictions rather than the any individual gene.

Transcriptional modules rich in biomarkers can be generated by measuring *coordinately expressed gene expression profiles* in *biofluids*. These "biomarker modules" can further be used to predict new *biomarker networks* in an iterative manner. Such biomarkers based on networks could also be abstracted from the integrative network models of the cellular networks, which are constructed from high throughput proteomics and genomics datasets.

The protein network biomarkers could be identified for the followings:

- Stratification of disease progression from one stage to another: For example, the protein networks obtained from expression profile and/or interaction data from a patient or diseased tissue could be mapped onto a network, which is derived from the expression profile or protein interaction data from healthy individuals and tissues.
- Tissue differentiation: Network biomarkers can help in identifying the process of tissue differentiation.
- Improved interpretation of genome-wide association studies (GWAS): Finally, protein networks may be the key in mining GWAS data to understand the complex diseases, which have multiple genetic loci to play the causal role. The researchers have recently used protein networks to translate GWAS into maps of functional interactions between protein complexes and pathways.

In near future, these concepts are going to be useful in medicine. It is proposed that such networks might be crucial multi-node drug targets for multiple diseases. Further, the new clinical trials with combination drugs should be encouraged to discover the effect toward the disease treatment (Erler and Linding 2010; Azuaje 2010).

Advantages

In case of a complex biological phenomena and diseases, it is very difficult to detect and quantitatively analyze the biomarkers specific to tissue and corresponding diseases by the conventional biomarker discovery approaches, which mostly identify the growth and progression of a single protein molecule. In such cases, network biology–based methods enable us to understand the complex disease mechanism at the system level and facilitate identification of network-based biomarkers.

All the proteins involved therein signify corresponding cellular state and biological functions, which could also be modeled computationally to

observe the dynamics and alteration in biological functions. Such markers are robust, predictive, and quantitative markers or signatures for complex diseases.

References

Azuaje F (2010) Disease biomarkers and biological interaction networks. In: Bioinformatics and biomarker discovery: "Omic" data analysis for personalized medicine. Wiley-Blackwell, Hoboken
Erler JT, Linding R (2010) Network-based drugs and biomarkers. J Pathol 220:290–296

Network-based Pathway Analysis

▶ Metabolic Pathway Analysis

Networks Comparison

Shihua Zhang[1] and Zhenping Li[2]
[1]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing, China
[2]School of Information, Beijing Wuzi University, Beijing, China

Synonyms

Network alignment; Network querying

Definition

Network comparison problem is to compare cellular networks by employing the network topological characteristics as well as biological information of molecules to uncover the similarity or dissimilarity information among networks (Przulj 2006; Rito et al. 2010; Sharan and Ideker 2006). This problem can be analogous to biological sequence comparison and structure comparison. It can be topologically coarse-level comparison such as the comparative analysis of degree distribution, clustering coefficient, diameter, and relative graphlet frequency distribution (Przulj 2006), or can be the discovery of conserved subgraphs among networks. The latter is known as network alignment problem in bioinformatics field (Sharan and Ideker 2006).

References

Przulj N (2006) Biological network comparison using graphlet degree distribution. Bioinformatics 23:e177–e183
Rito T et al (2010) How threshold behavior affects the use of subgraphs for network comparison. Bioinformatics 26(18): i611–i617
Sharan R, Ideker T (2006) Modeling cellular machinery through biological network comparison. Nat Biotechnol 24:427–433

Neurodegenerative Diseases

▶ Disease System, Parkinson's Disease

Niche-defining

Maureen A. O'Malley
Department of Philosophy, University of Sydney, Sydney, NSW, Australia

Definition

In discussions of metagenomics, "niche-defining" refers to the approach whereby metagenomically detected genes and pathways are used to infer biogeochemical conditions and how relevant organisms have adapted to them.

Cross-References

▶ Metagenomics

References

Fuhrman JA (2009) Microbial community structure and its functional implications. Nature 459:193–199
Kowalchuk GA, Speksnijder AGCL, Zhang K, Goodman R, Veen J (2007) Finding the needles in the metagenome haystack. Microb Ecol 53:475–485
Marco D (2008) Metagenomics and the niche concept. Theory Biosci 127:241–247

Nitrogen Fixation, Modeling

▶ Metabolism Nitrogen Fixation

NK Cells, M. Tb Infection

▶ Natural Killer Cells, Mycobacterial Infection

N-K Model

▶ Boolean Networks

NLP Problem

▶ Optimization Algorithms for Metabolites Production

Node Score

▶ Node Score, Graph Alignment

Node Score, Graph Alignment

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

Synonyms

Network alignment; Node score

Definition

A node score of a graph alignment (▶ Graph Alignment, Protein Interaction Networks) A evaluates the alignment quality with respect to the similarity of the nodes of the aligned networks (▶ Protein-Protein Interaction Networks) $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$. Together with the link score (▶ Link Score, Graph Alignment) it forms the scoring function of the graph alignment (▶ Scoring Function, Graph Alignment).

For a pair-wise graph alignment (▶ Graph Alignment, Protein Interaction Networks) A of the networks G_1 and G_2, the node score S_n rewards pairs of aligned nodes $i, j = A(i)$ with a high node similarity R_{ij} (e.g., level of homology, a BLAST bit score) and penalizes similarity between pairs of vertices not respected by the graph alignment:

$$S_n = \sum_{i \in V_1^A} \left[s_1(R_{ij}) + \sum_{j' \in V_2^A \setminus j} w_{ij'}\, s_2(R_{ij'}) + \sum_{i' \in V_1^A \setminus i} w_{i'j}\, s_2(R_{i'j}) \right] \tag{1}$$

The sums run over the aligned nodes only, that is, for a global alignment over all nodes, for a local alignment over subsets of all nodes in $V^A{}_1 \subset V_1$ and $V^A{}_2 \subset V_2$. The factor w_{ij}, which takes the value of 1, when only one of i and j is aligned, and 0.5 when both the nodes are aligned to different partners, prevents over-counting of the node score contributions (Kolář et al. 2008). The functions s_1 and s_2 parameterize the node score and must be set in advance or inferred from the dataset (▶ Parameter Estimation, Graph Alignment). In general, we expect s_1 to be an increasing function of the protein similarity measure R, so that it rewards alignment of more alike proteins, and s_2 to be a decreasing function of the protein similarity R.

For a multiple graph alignment (▶ Graph Alignment, Protein Interaction Networks), the equivalence classes (▶ Graph Alignment, Protein Interaction Networks) are considered instead of the pairs of aligned proteins. The similarity of the proteins within each equivalence class is estimated by inferring a phylogenetic tree relating the species in the alignment (▶ Parameter Estimation, Graph Alignment) and by calculating a weighted sum of pair-wise protein similarities (e.g., BLAST bit scores) (Flannick et al. 2006). The weights of the sum are inferred from the phylogenetic tree relating the species in the alignment (Weighted Sum-of-Pairs Scoring, Altschul et al. 1989).

Cross-References

- ▶ Graph Alignment, Protein Interaction Networks
- ▶ Link Score, Graph Alignment
- ▶ Parameter Estimation, Graph Alignment
- ▶ Protein-Protein Interaction Networks
- ▶ Scoring Function, Graph Alignment

References

Altschul SF, Carroll RJ, Lipman DJ (1989) Weights for data related by a tree. J Mol Biol 207:647–653

Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S (2006) Græmlin: general and robust alignment of multiple large interaction networks. Genome Res 16: 1169–1181

Kolář M, Lässig M, Berg J (2008) From protein interactions to functional annotation: graph alignment in Herpes. BMC Syst Biol 2:90

Noise in Metabolic Networks

- ▶ Stochastic Effects in Metabolic Networks

Noise in Metabolic Pathways

- ▶ Stochastic Effects in Metabolic Networks

Noise, Intrinsic and Extrinsic

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

Definition

Gene expression is a stochastic process. The origin of stochasticity in a cell can be attributed to random transitions among the discrete chemical states. The noise may come in two ways. First, the inherent stochasticity in biochemical processes such as binding, transcription, and translation generates the intrinsic noise. Second, variations in the amounts or states of cellular components or the external environment generate the extrinsic noise. Such noises are believed to play especially important roles when species are present at low copy numbers.

The two kinds of noise can be distinguished by comparing the variation in expression of genes, for example, two genes, cyan and yellow fluorescent proteins, within single cells with the variation in expression of these two between different cells. If there is little intrinsic noise, then the two levels of protein expression vary in concert, while if intrinsic noise is high, they vary essentially independently.

References

Elowitz M, Levine A, Siggle E, Swain P (2002) Stochastic gene expression in a single cell. Science 297:1183–1186

Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci USA 99:12795–12800

Non-canonical Pathway of MicroRNA Biogenesis

- ▶ MicroRNA Biogenesis, Regulation

Non-classical Computation

- ▶ Unconventional Computation

Non-coding Intergenic Sequences

- ▶ Genomic Databases

Non-coding RNA

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

Synonyms

Non-protein-coding RNA

Definition

A non-coding RNA (ncRNA) is a functional RNA molecule that is not translated into a protein. Non-coding RNA genes include highly abundant and functionally important RNAs such as transfer RNA (tRNA) and ribosomal RNA (rRNA), as well as RNAs such as snoRNAs, microRNAs, siRNAs, and piRNAs and the long ncRNAs that include examples such as Xist and HOTAIR.

Characteristics

Biological Roles of ncRNA

Non-coding RNAs belong to several groups and are involved in many cellular processes. These range from ncRNAs of central importance that are conserved across all or most cellular life through to more transient ncRNAs specific to one or a few closely related species. The more conserved ncRNAs are thought to be molecular fossils or relics from LUCA and the RNA world.

NcRNAs in Translation

Many of the conserved, essential, and abundant ncRNAs are involved in translation. Ribonucleoprotein (RNP) particles called ribosomes are the "factories" where translation takes place in the cell. The ribosome consists of more than 60% ribosomal RNA; these are made up of three ncRNAs in prokaryotes and four ncRNAs in eukaryotes. Ribosomal RNAs catalyze the translation of nucleotide sequences to protein. Another set of ncRNAs, Transfer RNAs, form an "adaptor molecule" between mRNA and protein. The H/ACA box and C/D box snoRNAs are ncRNAs found in archaea and eukaryotes, RNase MRP is restricted to eukaryotes, and both groups of ncRNA are involved in the maturation of rRNA. The snoRNAs guide covalent modifications of rRNA, tRNA, and snRNAs, and RNase MRP cleaves the internal transcribed spacer 1 between 18S and 5.8S rRNAs. The ubiquitous ncRNA, RNase P, is an evolutionary relative of RNase MRP. RNase P matures tRNA sequences by generating mature 5′-ends of tRNAs through cleaving the 5′-leader elements of precursor-tRNAs. Another ubiquitous RNP called SRP recognizes and transports specific nascent proteins to the endoplasmic reticulum in eukaryotes and the plasma membrane in prokaryotes. In bacteria, Transfer-messenger RNA (tmRNA) is an RNP involved in rescuing stalled ribosomes, tagging incomplete polypeptides and promoting the degradation of aberrant mRNA.

NcRNAs in RNA Splicing

In eukaryotes, the spliceosome performs the splicing reactions essential for removing intron sequences, and this process is required for the formation of mature mRNA. The spliceosome is another RNP often also known as the snRNP or tri-snRNP. There are two different forms of the spliceosome, the major and minor forms. The ncRNA components of the major spliceosome are U1, U2, U4, U5, and U6. The ncRNA components of the minor spliceosome are U11, U12, U5, U4atac, and U6atac.

Another group of introns can catalyze their own removal from host transcripts; these are called self-splicing RNAs. There are two main groups of self-splicing RNAs, these are the group I catalytic intron and group II catalytic intron. These ncRNAs catalyze their own excision from mRNA, tRNA, and rRNA precursors in a wide range of organisms.

In mammals, it has been found that snoRNAs can also regulate the alternative splicing of mRNA, for example snoRNA HBII-52 regulates the splicing of serotonin receptor 2C.

NcRNAs in Gene Regulation

The expression of many thousands of genes is regulated by ncRNAs. This regulation can occur in trans or in cis.

In higher eukaryotes, microRNAs regulate gene expression. A single miRNA can reduce the expression levels of hundreds of genes. The mechanism by which

mature miRNA molecules act is through partial complementary to one or more messenger RNA (mRNA) molecules, generally in 3′ UTRs. The main function of miRNAs is to downregulate gene expression.

A number of ncRNAs are embedded in the 5′ UTRs of protein coding genes and influence their expression in various ways. For example, a riboswitch can directly bind a small target molecule, and the binding of the target affects the gene's activity.

NcRNAs and Genome Defense

Piwi-interacting RNAs (piRNAs) are expressed in mammalian testes and somatic cells; they form RNA-protein complexes with Piwi proteins. These piRNA complexes (piRCs) have been linked to transcriptional gene silencing of retrotransposons and other genetic elements in germ line cells, particularly those in spermatogenesis.

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) are repeats found in the DNA of many bacteria and archaea. The repeats are separated by spacers of similar length. It has been demonstrated that these spacers can be derived from phage and subsequently help protect the cell from infection.

NcRNAs and Chromosome Structure

Telomerase is an RNP enzyme that adds specific DNA sequence repeats ("TTAGGG" in vertebrates) to telomeric regions, which are found at the ends of eukaryotic chromosomes. The telomeres contain condensed DNA material, giving stability to the chromosomes. The enzyme is a reverse transcriptase that carries Telomerase RNA, which is used as a template when it elongates telomeres, which are shortened after each replication cycle.

X-inactive-specific transcript (Xist) is a long ncRNA gene on the X chromosome of the placental mammals, which acts as major effector of the X chromosome inactivation process forming Barr bodies. An antisense RNA, Tsix, is a negative regulator of Xist. X chromosomes lacking Tsix expression (and thus having high levels of Xist transcription) are inactivated more frequently than normal chromosomes. In drosophilids, which also use an XY sex-determination system, the roX (RNA on the X) RNAs are involved in dosage compensation. Both Xist and roX operate by epigenetic regulation of transcription through the recruitment of histone-modifying enzymes.

Cross-References

▶ MicroRNA, Disease and Therapy
▶ MiRNA

References

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308(5725):1149–1154

Eddy SR (2001) Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2(12):919–929

Hirota K, Miyoshi T, Kugou K, Hoffman CS, Shibata T, Ohta K (2008) Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. Nature 456(7218):130–134

Hüttenhofer A, Schattner P, Polacek N (2005) Non-coding RNAs: hope or hype? Trends Genet 21(5):289–297

Poole AM, Jeffares DC, Penny D (1998) The path from the RNA world. J Mol Evol 46(1):1–17

Non-coding RNA Annotation

▶ Non-coding RNA, Classification

Non-coding RNA Databases

Orland Gonzalez and Haroon Naeem
Institute for Bioinformatics, Ludwig-Maximilians-University Munich, Munich, Germany

Synonyms

MiRNA databases; NcRNA databases; Non-protein-coding RNA databases; PiRNA databases; ScaRNA databases; SnoRNA databases; SRNA databases; TRNA databases

Definition

These are databases that provide information on ▶ non-coding RNA (ncRNA), including nomenclature, sequence data, genomic maps, and functional annotation (e.g., targets).

