B

5-Bromo-2-deoxyuridine (BrdU)

► Lymphocyte Labeling, Cell Division Investigation

B Cell Epitope

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

Definition

A B cell epitope is the region of the antigen recognized by soluble or membrane-bound antibodies. B cell epitopes are classified as either linear or discontinuous epitopes.

Cross-References

► B Cells
► Systems Immunology, Data Modeling and Scripting in R

B Cell Epitope Prediction

Yasser EL-Manzalawy and Vasant Honavar
Center for Computational Intelligence, Learning, and Discovery, Computer Science, Iowa State University, Ames, IA, USA

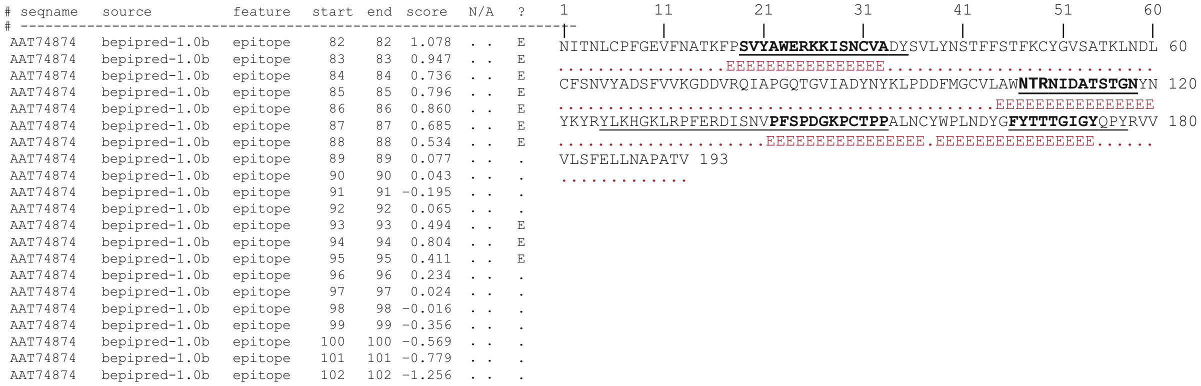
Synonyms

Antigen–antibody binding site prediction; Antigen–antibody interface residue prediction; Computational methods for mapping B cell epitopes

Definition

B cell epitopes, also known as antigenic determinants, are restricted parts of molecules that are recognized by immunoglobulin molecules (antibodies) either in their free form or as membrane-bound B cell receptors. B cell epitopes typically belong to one of two classes: linear (continuous or sequential) epitopes or conformational (discontinuous) epitopes. Linear epitopes are short peptides that correspond to a contiguous amino acid sequence fragment of a protein. Linear epitopes are usually identified using assays such as PEPSCAN. Consequently, current experimental methods offer

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



```
# seqname   source          feature   start  end  score   N/A   ?
# ----------------------------------------------------------------
 AAT74874   bepipred-1.0b   epitope    82    82   1.078   . .   E
 AAT74874   bepipred-1.0b   epitope    83    83   0.947   . .   E
 AAT74874   bepipred-1.0b   epitope    84    84   0.736   . .   E
 AAT74874   bepipred-1.0b   epitope    85    85   0.796   . .   E
 AAT74874   bepipred-1.0b   epitope    86    86   0.860   . .   E
 AAT74874   bepipred-1.0b   epitope    87    87   0.685   . .   E
 AAT74874   bepipred-1.0b   epitope    88    88   0.534   . .   E
 AAT74874   bepipred-1.0b   epitope    89    89   0.077   . .   .
 AAT74874   bepipred-1.0b   epitope    90    90   0.043   . .   .
 AAT74874   bepipred-1.0b   epitope    91    91  -0.195   . .   .
 AAT74874   bepipred-1.0b   epitope    92    92   0.065   . .   .
 AAT74874   bepipred-1.0b   epitope    93    93   0.494   . .   E
 AAT74874   bepipred-1.0b   epitope    94    94   0.804   . .   E
 AAT74874   bepipred-1.0b   epitope    95    95   0.411   . .   E
 AAT74874   bepipred-1.0b   epitope    96    96   0.234   . .   .
 AAT74874   bepipred-1.0b   epitope    97    97   0.024   . .   .
 AAT74874   bepipred-1.0b   epitope    98    98  -0.016   . .   .
 AAT74874   bepipred-1.0b   epitope    99    99  -0.356   . .   .
 AAT74874   bepipred-1.0b   epitope   100   100  -0.569   . .   .
 AAT74874   bepipred-1.0b   epitope   101   101  -0.779   . .   .
 AAT74874   bepipred-1.0b   epitope   102   102  -1.256   . .   .
```

```
1         11        21        31        41        51       60
|         |         |         |         |         |        |
NITNLCPFGEVFNATKFPSVYAWERKKISNCVADYSVLYNSTFFSTFKCYGVSATKLNDL 60
.................EEEEEEEEEEEEEEEE...........................
CFSNVYADSFVVKGDDVRQIAPGQTGVIADYNYKLPDDFMGCVLAWNTRNIDATSTGNYN 120
............................................EEEEEEEEEEEEEEEE
YKYRYLKHGKLRPFERDISNVPFSPDGKPCTPPALNCYWPLNDYGFYTTTGIGYQPYRVV 180
.....................EEEEEEEEEEEEEEEE.EEEEEEEEEEEEEEEE......
VLSFELLNAPATV 193
.............
```

B Cell Epitope Prediction, Fig. 1 Residue-based (*left*) and peptide-based (*right*) linear B cell epitope predictions. In the case of residue-based predictions, each predicted epitope residue is denoted with the letter "E." In the case of peptide-based predictions, the predicted epitopes are encoded with a string of "E"s

little direct evidence indicating that each residue in the epitope does in fact make contact with one or more residues in the paratope (the part in the antibody that binds to the antigen). The second class of B cell epitopes is called conformational (discontinuous) B cell epitopes that represent the vast majority of B cell epitopes found in proteins. These epitopes are composed of amino acids that, although not contiguous in the primary sequence, are brought into close proximity within the folded three-dimensional protein structure.

Characteristics

Predicting B Cell Epitopes

Identification of B cell epitopes is often a necessary first step in developing safe and effective vaccines. Characterization of sequence and structural features of B cell epitopes is important from the standpoint of understanding of pathogenicity and the adaptive immune response. Several experimental techniques are currently available for mapping B cell epitopes. However, with rapid increase in the number of fully or partially sequenced pathogen genomes and prohibitive cost and effort required for experimental identification of epitopes, there is an urgent need for cost-effective computational methods for reliable genome-wide identification of B cell epitopes.

B cell epitope predictors can be categorized based on the type of the B cell epitope predicted into linear and conformational B cell epitope predictors. Linear B cell epitope predictors accept as input a suitable representation of the amino acid sequence of a target antigen and output the predicted epitope(s). Some predictors, called peptide-based predictors, require the user to specify the length of the epitope as an input parameter. Peptide-based predictors are trained to classify amino acid fragments into epitopes and other non-epitopes. Residue-based predictors, on the other hand, are trained to score each residue in the query sequence; the higher the score, the more likely it is that the corresponding residue belongs to an epitope. Figure 1 illustrates peptide-based and residue-based B cell epitope prediction. In contrast to peptide-based predictors that can only identify epitopes of specified length or antigenic regions, residue-based predictors can be used to infer antigenic regions of unknown length or even conformational B cell epitopes. Conformational B cell epitope predictors take as input an antigen structure (actual or predicted PDB coordinate file) or antigen sequence, and extract some sequence- and/or structure-based feature representation of each residue in the input query antigen. The output of the predictor is a per residue scores assigned to each residue in the query antigen.

Predicting Linear B Cell Epitopes

Although a vast majority of B cell epitopes are believed to be conformational epitopes (Walter 1986), most of the existing experimental B cell-epitope mapping techniques and, perhaps consequently, most of the existing computational methods for B cell epitope prediction deal with linear B cell epitopes. Several computational-based methods for predicting linear B cell epitopes have been proposed in literature (EL-Manzalawy and Honavar 2010). B cell epitopes predicted using such methods can be easily synthesized and tested for their antibody-binding properties. Existing computational methods for predicting linear B cell epitopes fall into two major categories: (1) propensity scale–based methods; (2) machine learning–based methods.

Propensity Scale–Based Methods

Propensity scale–based methods (e.g., Parker and Guo 1986; Pellequer et al. 1993) rely on the observed correlations between specific physicochemical properties (e.g., hydrophilicity, flexibility, or solvent accessibility) of amino acids and the antigenic determinants in protein sequences to identify the location(s) of the linear B cell epitope(s) in the query protein sequence. The main idea is to assign a score to each amino acid in a query protein sequence. These propensity scores measure the tendency of an amino acid to be part of a B cell epitope (as compared to the background). The score for each target amino acid residue in a query sequence is computed as the average of the propensity values of the amino acids in a sliding window centered at the target residue. The propensity scores are then used as a basis of predicting whether a given amino acid sequence residue is likely to be part of a linear B cell epitope. Figure 2 shows the analysis of receptor-binding domain (RBD) of Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) spike protein using Parker's hydrophilic scale.

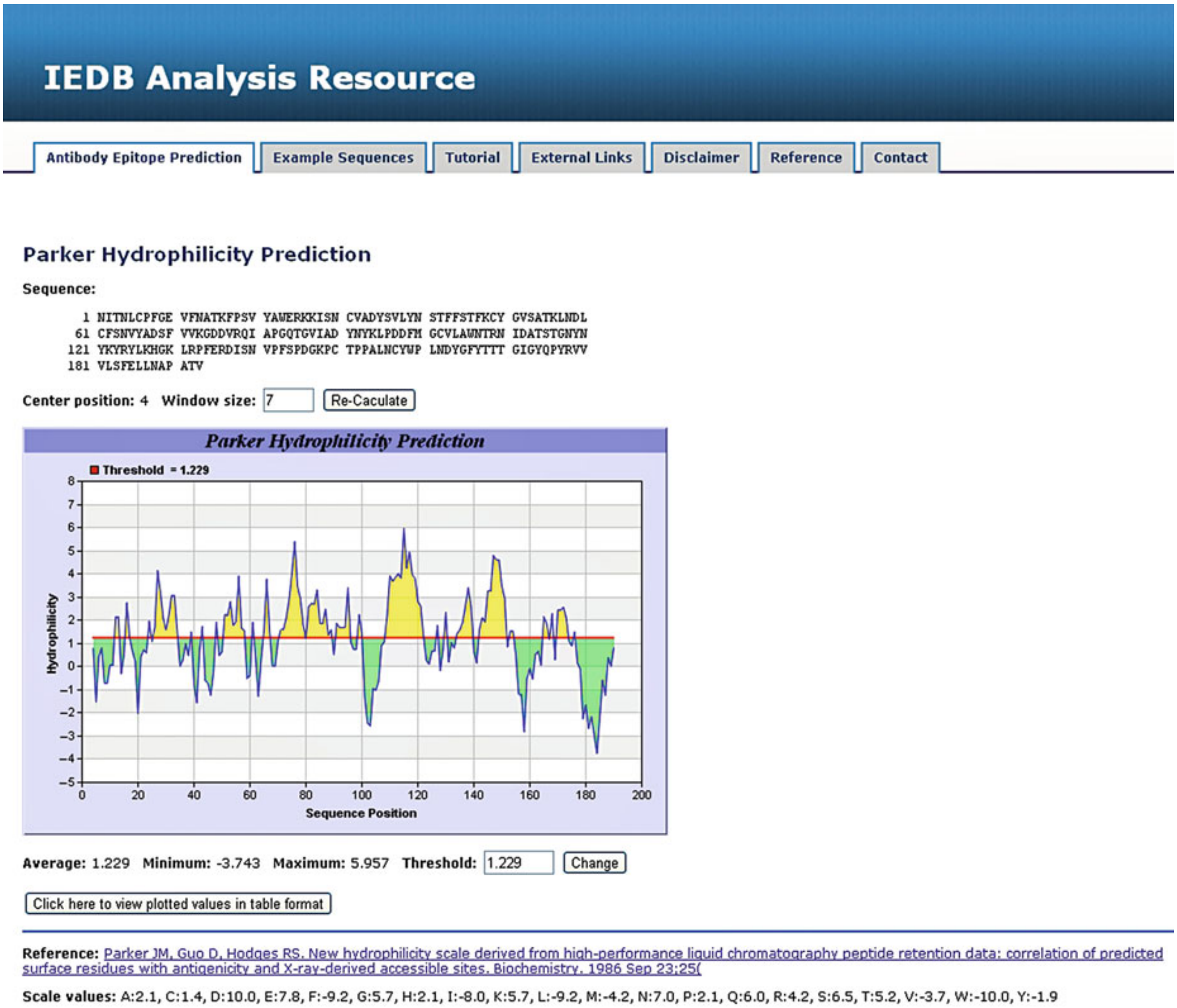
Recently, Blythe and Flower (Blythe and Flower 2005) evaluated 484 amino acid propensity scales on a data set of 50 proteins and concluded that the best achievable performance is only marginally better than random guessing. This result underscores the need for more sophisticated methods (e.g., those that use state-of-the-art machine learning algorithms together with appropriate data representations) for constructing improved linear B cell epitope predictors.

Machine Learning–based Methods

Machine learning currently offers one of the most cost-effective and hence widely used approaches to developing predictive models from data in bioinformatics applications (Baldi and Brunak 2001). Several machine learning–based linear B cell epitope predictors have been developed using Support Vector Machine, Artificial Neural Network, Decision Tree, k-Nearest Neighbor, and Ensemble classifiers. Such classifiers can be trained using examples that are amino acid peptide sequences with the corresponding binary labels (epitopes vs. non-epitopes) to obtain peptide-based B cell epitope predictors. Alternatively, they can be trained using examples that are fixed length amino acid sequence windows whose binary labels indicate whether or not the residue at the center of each sequence window is an epitope residue to obtain residue-based B cell epitope predictors. In either case, a variety of amino acid residue–based and amino acid sequence–based features for encoding the input to the classifiers (EL-Manzalawy and Honavar 2010) have been utilized. Despite recent advances in machine learning–based linear B cell epitope predictors, there is much room for improvement in the performance of the state-of-the-art in computational prediction of linear B cell epitopes (Greenbaum et al. 2007). This is due at least in part to the limited amount of experimentally characterized epitope data (and in particular, lack of reliable negative, i.e., non-epitope data). Consequently, the development of more reliable linear B cell epitope predictors remains a major challenge in computational immunology.

Predicting Conformational B Cell Epitopes

Several experimental techniques can be used for identifying conformational B cell epitopes. The most accurate method relies on the determination of the structure of antigen–antibody complexes using X-ray crystallography (Fleury et al. 2000). Progress of computational methods for conformational B cell epitope prediction has been hindered at least in part by the limited number of solved antigen–antibody complexes. As noted earlier, propensity scale–based methods can be used to predict both linear and conformational B cell epitopes using only amino acid sequence information of antigens. Two recently developed methods for predicting conformational B cell epitopes, DiscoTope and PEPITO, improve the accuracy of propensity scale–based methods by incorporating information derived from the structure of the antigens, e.g., solvent accessibility of residues. The task of predicting conformational B cell epitopes can be reduced to identifying protein–protein interface residues on the surface of a target protein (antigen). This opens up the possibility of adapting the state-of-the-art sequence and/or structure-based protein–protein interface residue prediction methods for developing conformational B cell epitope predictors. A recent study (Ponomarenko and Bourne 2007) compared the performance of six publicly available protein–protein



Parker Hydrophilicity Prediction

Sequence:

```
  1 NITNLCPFGE VFNATKFPSV YAWERKKISN CVADYSVLYN STFFSTFKCY GVSATKLNDL
 61 CFSNVYADSF VVKGDDVRQI APGQTGVIAD YNYKLPDDFM GCVLAWNTRN IDATSTGNYN
121 YKYRYLKHGK LRPFERDISN VPFSPDGKPC TPPALNCYWP LNDYGFYTTT GIGYQPYRVV
181 VLSFELLNAP ATV
```

Center position: 4 Window size: 7

Average: 1.229 Minimum: -3.743 Maximum: 5.957 Threshold: 1.229

Reference: Parker JM, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry. 1986 Sep 23;25(

Scale values: A:2.1, C:1.4, D:10.0, E:7.8, F:-9.2, G:5.7, H:2.1, I:-8.0, K:5.7, L:-9.2, M:-4.2, N:7.0, P:2.1, Q:6.0, R:4.2, S:6.5, T:5.2, V:-3.7, W:-10.0, Y:-1.9

B Cell Epitope Prediction, Fig. 2 Analysis of RBD domain of SARS-CoV Spike protein using Parker's hydrophilic scale. A sliding window of seven amino acids is used to assign a propensity score to the residue in the center of the window. The score is the sum of the epitope propensity of each amino acid in the window. The higher the score, the more likely it is that the corresponding residue is an epitope residue. Positive peaks along the sequence in the plot of the residue scores indicate the presence of an epitope in that region of the sequence

interface residue prediction tools on the conformational B cell epitope prediction task. The reported performance of such methods (with an average area under curve (AUC) no greater than 0.7) underscores the need for developing protein–protein interface predictors that are customized for the conformational B cell epitope prediction task.

A recently developed approach for identifying B cell epitopes uses a combination of both experimental and computational techniques. In this approach, a phage-display library of random peptides is scanned against an antibody of interest to obtain a panel of peptides (named mimotopes) that bind to the antibody with high affinity. It is assumed that this panel of mimotopes mimics the physicochemical properties and spatial organization of the genuine epitopes. Because the precise identification of the epitope mimicked by the set of mimotopes is not straightforward since the epitope is often discontinuous (conformational) and the epitope and mimotopes do not necessarily share a high degree of sequence similarity, several computational methods have been proposed for localizing the panel of affinity-selected peptides on the surface of a target antigen (e.g., Bublil et al. 2007).

References

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

Blythe M, Flower D (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci 14:246–248

Bublil E, Freund N, Mayrose I, Penn O, Roitburd-Berman A, Rubinstein N, Pupko T, Gershoni J (2007) Stepwise prediction of conformational discontinuous B cell epitopes using the Mapitope algorithm. Proteins Struct Funct Bioinform 68:294–304, Wiley

EL-Manzalawy Y, Honavar V (2010) Recent advances in B cell epitope prediction methods. Immunome Res 6:S2

Fleury D, Daniels R, Skehel J, Knossow M, Bizebard T (2000) Structural evidence for recognition of a single epitope by two distinct antibodies. Proteins 40:572

Greenbaum J, Andersen P, Blythe M, Bui H, Cachau R, Crowe J, Davies M, Kolaskar A, Lund O, Morrison S, Mumey B, Ofran Y, Pellequer J, Pinilla C, Ponomarenko J, Raghava G, van Regenmortel M, Roggen E, Sette A, Schlessinger A, Sollner J, Zand M, Peters B (2007) Towards a consensus on datasets and evaluation metrics for developing B cell epitope prediction tools. J Mol Recognit 20:75–82

Parker J, Guo D (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites. Biochemistry 25:5425–5432, American Chemical Society

Pellequer J, Westhof E, Van Regenmortel M (1993) Correlation between the location of antigenic sites and the prediction of turns in proteins. Immunol Lett 36:83–99

Ponomarenko J, Bourne P (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct Biol 7:64, BioMed Central Ltd

Walter G (1986) Production and use of antibodies against synthetic peptides. J Immunol Methods 88:149–161

B Cell-mediated Immune Response

Shoba Ranganathan
Department of Chemistry and Biomolecular Sciences and ARC Center of Excellence in Bioinformatics, Macquarie University, Sydney, NSW, Australia
Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Synonyms

Antibody-dependant immune response; Antibody-mediated immunity; B cell-mediated immunity

Definition

B cell-mediated immune response is defined as the immune response cascade triggered by the binding of antibodies (produced by the B cells) to the antigens and subsequent identification by the cell surface receptors of macrophages, neutrophils, or other cells of the B cell-mediated immunity to destroy the antigens. It is a type of adaptive immunity in vertebrates (Alberts et al. 2002).

Cross-References

- ▶ Major Histocompatibility Complex (MHC)
- ▶ TCR Recognition of MHC-Peptide Complexes

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) The adaptive immune system. In: Molecular biology of the cell, 4th edn. Garland Science, New York, pp 1363–1421

B Cells

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

Synonyms

B lymphocyte

Definition

B cells are a type of lymphocytes produced in the bone marrow of mammals, which later migrate to spleen and lymph nodes. B cells mainly differentiate into memory B cell and plasma cells. The plasma cells produce antibodies, thereby eliciting strong humoral immune response against antigens.

Cross-References

- ▶ B Cell Epitope
- ▶ Systems Immunology, Data Modeling and Scripting in R

B Lymphocyte

▶ B Cells

Backbone or Carbon-α Chain (Both with the Sidechain)

▶ Protein Structure Metapredictors

Bacterial Artificial Chromosome (BAC)

Myong-Hee Sung
Laboratory of Receptor Biology and Gene Expression, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

Definition

A bacterial artificial chromosome is a DNA construct used for cloning a relatively large piece of DNA (100∼700 kb), usually by transforming *E. coli*.

Bacterial Cell Cycle

▶ Cell Cycle, Prokaryotes

Bacterial Transcription

▶ Transcription in Bacteria

Bacterial Transcriptional Cascade

▶ Sigma Cascade

Bagged Predictors

▶ Bagging

Bagging

Celine Vens
Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium

Synonyms

Bagged predictors; Bootstrap aggregating

Definition

Bagging is an ▶ ensemble learning method. Each member of the ensemble is trained on a different ▶ bootstrap replicate of the training set. The outcomes of the individual learning models are aggregated to obtain the outcome of the ensemble. Bagging often uses ▶ decision tree learners to create the individual models, but it can be used with any unstable learning method.

Characteristics

Constructing Bagged Predictors

Bagging proceeds by applying the same learning algorithm to different bootstrap samples of the training set (Breiman 1996a). Given a training set D, M new training sets D_k are generated, by uniformly sampling examples from D, with replacement. Usually, the sampling procedure is terminated when each training set contains an equal amount of training examples as the original set D. Thus, each example may appear zero, one, or multiple times in each of the bootstrap samples. Algorithm 1 shows the pseudo-code to construct a bagged ensemble.

Algorithm 1 Pseudo-code for constructing an ensemble using bagging. D denotes the training set, M is the number of models in the ensemble.

```
for $k \Leftarrow 1$ to $M$ do
    $D_k \Leftarrow$ Bootstrap($D$)
    $h_k \Leftarrow$ Learning Algorithm ($D_k$)
end for
return $\mathrm{U}h_k$
```

The outputs of the base classifiers are aggregated (hence the name bagging, which is an acronym for *bootstrap aggregating*) to form the output of the bagging ensemble. Bagging is most popular in the context of ▶ supervised learning, in which case the output corresponds to a prediction of the target attribute. For ▶ classification tasks, the ▶ majority vote of the base classifiers' predictions is taken; for ▶ regression, the average is taken.

The number of models, M, is a parameter to be chosen by the user. Different values have been reported in the literature. Breiman (1996a) uses 50 models for classification tasks, and 25 models for regression.

Bagging improves the predictive performance of the base predictors if the base predictor is unstable. This means that small changes in the training set can yield large changes in the predictive model. ▶ Decision trees and artificial neural networks are examples of unstable predictors. Nearest neighbor methods, for instance, are not. In terms of the bias-variance tradeoff, bagging is known to reduce variance, while slightly increasing bias.

Out-Of-Bag Error Estimates

An advantage of using bagging is that out-of-bag error estimates (Breiman 1996b) can be used to estimate the generalization error of the ensemble, which removes the need for a set aside test set. If the ensemble comprises decision trees, then the out-of-bag error estimation proceeds as follows: for every example in the training set, a prediction is made, but only those trees for which the example was not in the bootstrap sample are used. The error rate of this resulting out-of-bag classifier is called the out-of-bag error estimate. In each resampled training set, about one third of the instances are left out (actually $1/e$ in the limit). As a result, out-of-bag estimates are based on combining only about one third of the total number of classifiers in the ensemble. This means that they might overestimate the error rate, certainly when a small number of trees are used in the ensemble.

Applications in Systems Biology

Dudoit and Fridlyand (2003) propose a bagging procedure to improve the clustering of gene expression data from cancer microarray studies. Schietgat et al. (2010) use bagging of ▶ decision trees to predict the functions of genes.

Implementations

Most data mining tools (e.g., Weka [Weka, Machine Learning Tool]) include an implementation of the bagging procedure.

Cross-References

▶ Bootstrapping
▶ Classification
▶ Decision Tree
▶ Ensemble
▶ Learning, Supervised
▶ Regression

References

Breiman L (1996a) Bagging predictors. Machine Learning 24(2):123–140

Breiman L (1996b) Out-of-bag estimation. Technical Report. University of California, Berkely ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps.Z. Accessed Aug 4, 2011

Dudoit S, Fridlyand J (2003) Bagging to improve the accuracy of a clustering procedure. Bioinformatics 19(9):1090–1099

Schietgat L, Vens C, Struyf J, Blockeel H, Kocev D, Džeroski S (2010) Predicting gene function using hierarchical multi-label decision tree ensembles. BMC Bioinformatics 11(2)

Basal Lamina

▶ Extracellular Matrix

Basal Transcription Factors

▶ General Transcription Factors

Basement Membrane

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

Definition

The basement membrane is a thin layer of extracellular components that lies underneath the epithelium of many organs, as well as the basal surface of the endothelium of the entire vasculature. Its main components include, but not limited to, laminin, fibronectin, entactin, proteoglycans, and collagen IV (Yurchenco and Patton 2009). The function of the basement membrane is to provide structural support for organs and for the vascular architecture. Degradation of the basement membrane by matrix metalloproteinases serves as a key step during tumor angiogenesis by facilitating endothelial cell sprouting and pericyte detachment (Jakobsson and Claesson-Welsh 2008).

Cross-References

▶ Extracellular Matrix
▶ Neovascularization

References

Jakobsson L, Claesson-Welsh L (2008) Vascular basement membrane components in angiogenesis – an act of balance. Scientific World Journal 8:1246–1249

Yurchenco PD, Patton BL (2009) Developmental and pathogenic mechanisms of basement membrane assembly. Curr Pharm Des 15(12):1277–1294

BAXS

▶ Bile Acid and Xenobiotic System

Bayes Rule

Katsuhisa Horimoto
Computational Biology Research Center, National
Institute of Advanced Industrial Science and
Technology, Koto-ku, Tokyo, Japan

Synonyms

Bayesian network model; Conditional independence

Definition

In probability theory, the definition of the conditional probability of A given B, P(A|B), is

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

where P(A) and P(B) are the probabilities of A and B, respectively. Also, P(B|A) is, by symmetry,

$$P(B|A) = \frac{P(A, B)}{P(A)}$$

Then Bayes rule is expressed as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In the situation where P(A|B) is difficult to compute directly but we have direct information about P(B|A), Bayes rule enables us to compute P(A|B) in terms of P (B|A). If the denominator P(B) in the above equation is a normalizing constant which can be computed by marginalization, then

$$P(B) = \sum_i P(B|A_i) = \sum_i P(B|A_i)P(A_i)$$

Thus, Bayes rule is also written as

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

Cross-References

- ▶ Bayesian Network Model
- ▶ Causal Relationship
- ▶ Conditional Independence
- ▶ Correlation Relationship

Bayesian

▶ Bayesian Decision Analysis

Bayesian Decision Analysis

Malcolm Farrow
School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne, UK

Synonyms

Bayesian; Decision theory

Definition

Bayesian decision analysis (Smith 2010) is concerned with making choices when the outcomes cannot be predicted with certainty. Probabilities are assigned to the various possible outcomes and so to values of the *reward* or *payoff*, under each possible choice. We choose between probability distributions of rewards. This is done by making the choice which maximizes the expected utility (▶ utility function), that is, by choosing the alternative which gives the probability distribution of rewards with the greatest expectation of utility, where utility is a function of the reward. For example, suppose that a person's utility for small financial rewards is an increasing linear function of the monetary value. Suppose that this person is given the choice between alternatives A and B with reward distributions as follows:

A: $1 with probability 0.6 or $2 with probability 0.4
B: $0 with probability 0.2 or $2 with probability 0.8

The optimal choice is then B, with expectation $1.6, while A has expectation $1.4.

See also ▶ Utility Function.

Cross-References

- ▶ Bayesian Inference
- ▶ Optimal Experiment Design

References

Smith JQ (2010) Bayesian decision analysis: principles and practice. Cambridge University Press, Cambridge

Bayesian Inference

Roger Higdon
Seattle Children's Research Institute, Seattle, WA, USA

Synonyms

Bayesian statistics

Definition

The principle of Bayesian inference is about assigning probability to "the state of knowledge" of parameters (θ) related to an experiment. To do so Bayesian inference assigns probability distribution to all unknown quantities in a statistical problem, parameters (θ) as well as the data (X). This is in contrast to frequentist interference (▶ Frequentist Approach) where parameters are fixed unknown quantities that define the frequency of the data.

Characteristics

The term "Bayesian" refers to Thomas Bayes (1702–1761), who proved a special case of what is now called Bayes' theorem (Bayes 1763). However, it was Pierre-Simon Laplace (1749–1827) who introduced a general version of the theorem and used it to approach problems in celestial mechanics, medical statistics, reliability, and jurisprudence.

Bayesian inference provides a standardized approach for analyzing statistical problems

(Berger 1985). It is based on the posterior distribution, $p(\theta/X)$, the distribution of parameters given in the data. The posterior can be represented as a constant times product of the prior distribution and the likelihood

$$p(\theta|X) = k\,p(X|\theta)p(\theta)$$

The likelihood is the joint distribution of the data as a function of θ. The prior distribution represents the state of knowledge about the experimental parameters before generating the data.

The ability to incorporate prior information is what gives Bayesian inference its power and makes it controversial. The prior distribution is a vehicle by which previous knowledge can be used to guide the statistical inference. One must be judicious in how the prior is specified since a very strong prior distribution can overwhelm the evidence given by the data. One can use an uninformative or reference prior if there is no strong initial belief about the parameters. In this case, Bayesian inference gives very similar results to maximum likelihood inference.

A simple example illustrating Bayesian inference is batting averages in the sport of baseball. Suppose, a player is observed to get six hits (successes) in ten at bats (trials) over the course of two games, the usual frequentist estimate of the player's batting average is .600. This is a very unrealistic estimate since throughout the history of the sport, only a few players have ever batted even .400. If one puts a prior distribution on the batting average p, where p is distributed beta (47,127), thus giving p a mean of .270 and standard deviation of .3 (this corresponds roughly to the typical distribution of batting averages for baseball players). So this implies that the posterior distribution is as follows.

$$p(\theta|X) \propto p^6(1-p)^4 p^{47}(1-p)^{127} = p^{53}(1-p)^{131}$$

A Bayesian estimate of the player's batting average can be found by taking the mean of the posterior distribution, which in this case is .288, a far different and likely more realistic estimate of the player's batting average than the frequentist estimate.

The use of the beta distribution (the conjugate prior for the binomial distribution) in the previous example makes calculations based on the posterior distribution quite easy. However, most Bayesian inference problems involve complex integration of the posterior distribution in order generate means or quantities such as highest posterior density regions (the Bayesian analog to the confidence interval). These calculations typically involve complex numerical approaches such as Markov Chain Monte Carlo (MCMC) (Brooks 1998).

A compromise between Bayesian and frequentist approaches are empirical Bayes methods (Carlin and Louis 2000). In empirical Bayes methods, the Bayesian distributional framework is used; however, instead of specifying the parameters in the prior distribution, the parameters are estimated from the data themselves. Empirical Bayes methods are quite commonly used in the analysis of microarrays and other high-throughput experimental data.

There are Bayesian analogs to most classical statistical approaches such as t-tests, linear regression, or chi-squared tests. Bayesian methods have become very popular in systems biology and bioinformatics because they offer a more standardized way of handling complex and noisy data. No specialized approaches are needed if models are not fully specified or if data are missing; this is not the case using traditional frequentist methods. Examples of Bayesian methods in systems biology are sequence alignment (Durbin et al. 1998), motif finding (Zhou and Liu 2004), structure prediction (Schmidler et al. 2000), microarray analysis (Gottardo et al. 2003), and Bayesian networks (Jansen et al. 2003).

Cross-References

▶ Bayesian Method
▶ Frequentist Approach

References

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Phil Trans 53:370–418

Berger JO (1985) Statistical decision theory and Bayesian analysis, 2nd edn, Springer series in statistics. Springer, New York

Brooks SP (1998) Markov chain Monte Carlo method and its application. Statistician 47:69–100

Carlin BP, Louis TA (2000) Bayes and empirical Bayes methods for data analysis, 2nd edn. Chapman & Hall/CRC Press, Boca Raton

Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge

Gottardo R, Pannucci JA, Kuske CR, Brettin T (2003) Statistical analysis of microarray data: a Bayesian approach. Biostatistics 4:597–620

Jansen R, Yu H, Greenbaum D et al (2003) Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302:449–453

Schmidler SC, Liu JS, Brutlag DL (2000) Bayesian segmentation of protein secondary structure. J Comput Biol 7:233–248

Zhou Q, Liu JS (2004) Modelling within-motif dependence for transcription factor binding site predictions. Bioinformatics 20:909–916

Bayesian Information Criterion (BIC)

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

Synonyms

Schwarz criterion; Schwarz information criterion (SIC)

Definition

In statistics, the Bayesian information criterion (BIC) (Schwarz 1978) is a model selection criterion. It is a selection criterion for choosing between different models with different numbers of parameters.

The BIC is an asymptotic result derived under the assumption that the data distribution belongs to the exponential family.

Suppose that:

1. $x =$ the observed data.
2. $n =$ the number of data points in x, or equivalently, the sample size.
3. $k =$ the number of free parameters to be estimated.
4. $p(x|k) =$ the probability of the observed data given the number of parameters.
5. $L =$ the maximized value of the likelihood function for the estimated model.

The formula for the BIC is:

$$-2\ln p(x|k) \approx BIC = -2\ln L + k\ln(n)$$

Given any two estimated models, the model with the lower BIC value is the one to be preferred.

References

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Bayesian Method

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

Synonyms

Bayesian inference; Bayesian network model

Definition

Bayesian method refers to a probability method to construct a knowledge structure used to support decision making, such as classification (▶ Classification; ▶ Identification of Gene Regulatory Networks, Machine Learning) and regression (▶ Regression Analysis) tasks, processes, or analyses. Bayesian methods are valuable whenever there is a need to extract information from data that are uncertain or subject to any kind of error or noise (including measurement error and experiment error, as well as noise or random variation intrinsic to the process of interest).

Characteristics

Traditional statistical techniques struggle to cope with complex nonlinear models that are only partially observed. Due to the fact that the Bayesian statistical paradigm is fully probabilistic, there is no fundamental distinction between any of the unknowns in a statistical model – parameters, hidden variables, and observations are all treated together in a consistent manner – and it is from this that the power of the methodology is derived. Provided that you can write down a statistical model relating the quantities you are interested in to the data you can observe (possibly via many unobserved intermediary variables), then you can carry out Bayesian method to extract the information in the data to give fully probabilistic information on all unobserved model variables.

Mathematical Formulation

In the Bayesian inference, we specify a sampling model $P(Z|\theta)$ (density or probability mass function) for our data given the parameters, and a prior distribution for the parameters $P(\theta)$ reflecting our knowledge about θ before we see the data. We then compute the posterior distribution

$$P(\theta|Z) = \frac{P(Z|\theta)P(\theta)}{P(Z)} = \frac{P(Z|\theta)P(\theta)}{\int P(Z|\theta)P(\theta)d\theta}, \quad (1)$$

The example shown in following illustrates the calculation of posterior probability for a hidden variable. Suppose there are two light bulb factories. Factory #1 has 40 qualified products and 10 unqualified products, while factory #2 has 30 qualified products and 20 unqualified products. One person chooses a factory at random, and then picks a light bulb at random. It is assumed that the two factories are treated equally, likewise for the light bulbs. The light bulb turns out to be an unqualified one. How probable is it that the light bulb is out of factory #1?

Intuitively, it seems clear that the answer should be less than a half, since there are less unqualified light bulbs in factory #1. The precise answer is given by Bayesian approach. Let H_1 correspond to factory #1, and H_2 to factory #2. It is given that the bowls are identical from the person's point of view, thus $P(H_1) = P(H_2)$, and the two must add up to 1, so both are equal to 0.5. The event E is the observation of an unqualified light bulb. From the products of the factories, we know that $P(E|H_1) = \frac{10}{50} = 0.2$ and $P(E|H_2) = \frac{20}{50} = 0.4$. Bayesian formula then yields

$$\begin{aligned} P(H_1|E) &= \frac{P(E|H_1)P(H_1)}{P(E|H_1)P(H_1) + P(E|H_2)P(H_2)} \\ &= \frac{0.2 \times 0.5}{0.2 \times 0.5 + 0.4 \times 0.5}, \\ &= 0.33 \end{aligned} \quad (2)$$

Before we observed the unqualified light bulb, the probability we assigned for the person having chosen factory #1 was the prior probability, $P(H_1)$, which was 0.5. After observing the unqualified light bulb, we must revise the probability to $P(H_1|E)$, which is 0.33.

Practical Application of Bayesian Methods

The main limiting factor in applying Bayesian methods is computational. For nontrivial problems, analytic approaches to Bayesian inference are not possible, and their numerical solution is often challenging due to the need to solve high-dimensional integration problems (which in the discrete case translate to combinatorial summation problems).

Advances in the speed of commodity computing hardware in recent decades have been paralleled by developments in computationally intensive algorithms for Bayesian inference. Arguably, the most important advance has been the development of a range of techniques based on ▶ Markov Chain Monte Carlo (MCMC). The ideas originate from statistical physics, but are now widely used for Bayesian inference. Although by no means a panacea, carefully crafted MCMC algorithms executed on fast computers are able to solve a phenomenal range of problems that would have been considered completely intractable only a few years ago. In the high-dimensional context, it is often necessary to decompose the full problem according to the underlying conditional independence structure of the model, and it is in this context that graphical model (also known as ▶ Bayesian Network Model) is particularly useful.

Gibbs sampler is just one of MCMC procedures for sampling from posterior distributions. It uses conditional sampling of each parameter given the rest, and is useful when the structure of the problem makes this sampling easy to carry out. Specifically, a Markov chain is constructed with ▶ equilibrium probability distribution $P(\theta|Z)$. Each iteration of the sampler involves cycling through each component of the K-dimensional vector θ in order and sampling from $P(\theta_i|\theta_{-i}, Z)$, $i = 1, \cdots, K$, where θ_{-i} denotes the vector of all components of θ except θ_i. Knowledge of the ▶ Bayesian network for the model can simplify the computation of these so-called full-conditional distributions. In many cases, the full-conditionals will be straightforward to sample directly, but in others, a Metropolis-Hastings method will be required. Here, a proposed new value is simulated from a largely arbitrary proposal distribution, $q(\theta_i^*|\theta_i)$ and accepted with a probability.

Bayesian Algorithms and Tools

There is a large variety of Bayesian algorithms and tools; open source tools widely used in computational biology include Matlab (Matlab, Machine Learning Tool and Matlab, Data Analysis Tool).

Cross-References

- ▶ Bayesian Network Model
- ▶ Equilibrium Probability Distribution
- ▶ Identification of Gene Regulatory Networks, Machine Learning
- ▶ Machine Learning
- ▶ Markov Chain Monte Carlo
- ▶ Regression Analysis

References

Bolstad WM (2010) Understanding computational bayesian statistics. Wiley, Hoboken

Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Wiley, Hoboken

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

Wilkinson DJ (2007) Bayesian methods in bioinformatics and computational systems biology. Brief Bioinform 8(2):109–116

Bayesian Network Model

Katsuhisa Horimoto[1] and Shigeru Saito[2]
[1]Computational Biology Research Center,
National Institute of Advanced Industrial and
Science and Technology, Koto-ku, Tokyo, Japan
[2]Infocom Corporation, Tokyo, Japan

Synonyms

Bayes rule; Causal relationship; Conditional independency

Definition

A Bayesian network model is a probabilistic graphical model of a set of variables for considering their conditional independencies in a directed acyclic graph (DAG).

Characteristics

Probabilistic Graphical Model

Let $\boldsymbol{X} = (X_1, X_2,...,X_n)$ be a set of random variables, and let x_i be a value of X_i, the i-th component of $\boldsymbol{X}$. Let $y = (x_i)_{Xi \in Y}$ be a value of $Y \subseteq X$. Then, a probabilistic graphical model for $\boldsymbol{X}$ is a graphical factorization of the joint generalized probability density function, $f(\boldsymbol{X} = \boldsymbol{x})$. The representation of this model is given by two concepts: a structure and a set of local generalized probability densities (Pearl 2000).