Non-coding RNA Databases, Table 1 Some ncRNA resources

Database	URL[a]	Focus	Taxonomy	Comments
Sequence databases[b]				
NCODE	noncode.org	General	General	Manual curation; tackles the nonuniform classification system of ncRNAs
RNAdb	research.imb.uq.edu.au/rnadb/	General	Mammals	Sequence annotations; expression data; some limited literature curation; focus on regulatory ncRNAs
lncRNAdb	lncrnadb.org	lncRNA	Eukaryotes	lncRNAs with demonstrated biological function; manual curation
miRBase	mirbase.org	miRNA	General	Primary online repository for miRNA sequence data and annotation; stable accessions; target prediction
PMRD	bioinformatics.cau.edu.cn/PMRD/	miRNA	Plants	Integrates data from public databases with in-house data; some limited literature curation; expression profiles
piRNABank	pirnabank.ibab.ac.in	piRNA	human, mouse, rat	Some literature curation; handles redundancy and repetition; cluster information (important for piRNAs)
Plant snoRNA-DB	bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home	snoRNA	Plants	Sequence; expression data; modification sites in targets
sno/scaRNAbase	bioinfo.fudan.edu.cn/snoRNAbase.nsf	snoRNA, scaRNA	General	Expert curation
snoRNA-LBME-db	www-snorna.biotoul.fr	snoRNA, scaRNA	Human	Extensive manual curation; most entries are experimentally verified in humans or close vertebrates; identifies modified nucleotides in target RNAs
tRNADB-CE	trna.nagahama-i-bio.ac.jp	tRNA	General	Computational prediction; partial manual curation; expert comments; includes results from metagenomics data
Vir-MirDB	alk.ibms.sinica.edu.tw	miRNA	Viruses	Computational prediction of miRNA sequences and host targets
Target databases				
NPInter	www.bioinfo.org.cn/NPInter/	General	Several	ncRNA interactions with proteins, mRNA, and genomic DNA; strict manual curation (counterchecked); requires experimental evidence
miRSel	services.bio.ifi.lmu.de/mirsel/	miRNA	Human, mouse, rat	Automated extraction of miRNA and target genes from literature; partial manual curation
TarBase	diana.cslab.ece.ntua.gr/tarbase/	miRNA	General	Curated collection of experimentally supported miRNA targets; info on differential expression in specific tissues
MicroCosm	www.ebi.ac.uk/enright-srv/microcosm/	miRNA	General	Computational prediction
Others				
miR2Disease	mir2disease.org	miRNA	Human	Manually curated associations between microRNA deregulation and diseases
Rfam	rfam.sanger.ac.uk	General	General	Grouping into RNA families

[a]Compiled Nov 17, 2011.
[b]Although classified as sequence databases, some of the following also provide more specialized information, such as functional targets.

Characteristics

From being regarded in the past as mere carriers of information in the process of gene expression, RNAs are now recognized to serve diverse and important non-messenger functions in biological systems. For example, ncRNAs have been shown to be directly involved in translation, splicing, regulation, genome defense, and

the cell cycle (see ▶ Cell Cycle Regulation, microRNAs). Indeed, with respect to the control of gene expression, numerous studies have even demonstrated that they play roles that are just as important as those played by protein transcription factors (Szymanski et al. 2007). This growing appreciation for the importance of ncRNAs has spurred the development of several dedicated databases (see Table 1 for a partial list).

Sequence Databases

Two examples of databases that collect ncRNAs are NONCODE (He et al. 2008) and ncRNAdb (Szymanski et al. 2007). Both derive their data primarily from GenBank, although ncRNAdb supplements this with sequences from the H-Invitational (Yamasaki et al. 2009) and FANTOM3 (Maeda et al. 2006) databases, and NONCODE with literature curation. All the entries in NONCODE are manually curated; more than 80% are from experiments. In addition, the database classifies ncRNAs based on the cellular process in which they take part (e.g., DNA imprinting, RNA editing, etc.) and annotates the molecular mechanism through which they exert their function (sequence base pairing, catalysis, etc.).

In contrast to NONCODE and ncRNAdb, which provide information on ncRNAs in general, some databases focus on either a particular class of ncRNAs (see ▶ Non-coding RNA, Classification) or on a taxonomic group. One example of the former is miRBase (Kozomora and Griffiths-Jones 2011), which is currently the primary online repository for ▶ miRNA sequence data and annotation. Entries in miRBase are either experimentally verified or predicted homologs of miRNAs verified in a related organism. Computational target prediction for the miRNAs (see ▶ MiRNA Target) is provided by its companion resource, MicroCosm (formerly miRBase Targets). In addition to providing data, miRBase also acts as an independent arbiter of miRNA gene nomenclature. Other databases that specialize in specific classes of ncRNA include piRNABank (Lakshmi and Agrawal 2008), snoRNA-LBME-db (Lestrade and Weber 2006), and tRNADB-CE (Abe et al. 2011), which focus on piwi-interacting RNA (piRNA), small nucleolar RNA (snoRNA), and transfer RNA (tRNA), respectively.

Target Databases

As with proteins, ncRNAs are controlled and exert their functions via interactions with other biological molecules. For example, ncRNAs have been shown to (1) regulate the expression of genes via binding to mRNAs, (2) act as factors affecting a protein's function, and (3) be regulated by proteins. Several resources specialize in compiling these interactions. For instance, a class of databases, which includes NPInter, TarBase, miRTarBase, and miRecords, works by manually curating the biomedical literature. Although the interactions contained in these databases, by virtue of the manner in which they were collected, have some form of experimental support, it has been noted that a considerable fraction of the data was derived only from large-scale experiments, where detailed validation of individual pairs was not performed. As a case in point, 75% and 58% of the miRNA-target pairs in human reported by TarBase and miRecords, respectively, originate from the supplementary materials of just two publications (Naeem et al. 2010). Other examples of databases that provide ncRNA target information are miRSel (Naeem et al. 2010), which employs text mining, and PITA and MicroCosm, which use computational prediction (see ▶ MicroRNA Target Prediction).

Cross-References

- ▶ Cell Cycle Regulation, microRNAs
- ▶ MicroRNA Target Prediction
- ▶ MiRBase
- ▶ MiRNA
- ▶ MiRNA Target
- ▶ Non-coding RNA
- ▶ Non-coding RNA, Classification

References

Abe T, Ikemura T, Sugahara J, Kanai A, Ohara Y et al (2011) tRNADB-CE 2011: tRNA gene database curated manually by experts. Nucleic Acids Res 39:D210–D213

He S, Liu C, Skogerbo G, Zhao H, Wang J et al (2008) NONCODE v2.0: decoding the non-coding. Nucleic Acids Res 36:D170–172

Kozomora A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39:D152–D157

Lakshmi S, Agrawal S (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. Nucleic Acids Res 36:D173–D177

Lestrade L, Weber MJ (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res 34:D158–D162

Maeda N, Kasukawa T, Oyama R, Gough J, Frith M et al (2006) Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet 2:e62

Naeem H, Kuffner R, Casaba G, Zimmer R (2010) miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. BMC Bioinformatics 11:135

Szymanski M, Erdmann VA, Barciszewski J (2007) Noncoding RNAs database (ncRNAdb). Nucleic Acids Res 35: D162–D164

Yamasaki C, Murakami K, Takeda J, Sato Y, Noda A et al (2009) H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. Nucleic Acids Res 38:D626–D632

Non-coding RNA Detection

▶ Non-coding RNA, Prediction

Non-coding RNA Gene Finding

▶ Non-coding RNA, Prediction

Non-coding RNA, Classification

Kay Nieselt and Alexander Herbig
Center for Bioinformatics Tübingen, Faculty of Science, University of Tübingen, Tübingen, Germany

Synonyms

Non-coding RNA annotation

Definition

The subject of non-coding RNA classification deals with assigning an unknown sequence to a family or class of RNAs. RNA classification methods can be used in automated categorization of single RNA molecules or in genome-wide annotation screens.

Characteristics

Overview

At the most general level, there are two types of RNA transcripts: messenger RNAs (mRNAs), also referred to as protein-coding RNAs, and non-protein-coding RNAs (ncRNAs). Functional ncRNAs are generally classified by their membership of an ▶ RNA family or an ▶ RNA class. RNA family membership is predominantly defined by sequence homology, while RNA class membership is defined via functional and/or structural similarities. On the other hand, a structured RNA can also be classified according to whether it is a transcript, that is, whether it has an independent promoter and transcription terminator, or whether it is a structural motif and part of an mRNA. Many methods applied to non-coding RNA classification are also used for non-coding RNA prediction (▶ Non-coding RNA, Prediction) (Bompfünewerer et al. 2007; Soldà et al. 2009).

RNA Transcript Classification

The discrimination of coding from non-coding RNAs is an important analytical task, in particular during the annotation of newly sequenced genomes. `RNAcode` is a program that detects and classifies coding regions in multiple sequence alignments based on a statistical model (Washietl et al. 2011). It is similar in spirit to the program `QRNA` of Rivas and Eddy (2001). Using explicit models for structural ncRNAs and protein coding RNAs, `RNAcode` reduces the number of falsely predicted and classified ncRNAs.

Assignment to Defined RNA Families

Assignment via Similarity Search

Members of an RNA family are homologous and, therefore, the sequences of the members of a transcript family are often recognizably similar. Approaches that classify a non-coding RNA on the basis of sequence homology either use alignment methods against a database of known non-coding RNAs or methods that are specific for a particular RNA family.

The most common and also most general approach to the characterization of homologous ncRNAs is to compute a pairwise local alignment of a target RNA sequence with a known RNA sequence. When aligning a given RNA sequence against annotated sequences in a large database, for example, Rfam, fRNAdb, RNAdb, NONCODE, and others (▶ Non-coding RNA Databases), a fast alignment method is needed. The best

known algorithm that allows for sensitive alignments is `blast`, a very fast heuristic of the Smith-Waterman algorithm.

Classification of an RNA sequence using `blast` or other sequence alignment methods is recommended for highly similar sequences. Members of the same RNA family may, however, share rather short regions of high-sequence similarity interrupted by regions with only structural similarity. Methods that allow one to search for short and quite similar sequence regions interrupted by regions of low conservation, which can vary in length, are more appropriate and yield better results than `blast` and similar methods. An example is `fragrep`, which is also one of the methods of choice for ncRNA class annotation.

Assignment via Structure Comparison

Since primary sequences within RNA families can also be poorly conserved, it is mostly impossible to determine membership of an RNA family purely on the basis of primary sequence homology. On the other hand, structural conservation is indirectly hidden in sequence alignments of members of an RNA family. The best indicators are base pairs. Double mutations preserving a base pair are known as compensatory mutations and the process of detecting them is called covariation analysis. Covariation analysis is a specific type of stochastic context-free grammars. Using covariation analysis, covariance models (CMs) are built from multiple sequence alignments of homologous non-coding RNA sequences and their consensus structure. Thus, a CM defines a consensus RNA structure which these sequences have in common. Each CM represents an RNA family (see Meyer 2007 for a practical guide). A database that is built on covariance models is Rfam (▶ Non-coding RNA Databases). Currently, Rfam stores covariance models for more than 2,000 RNA families, including non-coding RNA genes and structured cis-regulatory elements.

For the classification of an RNA sequence `CMsearch` is the most commonly used program. `CMsearch` is part of the toolkit `Infernal` that constructs CMs and finds new members of a family. A very convenient way of classifying an RNA sequence for a matching Rfam family is to use the Web interface of the database.

There also exist specialized methods that classify a sequence as a member of a specific RNA family. The most prominent methods include `tRNAscan-SE` to classify tRNAs, `BRUCE` to classify tmRNAs, and `SRPscan` to classify SRP RNAs.

Assignment to Defined RNA Classes

Determining the class membership of a given RNA sequence is a much harder problem than determining family membership, because members of an RNA class mostly share only structural properties and rarely sequence similarity.

snoRNAs are a class of RNAs whose canonical role is involvement in rRNA maturation. There are two major classes of snoRNAs that are determined by the formation of a local snoRNA-rRNA duplex: the C/D box snoRNAs direct 2-O-methylation of the ribose, while the H/ACA box snoRNAs guide the conversion of uridine nucleotides to pseudouridine. Members of each family are characterized by the presence of conserved sequence motifs in the snoRNA and short sequence tracts complementary to the cognate RNA target, also termed antisense elements. Several specific snoRNA classification programs have been developed (e.g., `snoScan`, `snoGPS`, `fisher`, and `snoReport`). They all have in common that they exploit structural features of snoRNAs.

Another important class of very small ncRNAs is the class of microRNAs (miRNAs). These are involved in the regulation of translation and degradation of mRNAs. Similar to the case of snoRNAs, the classification of miRNAs can be based either on their target or on their typical hairpin structure. For a detailed description the reader is referred to the entry on ▶ MicroRNA Gene Prediction.

Machine-Learning Methods

In order to classify whether a given primary sequence, or even alignment of several sequences containing a conserved structured RNA, belongs to a specific class of RNAs is a typical task for machine learning. Once a model is designed, it can be trained to distinguish either protein-coding from non-protein-coding RNAs or to classify a specific gene family from samples of existing genes.

Machine-learning methods applied to this task often use support vector machines (SVM). An SVM-based ncRNA classification approach needs a kernel function that computes similarity between two RNA molecules and takes the secondary structure into account.

An example for an SVM-based classification method is `snoReport`. The SVM is trained to

recognize the two major classes of snoRNAs, box C/D, and box H/ACA, in multiple sequence alignments. The classification of the multiple alignment is based solely on information about conserved sequence boxes and secondary structure constraints. In comparison with other snoRNA classification methods `snoReport` does not require any target information.

`GraPPLE` is an algorithm and a Web-based tool for classifying ncRNA sequences as functional as well as into Rfam families (Childs et al. 2009). `GraPPLE` uses graph properties derived from the consensus secondary structure of an RNA family, for which an SVM was trained. The SVM is then used to classify an unknown ncRNA.

RNA Motifs

RNA structural motifs play essential roles in RNA folding and interaction with other molecules, such as proteins or mRNAs.

Motifs can be classified into three broad classes based only on sequence, structure or a combination of sequence, and structure.

A prominent database resource for the structure-based classification is SCOR, focusing on internal and hairpin loops.

Many cis-regulatory motifs, such as riboswitches, are located in the 5′ or 3′ UTRs of mRNA sequences. Transterm is a database that provides access to mRNA sequences and associated cis-regulatory elements. Transterm allows, for example, users to search an RNA sequence for known regulatory elements.

Limitations

Generally, most transcripts are either clearly protein-coding or non-coding RNAs, and therefore also mostly easily distinguishable. However, the group of long non-coding RNAs, that is, those ncRNAs that are more than 200 nucleotides long, imposes limitations on currently available methods. For example, members of this class of non-coding RNAs often have neither a pronounced secondary structure nor an open-reading frame and have limited conserved sequence homology with sequences in other species.

Cross-References

- ▶ Markov Chain
- ▶ MiRNA
- ▶ MicroRNA Gene Prediction
- ▶ Model Training, Machine Learning
- ▶ Non-coding RNA
- ▶ Non-coding RNA Databases
- ▶ Non-coding RNA, Classification
- ▶ Non-coding RNA, Prediction
- ▶ RNA Secondary Structure
- ▶ Transcription

References

Athanasius F Bompfünewerer Consortium, Backofen R, Bernhart SH, Flamm C, Fried C, Fritzsch G, Hackermüller J, Hertel J, Hofacker IL, Missal K, Mosig A, Prohaska SJ, Rose D, Stadler PF, Tanzer A, Washietl S, Will S (2007) RNAs everywhere: genome-wide annotation of structured RNAs. J Exp Zool B Mol Dev Evol 308(1):1–25

Childs L, Nikoloski Z, May P, Walther D (2009) Identification and classification of ncRNA molecules using graph properties. Nucleic Acids Res 37(9):e66

Meyer IM (2007) A practical guide to the art of RNA gene prediction. Brief Bioinform 8(6):396–414

Rivas E, Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2:8

Soldà G, Makunin IV, Sezerman OU, Corradin A, Corti G, Guffanti A (2009) An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. Brief Bioinform 10(5):475–489

Washietl S, Findeiß S, Müller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA 17:578–594

Non-coding RNA, Prediction

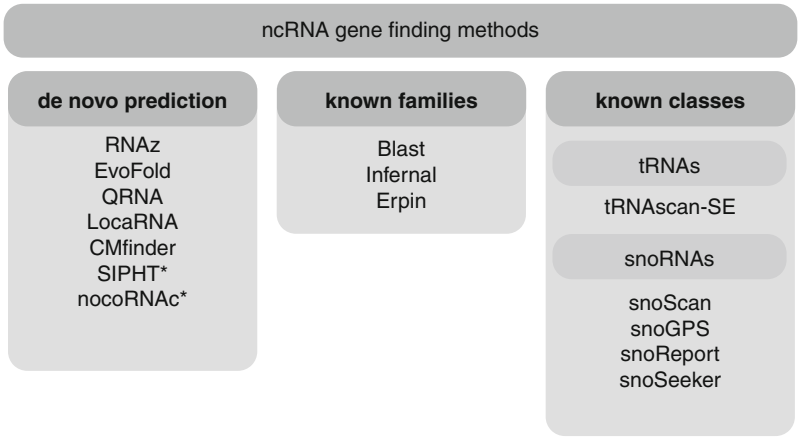
Alexander Herbig and Kay Nieselt
Center for Bioinformatics Tübingen, Faculty of Science, University of Tübingen, Tübingen, Germany

Synonyms

Non-coding RNA detection; Non-coding RNA gene finding

Definition

Non-coding RNA prediction refers to computational methods that predict ▶ Non-coding RNA elements in



Non-coding RNA, Prediction, Fig. 1 Overview of the three approaches to ncRNA prediction and examples of relevant programs. 1. De novo prediction, for which a subgroup of programs (*) is available that specifically predict ncRNA transcripts. 2. Searching for members of a specific family. 3. Searching for members of a specific class. These programs are usually specialized for single classes of ncRNAs

DNA/RNA sequence data, which may be transcribed as independent genes (ncRNA genes) or which are part of other transcripts (ncRNA elements). Methods can be applied to the detection of non-coding RNA elements in general, to the detection of specific RNA families, or to the detection of certain classes of ncRNAs.

Characteristics

Overview

There are three conceptually different approaches to the computational detection of ncRNAs in DNA/RNA sequence data (Bompfünewerer et al. 2007; Meyer 2007; Soldà et al. 2009) (see also Fig. 1).

1. The first approach concerns the de novo detection of ncRNAs. In this case, no specific features like distinct sequences or structural patterns of the elements of interest are known in advance. Most methods of this kind are based on comparative sequence analysis. They depend on the detection of conserved sequences and structures (▶ RNA Secondary Structure) in homologous sequences. In order to detect transcribed ncRNA genes these comparative methods are often combined with the prediction of transcriptional signals such as promoter and transcription terminator signals.
2. If the goal is to search for members of a certain RNA family and representatives of this family are known, it is possible to train a probabilistic model by using sequence and structure information as input. Further members of the same family can be detected by scanning sequence data using this model.
3. Finally, there are methods available that are designed to detect ncRNAs that belong to a certain class (▶ Non-coding RNA, Classification). Members of a class share common structural and/or functional features in the absence of strong sequence conservation. An example is the class of ▶ miRNA. Different miRNA genes differ significantly in their primary sequence, but all pre-miRNAs share several structural properties, which makes their computational prediction feasible.

Methods that are used for the prediction of ncRNAs can also be applied to their classification (▶ Non-coding RNA, Classification).

De Novo Prediction of Non-coding RNAs

Methods Based on Comparative Sequence Analysis

If the aim is the prediction of ncRNAs without the restriction to a certain RNA family or class, methods based on comparative sequence analysis can be applied. These methods do not rely on specific properties of the sequence or structure of the RNAs but search for conserved sequences and structures in general. All of these methods require a prior specification of homologous sequences, which are either aligned by the method itself or have to be aligned before the method is applied (e.g., using `ClustalW`). There are various ways in which the necessary homologous

sequences can be found. As the target of an ncRNA search is often a complete genome or chromosome, a whole-genome alignment with one or more related organisms can be generated, which then serves as input for the ncRNA prediction methods. If the search for ncRNAs is limited to specific regions of an organism's genome, homologous sequences can be searched in DNA sequence databases, using, for example, the fast local alignment search tool BLAST.

Based on these principles several computer programs are available that predict ncRNAs using a comparative approach. RNAz, for example, takes a multiple sequence alignment (MSA) as input and classifies it as containing a conserved ncRNA or not. If the input alignment length is large, it is sliced by a sliding window. RNAz uses an SVM to classify each input alignment. Basically two properties of the alignment are used as input for the classification. First, the so-called z-score:

$$z = \frac{m - \mu}{\sigma}, \tag{1}$$

where m is the average minimum free energy (MFE) of the structures of the sequences in the alignment and μ and σ are mean and standard deviation of the MFE values of a set of random sequences of similar length and base composition. The z-score of the alignment represents the stability of its structure compared to what is expected from a random sequence with similar properties. Here, the rationale is that functional RNAs have a more stable structure than other elements.

The assumption that functional RNAs tend to have a more conserved structure compared with other sequences leads to the second measure, the structure conservation index (SCI):

$$SCI = \frac{E_A}{\bar{E}}, \tag{2}$$

where E_A is the MFE of the consensus structure of the alignment as calculated by RNAalifold and $\bar{E}$ denotes the average MFE of the single sequences. The SCI is a measure of effective structural conservation. If the single sequences fold into a similar structure, their MFE values are close to the MFE value of the consensus structure, which results in an SCI close to 1. If the single sequences fold into dissimilar structures, the SCI is close to 0. The classifier SVM returns a value for the probability that the input alignment contains a structured RNA.

The program EvoFold is based on a comparative probabilistic model. It uses phylogenetic stochastic context-free grammars (phylo-SCFG) to distinguish functional RNA sequences from others. Again, a multiple sequence alignment is taken as input to determine whether the substitution patterns of the alignment columns fit a functional RNA model or a background model, which is also represented by a phylo-SCFG. In addition, EvoFold takes a phylogenetic tree as input providing information about the evolutionary distances between the organisms on which the alignment is based. Therefore, substitution patterns in the alignment can be weighted with respect to the phylogenetic information.

The program QRNA also makes use of SCFGs, but it does not take phylogenetic information into account. Also, the SCFG is only used for the RNA model. For the two background models (protein-coding, other) hidden Markov models (HMM, ▶ Markov Chain) are used. QRNA takes two aligned sequences as input and assigns the alignment to one of the three models.

An example of a program that takes unaligned sequences as input is LocARNA. LocARNA performs a sequence and structure alignment simultaneously and provides a base-by-base conservation profile for structure and sequence, which allows a precise prediction of structured RNAs that are potentially contained in the alignment.

The program CMfinder also takes unaligned sequences as input and produces an SCFG-based covariance model (CM) describing the structural motifs that are found therein. The advantage is that the sequences do not have to be completely homologous unless they do not share a common structural motif. The resulting CM can be used to scan genomic sequences for the discovered motifs.

Methods Based on Transcriptional Feature Detection

In order to predict transcribed ncRNAs several methods, applied mostly to prokaryotic genome sequences, make use of the prediction of transcriptional features like promoter regions and transcription terminator signals. These approaches are then combined with methods based on comparative sequence analysis to detect conserved structures within the predicted transcripts.

SIPHT (Livny et al. 2008) is a web-interface-based program, which combines various methods for finding sequences homologous to regions in the target genome,

predicting transcriptional features, and detecting structural conservation. In a first step, the intergenic regions of the selected bacterial target genome are compared with other bacterial genomes using BLAST. In homologous sequences transcription factor binding sites (TFBS, ▶ Transcription) are searched using position-specific weight matrices. In addition, different methods for the prediction of termination signals are applied. The program QRNA is used for the determination of structural conservation.

The program nocoRNAc (Herbig and Nieselt 2011) uses a slightly different approach. Starting from regions of structural conservation detected by RNAz in a whole-genome alignment of related eubacterial organisms, nocoRNAc uses a specific model for the calculation of DNA duplex stability to predict promoter regions independent of known transcription factor binding site patterns. This is combined with a method for transcription terminator prediction to distinguish transcribed ncRNAs from other structured RNA elements.

Searching for Members of an RNA Family

In the case that there are known representatives of the RNA family for which instances are sought in a target genome, it is possible to build a specific model for this family, which can then be used for searching. A database for such RNA family models is Rfam (▶ Non-coding RNA Databases). This database also offers the Infernal package, which contains software for training new models (CMbuild) or using existing family models for searching (CMsearch). The entries in the database consist basically of the trained CM (the SCFG-based covariance model of the RNA family) and the alignment and consensus structure of the representatives on which the model is based. The program CMsearch takes such a model and a target sequence as input and returns positions of matching regions as well as the respective scores as output. When properly calibrated models are used an e-value is calculated for each hit. An advantage of CMsearch is that it can use prefiltering approaches to limit the search space. It can derive an alignment-HMM (HMMER) from the MSA, which can be used for prefiltering or it can use the local alignment search tool BLAST for this purpose. These filtering steps only take the sequence information into consideration. The complete model, including structure information, is only applied to target regions that have not been filtered out. This approach results in a significant speedup of the search procedure.

Another program that can be used to search for instances of RNA families in complete genomes is Erpin. Erpin is not based on CMs but can still use the alignment and structure information in Rfam entries. It constructs position-specific weight matrices (PSWM) for helices and single-stranded regions found in the consensus structure of the family and calculates the entries of the matrix on the basis of the alignment. These PSWMs are then matched with the target sequences. The advantage of Erpin is that it takes a descriptor file for each RNA family as input. This input can be calculated automatically from an Rfam entry, for example, but it can also be changed manually, which is not feasible for CMs.

Searching for Members of an RNA Class

There are several methods which are designed to search for members of a specific class of ncRNAs. tRNAscan-SE, for example, is a program for searching tRNAs, which can also be considered as an ncRNA family. It can be used via a web interface but is also available as a standalone program. tRNAscan-SE offers a general tRNA model but also more specific models for bacteria, eukaryotes, etc.

Programs for the detection of snoRNAs focus on the localization of the sequence motifs in C/D box snoRNAs and H/ACA box snoRNAs and on the general secondary structure properties of this class. Also, the partial sequence complementarity to targeted snRNAs and rRNAs is considered in some applications. The program snoScan makes use of such target information for predicting C/D box snoRNAs and the program snoGPS specifically predicts H/ACA box snoRNAs. The program snoReport, however, does not need such target information. It predicts C/D box and H/ACA box snoRNAs in single sequences. The program snoSeeker combines CDseeker and ACAseeker to search for C/D box and H/ACA box snoRNAs, respectively, and is able to take whole-genome alignments and deep sequencing data as input.

For the prediction of miRNA genes several methods have been developed, which take the specific sequence and structure properties of this class of RNAs into account (▶ MicroRNA Gene Prediction).

Limitations

Most of the methods in the field of ncRNA prediction are based more or less on the detection of sequence

and structure conservation. This leads to certain limitations:

1. There are ncRNAs that do not exhibit a pronounced secondary structure. Such elements cannot be found by structure-based methods.
2. Many ncRNAs exhibit low sequence conservation, so that homologous sequences for comparative analysis are difficult to find.
3. Some ncRNAs are found only in one species, which limits the applicability of all comparative approaches in general.

Cross-References

- ▶ Markov Chain
- ▶ miRNA
- ▶ MicroRNA Gene Prediction
- ▶ Non-coding RNA
- ▶ Non-coding RNA, Classification
- ▶ Non-coding RNA Databases
- ▶ RNA Secondary Structure
- ▶ Transcription

References

Athanasius F Bompfünewerer Consortium, Backofen R, Bernhart SH, Flamm C, Fried C, Fritzsch G, Hackermüller J, Hertel J, Hofacker IL, Missal K, Mosig A, Prohaska SJ, Rose D, Stadler PF, Tanzer A, Washietl S, Will S (2007) RNAs everywhere: genome-wide annotation of structured RNAs. J Exp Zool B Mol Dev Evol 308(1):1–25

Herbig A, Nieselt K (2011) nocoRNAc: characterization of non-coding RNAs in prokaryotes. BMC Bioinformatics 12:40

Livny J, Teonadi H, Livny M, Waldor MK (2008) High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. PLoS 3(9):e3197

Meyer IM (2007) A practical guide to the art of RNA gene prediction. Brief Bioinform 8(6):396–414

Soldà G, Makunin IV, Sezerman OU, Corradin A, Corti G, Guffanti A (2009) An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. Brief Bioinform 10(5):475–489

Non-contextual Conditions

- ▶ Intrinsicality

Nondeterministic Polynomial-Time Hard

- ▶ NP-hard

Non-empirical Values

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

Definition

Nonempirical values serve to delineate specific distinctions of scientific knowledge beyond empirical adequacy. Such values express requirements of *significance* and *confirmation*. The former are influential on the choice of problems and the pursuit of theories, the latter contribute to assessing the bearing of evidence on theory. Nonempirical values maybe epistemic (i.e., truth related) or non-epistemic (i.e., pragmatic, ethical, or utilitarian).