The structure S for $\boldsymbol{X}$ is a directed acyclic graph (DAG) that describes a set of conditional independencies (Dawid 1979) about the variables on X. Pa_i^S represents the set of parents (variables from which an arrow is coming out in S) of the variable X_i in the probabilistic graphical model whose structure is given by S. The structure S for $\boldsymbol{X}$ assumes that X_i and its non-descendants are independent given Pa_i^S, $i = 2,..., n$. Therefore, the factorization can be written as follows:

$$f(x) = f(x_1,\ x_2,\ \ldots\ .\ x_n) = \prod_{i=1}^{n} f\left(x_i|pa_i^S\right) \tag{1}$$

A representation of the models of the characteristics described in Eq. 1 assumes that the local generalized probability densities depend on a finite set of parameters $\boldsymbol{\theta}_S \in \boldsymbol{\Theta}_S$, and as a result, Eq. 1 can be rewritten as follows:

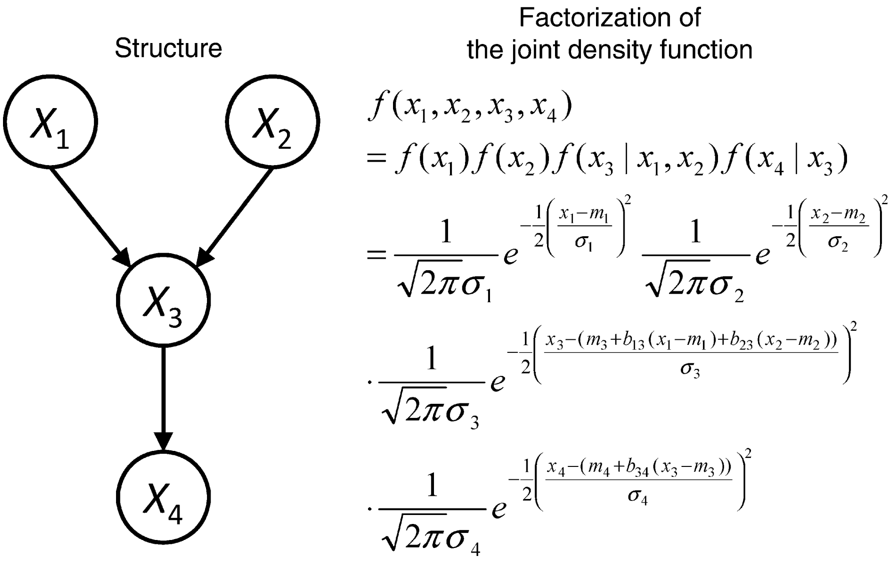
$$f(x|\theta_S) = \prod_{i=1}^{n} f\left(x_i|pa_i^S,\ \theta_i\right) \tag{2}$$

where $\theta_S = (\theta_1, \theta_2,..., \theta_n)$.

Bayesian Network Model

Bayesian network model is the probabilistic graphical model in the particular case of every variable $X_i \in X$

Bayesian Network Model, Fig. 1 Structure and resulting factorization for a Gaussian network with four variables


Structure
X_1
X_2
X_3
X_4
Factorization of the joint density function
$f(x_1,x_2,x_3,x_4)$
$= f(x_1)f(x_2)f(x_3 \mid x_1,x_2)f(x_4 \mid x_3)$
$= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{x_1-m_1}{\sigma_1}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{x_2-m_2}{\sigma_2}\right)^2}$
$\cdot \frac{1}{\sqrt{2\pi}\sigma_3} e^{-\frac{1}{2}\left(\frac{x_3-(m_3+b_{13}(x_1-m_1)+b_{23}(x_2-m_2))}{\sigma_3}\right)^2}$
$\cdot \frac{1}{\sqrt{2\pi}\sigma_4} e^{-\frac{1}{2}\left(\frac{x_4-(m_4+b_{34}(x_3-m_3))}{\sigma_4}\right)^2}$

being discrete. If the variable X_i has r_i possible values, $x_i^1, \ldots, x_i^{ri}$, the local distribution, $g\left(x_i | pa_i^{j,S}, \theta_i\right)$ is an unrestricted discrete distribution:

$$g\left(x_i^k \middle| pa_i^{j,S}, \theta_i\right) = \theta_{x_i^k | pa_i^j} \tag{3}$$

where $pa_i^{1,S}, \ldots, pa_i^{qi,S}$ denotes the values of Pa_i^S, that is, the set of parents of the variable X_i in the structure S; q_i is the number of different possible instantiations of the parent variables of X_i. The local parameters are given by $\theta_i = \left(\left(\theta_{ijk}\right)_{k=1}{}^{ri}\right)_{j=1}{}^{qi}\Big)$. In other words, the parameter θ_{ijk} represents the conditional probability that variable X_i takes its k-th value, knowing that its parent variables have taken their j-th combination of values. We assume that every θ_{ijk} is greater than zero.

Gaussian Network Model

Here, one example of Bayesian network model is illustrated, which assumes the joint density function to be a multivariate Gaussian density, Gaussian network model (Whittaker 1990).

An individual $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ consists of a continuous value in $\mathfrak{R}^n$. The local density function for the i-th variable X_i can be computed as the linear regression model

$$f\left(x_i \middle| pa_i^S, \theta_i\right) = N\left(x_i;\ m_i + \sum_{x_j \in pa_i} b_{ji}(x_j - m_j),\ v_i\right) \tag{4}$$

where $N\left(x_i;\ \mu_i,\ \sigma_i^2\right)$ is a univariate normal distribution with mean μ_i and variance $v_i = \sigma_i^2$ for the i-th variable.

Taking this definition into account, an edge missing from X_j to X_i implies $b_{ji} = 0$ in the former linear-regression model. The local parameters are given by $\boldsymbol{\theta}_i = (m_i, \boldsymbol{b}_i, v_i)$, where $b_i = (b_{1i}, b_{2i}, \ldots, b_{i-1i})^{\mathrm{T}}$ is a column vector. A probabilistic graphical model built from these local density functions is known as a Gaussian network (Shachter and Kenley 1989).

The components of the local parameters are as follows: m_i is the unconditional mean of X_i, v_i is the conditional variance of X_i given $\boldsymbol{Pa}_\mathrm{i}$, and b_{ji} is a linear coefficient that measures the strength of the relationship between X_j and X_i. Figure 1 is an example of a Gaussian network in a four-dimensional space.

For investigating how Gaussian networks and multivariate normal densities are related, the joint density function of the continuous n-dimensional variable $\boldsymbol{X}$ is by definition a multivariate normal distribution iff

$$\begin{aligned} f(x) &= N(x; \mu,\ \Sigma) \\ &= (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \end{aligned} \tag{5}$$

where $\boldsymbol{\mu}$ is the vector of means, Σ is covariance matrix $n \times n$, and $|\Sigma|$ denotes the determinant of Σ. The inverse of this matrix, $W = \Sigma^{-1}$, in which elements are denoted by w_{ij}, is known as the precision matrix.

This density can also be written as a product of n conditional densities using the chain rule, namely,

$$f(x) = \prod_{i=1}^{n} f(x_i | x_1,\ x_2,\ \ldots .x_{i-1}) \\ = \prod_{i=1}^{n} N\left(x_i;\ \mu_i + \sum_{j=1}^{i-1} b_{ji}(x_j - \mu_j),\ v_i\right) \quad (6)$$

where μ_i is the unconditional mean of X_i, v_i is the variance of X_i given X_1, X_2,..., X_{i-1}, and b_{ji} is a linear coefficient reflecting the strength of the relationship between variables X_j and X_i. This notation allows us to represent a multivariate normal distribution as a Gaussian network, where for any $b_{ji} \neq 0$ with $j < i$ this network will contain an edge from X_j to X_i.

Applications of Biological Network

The application of Bayesian network model to the biological network is first found in Friedman et al. (2000). One of the important research themes for applying Bayesian network model to biological network is to develop algorithm for obtaining accurate network structure in reasonable computational time, which is called "structure learning." Currently, approaches for structure learning are roughly divided into two categories. One is "score-based method," which provides a structure by maximizing some scoring function with respect to the posterior probability of the structure, such as Bayesian Dirichlet equivalent (BDe) and Bayesian information criterion (BIC), by using several maximization methods, such as greedy, K2, and Markov chain Monte Carlo. The other is "constraint-based method," which provides a structure by checking conditional independency among nodes in the graph, such as SGS/PC-algorithm and CI-algorithm. The score-based methods try to get an optimal structure in terms of likelihood (i.e., goodness of fit of the model). On the other hand, the constraint-based method focuses on consistency of conditional independencies in the graph (local Markov property).

According to improvement of measurement technology like DNA microarray, Bayesian networks have become quite a popular approach for genetic network inference. Nevertheless, although a lot of successful applications of Bayesian networks under various biological settings can be enumerated, Bayesian networks still have several limitations. One is that Bayesian networks accept only acyclic graphs, that is, they cannot represent feedback loops in the networks. It is well known that cyclic machinery is a common mechanism in various biological functions. Another is a more serious problem: Arbitral structure learning in Bayesian network is an NP-hard problem, which means that we are allowed to use only heuristic strategies for finding approximate solutions. But yet, Bayesian network is of certain worth in terms of its handiness and ease of interpretation.

Cross-References

- ▶ Bayes Rule
- ▶ Bayesian Method
- ▶ Causal Relationship
- ▶ Conditional Independency
- ▶ Probabilistic Graphical Model

References

Dawid AP (1979) Conditional independence in statistical theory. J Roy Stat Soc Ser B 41:1–31

Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7:601–620

Pearl J (2000) Causality: models, reasoning, and inference. Cambridge University Press, Cambridge/New York

Shachter R, Kenley C (1989) Gaussian influence diagrams. Manag Sci 35:527–550

Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, Chichester

Bayesian Statistics

- ▶ Bayesian Inference

B Cell-mediated Immunity

- ▶ B Cell-mediated Immune Response

BCR Receptor Diversity

- ▶ Immune Repertoire Diversity

Belief Elicitation

▶ Prior Elicitation

Bench-to-Bedside Research

▶ Translational Research

Benign

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

Definition

Localized disorganized tissue mass that does not spread and is amenable to local excision. The tissue characteristics are more similar to the tissue it is derived from.

Cross-References

▶ Cancer Pathology

Benjamini–Hochberg Method

Winston Haynes
Seattle Children's Research Institute, Seattle,
WA, USA

Definition

The Benjamini–Hochberg method controls the False Discovery Rate (FDR) using sequential modified ▶ Bonferroni correction for ▶ multiple hypothesis testing. While the ▶ Bonferroni correction relies on the Family Wise Error Rate (FWER), Benjamini and Hochberg introduced the idea of a FDR to control for multiple hypotheses testing. In the statistical context, discovery refers to the rejection of a hypothesis. Therefore, a false discovery is an incorrect rejection of a hypothesis and the FDR is the likelihood such a rejection occurs. Controlling the FDR instead of the FWER is less stringent and increases the method's power. As a result, more hypotheses may be rejected and more discoveries may be made.

In the Benjamini–Hochberg method, hypotheses are first ordered and then rejected or accepted based on their *p*-values. A *p*-value is a data point for each hypothesis describing the likelihood of an observation based on a ▶ probability distribution. The Benjamini–Hochberg method begins by ordering the *m* hypothesis by ascending *p*-values, where P_i is the *p*-value at the *i*th position with the associated hypothesis H_i. Let *k* be the largest *i* for which:

$$P_i \leq \frac{i}{m} q$$

Reject hypotheses $i = 1, 2, 3, \ldots, k$. The Benjamini–Hochberg method has been proven to control the FDR for all tests at a level of q.

References

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. J R Stat Soc B 57:289–300

Beta Workbench

▶ BlenX

BetaWB

▶ BlenX

Bias

Olga Vitek
Department of Statistics, Department of Computer
Science, Purdue University, West Lafayette, IN, USA

Definition

Bias is a property of experimental design defined as a systematic error in data-derived conclusions regarding the unknowns.

Cross-References

▶ Designing Experiments for Sound Statistical Inference

Bifan

▶ Canonical Network Motifs

Bifurcation

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

Definition

Bifurcation means the splitting of a main body into two parts. *Bifurcation theory* is the mathematical study of changes in the qualitative or topological structure of a given family, such as the integral curves of a family of vector fields, and the solutions of a family of differential equations. Most commonly applied to the mathematical study of dynamical systems, a *bifurcation* occurs when a small smooth change made to the parameter values (the bifurcation parameters) of a system causes a sudden "qualitative" or topological change in its behavior. Bifurcations can occur in both continuous systems (described by ▶ ODEs, DDEs, or PDEs) and discrete systems (described by maps).

Characteristics

Bifurcation Diagram

In mathematics, particularly in dynamical systems, a *bifurcation diagram* shows the possible long-term values (equilibria/fixed points or periodic orbits) of a system as a function of a bifurcation parameter in the system. It is usual to represent stable solutions with a solid line and unstable solutions with a dotted line.

An example is the bifurcation diagram of the logistic map:

$$x_{n+1} = rx_n(1 - x_n) \tag{1}$$

The bifurcation parameter r is shown on the horizontal axis of the plot and the vertical axis shows the possible long-term population values of the logistic function. Only the stable solutions are shown here; there are many other unstable solutions which are not shown in this diagram.

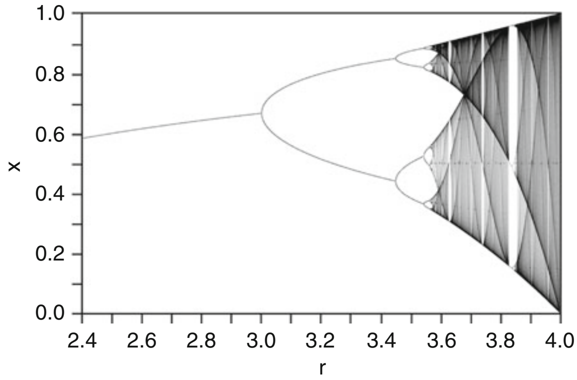
The bifurcation diagram nicely shows the forking of the possible periods of stable orbits from 1–2 to 4–8, etc. Each of these bifurcation points is a period-doubling bifurcation. The ratio of the lengths of successive intervals between values of r for which bifurcation occurs converges to the first Feigenbaum constant (Fig. 1).

Bifurcation Types

It is useful to divide bifurcations into two principal classes:

Local bifurcations, which can be analyzed entirely through changes in the local stability properties of ▶ equilibria, periodic orbits, or other invariant sets as parameters cross through critical thresholds.

Global bifurcations, which often occur when larger invariant sets of the system "collide" with each other, or with equilibria of the system. They cannot be detected purely by a stability analysis of the equilibria (fixed points).



Bifurcation, Fig. 1 Bifurcation diagram of the logistic map

Local Bifurcations

A local bifurcation occurs when a parameter change causes the stability of an equilibrium (or fixed point) to change. In continuous systems, this corresponds to the real part of an eigenvalue of an equilibrium passing through zero. In discrete systems (those described by maps rather than ODEs), this corresponds to a fixed point having a ▶ Floquet multiplier with modulus equal to one. In both cases, the equilibrium is *nonhyperbolic* at the bifurcation point. The topological changes in the phase portrait of the system can be confined to arbitrarily small neighborhoods of the bifurcating fixed points by moving the bifurcation parameter close to the bifurcation point (hence "local").

More technically, consider the continuous dynamical system described by the ODE:

$$\dot{x} = f(x; \lambda),\ f : R^n \times R \to R \tag{2}$$

A local bifurcation occurs at (x_0, λ_0) if the ▶ Jacobian matrix $Df(x_0; \lambda_0)$ has an ▶ eigenvalue with zero real part. If the eigenvalue is equal to zero, the bifurcation is a steady state bifurcation, but if the eigenvalue is nonzero but purely imaginary, this is a ▶ Hopf bifurcation.

For discrete dynamical systems, consider the system:

$$x_{n+1} = f(x_n; \lambda) \tag{3}$$

Then a local bifurcation occurs at (x_0, λ_0) if the matrix $Df(x_0; \lambda_0)$ has an eigenvalue with modulus equal to one. If the eigenvalue is equal to one, the bifurcation is either a saddle-node (often called fold bifurcation in maps), transcritical, or pitchfork bifurcation. If the eigenvalue is equal to -1, it is a period-doubling (or flip) bifurcation, and otherwise, it is a Hopf bifurcation.

Examples of local bifurcations include:

1. ▶ Saddle-node (fold) bifurcation
2. Transcritical bifurcation
3. Pitchfork bifurcation
4. Period-doubling (flip) bifurcation
5. ▶ Hopf bifurcation
6. Neimark (secondary Hopf) bifurcation

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

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

The normal form of a saddle-node bifurcation is:

$$\dot{x} = r + x^2 \tag{4}$$

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

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

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

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

Transcritical Bifurcation In ▶ bifurcation theory, a field within mathematics, a *transcritical bifurcation* is a particular kind of local bifurcation, meaning that it is characterized by an equilibrium having an eigenvalue whose real part passes through zero.

Both before and after the bifurcation, there is one unstable and one stable fixed point. However, their stability is exchanged when they collide. So the unstable fixed point becomes stable and vice versa.

The normal form of a transcritical bifurcation is:

$$\dot{x} = rx - x^2 \tag{5}$$

This equation is similar to logistic equation but in this case we allow r and x to be positive or negative (while in the logistic equation x and r must be nonnegative). The two fixed points are at $x = 0$ and $x = r$. When the parameter r is negative, the fixed point at $x = 0$ is stable and the fixed point $x = r$ is unstable. But for $r > 0$, the point at $x = 0$ is unstable and the point at $x = r$ is stable. So the bifurcation occurs at $r = 0$.

A typical example (in real life) could be the consumer-producer problem where the consumption is proportional to the (quantity of) resource.

For example:

$$\dot{x} = rx(1 - x) - px \tag{6}$$

where $rx(1 - x)$ is the logistic equation of resource growth; and px is the consumption, proportional to the resource x.

Pitchfork Bifurcation In ▶ bifurcation theory, a field within mathematics, a *pitchfork bifurcation* is a particular type of local bifurcation. Pitchfork bifurcations, like ▶ Hopf bifurcations, have two types – supercritical or subcritical.

In flows, that is, continuous dynamical systems described by ▶ ODEs, pitchfork bifurcations occur generically in systems with symmetry.

An ODE:

$$\dot{x} = f(x; r) \tag{7}$$

described by a one parameter function $f(x; r)$with $r \in R$ satisfying: $-f(x; r) = f(-x; r)$ (f is an odd function with regard to x), $\frac{\partial f}{\partial x}(0; r_0) = 0$, $\frac{\partial f^2}{\partial x^2}(0; r_0) = 0$, $\frac{\partial f^3}{\partial x^3}(0; r_0) \neq 0$, $\frac{\partial f}{\partial r}(0; r_0) = 0$, $\frac{\partial f^2}{\partial x \partial r}(0; r_0) \neq 0$ has a *pitchfork bifurcation* at $(x, r) = (0, r_0)$. The form of the pitchfork is given by the sign of the third derivative:

$$\frac{\partial f^3}{\partial x^3}(0; r_0) \begin{cases} < 0 & \text{super}critical \\ > 0 & \text{sub}critical \end{cases} \tag{8}$$

Period-Doubling Bifurcation In mathematics, a *period-doubling bifurcation* in a discrete dynamical system is a bifurcation in which the system switches to a new behavior with twice the period of the original system. Period-doubling bifurcations can also occur in continuous dynamical systems, namely, when a new ▶ limit cycle emerges from an existing limit cycle, and the period of the new limit cycle is twice that of the old one (Fig. 2).

Supercritical/Subcritical Hopf Bifurcations The limit cycle is orbitally stable if a certain quantity called the *first Lyapunov coefficient* is negative, and the bifurcation is supercritical. Otherwise it is unstable and the bifurcation is subcritical.

The normal form of a Hopf bifurcation is:

$$\dot{z} = z\left(\lambda + b|z|^2\right) \tag{9}$$

where z, b are both complex and λ is a parameter. Write $b = \alpha + i\beta$. The number α is called the first Lyapunov coefficient.

If α is negative, then there is a stable limit cycle for $\lambda > 0$:

$$z(t) = re^{i\omega t} \tag{10}$$

where

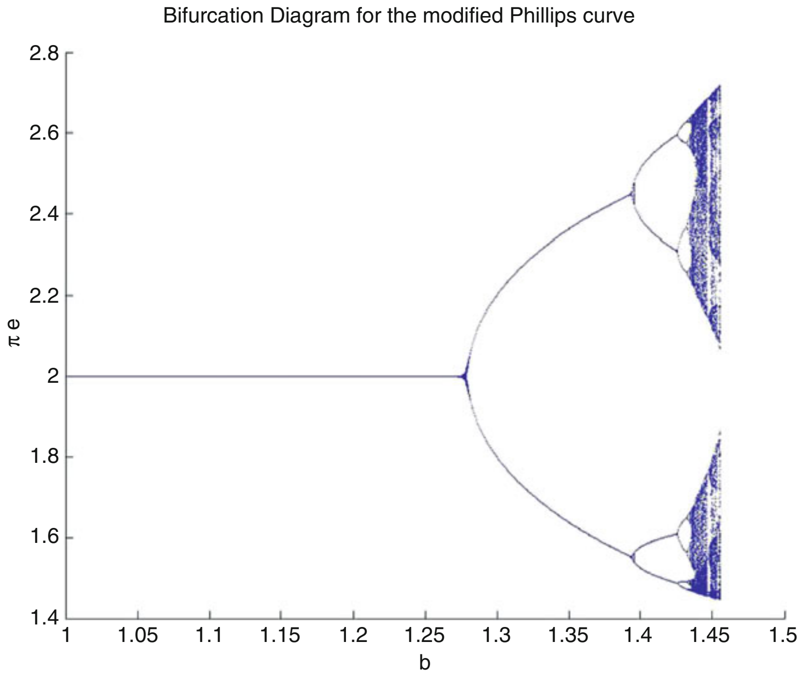
$$r = \sqrt{-\lambda/\alpha} \quad \text{and} \quad \omega = \beta r^2 \tag{11}$$

The bifurcation is then called *supercritical*.

If α is positive then there is an unstable limit cycle for $\lambda < 0$. The bifurcation is called *subcritical*.

Global Bifurcation

Global bifurcations occur when “larger” invariant sets, such as periodic orbits, collide with equilibria. This causes changes in the topology of the trajectories in the phase space which cannot be confined to a small neighborhood, as is the case with local bifurcations.



Bifurcation, Fig. 2 Bifurcation diagram for the modified Phillips curve

In fact, the changes in topology extend out to an arbitrarily large distance (hence "global").

Examples of global bifurcations include:

1. Homoclinic bifurcation in which a ▶ limit cycle collides with a saddle point (▶ Saddle-Node Bifurcation)
2. Heteroclinic bifurcation in which a limit cycle collides with two or more saddle points
3. Infinite-period bifurcation in which a stable node and saddle point simultaneously occur on a limit cycle
4. Blue sky catastrophe in which a limit cycle collides with a nonhyperbolic cycle

Global bifurcations can also involve more complicated sets such as chaotic attractors.

Homoclinic Bifurcation In mathematics, a *homoclinic bifurcation* is a global bifurcation which often occurs when a periodic orbit collides with a saddle point.

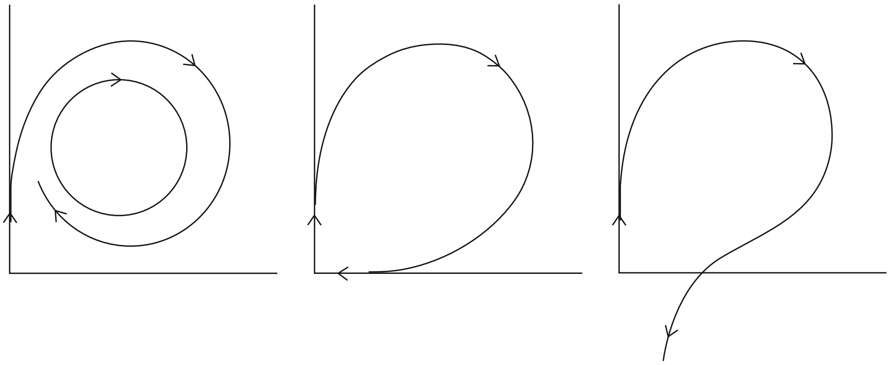
The image below shows a phase portrait before, at, and after a homoclinic bifurcation in 2D. The periodic orbit grows until it collides with the saddle point. At the bifurcation point the period of the periodic orbit has grown to infinity and it has become a homoclinic orbit. After the bifurcation there is no longer a periodic orbit (Fig. 3).

Heteroclinic Bifurcation In mathematics, particularly dynamical systems, a *heteroclinic bifurcation* is a global bifurcation involving a heteroclinic cycle. Heteroclinic bifurcations come in two types: resonance bifurcations and transverse bifurcations. Both types of bifurcations will result in the change of stability of the heteroclinic cycle.

At a resonance bifurcation, the stability of the cycle changes when an algebraic condition on the eigenvalues of the equilibria in the cycle is satisfied. This is usually accompanied by the birth or death of a periodic orbit.

A transverse bifurcation of a heteroclinic cycle is caused when the real part of a transverse eigenvalue of one of the equilibria in the cycle passes through zero. This will also cause a change in stability of the heteroclinic cycle.

Global Bifurcation In mathematics, an *infinite-period bifurcation* is a global bifurcation that can occur when two fixed points emerge on a limit cycle. As the limit of a parameter approaches a certain critical value, the speed of the oscillation slows down and the period approaches infinity. The infinite-period bifurcation occurs at this critical value. Beyond the critical value, the two fixed points emerge continuously from



Bifurcation, Fig. 3 A homoclinic bifurcation occurs when a periodic orbit collides with a saddle point. *Left panel*: For small parameter values, there is a saddle point at the origin and a limit cycle in the first quadrant. *Middle panel*: As the bifurcation parameter increases, the limit cycle grows until it exactly intersects the saddle point, yielding an orbit of infinite duration. *Right panel*: When the bifurcation parameter increases further, the limit cycle disappears completely

each other on the limit cycle to disrupt the oscillation and form two saddle points.

Blue Sky Catastrophe The *blue sky catastrophe* is a type of ▶ bifurcation of a periodic orbit. In other words, it describes a sort of behavior that stable solutions of a set of differential equations can undergo as the equations are gradually changed. This type of bifurcation is characterized by both the period and the length of the orbit approaching infinity as the control parameter approaches a finite bifurcation value, but with the orbit still remaining within a bounded part of the phase space, and without loss of ▶ stability before the bifurcation point. In other words, the orbit *vanishes into the blue sky*.

The bifurcation has found application in, among other places, slow-fast models of computational neuroscience. The possibility of the phenomenon was raised by David Ruelle and Floris Takens in 1971, and explored by R.L. Devaney and others in the following decade. A more compelling analysis was not performed until the 1990s.

This bifurcation has also been found in the context of fluid dynamics, namely, in double-diffusive convection of a small Prandtl number fluid. Double-diffusive convection occurs when convection of the fluid is driven by both thermal and concentration gradients, and the temperature and concentration diffusivities take different values. The bifurcation is found in an orbit that is born in a global saddle-loop bifurcation, becomes chaotic in a period-doubling cascade, and disappears in the blue sky catastrophe.

Symmetry Breaking in Bifurcation Sets
In a dynamical system such as:

$$\ddot{x} + f(x; \mu) + \varepsilon g(x) = 0 \tag{12}$$

which is structurally stable when $\mu \neq 0$, if a bifurcation diagram is plotted, treating μ as the bifurcation parameter, but for different values of ε, the case ε = 0 is the symmetric pitchfork bifurcation. When ε ≠ 0, we say we have a pitchfork with *broken symmetry*. This is illustrated in Fig. 4.

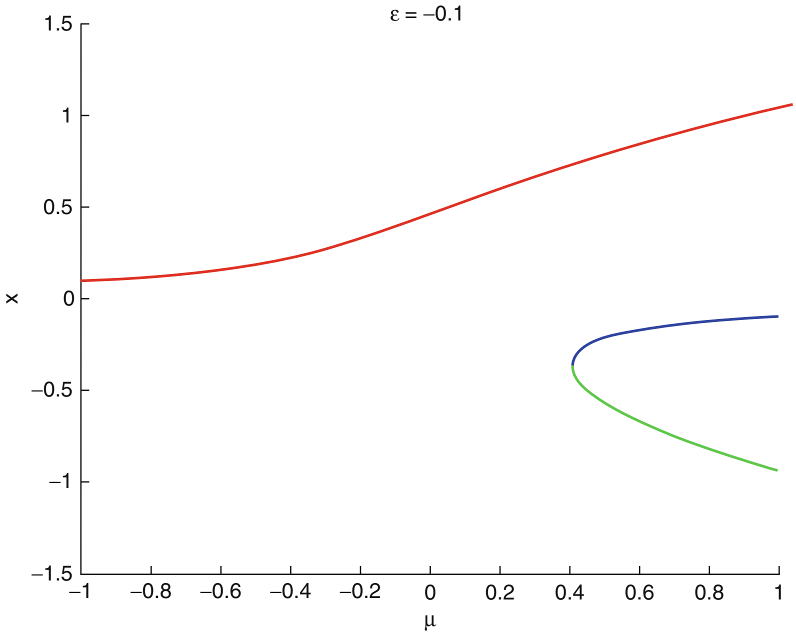
Codimension of a Bifurcation

The codimension of a bifurcation is the number of parameters which must be varied for the bifurcation to occur. This corresponds to the codimension of the parameter set for which the bifurcation occurs within the full space of parameters. Saddle-node bifurcations are the only generic local bifurcations which are really codimension-one (the others all having higher codimension). However, often transcritical and pitchfork bifurcations are also often thought of as codimension-one, because the normal forms can be written with only one parameter.

An example of a well-studied codimension-two bifurcation is the Bogdanov–Takens bifurcation.

Catastrophe Theory

In mathematics, *catastrophe theory* is a branch of bifurcation theory in the study of dynamical systems (▶ Dynamical Systems Theory, Bifurcation Analysis);



Bifurcation, Fig. 4 Symmetry breaking in pitchfork bifurcation as the parameter epsilon is varied. $\varepsilon = 0$ is the case of symmetric pitchfork bifurcation

it is also a particular special case of more general singularity theory in geometry.

Bifurcation theory studies and classifies phenomena characterized by sudden shifts in behavior arising from small changes in circumstances, analyzing how the qualitative nature of equation solutions depends on the parameters that appear in the equation. This may lead to sudden and dramatic changes, for example, the unpredictable timing and magnitude of a landslide.

Catastrophe theory, which originated with the work of the French mathematician René Thom in the 1960s, and became very popular due to the efforts of Christopher Zeeman in the 1970s, considers the special case where the long-run stable equilibrium can be identified with the minimum of a smooth, well-defined potential function (▶ Lyapunov Stability).

Small changes in certain parameters of a nonlinear system can cause equilibria to appear or disappear, or to change from attracting to repelling and vice versa, leading to large and sudden changes of the behavior of the system. However, examined in a larger parameter space, catastrophe theory reveals that such bifurcation points tend to occur as part of well-defined qualitative geometrical structures.

Catastrophe theory analyses *degenerate critical points* of the potential function – points where not just the first derivative but one or more higher derivatives of the potential function are also zero. These are called the germs of the catastrophe geometries. The degeneracy of these critical points can be *unfolded* by expanding the potential function as a Taylor series in small perturbations of the parameters.

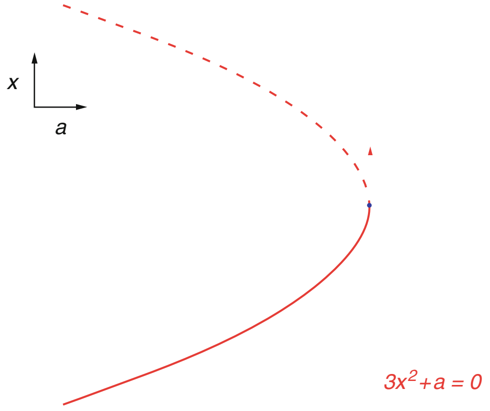
When the degenerate points are not merely accidental, but are structurally stable, the degenerate points exist as organizing centers for particular geometric structures of lower degeneracy, with critical features in the parameter space around them. If the potential function depends on two or fewer active variables, and four (respectively, five) or fewer active parameters, then there are only seven (respectively, eleven) generic structures for these bifurcation geometries, with corresponding standard forms into which the Taylor series around the catastrophe germs can be transformed by diffeomorphism (a smooth transformation whose inverse is also smooth). These seven fundamental types are now presented, with the names that Thom gave them.

Fold Catastrophe

The potential function of one active variable is:

$$V = x^3 + ax \tag{13}$$

At negative values of a, the potential has two extrema – one stable and one unstable. If the parameter a is slowly increased, the system can follow the stable minimum point. But at $a = 0$ the stable and unstable



Bifurcation, Fig. 5 Stable and unstable pair of extrema disappear at a fold bifurcation

extrema meet, and annihilate. This is the bifurcation point. At $a > 0$ there is no longer a stable solution. If a physical system is followed through a fold bifurcation, one therefore finds that as a reaches 0, the stability of the $a < 0$ solution is suddenly lost, and the system will make a sudden transition to a new, very different behavior. This bifurcation value of the parameter a is sometimes called the tipping point (Fig. 5).

Cusp Catastrophe

The potential function of one active variable is:

$$V = x^4 + ax^2 + bx \tag{14}$$

The cusp geometry is very common, when one explores what happens to a fold bifurcation if a second parameter, b, is added to the control space. Varying the parameters, one finds that there is now a *curve* (blue) of points in (a, b) space where stability is lost, where the stable solution will suddenly jump to an alternate outcome.

But in a cusp geometry the bifurcation curve loops back on itself, giving a second branch where this alternate solution itself loses stability, and will make a jump back to the original solution set. By repeatedly increasing b and then decreasing it, one can therefore observe hysteresis loops, as the system alternately follows one solution, jumps to the other, follows the other back, then jumps back to the first.

However, this is only possible in the region of parameter space $a < 0$. As a is increased, the hysteresis loops become smaller and smaller, until above $a = 0$ they disappear altogether (the cusp catastrophe), and there is only one stable solution.

One can also consider what happens if one holds b constant and varies a. In the symmetrical case $b = 0$, one observes a pitchfork bifurcation as a is reduced, with one stable solution suddenly splitting into two stable solutions and one unstable solution as the physical system passes to $a < 0$ through the cusp point $(0, 0)$ (an example of spontaneous symmetry breaking). Away from the cusp point, there is no sudden change in a physical solution being followed: when passing through the curve of fold bifurcations, all that happens is an alternate second solution becomes available.

A famous suggestion is that the cusp catastrophe can be used to model the behavior of a stressed dog, which may respond by becoming cowed or becoming angry. The suggestion is that at moderate stress ($a > 0$), the dog will exhibit a smooth transition of response from cowed to angry, depending on how it is provoked. But higher stress levels correspond to moving to the region ($a < 0$). Then, if the dog starts cowed, it will remain cowed as it is irritated more and more, until it reaches the “fold” point, when it will suddenly, discontinuously snap through to angry mode. Once in “angry” mode, it will remain angry, even if the direct irritation parameter is considerably reduced.

Another application example is for the outer sphere electron transfer frequently encountered in chemical and biological systems (Xu 1990).

Fold bifurcations and the cusp geometry are by far the most important practical consequences of catastrophe theory. They are patterns which reoccur again and again in physics, engineering, and mathematical modeling. They are the only way we currently have of detecting black holes and the dark matter of the universe, via the phenomenon of gravitational lensing producing multiple images of distant quasars.

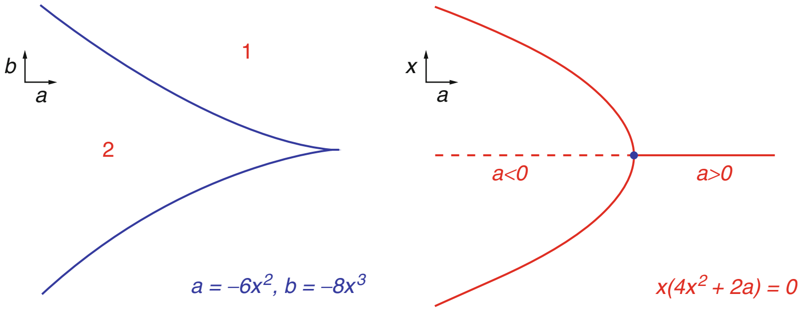
The remaining simple catastrophe geometries are very specialized in comparison, and presented here only for curiosity value (Fig. 6).

Swallowtail Catastrophe

The potential function of one active variable is:

$$V = x^5 + ax^3 + bx^2 + cx \tag{15}$$

The control parameter space is three dimensional. The bifurcation set in parameter space is made up of three surfaces of fold bifurcations, which meet in two lines of cusp bifurcations, which in turn meet at a single swallowtail bifurcation point.



Bifurcation, Fig. 6 (*Left*) Cusp shape in parameter space (*a*,*b*) near the catastrophe point, showing the locus of fold bifurcations separating the region with two stable solutions from the region with one; (*right*) pitchfork bifurcation at $a = 0$ on the surface $b = 0$

As the parameters go through the surface of fold bifurcations, one minimum and one maximum of the potential function disappear. At the cusp bifurcations, two minima and one maximum are replaced by one minimum; beyond them the fold bifurcations disappear. At the swallowtail point, two minima and two maxima all meet at a single value of x. For values of $a > 0$, beyond the swallowtail, there is either one maximum-minimum pair, or none at all, depending on the values of b and c. Two of the surfaces of fold bifurcations, and the two lines of cusp bifurcations where they meet for $a < 0$, therefore disappear at the swallowtail point, to be replaced with only a single surface of fold bifurcations remaining. Salvador Dalí's last painting, *The Swallow's Tail*, was based on this catastrophe.

Butterfly Catastrophe

The potential function of one active variable is:

$$V = x^6 + ax^4 + bx^3 + cx^2 + dx \qquad (16)$$

Depending on the parameter values, the potential function may have three, two, or one different local minima, separated by the loci of fold bifurcations. At the butterfly point, the different three surfaces of fold bifurcations, the two surfaces of cusp bifurcations, and the lines of swallowtail bifurcations all meet up and disappear, leaving a single cusp structure remaining when $a > 0$.

References

Blanchard P, Devaney RL, Hall GR (2006) Differential equations. Thompson, London, pp 96–111

Courtney M et al (1995) Closed orbit bifurcations in continuum stark spectra. Phys Rev Lett 74:1538–1541

Founargiotakis M, Farantos SC, Skokos CH, Contopoulos G (1997) Bifurcation diagrams of periodic orbits for unbound molecular systems: FH2. Chem Phys Lett 277(5–6):456–464

Galan J, Freire E (1999) Chaos in a mean field model of coupled quantum wells; bifurcations of periodic orbits in a symmetric hamiltonian system. Rep Math Phys 44(1):81–86

Gao J, Delos JB (1997) Quantum manifestations of bifurcations of closed orbits in the photoabsorption spectra of atoms in electric fields. Phys Rev A 56:356–364

Gutzwiller MC (1990) Chaos in classical and quantum mechanics. Springer, New York. ISBN 0-387-97173-4

http://www.scholarpedia.org/article/Quantum_chaos

http://www.scholarpedia.org/article/User:Gutzwiller

Kleppner D, Delos JB (2001) Beyond quantum mechanics: insights from the work of Martin Gutzwiller. Found Phys 31(4):593–612

Monteiro TS, Saraga DS (2001) Quantum wells in tilted fields: semiclassical amplitudes and phase coherence times. Foundations Phys 31(2):355

Peters AD, Jaffé C, Delos JB (1994) Quantum manifestations of bifurcations of classical orbits: an exactly solvable model. Phys Rev Lett 73:2825–2828

Stamatiou G, Ghikas DPK (2007) Quantum entanglement dependence on bifurcations and scars in non-autonomous systems: the case of quantum kicked top. Phys Lett A 368(3–4):206–214

Wieczorek S, Krauskopf B, Simpson TB, Lenstra D (2005) The dynamical complexity of optically injected semiconductor lasers. Phys Rep 416(1–2):1–128

Xu F (1990) Application of catastrophe theory to the $\Delta G^{\neq}$ to $-\Delta G$ relationship in electron transfer reactions. Zeitschrift für Physikalische Chemie Neue Folge 166:79–91

Bifurcation, Supercritical and Subcritical

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

Definition

In bifurcation theory, a field within mathematics, a *pitchfork bifurcation* is a particular type of local bifurcation. Pitchfork bifurcations, like Hopf bifurcations, have two types – supercritical and subcritical.

In flows, that is, continuous dynamical systems described by ODE, pitchfork bifurcations occur generically in systems with symmetry.

An ODE

$$\frac{dx}{dt} = f(x;r)$$

described by a one parameter function $f(x;r)$ with $r \in R$ satisfying: $-f(x;r) = f(-x;r)$ (f is an odd function with regard to x), $\frac{\partial f}{\partial x}(0;r_0) = 0$, $\frac{\partial f^2}{\partial x^2}(0;r_0) = 0$, $\frac{\partial f^3}{\partial x^3}(0;r_0) \neq 0$, $\frac{\partial f}{\partial r}(0;r_0) = 0$, $\frac{\partial f^2}{\partial x \partial r}(0;r_0) \neq 0$ has a *pitchfork bifurcation* at $(x,r) = (0,r_0)$. The form of the pitchfork is given by the sign of the third derivative:

$$\frac{\partial f^3}{\partial x^3}(0;r_0)\begin{cases} <0 & \text{supercritical} \\ >0 & \text{subcritical} \end{cases}$$

Bigraph

▶ Bipartite Graph

Bile Acid and Xenobiotic System

Noel Kennedy[1], Paul Thompson[1], Oliver Schmidt[1], Werner Dubitzky[2] and Huiru Zheng[3]
[1]School of Biomedical Sciences, University of Ulster, Coleraine, UK
[2]Biomedical Sciences Research Institute, University of Ulster, Coleraine, UK
[3]School of Computing and Mathematics, Computer Science Research Institute, University of Ulster, Jordanstown, UK

Synonyms

BAXS; Bile acid system

Definition

The bile acid and xenobiotic system (BAXS) defines an intricate physiological network of chemoprotective and transporter-related functions that ensure the detoxification and removal from the body of harmful xenobiotic and endobiotic compounds while ensuring that primary bile acids (essential for the emulsification and absorption of dietary fats and fat-soluble vitamins) are not eliminated and can be reused. Xenobiotics are chemical compounds which are foreign to the living organism. They are not naturally found or expected to be present in the organism as they are neither produced by it nor are they part of the organism's natural diet. The process of xenobiotic metabolism ensures they are broken down and either put to good use or detoxified and removed from the organism. Examples of xenobiotics are drugs, pesticides, and carcinogens. Endobiotics are chemicals produced by an organism, such as steroid hormones or bile acids. They are produced to carry out a variety of functions, e.g., acting as a signaling molecule or facilitate absorption of dietary fats; however, their concentrations need to be strictly regulated as they can become toxic. The overall ▶ metabolic flux of the BAXS is primarily achieved through the activities of nuclear receptors, which have the ability to directly bind to DNA and regulate ▶ gene expression. Nuclear receptors can be thought of as metabolic sensors of exogenous and endogenous toxins. Detailed knowledge of the factors that govern the activity of nuclear receptors is required to understand a range of physiological processes, such as drug-drug interactions, intracrine hormone metabolism, ▶ xenobiotic clearance, and cholesterol homeostasis and lipid homeostasis.