Background

The background of the claim that nonempirical values contribute to shaping the system of scientific knowledge is constituted by the Duhem–Quine underdetermination thesis (underdetermination). This thesis says that the agreement of the empirical consequences of a theory with the available observations is not a sufficient reason for accepting the theory. In other words, logic and experience leave room for conceptually incompatible but empirically equivalent explanatory alternatives. Consider the example of explaining a bunch of phenomena by a plethora of hypotheses, using, say, one hypothesis for each phenomenon, in contrast to accounting for the same class of phenomena by one overarching principle. Both approaches come out empirically equivalent, yet the scientific community would unanimously pick the latter. This trivial case reveals that the choice of hypotheses in science is governed by values transcending empirical adequacy.

Characteristics

Nonempirical values serve two chief purposes in science: they contribute to establishing the significance

and the confirmation of knowledge claims. Nonempirical values may be epistemic or non-epistemic. Epistemic values are supposed to characterize the merit inherent in certain kinds of knowledge, while non-epistemic values express the usefulness of kinds of knowledge for accomplishing certain practical ends. Epistemic values distinguish knowledge intrinsically worth knowing, while non-epistemic values determine the appropriateness of knowledge for instrumental use. Truth is among the pivotal epistemic values of science, but is of limited suitability for directing scientific research. First, science does not strive for truth simpliciter, but for relevant or significant truth. Second, truth is difficult to recognize, and thus more easily accessible indicators need to be employed. Epistemic values offer guidance in both respects. They provide measures of epistemic significance and standards of credibility that hypotheses need to satisfy in order to pass as acceptable.

Significance Relations

Epistemic values delineate the goals attributed to science as a knowledge-seeking enterprise. For instance, scientists strive for knowledge that is valid in a wide domain; they appreciate universal principles. At the same time, they rate highly precision and correspondingly hold quantitative relations in esteem. Further, scientists search for understanding which is often expounded in terms of the coherence of the views entertained. Knowledge may encompass isolated pieces of information, but understanding demands relations of fit or mutual support among the knowledge elements.

In science, such requirements are typically made more concrete by highlighting causality and unification. That is, relations of cause and effect and of being of the same kind are often taken to be relevant. Speaking more generally, the epistemic significance of a proposition is influenced by its logical content, i.e., by the set of propositions whose validity depends on the truth value of the proposition in question. This is why establishing logically isolated propositions (such as ascertaining the number of leaves of a given tree at a certain time) is considered pointless. By contrast, examining more fundamental or more universal claims is supposed to possess epistemic value. The critical feature is the set of propositions whose acceptability hinges on the truth value of the assumption at issue. Epistemic significance serves as a nontriviality condition of knowledge claims and establishes relevance relations.

Epistemic significance affects the choice of problems and thus contributes to setting up the research agenda. Non-epistemic values are also influential in shaping lines of research. Such values may be pragmatic (such as simplicity in the sense of easy handling of a theory), ethical (such as not to conduct experiments that involve a violation of human rights), utilitarian (taking the technological usefulness of a theory or the prospect of economic benefit tied to it as reasons for working on it), or social (in that knowledge that can be used to the detriment of social groups is required not to be gained (Kitcher 2001)).

Confirmation Relations

Epistemic values are employed in assessing how well a hypothesis is confirmed by the available evidence. Hypotheses need to exhibit certain virtues over and above fitting the phenomena in order to be included in the system of knowledge. For instance, Karl Popper demanded of any confirmed hypothesis that it withstand severe attempts to refute it. It is not enough that the hypothesis agrees with the data. Such an agreement only counts as confirmation if tests that can be expected to reveal mistakes did not produce anomalies or counterinstances. Only when a hypothesis was likely to fail but bears critical scrutiny does the resulting agreement with the observations count as empirical backing. In the same vein, Popper emphasized the importance of predictive success. In contrast to the mere derivation of data from a theory, it is the successful anticipation of novel phenomena, unobserved and unexpected before, that is rightly regarded as support of the theory (Popper 1963). In empirical respect, the explanation of a known fact and the successful prediction of a new fact do not make a difference. Regarding confirmation, appeal to nonempirical values amounts to favoring certain forms of agreement with the observations over other forms.

The assertion that values play a role in testing and confirming theories was prominently defended by Thomas Kuhn (1977). Kuhn claimed that theories are assessed in light of virtues such as accuracy, broad scope, or fruitfulness. Ernan McMullin (1983) coined the term "epistemic values," which was intended to express that the implementation of such values can be presumed to promote the truth-like character of the theory in question. A variety of lists of epistemic values have been proposed (e.g., Longino 1995) that all suffer from their lack of stringency: Items can be abandoned or replaced by others without creating

inconsistencies. An alternative approach to judging scientific theories is to devise more systematic methodological theories that identify such features of excellence from a unified point of view. An example is Bayesianism, which invokes Bayes' rule to evaluate the probability (or "rational credibility") of a hypothesis in light of the evidence. Bayesianism claims to give a coherent account of methodological excellence and thus to provide a rationale as to why these features in contrast to others are to be preferred. Bayesianism also includes epistemic values: high hypothesis probability is an epistemic value, and the features that promote it, namely, high prior probability (Bayesian method) and the increase of the likelihood of the evidence by the adoption of a hypothesis, qualify as epistemic values as well. Such values feature cognitive or explanatory achievements rather than social interests or ethical concerns. They can be linked up with the notion of science as a knowledge-seeking enterprise. After all, a theory that takes account of a wide realm of phenomena in a unified and precise fashion and coheres well with other accepted beliefs is what we take scientific knowledge to be all about (Carrier 2008).

Social epistemology focuses on the procedures within the scientific community that govern the assessment of theories. An early example is Robert Merton's "ethos" of scientists. Values like "universalism," i.e., reliance on impersonal, preestablished criteria of evaluation, or "organized skepticism," are demanded to guide the behavior of scientists (Merton 1942). Such values are social in that they are supposed to be inherent to the scientific community, but they also have an epistemic bearing in that their adoption is assumed to promote the quest for truth or understanding. Giving up universalism by excluding social groups from research or abandoning skeptical scrutiny favors the acceptance of ill-supported claims or the premature rejection of possibly true ones.

Epistemic values are invoked for resolving the Duhem–Quine underdetermination by narrowing the spectrum of theoretical alternatives worthy of scientific examination and pursuit. However, the multiplicity of epistemic values may create tensions among them with the result that they fail to support a clear-cut ranking among theoretical competitors. Such values tend to conflict with one another when applied to particular cases and they are too imprecise to guide theory choice unambiguously. As a result, one of the competing theories may appear superior according to some such standards and inferior according to others. This uncertainty of judgment is known as Kuhn-underdetermination or methodological incommensurability. For instance, as judged in the 1910s, classical electron theory had a larger domain of application than special relativity but the latter excelled in explanatory power in that a few principles covered a wide range of phenomena. Epistemic values often do not provide a basis for unambiguously rating one rival account over the other (Kuhn 1977; Carrier 2008).

Non-epistemic Values

The practical relevance of science suggests the importance of non-epistemic values (Carrier 2010). Richard Rudner (1953) argued that non-epistemic considerations should play an essential role in judging hypotheses. Any hypothesis appraisal is fallible and may thus always produce false positives or false negatives (Error of type I and type II). A high threshold level of acceptance reduces the risk of false positives, but increases the odds of false negatives. Rudner's suggestion is that weighing the non-epistemic consequences of these potential errors should bear on the threshold of acceptance. However, it was pointed out in the subsequent debate that accepting a hypothesis is not tantamount to acting on the basis of this hypothesis. The practical impact of research, upon which Rudner's argument draws, only emerges by the decision to take certain action by relying on the relevant beliefs. Yet in general, the same set of beliefs leaves room for a variety of actions with different practical aftermath.

The more general point is that the assessment of hypotheses requires balancing the risks of false positives and false negatives. Heather Douglas (2000) emphasized that many factors in the design of a study affect its sensitivity in detecting false positives or false negatives, respectively. It is not solely the choice of a threshold of acceptance, but a lot of decisions about procedures used for providing relevant materials or classifying results that affect how suitable tests are for detecting mistakes of either kind. Adjusting sensitivity such that certain errors are more probably revealed than others affects the bearing of the data on the assessment of the hypothesis. Since large parts of research today have serious practical ramifications, Rudner's basic claim that finding the appropriate balance between false positives and false negatives demands the appeal

to non-epistemic values is taken seriously in many quarters. This amounts to granting non-epistemic values some influence in the context of justification.

Cross-References

- ▶ Bayes Rule
- ▶ Bayesian Method
- ▶ Error of Type I and Type II
- ▶ Social Epistemology
- ▶ Underdetermination

References

Carrier M (2008) The aim and structure of methodological theory. In: Soler L, Sankey H, Hoyningen-Huene P (eds) Rethinking scientific change and theory comparison: stabilities, ruptures, incommensurabilities? Springer, Dordrecht, pp 273–290

Carrier M (2010) Knowledge, politics, and commerce: science under the pressure of practice. In: Carrier M, Nordmann A (eds) Science in the context of application. methodological change, conceptual transformation, cultural reorientation. Springer, Dordrecht, pp 11–30

Douglas H (2000) Inductive risk and values. Philos Sci 67:559–579

Kitcher P (2001) Science, truth, democracy. Oxford University Press, Oxford

Kuhn TS (1977) Objectivity, value judgment, and theory choice. The essential tension. Selected studies in scientific tradition and change. University of Chicago Press, Chicago, pp 320–339

Longino H (1995) Gender, politics, and the theoretical virtues. Synthese 104:383–397

McMullin E (1983) Values in science. In: Asquith P, Nickles T (eds) PSA 1982 II. Proceedings of the 1982 biennial meeting of the philosophy of science association: symposia, philosophy of science association, East Lansing, pp 3–28

Merton RK (1942) The normative structure of science. The sociology of science. Theoretical and empirical investigations. University of Chicago Press, Chicago, pp 267–278, 1973

Popper KR (1963) Conjectures and refutations. The growth of scientific knowledge. Routledge, London, 2002

Rudner R (1953) The scientist qua scientist makes value judgments. Philos Sci 20:1–6

Nonlinear Dynamical Systems Theory

- ▶ Cell Cycle Model Analysis, Bifurcation Theory

Nonlinear Dynamics, miRNA Circuits

Julio Vera[2], Svetoslav Nikolov[1] and Xin Lai[2]
[1]Institute of Mechanics and Biomechanics-BAS, Bulgarian Academy of Science, Sofia, Bulgaria
[2]Department of Systems Biology and Bioinformatics, Institute of Computer Science, University of Rostock, Rostock, Germany

Synonyms

ODE modeling of *miRNA* regulation

Definition

MicroRNAs are small regulatory RNAs of ~22nt length that bind to specific messenger RNAs regulating their activity and stability and, therefore, the availability of the translated protein. Together with ▶ transcription factors or other proteins, messenger RNAs and small molecules, *miRNAs* are embedded in regulatory networks that are rich in nonlinear dynamical mechanisms. Mathematical models in ordinary differential equations (a.k.a. kinetic models) are useful tools to investigate the role of microRNAs in the organization and functioning of those networks.

Characteristics

miRNAs are kinds of non-coding RNAs which can regulate the activity and stability of the target *mRNAs* through base-pair matching. When a *miRNA* binds to a target messenger RNA target, it can induce deadenylation of the target which is typically followed by degradation of the *mRNA* (▶ Target Cleavage), or it can repress protein synthesis of the target by blocking translation initiation or elongation, or by causing early translation termination (▶ MicroRNA Target Regulation). Rarely observed in animals but common in plants, *miRNAs* can direct argonaute-catalyzed cleavage of the target *mRNA* when the ▶ target site exhibits extensive sequence complementarity (target cleavage).

MicroRNAs are embedded in regulatory networks rich in complex nonlinear dynamical mechanisms that are conducted by positive and negative feedback loops

and all kinds of transcriptional and post-transcriptional regulation. These mechanisms modulate networks in a time-dependent manner. In order to understand the organization and functioning of these networks, they can be represented mathematically using models in ordinary differential equations. These models reflect rates of changes of molecular quantities over time, their activation status, compartmentalization, and interaction with other partners. A simplified kinetic model in ordinary differential equations for the *miRNA* regulation of a target gene has the following structure:

$$\frac{d}{dt}mRNA = k_{syn\_mRNA} \cdot TF_g - mRNA \cdot (k_{deg\_mRNA} + k_{ass\_miR} \cdot miR)$$

$$\frac{d}{dt}[mRNA|miR] = k_{ass\_miR} \cdot mRNA \cdot miR - k_{deg\_CpX} \cdot [mRNA|miR]$$

$$\frac{d}{dt}miR = k_{syn\_miR} \cdot TF_{miR} - miR \cdot k_{deg\_miR} - k_{ass\_miR} \cdot mRNA \cdot miR$$

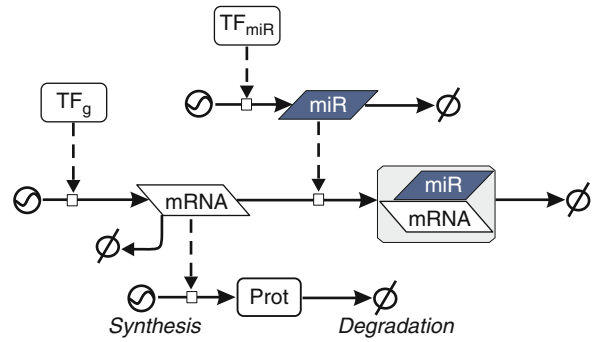
$$\frac{d}{dt}Prot = k_{syn\_prot} \cdot mRNA - k_{deg\_prot} \cdot Prot + \sum_j \cdot Z_j(Prot, \overline{P})$$

The model accounts for the evolution on time of the gene expression levels, *mRNA*, protein product concentration, Prot, the free cytosolic fraction of the targeting *miRNA*, *miR*, and complexes integrating *mRNA* and *miRNA* molecules, $[mRNA|miR]$ (Fig. 1). The following processes are modeled:

For the messenger RNA (*mRNA*): (1) basal synthesis (modulated by the kinetic constant k_{syn_mRNA}), mediated by transcription factors promoting gene expression (TF_g); (2) basal degradation (k_{deg_mRNA}); and (3) association with the *miRNA* into the complex $[mRNA|miR]$ (k_{ass_miR}).

For the *miRNA* (*miR*): (1) basal synthesis (k_{syn_miR}), mediated by its transcription factor (TF_{miR}); (2) basal degradation (k_{deg_miR}); and (3) association with the *mRNA* into the complex $[mRNA|miR]$ (k_{ass_miR}).

For the complex ($[mRNA|miR]$): (1) association of microRNA and messenger RNA that forms a complex (k_{ass_miR}); and (2) degradation/inactivation (k_{deg_CpX}) of the complex.



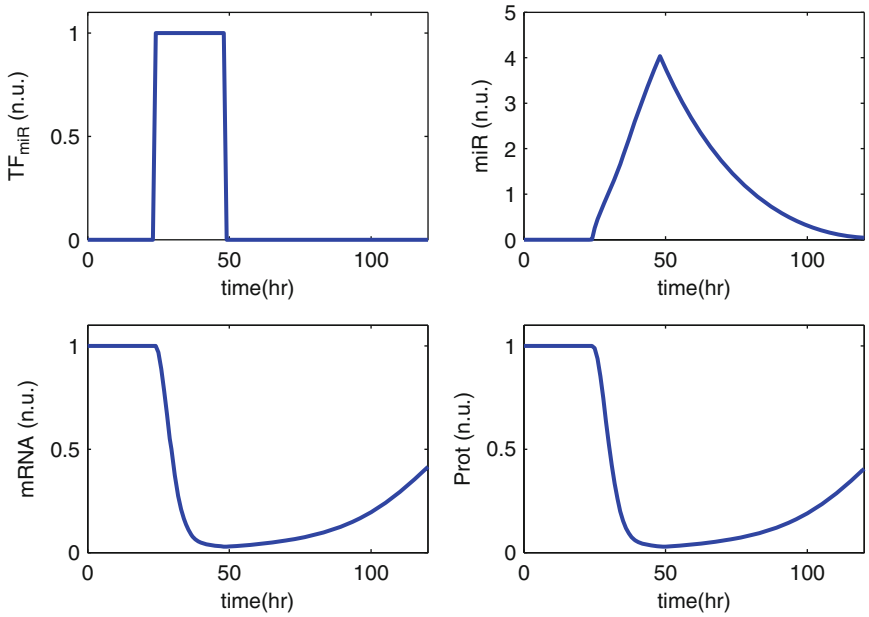
Nonlinear Dynamics, miRNA Circuits, Fig. 1 Schematic representation of a microRNA regulation network. RNA species appear as *parallelograms*, proteins appear as *rectangles*. Legend: *mRNA* messenger RNA, *miR* micro RNA, *Prot* transcribed protein, TF_g Transcription factor promoting gene expression, TF_{miR} Transcription factor promoting *miRNA* expression. The *gray box* represents the complex integrated by miR and the targeted *mRNA*

For the protein levels (Prot): (1) *mRNA*-mediated synthesis of protein (k_{syn_prot}); and (2) basal degradation (k_{deg_prot}). In addition, the protein may undergo further regulatory processes induced by other proteins ($\overline{P}$), which are not considered here in detail. In this simple model, the total measurable amounts for the two RNA types are defined by the equations:

$$mRNA_{TOTAL} = mRNA + [mRNA|miR]$$
$$miR_{TOTAL} = miR + [mp21|miR_i]$$

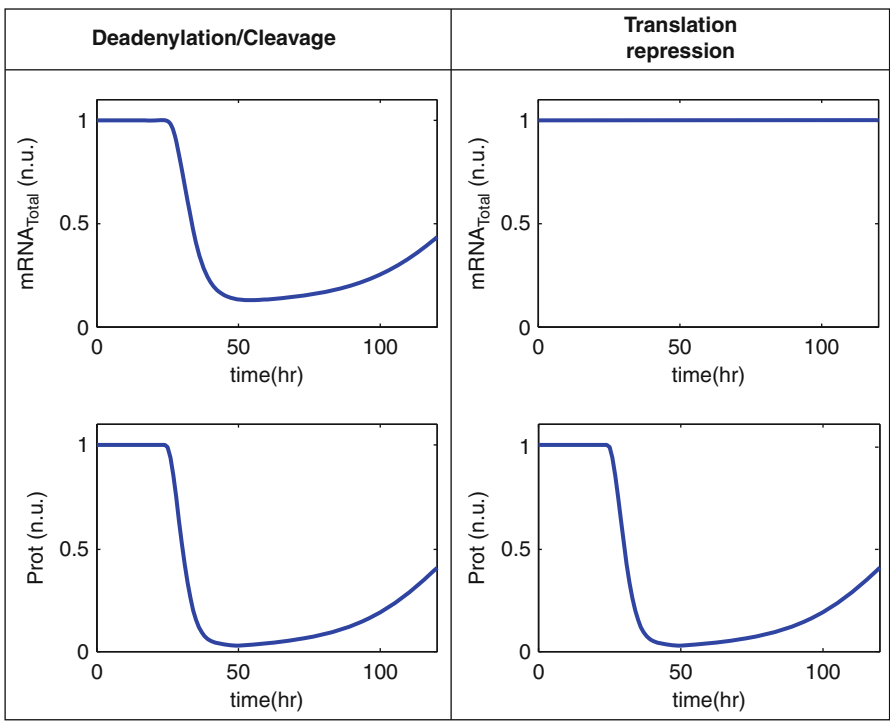
Upon activation of *miRNA* expression, levels of translationally active messenger RNA are reduced, which in turn downregulates protein synthesis and provokes a delayed reduction of protein levels (Fig. 2). Since *miRNAs* have a rather long half-life, the repression exerted by them can last long after TF_{miR} signal termination.

The extent of the *miRNA* post-transcriptional repression is tightly controlled by the efficiency of some of the molecular events described here, especially the *mRNA* and *miRNA* basal turnover and the association of *miRNA* and messenger RNA. The model here described permits to discriminate between microRNA-mediated *mRNA* deadenylation or cleavage (Fig. 3 left) and *miRNAs*-induced inhibition via translation repression (Fig. 3 right). As can be seen, differences between both



Nonlinear Dynamics, miRNA Circuits, Fig. 2 *Simulation of transient miRNA-mediated post-transcriptional regulation.* In the initial configuration of the system, *mRNA* and protein are at basal levels, while *mRNA* is not expressed: $mRNA(0) = 1$; $\text{Prot}(0) = 1$; $\text{miR}(0) = 0$. At t = 24 h, *miRNA* expression is promoted by a pulse-like TF_{miR} activation. The dynamics of the system afterward is simulated. Since *miRNAs* have a rather long half-life, the repression exerted by them can last long after TF_{miR} signal termination. Model parameters used, representative of that kind of systems: $k_{ass\_miR} = 0.25$; $k_{deg\_CpX} = 0.0289$; $k_{deg\_Prot} = 1.3863$; $k_{deg\_mRNA} = 0.0289$; $k_{deg\_miR} = 0.0289$; $k_{syn\_mRNA} = 0.0289$; $k_{syn\_miR} = 0.28910$; $k_{syn\_Prot} = 1.3863$

N



Nonlinear Dynamics, miRNA Circuits, Fig. 3 *Different mechanisms of miRNA-mediated silencing. Left*: *miRNA*-mediated *mRNA* decay through deadenylation or cleavage. *Right*: *mRNA*-mediated translational repression. Initial conditions and parameters are identical to those in Fig 2. The exception is k_{deg_CpX} for *miRNA*-mediated *mRNA* deadenylation ($k_{deg\_CpX} = 0.289$)

mechanisms pertain to the expression level of translationally active messenger RNA.

The model can be further modified to account for additional regulatory properties in *miRNA* regulation. For example, Khanin and Vinciotti (2008) proposed to consider saturation in the target regulation process by using a Michaelis-Menten equation for the rate of association between *mRNA* and *miRNA*:

$$k_{ass\_miR} \cdot mRNA \cdot miR \Rightarrow k_{ass\_miR} \cdot mRNA \cdot \frac{miR}{k_{st} + miR}$$

In case of multiple binding sites to which the same *miRNA* can bind to, the same authors proposed to transform the equation including a Hill coefficient (▶ Hill Equation) to account for cooperativity in the post-transcriptional regulation:

$$k_{ass\_miR} \cdot mRNA \cdot miR \Rightarrow k_{ass\_miR} \cdot mRNA \cdot \frac{miR^h}{k_{st}^h + miR^h}$$

Design Principles in *miRNA* Regulation

Several studies have used kinetic models to investigate basic properties and design principles of *miRNA* regulation. One of the earliest mathematical models that has focused on *miRNA* gene silencing was constructed by Levine and coauthors (2007). They developed a simple quantitative model for microRNA-mediated silencing and used it to investigate how distinctive responses in the *mRNA* and protein expression levels of the target gene are affected by parameters settings, which are thought as target gene specific. Interestingly, they also proposed that global effectors, related to different cellular conditions, distinctively affect different *miRNA* targets. Xie and coauthors (2007) developed a kinetic model including delayed negative feedback to investigate the effect of *miRNAs* on the oscillatory pattern of gene expression, which are caused by the feedback loop structure. Xu and colleagues (2009) combined theoretical analysis and numerical simulations to analyze the effect of microRNA regulatory motifs on the system's robustness to external and stochastic perturbations. Another model accounting for *miRNA* repression at the initial protein translation processes was set up by Nissan and Parker (2008). Zinovyev and coauthors (2010) applied an asymptotic analysis to the same model to investigate whether dynamical data allow for distinguishing between different mechanisms of microRNA regulation. Wang and colleagues (2010) derived a quantitative model of generic *miRNA* pathway, which was implemented deterministically and stochastically, and used it to identify the critical processes in the *miRNA* pathway via sensitivity analysis. In addition, they verified that a microRNA regulation pathway with that structure induces noise reduction and exhibits robustness. Whichard and coauthors (2011) derived a model and used it to investigate which biochemical events in the *miRNA*-mediated post-transcriptional regulation are more prominent for the modulation of gene expression. Using sensitivity analysis, they found that *miRNA* synthesis acts to fine-tune protein concentration. In addition, their model shows that *miRNAs* can exert potent target repression even in low copy number. Nikolov and coworkers (2011) showed how to combine bioinformatics algorithms for microRNA target prediction, database knowledge, and kinetic modeling to investigate in detail the dynamics and function of regulatory networks embedding *miRNAs*, their targets, and transcription factors.

Modeling *miRNA* Regulation in Biomedicine

In recent years, some works have shown how to employ kinetic modeling to investigate the regulation of *miRNA*-regulated networks with biomedical interest. Aguda and coauthors (2008) derived a kinetic model describing the feedback loop system integrated by the ▶ *miRNA* cluster miR-17-92, E2F and Myc. E2F and Myc are two transcription factors involved in the regulation of cell proliferation and apoptosis, which can shift their roles from being oncogenic to tumor suppressor depending on their expression levels. They found that the *miRNA* cluster plays a critical role in regulating the ▶ bistability (off–on switch-like behavior) exhibited by the system. Vohradsky and colleagues (2010) built a model of *miRNA* regulation using microarray data of HepG2 cells transfected with *miRNA*-124a. They identified the genes in those cells repressed by *miRNA*-124a and computed the model parameter values for all the *mRNAs* affected by *miRNA*-mediated regulation. Based on their model, they identified a digital switch-like mechanism of microRNA regulation.

Cross-References

- ▶ Bistability
- ▶ Hill Equation
- ▶ MicroRNA Target Regulation

- ▶ miRNA Cluster
- ▶ Target Cleavage
- ▶ Target Site
- ▶ Transcription Factor

References

Aguda B, Kim Y, Piper-Hunter M, Friedman A, Marsh C (2008) MicroRNA regulation of a cancer network: consequences of the feedback loops involving miR-17-92, E2F, and Myc. Proc Natl Acad Sci USA 105(50):19678–19683

Khanin R, Vinciotti V (2008) Computational modeling of post-transcriptional gene regulation by microRNAs. J Comput Biol 15:305–316

Levine E, Ben Jacob E, Levine H (2007) Target-specific and global effectors in gene regulation by MicroRNA. Biophys J 93:L52–L54

Nikolov S, Vera J, Schmitz U, Wolkenhauer O (2011) A model-based strategy to investigate the role of microRNA regulation in cancer signalling networks. Theory Biosci 130(1):55–69

Nissan T, Parker R (2008) Computational analysis of miRNA-mediated repression of translation: implications for models of translation initiation inhibition. RNA 14:1480–1491

Vohradsky J, Panek J, Vomastek T (2010) Numerical modelling of microRNA-mediated mRNA decay identifies novel mechanism of microRNA controlled mRNA downregulation. Nucleic Acids Res 38:4579–4585

Wang X, Li Y, Xu X, Wang YH (2010) Toward a system-level understanding of microRNA pathway via mathematical modeling. Biosystems 100(1):31–38, Epub 28 Dec 2009

Whichard ZL, Motter AE, Stein PJ, Corey SJ (2011) Slowly produced microRNAs control protein levels. J Biol Chem 286:4742–4748

Xie Z, Yang H, Liu W, Hwang M (2007) The role of microRNA in the delayed negative feedback regulation of gene expression. Biochem Biophys Res Commun 358:722–726

Xu F, Liu Z, Shen J, Wang R (2009) Dynamics of microRNA-mediated motifs. IET Syst Biol 3:496–504

Zinovyev A et al (2010) Dynamical modeling of microRNA action on the protein translation process. BMC Syst Biol 4:13

Nonlinear Model

Maria Rodriguez-Fernandez and Francis J. Doyle III
Department of Chemical Engineering, Institute for Collaborative Biotechnologies, University of California, Santa Barbara, CA, USA

Definition

A nonlinear model is a mathematical model which is not linear, that is, a model structure whose outputs do not satisfy the superposition principle with respect to its inputs, or whose outputs are not directly proportional to its inputs.

Control engineers usually speak of nonlinear models referring to nonlinearity in the inputs (non-LI). However, when statisticians speak of nonlinear models, they usually refer to nonlinearity in the parameters. Analogously to the definition of non-LI, a model structure is said to be nonlinear in its parameters (non-LP) if its outputs do not satisfy the superposition principle with respect to its parameters (Walter and Pronzato 1997).

Cross-References

- ▶ Optimal Experiment Design

References

Walter E, Pronzato L (1997) Identification of parametric models from experimental data. Springer, Berlin

Nonlinear Optimization

- ▶ Mathematics, Nonlinear Programming
- ▶ Nonlinear Programming

Nonlinear Programming

Maria Rodriguez-Fernandez and Francis J. Doyle III
Department of Chemical Engineering, Institute for Collaborative Biotechnologies, University of California, Santa Barbara, CA, USA

Synonyms

Nonlinear optimization

Definition

Nonlinear programming (NLP) deals with the problem of optimizing an objective function in the presence of

a system of equality and inequality constraints over a set of unknown real variables, where the objective function or some of the constraints are nonlinear (Banga 2008).