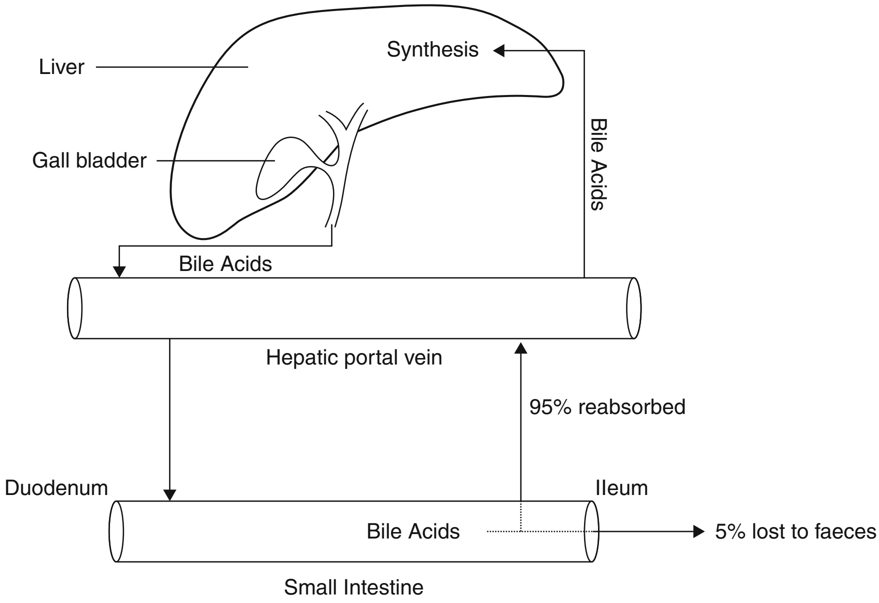
Characteristics

A System View

▶ Bile acids represent essential but also toxic biological reagents whose concentrations within the body require critical maintenance. Many of the genetic factors that dictate bile acid concentration also govern the detoxification and removal from the body of many drugs and foreign compounds. These overlapping biological processes define a network or system termed the bile acid and xenobiotic system which involves the coordinated activities of many genes in different tissues.

Bile acids are necessary for the emulsification and absorption of dietary fats and the regulation of cholesterol homeostasis. They are synthesized in the liver

Bile Acid and Xenobiotic System, Fig. 1 Schematic illustration of enterohepatic circulation, illustrating the circulation of biliary acids from the liver



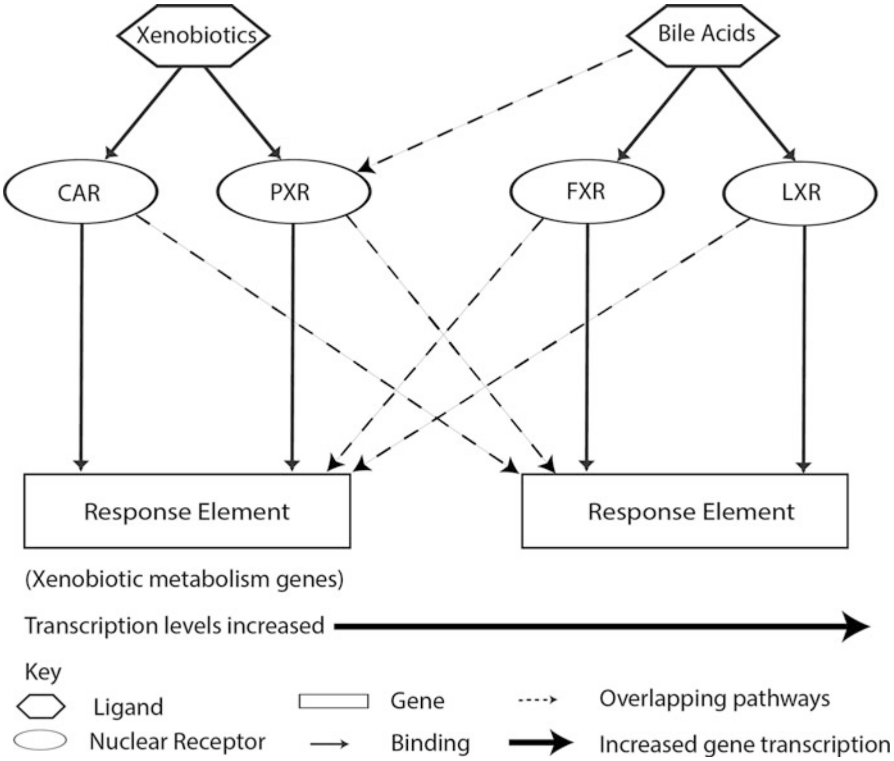
from the catabolism of cholesterol forming the primary bile acids, cholic and chenodeoxycholic acids. Bacterial flora dehydroxylates a portion of the primary bile acids in the intestinal lumen, resulting in secondary bile acids, deoxycholic and lithocholic acids. These four bile acids possess detergent-like properties necessary for the absorption of dietary lipids and fat-soluble vitamins. When present at high concentrations, they can become toxic; therefore, ▶ bile acid concentrations need to be appropriately regulated and recycled (Lefebvre et al. 2009).

Similarly, the BAXS detects any accumulation of xenobiotic and endobiotic compounds and facilitates their detoxification and removal from the body. This is accomplished through a complex network of sensors in the form of nuclear receptors that function as ligand-activated transcription factors (Kliewer and Willson 2002). The process of enterohepatic circulation (as depicted in Fig. 1) ensures 95% of bile acids are recycled.

The BAXS involves the coordinated activities of a number of genes across multiple temporal and spatial scales. Basic BAXS processes and their time scales include the binding of ▶ ligands to nuclear receptors (seconds), ▶ gene expression and ▶ gene regulation (hours), ▶ transporter protein activity (minutes), and metabolic enzyme activity (seconds). Spatially, BAXS components range from the dimension of molecules (e.g., nuclear receptors) to organs (e.g., liver). Given the multi-scale nature of the BAXS, it is difficult to assess the exact role of individual receptors and their activating/deactivating ▶ ligands with respect to the overall BAXS flux and its variation throughout a number of participating organs. A comprehensive description of the interacting components that govern BAXS ▶ gene expression would enable the identification of regulatory "nodes" as targets for treatment regimes, facilitate a deeper understanding of the components impacting drug-drug interactions, and provide a framework for the design of large-scale, integrated prediction studies.

Nuclear Receptors

Nuclear receptors are a superfamily of proteins which can bind directly to DNA and upregulate or downregulate the transcription of a gene. Within the BAXS, nuclear receptors act like a network of sensors detecting compounds such as hormones or xenobiotics. These compounds, referred to as ligands, bind to the nuclear receptors and can either activate them, leading to increased transcription, or deactivate them, leading to repression of gene transcription. A bound nuclear receptor may also dimerize with another nuclear receptor to form a complex which then binds to response elements located in the promoter region of the gene. This activates the gene and transcription is increased considerably. Nuclear receptors can also repress gene expression through competitive inhibition, whereby



Bile Acid and Xenobiotic System, Fig. 2 The role of nuclear receptors, illustrating the competition between nuclear receptors for the ligands they bind and the genes targeted

the nuclear receptor competes with other receptors for the ligands they bind, or by binding to the promoter region of the gene, thus reducing the efficacy of activation and therefore gene transcription. Nuclear receptors are classified as ▶ transcription factors and their combined inhibitory and activating effects regulate gene expression. This is vital in controlling development, metabolism and maintaining adult homeostasis (▶ Homeostasis).

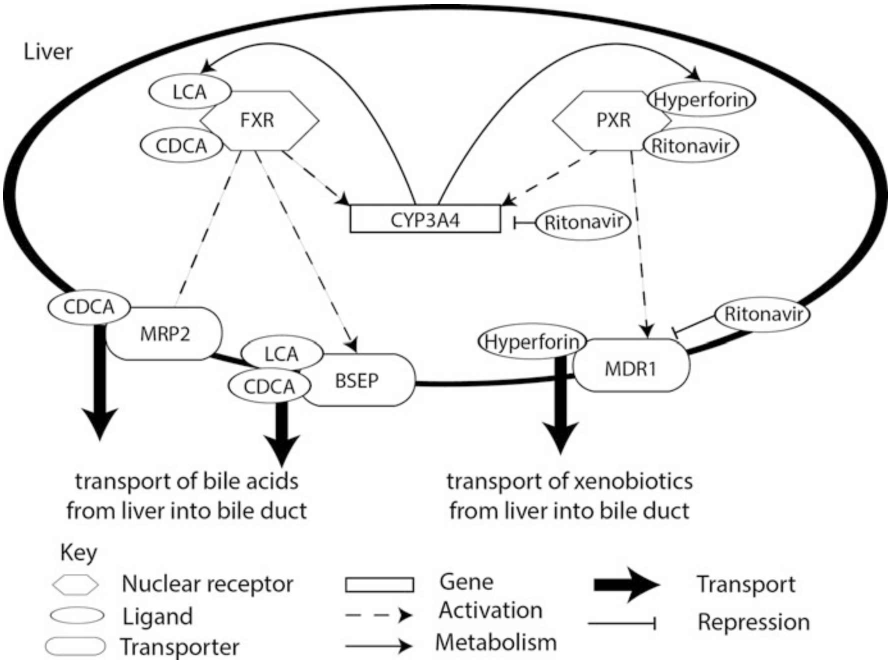
Nuclear receptors serve to detect fluctuations in concentration of many compounds and initiate a physiological response by regulating the BAXS. ▶ Transcriptional regulation by nuclear receptors involves both activating and repressive effects upon specific groups of genes. Figure 2 illustrates the overlap present between nuclear receptors and the genes they target and also the ligands that bind to and activate them. It is these factors that contribute to the phenomenon of drug-drug interactions, e.g., between St. John's wort and Cyclosporine (Barone et al. 2000), or St. John's wort and an oral contraceptive (Hall et al. 2003). Positive feed-forward and negative feedback loops can also occur, e.g., within the cholesterol ▶ metabolic pathway (Eloranta and Kullak-Ublick 2005).

The BAXS main process revolves around the circulation and critical maintenance of bile acid concentration and the detection and removal of harmful compounds. This process is composed of multiple subprocesses (e.g., transactivation, interactions with co-activators or corepressors, heterodimerization) operating on different time and space scales.

In the BAXS, the initial stimuli leading to a physiological response would be the binding of a ligand by a nuclear receptor. Subsequently, the bound nuclear receptor binds to response elements in the target genes, leading to increased ▶ gene expression and the cascading effects that would ensue. Further processes include conjugation and transporter functions (Stieger and Meier 1998).

Examples of nuclear receptors within the BAXS include the pregnane X receptor (PXR), farnesoid X receptor (FXR), constitutive androstane receptor (CAR), retinoid X receptor (RXR), and vitamin D receptor (VDR). A ligand can be either endogenous, e.g., a hormone or bile acid, or exogenous such as a drug, e.g., Rifampicin, St. John's wort, Dexamethasone. A ligand binds to the nuclear receptor which may also bind to another receptor of the same type to form

Bile Acid and Xenobiotic System, Fig. 3 Illustration of PXR-mediated metabolism of Hyperforin in the liver inhibited by Ritonavir, FXR-mediated bile acid metabolism, and the transport process



a homodimer (homodimerization) or with other nuclear receptor complexes to form a heterodimer (heterodimerization). The overall complex then binds to the ▶ promoter or ▶ enhancer of a specific gene and this either upregulates or downregulates expression of that gene depending on the complex formed.

BAXS Process and Example

The example in Fig. 3 shows the effects of Ritonavir (an antiretroviral drug from the protease inhibitor class used to treat HIV infection and AIDS) on the metabolism of Hyperforin (a phytochemical produced by some of the members of the plant, e.g., *Hypericum perforatum* which is commonly known as St John's wort) in the liver and the overlap of this process with FXR-mediated primary and secondary bile acid metabolism. Both Ritonavir and Hyperforin are activating ligands for PXR (Chang 2009; Willson and Kliewer 2002). CYP3A4 is a member of the cytochrome P450 superfamily of enzymes and one of the most important involved in xenobiotic metabolism. Multidrug resistance protein 1 (MDR1) is a member of the ATP-binding cassette (ABC) transporter superfamily of membrane-associated proteins responsible for transporting a wide variety of substrates from the cell (Koudriakova et al. 1998). Once PXR is activated through ligand binding, expression of CYP3A4 and MDR1 are considerably increased, resulting in the metabolism of Hyperforin and its transport from the cell into the intestinal lumen (Lin et al. 2009). In the absence of receptor binding, Ritonavir inhibits transcription of CYP3A4 and MDR1. In theory, this could lead to accumulated levels of Hyperforin in the liver as unbound Ritonavir inhibits the metabolism mechanism.

FXR is activated by primary and secondary bile acids, chenodeoxycholic acid (CDCA) and lithocholic acid (LCA). It upregulates transcription of CYP3A4, multidrug-resistant protein 2 (MRP2) and bile salt efflux pump (BSEP), both members of the ATP-binding cassette (ABC) transporter superfamily of membrane-associated proteins responsible for transporting bile acids into the bile duct. In both processes, an overlap occurs at CYP3A4. A patient taking Hyperforin will have increased expression of CYP3A4 which may lead to a deficiency in ▶ bile acid concentration as this gene metabolizes bile acids. Similarly, a patient with high bile acid concentrations may reduce the efficacy of Hyperforin (if taken) as transcription of CYP3A4 is increased. If Ritonavir is added, then bile acids and Hyperforin could accumulate to toxic levels in the liver.

Cross-References

▶ Bile Acid and Xenobiotic System

▶ Enhancer

▶ Gene Expression

▶ Gene Regulation

▶ Ligand

▶ Metabolic Flux Analysis

▶ Metabolic Networks, Databases

▶ Metabolic Pathway Analysis

▶ Promoter

▶ Transcription Factor

▶ Transcriptional Regulation

▶ Xenobiotics

References

Barone G, Gurley B, Ketel B, Lightfoot M, Abul-Ezz S (2000) Drug interaction between St. John's wort and cyclosporine. Ann Pharmacother 34(9):1013–1016

Chang T (2009) Activation of pregnane X receptor (PXR) and constitutive androstane receptor (CAR) by herbal medicines. AAPS J 11(3):590–601

Eloranta JJ, Kullak-Ublick GA (2005) Coordinate transcriptional regulation of bile acid homeostasis and drug metabolism. Arch Biochem Biophys 433(2):397–412

Hall SD, Wang Z, Huang S, Hamman MA, Vasavada N, Adigun AQ, Hilligoss JK, Miller M, Gorski JC (2003) The interaction between St John's wort and an oral contraceptive [ast]. Clin Pharmacol Ther 74(6):525–535

Kliewer SA, Willson TM (2002) Regulation of xenobiotic and bile acid metabolism by the nuclear pregnane X receptor. J Lipid Res 43(3):359–364

Koudriakova T, Iatsimirskaia E, Utkin I, Gangl E, Vouros P, Storozhuk E, Orza D, Marinina J, Gerber N (1998) Metabolism of the human immunodeficiency virus protease inhibitors Indinavir and Ritonavir by human intestinal microsomes and expressed cytochrome P4503A4/3A5: mechanism-based inactivation of cytochrome P4503A by Ritonavir. Drug Metab Dispos 26(6):552–561

Lefebvre P, Cariou B, Lien F, Kuipers F, Staels B (2009) Role of bile acids and bile acid receptors in metabolic regulation. Physiol Rev 89(1):147–191

Lin YS, Yasuda K, Assem M, Cline C, Barber J, Li C, Kholodovych V, Ai N, Chen JD, Welsh WJ, Ekins S, Schuetz EG (2009) The major human pregnane X receptor (PXR) splice variant, PXR.2, exhibits significantly diminished ligand-activated transcriptional regulation. Drug Metab Dispos 37(6):1295–1304

Stieger B, Meier PJ (1998) Bile acid and xenobiotic transporters in liver. Curr Opin Cell Biol 10(4):462–467

Willson TM, Kliewer SA (2002) PXR, CAR and drug metabolism. Nat Rev Drug Discov 1(4):259–266

Bile Acid System

▶ Bile Acid and Xenobiotic System

Bimodality

▶ Bistability

Binding Affinity

Xiujun Zhang
Institute of System Biology, Shanghai University, Shanghai, China

Definition

Binding affinity is a measure of the tendency or strength of interactions between molecules. The molecules that can bind together include proteins, DNA, antibodies, enzymes, and some other organic molecules such as drugs. The result of molecular binding is formation of a molecular complex such as protein-protein, protein-DNA, and protein-drug complex.

Binding affinity can be quantified and detected by some physical techniques (Karney et al. 2005). For example,

$$L + P \leftrightharpoons LP \qquad (1)$$

where ***L*** is a ligand, ***P*** is a protein, ***LP*** is the ligand-protein complex.

The concentration of the ligand-protein complex is given by the dissociation constant:

$$\mathrm{K}_d = \frac{[L][P]}{[LP]} \qquad (1)$$

The binding affinity is defined as:

$$pK_d = -\log_{10}\left(\frac{\frac{k_d}{N_A}}{1\ \mathrm{kmolm}^{-3}}\right) \qquad (2)$$

where $\boldsymbol{N_A}$ is the Avogadro constant.

References

Karney CF, Ferrara JE, Brunner S (2005) Method for computing protein binding affinity. J Comput Chem 26(3):243–251

Binomial Distribution Without Replacement

► Hypergeometric Distribution

Bioactivity

► Biological Activity

Bioassay

► Biological Assay

Biochart Diagram

► Modeling Formalisms, Lymphocyte Dynamics and Repertoires

Biochemical Kinetics

► Mass Action Stochastic Kinetics

Biochemical Network

► Metabolic Networks

Biochemical Pi-Calculus

► Stochastic pi-Calculus

Biochemical Systems Optimization Through Mathematical Programming

Julio Vera[1] and Néstor V. Torres[2]
[1]Department of Systems Biology and Bioinformatics, Institute of Computer Sciences, University of Rostock, Rostock, Germany
[2]Department of Biochemistry and Molecular Biology, University of La Laguna, San Cristóbal de La Laguna, Islas Canarias, Tenerife, Spain

Synonyms

Metabolic engineering

Definition

Mathematical models can be used to predict the effect of costly complex interventions over biochemical systems with technological or biomedical purposes. Toward this end, mathematical modeling can be combined with mathematical optimization techniques giving support to microbiologists and biomedical researchers in the biotechnological improvement of microorganisms with industrial applications ("what is the minimum set of enzymes that should be modified in order to maximize the production of a desired end product?") or in the design of new therapeutic approaches ("what is the minimum set of interactions in a biochemical network that must be inhibited by specific drugs to subvert a given pathological condition?") (Torres and Voit 2002; Vera et al. 2007).

The underlying idea is that a well-characterized and calibrated mathematical model, in ordinary differential equations, has predictive abilities that can be used to obtain optimal configurations of the biochemical system regarding a biotechnological (or biomedical) problem. This idea has been exploded by a number of groups in the past decade (Voit 1992; Torres et al. 1997; Hatzimanikatis et al. 1998). In a nutshell, simulations with the mathematical model are used to predict the behavior of the biochemical system. An optimization program, according to the biotechnological problem under investigation, is established including the following: (1) one or more objectives are established, describing in mathematical terms the properties of the system that are to be improved; (2) a set of constrains is established, accounting for physiological or technological limitations; and (3) a set of potential interventions on the system is established, which may include modulation of the system inputs, overexpression, or repression of enzymes, but also inhibition of given biochemical processes. Based on this program, optimization algorithms are used to estimate new configurations of the system that optimized the established objectives via the choice of the appropriate interventions, which should be experimentally tested.

In many cases, the methodology here discussed involves highly nonlinear optimization problems, difficult to solve in an efficient manner. Some of these difficulties are circumvented for steady-state optimization when canonical models like S-systems and power models are used. To illustrate the characteristics of this methodology and some of the advantages associated to the use of canonical models, we will focus here on the so-called indirect optimization method with S-systems models.

Characteristics

Linear Programming and the IOM Method

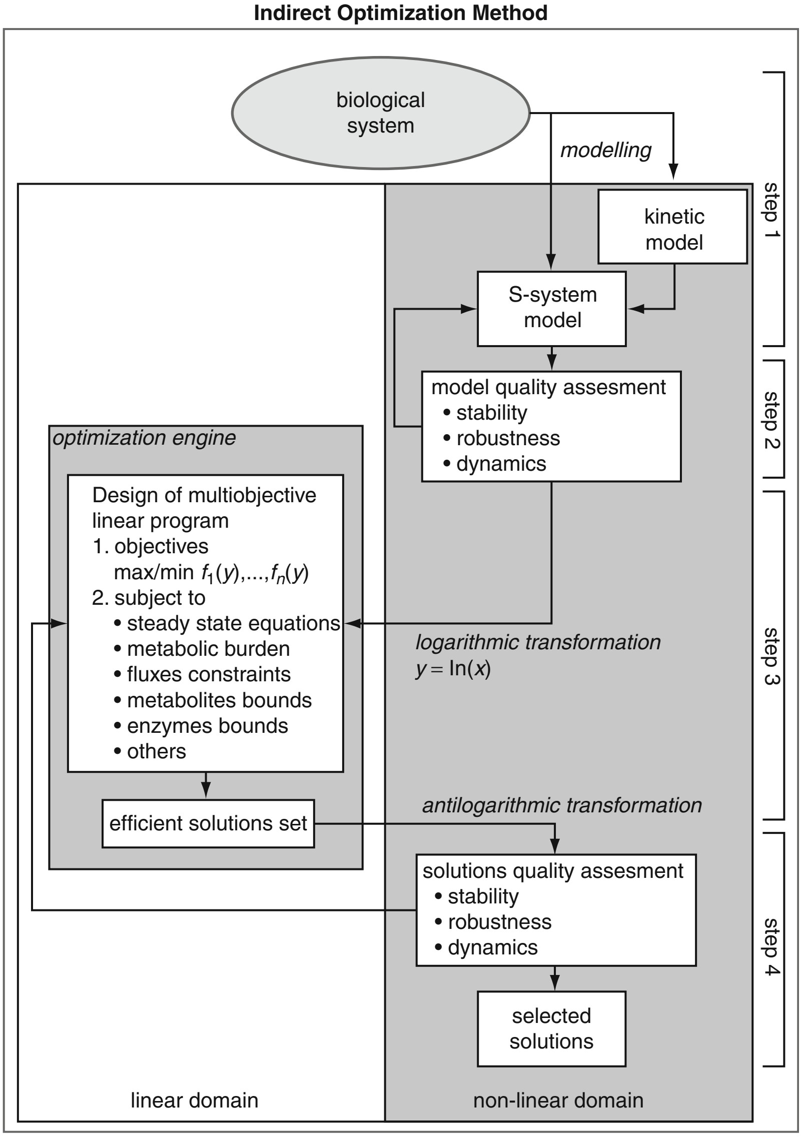
The Indirect Optimization Method (IOM) is based on the fact that although S-system models are nonlinear, their steady-state equations are linear when the variables are expressed in logarithmic coordinates. Since a function and its logarithm assume their maxima for the same argument, yields or fluxes can thus be optimized with linear programs expressed in terms of the logarithms of the original variables. Also, typical constraints that the optimized system has to satisfy reduce to linear equations in a logarithmic coordinate system. Thus, transporting the steady-state system and the constraints into a logarithmic space reduces the nonlinear optimization problem to a problem of straightforward linear programming (Voit 1992; Vera et al. 2003). The method is a procedure that systematizes the steps to follow for optimum results. Figure 1 illustrates the stages of the method in an outline for the most general case.

Step 1: *Getting the model*. The first step is the development of a model system. This can be modeled directly on the S-system formalism or in the form of kinetic model or *GMA*, and then transferred to S-system.

Step 2: *Analysis of quality of the model*. The S-system modeling formalism provides us with tools to analyze its stability, robustness, and dynamic response. It is important to ensure the validity of the model, i.e., that it properly describes the known behavior of the system. Otherwise, it is necessary to refine the model prior to its optimization.

Step 3: *Construction and resolution program optimization*. This is the core of the procedure. All information gathered on the system and the limitations of biological or technological constraints on the solutions are converted into the mathematical equations that make up the program optimization. After ensuring the validity of the model and the adequacy of the components of the program optimization, the stated objectives and the conditions imposed on the system are moved into logarithmic space. In the logarithmic space, the optimization problem becomes a linear program and once the optimal solution is obtained it is transferred to the area of the original variables. In addition to the above, the optimized system must be at a steady state and fulfill some constraints. These constraints, as well as the objective functions, are readily translated into linear functions and inequalities, and the optimization task becomes a linear program.

In the more general case (multiobjective optimization), after this point we can choose among three possible options (Vera et al. 2003, 2010): (1) pure multicriteria problem, which is the formulation of a multiple objective linear program, a direct extension of the linear programming case; (2) weighted sum approach, where we assign a weight to each objective or function to be optimized (usually these weights are chosen between zero and one, such that they sum up to one); and (3) goal programming, where a goal level of achievement is established for each objective. These goal levels are soft constraints that are included in the definitions of the objective functions. Available software packages can produce the efficient solution set for multiobjective linear programming tasks such as ADBASE or the Matlab optimization toolbox for the goal programming and weighted sum approach.



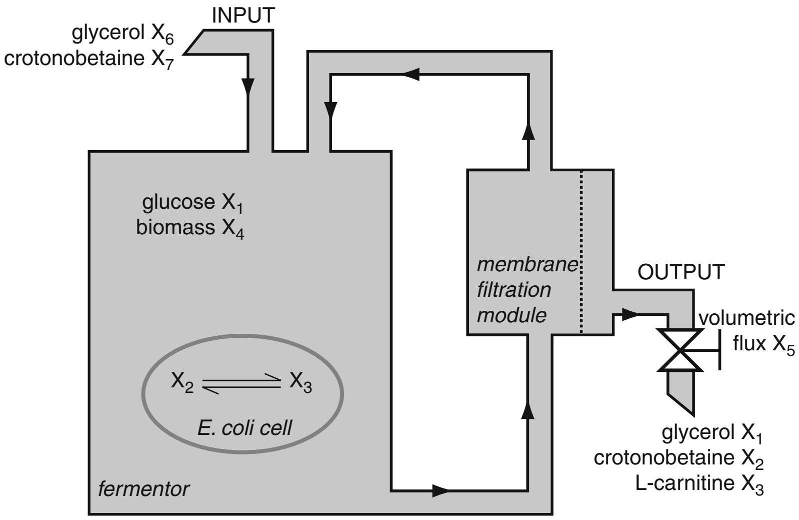
Biochemical Systems Optimization Through Mathematical Programming, Fig. 1 Sketch of the indirect optimization method

Step 4: *Analysis and refinement of the solution.* Here the system's properties at the optimal state (stability, robustness, and dynamics) are evaluated. If the optimal solution is sound, it is chosen as a solution; otherwise the imposed restrictions or the objective function are refined and the optimization process repeated until a satisfactory solution is reached. In the case where an S-system model was built from a previous kinetic model, the S-system solution must be transferred to the original model. In the following we will illustrate the application of this approach to two case studies.

Case Study 1. The Monoobjective Version of the IOM: Optimization of the L-(−)-Carnitine Biosynthesis by *Escherichia Coli*

L-(−)-carnitine is a chiral compound widely distributed in nature and with a wide range of medical applications. A great deal of the L-carnitine is produced by chemical synthesis, with the disadvantage of producing a racemic mixture that is necessary to separate in a costly process.

Figure 2 shows the experimental set up for the biotransformation of crotonobetaine into L-carnitine by an overproducing *E. coli* strain in a cell recycle



Biochemical Systems Optimization Through Mathematical Programming, Fig. 2 Experimental set up for the biotransformation of crotonobetaine into L-(−)-carnitine by *E. coli* strains

bioreactor. The model includes the metabolic transformation of crotonobetaine into γ-butyrobetaine, through the activity of a previously described crotonobetaine reductase. In the optimization procedure the dependent variables are X_1 to X_4, while X_5 to X_9 are the parameters. The original model set up was translated to the S-system representation yielding the following S-system representation (Alvarez-Vasquez et al. 2002):

$$\frac{dX_1}{dt} = X_5X_6 - 1.0213\, X_1^{0.9162}X_4^{0.5675}X_5^{0.4324}X_8^{0.5675}$$

$$\frac{dX_2}{dt} = 0.2438\, X_3^{0.3538}X_4^{0.9394}X_5^{0.0605}X_7^{0.0605}X_9^{0.9292} - 0.464\, X_2^{0.1424}X_4^{0.9643}X_5^{0.0356}X_9^{0.9537}$$

$$\frac{dX_3}{dt} = 0.3844\, X_2^{0.1106}X_4X_9^{0.989} - 0.2058\, X_3^{0.3926}X_4^{0.9742}X_5^{0.0257}X_9^{0.9636}$$

$$\frac{dX_4}{dt} = 0.0053\, X_1^{0.8524}X_4X_8 - 0.08\, X_4.$$

Subsequently the (mono)objective function is defined, namely, the net rate of L-carnitine biosynthesis, expressed as $V_c = X_3{\cdot}X_5$. In logarithmic coordinates this function becomes $Y_3 + Y_5$, where Y_i is $\ln(X_i)$. The steady-state condition, once linearized in logarithmic coordinates, obtained was:

$$\begin{aligned}
0.0211 &= -\,0.9162\,Y_1 - 0.5675\,Y_4 + 0.5676\,Y_5 \\
&\quad + Y_6 - 0.5675\,Y_8 \\
0.6433 &= -\,0.1424\,Y_2 + 0.3538\,Y_3 - 0.0249\,Y_4 \\
&\quad + 0.0249\,Y_5 + 0.0605\,Y_7 - 0.0245\,Y_9 \\
-0.6246 &= 0.1106\,Y_2 - 0.3926\,Y_3 + 0.0257\,Y_4 \\
&\quad - 0.0257\,Y_5 + 0.0254\,Y_9 \\
2.706 &= 0.8524\,Y_1 + Y_8
\end{aligned}$$

where Y_i is $\ln(X_i)$ for $i = 1,\ldots, 9$. At this point it is important to define the extent to which changes in the variables and parameters are allowed. Generally this variation range will be between 0.5 and 1.5 times the baseline values. An exception here is X_4, the biomass, which can be allowed to increase up to two times the baseline. Accordingly:

$$\begin{array}{ll}
3.0746 \le Y_1 \le 4.1732; & 2.6909 \le Y_2 \le 3.7895; \\
2.3277 \le Y_3 \le 3.4263; & 1.8068 \le Y_4 \le 3.1931; \\
-0.6931 \le Y_5 \le 0.4054; & 3.9120 \le Y_6 \le 5.0106; \\
3.2188 \le Y_7 \le 4.3174; & -1.1988 \le Y_8 \le -0.1002; \\
4.1214 \le Y_9 \le 6.424. &
\end{array}$$

With the defined linear program two optimization tasks were carried out: with one or two decision (independent) variables. The obtained optimal profiles were translated from the S-system to the original model (K-M) and then implemented in the bioreactor. Table 1 shows the experimental results obtained which are in good agreement with the theoretical predictions.

Biochemical Systems Optimization Through Mathematical Programming, Table 1 Comparison between predicted (K-M) and actual experimental values (Exp.) of L-(−)-carnitine production rate by *E. coli* in a cell recycle crotonobetaine biotransformation system. The optimized parameter profiles for changes in one (1; dilution rate, X_5) and two parameters (2; dilution rate and crotonobetaine concentration in the input, X_7) are shown. Results are given as the values divided by the basal

		1		2	
Variables	Basal	K-M	Exp.	K-M	Exp.
Dependent					
Glycerol, X_1	43.28 mM	1	0.96	1	0.95
Crotonobetaine, X_2	29.49 mM	1	0.97	1.49	1.77
Carnitine, X_3	20.51 mM	1	1.01	1.11	1.13
Biomass, X_4	12.18 g/L	1.5	1.53	1.5	1.53
Independent					
Dilution rate, X_5	1 L/h^{-1}	1.5		1.5	
Crotonobetaine input, X_7	50 mM	1		1.33	
Production rate, V_c	21.1 mM/h^{-1}	1.5	1.54	1.65	1.74

In both cases assayed, the experimental steady state was the same as that predicted by the model and the increase in the rate of L-(−)-carnitine production was almost coincident with the predicted increase.

Case Study 2. The Multiobjective Version of the IOM: Application to the Ethanol Production by *Saccharomyces Cerevisiae*

We have used a mathematical model of the ethanol production by *Saccharomyces cerevisiae* presented by Schlosser et al. (1994) as a case study to apply the multiobjective version of the IOM procedure.

As it can be seen in Fig. 3, glucose is converted in ethanol, polysaccharides, and glycerol. The model refers to the ethanol production under anaerobic, no growing conditions, with glucose as the sole carbon source and absence of nitrogen. The preliminary model by Schlosser et al. (1994) was converted into an S-system model in Vera et al. (2003) with the following structure:

$$\frac{dX_1}{dt} = 1.0006\ X_2^{-0.0492}\ X_6 - 1.6497\ X_1^{0.5582}\ X_5^{0.0465}\ X_7$$

$$\frac{dX_2}{dt} = 1.6497\ X_1^{0.5582}\ X_5^{0.0465}\ X_7 - 0.5793\ X_2^{0.5097}\ X_5^{-0.2218}\ X_8^{0.8322}\ X_{11}^{0.1678}$$

$$\frac{dX_3}{dt} = 0.4536\ X_2^{0.4407}\ X_5^{-0.2665}\ X_8 - 0.2456\ X_3^{0.4506}\ X_4^{0.0441}\ X_5^{0.092}\ X_9^{0.8547}\ X_{12}^{0.1453}$$

$$\frac{dX_4}{dt} = 0.2365\ X_3^{0.5285}\ X_5^{0.0994}\ X_9 - 2.0892\ X_3^{-0.0075}\ X_4^{0.304}\ X_5^{0.0484}\ X_{10}$$

$$\frac{dX_5}{dt} = 1.406\ X_3^{0.2605}\ X_4^{0.152}\ X_5^{0.0739}\ X_9^{0.5}\ X_{10}^{0.5} - 2.9437\ X_1^{0.1962}\ X_2^{0.1791}\ X_5^{0.2354}\ X_7^{0.3514}\ X_8^{0.2925}\ X_{11}^{0.0589}\ X_{13}^{0.297}.$$

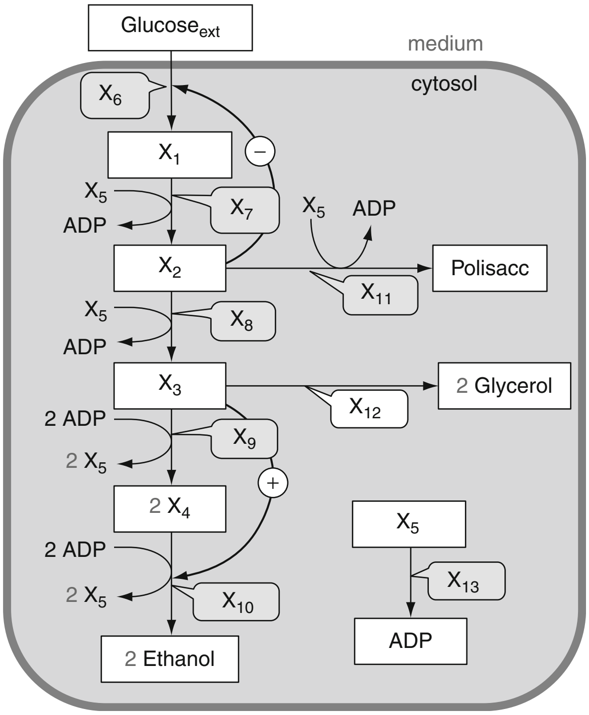
Three biotechnologically relevant optimization objectives are considered (Vera et al. 2003): (1) the maximization of the ethanol production, F_{prod}; (2) the minimization of the total internal metabolite concentrations, F_{int}; and (3) the minimization of the total enzyme activities, F_{enz}. The first objective relates to the enhancement of productivity, measured in terms of ethanol production. The other two objectives refer to the minimization of the amount of glucose transformed into other products different from ethanol and to the cell viability by reducing the osmolarity stress (objective 2) and the metabolic burden (objective 3).

Being $Y = \ln(X)$, they are defined as:

$$\min\ Z(Y)\ ,\qquad Z(Y) = \left(-F_{prod}(Y),\ F_{int}(Y),\ F_{enz}(Y)\right)$$

The maximization objective becomes a minimization objective by changing the sign of the objective function. These objective functions take the following form in the logarithmic transformed space:

$$F_{prod}(Y) = -0.0075\ Y_3 + 0.304\ Y_4 + 0.0483\ Y_5 + Y_{10},$$



Biochemical Systems Optimization Through Mathematical Programming, Fig. 3 Glucose fermentation pathway to ethanol, glycerol, and polysaccharides in *Saccharomyces cerevisiae*. Five dependent concentrations and eight fluxes are involved: Intracellular Glucose (X_1); Glucose-6-phosphate (X_2); Fructose-1,6-bisphosphate (X_3); Phosphoenol-pyruvate (X_4); ATP (X_5); Polysaccharides; Glycerol; Ethanol (X_{16}), and ADP. Enzymes/pathway steps: V$_{\text{in}}$, Glucose uptake (X_6); V$_{\text{HK}}$, Hexokinase (X_7); V$_{\text{PFK}}$, Phosphofructokinase (X_8); V$_{\text{GAPD}}$, Glyceraldehyde 3-phosphate dehydrogenase (X_9); V$_{\text{PK}}$, Pyruvate kinase (X_{10}); V$_{\text{POL}}$, Total disaccharide and polysaccharide storage (X_{11}); V$_{\text{GOL}}$, Glycerol production (X_{12}); V$_{\text{ATPase}}$ (X_{13}), Generalized rate for all ATP-utilizing process, with the exception of V$_{\text{HK}}$, V$_{\text{PFK}}$, and V$_{\text{POL}}$

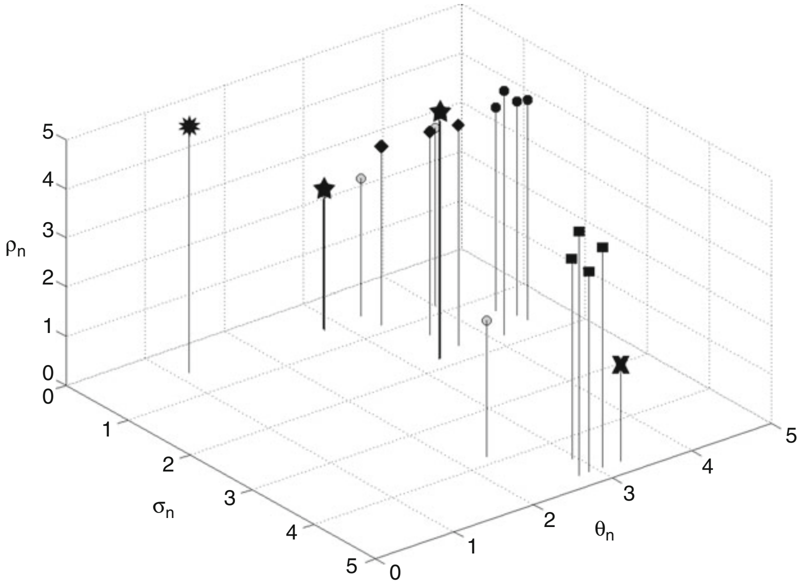
$$F_{int}(Y) = \sum_{i=1}^{5} Y_i \quad \text{and} \quad F_{enz}(Y) = \sum_{i=6}^{13} Y_i.$$

In the optimization program three types of constraints are considered (Vera et al. 2003): (a) those that guarantee that solutions correspond to a steady-state solution of the system; (b) those setting the physiologically feasible lower and upper boundaries for the model variables; and (c) additional constraints related to special features of the considered biosystem. The constraints that set up the boundaries for the model variables (metabolite or enzyme concentrations) configure the properties of the solutions; their values are dictated by the experience and are formulated for each variable. Experience dictates (Alvarez-Vasquez et al. 2000) that a reasonable lower boundary is the 50% of the original basal steady-state value and five times this basal value for most of the variables $\left(0.5X_i^0 < X_i < 5X_i^0\right)$ for the upper boundary. The exception to this rule were the ATPase activity $\left(X_5 < 1.5X_5^0\right)$, limited by additional physiological constrains, and the side-products polysaccharides and glycerol, which were maintained at their basal values $\left(X_{11} = X_{11}^0; X_{13} = X_{13}^0\right)$. A constrain, limiting the value of the flux ratio through the enzymes X_9 and X_{10} in order to prevent dynamical instability in the solutions ($X_9/X_{10} \leq 1.75$) was introduced. A last constraint to ensure that any computed solution doubles at least the ethanol production in the original unmodified microorganism, $V_{PK}(X) \geq 2\, V_{PK}(X^0)$, was also imposed.