A general nonlinear programming problem has the form (Horst et al. 2000):

$$\begin{aligned} &\text{Minimize or maximize } f(x) \\ &\text{subject to} \quad g_i(x) \le 0 \quad for \quad i = 1, \ldots m \\ &\qquad\qquad h_i(x) = 0 \quad for \quad i = 1, \ldots l \\ &\qquad\qquad x \in X \end{aligned}$$

where f, $g_1, \ldots, g_m$, $h_1, \ldots, h_l$ are functions defined on $\mathbb{R}^n$, X (feasible or admissible domain) is a subset of $\mathbb{R}^n$, and x is a vector of n components $x_1, \ldots, x_n$ that satisfy the restrictions and meanwhile minimize or maximize the so-called objective function or cost function f.

Cross-References

▶ Optimal Experiment Design

References

Banga JR (2008) Optimization in computational systems biology. BMC Syst Biol 2:47

Horst R, Pardalos PM, Thoai NV (2000) Introduction to global optimization, 2nd edn. Kluwer Academic, Boston

Nonparametric Tests

▶ Hypothesis Testing, Parametric vs Nonparametric

Non-Protein-Coding RNA

▶ Non-coding RNA

Non-Protein-Coding RNA Databases

▶ Non-coding RNA Databases

Nonstandard Computation

▶ Unconventional Computation

Non-Synthetic Reactions

▶ Phase I Enzymes

NoSQL Databases

▶ Schemaless Databases

NP-hard

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

Synonyms

Nondeterministic polynomial-time hard

Definition

The class of NP-hard problems is the subset of decision problems $\boldsymbol{P}$ such that for all $\boldsymbol{Q} \in \mathbf{NP}$ (nondeterministic polynomial), $\boldsymbol{Q}$ is polynomially reducible to $\boldsymbol{P}$.

References

Wolsey LA (1998) Integer programming. Wiley, New York

Nuclear Pore

Yota Murakami
Department of Chemistry, Hokkaido University, Sapporo, Japan

Definition

Nuclear pore is a huge protein complex that is embedded in nuclear membrane and makes a “hole”

for active transport of proteins and RNAs between nucleus and cytosol.

Cross-References

▶ Heterochromatin and Euchromatin

Nucleosome

▶ Histones

Nucleosome Acting Factors

Masayuki Seki
Graduate School of Pharmaceutical Sciences, Tohoku University, Sendai, Miyagi, Japan

Synonyms

Factors modulating nucleosome structure

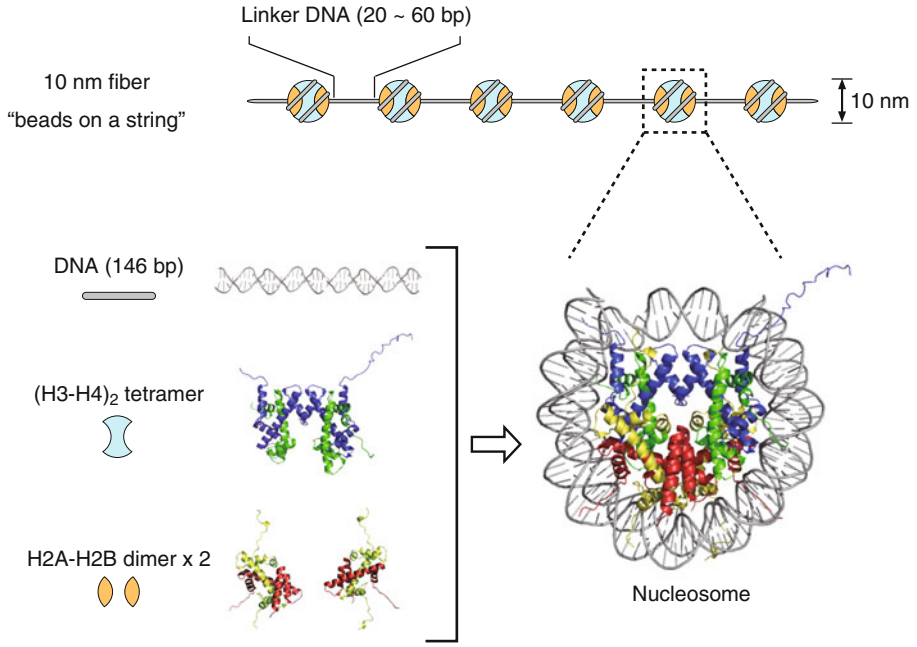
Definition

Nucleosome ▶ Nucleosome Structure acting factors indicate three classes of chromatin factors (▶ ATP-dependent Nucleosome-Remodeling Factors, and ▶ Histone Chaperones). The basic repeated unit of eukaryotic chromatin within the nucleus is an array of nucleosomes (Fig. 1). Nucleosomes comprise a core histone octamer (two histone H2A-H2B dimers and a histone $(H3\text{-}H4)_2$ tetramer) surrounded by approximately 146 bp of DNA (Fig. 1). The nucleosome structure negatively regulates a variety of DNA-mediated reactions including transcription, DNA replication, and DNA repair. It is required for executing these reactions to alter nucleosome structure and dynamics by nucleosome acting factors.

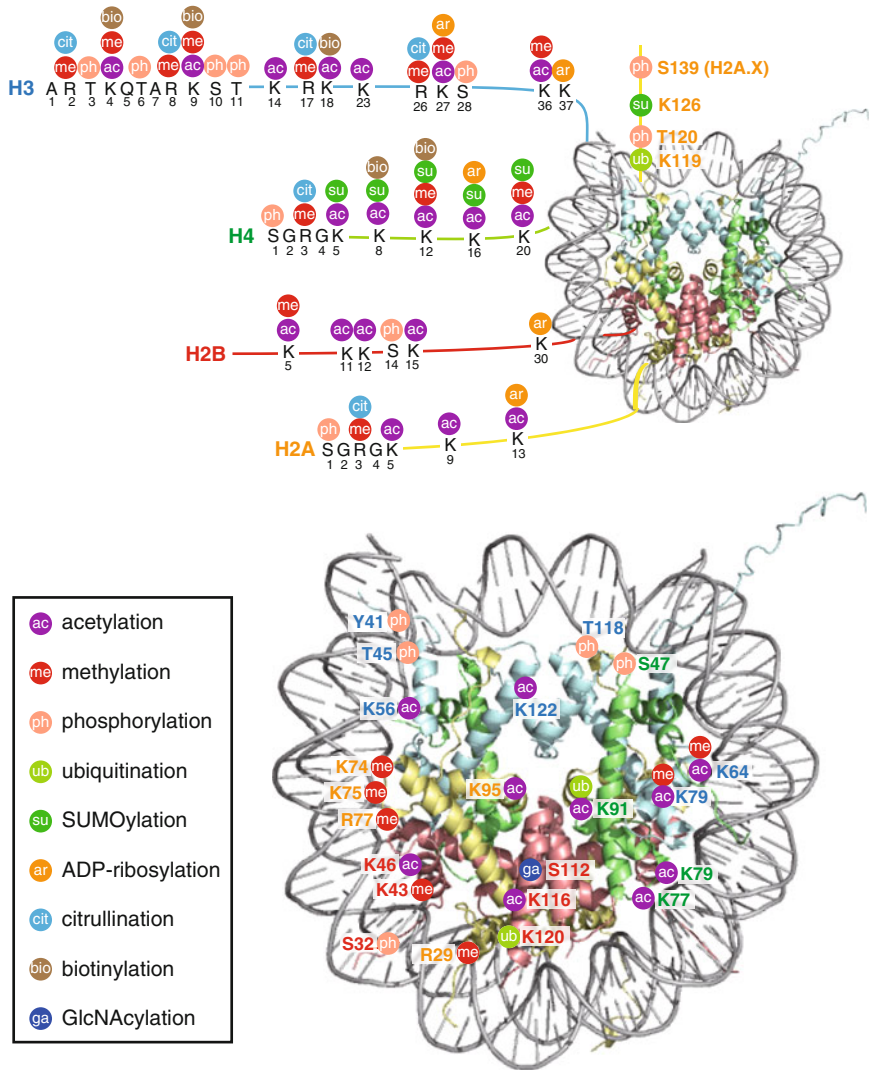
Characteristics

Histone Modification Enzymes and Effectors

A number of post-translational covalent modifications of histones (▶ Histone Post-translational Modification to Nucleosome Structural Change) (e.g., acetylation [ac], methylation [me], phosphorylation [ph],



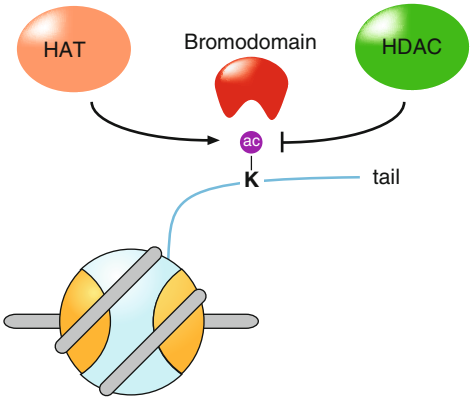
Nucleosome Acting Factors, Fig. 1 The basic unit of chromatin: (*Upper panel*) Arrays of nucleosomes. A 10-nm fiber is equivalent to an array of nucleosomes along the DNA strand, much like beads on a string. Nucleosomes, whose diameter is about 10 nm, are separated by approximately 20–60 base pairs of linker DNA. (*Lower panel*) Composition and structure of nucleosome. Histones H2A, H2B, H3, and H4 are colored in *yellow*, *red*, *blue*, and *green*, respectively



Nucleosome Acting Factors, Fig. 2 Post-translational covalent modifications of histones (histone PTMs): (*Upper panel*) A variety of histone PTMs are found in the intrinsically disordered histone tails and (*Middle panel*)the structured nucleosome core. (*Lower panel*) Representative histone modification enzymes and de-modification enzymes. Among a variety of histone PTMs, acetylation [ac], methylation [me], phosphorylation [ph], and ubiquitination [ub] related enzymes are listed

PTMs	Representative histone modification and de-modification enzymes	
ac	Histone acetyltransferase (HAT)	GCN5 TIP60 CBP/p300 etc.
	Histone deacetylase (HDAC)	RPD3 SIRT1 HDAC1 etc.
me	Histone methyltransferase (HMT)	SET1 SUV39H1 CARM1 etc.
	Histone demethylase	LSD1 UTX JMJD2 etc.
ph	Histone kinase	Aurora B Haspin PKC etc.
	Phosphatase	GLC7 PPH3 etc.
ub	E3 ubiquitin ligase	RAD6/BRE1 RNF20 Ring1B etc.
	Deubiquitinase	UBP8 USP16 MYSM1 etc.

ubiquitination [ub], SUMOylation [su], ADP-ribosylation [ar], citrullination [cit], and biotinylation [bio]) have been identified. The nucleosome contains intrinsically disordered histone tails and a structured nucleosome core. The majority of ▶ histone Posttranslational Modifications (PTMs) occur not only on these intrinsically



Effectors	Histone PTMs
Bromodomain	Acetylated lysine
Chromodomain Double chromodomain Chromo barrel	H3 K9me2/3 H3 K27me2/3 H3 K4me1 H3 K36me2/3
Tudor	H3 K4me3 H3 K9me3 H4 K20me2/3
MBT	H3 K4me1 H3 K9me1/2/3 H4 K20me1/2
PHD finger	H3 K4me2/3 H3 K9me3
WD40	H3 R2me0 H3 K4me2/3
PWWP	H3 K36me3 H4 K20me1
14-3-3	H3 S10ph H3 S28ph
BRCT	H2A.X S139ph

Nucleosome Acting Factors, Fig. 3 Representative effectors: (*Upper panel*) Acetylated histone lysine residues, which are modified by histone acetyltransferase (HAT) and de-modified by histone deacetylase (HDAC), recruit bromodomain-containing effectors to the chromatin. (*Lower panel*) Among a variety of histone PTMs, effectors for acetylation [ac], methylation [me], and phosphorylation [ph] are indicated. Mono-, di-, and tri-methylation are represented as me1, 2, and 3, respectively

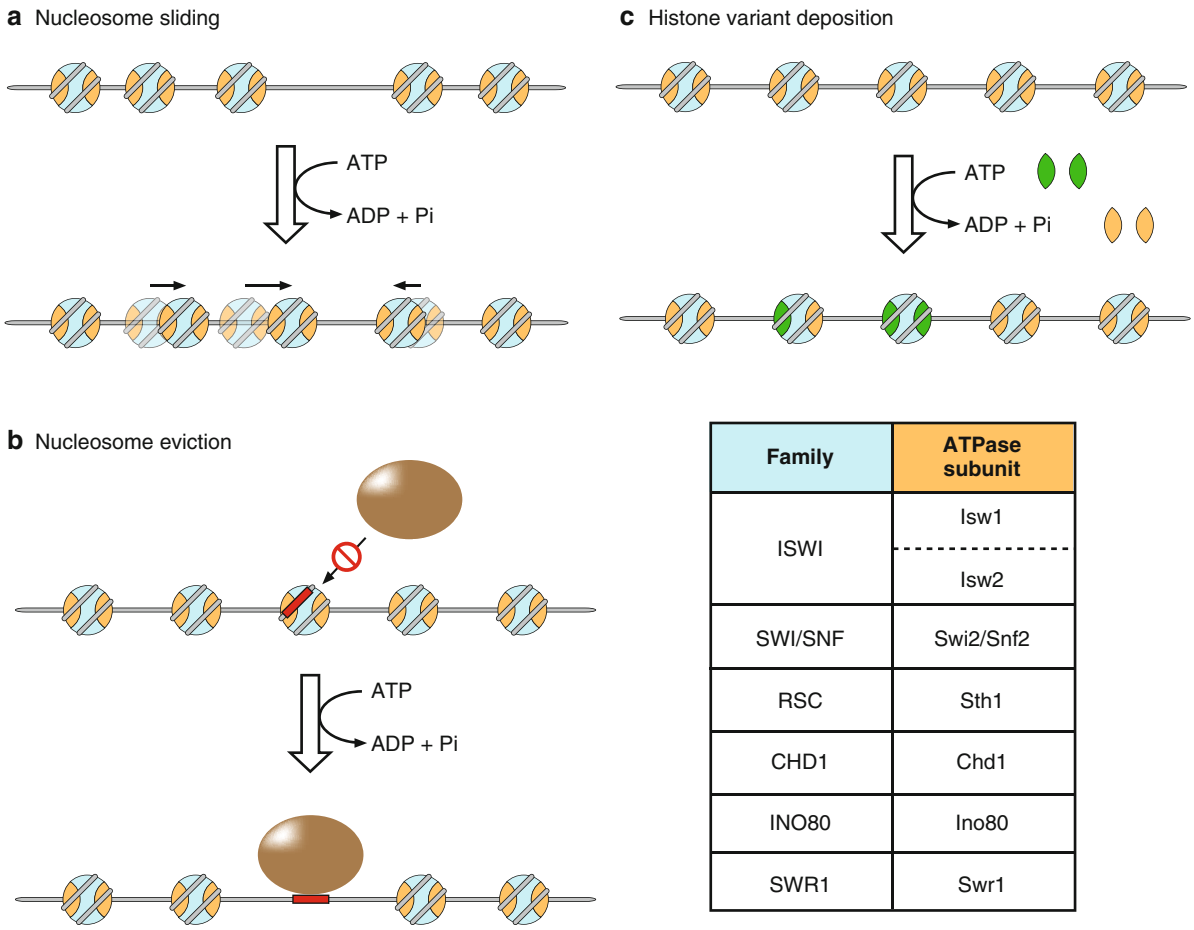
disordered histone tails but also on the structured nucleosome core (Fig. 2). Each histone PTM is modified or de-modified by histone PTM-related enzymes. A representative list of modification enzymes and de-modification enzymes of histone PTMs is shown in Fig. 2. It is noteworthy that relationships between histone PTMs have emerged to be the subject of study within the field of Systems Biology (Hayashi et al. 2009).

Histone PTMs play a pivotal role in regulating the structure and function of chromatin. So-called effectors recognize each histone PTM and mediate subsequent reactions on chromatin, such as recruitment of a chromatin factor and/or modification of another histone residue, during DNA-mediated reactions (Fig. 3).

Single genome is converted into hundreds of ▶ epigenomes (Epi = outside), whose chromatin changes in the genome are heritable by mechanisms other than changes in DNA sequence, such as ▶ DNA methylation, histone variants, and a variety of histone PTMs (Allis et al. 2006). The pattern of histone PTMs within an individual cell dictates the phenotype of that cell. Recently, techniques such as ChIP-seq (ChIP-sequencing) (▶ RIP-Chip and RIP-Seq) enable the genome-wide analyses of the epigenome (Park 2009). ChIP-seq is used to identify the genome sites to which a protein of interest localizes. The technique combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify DNA-associated proteins, leading to precise mapping of the global binding sites for a protein. ChIP-seq using an antibody against the histone PTM of interest can identify the pattern of histone PTMs along the whole genome. ChIP-seq has led to a rapid increase in the amount of information pertaining to the pattern of histone PTMs in individual cells during development, cell differentiation, the maintenance of embryonic stem cell status, and numerous diseases (including cancers). Elucidating massive information about the ▶ epigenome from the Systems Biology point of view is an emerging and challenging research area (Dodd et al. 2007).

ATP-Dependent Nucleosome Remodeling Complexes

Nucleosomes are regularly spaced along a 10-nm fiber of chromatin; an arrangement akin to beads (Nucleosomes) on a string (DNA) (Fig. 1). Sliding of the nucleosomes along the DNA (Nucleosome remodeling) is necessary for the proper spacing of these "beads" (Fig. 4). Nucleosome remodeling complex slides or evicts nucleosome to allow other factors required for a variety of DNA-mediated reactions access the chromatin (Fig. 4). Nucleosome remodeling is catalyzed by ▶ ATP-dependent nucleosome remodeling complexes, which comprise a catalytic ATPase subunit and multiple non-ATPase subunits. Several distinct families of ATP-dependent nucleosome remodeling complexes have been identified and are listed in Fig. 4. ATP-dependent nucleosome remodeling complexes can be recruited onto chromatin via their subunits, which recognize histone PTMs and/or



Family	ATPase subunit
ISWI	Isw1 / Isw2
SWI/SNF	Swi2/Snf2
RSC	Sth1
CHD1	Chd1
INO80	Ino80
SWR1	Swr1

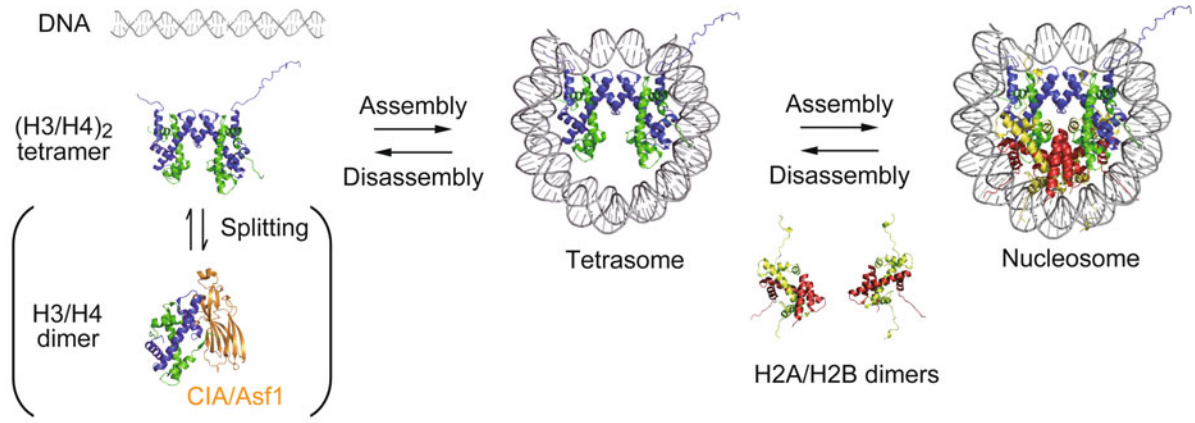
Nucleosome Acting Factors, Fig. 4 The actions of ATP-dependent nucleosome remodeling complexes: Representative budding yeast ATP-dependent nucleosome remodeling complexes are shown (*lower right panel*). (**a**, **b**) The ATP-dependent nucleosome remodeling complexes slide or evict nucleosomes. (**c**) In some case, histone variants are incorporated into the nucleosome instead of canonical histones by ATP-dependent nucleosome remodeling complexes. The four core histones, H2A, H2B, H3, and H4, are referred to as "canonical histones." However, there are variants of histones (histone variants) that are different from these core histones in terms of amino-acid sequence

other chromatin-bound factors. The ATP-dependent nucleosome remodeling complexes then destabilize the histone-DNA interaction, slide the nucleosomes along the DNA using energy derived from ATP hydrolysis, and thereby facilitate a variety of DNA-mediated reactions on the chromatin. Furthermore, the ATP-dependent nucleosome remodeling complex such as Swr1 complex replaces a canonical histone with its histone variant (Fig. 4).

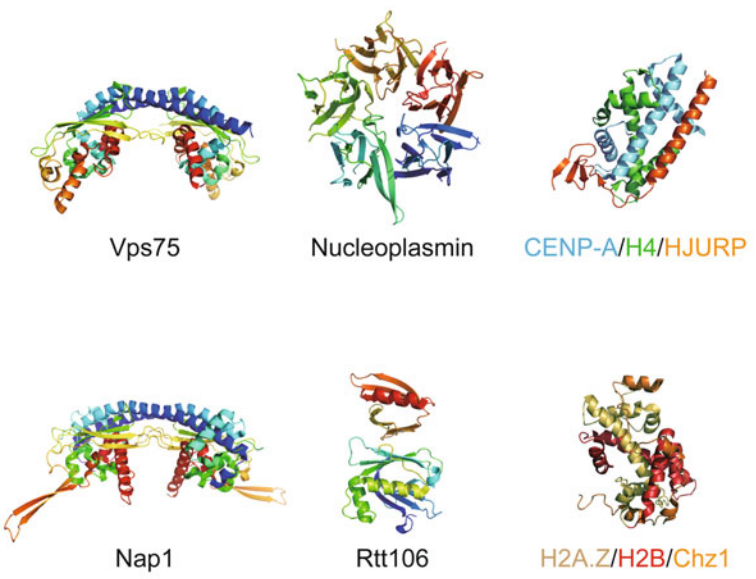
All ATP-dependent nucleosome remodeling complexes show similar biochemical properties, i.e., they remodel nucleosomes in vitro; however, the in vivo function of each individual ATP-dependent nucleosome remodeling complex is nonredundant (Clapier and Cairns 2009). In mammals, a defect in any of the ATP-dependent nucleosome remodeling complexes leads to severe effects during early embryonic development. Furthermore, a single copy of the genes encoding these ATPases or their subunits in mammals is often incapable of providing sufficient protein production so as to assure normal function in vivo.

Histone Chaperones

Nucleosome assembly and disassembly are essential steps for the successful execution of many DNA-mediated reactions on chromatin. ▶ Histone chaperones are histone-binding proteins that facilitate nucleosome assembly and disassembly in vitro without



Canonical histone chaperone	
H3/H4	CIA/Asf1
	CAF-1
	HIRA
	N1/N2
	Rtt106
	Spt6
	Vps75
	Fpr3/4
H2A/H2B	Nucleoplasmin
	Nap1
H3/H4 & H2A/H2B	FACT
Variant histone chaperone	
CENP-A/H4	HJURP
H2A.Z/H2B	Chz1



Nucleosome Acting Factors, Fig. 5 Histone chaperones and their histone-binding preferences: Representative histone chaperones, their histone-binding preferences, and their structures are indicated. Histones H2A, H2B, H3, and H4 are colored in *yellow*, *red*, *blue*, and *green*, respectively

using energy derived from ATP hydrolysis (Avvakumov et al. 2011). Under physiological conditions, histones and DNA do not easily self-assemble into nucleosomes because the positively charged histones nonspecifically aggregate with the negatively charged DNA. Histone chaperones assemble histones and DNA into the nucleosome by preventing the nonspecific interactions between histones and DNA and promoting the specific ones. Histone chaperones also reversibly facilitate the disassembly of the nucleosome into its individual components (Fig. 5). There are numerous types of histone chaperones, which are categorized by their primary and three-dimensional structures. Representative histone chaperones and their specific histone-binding preferences are shown in Fig. 5. Since the assembly and disassembly of nucleosomes occurs in a stepwise fashion (Fig. 5), each histone chaperone regulates a particular assembly and disassembly process depending on its histone-binding preference. For instance, histone chaperone CIA/Asf1 splits the histone $(H3\text{-}H4)_2$ tetramer into two histone H3-H4 dimers.

Histone chaperones interact not only with canonical histones but also with other chromatin factors.

The chromatin factors with which histone chaperones interact specify the location within the chromatin at which nucleosome assembly and disassembly take place during DNA-mediated reactions. Furthermore, some of histone chaperones interact with histone variants (Fig. 5) and linker histones. Linker histones (basic proteins within eukaryotic cell nuclei are not components of the Nucleosome) interact with the spacer (linker) DNA (Fig. 1) between adjacent nucleosomes. It is noteworthy that histone chaperones are also involved in a variety of histone-related activities in addition to nucleosome assembly and disassembly, histone variant exchange, and linker histone deposition such as regulation of histone PTMs, nucleosome sliding, histone shuttling, and histone storage.

Cross-References

- ▶ ATP-dependent Nucleosome-Remodeling Factors
- ▶ Epigenome
- ▶ Histone Chaperones
- ▶ Histone Post-translational Modification to Nucleosome Structural Change
- ▶ RIP-Chip and RIP-Seq

References

Allis CD, Jenuwein T, Reinberg D, Caparros ML (2006) Epigenetics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor

Avvakumov N, Nourani A, Côté J (2011) Histone chaperones: modulators of chromatin marks. Mol Cell 41(5):502–514

Clapier CR, Cairns BR (2009) The biology of chromatin remodeling complexes. Annu Rev Biochem 78:273–304

Dodd IB, Micheelsen MA, Sneppen K, Thon G (2007) Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell 129(4):813–822

Hayashi Y, Senda T, Sano N, Horikoshi M (2009) Theoretical framework for the histone modification network: modifications in the unstructured histone tails form a robust scale-free network. Genes Cells 14(7):789–806

Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10(10):669–680

Nucleosome Core Particle

- ▶ Nucleosome Structure

Nucleosome Structure

Hitoshi Kurumizaka
Graduate School of Advanced Science and Engineering, Waseda University, Tokyo, Japan

Synonyms

Nucleosome core particle

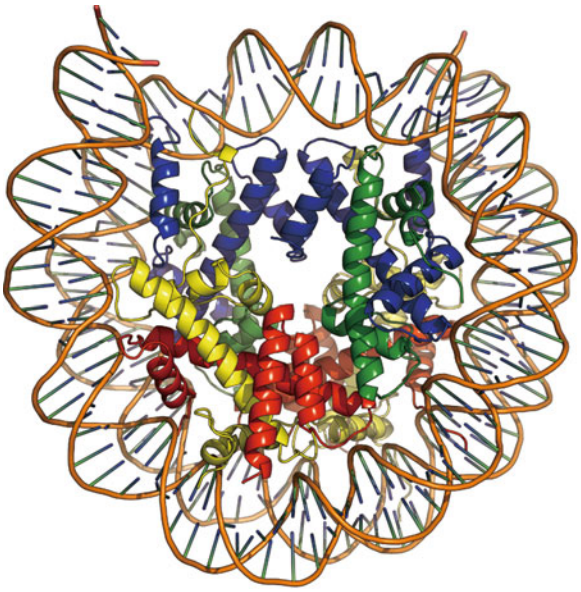
Definition

In eukaryotes, genomic DNA is packaged into "chromatin," for its accommodation within the nucleus. The core histones H2A, H2B, H3, and H4 are the major protein components of chromatin. The fundamental repeating unit of chromatin is the "nucleosome core particle," consisting of 146 base pairs of DNA wrapped in 1.65 left-handed superhelical turns around the histone octamer (Fig. 1) (Luger et al. 1997). The histone octamer is composed of two of each histone, H2A, H2B, H3, and H4, as two H2A/H2B dimers and one H3/H4 tetramer. The nucleosome core particles are connected by short linker DNA segments (roughly 20–50 base pairs), which do not directly bind to the histone octamer surface, and form the nucleosome array. The nucleosome and the nucleosome core particle are distinguished by the inclusion of the linker DNAs. Linker histones, such as histones H1 and H5, bind to the linker DNA within chromatin, and the mononucleosome containing one linker histone is defined as the "chromatosome" (Simpson 1978). The nucleosome array is folded into higher-ordered chromatin structures. These higher-ordered structures are dictated by the post-translational modifications of core histones and the specific incorporation of histone variants, and function to carry epigenetic information.