After computing the multiobjective program by using ADBASE, 22 efficient solutions were obtained. Only the efficient vertexes of the problem were considered. The analysis of the solutions showed that solutions with a high ethanol production are only possible when the maximum of ATP ($X_5 = 1.5$) and ATPase ($X_{13} = 5$) are provided to the system. This indicates the extreme importance of the ATP turnover to increase productivity. To facilitate the analysis and classification of the generated solutions, some auxiliary biologically relevant parameters were defined as follows:

$$\sigma(X) = \frac{\sum_{i=1}^{5} X_i}{\sum_{i=1}^{5} X_i^0}, \quad \theta(X) = \frac{\sum_{i=6}^{13} X_i}{\sum_{i=6}^{13} X_i^0}, \quad \rho(X) = \frac{V_{PK}(X)}{V_{PK}(X^0)}.$$

σ is the ratio between the total intermediate concentration of the current solution and the one in the original basal solution of the system; θ is the ratio between the total enzyme activities at the optimum solution and the ones at the basal steady state, while ρ describes the same ratio for the ethanol production of the system. In order to support the biotechnologists making the proper decision, additional criterions were introduced. We ruled out all solutions with the lowest admissible ethanol production ($\rho^{\min} = 2.0$) and ranked the remaining ones according with their parametric distance to the so-called utopian point, a fictitious



Biochemical Systems Optimization Through Mathematical Programming, Fig. 4 Efficient solutions for the multicriteria optimization of ethanol production by *S. cerevisiae*. (●): Set I, solutions with the largest D(X,X^{utp}) value. (♦): Set II, solutions closer to the utopian point (star). (■) Set III, solutions with intermediate values of D(X,X^{utp}). (**X**) represents the anti-ideal point

solution constructed with the optimal values of ethanol production (ρ^{utp}), total intermediate concentration (σ^{utp}), and total enzyme concentration (θ^{utp}) from the computation of separated monoobjective programs:

$$\rho^{utp} = 4.986 \quad \sigma^{utp} = 0.5 \quad \theta^{utp} = 1.0$$

In a similar way, we defined the anti-utopian solution, extracting the worst value from every column in the payment matrix for the separated monoobjective programs (Vera et al. 2003):

$$\rho^{fun} = 2.0 \quad \sigma^{fun} = 4.76 \quad \theta^{fun} = 3.30$$

With this information we defined a weighted parametric distance from every solution to the utopian point, which is described by the following equation:

$$D(X,X^{utp}) = \sqrt{\frac{1}{3}\left[\left(\frac{\rho^{utp}-\rho(X)}{\rho^{utp}-\rho^{fun}}\right)^2 + \left(\frac{\sigma^{utp}-\sigma(X)}{\sigma^{utp}-\sigma^{fun}}\right)^2 + \left(\frac{\theta^{utp}-\theta(X)}{\theta^{utp}-\theta^{fun}}\right)^2\right]}$$

Figure 4 shows the solutions grouped in three sets according with their D(X,X^{utp}) value. Set I corresponds to those solutions with the largest D(X,X^{utp}) value. This group is characterized by a high productivity (ρ), but also by high total intermediate concentration (σ), an undesirable property. Set II includes the closer solutions to the utopian point D(X,X^{utp}), characterized by high productivity and enzyme levels but low metabolite concentrations (0.8–1.85). Finally, Set III represents those with intermediate values of D(X,X^{utp}), high productivity and enzyme levels but medium values of total metabolites concentration. When the parametric distance is considered the criterion to choose the best solution two solutions were found (five-pointed star in Fig. 4: Table 2).

These solutions represent opposite alternatives from a biotechnological perspective. Solution I prefers the increase in the ethanol production at the price of high cell resources consumption, while solution II establishes a biotechnological compromise between satisfactory increasing in the ethanol production and non-excessive cell resources consumption.

It should be noted that the heuristic nature of the last step in the IOM method may provoke a loss of efficiency in the solutions when the behavior of the original model departs from their S-system expansion. This may result in the violation of some of the restrictions imposed. However, experience shows that these violations are often negligible and the steady-state solutions usually stay in the vicinity of the real efficient solution. Moreover, the method incorporates tools to estimate solutions closer to the optimum in the nonlinear domain. This latter step consumes considerable computational effort, but even more important, the solution obtained ceases to be optimal. The practical application of the method indicates however that in most cases it remains a high-quality solution.

Biochemical Systems Optimization Through Mathematical Programming, Table 2 Best two efficient solutions for the multicriteria optimization of ethanol production by *S. cerevisiae*

Sol	ρ	σ	θ	D(X,X^{utp})
I	4.987	1.812	3.301	0.604
II	2.791	0.596	2.798	0.613

Case Study 3. Combining Mathematical Modeling and Optimization to Detect Potential Drug Targets in Human Diseases

The methodology discussed here can be used to detect potential drug targets in diseases that are originated by dysfunctions of biochemical networks. The strategy is to point out one or more enzymes in the biochemical network, whose modulation via specific drugs permits to redirect some biochemical fluxes in the network. In this manner, the critical fluxes and metabolites for the system are restored to values similar to the ones found in healthy subjects (Vera et al. 2007). The method requires the derivation of a mathematical model describing the dynamics of the metabolic network under analysis. It is also necessary to retrieve biomedical knowledge required to select the critical fluxes and metabolites unbalanced in the disease condition, as well as their values in healthy subjects. Model optimization is used to determine which biochemical processes in the network must be modulated, via drug-mediated inhibition or activation, in order to move the system from the current values of critical metabolites and fluxes toward the healthy ones. The simulation-derived therapeutic treatments may consider the modulation of one reaction in the network at a time, but also suggest strategies to develop multifactorial treatments. The optimization program suggested contains at least the following elements.

*Objective***.** An objective function, which translated in mathematical terms the minimization of the distance between the values of the critical metabolites and fluxes in the current systems state with respect to those in the desired health state:

$$Min\left[\sum_{j=1}^{l}\lambda_j\left|\frac{X_j-X_j^{HS}}{X_j^{HS}}\right|+\sum_{i=1}^{P}\lambda_i\left|\frac{J_i-J_i^{HS}}{J_i^{HS}}\right|\right]$$

where X_j and X_j^{HS} denote the current and the health values of the metabolites and J_i and J_i^{HS} are the corresponding values for the fluxes, respectively. The values assigned to the weights λ_j and λ_l are proportional to the relative importance of each key metabolite and flux. Every term in the equation is scaled by its value in the healthy condition such that all contributions have comparable weights.

Dysfunction description. A mathematical model description of the functional origin of the disease, which is obtained by imposing a characteristic value to the enzyme activities whose deregulation originates the pathology:

$$X_j = X_j^{PS}$$

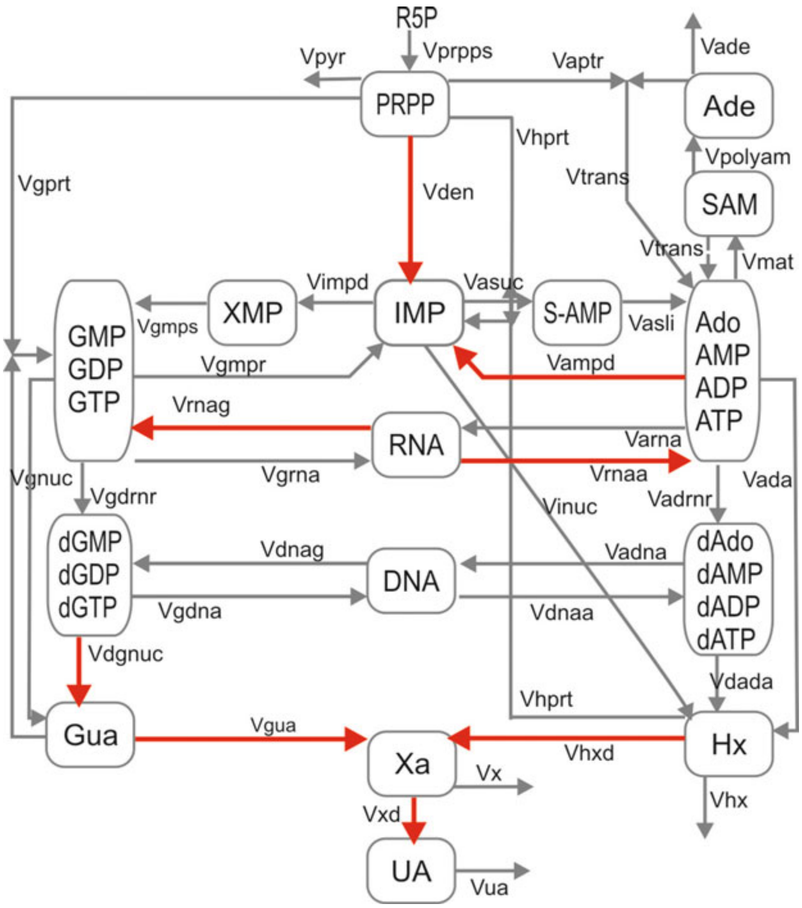
Physiological constrains. Some additional equations grating that the biochemical network configuration obtained is physiologically acceptable. These equations can be steady-state conditions and upper and lower bounds in the concentrations of the metabolites, respectively:

$$\frac{dX_i}{dt}=0,\quad i=1,\ldots,n.$$

$$X_i^{LB}\leq X_i\leq X_i^{UB}$$

The solutions generated by the method are sets of computationally predicted values for metabolites and substrates, as well as the required modulation level of the targeted enzyme. The selected solutions are those for which the computed values of critical metabolites and fluxes are sufficiently close to the ones in the healthy condition without compromising the values of other metabolites. In Vera et al. (2007) this methodology was used to identify potential enzyme drug targets for hyperuricemia, a disease associated with a defect in phosphoribosyl pyrophosphate synthetase, an enzyme that regulates the de novo metabolic synthesis of purines. The final pathological effect of this deregulation is an abnormal level of uric acid, which triggers arthritic pain and nephropathy. We used a power-law mathematical model of purine metabolism (Curto et al. 1997) and defined a mathematical optimization program with the structure described above. We defined as critical metabolite the uric acid and computed solutions of the optimization program consisting in the inhibition of enzymes and the potential application of a diet with low levels of purine

Biochemical Systems Optimization Through Mathematical Programming, Fig. 5 Sketch of the mathematical model for purine metabolism used in our analysis (see Vera et al. 2007 for complete description). In red we highlight the metabolic fluxes whose inhibition in combination with a diet low in purine precursors reduces in our simulations the levels of uric acid. Precisely, those solutions propose the drug-mediated inhibition of one of the following enzymes: Adenine phosphoribosyltransferase (V_{den}), AMP deaminase (V_{ampd}), RNases to AMP and GMP (V_{rgna}, V_{rnaa}), 5′ (3′)-Nucleotidase (V_{dgnuc}), Guanine hydrolase (V_{gua}), or xanthine oxidase (V_{xd}, V_{hxd}).



precursors, as well as combinatorial treatments consisting in the parallel inhibition of two enzymes. With the method we detected up to six potential single enzyme targets, including an analogous of the conventional clinical treatment using the drug allopurinol, but also two other totally unexpected potential drug targets. When considering potential multifactorial treatments, numerous possible solutions were detected (Fig. 5).

Cross-References

- Flux Balance Analysis
- Parameter Estimation, Metabolic Network Modeling
- Power-Law Functions
- S-System
- Synthetic Biology, Predictability and Reliability

References

Alvarez-Vasquez F, González-Alcón CM, Torres NV (2000) Metabolism of citric acid production by Aspergillus niger. Model definition, steady state analysis, dynamic behavior and constrained optimization of citric acid production rate. Biotechnol Bioeng 70(1):82–108

Alvarez-Vasquez F, Cánovas M, Iborra JL, Torres NV (2002) Modeling, optimization and experimental assessment of continuous L-(−)-carnitine production by *Escherichia coli* cultures. Biotechnol Bioeng 80(7):794–805

Curto R, Voit EO, Sorribas A, Cascante M (1997) Validation and steady-state analysis of a power-law model of purine metabolism in man. Biochem J 324(Pt 3):761–775

Hatzimanikatis V, Emmerling M, Sauer U, Bailey JE (1998) Application of mathematical tools for metabolic design of microbial ethanol production. Biotechnol Bioeng 58(2–3):154–161 Review

Schlosser PM, Riedy G, Bailey J (1994) Ethanol production in baker's yeast: application of experimental perturbation techniques for model development and resultant changes in flux control analysis. Biotechnol Prog 10(2):141–154

Torres NV, Voit EO (2002) Pathway analysis and optimization in metabolic engineering. Cambridge University Press, Cambridge

Torres NV, Rodríguez F, González-Alcón C, Voit EO (1997) An indirect optimization method for biochemical systems: description of the method and application to ethanol, glycerol and carbohydrates production in *Saccharomyces cerevisiae*. Biotechnol Bioeng 55(5):758–772

Vera J, De Atauri P, Cascante M, Torres NV (2003) Multicriteria optimization of biochemical systems by linear programming. application to the ethanol production by *Saccharomyces Cerevisiae*. Biotechnol Bioeng 83(3):335–343

Vera J, Curto R, Cascante M, Torres NV (2007) Detection of potential enzyme targets by metabolic modelling and optimization. Application to a simple enzymopathy. Bioinform 23(17):2281–2289

Vera J, González-Alcón C, Marín-Sanguino A, Torres N (2010) Optimization of biochemical systems through mathematical programming: methods and applications. Comput Operat Res 37(8):1427–1438

Voit EO (1992) Optimization in integrated biochemical systems. Biotechnol Bioeng 40:572–582

Biochemical Systems Theory (BST)

Eberhard O. Voit
The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Definition

BST is a fully dynamic modeling framework, based on ordinary differential equations, in which all processes are represented with products of power-law functions. This representation leads directly to specific rules and guidelines for the design, diagnostics, analysis, and application of models in various fields of biology. BST permits several variants, among which *Generalized Mass Action* (GMA) systems (▶ Generalized Mass Action System) and *S-systems* (▶ S-System) are most important. Models within BST are called *canonical* (▶ Canonical Model) in reference to their strict structure and strong modeling guidelines.

Cross-References

▶ Dynamic Metabolic Flux Analysis

BioCreative II.5 and the FEBS Letters Experiment on Structured Digital Abstracts

Florian Leitner, Martin Krallinger and Valencia Alfonso
Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

Definition

BioCreative is a community challenge to evaluate applied systems in biomedical text-mining. In the context of the third installment of this challenge, BioCreative II.5, the feasibility of using automated text-mining systems for protein–protein interaction (PPI) database curation was evaluated. In parallel, FEBS Letters asked manuscript authors to annotate their manuscripts with protein interaction information. The author information then was further utilized by MINT PPI curators to compare the impact of basing their work on the author data against their performance when starting from scratch. The BioCreative organizers evaluated the performance of the curators, authors, and automated systems individually. Then, curator annotations based on author data, as well as combined annotations from all text mining systems, and combining author and automated systems' data was compared to those individual results. Finally, the annotation overlap between curators, authors, and the best performing automated system was measured to establish to what extent each of the three sources would be complementary to the others.

Introduction

The Structured Digital Abstract (SDA) initiative (Ceol et al. 2008) is an ongoing effort by FEBS Journal and FEBS Letters to add annotations describing protein–protein interactions (PPIs) that are reported with experimental verification to publications. BioCreative (Blaschke et al. 2003; Krallinger et al. 2008) is a community evaluation of text-mining systems where the BioCreative organizers ask participants to perform biologically relevant tasks and then evaluate the submitted results on a blind test set. In the context of BioCreative II.5, the third installment of this challenge, the organizers asked participants to automatically generate parts of the SDAs from FEBS Letters publications (Leitner 2010a). In total, 15 research groups worldwide followed this call and participated in BioCreative II.5. The aim was to provide a quantitative insight on how well automated systems can reproduce human annotations. Additionally, annotations were provided by the authors of the papers themselves as well as by expert bio-curators of the MINT PPI database (Ceol et al. 2010; Chatr-aryamontri et al. 2007) in the context of the associated FEBS Letters experiment. All these results were evaluated then by the BioCreative organizers (Leitner 2010b). Together with the author and curator annotations it was possible to compare the results of the automated systems to the manual annotations.

The SDA initiative and BioCreative II.5 are the first quantitative approach to directly compare the impact of generating annotations for scientific manuscripts by various plausible approaches, namely, database curators (Howe et al. 2008), manuscript authors, and automated text-mining systems. At current funding levels, PPI databases can only keep up with a fraction of the data that is published by the biosciences. This leaves the larger part of generated scientific data “dormant” in written text: This data is neither trivial to retrieve or locate by researchers looking for a particular information, nor is this format accessible for large-scale data manipulation as would be required by Systems Biology or other data mining approaches in Computational Biology. Therefore, BioCreative II.5 intended to investigate the feasibility of pursuing new avenues to increase the coverage of data deposited in structured repositories as a result of scientific publications, using PPIs as the exemplary data in this setting. For example, in the realm of PPIs, the united consortium of all major PPI databases, IMEx, only manages to cover an estimated 10–20% of the existing interactions, limiting the reconstruction of interactive graphs for protein interaction maps to a fraction of the existing data and making it difficult for biologists to determine all known, proven interaction partners with their proteins of interest.

Methods

The most crucial part of this effort was to produce a high-quality dataset with near-perfect annotations that could be used as common basis for the evaluation, a so-called gold standard. To this end, article annotations that had initially been made by authors were refined by three independent MINT database bio-curators, who continued refining the results until they reached a consensus annotation for those articles that matches the MIMIx standard for annotating PPIs (Orchard et al. 2007). In a second round of pruning – after the BioCreative participants had submitted their results – the BioCreative organizers reevaluated the annotations with the results of the automated systems and identified potentially erroneous annotations by examining results where the automated systems unanimously reported annotations that were inconsistent with the curator data; Requesting the bio-curators counsel on the found inconsistencies and making corrections where necessary, this third step produced the final “gold standard” annotations for the evaluation.

The applied performance measures are intended to reflect (a) the overall quality of the annotations made by any of the sources (i.e., systems, authors, and curators) and (b) the performance of the automated systems when ranking their results by a probability score scheme, so-called “confidence values” (i.e., probability values in the (0,1] range) that participants were asked to report together with their annotations. This latter evaluation method was used as the primary evaluation target, as it measures the systems’ ability to produce result lists that can be used by human annotators as initial dataset to base their own annotation effort on. Therefore, two statistical measures were used:

(a) The harmonic F-measure (aka. F1-score) for the overall quality of the result sets (both human and system annotations)
(b) The area under the interpolated precision/recall curve (aka. AUC iP/R) for the quality of automated results given the ranking

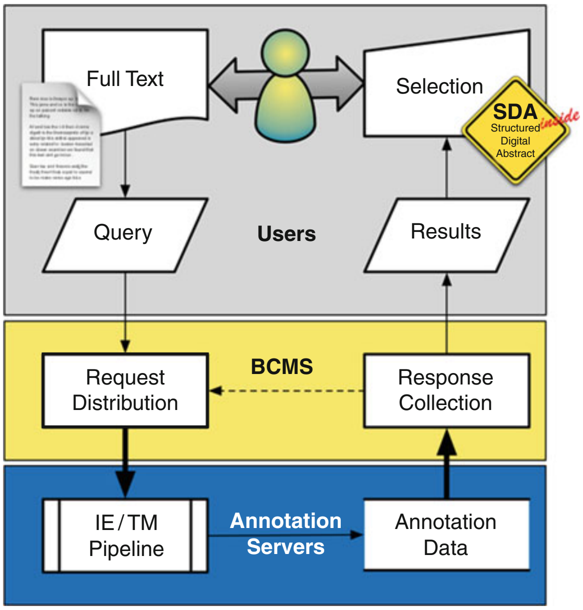
The competition itself was carried out online using an extended version of the BioCreative Meta-Server (BCMS), a special framework that interacts via web services with the automated systems provided by the challenge’ participants (Leitner et al. 2008) (also see the entry ▶ BioCreative Meta-Server and Text-Mining Interoperability Standard in this encyclopedia). Contrary to the initial version, the extended BCMS allows participants to take direct control of and monitor the communication between their systems and the BCMS. This tool set the whole challenge in an environment that simulated the applied use case in which these annotations would be generated online and provided to a human annotator (see Fig. 1). Furthermore, this also made it possible to time the systems and factor that variable into the evaluation's conclusions.

Finally, in an effort to bring together all sides of this effort, namely, the publishers, the curators, and the scientists, a workshop was held in Madrid (Spain), in 2009, to present the results and discuss their implications for both the efforts in text-mining research, the feasibility of adding human annotation efforts with automated systems, and the outlook of possible avenues to add annotations to publications in general.

Results

Detailed results are found in the relevant publications reported across three journals – Nature Biotechnolgy, FEBS Journal, and IEEE's Transactions on Computational Biology and Bioinformatics (TCBB); all (Leitner 2010a, b, c). They describe the outcome and conclusions in meticulous detail and also spurred research of independent organizations – the challenge's participants – resulting in nine additional research articles that formed part of the TCBB special issue that contains the main BC II.5 article.

The texts this effort is based on, 122 PPI publications, together with more than 1,000 negative example articles (i.e., articles that explicitly do not contain PPI descriptions with experimental evidence) from the same period and journal (FEBS Journal 2008 and 2009), were provided by the FEBS Journal. With the friendly permission of Elsevier, they are now accessible as a free and open resource for further research projects in text mining (Leitner 2010c). This resource is available through the BioCreative homepage (www.biocreative.org), as a so-called corpus. This alone presents an important achievement, as text resources with high-quality annotations are – mostly due to copyright and other legal issues – scarce and very hard to come by; Furthermore, this resource is in XML format, which is simpler to read for machines as opposed to the usual PDF format scientific manuscripts are made available.



BioCreative II.5 and the FEBS Letters Experiment on Structured Digital Abstracts, Fig. 1 The simulated online setting of BioCreative II.5 reproduced by the BioCreative Meta-Server. The gray box "Users" represents the hypothetical human annotator creating a SDA, while the "BCMS" and "Annotation Servers" form the technical framework created for the challenge

The main comparison of the results of both the FEBS experiment and the BioCreative II.5 challenge dealing with each source of annotations (authors, curators, and text-mining systems) has been published in Nature Biotechnology; this is also the lead article of this joint effort. The comparison focused on the correct annotation of protein [database] identifiers and of the [binary] interaction pairs found in the SDAs, which were the main two tasks for the automated systems in BioCreative II.5. The main conclusions from this work are:

(a) At least the two FEBS magazines (FEBS Journal and Letters), Cell (their graphical abstracts), and the ScienceDirect annotations (provided by NeXT Bio) – but potentially other publishers, too – are very interested in adding annotations to their manuscripts.
(b) Although authors and systems might perform reasonably well, none of the two produced results that are sufficient for database standards.
(c) The time requirement of systems is significantly lower than that of manual annotations

BioCreative II.5 and the FEBS Letters Experiment on Structured Digital Abstracts, Table 1 Individual evaluation results of the three sources and the combination of authors and curators (*left*) and results of individual text-mining systems, including their performance when taking their ranking into account (*AUC iP/R, right*)

Task	Class	Precision (%)	Recall (%)	F-Score	Task	Class	Precision (%)	Recall (%)	F-Score	AUC iP/R
Protein identifiers	Systems	74	55	0.59	Protein identifiers	Best F-score (T42, S1)	74	55	0.59	0.53
	Authors	84	66	0.71						
	Curators	96	89	0.91		Best AUC iP/R (T10, R5)	14	73	0.21	0.57
	Authors + curators	96	94	0.95						
Interaction pairs	Systems	53	34	0.37	Interaction pairs	Best F-score and AUC iP/R	53	34	0.37	0.31
	Authors	72	57	0.59						
	Curators	93	83	0.86		Incl. MINT data (T18, R1)	64	61	0.58	0.58
	Authors + curators	93	89	0.90						

(averaging at 2 min/articles, vs. almost 1 h for the human annotations) and the quality of systems seems to be sufficient to at least aid human annotators.

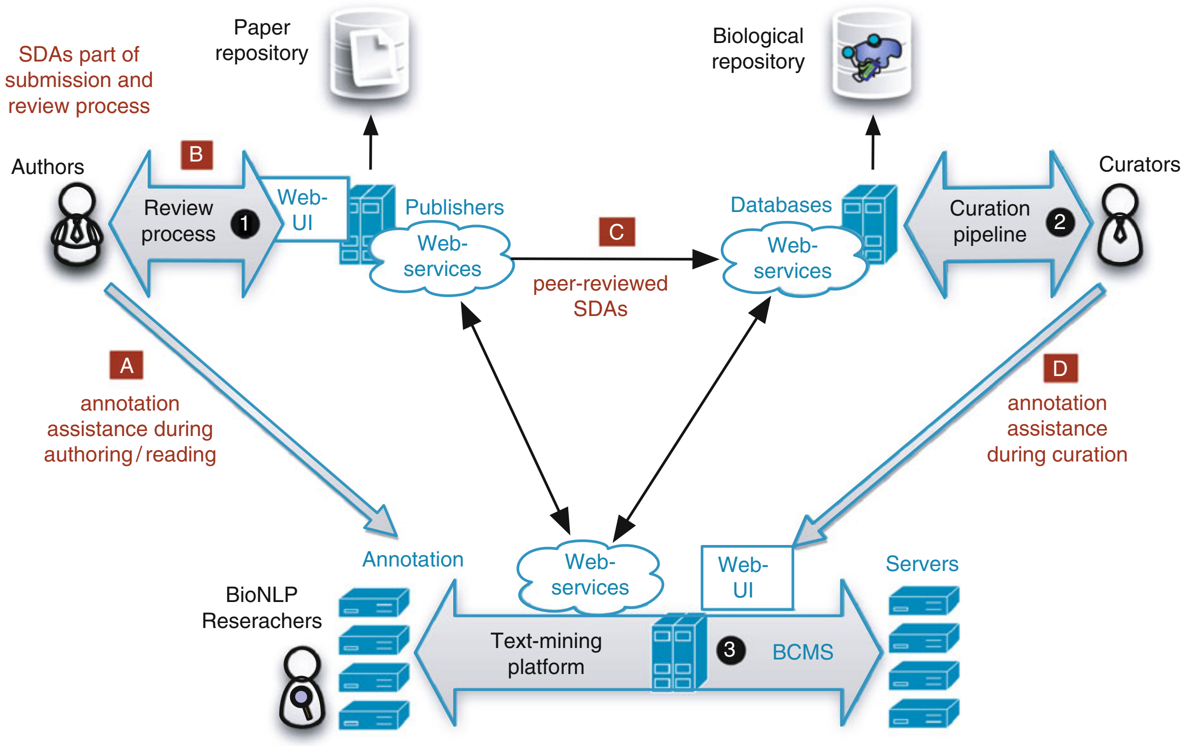
(d) Systems and authors, although producing lower quality results than curators, were able to identify annotations the curators missed, and in the evaluation it became apparent that when one source based its annotations on another, the quality of the resulting annotations was higher. For example, curators basing their annotations on author results performed better than if the curators started annotating from scratch. A similar correlation is shown in annotations in the publications between systems and authors.

Table 1 shows the evaluation results of the three sources as well as the performance of curators when basing their annotations on author data (*left*), and the detailed results of the best submissions of text-mining systems from the BioCreative II.5 participants (*right*).

Last but not least, a detailed analysis of the core text-mining part, that is, BioCreative II.5 itself, that includes a comparison of the systems and applying various statistical approaches to the 134 result sets produced by the 15 participant teams is part of the third publication in Transactions on Computational Biology and Bioinformatics. This publication documents the evaluation metrics (F1-Score, AUC iP/R), the online setting of the challenge with the BioCreative Meta-Server, and shows that combining the results of the text-mining systems with the author annotations would produce better results than the two approaches by themselves did.

Perspectives

With the results of the BioCreative II.5 challenge and the insight created by evaluating it in comparison with the FEBS Letters SDA experiment, it is clear that neither curators (databases), nor authors, or text-mining systems can be tasked with creating the needed annotations on their own. Our proposal is to combine the approaches (as combining sources at least increased performance in terms of quality and it is likely that adding automated systems will decrease annotation times) and distribute the effort between all participants. A theoretical framework for combining human and machine annotations is presented in Leitner and Valencia (2008). A potential outline of this architecture is shown in Fig. 2 and explained in deep detail in the supplementary material of the Nature Biotechnology publication. Therefore, the further development of the BioCreative Meta-Server from a demonstration server to a production-ready public service representing the main resource provided by the text-mining community is a major objective, and this also encompasses the generation of an international standard for annotating scientific texts, which will increase the interoperability of text-mining systems and make it easier to interface with tools humans use to annotate the articles.



BioCreative II.5 and the FEBS Letters Experiment on Structured Digital Abstracts, Fig. 2 The target architecture to integrate authors and text-mining systems in the process of extracting annotations from scientific manuscripts (see the chapter ▶ BioCreative Meta-Server and Text-Mining Interoperability Standard in this book, too)

Cross-References

▶ BioCreative Meta-Server and Text-Mining Interoperability Standard

References

Blaschke C, Hirschman L, Yeh A, Valencia A (2003) Critical assessment of information extraction systems in biology. Comp Funct Genomics 4:674–677

Ceol A, Chatr-aryamontri A, Licata L, Cesareni G (2008) Linking entries in protein interaction database to structured text: the FEBS Letters experiment. FEBS Lett 582:1171–1177

Ceol A, Chatr-aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G (2010) MINT, the Molecular INTeraction database: 2009 update. Nucleic Acids Res 38:D532–D539

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res 35: D572–D574

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, Rhee SY (2008) Big data: the future of biocuration. Nature 455:47–50

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, Valencia A (2008) A text-mining perspective on the requirements for electronically annotated abstracts. FEBS Lett 582(8):1178–1181

Leitner F et al (2008) Introducing meta-services for biomedical information extraction. Genome Biol 9(suppl 2):S6

Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman L, Valencia A (2010a) An overview of BioCreative II.5. IEEE/ACM Trans Comput Biol Bioinform 7(3):385–399

Leitner F, Chatr-aryamontri A, Mardis SA, Ceol A, Krallinger M, Licata L, Hirschman L, Cesareni G, Valencia A (2010b) The FEBS Letters/BioCreative II.5 experiment: making biological information accessible. Nat Biotechnol 28(9):897–899

Leitner F, Krallinger M, Cesareni G, Valencia A (2010c) The FEBS Letters SDA corpus: a collection of protein interaction articles with high quality annotations for the BioCreative II.5 online challenge and the text mining community. FEBS Lett 584(19):4129–4130

Orchard S et al (2007) The minimum information required for reporting a molecular interaction experiment (MIMIx). Nat Biotechnol 25:894–898

BioCreative Meta-Server and Text-Mining Interoperability Standard

Florian Leitner, Martin Krallinger and
Valencia Alfonso
Structural Biology and BioComputing Programme,
Spanish National Cancer Research Centre (CNIO),
Madrid, Spain

Definition

Over the past decade, text-mining systems have matured to serve both bioinformaticians and biologists in a variety of ways. However, this also has the negative impact that a plethora of protocols, sites, and tools exist that are not compatible between each other. With the BioCreative Meta-Server (BCMS) the first prototype system to counterbalance this development was published. Because of uniting several text-mining systems into one service using one protocol, both access and use of those resources could be significantly simplified. In addition, the BCMS provides a straightforward means to combine results from multiple text-mining systems, contributing an added benefit that can lead to an even higher quality result than the individual annotations of the systems available through the BCMS. To push the capabilities of text-mining systems even further, a project was recently launched to create a text-mining community consensus on an interoperability standard for the annotation of texts. This project should facilitate the use of annotation systems and annotated texts for both text miners and consumers of text-mining services.

Characteristics

Over the past decade, many text-mining and information extraction systems for biomedical texts have been developed (Krallinger et al. 2008a). Some of them have matured to industrial-standard quality sites that are frequented by many users, such as iHOP (Hoffmann and Valencia 2004) or Reflect (Pafilis et al. 2009). Others have developed powerful web service APIs (application programming interfaces), as, for example, WhatIzIt (Rebholz-Schuhmann et al. 2008), or the NaCTeM Service Systems (Kano et al. 2010). Many more are available as individual applications that can be downloaded and installed. However, all this wealth of available software, tools, services, and sites does not come without cost. Due to no real community agreement on annotation standards, document formats, and the semantics of the generated data, it is painful to interface between tools, especially if they are not from the same organization. Also, the power of even the best systems can be still outstripped by a consensus result created from many different systems that all provide the same kind of annotations (e.g., Smith et al. 2008). In addition, to use one tool, it often is necessary to rely on results of another. For example, before identifying gene names in an article, such a gene name “tagger” can significantly gain performance if its annotations are based on the results of a pipeline that first tags the “part-of-speech” (PoS) of the words; PoS taggers annotate the grammatical sense of words, such as verb, noun, adjective, etc.. However, while it might be easy to find high-quality tools for one specific task, another area often can be lacking in terms of the number of systems that are available. Many tools only come with very basic command-line interfaces or are system libraries that only computer experts know how to make use of. This creates a very high entry barrier for many potential users. Even worse, in a few cases research groups decide to not make their tools publicly available; Yet, they might agree to make their pipeline available as a web service, so that users can gain access to them but nobody can gain access to the source code the system relies on.

All these issues led to a community understanding that the current status quo is ripe for improvement. In the area of sequence annotation and structure prediction, these bioinformaticians have years ago already begun to build distributed systems and meta-services. The idea is fairly simple: For distributed systems, instead of having many different protocols and formats, a community agrees to certain standards and then builds their tools to those specifications (nowadays, this design principle is termed "service-oriented architecture"). The earliest example of such as system is (Bio)DAS, a distributed sequence annotation system (Dowell et al. 2001). On the other hand, as combining results leads to often superior consensus predictions, structural biology researchers at the same time begun to create so-called meta-servers (Bujnicki et al. 2001). In this case, the user contacts a central "service broker" or "meta-server" with the protein sequence for which he wishes to receive a structure prediction. The meta-server, in turn, instead of calculating a protein structure on its own, forwards this request to many different prediction servers via web services. Then the broker waits for the results to return form those prediction servers, collecting each server's result. Once all results are in, the meta-server creates a superimposed consensus prediction from the individual the results, which usually tends to be a better estimate of the true protein structure than any single result. Then the user is informed that his result is ready and he can fetch the structure and accompanying information directly from the meta-server.

Both of these approaches would also solve plenty of the issues mentioned with current approaches in text-mining: The (Bio)DAS approach shows how to reduce interoperability problems by defining a standard for the data exchange. This makes it easy to chain the outputs of one tool to another and to join or merge data from several sources. The general web service approach allows organizations to make the fruits of their labor available to the public without having to worry about loosing control over who has access to their source code. By using the web, a user also does not need to know anything about command-line interaction or programming and can run sequence annotation tools and structure prediction services directly from his browser in a graphical environment he is familiar with. Last, meta-servers reap the fruits of consensus results, providing the user with possibly improved results.

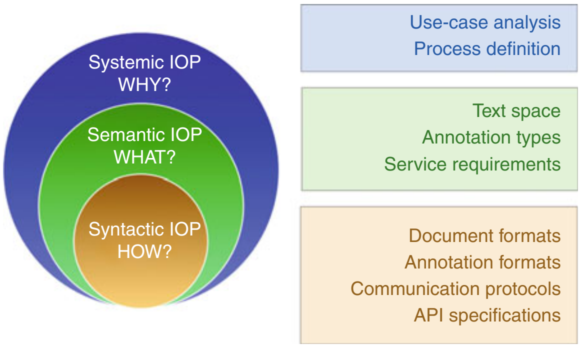
All this insight created the motivational basis to develop the BioCreative Meta-Server (BCMS) (Leitner et al. 2008). This first prototype was directly created after BioCreative II, a text-mining community challenge where information extraction systems are evaluated by an independent group of judges to measure the performance of those systems on some topic in the field of molecular biology (Krallinger et al. 2008b). This version of the BCMS became the first attempt to provide true "meta-services" for text-mining to the scientific community. In essence, it is similar to a structure prediction meta-server, but instead of adapting the interface to every possible text-mining system, it defined a standard exchange format that any system plugged into the BCMS had to follow. Thereby, it inherently has advantages similar to the ones of the BioDAS system – namely, a uniform protocol and data model. The BCMS prototype requests and manages annotations of four basic types and works on PubMed abstracts:

1. Classifying the abstract as to whether it describes protein–protein interactions
2. Identifying the mentions of genes and proteins in the abstract
3. Listing of possible mappings to UniProt accessions for the mentions
4. Listing the most likely organisms the abstract is discussing

However, the current, publicly available version of the BCMS is a prototype. It only offers these annotations in a very limited scope: This particular service broker is limited to the approximately 22,000 PubMed abstracts used during the BioCreative II challenge. Therefore, only queries for that data set can be made and are distributed to the connected servers. In other words, it is not possible to enter any new text to annotate, request annotations other than the above mentioned four base types, or even just query for any other PubMed abstract.

This prototype project, however, enabled the participants and developers of the servers to glean some very important insights into the endeavor of setting up a distributed and interoperable annotation system for biomedical texts to and providing a public, automated text-mining platform. These problems can be related to the three layers of interoperability (see Fig. 1): First, there are low-layer syntactic issues that need appropriate solutions, such as:

- Concurrency and thread management on all servers; a fair queueing policy of requests on the broker/meta-server side
- Fail-safe and nonrestrictive handling of the various possible representations of text in computers (so-called encoding schemes) on all participating servers
- Compatibility issues with web service implementations provided by different platforms and programming languages ("service agnosticism")



BioCreative Meta-Server and Text-Mining Interoperability Standard, Fig. 1 The three layers of interoperability (IOP). The *syntactic layer* relates to technical problems such as specifications and implementation details. The *semantic layer* treats issues related to data content and meaning, and interfaces. Finally, the *systemic layer* determines the use-case and process requirements and outlines the strategy for solving issues on the two inner layers

These are just a few of the many technical details that need attention. Then there are concerns that are related to the content and meaning of the annotation data, such as:

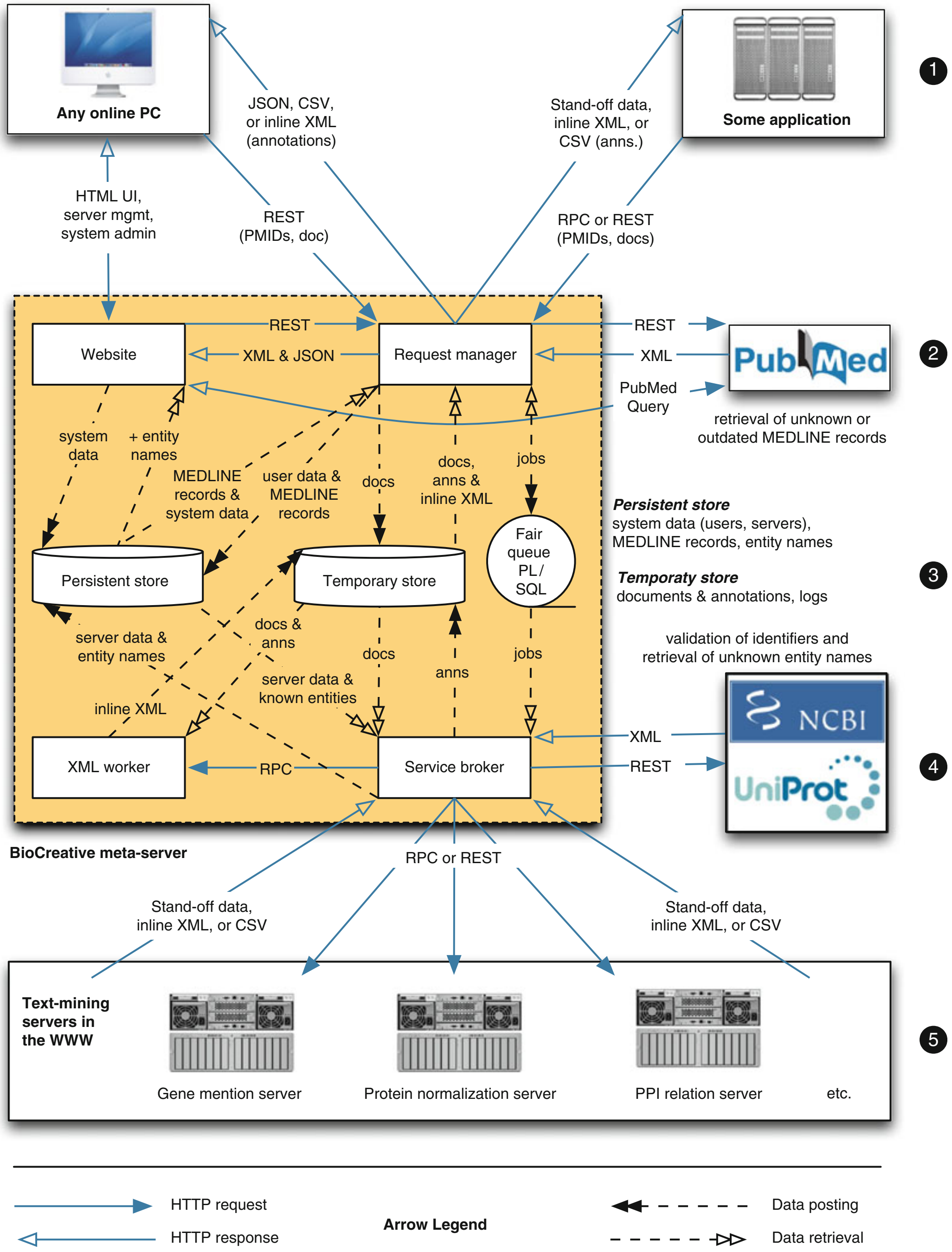
- What kind of attribute values are permitted; for example, that reporting a probability value annotation of zero for an annotation is a violation or how to report the position of a mention inside a text: either by counting the number of characters or the number of bytes up to that mention, which is additionally obfuscated by the variable multibyte character nature of many encoding schemas, or via inline annotations such as XML tags that come with distinct and possibly larger set of disadvantages for interoperability
- Which databases to use for the mappings to and how (e.g., which databases can be used for a gene annotation or a protein, and, e.g., for UniProt, whether to map to entry names or to the UniProt accessions; or how to treat obsolete database identifiers; etc.)

Are the most prominent issues to agree upon at the semantic layer. Finally, there are high-level, syntactic problems that have to be considered, such as:

- Establishing a consensus annotation, runtime requirements of the servers (if they take too long, users would get annoyed, but if servers tune the text processing too much, annotation quality suffers)
- The impact of copyright issues (e.g., only granting access to a text to which the user has access, especially for publications, once the BCMS allows to annotate any text resource)
- Designing an interactive process with the users, for example, by allowing the user to provide initial annotations when sending a request, such as the organism(s) the text treats, which would significantly reduce the possible number of protein IDs the systems need to concern themselves when generating ID mappings (Leitner and Valencia 2008)

From these insights into issues that have to be addressed for an "ideal" distributed text annotation system, and by designing solutions for these problems, the next stage of the BCMS is now under development (Fig. 2). The goal will be to provide a framework that will allow users to request annotations for text in many of the common document formats (e.g., plaintext, HTML, XML, Word, or PDF files), permitting the use of any encoding schema to represent those texts, having a universal entity annotation schema that is not limited to a predefined set of types, avoiding a lock-in into one specific service protocol by providing the most common ones "out of the box" (i.e., JSON-RPC, XML-RPC, SOAP, and REST), providing the ability to query for and annotate any PubMed abstract, interactively allowing both human and machine annotations, etc. An important milestone on this path is to establish a community agreement on a biomedical text-mining annotation interoperability standard (currently, termed "BAIS" and accessible via web at http://bais.bioinfo.cnio.es/) that will provide the essential guidelines for any information extraction tool, even for systems that do not necessarily have to be coupled with the BCMS platform.

At the current state (October 2010), all the development is by and large still in very early development stages. Hopefully, a fully functional BCMS will replace the prototype version that is currently found at http://bcms.bioinfo.cnio.es/, running at the Spanish National Cancer Research Center in Madrid in the near future. Also, the interoperability standard (BAIS) should have matured by that time and currently resides at http://bais.bioinfo.cnio.es/.



BioCreative Meta-Server and Text-Mining Interoperability Standard, Fig. 2 The current design proposal for the final BCMS. Essentially, there are five units, from *top* to *bottom*: (**1**) Clients accessing the platform, either from a browser or programmatically via a web service; (**2**) The front end that provides the web site and services, accepts documents, queues jobs, returns annotations, and communicates with PubMed; (**3**) The storage layer that is responsible for the data management and persistency; (**4**) The back end that processes inline annotations, manages communication with the text-mining servers, and validates database identifiers, adding supplementary information (e.g., the protein or gene names) to the annotations; (**5**) And the text-mining servers, providing the actual annotations of any type agreed upon by the community (i.e., BAIS) via web services again. *Blue arrows* symbolize communication channels over the Internet, *black dashed arrows* outline internal communication pathways for the BCMS components

Cross-References

▶ BioCreative II.5 and the FEBS Letters Experiment on Structured Digital Abstracts

References

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001) Structure prediction meta server. Bioinformatics 17(8):750–751

Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L (2001) The distributed annotation system. BMC Bioinform 2:7

Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nat Genet 36(7):664

Kano Y, Dobson P, Nakanishi M, Tsujii J, Ananiadou S (2010) Text mining meets workflow: linking U-compare with Taverna. Bioinformatics 26(19):2486–2487

Krallinger M, Valencia A, Hirschman L (2008a) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol 9(Suppl 2):S8

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

Leitner F, Valencia A (2008) A text-mining perspective on the requirements for electronically annotated abstracts. FEBS Lett 582(8):1178–1181

Leitner F et al (2008) Introducing meta-services for biomedical information extraction. Genome Biol 9(Suppl 2):S6

Pafilis E, O'Donoghue SI, Jensen LJ (2009) Reflect: augmented browsing for the life scientist. Nat Biotechnol 27(6):508–510

Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A (2008) Text processing through Web services: calling Whatizit. Bioinformatics (Oxford, England) 24(2):296–298

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

Bioinformatics

▶ Disease System, Malaria

Biological Activity

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

Synonyms

Bioactivity; Pharmacological activity

Definition

Biological activity is "the capacity of a specific molecular entity to achieve a defined biological effect" on a target (Jackson et al. 2007). It is measured in terms of potency or the concentration of the molecular entity needed to produce the effect (Pelikan 2004). A biological activity is determined by means of a biological assay.

Cross-References

▶ Biological Assay
▶ Drug Target
▶ Natural Product Resources

References

Jackson MJ, Esnouf MP, Winzor D, Duewer D (2007) Defining and measuring biological activity: applying the principles of metrology. Accredit Qual Assur 12(6):283–294

Pelikan EW (2004) Glossary of terms and symbols used in pharmacology. University School of Medicine, Department of Pharmacology and Experimental Therapeutics, Boston. http://www.bumc.bu.edu/busm-pm/resources/glossary

Biological Applications of Network Modules

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

Definition

A protein-protein interaction (PPI) network of a certain organism is a network that contains proteins and physical interactions between protein pairs. In the network, each node/vertex is a protein and each edge/link is a physical interaction.

An interactome refers to a complete PPI network that contains all protein physical interactions in a certain organism. An interactome is usually approximated by combining all known PPI data from large- and small-scale experiments, expert curations, and possibly computational predictions.

A general greedy search is any algorithm or method that makes a locally optimal decision in order to retrieve or approximate a globally optimal solution. A greedy search method often contains a candidate set, a selection criterion, a feasibility criterion, an objective function, and a solution criterion. The candidate set contains all candidates that can be added to the solution. The selection criterion or function determines the best candidate to add next into the solution. The feasibility criterion determines whether a candidate can be added to contribute for the solution. The objective function assigns and calculates a heuristic score for a solution or partial solution. The solution criterion indicates whether the final solution is achieved.

Characteristics

Network modules identified by network clustering are widely suspected to correspond to ▶ functional modules and complexes. Therefore, module identification is directly applied for functional module discovery. The evaluation of identified modules mainly includes the comparison with known protein complexes and pathways, as well as ▶ functional enrichment analysis in modules. Protein complexes data of yeast can be retrieved from Munich Information Center for Protein Sequences (MIPS) database (Mewes et al. 1999), and Kyoto Encyclopedia of Genes and Genomes (▶ KEGG PATHWAY) is the most comprehensive database for pathway-related information (Kanehisa and Goto 2000). On one hand, identified modules are likely biologically meaningful if the proteins/genes in the module overlap with certain protein complexes or pathways. On the other hand, network modules are likely functional modules if many of identified modules have overrepresented functions by their protein members.

Further applications of identified network modules, especially in ▶ protein-protein interaction (PPI) networks, mainly include protein function prediction by assigning overrepresented functions in a module to protein members with unknown functions, as well as ▶ network-based biomarker discovery for disease status or outcome (Chuang et al. 2007; Zhang et al. 2010).

Protein Function Prediction Based on Modular Analysis

One application of module identification is to predict functions for proteins in modules with unknown functional annotations. Such protein function predictions are based on identified modules as well as known functional annotations, such as ▶ gene ontology or *Saccharomyces* Genome Database annotations (Ashburner et al. 2000; Cherry et al. 1998).

A straightforward method to predict protein functions is to assign a function shared by the majority of proteins/genes in a module to other unannotated protein/gene in the same module, also considering each of the assigned functions as a function of the whole module. In this way, proteins with unknown functions are assigned these shared functions within the same module. Alternatively, a hypergeometric test (or one-tailed ▶ Fisher's test) can be performed for each function to calculate a p value indicating whether or not a function is overrepresented by genes/proteins in a module:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{j}{i}\binom{n-j}{m-i}}{\binom{n}{m}},$$

where n is the number of nodes in the whole network, j is the number of nodes in the network annotated with the function, and m is the module size. The p value indicates the probability of observing at least k proteins from a module or cluster of size m by chance to have the tested function, and small p values imply possible enrichment of certain functions in modules. Given a threshold for the p value, the significantly overrepresented functions are then predicted for all proteins/genes in the module and assigned to proteins with unknown functions. For example, suppose a module with 20 nodes is identified from a curated human ▶ interactome network from Human Protein Reference Database (HPRD; the 9th release) (Keshava Prasad et al. 2009), where each node is a human protein and each edge indicates a physical interaction between two proteins, and 18 out of the 20 proteins have ▶ gene ontology annotations. If 16 out of the 18 annotated proteins all have a specific function, for example GO:0006950 (response to stress), while among other 9,462 human proteins in the interactome, 1,924 proteins are annotated with this function, then this function is significantly overrepresented by proteins in this module, with p value = 1.7e-8 by a ▶ Fisher's exact test. Considering that the module is very likely to be involved in a function related to response to stress, the other four proteins including the two proteins with unknown functions can be predicted or assigned to have this function as well.

Biomarker Discovery Based on Modular Analysis

Another application of modular analysis is the discovery of ▶ network-based modular biomarkers for disease status or outcome. Such discovery is performed by combining network modular analysis with disease-related genomic data and often involves several steps of scoring and search. Specifically, overlaying genes from gene expression data of two different phenotypes (e.g., disease versus normal control, or cancer metastatic samples versus primary tumor samples) onto molecular networks, for example PPI networks, will result in a set of connected components, or subnetworks with disease-related genes. For each sample from one of the two phenotypes, an activity score can be calculated for each subnetwork by averaging the normalized gene expression levels of all genes in the subnetwork. Correlation between the subnetwork gene activity score and the phenotype can be assessed by the ▶ mutual information measure. Such ▶ mutual information between genes in a subnetwork and the resulting phenotype can be optimized based on a greedy search method by seeding with one gene and growing the subnetwork by adding connected genes based on the PPI network. Finally, a set of subnetworks that can best classify the two phenotype outcomes are output as ▶ network-based modular biomarkers for disease status or outcome.

Cross-References

- ▶ Biomarkers
- ▶ Fisher's Test
- ▶ Functional Enrichment Analysis
- ▶ Functional Modules and Complexes
- ▶ Functional/Signature Network Module for Target Pathway/Gene Discovery
- ▶ KEGG Pathway Database
- ▶ Modules in Networks, Algorithms and Methods
- ▶ Mutual Information
- ▶ Network-based Biomarkers
- ▶ Ontology
- ▶ Ontology Analysis of Biological Networks

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS et al (1998) SGD: Saccharomyces genome database. Nucleic Acids Res 26(1):73–79

Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S et al (2009) Human protein reference database–2009 update. Nucleic Acids Res 37(Database issue): D767–D772

Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F et al (1999) MIPS: a database for genomes and protein sequences. Nucleic Acids Res 27(1):44–48

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

Biological Assay

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

Synonyms

Assay; Bioassay

Definition

A biological assay is an experiment that determines a substance's biological activity based on its effect on a specific drug target, relative to that of a standard preparation (IUPAC 1997). It is also the process by which the potency of an agent is measured in terms of the reactions of a specific drug target (Pelikan 2004).

Cross-References

▶ Biological Activity
▶ Drug Target
▶ Natural Product Resources

References

IUPAC (1997) Compendium of chemical terminology. In: McNaught AC, Wilkinson A (eds) The "gold book", 2nd edn. Blackwell Scientific, Oxford

Pelikan EW (2004) Glossary of terms and symbols used in pharmacology. University School of Medicine, Department of Pharmacology and Experimental Therapeutics. Boston. http://www.bumc.bu.edu/busm-pm/resources/glossary

Biological Clock

▶ Circadian Rhythm

Biological Disease Mechanism Networks

Shipra Agrawal[1] and M. R. Satyanarayana Rao[2]
[1]BioCOS Life Sciences Pvt. Limited, Institute of Bioinformatics and Applied Biotechnology, Bangalore, Karnataka, India
[2]Chromatin Biology Laboratory, Molecular Biology and Genetics Unit, Jawaharlal Nehru Center for Advanced Scientific Research, Bangalore, Karnataka, India

Synonyms

Diseasome; Gene network; Intercatome; Protein network; Scale-free network; Transcriptional regulatory network

Definition

The biological networks are constructed from any type of molecular/biochemical and biological data to interpret the overall functional and regulatory schema of any biological system (cell/tissue/organ and organism). These networks are represented through wiring diagram of nodes (genes/proteins) and edges (directed/undirected connections) and classified into different subtypes.

In other words, the complex biological processes are presented through the precise interaction and regulation of 1,000 of molecules. These biological networks are significantly different from any random networks and often exhibit ubiquitous properties in terms of their structure and organization.

Biological networks are developed and analyzed to understand the pathophysiology of different human diseases. Such networks also help in identifying novel biomarkers, candidate genes, pathway cross talk pattern, etc., leading to the progression/molecular characterization of complex diseases, i.e., type 2 Diabetes and Cancers.

Characteristics

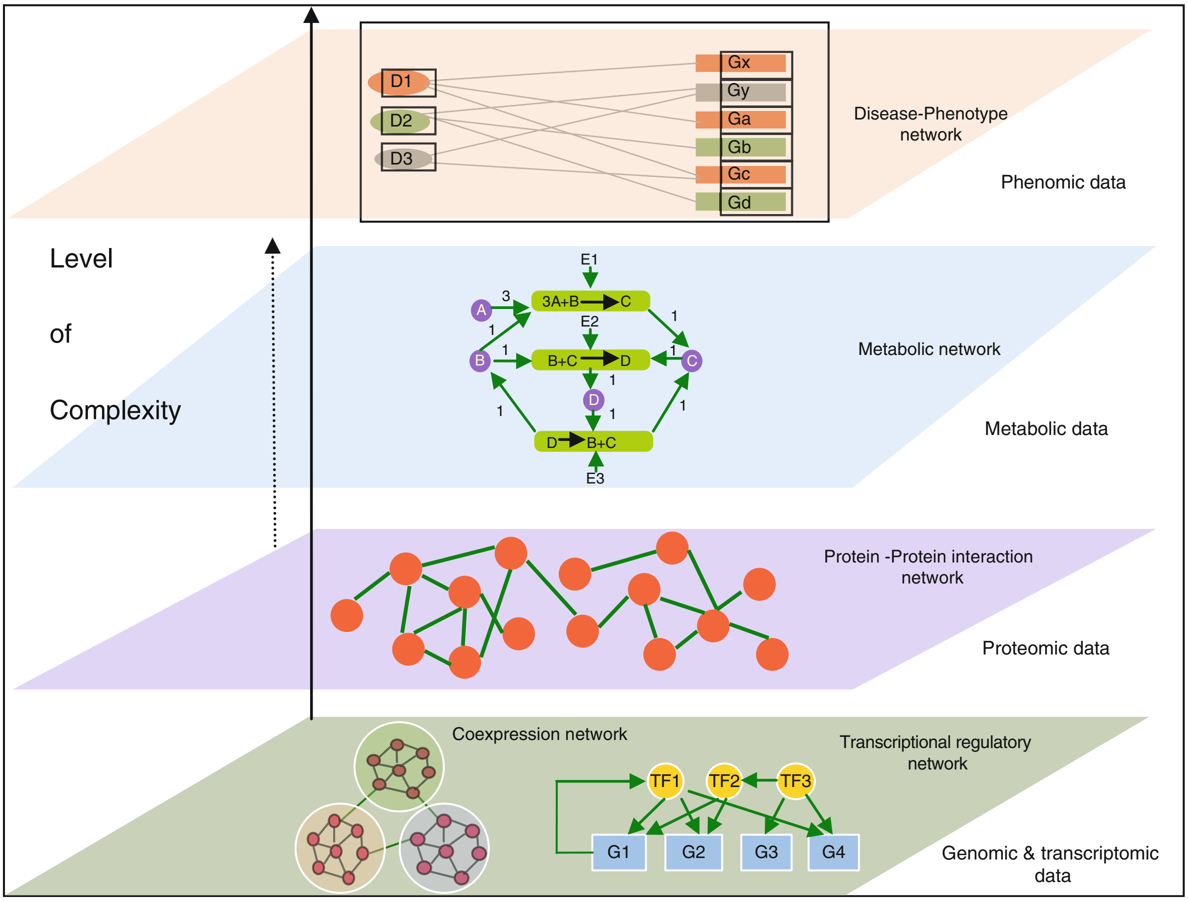
Network Classification/Network Types

The high-throughput data-collection techniques like microarrays, protein chips, or yeast two-hybrid screens determine how and when these molecules interact with each other. Various types of interaction networks, including protein–protein interaction, and metabolic, signaling, and transcription-regulatory networks emerge by integrating these data and interactions. Figure 1 describes interrelation across different types of biological networks, which are mostly constructed from high-throughput molecular data. The basic networks are required to study any biological phenomena or disease includes protein network, signaling and metabolic network, co-expression and transcriptional regulatory network, and protein-DNA interaction network (Fig. 1). All these networks are analyzed to identify biologically relevant modular structure of the networks (Please see the box for Modular Network).

The classifications of such biological networks are mostly based on types of input data, information connectivity, architecture topology, and overall functional interpretations.

Based on aforementioned criteria, some of the important networks are defined as follows:

1. Protein–Protein Interaction Network
   - It is a network of interacting proteins. This is undirected network, in which proteins are represented by nodes and the interaction between the proteins are represented by edges.



Biological Disease Mechanism Networks, Fig. 1 *Biological/molecular networks and relative complexity* It shows sequential increase in complexity in terms of both data and information. The layers represent different network types and cross-correlation of different types of molecular networks, which are finally merged to construct complex disease networks (Please refer Network Complexity Box for details) (Note: D1, D2, D3 exemplify related diseases and Gx, Gy, Ga, etc. represent reported candidate genes for corresponding diseases)

2. Signaling and Metabolic Networks
   - Signaling network is molecular bridge of events between the cell and the outside environment. It is directed and is composed of proteins or enzymes or factors that trigger or suppress the expression of a number of genes. The signaling event involves various proteins localizing to different compartments in the cell and finally controlling gene expression in the cell nucleus (Hughey et al. 2009).
   - The signaling networks are mostly integrated with metabolic networks, which are biochemical reactions along of substrates, metabolites, and enzymes. The nodes are either metabolites or enzymes or reactions. It is a directed and weighted network. It is constructed from the literature information on reactions, enzymes, genes, substrate–enzyme concentration, product concentration, etc. (Ma and Zeng 2003; Albert 2005).
3. Co-expression Network and Transcriptional Regulatory Network
   - *Co-expression networks* are constructed from the *co-expression measures* of genes across various tissue samples. Here, the nodes correspond to genes and edges represent the connection

strength which is obtained from the co-expression similarity. It is also called as correlation or association networks (Horvath and Dong 2008).

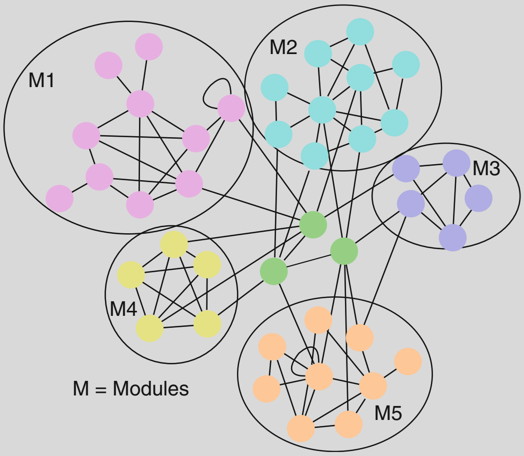
- The co-expression data is modeled to generate *transcriptional regulatory networks*, in which one gene may regulate transcription of another gene either directly or indirectly. Here, the nodes are genes that are connected through directed edges. Arrows in the network topology diagram indicate that they trigger the target gene activation, while bars in the network indicate a repressive effect on the target gene (Keurentjes et al. 2007). This network presents both physical and functional interactions composed of genes controlled by transcription factors. The network is directed and the node types are genes and transcription factors. The transcriptional regulatory networks are having several signature motifs including feed-forward loop motif, bifan, and autoregulatory loops (Dobrin et al. 2004). Descriptions of these motifs are given in the following paragraphs.
- Feed-forward loop motif – It consists of a pattern of three genes which is composed of two transcription factors and a target gene. One of the transcription factors regulates the other while both together regulate the target gene.
- Autoregulatory loops – It consists of a regulator gene that in turn binds to its own gene promoter, thereby bringing autoregulation.
- Bifan motifs – Bifan consists of two source nodes, which directly cross regulates two target nodes. Such motifs are common in mammalian cell signaling and in transcriptional networks.

4. Protein–DNA interaction (PDI) network – It is constructed from interactions between proteins/transcription factors and gene regulatory elements such as promoters, cis- and trans-regulatory elements. Such interactions lead to temporal gene expression during development, and other physiological status. Here, the nodes represent “interactor proteins” and promoters as “DNA targets” while the edges represent the interactions between them. The complexity of PDI is dependent on number of promoter targets per interactor/protein (outgoing connectivity) and the number of interactors per promoter (in coming connectivity). The network is undirected (Deplancke et al. 2006).

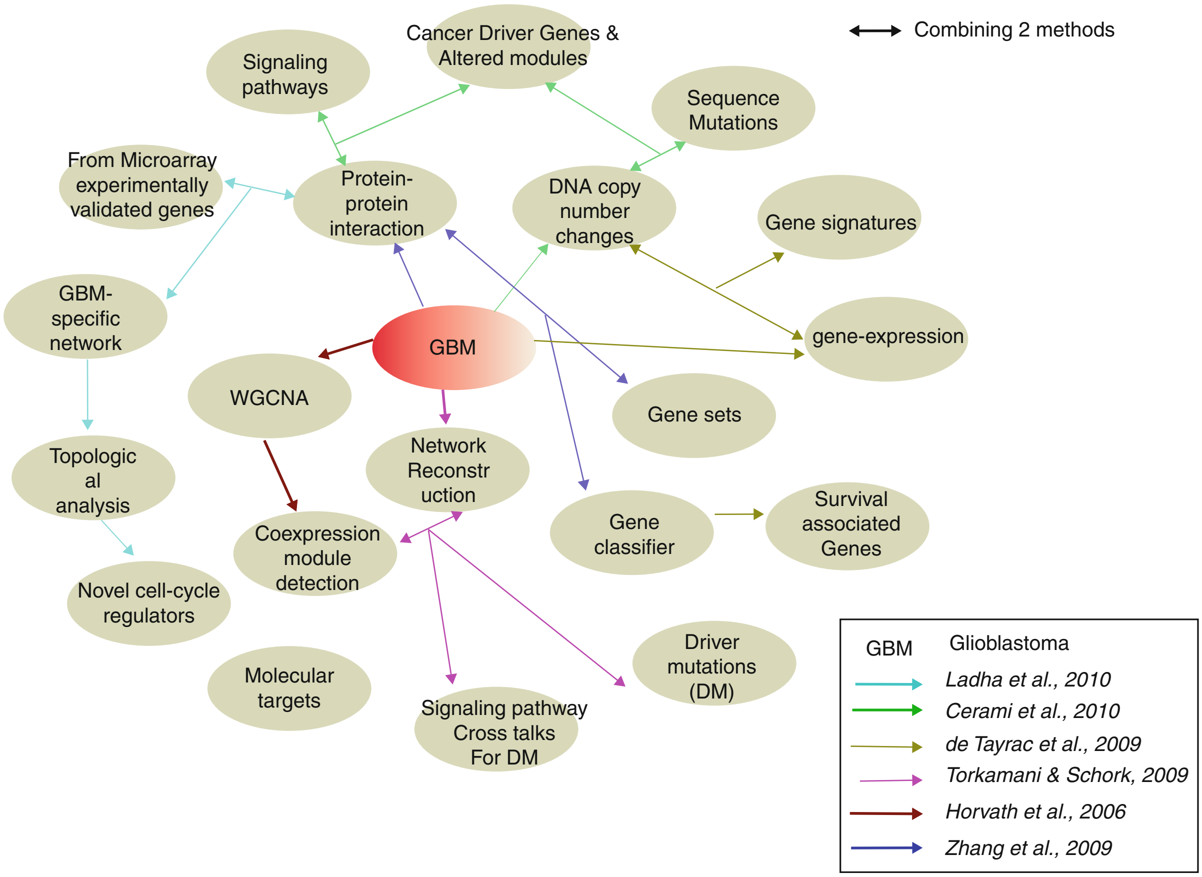
Network Complexity

A rational integration of molecular networks describes substantial understanding of the complexity of a biological phenomena/molecular mechanism. For example, a researcher constructs biologically important networks separately from genomics and proteomics data from a diseased system. Then, these networks are correlated and integrated to develop a single and meaningful disease model or network. The biological significance and complexity of a molecular network increases by integrating various types of molecular data/networks as one base and holistic network.

Modular Network



Modular networks are created due to interactions across smaller subnetwork modules. The biological networks are functionally organized into modules, which are group of genes forming hubs. Interactions between the modules represent the cross talk between them. Each module’s function is determined by the module organizer genes or proteins. This modular fashion of networks reduces the complexity into a small number of connected structures and function (Rives and Galitski 2003).



Biological Disease Mechanism Networks, Fig. 2 Graphical representation of different approaches, which have been used to construct systems biology models in GBM. The *arrows* in different colors correspond to specific research methods and results available in the literature

Disease Mechanism Network

High-throughput molecular, biochemical, and genetic data are used to construct disease specific networks in human. Such networks hold high relevance in terms of elucidating disease mechanism, identification of important candidate genes and pathways, biomarkers etc.

There are some landmark developments in the area, which signify understanding on disease mechanism and description of human diseasome through the correlation of candidate genes and corresponding phenotypic properties. The following network describes a diseasome.

Phenotypic Disease Network (PDN)

It is the network of interactions between diseases and the candidate genes. The candidate genes are linked to a disorder/disease from the known disease–gene relationship. The nodes correspond to candidate genes or disease names while the edges represent the relation between them. This is a type of undirected network (Goh et al. 2007). The phenotypic disease network displays a relation across multiple diseases through overlapping candidate genes.

Scientists working on specific complex diseases have used various approaches to build disease networks to understand the disease mechanism at system wide scale. Studies on complex diseases like type 2 diabetes, prostate cancer, lung cancer, colon cancer, glioma, malaria, and tuberculosis led to the identification of candidate genes, biomarkers, and novel cross talk and protein interaction points. The network biology approach *has led to the discovery of causative pathways* in several of these complex diseases like androgen receptor (AR) pathway for metastatic prostate cancer (Vellaichamy et al. 2010), role of TGFB pathway in inducing oxidative stress and MAPK pathways to finally facilitate vascular complications in type 2 diabetics (Sengupta et al. 2009). Network-based analysis has also facilitated discovery of markers in brain cancer, i.e., identification of CSK21 and PP1A as progression markers in the pathogenesis of glioblastoma (GBM) (Ladha et al. 2010).

Glioblastoma

GBM is the most common primary brain tumor occurring in adult population. They are highly malignant and in turn show several subtypes. System biology approaches have identified potential tumor suppressor genes, prognostic genes, gene signatures, tumor subtypes, survival-associated genes, and cell-cycle regulators.

The functionally relevant weighted gene coexpression network analysis (*WGCNA*) method has identified some key molecular targets for glioblastoma (Horvath et al. 2006). A similar approach has detected rare cancer driver mutations, which could drive late tumorigenesis (Torkamani and Schork 2009).

An *integrated approach* based on DNA copy number changes and gene expression changes was used to identify targeted gene signatures, tumor subgroups, and potential tumor suppressor genes for glioblastoma (de Tayrac et al. 2009). *Systems approach* based on combined gene sets and protein interaction network has been used to develop a prognostic gene classifier that can predict survival-associated genes (Zhang et al. 2009).

Another *approach* combining DNA copy numbers and mutations, associated with sequences, along with human interaction networks has helped in identifying the cancer driver genes and potentially altered modules (Cerami et al. 2010).

Our own investigation from experimentally validated upregulated genes and corresponding protein–protein interaction information has led to identification of novel and important connecting proteins, CSK21 and PP1A, which are implicated in cell-cycle regulation (Ladha et al. 2010) (Fig. 2).

References

Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118(Pt 21):4947–4957

Cerami E, Demir E, Schultz N, Taylor BS, Sander C (2010) Automated network analysis identifies core pathways in glioblastoma. PLoS One 5(2):e8918.

Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, Sequerra R, Doucette-Stamm L, Reece-Hoyes JS, Hope IA, Tissenbaum HA, Mango SE, Walhout AJ (2006) A gene-centered *C. elegans* protein-DNA interaction network. Cell 125(6):1193–1205

Dobrin R, Beg QK, Barabási AL, Oltvai ZN (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. BMC Bioinformatics 5:10

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL (2007) The human disease network. Proc Natl Acad Sci USA 104(21):8685–8690

Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Qi S, Chen Z, Lee Y, Scheck AC, Liau LM, Wu H, Geschwind DH, Febbo PG, Kornblum HI, Cloughesy TF, Nelson SF, Mischel PS (2006) Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc Natl Acad Sci USA. 103 (46):17402–17407

Horvath S, Dong J (2008) Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 4(8): e1000117

Hughey JJ, Lee TK, Covert MW (2009) Computational modeling of mammalian signaling networks. Wiley Interdiscip Rev Syst Biol Med 2(2):194–209

Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, Peeters AJ, Vreugdenhil D, Koornneef M, Jansen RC (2007) Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. Proc Natl Acad Sci USA 104(5):1708–1713

Ladha J, Donakonda S, Agrawal S, Thota B, Srividya MR, Sridevi S, Arivazhagan A, Thennarasu K, Balasubramaniam A, Chandramouli BA, Hegde AS, Kondaiah P, Somasundaram K, Santosh V, Rao MR (2010) Glioblastoma-specific protein interaction network identifies PP1A and CSK21 as connecting molecules between cell cycle–associated genes. Cancer Res 70(16):6437–6447

Ma H, Zeng AP (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 19(2):270–277

Rives AW, Galitski T (2003) Modular organization of cellular networks. Proc Natl Acad Sci USA 100(3):1128–1133

Sengupta U, Ukil S, Dimitrova N, Agrawal S (2009) Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications. PLoS One 4(12):e8100

Torkamani A, Schork NJ (2009) Identification of rare cancer driver mutations by network reconstruction. Genome Res 19 (9):1570–1578

Vellaichamy A, Dezso Z, Lellean JeBailey AM, Chinnaiyan ASreekumar, Nesvizhskii A, Omenn GS, Bugrim A (2010) "Topological significance" analysis of gene expression and proteomic profiles from prostate cancer cells reveals key mechanisms of androgen response. PLoS One 5(6):e10936

Zhang J, Liu B, Jiang X, Zhao H, Fan M, Fan Z, Lee JJ, Jiang T, Jiang T, Song SW (2009) A systems biology-based gene expression classifier of glioblastoma predicts survival with solid tumors. PLoS One 4(7):e6274

Biological Marker

► Biomarkers

Biological Module

► Organelle and Functional Module Resources

Biological Network Model

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

Definition

Biological network model is an abstract, graphical, or mathematical representation of a biological system using nodes and connecting edges to denote biomolecules and biomolecular interactions, respectively.

Biological System Model

Eberhard O. Voit
The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Definition

A biological system model is a mathematical representation of a biological network, its regulation, and dynamics.

Biomarker

► Biomarker Discovery, Knowledge Base

Biomarker Discovery, Knowledge Base

Donna L. Mendrick and Weida Tong
Division of Systems Biology, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR, USA

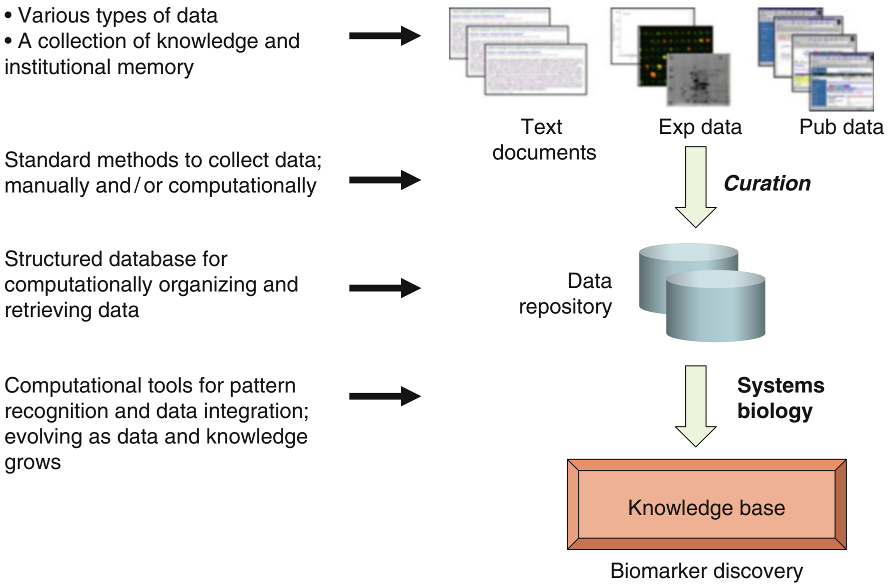
Synonyms

Adverse events; Biomarker; Knowledge base; Systems biology

Definition

The current challenges as well as opportunities in biomarker discovery lie in the integration of in-house generated data of multiple types together with diverse data in the public domain to assess disease and toxicity at the systems level (systems biology). The knowledge base approach (Fig. 1) takes advantage of current advances in computer technology and bioinformatics to integrate diverse data sets into a content-centric resource for evaluation of molecular biomarkers in determining efficacy or adverse events in animals and humans. Such a knowledge base will spawn hypotheses to develop new studies to address the current gaps and lead to further improvements in biomarker discovery. New data and information generated from these studies, in turn, will further enrich the knowledge base. Knowledge bases are not only important in biomarker discovery for biomedical research and drug development, but also essential for the regulatory agencies for use as a first tier of information to review drug submissions supported by biomarkers for efficacy or safety concerns.

Disclaimer: The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration.


• Various types of data
• A collection of knowledge and institutional memory
Standard methods to collect data; manually and / or computationally
Structured database for computationally organizing and retrieving data
Computational tools for pattern recognition and data integration; evolving as data and knowledge grows
Text documents
Exp data
Pub data
Curation
Data repository
Systems biology
Knowledge base
Biomarker discovery

Biomarker Discovery, Knowledge Base, Fig. 1 A knowledge base schema for biomarker discovery

Biomarker Discovery, Typical Process

Emily A. Moon and Marquis P. Vawter
Functional Genomics Laboratory, Department of Psychiatry and Human Behavior, University of California, Irvine, CA, USA

Definition

Biomarker – a specific physical measure or indicator of healthy biological processes, disease states, or pharmacologic responses to treatment. Can be a predictor of disease severity, onset, or recovery; does not have to be related to pathophysiology, may be an indirect marker of pathophysiology.

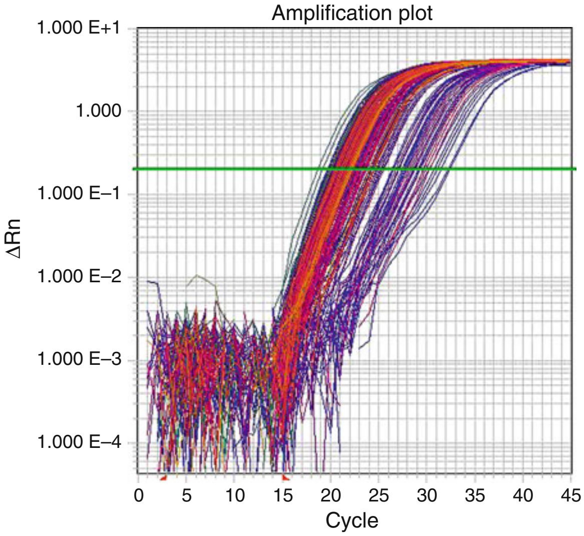
Characteristics

The discovery of novel biomarkers represents a lucrative future of preventative, predictive, and personalized medicine (▶ Biomarkers, Clinical Relevance). A great subset of funding from the pharmaceutical and clinical diagnostic industries and government sponsors has resulted in the detection and validation of biomarkers relevant to disease state, point of care monitoring, drug-metabolism, drug-efficacy, and drug-toxicity biomarkers. The processes by which biomarkers are discovered vary, depending on the type of biological materials and health or disease states being investigated. Typically, the process begins when researchers discover an association between a specific indicator and a disorder, diagnosis, or drug response, usually out of a large number of indicators that are analyzed in the initial phase of the study. These significant indicators are then verified through a range of tests and analyses, depending upon the biomarker being studied, and then validated in a larger sample set that seeks to replicate the biomarker association across multiple populations. Regardless of the course that researchers take in the biomarker discovery process, however, this process is only the initial step in a path toward clinical validation and commercialization of novel biomarkers.

Perhaps the best way to understand the biomarker discovery process is to look at a few examples within the discovery of blood-derived biomarkers.

RNA

RNA biomarkers are typically RNA measurements that are indicators of normal biologic processes, disease states, toxicological reaction, or therapeutic response to treatment. RNA species can include microRNA, message RNA, ribosomal RNA, and transfer RNA levels, as well as RNA editing and splicing of message RNA. Alterations in the sequences and expression of the RNA molecules are commonly used biomarkers. The RNA biomarker discovery process begins with the recruitment of individuals for a biomarker study and the isolation of RNA from subjects' whole blood or peripheral blood mononuclear cells (PBMCs), or generation of lymphoblastic cell lines (LCLs). Researchers typically screen the known set of genes with candidate genes generated either from published studies or unpublished pilot studies done by the researcher that repeatedly show expression alterations in the state of interest, as compared to normal expression. These putative biomarkers are usually compiled via gene-focused microarray analysis of whole blood, LCL, or PBMCs. Once a gene, gene list, or network of genes of interest is compiled, high quality RNA from subjects that meet the researcher's inclusion criteria is transcribed into cDNA and expression data is generated via quantitative PCR, normalized with an appropriate reference gene. The qPCR data can be generated with real-time amplification plots (Fig. 1) and is analyzed to evaluate the ability of individual and composite gene expression markers to differentiate cases from controls based on transcript abundance as well as to validate the microarray results. The primary endpoint of the RNA biomarker discovery process is the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curve (▶ Area Under the ROC Curve, ▶ Receiver Operating Characteristic (ROC) Curve). The ROC is a plot that measures classification performance of a model across the entire range of thresholds (Dwinnell 2010). The ROC curve is generated by graphed data points of the true positive rate (sensitivity) and the false positive rate (100 minus specificity). The more the ROC curve breaks toward the top-left of the graph, the better the model is at separating cases from controls, as opposed to a random prediction shown by the dashed line (Fig. 2). Thus, AUC of the ROC represents the predictive ability of the gene expression markers, and a greater AUC correlates to a stronger predictive marker. If an expression profile based on the ROC can be developed for cases versus controls, the biomarker can be used for early detection, monitoring, and treatment decisions with predictive medicine for disorders, and with the added benefit of greater cost effectiveness.

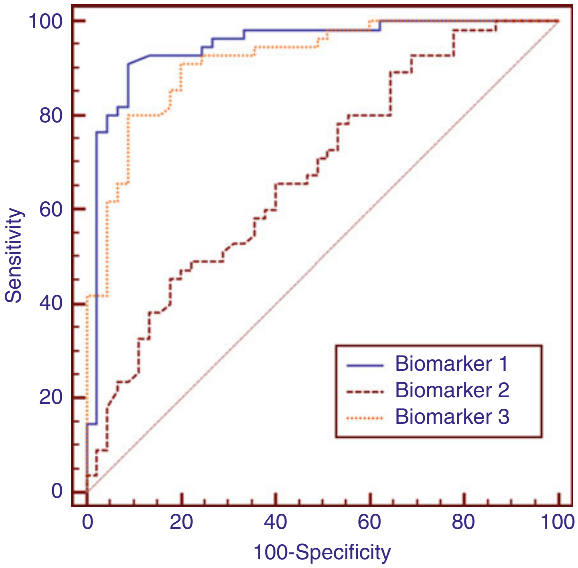


Biomarker Discovery, Typical Process, Fig. 1 A qPCR amplification plot is the plot of fluorescence signal versus cycle number. Figure 1 is an amplification plot of the HLA-CD74 gene in brain, showing the change in fluorescence of SYBR Green dye (*Y axis*) plotted versus cycle threshold (*X axis*). Lower Ct number represents greater expression of the HLA-CD74 gene

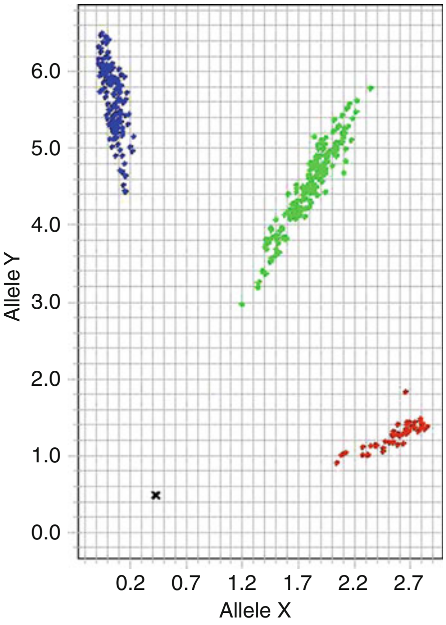
DNA

Genomic biomarkers can be characterized by variations in DNA: single nucleotide polymorphisms (SNPs), haplotypes, insertions, copy number variation, inversions, deletions, or methylated cytosine residues. Relevant SNPs can be determined from a subject's isolated DNA by direct sequencing or SNP genotyping assays which employ fluorescence technology to assess genotypes. Figure 3 shows an allelic discrimination plot generated from a TaqMan SNP genotyping assay from Applied Biosystems (Foster City, CA). ▶ Fluorescent markers are detected at the locus of interest, depending on genotype of the sample. The blue cluster is a software call of homozygous Allele Y, marked with FAM fluorescence, and the red cluster is a call of homozygous Allele X, marked with VIC fluorescence. Dual fluorescence, an indicator of heterozygosity, is seen as the green cluster centered between the two fluorescent extremes. The results of multiple SNP genotyping assays or sequencing studies can be analyzed together to find haplotype biomarkers.