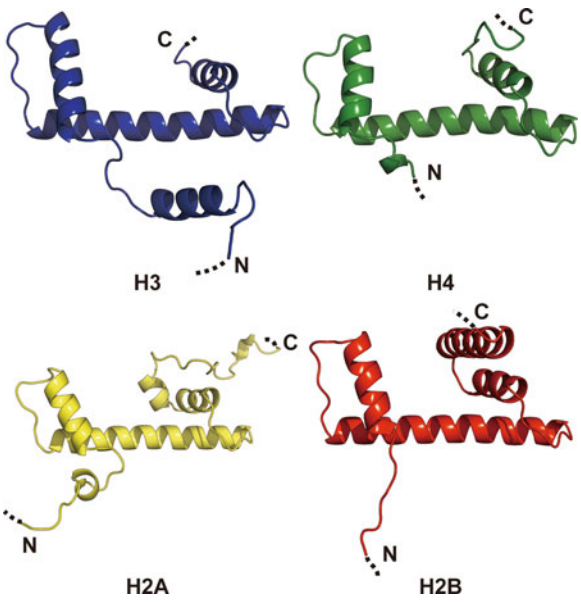
Characteristics

Core Histones

The four core histones, H2A, H2B, H3, and H4, share a common structural motif, consisting of N- and/or C-terminal tails and the histone-fold domain (Fig. 2) (Arents and Moudrianakis 1995).



Nucleosome Structure, Fig. 1 Structure of the nucleosome core particle containing human histones H2A, H2B, H3.1, and H4 (Tachiwana et al. 2010) RCSB ID code 3AFA



Nucleosome Structure, Fig. 2 Structures of the four core histones, H2A, H2B, H3, and H4, in the nucleosome core particle

The histone tails project from the nucleosome surface, and flutter around the nucleosome core particle in the solvent. The flexible histone tails may be the targets for trans-acting factors, such as histone modification enzymes, including histone acetyltransferase, histone methyltransferase, and histone kinase, and are actually enriched with amino acid residues targeted for post-translational modifications. The histone tails may also function as sites for inter-nucleosome interactions for higher-ordered chromatin formation.

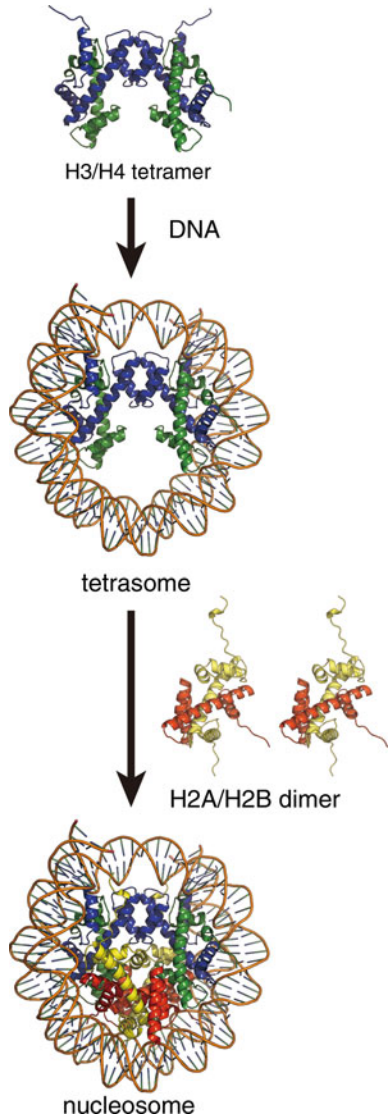
The histone-fold domains consist of a long central α-helix (α2) bordered by two short α-helices (α1 and α3). The loop 1 (L1) and loop 2 (L2) regions connect α1-α2 and α2-α3, respectively. In the histone octamer, H2A and H2B form a heterodimer, and the histone-fold domains interact in a handshake-like manner. The H3/H4 heterotetramer is composed of two H3/H4 heterodimers, in which the histone-fold domain of H3 interacts with that of H4, in a similar manner to the H2A/H2B heterodimer. The C-terminal α3 loop of histone H3 directly interacts with the other histone H3 molecule in the H3/H4 heterotetramer.

Nucleosome Core Particle

In the nucleosome core particle, 146 base pairs of DNA are left-handedly wrapped 1.65-times around the histone octamer, and form a disk-like structure with dimensions of 6 nm in height and 11 nm in diameter (Fig. 1). The DNA is bound to the lateral surface of the histone octamer. In the nucleosome core particle, the L2 regions of two histone H3 molecules bind to the backbone phosphates of the central region (nucleosomal dyad) of the 146 base pair DNA. The DNA segments at the entrance and exit of the nucleosome bind to the histone H3 αN-helix. The αN-helix of histone H3 is located outside the histone-fold domain, and just precedes the N-terminal region of the histone-fold α1-helix. The histone-DNA interactions are intermittently formed within the entire region of the nucleosome core particle. The backbone phosphates of the DNA bind to the histone octamer surface without sequence specificity, and the DNA is bent to fit the lateral surface of the histone octamer.

Nucleosome Assembly

The nucleosome is considered to be assembled in a stepwise manner. Two H3/H4 heterodimers first bind to the DNA, and the newly formed H3/H4 heterotetramer wraps the DNA around it. This subnucleosome structure is called the "tetrasome." Two H2A/H2B dimers then bind to the tetrasome, thus forming the mature nucleosome (Fig. 3). This sequential assembly of the nucleosome may be promoted by the actions of numerous



Nucleosome Structure, Fig. 3 Stepwise assembly of the nucleosome

histone chaperones in the nucleus. In the nucleosome, the H2A/H2B dimers interact with the loop regions of two histone H2A molecules, and the H3/H4 dimers form a bundle of four helices, containing the C-terminal portions of the α2 and α3 helices of histone H3 at the dimer interface. At the H3-H3′ interface, van der Waals contacts and a hydrogen bond are formed between two histone H3 molecules. Since the H2A/H2B dimers are relatively mobile in the nucleus, the H2A-H2A′ interaction in the nucleosome may not be strong. In contrast, the H3/H4 tetramers are stably incorporated into chromatin, and their mobility in the nucleus is extremely slow. A nucleosome lacking one H2A/H2B dimer may be formed for the promotion of transcription and/or replication.

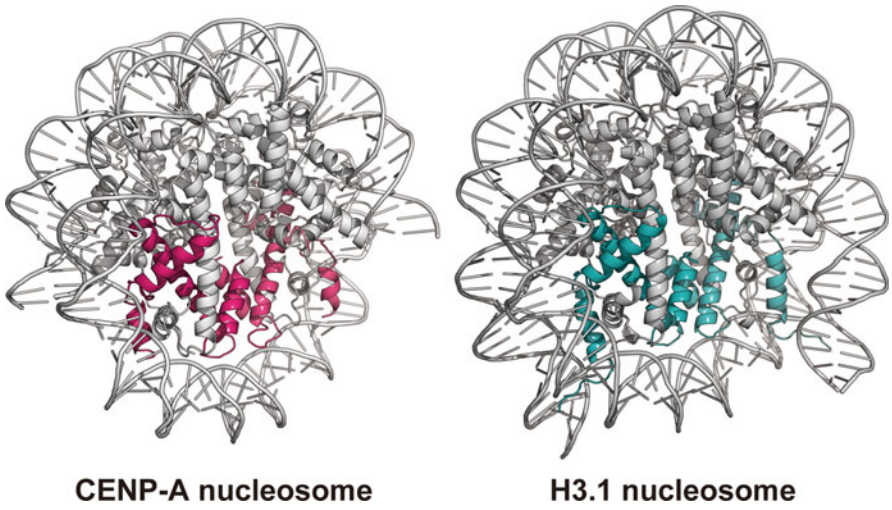
Histone Variants

Except for histone H4, nonallelic isoforms of histones H2A, H2B, and H3 exist in higher eukaryotes (Franklin and Zweidler 1977). These histone isoforms are called "histone variants."

Five human histone H2A variants have been identified: H2A, H2A.X, H2A.Z, H2A.Bbd, and macroH2A. H2A.Z and macroH2A also have sub-isoforms, such as H2A.Z-1, H2A.Z-2, macroH2A1.1, macroH2A1.2, and macroH2A2. H2A, H2A.X, and H2A.Z share similar structural features to those in other histones. In contrast, macroH2A contains a nonhistone-fold domain, called the macrodomain, at its C-terminus. H2A.Bbd lacks the C-terminal region, as compared to the canonical H2A (129 amino acid residues), and is composed of 114 amino acid residues. H2A.Z is found within the functional regions of chromosomes, and exhibits multiple functions. H2A.X is recruited to DNA double strand break sites, suggesting that it functions in the repair processes of these lesions. H2A.Bbd is thought to exist in transcriptionally active regions (euchromatin) of chromosomes. In contrast, macroH2A predominantly exists in transcriptionally inactive regions (heterochromatin) of chromosomes.

Three histone H2B variants, spH2B, hTSH2B, and H2BFWT, have been identified in humans. These H2B variants are highly expressed in testes, but not in somatic cells. Therefore, they may function during spermatogenesis and/or oogenesis.

Eight histone H3 variants, H3.1, H3.2, H3.3, H3T, H3.5, H3.X, H3.Y, and CENP-A, have been identified in humans. Among them, H3.1, H3.2, and H3.3 are abundantly produced, and commonly exist in all types of tissues and cells. H3.1 and H3.2, which are expressed during S phase, are incorporated into the chromatin in a replication-dependent manner. In contrast, H3.3 is constitutively expressed in a replication-independent manner. H3.3 functions as a replacement for histone H3, and seems to be predominantly incorporated into transcriptionally active chromatin regions and the telomeres of chromosomes. H3T and H3.5 are highly expressed in testes, but not in somatic cells. H3.X and H3.Y are novel histone variants that may be involved in the regulation of cellular responses to outside stimuli. CENP-A, a centromere-specific H3 variant, is an



Nucleosome Structure, Fig. 4 Structure of the human CENP-A nucleosome

essential component of active centromeres. CENP-A is a strong candidate for an epigenetic marker of kinetochore formation sites.

Nucleosome Structures with Frog, Fly, Yeast, and Human Histones

The crystal structures of nucleosomes have been determined with histones from frog, fly, yeast, and human. The frog *Xenopus laevis* and *Drosophila melanogaster* have H3.2 as the canonical replication-dependent histone H3. The yeast *Saccharomyces cerevisiae* has the H3.3-type histone H3 and the H2A.X-type histone H2A as the canonical core histones. The crystal structures of the frog, fly, yeast, and human nucleosomes revealed that the DNA-binding path of the human nucleosome differs from those of the frog and yeast nucleosomes (Luger et al. 1997; White et al. 2001; Tsunaka et al. 2005; Clapier et al. 2008). Since no obvious structural variations are apparent among the human H3.1, H3.2, and H3.3 nucleosome structures (Tachiwana et al. 2011a), the difference in the DNA-binding path may not be due to the distinct histone variants.

Nucleosome Structures Containing Histone Variants

The structures of nucleosomes containing histone H2A.Z or histone H3T have been reported, thus revealing their specific physical features (Suto et al. 2000; Tachiwana et al. 2010). A biochemical study showed that the human nucleosome containing histone H3T is extremely unstable, as compared to the canonical H3.1 nucleosome. This instability of the H3T nucleosome may play an important role during spermatogenesis. The centromere-specific nucleosome containing CENP-A has been determined (Tachiwana et al. 2011b). In the CENP-A nucleosome, only 121 base pairs of DNA are wrapped around the histone octamer, and 13 base pairs are detached from the histone surface at the entrance and exit of the nucleosome (Fig. 4). In the CENP-A nucleosome, the L1 region of the histone-fold domain is longer, by two amino acid residues, as compared to the H3 L1 region, and the L1 loop protrudes from the CENP-A nucleosome. The tip of the CENP-A L1 loop is exposed to the solvent, and CENP-A deletion mutants in the L1 loop have reduced stability at the centromeres of human cells (Tachiwana et al. 2011b).

Cross-References

- ▶ Chromatin
- ▶ Epigenetics
- ▶ Histones

References

Arents G, Moudrianakis EN (1995) The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. Proc Natl Acad Sci USA 92(24):11170–11174

Clapier CR, Chakravarthy S, Petosa C, Fernández-Tornero C, Luger K, Müller CW (2008) Structure of the *Drosophila* nucleosome core particle highlights evolutionary constraints on the H2A-H2B histone dimer. Proteins 71(1):1–7

Franklin SG, Zweidler A (1977) Non-allelic variants of histones 2a, 2b and 3 in mammals. Nature 266(5599):273–275

Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature 389(6648):251–260
Simpson RT (1978) Structure of the chromatosome, a chromatin particle containing 160 base pairs of DNA and all the histones. Biochemistry 17(25):5524–5531
Suto RK, Clarkson MJ, Tremethick DJ, Luger K (2000) Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. Nat Struct Biol 7(12):1121–1124
Tachiwana H, Kagawa W, Osakabe A, Kawaguchi K, Shiga T, Hayashi-Takanaka Y, Kimura H, Kurumizaka H (2010) Structural basis of instability of the nucleosome containing a testis-specific histone variant, human H3T. Proc Natl Acad Sci USA 107(23):10454–10459
Tachiwana H, Osakabe A, Shiga T, Miya Y, Kimura H, Kagawa W, Kurumizaka H (2011a) Structures of human nucleosomes containing major histone H3 variants. Acta Crystallogr D Biol Crystallogr 67(Pt 6):578–583
Tachiwana H, Kagawa W, Shiga T, Osakabe A, Miya Y, Saito K, Hayashi-Takanaka Y, Oda T, Sato M, Park SY, Kimura H, Kurumizaka H (2011b) Crystal structure of the human centromeric nucleosome containing CENP-A. Nature 476(7359):232–235
Tsunaka Y, Kajimura N, Tate S, Morikawa K (2005) Alteration of the nucleosomal DNA path in the crystal structure of a human nucleosome core particle. Nucleic Acids Res 33(10):3424–3434
White CL, Suto RK, Luger K (2001) Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. EMBO J 20(18):5207–5218

Nucleosome-Remodeling Factor

► ATP-dependent Nucleosome-Remodeling Factors

Nucleosomes

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

Synonyms

Chromatin

Definition

The basic unit of ► genome is the nucleosome. It is made of two subunits each of the four core histones H2A, H2B, H3, and H4 forming an octamer on which two full turns of DNA (146 bp) is wrapped, and another ~50 bp of DNA complexed with histone H1 serves as a linker DNA connecting each nucleosome to the other forming the "beaded DNA" structure or the ► chromatin. The length of the linker DNA varies between cells. The H2A-H2B and H3-H4 heterodimers form the octamer which associate, and, hence, approximately 200 base pairs of DNA are wrapped on the surface of one nucleosome unit (Fig. 1 ► Epigenetics).

Cross-References

► Epigenetics
► Genome
► Chromatin

Numerical Continuation

► Dynamical Systems Theory, Bifurcation Analysis

Numerical Methods

► Partial Differntial Equations, Numerical Methods and Simulations

Numerical Optimization

► Optimization and Parameter Estimation, Genetic Algorithms

Numerical Taxonomy Methods

► Clustering Methods