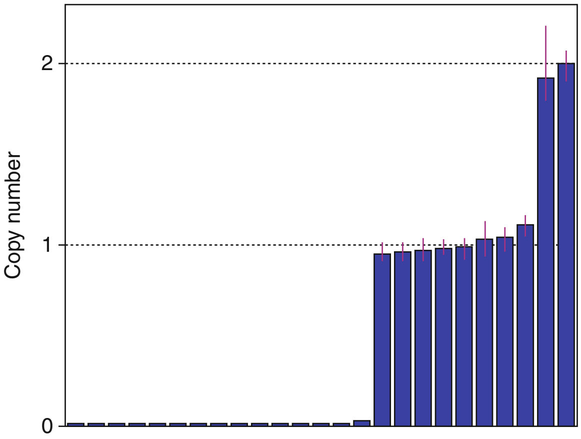
Copy number variation (CNV) assays consisting of a primer/probe reagent mixture can be purchased from



Biomarker Discovery, Typical Process, Fig. 2 Three biomarkers showing the degree of sensitivity and specificity for predicting cases and controls; biomarker 1 and 3 are nearly equal, and both are better predictors than biomarker 2



Biomarker Discovery, Typical Process, Fig. 3 An allelic discrimination plot of genotype calls at Neuregulin1 SNP rs3924999, showing homozygotes (*blue* and *red*) and heterozygotes (*green*)



Biomarker Discovery, Typical Process, Fig. 4 Bar graph output from CopyCaller, demonstrating 0, 1, and 2 copy number variants in DNA. Each bar represents one sample

bioscience companies, or can be created in the laboratory using primer and probe and standard dilutions of calibrated DNA references. Figure 4 shows a table generated by CopyCaller after an Applied Biosystems CNV assay run on an ABI 7900HT PCR system (Applied Biosystems, Foster City, CA). The assay uses a multiplex reaction that includes a reference assay and a target assay, and the cycle numbers of the two assays are compared for each sample, and a copy call is generated. CNV calls are analyzed to determine their predictive power of the characteristic of interest.

Methylation of cytosine deoxyribonucleotides in DNA (▶ DNA Methylation) has been added to this critical list of genomic biomarkers, and there has been a flood of methylation research in all fields of medicine to discover valid epigenetic markers. DNA cytosine methylation is stable, and although not permanent, could mark transition states from health to disease as well as environmental alterations to the epigenome. Biomarkers of methylated DNA are typically discovered via PCR techniques, as methylated DNA is readily amplifiable and easily detectable. Typically, DNA is treated with sodium-bisulfite, which converts cytosine residues to uracil. PCR and nucleotide sequencing is then used to detect the converted base pair and confirm the presence of methylation at cytosine residues (Shiraishi and Hayatsu 2004). DNA hypermethylation usually occurs near or within the promoter region of genes, although microarray probes can measure methylation changes across all regions of genomic DNA. Biomarkers of methylation are particularly useful, as screening in at-risk populations can occur with minimally invasive techniques. Multiple studies show that DNA methylation of certain loci can be detected in blood, sputum, bronchoalveolar lavage, and, potentially, exhaled breath-condensate (Anglim et al. 2008).

Proteins

Protein biomarker discovery (▶ Biomarkers, Protein Expression) can be defined as the process by which the differential expression of proteins in diseased versus healthy or transitional states is first identified. Protein biomarker discovery can use model systems (i.e. mouse models, cell lines) or the analysis of a variety of human biological fluids including blood, urine, tissue lysates, and cerebrospinal fluid. Protein biomarker discovery commonly employs two-dimensional gel electrophoresis and mass spectrometry technology to separate and identify differential expression of proteins in diseased and normal states, via measurement of peptide abundance.

From this first pass separation and identification analysis, a list of candidate biomarker proteins is generated. These lists generally consist of hundreds of candidate biomarkers, with a high rate of false positives. A biomarker candidate verification phase reduces the number of false positives and ensures that only the most promising putative biomarkers found in discovery go on to the validation phase. Human plasma (if not used in the discovery process) is typically analyzed during verification, as it is considered the most comprehensive illustration of the human proteome and is in contact with all tissues and could be a sentinel of the disease processes (Anderson and Anderson 2002). Once a candidate biomarker has been verified, it can move on to the clinical validation and commercialization process.

Caveats of Biomarker Discovery: DNA Versus RNA Versus Protein

DNA diagnostics such as SNPs are useful but often do not reflect variants that have a biological role. Moreover, DNA variants do not provide critical information on the environmental factors that are important for pathogenesis of these diseases except in cases of epigenetic biomarkers. Protein biomarker discovery, though rich in promise, is not always a particularly fruitful endeavor. Several limiting factors to protein biomarker discovery exist and include:

- The relatively low abundance of some biomarkers in the proteome that must be detected in a complex matrix and the lack of means to amplify these infrequent markers
- The post-translational complexity of protein biomarkers in human blood and biological fluids making detection somewhat ambiguous
- The variability in human population and pathologies that make valid biomarkers very difficult to detect and apply to disease states (Rifai et al. 2006)

Further, on a genome-wide basis, protein or multiproteins levels are quite difficult to obtain and technically not feasible at this time. However, this is not as large of a problem for RNA diagnostics, which evaluate the expression of the genes in question. Because gene expression can reflect both genetic and environmental influences, it may be particularly useful for identifying risk factors for complex disorders which are thought to have a multi-factorial polygenic etiology in which many genes and environmental factors interact. In addition, multiple gene biomarkers can be used to identify disease risk and to predict or monitor drug response. Lastly, novel RNA diagnostic tools can be employed with PBMCs, whole blood or LCLs, providing more options for researchers and practitioners.

Other Biomarkers

Besides the trilogy of RNA, DNA, and protein, the field of biomarkers routinely engages the use of other diverse measures such as metabolomics, lipids, small molecules such as steroids, and toxic environmental residues in bodily tissues. Also to be considered as biomarkers are the following imaging techniques: computed tomography, magnetic resonance imaging, positron emission tomography, and ultrasound. Electrophysiological measures such as electroencephalogram, electrocardiogram, and electromyogram are considered informative biomarkers, to accurately measure tissues of interest in disease or transitional states. Biomarker discovery in these categories follows a process similar to that of blood-derived biomarkers, but with verification and validation stages and analyses that suit the biomarker in study.

Cross-References

- ▶ Area Under the ROC Curve
- ▶ Biomarkers, Clinical Relevance
- ▶ Biomarkers, Protein Expression
- ▶ DNA Methylation
- ▶ Fluorescent Markers
- ▶ Receiver Operating Characteristic (ROC) Curve

References

Anderson NL, Anderson NG (2002) The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics 1(11):845–867

Anglim PP, Alonzo TA, Laird-Offringa IA (2008) DNA methylation-based biomarkers for early detection of non-small cell lung cancer: an update. Mol Cancer 7:81

Dwinnell W (2010) ROC curves and AUC. http://matlabdatamining.blogspot.com/. Accessed 1 June 2010

Rifai N, Gillette MA, Carr SA (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol 24(8):971–983

Shiraishi M, Hayatsu H (2004) High-speed conversion of cytosine to uracil in bisulfite genomic sequencing analysis of DNA methylation. DNA Res 11(6):409–415

Biomarkers

Donna L. Mendrick and Weida Tong
Division of Systems Biology, National Center for
Toxicological Research, US Food and Drug
Administration, Jefferson, AR, USA

Synonyms

Biological marker; Molecular marker; Surrogate endpoint

Definition

A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Biomarkers Definitions Working Group 2001). A biomarker can be used as an indicator of a change or, if it meets the highest standard of proof, a surrogate endpoint that is expected to predict clinical benefit (or harm, or lack of benefit or harm) based on epidemiologic, therapeutic, pathophysiologic, or other scientific evidence.

Biomarkers are widely used in medicine, drug discovery and development, and safety assessment. They serve as indicators for disease progression (diagnosis and prognosis), therapeutic responses (efficacy), and adverse side effects (safety) in organisms, organs, and/or cells. An ideal clinical biomarker should contain the following characteristics (Lesko and Atkinson 1999): (1) clinical relevance, (2) sensitivity and specificity to treatment effects, (3) reliability, (4) practicality, and (5) simplicity.

Biomarkers need to be qualified to define their specific use and context (fit for purpose). However, most biomarkers in use today have not been evaluated in a comprehensive qualification process (Goodsaid et al. 2008). This, unfortunately, means that new biomarkers are being compared to older ones for which a full understanding of their context for use is not known. Currently, there is no widely accepted framework for biomarker qualification. The FDA has taken an initiative to develop a consistent and standardized qualification framework for the acceptance of biomarkers for regulatory use. Such a regulatory driven effort will facilitate communication between regulatory agencies, pharmaceutical companies, the research community, clinical practice, and consumer participants for evaluation of the biomarker-surrogate-clinical endpoint relationship in different settings and applications.

Current biomarker discovery increasingly is relying on emerging molecular technologies that aim to determine the causal and mechanistic relationships of molecular markers with clinical endpoints. The representative technologies and associated biomarker types are summarized in Table 1, and most of them are high throughput or high content in nature. These technologies can be applied independently or in parallel; the latter is able to identify multiple biomolecules at different level of biological complexity.

Disclaimer: The views presented in this article do not necessarily reflect those of the US Food and Drug Administration.

Biomarkers, Table 1 An emerging molecular technology landscape in biomarker discovery

Different biological levels	Scientific disciplines and their representative molecular technologies	Data type	Biomarkers
DNA	Genetics/epigenetics: Genome Wide Association Study (GWAS), next generation sequencing	SNP variation	Genetic markers
RNA	Genomics: microarrays; next generation sequencing	Gene expression	Genomic markers
Protein	Proteomics: 1D/2D gel coupled with MS or MS/MS	Protein profiling	Protein markers
Metabolite	Metabolomics: NMR and MS	Metabolites	Metabolomics markers
Cell	Drug screening: cell-based assays	Multiple mechanistically relevant parameters	Cellular biomarkers

Cross-References

- ► Biomarkers, Protein Expression
- ► Biomarkers, Solid Tissue

References

Biomarkers Definitions Working Group (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 69:89–95

Goodsaid FM, Frueh FW, Mattes W (2008) Strategic paths for biomarker qualification. Toxicology 245:219–223

Lesko LJ, Atkinson AJ Jr (1999) Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. Annu Rev Pharmacol Toxicol 41:347–366

Biomarkers, Blood Derived

Marquis P. Vawter and Emily A. Moon
Functional Genomics Laboratory, Department of
Psychiatry and Human Behavior, University of
California, Irvine, CA, USA

Definition

Human blood provides a source for multiple types of biomarkers. Within blood, biomarkers can be found in multiple components (plasma, serum, cellular) at the following levels.

Functional Genomic

A functional genomic biomarker can be defined as an RNA characteristic that is an indicator of normal biologic processes, disease states, toxicological reaction, or therapeutic response to treatment. RNA species can include microRNA, message RNA, ribosomal RNA, transfer RNA levels, as well as RNA editing and splicing of message RNA. Alterations in the sequences and expression of the RNA molecules are commonly used biomarkers.

Genomic

Genomic biomarkers can be characterized by variations in DNA (single nucleotide polymorphisms, insertions, copy number variation, translations, inversions, haplotype effects or deletions). A recently discovered single nucleotide biomarker predicting response to Hepatitis C treatment is a relevant example of a DNA level biomarker of therapeutic response to pegylated interferon-alpha-2b. The CC genotype at rs12979860, located 3 kb from the IL28B gene on chromosome 19, is associated with a twofold greater rate of sustained virological response (the absence of detectable Hepatitis C virus) at the end of pegylated interferon-alpha-2b treatment, over the TT genotype (Ge et al. 2009). Genomic screening also increasingly includes mitochondrial DNA as a biomarker for mitochondrial diseases, and there are databases developed that list all mitochondrial DNA mutations associated with disease such as MITOMAP (http://www.mitomap.org/MITOMAP).

Epigenomic

DNA methylation involves the addition of a methyl group to DNA and is essential for normal development. Mammalian DNA methylation occurs mostly at the number 5 carbon of the cytosine of a CpG dinucleotide. Approximately 75% of all CpGs are methylated, and unmethylated CpGs are clustered in "CpG islands," generally located at the 5′ end of a gene. In many disease processes, CpG islands undergo abnormal hypermethylation. Hypermethylation has been found to be a powerful biomarker for prostate cancer screening, detection, and diagnosis (Nakayama et al. 2004), as well as early detection and monitoring of lung cancer (Belinsky 2004). DNA methylation may also be useful for monitoring adverse environmental effects upon DNA.

Peptides and Proteins

Proteins and peptides are often studied in the biomarker discovery process, as proteomics gives a much better dynamic understanding of an organism than genomics. Human plasma is thought to be the most comprehensive representation of the proteome and of all body tissues and processes. Protein biomarker discovery faces many challenges given the complexity of the proteome; the difficulties in studying low abundance proteins, where many biomarkers are thought to exist; and the variation within populations and pathologies. Protein biomarker discovery will become a more fruitful endeavor as high throughput proteome technologies and data mining continue to advance.

Metabolomics

Metabolomics, the study of the small molecules (such as metabolites) that chemical processes leave behind, is a rapidly emerging field of study and rather promising source of novel biomarkers. Studying these molecules, via mass spectrometry or capillary zone electrophoresis, is particularly useful in biomarker discovery, as metabolomics takes into account the way lifestyle, diet, and environment affect the health of an individual, in addition to genetics. Additionally, as compared to the genome, transcriptome, and proteome, the metabolome is very small, and with a high translatability across species and eukaryotic and prokaryotic cells. One caveat of metabolomic biomarker discovery, however, is the enormous amount of data that is generated from metabolomic mining studies. Analytical technologies must advance in order to fully elucidate the potential of the information-rich field of metabolomics.

References

Belinsky SA (2004) Gene promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer 4:707–717

Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, Heinzen EL, Qiu P, Bertelsen AH, Muir AJ, Sulkowski M, McHutchison JG, Goldstein DB (2009) Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature 461:399–401

Nakayama M, Gonzalgo ML, Yegnasubramanian S, Lin X, De Marzo AM, Nelson WG (2004) GSTP1 CpG island hypermethylation as a molecular biomarker for prostate cancer. J Cell Biochem 91:540–552

Biomarkers, Clinical Relevance

Marquis P. Vawter and Emily A. Moon
Functional Genomics Laboratory, Department of Psychiatry and Human Behavior, University of California, Irvine, CA, USA

Definition

Biomarkers have significant clinical relevance as diagnostic tools. Finding diagnostic biomarkers for diseases that can only be identified by observation and patient self-reports, like schizophrenia and Alzheimer's, can direct proper treatment and, for diseases where cure or improvement in prognosis is related to early detection and intervention, potentially extend the lives of affected patients.

Biomarker discovery represents a plausible future of preventative, predictive, and personalized medicine.

Preventative

An example of a preventative biomarker is blood pressure, used to assess risk for coronary heart disease (CHD) in a population. A meta-analysis of nine studies of over 400,000 subjects found that there was a positive, continuous, log-linear relationship between diastolic blood pressure (DBP) and stroke/CHD events (MacMahon et al. 1990). Blood pressure screening represents a preventative biomarker that, when indentified, can be used to direct treatment for lowering DBP, either via drug therapies or changes in diet and exercise.

Predictive and Personalized

Biomarkers can also be used to predict the efficiency of drug treatment in affected patients. One example of a drug-response predictive biomarker is cytochrome P450 2D6 (*CYP2D6*), a gene predominantly expressed in the liver and involved in the metabolism of many drugs, including antidepressants and tamoxifen, which is biotransformed to the estrogen receptor blocker, endoxifen, by the CYP2D6 enzyme. Among women with estrogen-dependent breast cancer, tamoxifen is the most widely used drug therapy in tumor suppression. Studies on *CYP2D6* polymorphisms suggest that individuals carrying functional variants associated with deficient *CYP2D6* metabolism receive less curative benefit from tamoxifen and are at a higher risk of relapse than those without these variants (Goetz et al. 2007). Screening breast cancer patients for these variants can guide practitioners in developing a more personalized and effective treatment plan. In this example, prediction of better survival rates for patients with estrogen-dependent malignancy treated with tamoxifen is conditioned on the genetic variant that increases tamoxifen bioavailability.

Biomarker discovery also plays a role in clinical economics; as health care moves more toward individualized care, the market for biomarker commercialization has grown substantially. The increased allocation of health-care resources to molecular diagnostics has created a $23 billion dollar potential global market of high-cost, high-profit products (Wilson et al. 2007). As an example, prescreening patients in a drug trial based upon a genetic variant can potentially limit the cost of achieving positive results by reducing the number of non drug responders enrolled.

References

Goetz MP et al (2007) The impact of cytochrome P450 2D6 metabolism in women receiving adjuvant tamoxifen. Breast Cancer Res Treat 101:113–121

MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, Abbott R, Godwin J, Dyer A, Stamler J (1990) Blood pressure, stroke, and coronary heart disease. Part 1, Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. Lancet 335:765–774

Wilson C, Schulz S, Waldman SA (2007) Biomarker development, commercialization, and regulation: individualization of medicine lost in translation. Clin Pharmacol Ther 81:153–155

Biomarkers, Independent Validation

Péter Horvatovich
Department of Pharmacy, Analytical Biochemistry
Research Group, University of Groningen, Groningen,
The Netherlands

Synonyms

Machine learning; Model cross-validation; Model testing; Model validation

Definition

Biomarker discovery is a complex procedure consisting of discovery and validation phases. In cases where there are no assumptions about the underlying molecular mechanism of a disease, the discovery phase consists of comprehensive profiling using the biological fluid, which will be used for the final diagnostic test. Analytical methods applied for comprehensive profiling are time-consuming and have low sample throughput, while they provide quantitative information on several hundreds or thousands of proteins or metabolites. The aim of statistical analysis is to select from these compounds those that may have an association with the biological states in question, such as diseased vs. healthy. However, due to the low sample size relative to the number of measured compounds, the outcome of statistical analysis may lead to the selection of spurious biomarker candidates that do not have a strong correlation with the biological state that is being monitored. Supervised statistical classification methods have different assumptions, which may not always be true, such as even distribution of features across the data set, similar variance in case control sample groups. There are assumptions about the generalization of findings and that a real association exists between the identified biomarkers and the investigated conditions. In addition, statistical methods in cases of low sample size and large quantified compounds tend toward overfitting, especially in cases where the outcome of classification is a combination of multiple compounds. Latent parameters not taken into consideration at sample selection, collection, storage, or analysis may influence the outcome of biomarker discovery and may result in poor classification performance to classify new sample sets using the model built with previously identified biomarkers (Mischak et al. 2010).

To test the association strength between the selected biomarkers and the studied biological state as well the generalizability of the classification, validation with an independent sample set is compulsory. The selection of a new sample set should be based on the same clinical criteria that were used in the study to identify the biomarker panel, and these clinical criteria should also be applied in the final diagnostic test. For example, if samples with a certain age range are used to perform discovery and validation studies, the final diagnostic test is only valid for that age range. In order to minimize the influence of latent variables it is advisable to perform the validation in a multi-center setting with blinded samples. Application of a faster analytical method targeting only the selected compounds is recommended for the validation phase to increase the statistical power. For example, when comprehensive profiling using label-free LC-MS is applied in the discovery phase, targeted Multiple-Reaction-Monitoring (MRM) may be used to quantify the identified discriminating compounds in the independent data sets. MRM acquisition enables faster analysis, resulting in higher sample throughput. Both the discovery and validation data sets should have a large enough sample size to provide meaningful statistical outcome (Mischak et al. 2010; Feng et al. 2004). However, use of an independent validation data set with enough power is a necessary but not sufficient condition to identify a clinically successful biomarker panel, but other parameters such as sample selection, collection, characterization and handling, and a clear statement of the clinical objectives are also important. Predictive value of the diagnostic test should be determined on the independent validation set as it provides unbiased results compared to the more optimistic predictive value determined using only the data set used for discovery (Azuaje 2010).

References

Azuaje F (2010) Bioinformatics and biomarker discovery: "Omic" data analysis for personalized medicine. Wiley, Chichester

Feng Z, Prentice R, Srivastava S (2004) Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. Pharmacogenomics 5(6):709–719

Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, Benigni A, Bennett SE, Bischoff R, Bongcam-Rudloff E, Capasso G, Coon JJ, D'Haese P, Dominiczak AF, Dakna M, Dihazi H, Ehrich JH, Fernandez-Llama P, Fliser D, Frokiaer J, Garin J, Girolami M, Hancock WS, Haubitz M, Hochstrasser D, Holman RR, Ioannidis JP, Jankowski J, Julian BA, Klein JB, Kolch W, Luider T, Massy Z, Mattes WB, Molina F, Monsarrat B, Novak J, Peter K, Rossing P, Sánchez-Carbayo M, Schanstra JP, Semmes OJ, Spasovski G, Theodorescu D, Thongboonkerd V, Vanholder R, Veenstra TD, Weissinger E, Yamamoto T, Vlahou A (2010) Recommendations for biomarker identification and qualification in clinical proteomics. Sci Transl Med 2(46ps42):46

Biomarkers, Protein Expression

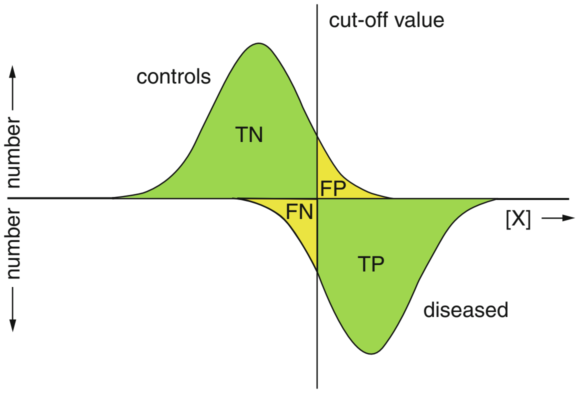
Péter Horvatovich
Department of Pharmacy, Analytical Biochemistry Research Group, University of Groningen, Groningen, The Netherlands

Synonyms

Biomarkers; Independent validation; Protein expression; Solid tissue

Definition

Biomarkers are characteristics that are objectively measured and evaluated as indicators of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Biomarkers Definitions Working Group 2001). In most cases, biomarkers are related to compounds such as proteins or metabolites, whose concentrations are specifically correlated to the biological state(s). Biomarkers are mostly used to diagnose diseases and to monitor disease progression and treatment efficiency. Other physical measures such as MRI measurements, PET images, or imaging mass spectrometry data may be used as biomarkers; however, in most cases, these measures are strongly correlated with the molecular change due to disease. A surrogate marker (Katz 2004) is a laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a



Biomarkers, Protein Expression, Fig. 1 Concentration distribution of hypothetical protein biomarker in group of samples obtained from healthy control group and diseased individuals. The difference of the concentration distribution and the cut-off value define the number of true positive (TN), false positive (FP), true positive (TP) and the false negative (FN)

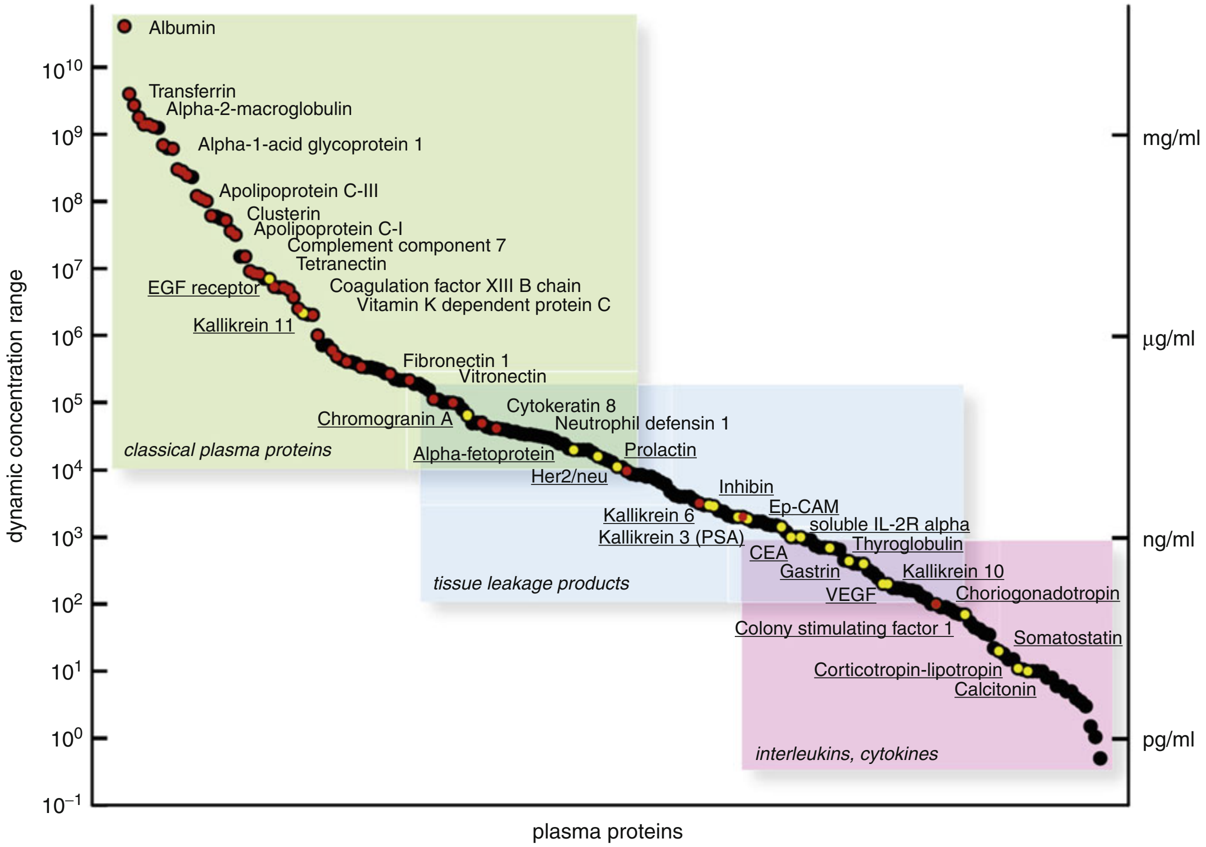
clinically meaningful endpoint, and it is a direct measure of how a patient feels, functions, or survives and is expected to predict the effect of the therapy. The main difference between a surrogate marker and a biomarker is that a biomarker is a meaningful "candidate" for a surrogate marker, and a surrogate marker is a test used as a measure of the effects of a specific treatment.

Proteins are most suitable for biomarkers as their expression is regulated in correlation with the general state of living organisms and contributes considerably to the display of the phenotype. In an experimental context, this means measuring protein concentration distribution in samples obtained at different biological states. Figure 1 shows different concentration distributions of a hypothetical biomarker between two stages, corresponding to the healthy and diseased stages. Overlap between the two concentration distributions and the cut-off value define the Type I (number false positive or FN) and Type II errors (number false negative or FP) and specify the sensitivity and specificity of the test. The area under the receiving operation curve (AUROC) can be used to compare efficiency of different diagnostic tests.

Proteins have a complex life cycle from synthesis on the ribosome to the ubiquitin degradation mechanism. During the life span of a protein it may undergo several chemical, isomerization, and sterical modifications, and modified proteins may perform different biological roles. This may lead to an enormous explosion in the number of protein forms, and sophisticated analytical methods are required to distinguish between the different forms. The situation is further complicated because the activity of proteins may be influenced by many factors, such as cofactors, formation of protein complexes, local pH conditions, and the presence of natural inhibitors, and often in is not the concentration but the activity of the protein that is in strong relation with the disease state. This has lead to the development of analytical methods providing activity profiles of classes of proteins (Fig. 2).

For diagnosis and treatment, follow-up biomarkers are mostly detected from easily accessible body fluids such as blood, urine, and saliva or from the cerebrospinal fluid. However, in body fluids, the difference between the concentration of the least and most abundant compounds is large (11–12 orders of magnitude in the case of human blood) (Schiess et al. 2009). Interesting biomarker candidates originating from the diseased tissue(s) or organ(s) are mixed with compounds from all other parts of the body, providing a huge challenge for detection by analytical chemistry. Other biomarker discovery approaches first identify biomarker candidates directly in the diseased cells, tissues, or organs, followed by subsequent validation of the identified candidate compounds or their specific metabolites if they are present in sufficient amount in easy accessible body fluids, where the final diagnostic test is applied.

The most widely used techniques are antibody-based ELISA tests, protein arrays, and specific LC-MS analysis. Research to identify new biomarkers is composed from biomarker discovery process, where a high number of proteins are identified in low number of samples using comprehensible quantitative analytical techniques such as protein arrays, LC-MS, or SELDI measurements. This is followed by identification of the most discriminating peaks between predefined class of samples (e.g., healthy and disease) and validation of the biomarkers on large number of samples using fast, targeted analytical methods such as MRM-based LC-MS or ELISA tests. The final validation comprises the exploration of the role of the biomarker compound(s) in the mechanism of the disease.



Biomarkers, Protein Expression, Fig. 2 Plasma protein concentration showing three main categories (classical plasma proteins, tissue leakage products, signaling compounds e.g., cytokines). *Red dots* indicate proteins that were identified by the HUPO plasma proteome initiative and *yellow dots* represent currently utilized biomarkers (Figure taken from Schiess et al. (2009))

Affinity-based protein arrays or two-dimensional gel electrophoresis with mass spectrometric identification are able to quantify whole intact proteins. Due to the difficulties to separate intact proteins with reverse-phase chromatography compatible with MS detection, LC-MS methods use protein fragments obtained with cleavage of protease enzymes such as trypsin. In this approach, which is also called as shotgun proteomics, the better chromatographic separation is at the expense of obtaining higher sample complexity, which results in difficulty to obtain exact quantification of the intact protein in the presence of proteins with high homology having few peptides in common or proteins having partly different type of post-translation modifications. Quantification of compound using mass spectrometry can be also performed in several ways. Methods without using any chemical modifications are called label-free, and methods using chemical reaction or incorporation of amino acids with non-natural isotope compositions form one other type of protein quantification method.

References

Biomarkers Definitions Working Group (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 69(3):89–95

Katz R (2004) Biomarkers and surrogate markers: an FDA perspective. NeuroRx 1(2):189–195

Schiess R, Wollscheid B, Aebersold R (2009) Targeted proteomic strategy for clinical biomarker discovery. Mol Oncol 3(1):33–44

Biomarkers, Ranking

Ronnie Alves
Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

Synonyms

Gene ranking; Order statistics; Transcriptomic data analysis

Definition

Biomolecular markers (Biomarkers) include altered or mutant genes, RNA, proteins, lipids, carbohydrates, small metabolites molecules, and changed expression states of such markers that can be correlated with biological behavior or a clinical outcome. Most of the discovered biomarkers are based on changes in gene expression patterns from profiling (Transcriptomic) studies (Ewens and Grant 2004).

In order to make a clear definition of the term "biomarkers" solely based on gene expression data, let us define gene expression data as a pair GED = (g,c), where the $m \times n$ matrix $g = (g_{ij})_{i\ =\ 1,...,m;\ j\ =\ 1,...,n}$ contains m observations of the random vector $(G_1,...,G_n)$ (as an example, the expression levels of m genes), and $c = (c_1,...,c_m)$ stores either an experimental condition C fixed by design or the response variable of interest for those m observations. Let us define a ranking of the variables $G_1,...,G_n$ as a permutation $r = (r_j)_{j\ =\ 1,...,n}$ of $(1,...,n)$, where r_j is the rank of the variable G_j with respect to its association between the considered gene and C, either positive or negative. A ranking provides an ordered gene list $gl = (gl_k)_{k\ =\ 1,...,n}$ defined by

$$\mathrm{Gl_k} = \mathrm{j} \Leftrightarrow \mathrm{r_j} = \mathrm{k} \text{ for all } \mathrm{j}, \mathrm{k} = 1, \ldots, \mathrm{n}. \qquad (1)$$

Taking the differential expression of a gene G_{54}, the variables $\mathrm{r}_{54} = 1$ and $\mathrm{gl}_1 = 54$ mean that G_{54} is pointed as the most differentially expressed gene. Since gl is an ordered list, the k top genes in the list $\mathrm{gl}_1,...,\mathrm{gl}_\mathrm{k}$ form the Top-K gene list (k << n). Biomedical reports usually report gene lists ranging from the Top-10 to the Top-50 (Lockhart and Winzeler 2000; Parmigiani et al. 2003).

References

Ewens WJ, Grant GR (2004) Statistical methods in bioinformatics: an introduction (statistics for biology and health), 2nd edn. Springer, New York

Lockhart DJ, Winzeler EA (2000) Genomics, gene expression and DNA arrays. Nature 405(6788):827–836

Parmigiani G, Garett ES, Irizarry RA (2003) The analysis of gene expression data. Springer, New York

Biomarkers, Solid Tissue

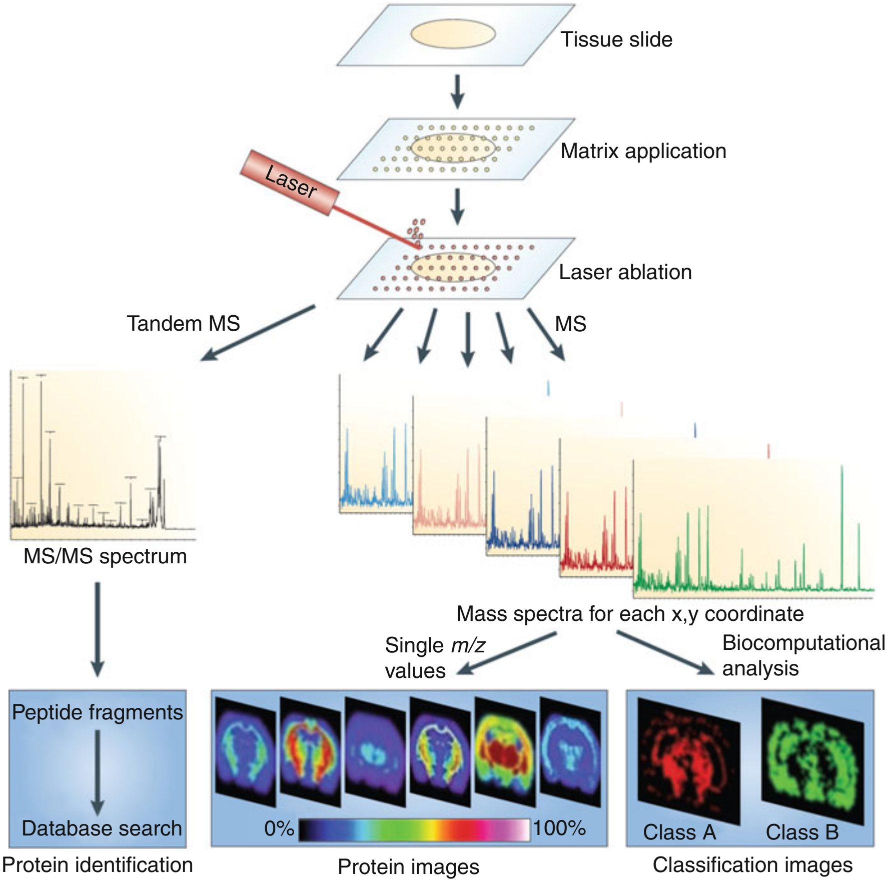
Péter Horvatovich
Department of Pharmacy, Analytical Biochemistry Research Group, University of Groningen, Groningen, The Netherlands

Synonyms

Biomarkers; Protein expression

Definition

Biomarker discovery approaches based on case-control studies of solid tissues or organs are best placed to provide compounds with a close relation to the disease. However, diagnostic tests are generally applied to easy accessible body fluids such as blood, urine, saliva, or cerebrospinal fluid, and, in most cases, sampling solid tissues and organs affected by pathological alterations is too invasive for the patient. Therefore, compounds strongly associated with the disease indentified in solid tissues and organs need to find their way directly or indirectly to the body fluid sampled for the diagnostic test. A further complication is that such compounds, after being transported to the target body fluid, are surrounded by a large number of other compounds having a large dynamic



Biomarkers, Solid Tissue, Fig. 1 Workflow for MALDI imaging mass spectrometry analysis. Schematic outline of a typical workflow for fresh frozen tissue samples. Sample pre-treatment steps include cutting and mounting the tissue section on a conductive target. Matrix is applied in an ordered array across the tissue section and mass spectra are acquired at each x, y coordinate. Single stage mass spectra are used to found discriminating between diseased and healthy area with statistical methods. Tandem MS (MS/MS) spectra are used for peptide and protein identification. Further data analysis steps include the visualization of the distribution of a single peptide or protein within the tissue or to visualize the image of discriminating peaks. The *scale* represents the relative intensity of the protein (Figure taken from ref (Schwamborn and Caprioli 2010))

concentration range (often 11–12 orders of magnitude). This presents an enormous challenge for analytical chemistry and statistical methods to select compounds with a real association with the disease. Tissue- or organ-derived discriminating compounds should therefore be validated so that they reach peripheral body fluids either in intact or modified forms. One promising approach is the measurement of the secretome molecular profile of in vitro cultivation of removed tissue samples. Generally, samples obtained from solid tissues are pure, but inevitable contamination from blood may result in alteration of the measured tissue secretome molecular profile.

The most widely used sampling method for solid tissues is laser capture microdissection (Murray and Curran 2005; Liu 2010), which enables highly accurate cutting of diseased altered and healthy tissue pieces using precise laser shots. This technique enables case-control sampling of the same tissue and therefore excludes the majority of the biological variance from the analysis. However, disease cells may influence the surrounding healthy cells, e.g., with proteins secreted by the disease-altered cells. Therefore, validation by comparing disease-altered cells, the surrounding healthy tissue in diseased tissue, and healthy cells in healthy samples is required.

MALDI imaging mass spectrometry (Schwamborn and Caprioli 2010) is one other promising technique for finding tissue-derived compounds closely related to the disease. It enables direct in situ or cross-sample comparison of the molecular profile of diseased and healthy cells in tissue sections (Schwamborn and Caprioli 2010). Figure 1 presents the outline of MALDI imaging mass spectrometry analysis from tissue section preparation to the statistical analysis of the acquired data.

References

Liu A (2010) Laser capture microdissection in the tissue biorepository. J Biomol Tech 21(3):120–125

Murray GI, Curran S (2005) Laser capture microdissection: methods and protocols, Methods in Molecular Biology series, vol 293. Humana Press, Totowa

Schwamborn K, Caprioli RM (2010) Molecular imaging by mass spectrometry–looking beyond classical histology. Nat Rev Cancer 10(9):639–646

Biomedical Decision Support Systems

Guy Tsafnat
Centre for Health Informatics, Australian Institute for Health Innovation, University of New South Wales, Sydney, NSW, Australia

Synonyms

Clinical decision support systems

Definition

A biomedical decision support system (DSS) is any computational system that aids decisions made by clinicians, medical researchers, and medical biologists when interpreting large amounts of information relevant to a particular decision. Some DSS may support decision making by retrieving (only) the information required to make the decision, while others also summarize the information and/or predict the outcomes of decisions.

Examples of biomedical decision support systems are:

- Information retrieval systems that help clinicians find clinical guidelines accurately and quickly
- Intelligent systems that suggest diagnoses based on patient symptoms
- Simulation systems that calculate the outcomes of complex biochemical pathway to predict treatment outcomes
- Data analysis systems that summarize, interpret, and visualize hundreds of microarray assays identifying genes that are under- or over-expressed in cancer tissue

Biomedical Named Entity Recognition, Whatizit

Dietrich Rebholz-Schuhmann
European Bioinformatics Institute, Hinxton, UK

Definition

Named entities in the biomedical research domain comprise genes and proteins, diseases, species, chemical entities, anatomical components, and other semantic types. A large number of biomedical named entities have to be included into text-mining solutions due to the descriptive nature of biomedical research.

Introduction

Ready access to text-mining solutions requires the integration of various resources and specialized technologies into an open access infrastructure. In the past, solutions have been proposed that incorporate these resources into standalone applications (Friedman et al. 2001; Kano et al. 2009). Unfortunately, such solutions have an architecture that does not support well integration of bioinformatics services similar to open IT solutions such as Taverna (Hull et al. 2006). Other systems such as iHOP provide special interfaces for programmatic access, but iHOP does not allow to process other documents than Medline abstracts with new means (Hoffmann et al. 2005).

A Web service–based TM solution centralizes and harmonizes crucial tasks and thus solves a number of difficulties reducing maintenance for users. A server-based solution can incorporate large terminology sets from biomedical data resources: updates to these resources are efficiently propagated through the server. The end user profits from a harmonized schema, including coupling of text processing services to bioinformatics data resources.

Implementation

Whatizit is a modular infrastructure that delivers TM services to the public. Each module processes and annotates text, for example, identifies named entities and introduces links to database entries. Individual modules can be composed of a number of internal modules. All terminologies are based on publicly available resources (e.g., UniProtKb/Swiss-Prot, gene ontology, DrugBank, see below). Terms are matched to the text taking morphological variability into consideration (Kirsch et al. 2006).

The user submits the name of a pipeline and the text to be annotated, and Whatizit returns the text with all the annotations contained. As an alternative, the user can access Whatizit services without building a client application. On the Web interface, Whatizit provides a text input area where user can submit any kind of Unicode encoded text or retrieve Medline abstracts for subsequent analysis.

Different types of modules are available through the Whatizit infrastructure. One set of modules annotates named entities. *WhatizitChemical* searches for chemical entities based on the terminology from ChEBI and the identification of chemical terms by OSCAR3 (Corbett et al. 2006). *WhatizitDisease* identifies disease terms using a controlled vocabulary (CV) extracted from MedlinePlus, whereas *whatizitDiseaseUMLS* allows access to MetaMap (Aronson 2001). For *whatizitDrugs*, the CV has been extracted from DrugBank (http://redpoll.pharmacy.ualberta.ca/drugbank/). *WhatizitGO* is a pipeline searches for gene ontology terms using exact matching and considering morphological variability (Ashburner et al. 2000). Last, *whatizitOrganism* identifies species names extracted from the NCBI taxonomy (NLM).

Other annotation pipelines represent solutions that are more complex, i.e., they identify combinations of semantic types: for example, *whatizitSwissprotGo* for UniProtKb/Swiss-Prot in conjunction with GO annotations and *whatizitEbiMed* for the annotation pipeline from EbiMed (Rebholz-Schuhmann et al. 2007). The retrieval engine for Medline abstracts is accessible via the module *whatizitQbmarsdf*. For the retrieval, the user has to submit query terms or PubMed IDs.

Conclusion

Whatizit is a service that copes with large terminological resources, is aligned with updates from the primary resource, and is available through a centralized service that scales with the amount of integrated resources, with the demands of different extraction methods and with the amount of literature processed over time.

Acknowledgments This work is funded by the Network of Excellence "Semantic Mining" (NoE 507505). Medline abstracts are provided from the National Library of Medicine (NLM, Bethesda, MD, USA). Sylvain Gaudan is supported by an "E-STAR" fellowship (EC's FP6 Marie Curie Host fellowship, MESTCT-2004-504640).

Cross-References

- ▶ Applied Text Mining
- ▶ BioCreative Meta-Server and Text-Mining Interoperability Standard
- ▶ Bio-Ontologies
- ▶ Information Extraction
- ▶ Named Entity Recognition
- ▶ Natural Language Processing
- ▶ Ontology Lookup Service for Controlled Vocabularies and Data Annotation

References

Aronson AR (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium, vol 5. Hanley & Belfus, Philadelphia, pp 17–21

Ashburner M et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29

Corbett P, Murray-Rust P (2006) High-throughput identification of chemistry in life science texts. In: Proceedings of the CompLife. Lecture notes in bioinformatics, vol 4216. Cambridge, pp 107–118

Friedman C et al (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 17(Suppl 1):S74–S82

Gaizauskas R et al. (1996) GATE: an environment to support research and development in natural language engineering. In: Proceedings of the 8th IEEE international conference on tools with artificial intelligence, Toulouse, pp 58–66, ISBN 0-8186-7686-8

Hatcher E, Gospodnetic O (2004) Lucene in action. Manning, Greenwich

Hirschman L et al (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics 6(Suppl 1):S11

Hoffmann R, Valencia A (2005) Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21(Suppl 2):ii252–ii258

Hull D et al (2006) Taverna: a tool for building and running workflows of services. Nucleic Acids Res 34:W729–W732, Web server issue

Kano Y, Baumgartner WA Jr, McCrohon L, Ananiadou S, Cohen KB, Hunter L, Tsujii J (2009) U-Compare: share and compare text mining tools with UIMA. Bioinformatics 25(15):1997–1998

Kirsch H et al (2006) Distributed modules for text annotation and IE applied to the biomedical domain. Int J Med Inform 75(6):496–500

Rebholz-Schuhmann D et al (2006a) Annotation and Disambiguation of semantic types in biomedical text: a cascaded approach to named entity recognition. In: Workshop on "Multi-dimensional markup in NLP", EACL 2006. ACL, USA

Rebholz-Schuhmann D et al (2006b) IeXML: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules. SIG BioLink, ISMB 2006, Fortaleza, Brasil

Rebholz-Schuhmann D et al (2007) EBIMed–text crunching to gather facts for proteins from medline. Bioinformatics 23(2): e237–e244

Schmid H (1994) Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the international conference on new methods language processing, vol 12, Manchester

Biomedical Natural Language Processing (BioNLP)

► Natural Language Processing

Biomedical Ontologies

► Bio-Ontologies

Biomedical Web Communities

► Collaborative and Distributed Biomedical Applications

BioModels

► BioModels Database: A Repository of Mathematical Models of Biological Processes

BioModels Database: A Repository of Mathematical Models of Biological Processes

Vijayalakshmi Chelliah[1], Camille Laibe[2] and Nicolas Le Novère[3]
[1]EMBL European Bioinformatics Institute, Hinxton, Cambridgeshire, UK
[2]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, UK
[3]EMBL European Bioinformatics Institute and Babraham Institute, Cambridge, UK

Synonyms

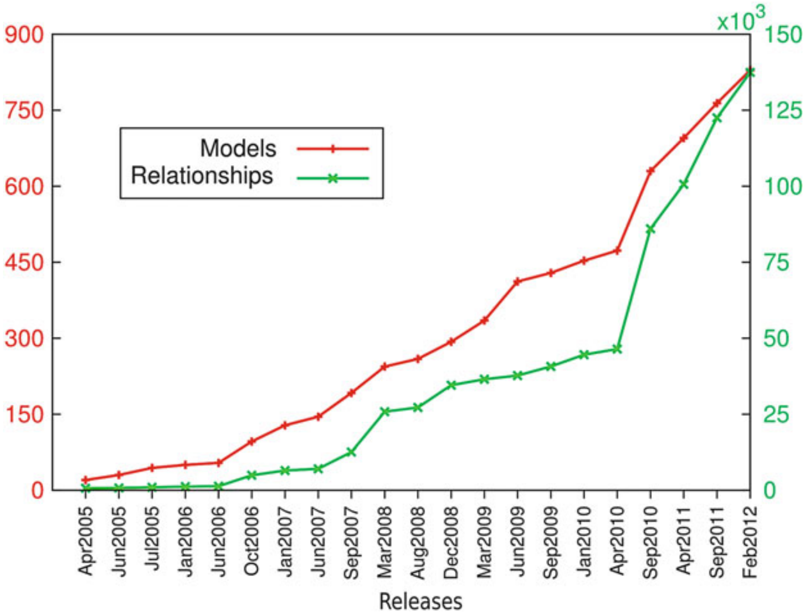
BioModels; Model repository

Definition

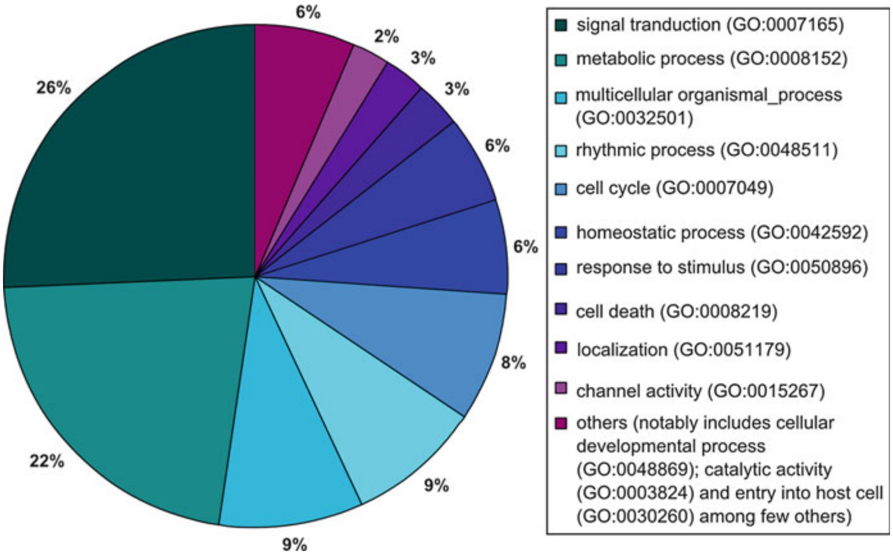
BioModels Database is a public online resource that allows storing and sharing of published, peer-reviewed quantitative, dynamic models of biological processes. The model components and behavior are thoroughly checked to correspond the original publication and manually curated to ensure reliability. Furthermore, the model elements are annotated with terms from controlled vocabularies as well as linked to relevant external data resources. This greatly helps in model interpretation and reuse. Models are stored in SBML format, accepted in SBML and CellML formats, and are available for download in various other common formats such as BioPAX, Octave, SciLab, VCML, XPP and PDF, in addition to SBML. The reaction network diagram of the models is also available in several formats. BioModels Database features a search engine which provides simple and more advanced searches. Features such as online simulation and creation of smaller models (submodels) from the selected model elements of a larger one are provided. BioModels Database can be accessed both *via* a web interface and programmatically *via* web services. New models are available in BioModels Database at regular releases, about every 4 months.

Characteristics

BioModels Database (http://www.ebi.ac.uk/biomodels/) (Le Novère et al. 2006; Li et al. 2010a) hosts a collection



BioModels Database: A Repository of Mathematical Models of Biological Processes, Fig. 1 *Growth of BioModels Database:* The total number of models (*red*) and the total number of reactions (*green*) in all models are plotted here. The number of relationships includes SBML "reactions," "rate rules," "assignment rules," and "events." There has been approximately a 20-fold increase in the number of models since the launch of the resource in 2005, with an average increase in complexity of the models (measured by the number of mathematical relationships) being increased five times in the same period



BioModels Database: A Repository of Mathematical Models of Biological Processes, Fig. 2 *Type of models:* Categorization of models in the curated branch of BioModels Database based on the GO terms present in the annotation of the models. This chart was generated by enumerating models in the database, whose annotations refer either the GO terms listed here or the child of the GO terms listed here

of mathematical models of biological processes, described in peer-reviewed scientific literature. It is part of the BioModels.net initiative (http://biomodels.net/) (Le Novère 2006), which aims to help researchers in the modeling field to exchange and build upon each other's work with greater ease and accuracy.

It has significantly grown both in size and number of the models since its origin in 2005 (Fig. 1) and serves as an efficient model sharing platform.

Diversity of Models Hosted

BioModels Database covers a wide range of models from several biological categories. It hosts models from simple biochemical reaction systems to larger and complex dynamic models, metabolic network models, and FBA models. Figure 2 represents the categorization of models in the curated branch using terms from Gene Ontology (GO) present in the model annotation.

Models Provenance

Models are implemented from articles published in peer-reviewed scientific journals by a team of curators. Also, there are an increasing number of models which are directly submitted by the modelers themselves and this demonstrates the popularity and recognition of BioModels Database in the community. In the past, some models came from collaboration with other repositories, such as the former SBML model repository (Caltech, USA), JWS Online, (http://jjj.biochem.sun.ac.za/) the Database Of Quantitative Cellular Signaling (DOQCS), (http://doqcs.ncbs.res.in/) and the CellML Repository (http://models.cellml.org/cellml). Several scientific journals recommend deposition of models to BioModels Database in their instructions for authors. These include journals that are published by Nature Publishing Group (NPG), Public Library of Science (PLoS), Royal Society of Chemistry (RSC), and BioMed Central (BMC).

Model Submission, Curation, Annotation, and Publication

Submission of models to BioModels Database is free and is open to everyone. Models can be submitted *via* an online interface and are accepted in two formats, SBML (Systems Biology Markup Language) (http://sbml.org/) and CellML. While BioModels Database only distributes models that have been described in the peer-reviewed scientific literature, models can be submitted prior to the publication of their associated paper. However, these models are made publicly available only after the publication of their corresponding paper. At the time of submission, each model is assigned a unique and perennial identifier which allows users to access and retrieve it. This identifier can be used by authors as a reference in their publications.

The models that are submitted pass through several steps prior to get published. BioModels Database is composed of a curated and a non-curated branch. Models in both these branches are fully SBML compliant. Depending on their curation status, models are moved to one of the two branches. Models that satisfy the MIRIAM (Minimum Information Required in the Annotation of Models) guidelines (Le Novère et al. 2005) progress to the curated branch. These models are thoroughly checked and corrected for accuracy as they must match and reproduce the results published in the reference publication. A representative figure or table reproduced by the model (which is present in the reference publication) together with a description of how it was obtained are available for each model. This ensures that the encoded form of the model that is provided corresponds to what was described in the paper.

There are several reasons for models to be in the non-curated branch. These are models that either do not satisfy the full requirements for MIRIAM compliance or have not been curated yet due to limited time and resources. Many of these models are pathway maps or models for networks and Flux Balance Analysis (FBA), without sufficient quantitative results provided for validation. Some others are only the subsets of the whole model described in the article, as some parts cannot be encoded in SBML. And finally, a small number of models could not be made to reproduce the published results due to untraceable errors in the implementation or typos in the publication (often even after contacting the authors of the article).

All model elements in the curated branch are furthermore thoroughly annotated with cross-references to other database records and ontology terms. The annotations are included in the models using MIRIAM URIs (Juty et al. 2012). As model elements are not always named precisely to relate directly to the corresponding biological processes or physical entity, annotations are necessary to enhance interpretability by both users and software tools. To date, model elements in BioModels Database are annotated using around 40 different external data resources. Some of the predominantly used external resources for model annotations are Gene Ontology, ChEBI ontology, Brenda Tissue Ontology, Systems Biology Ontology (SBO), Taxonomy, Reactome, KEGG, and UniProt. A collection of data resources and their URIs can be obtained from MIRIAM Registry (http://www.ebi.ac.uk/miriam/) (Laibe and Novère 2007).

Once the curation and annotations are completed, the model is tagged as ready for publication, and becomes available online during the next release of BioModels Database. New releases happen two to four times a year.

Model Browsing, Searching, and Retrieval

There are several features available through the web interface which facilitate efficient usage of the models. They can be browsed by branches. They can also be located by using a tree-structured browser based on the Gene Ontology (GO) terms used in the annotation of the models. BioModels Database incorporates a powerful search engine that allows users to retrieve models of their interest. The search can either be a simple keyword search or an advanced one.

Model Display

Each model in the curated branch is presented in a tabbed form, providing access to all the information stored about the model. The model elements are hyperlinked between its entries in different tabs and in addition, the annotations are hyperlinked to their detailed resource page. The detailed description about the model is separated into six categories namely: *Model*, *Overview*, *Math*, *Physical entities*, *Parameters*, and *Curation*, which are all accessible *via* a dedicated tab.

Model Exports

For convenience, as certain simulation tools support only specific Levels or Versions of SBML, the *Download SBML* menu allows users to download the model in various versions of SBML, including the version that was checked by the curators. Models can be downloaded in various other formats such as BioPAX, (http://www.biopax.org/) the Virtual Cell Markup Language (VCML), (http://vcell.org/) XPPAUT, (http://www.math.pitt.edu/~bard/xpp/xpp.html) SciLab, (http://www.scilab.org/) Octave (m-file), (http://www.gnu.org/software/octave/) and PDF (generated using the SBML2LaTeX tool) through the "*Other formats (auto-generated)*" menu.

The reaction network of the model that follows the Systems Biology Graphical Notation (SBGN) is available in PNG and SVG formats *via* the *Action* menu. The reaction networks are also provided in a dynamic way *via* an interactive Java applet.

The *Actions* menu also provides access to the online simulation tools. BioModels Database embeds SOSlib (Machné et al. 2006) to provide a basic online simulation tool. The simulation results are returned both in graphical and textual form. For many models, an additional and more flexible simulation tool is available using JWS Online.

The *Action* menu provides an additional feature for some model, called "*Model of the Month*" which is a brief article that discusses the biological background, significance, structure, and results of the model. The "*Model of the Month*" is also accessible through BMC Systems Biology Gateway.

Web Services

For programmatic access, BioModels Database features web services (http://www.ebi.ac.uk/biomodels-main/webservices) (Li et al. 2010b), allowing, for example, direct retrieval of complex searches for models, and the creation of submodels. The available services are described in a Web Services Description Language (WSDL) file that enables software to understand available functions and their usage. The web services use the Simple Object Access Protocol (SOAP) to encode requests and responses. The complete list of available methods, as well as a Java library and the associated documentation, is provided on the BioModels Database website.

BioModels Database is developed under the GNU General Public License and the software is freely available from its SourceForge repository (http://sourceforge.net/projects/biomodels/).

Usage

BioModels Database has significantly grown both in size and number of the models since its origin in 2005 and serves as an efficient model sharing platform. BioModels Database announced its 21st release on 8 February 2012, with a total of 829 models provided. The submission of models by modelers/authors themselves is increasing rapidly, and this demonstrates the popularity and recognition of BioModels Database in the community. The resource helps modelers to reuse already existing models or model components, to modify them by implementing their own theory, to publish articles describing new models, and submit those new models to BioModels Database. For example, BIOMD0000000176 and BIOMD0000000177 are derived from BIOMD0000000172 which in turn is derived from BIOMD00000000064. BioModels Database is also used as a source of trusted models for benchmarking model simulation software packages.

BioModels Database has the potential to serve as a comprehensive repository for computational systems biology models, similar to the functionality of GenBank and Protein Data Bank (PDB), the data resources for genes and protein three-dimensional structures.

Cross-References

- ▶ MIRIAM Guidelines
- ▶ MIRIAM URI
- ▶ SBGN
- ▶ Systems Biology Markup Language (SBML)
- ▶ Systems Biology Ontology

References

Juty N, Le Novère N, Laibe C (2012) Identifiers.org and MIRIAM registry: community resources to provide persistent identification. Nucleic Acids Res 40:D580–D586

Laibe C, Novère NL (2007) MIRIAM resources: tools to generate and resolve robust cross-references in systems biology. BMC Syst Biol 1:58

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

Le Novère N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro B, Snoep JL, Spence HD, Wanner BL (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). Nat Biotechnol 23:1509–1515

Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res 34: D689–D691

Li C, Courtot M, Le Novère N, Laibe C (2010a) BioModels.net Web Services, a free and integrated toolkit for computational modelling software. Brief Bioinform 11:270–277

Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, Li L, He E, Henry A, Stefan MI, Snoep JL, Hucka M, Le Novère N, Laibe C (2010b) BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol 4:92

Machné R, Finney A, Müller S, Lu J, Widder S, Flamm C (2006) The SBML ODE solver library: a native API for symbolic and fast numerical analysis of reaction networks. Bioinformatics 22:1406–1407

Biomolecular Interaction Networks

- ▶ Interaction Networks

BioNLP Shared Task

Jin-Dong Kim[1] and Sampo Pyysalo[2]
[1]Database Center for Life Science, Tokyo, Japan
[2]Department of Computer Science, University of Tokyo, Tokyo, Japan

Definition

BioNLP Shared Task (BioNLP-ST, hereafter) is a series of shared evaluations and workshops focused on biomolecular event extraction from literature. The first BioNLP-ST evaluation was organized in 2009 by the Tsujii Laboratory of the University of Tokyo, with a workshop held under the auspices of Biomedical Natural Language Processing Special Interest Group (SIGBIOMED) of the Association for Computational Linguistics (ACL). The task targeted the extraction of *biomolecular events* (*bio-events*), defined as *changes in the states of biomolecular objects* following the definition and event representation of the GENIA project. The task data was based on the GENIA corpus (Kim et al. 2008) human-curated annotations of bio-events in PubMed abstracts. BioNLP-ST 2009 was the first community-wide effort for the automatic extraction of bio-events from literature (Kim et al. 2009).

The event extraction task was quite new to much of the community, as most bio-text mining (bio-TM) targeted simple binary relationships between bio-entities, such as protein–protein interactions (PPI) (Bunescu et al. 2004) and disease-gene associations (DGA) (Chun et al. 2006). However, BioNLP-ST'09 met community-wide participation, with 42 teams signing up for initial registration and 24 teams submitting final results.

At the time of writing, BioNLP-ST 2011, the second event of the series, is being organized as a joint effort of several groups. The evaluation seeks to provide a variety of tasks and annotations centered around event extraction.

Characteristics

The MUC (1987–1997) (Chinchor 1998), TREC (1992–) (Voorhees 2007), and ACE (1999–) (Strassel et al. 2008) shared evaluation initiatives have significantly motivated research in information retrieval (IR), information extaction (IE), and text mining (TM) for general domain text. There have been similar efforts organized for bio-IR, -IE or -TM. The TREC Genomics track (2004–2007) (Hersh et al. 2007) was organized to address bio-IR, and the JNLPBA shared task (2004) (Kim et al. 2004) to address named entity recognition (NER). (Organized by the GENIA project using the same corpus, JNLPBA can be regarded a predecessor to the BioNLP-ST series.) The LLL challenge (2005) (Nédellec 2005) focused on interactions of proteins or genes (IE), and BioCreative (2005–) (Hirschman et al. 2007) addressed a variety of tasks, including named entity normalization and protein–protein interactions.

While LLL ran a typical relation extraction task to find interacting pairs of proteins or genes, BioCreative and BioNLP-ST have taken definitive steps forward, although in different directions: BioCreative toward user-oriented task settings and extrinsic evaluation and BioNLP-ST toward fine-grained IE. The difference in direction is motivated in part by different applications envisioned as being supported by the IE methods. For example, BioCreative aims to support curation of PPI databases, such as MINT (Chatr-aryamontri et al. 2007), for a long time one of the primary tasks in the domain. BioNLP-ST aims to support the development of more detailed and structured databases, e.g., pathway (Bader et al. 2006) or Gene Ontology Annotation (GOA) (Camon et al. 2004) databases, which are gaining increasing interest in bioinformatics research in response to recent advances in molecular biology.

BioNLP-ST 2009

The first event of the series, BioNLP-ST 2009 (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/), had the following characteristic features which made it different from and complementary to other shared evaluation efforts.

1. Events were represented as typed *n*-ary associations of proteins and other events in various roles, allowing complex, structured associations of multiple entities to be captured.
2. The event extraction task was divided into three subtasks: core event extraction, event enrichment, and negation and speculation recognition.
3. Detailed evaluations at each subtask were provided.
4. Human-curated corpus annotations corresponding to each subtask were prepared for system training, tuning, and evaluation purposes.
5. The basic IE task of named entity recognition was excluded to encourage the participants to concentrate on event extraction, the novel challenge of the task. Accordingly, the gold standard annotations for protein references were provided to the participants also for test data.
6. Publicly available natural language processing (NLP) tools adapted for the BioNLP-ST data format were provided in readily available packages to encourage the adoption of NLP tools and to further allow participants to focus on event extraction.

Task Definition

Table 1 shows the event types targeted at BioNLP-ST 2009. The event types are selected from the GENIA ontology, with consideration given to their importance and the number of annotated instances in the GENIA corpus. The selected event types all concern protein biology, implying that they take proteins as their theme. The first three types concern protein metabolism, i.e., protein production and breakdown. The fourth, Phosphorylation, is one type of protein modification, selected because it appears frequently in the GENIA corpus. Localization and Binding are representative fundamental molecular events. Regulation (including its subtypes, Positive and Negative_regulation) represents regulatory events and general causal relations. The last five event types are universal, but frequently occur on proteins. For the biological interpretation of the event types, readers are referred to Gene Ontology (GO) and the GENIA ontology.

As shown in Table 1, the theme or themes of all events are considered primary arguments, that is, arguments that are critical to identifying the event. For regulation events, the entity or event stated as the *cause* of the regulation is also regarded as a primary argument. For some event types, further arguments detailing the events are also defined (*Secondary Arg.* in Table 1). From a computational point of view, the event types represent different levels of complexity. When only primary arguments are considered, the first five event types require only a single argument, and the task can be cast as binary relation extraction between

BioNLP Shared Task, Table 1 The event types and their arguments targeted at BioNLP-ST 2009. Arguments that may be filled more than once per event are marked with "+." *Site* means a site of a theme entity, and *CSite* a site of a cause entity

Type	Primary arguments	Secondary Arg.
Gene_expression	Theme(Protein)	
Transcription	Theme(Protein)	
Protein_catabolism	Theme(Protein)	
Phosphorylation	Theme(Protein)	Site
Localization	Theme(Protein)	AtLoc, ToLoc
Binding	Theme(Protein)+	Site+
Regulation	Theme(Protein/Event), Cause(Protein/Event)	Site, CSite
Positive_regulation	Theme(Protein/Event), Cause(Protein/Event)	Site, CSite
Negative_regulation	Theme(Protein/Event), Cause(Protein/Event)	Site, CSite

a predicate (event trigger) and an argument (Protein). The Binding type is more complex in requiring the detection of an arbitrary number of arguments. Regulation events always take a Theme argument and, when expressed, also a Cause argument. Further, a Regulation event may take an event as its primary argument (theme or cause), creating structures linking multiple events that represent causal chains. Such connections between events are a unique feature of the BioNLP task compared to other event extraction tasks such as ACE.

The failure of p65 translocation to the nucleus ...

T1 (Protein, 15-18)
T2 (Localization, 19-32)
E1 (Type:T2, Theme:T1, ToLoc:T3)
T3 (Entity, 40-46)
M1 (Negation E1)

BioNLP Shared Task, Fig. 1 Example event annotation. The protein annotation T1 is given as a starting point. The extraction of annotation in bold is required for Task 1, the *Entity* type t-entity *T3* and the secondary argument *ToLoc:T3* for Task 2, and *M1* for Task 3

Format and Example

In the BioNLP-ST data sets, annotations to text are provided in a stand-off style, as shown in the example in Fig. 1. The annotations are of two general types: ones identifying text spans referring to bio-entities (e.g., T1 and T3) and event triggers (e.g., T2), and ones expressing events (e.g., E1) and their modifications (e.g., M1). The former type of annotations are represented by pairs, (type-specification, span-specification), and the latter by *n*-tuples resembling predicate–argument structures, (predicate, argument1, argument2, etc.).

Results

The final results enabled to observe the state-of-the-art performance of the community on the bio-event extraction task. It showed that the automatic extraction of simple events – those with unary arguments, e.g., gene expression, localization, phosphorylation – could be achieved at the performance level of 70% in F-score, but extraction of complex events, e.g., binding and regulation, was a lot more challenging, achieving 40% of performance level.

Continuation

After BioNLP-ST 2009, all the resources from the event were released to public, to encourage continuous efforts for further advancement. Since then, several improvements have been reported (Björne et al. 2010; Miwa et al. 2010a, b; Poon and Vanderwende 2010; Vlachos 2010). For example, Miwa et al. (2010b) reported a significant improvement with binding events, achieving 50% of performance level.

BioNLP-ST 2011

The second event of BioNLP-ST series is organized for 2011 (https://sites.google.com/site/bionlpst/). While its predecessor, BioNLP-ST 2009, relied on the GENIA corpus which only contained PubMed abstracts on *transcription factors in human blood cells*, the main theme of BioNLP-ST 2011 is *generalization*, which was pursued to three directions: text types, event types, and subject domains. To achieve that, various tasks and annotations were collected through contributions from several groups, and five event extraction tasks were arranged in four tracks.

- GENIA (GE)
- Epigenetics and Post-translational Modifications (EPI)
- Infectious Diseases (ID)
- Bacteria Track
  - Bacteria Biotopes (BB)
  - Bacteria Interaction (BI)

The *GE* task remains similar to the task of BioNLP-ST 2009, to play the role of *pivot* for generalization and to measure the progress of community with the previous task. However, to avoid overfitting to the evaluation data set, the GE task is planned to arrange additional text sets, most of them coming from full paper articles, so that generalization from abstracts to full papers can be measured. The *EPI* task is arranged to evaluate generalization of event types with focus on events relevant to epigenetics and post-translational protein modification without further subdomain restruction. The *ID* and *BB* tasks involve generalization to new domains, ID targeting events relevant to the biomolecular mechanisms of infectious diseases and BB localization events of bacteria.

BioNLP-ST 2011 also includes three supporting tasks: Coreference, Entity relation, and Gene renaming. Although these are not event extraction tasks themselves, according to the analysis on the results of BioNLP-ST 2009, coreference resolution and entity relation detection are expected to play an important role for making a breakthrough in improving event extraction performance. Gene naming has similar features with coreference; thus it is included as a supporting task.

References

Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. Nucleic Acids Res 34(Suppl 1):D504–506

Björne J, Ginter F, Pyysalo S, Tsujii J, Salakoski T (2010) Complex event extraction at PubMed scale. Bioinformatics 26(12):i382–390

Bunescu R, Ge R, Kate RJ, Marcotte EM, Mooney RJ, Ramani AK, Wong YW (2004) Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med 33(2):139–155

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R (2004) The gene ontology annotation (GOA) database: sharing knowledge in uniprot with gene ontology. Nucl Acids Res 32(Suppl 1): D262–266

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the molecular INTeraction database. Nucleic Acids Res 35(Suppl 1):D572–574

Chinchor N (1998) Overview of MUC-7/MET-2. In: Message understanding conference (MUC-7) proceedings, Fairfax

Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, Tsujii J (2006) Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. BMC-Bioinformatics 7(Suppl 3):S4

Hersh W, Cohen A, Lynn R, Roberts P (2007) TREC 2007 genomics track overview. In: Proceeding of the sixteenth text retrieval conference, Gaithersburg

Hirschman L, Krallinger M, Valencia A (eds) (2007) Proceedings of the second BioCreative challenge evaluation workshop. CNIO Centro Nacional de Investigaciones Oncológicas, Madrid

Kim JD, Ohta T, Tsuruoka Y, Tateisi Y, Collier N (2004) Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (JNLPBA), Geneva, pp 70–75

Kim JD, Ohta T, Tsujii J (2008) Corpus annotation for mining biomedical events from literature. BMC Bioinformatics 9(1):10

Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J (2009) Overview of BioNLP'09 shared task on event extraction. In: Proceedings of natural language processing in biomedicine (BioNLP) NAACL 2009 workshop, Colorado, pp 1–9

Miwa M, Pyysalo S, Hara T, Tsujii J (2010) A comparative study of syntactic parsers for event extraction. In: Proceedings of BioNLP'10, Uppsala, pp 37–45

Miwa M, Sætre R, Kim JD, Tsujii J (2010b) Event extraction with complex event classification using rich features. J Bioinform Comput Biol (JBCB) 8(1):131–146

Nédellec C (2005) Learning language in logic – genic interaction extraction challenge. In: Cussens J, Nédellec C (eds) Proceedings of the 4th learning language in logic workshop (LLL05), Bonn, pp 31–37

Poon H, Vanderwende L (2010) Joint inference for knowledge extraction from biomedical literature. In: Proceedings of NAACL-HLT'10, Los Angeles, pp 813–821

Strassel S, Przybocki M, Peterson K, Song Z, Maeda K (2008) Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In: Proceedings of the 6th international conference on language resources and evaluation (LREC 2008), Marrakesh

Vlachos A (2010) Two strong baselines for the BioNLP 2009 event extraction task. In: Proceedings of BioNLP'10, Uppsala, pp 1–9

Voorhees E (2007) Overview of TREC 2007. In: The sixteenth text REtrieval conference (TREC 2007) proceedings, Gaithersburg

Bio-Ontologies

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

Synonyms

Biomedical ontologies; Ontologies for the life sciences

Definition

Bio-ontologies have come to play a crucial role in the dissemination of results across research contexts and in the extraction of inferences and testable hypotheses from available datasets (▶ Data-intensive Research). They are essentially classification tools, whose immediate function is to store, organize, and retrieve data via databases accessible through the Internet. They consist of networks of terms that are linked to existing databases. Each term in the network is precisely defined as describing specific biological entities or processes; it is used as a keyword with which to identify and retrieve datasets that constitute relevant evidence toward the investigation of the entities or processes described by its definition. The terms are related to each other through simple structuring rules, such as X "is_a" Y and X is "part_of" Y.

Bio-ontologies classify datasets in terms of their potential usefulness to research, as documented by previous empirical studies. This is different from a classification based on information about the provenance of data (the place or group in which they were produced), the model organism on which they were obtained, or the instrument with which they were created. Bio-ontologies do incorporate these other types of information through ▶ metadata; yet the key property of datasets annotated through bio-ontologies – the official criterion for data retrieval in this system – remains their association to terms defining their relevance to understanding biological objects and processes.

Because of their descriptive nature, bio-ontologies are often referred to as "representations of biological knowledge" (e.g., Bard and Rhee 2004). They represent the knowledge needed to share resources for further research, which is thus meant to be as universal and basic as possible. Rhee et al. (2006, p. 345) point to the quality of definitions in bio-ontologies as their most important characteristics: "data and knowledge need to be described in explicit and unambiguous ways that must be comprehensible to both human beings and computer programs." Indeed, bio-ontologies are constructed to work across different research contexts, disciplines, and ▶ model organism communities. The Gene Ontology, for instance, was developed to annotate gene products across community databases in ▶ model organism biology. To guarantee the interoperability of bio-ontologies across biological domains, the Open Biomedical Ontologies Consortium has proposed a set of simple rules for the development of ontologies, such as, for instance, the rule of univocity: Terms should mean the same wherever they are used, so the same term cannot be used to indicate two different processes (Smith et al. 2007). Also, bio-ontology terms and relations are selected and defined through consensus-seeking mechanisms implemented on a global scale, such as consultations and workshops with prospective users (Leonelli 2008).

Characteristics

Bio-ontologies are an achievement of the bioinformatic effort toward an efficient organization and distribution of data produced by genomic research. They provide a framework through which heterogeneous sets of biological data can be classified, stored, and retrieved through freely available, online databases (Rubin et al. 2008). For the purposes of this entry, I restrict my examination of bio-ontologies to the ones collected by the Open Biomedical Ontologies Consortium (http://www.obofoundry.org), an organization founded to facilitate communication and coherence among bio-ontologies with broadly similar characteristics (Ashburner et al. 2003), and particularly to the Gene Ontology, which is widely regarded as the most successful case of bio-ontology construction to date and used as a template for several other prominent bio-ontologies (Ashburner et al. 2003; Brazma et al. 2006).

Structure

Bio-ontologies have three defining features: the use of precisely defined terms to refer to biological entities or processes; the use of precisely defined relations among terms; and the association of each term with datasets.

Terms

The crucial feature of bio-ontology terms is their substantive meaning. Bio-ontology terms are not purely conventional signs. They are intended to be descriptive of existing biological phenomena – in other words, to capture commonly agreed knowledge about the features and components of biological entities and processes. The meaning of each term is fixed via a precise definition, in which curators specify the characteristics of the phenomenon which the term is intended to designate, sometimes including species-specific exceptions (Baclawski and Niu 2006, p. 35). Each ontology represents knowledge from one or more specific domains. For instance, the Gene Ontology includes three classifications, each of which addresses different groups of phenomena (Ashburner et al. 2003): a *process* ontology describing biological objectives to which the gene or gene product contributes, such as metabolism or signal transduction; a *molecular function* ontology representing the biochemical activities of gene products, such as the biological functions of specific proteins; and a *cellular component* ontology, referring to the places in the cell where a gene product is active (see also entry on ▶ Disease Ontology). At the same time, to guarantee interoperability among ontologies, curators strive to make sure that terms used in one ontology are compatible with terms used in other ontologies (Smith et al. 2007).

Relations

Bio-ontology terms are related through a network structure whose basic features are determined by the programming language through which bio-ontologies are implemented. Within the eXtensible Markup Language (XML), the standard language used for bio-ontologies, the basic relationship between objects is called containment and involves a "parent" term and a "child" term. The child term is "contained" by the parent term when the child term represents a subclass of the parent term. This relationship is fundamental to the organization of the bio-ontology network, as it supports a hierarchical ordering of the terms used. The criteria used for the hierarchical ordering of terms are chosen in relation to the types of phenomena described by the terms in each bio-ontology. Each subset of the Gene Ontology uses the same three types of relations among terms: "is_a," "part_of," and "regulates." The first category denotes relations of identity, as in "the nuclear membrane is a membrane"; the second category denotes mereological relations, such as "the membrane is part of the cell." The third category has been implemented in 2008 to signal regulatory roles. In other bio-ontologies, for instance, the ones employed to gather data about phenotypes, the categories of relations available are more numerous and complex: for instance, including relations signaling measurement ("measured_as") or belonging ("of_a"). To help with standardization and interoperability, an ontology of relationship types (▶ Relationship Type Ontology) is currently under development.

Data Association

Terms in bio-ontologies function as keywords through which existing datasets can be ordered and retrieved. The criterion used for data classification is the evidential significance of data, i.e., their role as evidence toward establishing the structure and function of a specific entity or process. In the Gene Ontology, data on a specific gene product are categorized in terms of its known functional significance toward the development of specific traits (e.g., gene FFO2 is associated with "meristem growth"). Datasets are extracted (▶ Data Mining) either from digital repositories containing all available data of a specific type (e.g., GenBank), or from scientific publications. Once curators have selected a set of relevant data as well as a set of bio-ontology terms to which data should be associated, they label the data with a unique identifier, a machine-readable symbol that allows for the automatic analysis of data in cross-reference to other datasets. This process is called annotation (Hill et al. 2008). Unique identifiers effectively enable bio-ontologies to function as tools for data analysis. For instance, functional annotations made within the Gene Ontology are used to analyze and correlate microarray data on gene expressions and thus to ground statistical evaluations of clusters of co-expressed genes (Rubin et al. 2008).

Curation

Bio-ontology terms have the same tendency of other classificatory categories: that of stabilizing objects of knowledge in ways that enable, but at the same time constrain, future research. At the same time, the knowledge captured by bio-ontologies is bound to change with further research, as well as manifesting themselves differently in each research context. Resolving the tension between stability and flexibility of classificatory categories is crucial to bio-ontologies' success and is a core responsibility of curators, who engage in adapting and updating bio-ontologies so that they mirror the research practices and knowledge of their users. The following activities are good examples (if not exhaustive) of steps in bio-ontology development that require expert judgment and intervention by curators.

Data Mining from Publications

When extracting data from publications, curators have to single out publications that they consider to be reliable, updated, and representative for specific datasets. For instance, when gathering available data on a specific gene, curators need to choose one or two publications that best represent data relevant to a given gene product for the purposes of classification. They cannot compile data from each relevant publication, as it would be too time consuming as well as generating inconsistent annotations. Thus, curators choose what they see as the most up-to-date and accurate publications on a specific gene product, which as a consequence become "representative" publications for that entity. Further, once curators settle on a specific publication, they have to assess which data therein contained should be extracted and/or how the interpretation given within the paper matches the terms and definitions already contained in the bio-ontology. Does the content of the paper warrant the classification of given data under a new bio-ontology term? Or can the contents of the publication be associated to one or more existing terms? These choices are unavoidable when extracting data from a publication and they are impossible to regulate through fixed and objective standards. The reasons why the process of data extraction requires manual curation are the same reasons why it cannot be divorced from subjective judgment: The choices involved are informed by a curator's expertise and his or her ability to bridge between the original context of publication and the context of bio-ontology classification.

Data Formatting

Without a minimal degree of homogeneity across formats, there is no chance of making data searchable and displaying them through the visualization tools used within databases. Formatting is also geared toward making data computationally manageable: Data generated through high-throughput technologies are already in a machine-readable format, while other types of data (such as photographs) might need manipulation to be stored and retrieved through digital means. The extent to which curators need to manipulate data depends on the format in which data are produced and extracted, which is not always compatible with the software and format supported by any given bio-ontology.

Including Information About Data Provenance

Curators use metadata to classify information about data provenance. This procedure enables database users to search through datasets on the basis of context-independent criteria such as bio-ontology terms, while still allowing them to retrieve information about the production of data when needed. Curators have the responsibility of determining which information about provenance are most useful to the classification of data for circulation and reuse.

Defining Terms

Definitions that befit bio-ontology terms are not often found in biology textbooks or other publications, since those are usually steeped within specific research traditions or communities. Constructing a definition that will be acceptable to all potential users of bio-ontology, no matter which tradition they come from and which organism they work on, constitutes a challenging conceptual task for curators (Leonelli 2012).

Through activities such as the above, curators mediate between the diverse assumptions and practices characterizing the work of bio-ontology users and the need for bio-ontologies to conform to universal requirements such as consistency, computability, ease of use, and wide intelligibility. Curators' interventions are crucial to the good functioning of bio-ontologies, and ideally need to be informed by a wide range of expertise, including IT and programming skills, training in more than one biological discipline (allowing them to bridge between different scientific contexts) and familiarity with experimentation at the bench (so that they understand observational statements made in the context of specific experimental settings, as well as anticipating the expectations of bio-ontologies users).

Cross-References

- ▶ Data Mining
- ▶ Data-intensive Research

- ▶ Disease Ontology
- ▶ Model Organism
- ▶ Ontology Lookup Service for Controlled Vocabularies and Data Annotation
- ▶ Relationship Type Ontology

References

Ashburner M, Mungall CJ, Lewis SE (2003) Ontologies for biologists: a community model for the annotation of genomic data. Cold Spring Harb Symp Quant Biol 68:227–236

Baclawski K, Niu T (2006) Ontologies for bioinformatics. The MIT Press, Cambridge, MA

Bard JBL, Rhee S (2004) Ontologies in biology: design, applications and future challenges. Nat Rev Genet 5:213–222

Brazma A, Krestyaninova M, Sarkans U (2006) Standards for systems biology. Nat Rev Genet 7:593–605

Hill DP, Smith B, McAndrews-Hill MS, Blake JA (2008) Gene ontology annotations: what they mean and where they come from. BMC Bioinformatics 9(5):S2

Leonelli S (2008) Bio-ontologies as tools for integration in biology. Biol Theory 3(1):8–11

Leonelli S (2012) Classificatory theory in data-intensive science: the case of open biomedical ontologies. International Studies in the Philosophy of Science 26(1):47–65

Rhee SY, Dickerson J, Xu D (2006) Bioinformatics and its applications in plant biology. Annu Rev Plant Biol 57:335–360

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

Smith B et al (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25(11):1251–1255

Bio-PEPA

Alida Palmisano[1] and Corrado Priami[2]
[1]Department of Biological Sciences and Department of Computer Science, Department of Biological Sciences Virginia Tech, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[2]Microsoft Research-University of Trento Centre for Computational and Systems Biology and DISI, University of Trento, Povo, Trento, Italy

Synonyms

PEPA

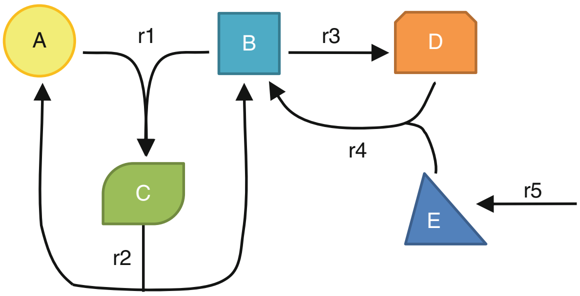
Definition

Bio-PEPA (Ciocchetta and Hillston 2009) is an extension of the PEPA language (Hillston 1996) for dealing with biochemical signaling pathways. PEPA is a formal language for describing continuous time Markov chain originally defined for the performance analysis of computer systems. PEPA allows to quantitatively model and analyze large pathway systems and it is supported by a lot of software tools for analysis and stochastic simulations are available. The combined use of PEPA and the probabilistic model checker PRISM describe, simulate, and analyze biochemical signaling pathways.

A Bio-PEPA system is a formal, intermediate, and compositional representation of biochemical systems, on which different kinds of analysis can be carried out: deterministic analysis of the ODE system, stochastic simulations (with the ▶ Stochastic Simulation Algorithm), generation and analysis of the underlying continuous time Markov chain, and generation of the input code for the PRISM model checker. Each of the analysis carried on in the different derived formalisms can be of help for studying different aspects of the biological model. Moreover, they can be used in conjunction in order to have a better understanding of the system.

The Bio-PEPA extension modifies PEPA to deal with some features of biological models that are peculiar of those kinds of systems (Calder and Hillston 2009):

- Functional rates: In contrast to PEPA, individual processes are not able to define their own rates for actions. Instead the rate associated with an action is specified once, independently of the processes in which the action occurs. The value of this rate can be specified to be a function that depends on the current state of the system.
- Stoichiometry: For each action, as well as its type, the stoichiometry or degree of involvement is also specified.
- Parameterized processes: Bio-PEPA has been designed to support the population-based reagent-centric style of modeling and so a model consists of a number of sequential components each representing a distinct species which evolve quantitatively (increasing or decreasing amounts).



$$A \overset{def}{=} (r1, 1){\downarrow}A + (r2, 1){\uparrow}A$$

$$B \overset{def}{=} (r1, 1){\downarrow}B + (r2, 1){\uparrow}B + (r3, 1){\downarrow}B$$

$$C \overset{def}{=} (r2, 1){\uparrow}C$$

$$D \overset{def}{=} (r4, 1){\downarrow}D + (r3, 1){\uparrow}D$$

$$E \overset{def}{=} (r4, 1){\downarrow}E + (r5, 1){\uparrow}E$$

$$System \overset{def}{=} A(A_0) \underset{*}{\bowtie} B(B_0) \underset{*}{\bowtie} C(C_0) \underset{*}{\bowtie} D(D_0) \underset{*}{\bowtie} E(E_0)$$

Bio-PEPA, Fig. 1 A small synthetic pathway of five reactions: A + B → C, C → A + B, B → D, D + E → B, → E. The equations exhibit various combinations of increasing/decreasing/preserved reagents between the left- and right-hand sides. The meaning of the code in the lower part of the figure is the following: in each term in the form of "(r, k) op S", r is an action name and can be viewed as the name or label of a reaction, k is the stoichiometry coefficient of the species, the combinator "op" represents the role of the element in the reaction (specifically, *down arrow* denotes the role of reactant, *up arrow* product), and S the state after the reaction is fired. The operator + expresses the choice between possible actions. The system is defined as the synchronization between components A, B, C, D, and E on the whole set of common action names (" * " symbol). $S(S_0)$ represent the initial quantities of each species

Thus in order to capture the state of a system each component is parameterized recording its current level.

- Differentiated prefix: For each action (reaction) that a component is involved in it records its role within that reaction, e.g., reactant, product, inhibitor etc. This enables the appropriate values to be used in the functional rate associated with this reaction.

Recently, some extensions of Bio-PEPA have been defined in order to represent some specific features of some biochemical networks. Specifically, the language has been extended to support SBML-events that represent changes in the system due to some trigger conditions and to support the definition of a hierarchy of compartments that have a fixed structure, but a dynamic varying size.

In Fig. 1, an example pathway modeled with Bio-PEPA is shown.

Bio-PEPA language has been used for modeling many biological case studies (for a complete list see http://biopepa.org/).

Cross-References

► Cell Cycle Modeling, Process Algebra

References

Calder M, Hillston J (2009) Process algebra modelling styles for biomolecular processes. LNCS 5750:1–25

Ciocchetta F, Hillston J (2009) Bio-PEPA: a framework for the modelling and analysis of biological systems. Theor Comput Sci 410:3065–3084

Hillston J (1996) A compositional approach to performance modelling. Cambridge University Press, Cambridge

BioPortal

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

Definition

BioPortal is an online repository of biomedical terminologies and ontologies maintained by the *National Center for Biomedical Ontology*. BioPortal contains the world's largest collection of biomedical ontologies, which it makes available using a standard Web interface and a standard API at http://bioportal.bioontology.org.

Cross-References

▶ Protégé Ontology Editor

BioSPI

Alida Palmisano[1] and Corrado Priami[2]
[1]Department of Biological Sciences and Department of Computer Science, Department of Biological Sciences Virginia Tech, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[2]Microsoft Research-University of Trento Centre for Computational and Systems Biology and DISI, University of Trento, Povo, Trento, Italy

Synonyms

Stochastic pi-calculus simulator

Definition

BioSPI is a computer application developed for simulating the behavior of biochemical systems specified in the stochastic version of pi-calculus. It is based on the Logix system, which implements flat concurrent prolog (FCP). The use of FCP allows both mobility and synchronized communication, two of the major features of the pi-calculus. This framework includes several debugging tools for tracing stepwise execution of programs, including tree traces and step-by-step execution mode. In stochastic simulations, the number of processes is monitored and recorded, as the internal clock progresses, to produce a fully ordered trace of all events (Regev et al. 2001).

This framework has been used for modeling many biological systems like circadian clock (Barkai and Leibler 2000), cell cycle (see ▶ Cell Cycle Modeling, Process Algebra), metabolic pathways, signal transduction networks, etc.

All the software tools and examples related to BioSPI can be found at its official Web site (http://www.wisdom.weizmann.ac.il/~biospi/index_main.html).

The BioSPI project was the main inspiration for another framework designed around the pi-calculus language: SPiM (Phillips and Cardelli 2007). The simulation algorithm is based on the ▶ stochastic simulation algorithm and the language features a simple graphical notation for modeling a range of biological systems.

Cross-References

▶ Cell Cycle Modeling, Process Algebra

References

Barkai N, Leibler S (2000) Biological rhythms: circadian clocks limited by noise. Nature 403:267–268

Phillips A, Cardelli L (2007) Efficient, correct simulation of biological processes in the stochastic pi-calculus. In: CMSB. LNCS, vol 4695. Springer, p 184

Regev A, Silverman W, Shapiro E (2001) Representation and simulation of biochemical processes using the pi-calculus process algebra. In: Proceedings of the pacific symposium of biocomputing 2001 (PSB2001), vol 6. pp 459–470

Bipartite Graph

Marie Lisandra Zepeda-Mendoza and Osbaldo Resendis-Antonio
Center for Genomics Sciences-UNAM, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

Synonyms

Bigraph

Definition

A bipartite graph is one whose vertices, V, can be divided into two independent sets, V_1 and V_2, and every edge of the graph connects one vertex in V_1 to one vertex in V_2 (Skiena 1990). If every vertex of V_1 is connected to every vertex of V_2 the graph is called a complete bipartite graph. If V_1 and V_2 have equal cardinality, meaning they have same number of vertices, the graph is called a balanced bipartite graph.

Another way to view a bipartite graph is by coloring the two vertices with different colors. Say all vertices of set V_1 will be colored green and all vertices of set V_2 will be colored red, then each edge will connect vertices of different colors.

This view helps to understand the fact that a graph that does not contain odd-length circles is a bipartite graph, because if it had an odd number of vertices, one of the edges would have endpoints of the same color.

Cross-References

▶ Modules, Identification Methods and Biological Function

References

Skiena S (1990) Coloring bipartite graphs, §5.5.2. In: Skiena S (ed) Implementing discrete mathematics: combinatorics and graph theory with mathematica. Addison-Wesley, Reading, p 213

Bipartite Network

▶ MicroRNA-mRNA Regulation Networks

Bipolar Attachment

Rosella Visintin
IEO, European Institute of Oncology,
Milan, Italy

Definition

Bipolar attachment is achieved when sister chromatids bind to microtubules emanating from the opposite poles.

Cross-References

▶ Mitosis

Bistability

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

Synonyms

Bimodality

Definition

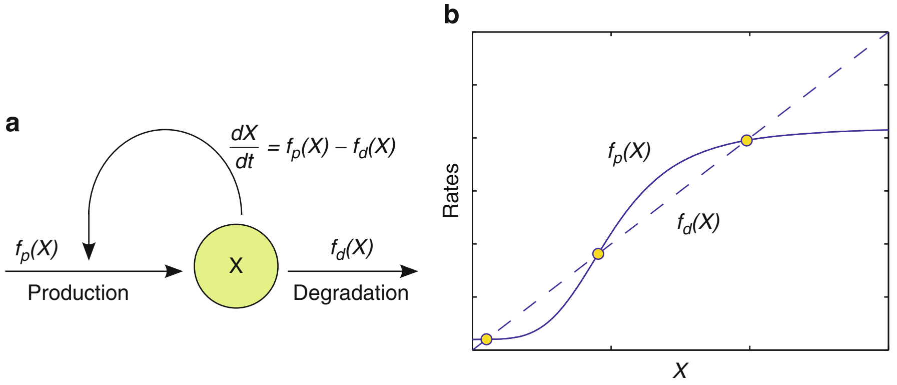
Bistability is a fundamental phenomenon in nature by which a system can be resting in two states. In biological systems, bistability is a situation in which two stable states coexist in a population of interest.

Bistability is key for understanding basic phenomena of cellular functioning, such as decision-making processes in cell cycle progression, cellular differentiation, and apoptosis. In a population with bistability, we typically observe bimodal distribution.

From a physical point of view, the bistability of a system comes from the fact that the free energy of the whole system possesses two local minimums that are separated by a peak (maximum).

In genetic network, a ▶ positive feedback with cooperative binding is a necessary condition to achieve bistability. The positive feedback often appears as a double-▶ negative feedback.

Figure 1 shows the basic characteristics of a bistable genetic system. In a bistable system, a key regulator protein *X* activates its own expression to form a positive feedback. The feedback can be formed either directly, through a cooperative binding site of protein *X* to its own promoter, or indirectly, through a signal-transduction cascade via other regulators.



Bistability, Fig. 1 Characteristics of bistable systems (Smits et al. 2006). (**a**) Schematic illustration of a bistable gene regulation. The functions $f_p(X)$ and $f_d(X)$ represent the production and degradation rates of X, respectively. The differential equation shows the change of X over time (dX/dt). (**b**) Plot of the functions $f_p(X)$ and $f_d(X)$ within the same graph. The intersection points show the steady states

The production rate of X can be expressed as a Hill function:

$$f_p(X) = f_o + \frac{f_1 X^n}{K^n + X^n}, \qquad (1)$$

Here n is the Hill coefficient. The degradation (or dilution) of the protein can be described by a linear-type function. The change of X over time is described by a differential equation combining the production and degradation rates. The interaction of these two functions gives the steady states of the system in which the change of X is equal to 0 (Fig. 1b). When the positive feedback is cooperative, i.e., the Hill coefficient is $n > 1$, there are three steady states, two of which are stable and separated by an unstable state.

Cross-References

▶ Cell Cycle Dynamics, Irreversibility

References

Smits WK, Kuipers OP, Veening JW (2006) Phenotypic variation in bacteria: the role of feedback regulation. Nat Rev Microbiol 4:259–271

Bistable Switch

▶ Toggle Switch, Switching Network

Bisulfite Conversion

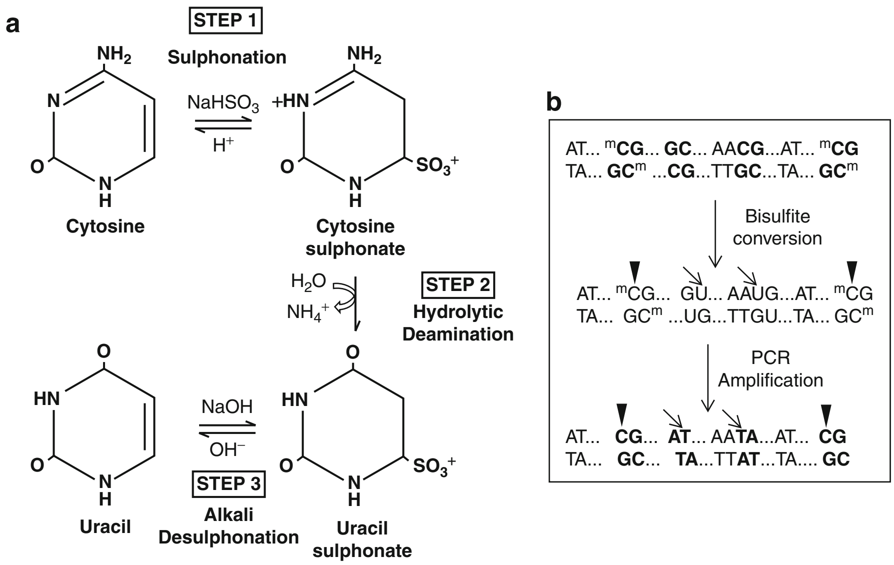
Vani Brahmachari and Shruti Jain
Dr. B. R. Ambedkar Center for Biomedical Research,
University of Delhi, Delhi, India

Synonyms

Bisulfite sequencing: methylation status analysis

Definition

Bisulfite conversion is a frequently used technique to directly identify the methylated status of cytosine residues in DNA, besides the method based on using ▶ methylation-sensitive restriction endonucleases (Frommer et al. 1992). On treatment of DNA with sodium bisulfite, unmethylated cytosine residues are converted to uracil, and on replication of this DNA, thymine is incorporated at these positions. On the other



Bisulfite Conversion, Fig. 1 Schematic representation of bisulfite sequencing. (**a**) Steps of chemical conversion. (**b**) The sequence obtained after conversion is shown; cytidine (C) is replaced by thymidine (T) at positions where there is no methylation (*arrow*) while methylated cytidine remains unchanged (*arrowhead*). The relevant dinucleotides are shown in *bold font*

hand, methylated cytosine (5-methylcytosine) remains unaffected by the bisulfite treatment (Fig. 1). When the sequence of bisulfite converted DNA is determined, unmethylated cytosine is read as thymine while 5-methyl cytosine remains as cytosine. On comparison with DNA sequence without bisulfite conversion, the methylation pattern of the DNA can be deciphered.

Cross-References

- ▶ Epigenetics
- ▶ Methylation-sensitive Restriction Endonucleases

References

Frommer M, Mcdonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Nat Acad Sci USA 89:1827–1831

Bisulfite Sequencing: Methylation Status Analysis

▶ Bisulfite Conversion

BlenX

Alida Palmisano[1] and Corrado Priami[2]
[1]Department of Biological Sciences and Department of Computer Science, Department of Biological Sciences Virginia Tech, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[2]Microsoft Research-University of Trento Centre for Computational and Systems Biology and DISI, University of Trento, Povo, Trento, Italy

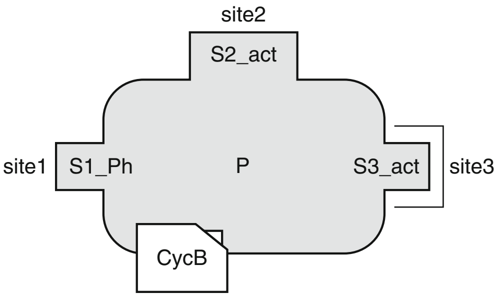
Synonyms

Beta workbench; BetaWB; CoSBiLab

Definition

BlenX (Dematté et al. 2008, 2010) is a stochastic language implementing the β-binders process algebra (Priami and Quaglia 2005) explicitly designed for modeling biological systems. A BlenX program is made up of a program file for the system structure, an interfaces file for the quantitative information about the system, and an optional declaration file for the user-defined variables and functions. A BlenX program can be executed by the Beta Workbench software tool (which can be freely downloaded at http://www.cosbi.eu/index.php/research/prototypes together with the CoSBiLab companion tools) that implements an efficient variant of the ▶ stochastic simulation algorithm.

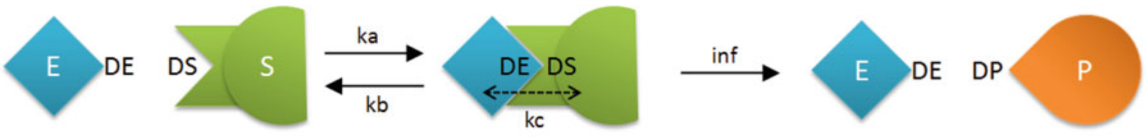
The basic metaphor of BlenX is that a biological entity (i.e., a component that is able to interact with other components to accomplish some biological functions) is represented by a box. A box has interfaces (also called binders) and an internal structure that drives its behavior (see Fig. 1). For example, in a box modeling a protein, binders may represent sensing and effecting domains. Sensing domains are the places where the protein receives signals, effecting domains are the places that a protein uses for propagating signals, and the internal structure codifies for mechanism that transforms an input signal into a protein conformational change, which can result in the activation or deactivation of another domain. The exchanging of signals can happen between boxes whose binders have a certain degree of affinity, which codes the strength of their interaction.



BlenX, Fig. 1 Graphical notation of a BlenX box. The small squares on the border of the box are the binders; site1, site2, and site3 are the interfaces subjects (omitted when not necessary); S1_Ph, S2_act, and S3_act are the interfaces types; the line surrounding the interface with type S3_act indicates that the interface is hidden; P is the internal program and CycB is the name of the box

The internal program of a box is described by a process built on top of the a set of primitives, derived from the classical process algebra primitives constructs (i.e., input/output actions on communication channels, sequential/alternative/parallel composition of actions) and some other primitives specific for the BlenX language (i.e., changing the type/state of an interface and conditional statements).

In addition, BlenX allows the definition of events which are statements, or verbs, that are executed with a specified rate and/or when some conditions of the system are satisfied. Events can split an entity into two entities, join two entities into a single one, and add or remove entities into or from the system. The final peculiar characteristic of BlenX is the possibility of



Program file	Interfaces file	Declaration file

[steps = 1000]
<< BASERATE:inf >>
 	
let E : bproc = #(x,DE) [rep x!().nil];
let S : bproc = #(y,DS) [y?().ch(y,DP).nil];
Let P : bproc = #(y,DP) [nil];
 	
run 1 E || 100 S
	{ DE, DS, DP }
	%%
	{
	 (DE,DS, rate(ka), rate(kb), rate(kc))
	 (DE,DP,0,inf,0)
	}
		let ka : const = 1.0;
		let kb : const = 1.0;
		let kc : const = 1.0;

BlenX, Fig. 2 Coding a basic enzymatic reaction in BlenX. *Upper part*: boxes are pictured as different shapes where corners represent molecule interfaces (with their associate type). *Lower part*: The BlenX model of this reaction network is coded in three files: the program file (containing the definitions of the species), the interfaces file (containing the affinities between the types), and the declaration file (containing the definitions of the constants). The complexation between E and S happens because of the affinity of DE and DS types at rate ka. This reaction is reversible, so the decomplexation rate between the same types is kb. If E and S are connected, they share a private channel through which the synchronization of the output action x!() and input action y?() can happen at rate kc. If this happens the binder of S changes (at infinite rate) its type from DS to DP (see the internal code of S), allowing the decomplexation of the two boxes (because the affinity between DE and DP is infinite for the decomplexation event, see the interfaces file). The final state of the system is the E species back in its initial state and a new product species P creating complex-on-the-fly (i.e., during the simulation): the complexation event is performed when two boxes have interfaces with compatible types and it is used for modeling the creation of a private and permanent (until the complementary decomplexation event) communication channel between two specific boxes. Those two boxes will maintain their individual internal behavior but they will be able to communicate on the shared channel: the only thing that the user needs to define is a rate of complexation/decomplexation between two binder types.

An example of a basic enzymatic reaction modeled in BlenX is shown in Fig. 2.

Cross-References

▶ Cell Cycle Modeling, Process Algebra

References

Dematté L, Priami C, Romanel A (2008) The BlenX language: a tutorial. In: SFM 2008. LNCS, vol. 5016, Springer, pp 313–365

Dematté L, Larcher R, Palmisano A, Priami C, Romanel A (2010) Programming biology in BlenX. In: Sangdun Choi (ed) Systems biology for signaling networks, vol. 1. Springer, pp 777–821

Priami C, Quaglia P (2005) Beta binders for biological interactions. In: Proceedings of the second international workshop on computational methods in systems biology (CMSB04). LNBI, vol. 3082. Springer, pp 21–34

BLOcks SUbstitution Matrix (BLOSUM)

Joo Chuan Tong
Data Mining Department, Institute for Infocomm Research, Singapore, Singapore
Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Definition

BLOSUM matices are calculated from blocks of highly conserved amino acid sequences with a certain degree of similarity between these sequences. Matrix constructed from no more than x% similarity is called BLOSUM-x matrix. The probability $p_{i,j}$ of a unique pair of amino acids at a site (A_i and A_j) is computed as well as the probability p_i of the unique amino acid to be A_i. Then the log odd ratio $\log \frac{p_{i,j}}{p_i}$ is computed and recorded in the (i,j) entry of the matrix. The BLOSUM matrix comprises of information regarding the 20 amino acid properties similarity and yields delicate evolutional and chemical association among the 20 amino acids.

Cross-References

▶ TAP Translocator, In Silico Prediction

References

Henikoff S, Henikoff JG (1992) Amino acid substitution matrics from protein blocks. Proc Natl Acad Sci USA 89:10915–10919

Bone Marrow-derived Cells

Mary Helen Barcellos-Hoff
Department of Radiation Oncology and Cell Biology,
New York University School of Medicine,
New York, NY, USA

Definition

Bone marrow-derived cells (BMDC) are derived from hematopoietic stem cells that give rise to diverse cell types present in blood and lymphatic organs, some of which are recruited to other organs.

Characteristics

The term, bone marrow-derived cells (BMDC), encompasses a broad class of undifferentiated cells known to originate from the bone marrow (▶ Bone Marrow-derived Cells) that are dispersed among tissues. The plasticity of these cells has been implicated in diverse processes from development to pathology but remains poorly understood. Bone marrow transplantation and lineage-tracing experiments have provided strong experimental evidence that BMDC can generate distinct cell types, including cardiac muscle, liver cell types, neuronal and nonneuronal cell types of the brain, as well as endothelial cells and osteoblasts. These multiple cell types could have originated from either HSC or MSC within bone marrow or from other less well-defined precursors. Injury may mobilize BMDC into circulation and recruitment and differentiation at damage sites (Orlic et al. 2001). Inflammation can also mobilize BMDC, and chronic inflammation may drive BMDC to actively participate in pathological process, including cancer (Houghton et al. 2004).

Bone marrow, a tissue that replenishes the circulating cells of the blood and immune system, contains many cell types, including stroma, vascular cells, adipocytes, osteoblasts, and osteoclasts. These cells form the microenvironment, or niche, in which mesenchymal stem cells (MSC) and hematopoietic stem cells (HSC) reside. Both stem cells originate within the bone marrow, but HSC mostly function to regenerate hematopoietic cells within the marrow and the circulation, while MSC mostly relocate to produce a variety of cell types found in diverse organs.

MSC replicate as undifferentiated cells and have the potential to differentiate to lineages of mesenchymal tissues, including bone, cartilage, fat, tendon, muscle, and marrow ▶ stroma (Pittenger et al. 1999). A MSC population resides in bone marrow, but cells with similar molecular, phenotypic, and functional properties, at least in vitro, have been identified in a wide range of tissues. MSC populations in adult tissues and organs contribute to tissue cell turnover and also respond to tissue damage, for example, in wound healing (▶ Wound Response).

HSC give rise to three main lineages, erythroid, myeloid and lymphoid, which respectively form red blood cells, monocytes, and immune cells (Morrison et al. 1997). Extensive single-cell transplantation of highly purified stem cells demonstrated that the clonal contribution to the different blood cell lineages varies significantly and can be stably maintained through serial passaging, providing evidence that the pool of HSCs comprises at least two and possibly more distinct clonal subtypes imbued with differential lineage and self-renewal potential (Dykstra et al. 2007). A great many functional and characterization studies have established a functional hierarchy of HSC using assays for growth potential under restrictive conditions and the ability to rescue lethally irradiated mice. The uncommitted, or pluripotent, gives rise to five progenitor cells that give rise to seven distinct cell types: erythrocyte, lymphocyte, neutrophil, eosinophil, basophil, monocytes, and platelets. For many of these cells differentiation occurs within the bone marrow, but for lymphocytes and monocytes, maturation occurs after release into circulation and residence in tissues. In the case of lymphocytes, this occurs in the tissues of the immune system. Monocytes circulate for 3 days before migrating into tissues where they become tissue-resident macrophages. Dendritic cells, an important antigen-presenting cell residing in tissues, can be either lymphoid or myeloid in origin.

BMDC are mobilized in cancer-bearing animals and in humans (Wels et al. 2008). Many studies were initiated following the observation that recruitment of a subset of hematopoietic and vascular progenitors were required to assemble new vessels in tumors (Lyden et al. 2001). The key concepts are that BMDC have temporally and spatially restricted roles in supporting in specific types of tumors, that recruitment of even small numbers of BMDC can play a crucial role in catalyzing tumor progression and that BMDC may precede or presage cancer, implicating them in the very earliest manifestations of cancer (Kaplan et al. 2005).

The variety, rarity, and plasticity of different BMDC subsets complicate their analysis and respective importance and specific functions in response to injury, cancer, and pathology at large. An important tool has been implementation of bone marrow transplantation in conjunction with various genetic reporter systems and immunologic markers. One idea is that BMDC can be used to treat injured tissues and promote repair, and another is that blocking BMDC recruitment and/or action can prevent disease, but there is still much to be learned about the recruitment, behavior, and stability of these cells across development, during homeostasis, and in disease.

Cross-References

▶ Adaptive Immune System
▶ Bone Marrow-derived Cells
▶ Fibroblasts
▶ Stroma

References

Dykstra B, Kent D, Bowie M, McCaffrey L, Hamilton M, Lyons K,Lee S-J, Brinkman R, Eaves C (2007) Long-term propagation of distinct hematopoietic differentiation programs in vivo. Cell Stem Cell 1(2):218–229

Houghton J, Stoicov C, Nomura S, Rogers AB, Carlson J, Li H, Cai X, Fox JG, Goldenring JR, Wang TC (2004) Gastric cancer originating from bone marrow-derived cells. Science 306(5701):1568–1571. doi:10.1126/science.1099513

Kaplan RN, Riba RD, Zacharoulis S, Bramley AH, Vincent L, Costa C, MacDonald DD, Jin DK, Shido K, Kerns SA, Zhu Z, Hicklin D, Wu Y, Port JL, Altorki N, Port ER, Ruggero D, Shmelkov SV, Jensen KK, Rafii S, Lyden D (2005) VEGFR1-positive haematopoietic bone marrow progenitors initiate the pre-metastatic niche. Nature 438:820–827

Lyden D, Hattori K, Dias S, Costa C, Blaikie P, Butros L, Chadburn A, Heissig B, Marks W, Witte L, Wu Y, Hicklin D, Zhu Z, Hackett NR, Crystal RG, Moore MAS, Hajjar KA, Manova K, Benezra R, Rafii S (2001) Impaired recruitment of bone-marrow-derived endothelial and hematopoietic precursor cells blocks tumor angiogenesis and growth. Nat Med 7(11):1194–1201

Morrison SJ, Wandycz AM, Hemmati HD, Wright DE, Weissman IL (1997) Identification of a lineage of multipotent hematopoietic progenitors. Development 124(10):1929–1939

Orlic D, Kajstura J, Chimenti S, Limana F, Jakoniuk I, Quaini F, Nadal-Ginard B, Bodine DM, Leri A, Anversa P (2001) Mobilized bone marrow cells repair the infarcted heart, improving function and survival. Proc Natl Acad Sci 98(18):10344–10349. doi:10.1073/pnas.181177898

Pittenger MF, Mackay AM, Beck SC, Jaiswal RK, Douglas R, Mosca JD, Moorman MA, Simonetti DW, Craig S, Marshak DR (1999) Multilineage potential of adult human mesenchymal stem cells. Science 284(5411):143–147. doi:10.1126/science.284.5411.143

Wels J, Kaplan RN, Rafii S, Lyden D (2008) Migratory neighbors and distant invaders: tumor-associated niche cells. Genes Dev 22(5):559–574

Bonferroni Correction

Winston Haynes
Seattle Children’s Research Institute, Seattle,
WA, USA

Definition

In ▶ Multiple Hypothesis Testing, the Bonferroni correction is a conservative method for probability thresholding to control the occurrence of false positives.

When deciding whether to accept or reject an individual null hypothesis, a probability threshold, α, is utilized to control the likelihood of false positives. In multiple hypothesis testing, an increased number of samples, n, in a given family increases the probability that false positives will arise within that family at the same probability threshold, α. Thus, the threshold, α, should be lowered to control the total number of false positives.

The Bonferroni correction controls the number of false positives arising in each family by using a probability threshold of α/n for each observation within the family. By guaranteeing that the probability of a test being accepted within a family is the same as or less than the probability of any individual test being accepted, the Bonferroni correction is extremely conservative. When the number of comparisons is large, use of the Bonferroni correction should be limited and replaced by methods like the ▶ Benjamini-Hochberg Method.

References

Abdi H (2007) The Bonferroni and Sidak corrections for multiple comparisons. In: Salkind NJ (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilit ‘a. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8:3–62

Boolean Function

Xiujun Zhang
Institute of Systems Biology, Shanghai University,
Shanghai, China

Synonyms

Switching function

Definition

In mathematics, a Boolean variable $\boldsymbol{B}$ has the value 0 or 1. A Boolean function is a function with the form

f: $\boldsymbol{B}^n \rightarrow \boldsymbol{B}$, where $\boldsymbol{B}^n = \boldsymbol{B} \times \boldsymbol{B} \times \ldots \times \boldsymbol{B} = \{0, 1\} \times \{0, 1\} \times \ldots \times \{0, 1\}$, the Cartesian product of $\boldsymbol{B}$ with itself to n factors (Chatterjee 2005).

For example, for $n = 2$, $\boldsymbol{B}^2 = \{0, 1\} \times \{0, 1\} = \{00, 01, 10, 11\}$. Thus, a Boolean function of two variables is a function from $\boldsymbol{B}^2$ to $\boldsymbol{B}$ which assigns every ordered pair of elements of $\boldsymbol{B}$ to a unique element of $\boldsymbol{B}$. “AND” function and “OR” function are two important Boolean functions of two variables (Koshy 2004).

References

Chatterjee D. Abstract algebra. New Delhi: Prentice Hall; 2005.
Koshy T. Discrete mathematics with applications. Burlington: Elsevier; 2004

Boolean Model

Xi Chen, Wai-Ki Ching and Nam-Kiu Tsing
Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, University of Hong Kong, Hong Kong, China

Synonyms

Boolean networks; Probabilistic boolean networks

Definition

Boolean network (BN) is known as a popular mathematical model for modeling genetic regulatory networks. The BN model was first proposed by Kauffman (1969). In a BN model, the gene expression states are quantized into only two levels: on and off (represented as 1 and 0). The target gene is determined by several other genes called its input genes according to regulation rules (given as Boolean functions). A BN is said to be well defined when all the input genes and Boolean functions are given (Kauffman (1993)). There are two types of BN models: synchronous BNs and asynchronous BNs, depending on whether or not the states of nodes are updated synchronously. Synchronous model is more popular and easier to analyze and therefore we adopt it in our discussion. We note that a BN model is a deterministic model and the only randomness comes from its initial state. Considering the inherent deterministic directionality in BNs as well as only a finite number of possible states, it is easy to see that the BN will eventually enter into a set of state(s) and stay there forever. These states are called attractors and the states that lead to them comprise their basins of attraction. The number of transitions needed to return to a given state in an attractor is called cycle length.

Since a BN is a deterministic model, to better model a genetic regulatory network, one has to overcome the deterministic rigidity of a BN. It is therefore natural to extend BNs to a stochastic setting and this results in a new class of mathematical models, namely, probabilistic Boolean networks (PBNs). The basic idea is described as follows. In a PBN, each gene can have more than one Boolean function with a certain selection probability assigned to it. The dynamics of a PBN can be studied by using Markov chain theory, and the network behavior is characterized by its transition probability matrix and its steady-state probability distribution. One can understand a genetic regulatory network and identify the influence of different genes via such a network. Then, therapeutic gene intervention and gene control policies can be developed and studied. Similar to BN, there are two classes of PBNs : synchronous PBNs and asynchronous PBNs. Synchronous PBNs are more popular as they are easier to be analyzed. Moreover, there are two types of PBNs: instantaneously random PBNs and context-sensitive PBNs. The instantaneously random PBN is essentially a collection of Boolean networks in which, at any discrete time point, the gene state vector transforms according to the rules of one of the constituent networks. That is, the rule for updating each gene is randomly chosen at each time step from several possible rules in accordance with a fixed probability distribution. The context-sensitive PBN is an extension of PBN. It differs from instantaneously random PBNs for two reasons: (1) each gene is allowed to change its activity with a small probability at each time instant, regardless of which constituent BN is active at the moment, and (2) switching between constituent BNs occurs with a small probability, such that a PBN may remain in the same constituent BN for a longer time interval (Pal et al. 2005).

Characteristics

A Boolean Network (BN) $G(V, F)$ actually consists of a set of n nodes (where each node corresponds to a gene):

$$V = \{v_1, v_2, \ldots, v_n\}.$$

and a list of Boolean functions (which represent the regulatory rules for nodes):

$$F = \{ f_1, f_2, \ldots, f_n\}.$$

where $f_i : \{0, 1\}^n \to \{0, 1\}$. Define $v_i(t)$ to be the state (0 or 1) of the node v_i at time t. The rules of the regulatory interactions among the genes are then represented by

$$v_i(t+1) = f_i(\mathbf{v}(t)), \quad i = 1, 2, \ldots, n.$$

Here, we let $\mathbf{v}(t) = (v_1(t), v_2(t),\ldots, v_n(t)\ )^T$, which is called the Gene Activity Profile (GAP). The GAP can take any possible form (state) from the set

$$S = \left\{ (v_1, v_2, \ldots, v_n)^T : v_i \in \{0, 1\} \right\} \tag{1}$$

and thus totally there are 2^n possible states in the network. Hence, $\mathbf{v}(t + 1)$ is determined from $\mathbf{v}(t)$ in a BN. As mentioned before, given an initial state $\mathbf{v}(0)$, the BN will enter into a cycle called attractor. An attractor consisting of only one global state is called a singleton attractor. Otherwise, it is called a cyclic attractor with period p if it consists of p states.

The following is an example of a BN having two genes with the truth table given in Table 1.

Boolean Model, Table 1 The truth table

State	$v_1(t)$	$v_2(t)$	$f^{(1)}$	$f^{(2)}$
1	0	0	1	1
2	0	1	1	0
3	1	0	1	0
4	1	1	0	0

From the truth table, there are four states and they are (0, 0), (0, 1), (1, 0), and (1, 1). Let us label them as 1, 2, 3, and 4, respectively. We note that if the current state of the BN is 3, the network will stay with State 3 in the next step (with probability one). Suppose the current state is 2, the network will go to State 3 in the next step (with probability one). Similarly if the current state is 1, the network will go to State 4 in the next step (with probability one). Finally, if the current network state is 4, the BN will then go to State 1 in next step (with probability one). The transition-probability matrix (Boolean network matrix) of the 2-gene BN is then given by

$$B_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \tag{2}$$

The truth table provides the one-step transition probability (0 or 1 in the case of BN) between any two states. We observe that there are two cycles and they are given as follows: (a) (0, 0) ↔ (1, 1), and (b) (1, 0) ↔ (1, 0). Thus, State 3 is an attractor cycle of period (length) one and States 1 and 4 form an attractor cycle of period two. We remark that there is a one-to-one relation between a BN and its corresponding BN matrix.

Since BN is a deterministic model, it may not be able to capture the properties of a biological system which processes certain randomness. Moreover, the microarray data sets used to infer the network structure are usually not accurate because of experimental noise in the complex measurement process. To address this, a stochastic model, namely, probabilistic Boolean network (PBN) was developed by Shmulevich, Dougherty, Kim, and Zhang (2002). In a PBN, each gene instead of having only one Boolean function, there are a number of Boolean functions (predictor functions) $f_j^{(i)}(\,j = 1, 2, \ldots, l(i))$ to be chosen for determining the state of gene v_i. The probability of choosing $f_j^{(i)}$ as the predictor function is $c_j^{(i)}$ (Shmulevich and Dougherty 2007),

$$0 \le c_j^{(i)} \le 1 \quad \text{and} \quad \sum_{j=1}^{l(i)} c_j^{(i)} = 1 \quad \text{for} \\ j = 1, 2, \ldots, l(i), \quad i = 1, 2, \ldots, n. \tag{3}$$

The probability $c_j^{(i)}$ can be estimated by using the statistical method Coefficient of Determination (COD) developed by Dougherty, Kim, and Chen (2000) with real gene expression data sets.

Cross-References

- ▶ Global State, Boolean Model
- ▶ Local State, Boolean Model
- ▶ Probabilistic Boolean Network
- ▶ Regulation Function
- ▶ Synchronous Model

References

Dougherty E, Kim S, Chen Y (2000) Coefficient of determination in nonlinear signal processing. Signal Process 80:2219–2235

Kauffman S (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22:437–467

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

Pal R, Datta A, Bittner M, Dougherty E (2005) Intervention in context-sensitive probabilistic boolean networks. Bioinformatics 21:1211–1218

Shmulevich I, Dougherty E, Kim S, Zhang W (2002) From boolean to probabilistic boolean networks as models of genetic regulatory networks. Proc IEEE 90:1778–1792

Shmulevich I, Dougherty E (2007) Genomic signal processing. Princeton University Press, New Jersey

Boolean Modeling of Cell Cycle Control

▶ Cell Cycle Modeling Using Logical Rules

Boolean Networks

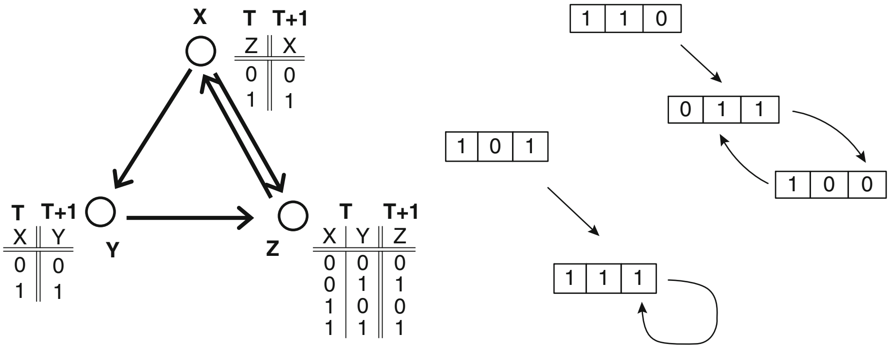
Zhong-Yuan Zhang
School of Statistics, Central University of Finance and Economics, Beijing, China

Synonyms

Kauffman network; N-K model

Definition

A Boolean network model is composed of two parts: (1) A directed and loops-allowed graph with N nodes representing the genes where each gene is connected with only K other genes (hence, Boolean network is also called N-K model) (Jong 2002; Kauffman 1969; Lähdesmäki et al. 2003). Each gene only has two states: 1 (on) means expressed and 0 (off) means silent. (2) Boolean functions and states of genes, where the state of each gene i, $i = 1, 2, \cdots, N$, are modeled as the output of Boolean function f_i with K (or at most K) inputs specified by the graph. Once the network and the Boolean functions are determined, the model's state at any discrete time step is uniquely determined by the initial assignment of the states of the genes. Since the number of the possible states is finite, the model will inevitably fall into an attractor at last (point attractor or cyclic attractor; see Fig. 1).



T	T+1
Z	X
0	0
1	1

T	T+1
X	Y
0	0
1	1

T		T+1
X	Y	Z
0	0	0
0	1	1
1	0	0
1	1	1

Boolean Networks, Fig. 1 *Left*: An example of Boolean network. X, Y, and Z are three genes. The Boolean functions of them are given by the truth tables placed next to them. For example, the Boolean function f_z of gene Z at time T is $f_z(0, 0) = 0$; $f_z(1, 0) = 0; f_z(0, 1) = 0; f_z(1, 1) = 1$. *Right*: The states trajectories of two particular realizations of the Boolean network. The top trajectory falls into a cyclic attractor, and the bottom one falls into a point attractor

To find the Boolean networks for real gene expression dataset, many algorithms have been developed.

Cross-References

- ▶ Boolean Model
- ▶ Identification of Gene Regulatory Networks, Machine Learning

References

Jong HD (2002) Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol 9(1):67–103

Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22:437–467

Lähdesmäki H, Shmulevich I, Yli-Harja O (2003) On learning gene regulatory networks under the Boolean network model. Mach Learn 52:147–167

Bootstrap Aggregating

- ▶ Bagging

Bootstrap Resampling

- ▶ Bootstrapping

Bootstrapping

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

Synonyms

Bootstrap resampling

Definition

The bootstrap is a data resampling strategy (Efron 1983; Efron and Tibshirani 1997; Duda et al. 2001). This resampling provides an estimate for an unknown population parameter θ. Let a data set D be a sample of n data points (or cases) $\mathbf{x}_i$, $i = 1..n$, from the population under study. The values of these cases are assumed to be the outcomes of independent and identically distributed random variables with (unknown) probability density function. Without specific assumptions or a particular model for this probability function, the bootstrap is non-parametric, otherwise parametric. The parameter θ is a function of the values in the population, $\theta = T(x)$, for example, the population mean (Dixon 2006). This function also determines the sample statistic, $\hat{\theta} = T(\mathbf{x})$, for example, the mean of all values in D (i.e., the sample mean). To estimate the sampling distribution $F_\theta(x)$ of the function T, a bootstrap sample is generated by randomly and uniformly selecting n cases from D with replacement. This sampling is repeated B times to generate B bootstrap samples (Fig. 1). The statistic $\hat{\theta}$ is calculated for all B bootstrap samples individually. Then the bootstrapped estimate, $\hat{\theta}^{boot}$, of the population parameter is calculated based on all $b = 1..B$ individual statistics $\hat{\theta}_b$.Thus, the empirical distribution of all $\hat{\theta}_b$ provides an estimate for the theoretical sampling distribution $F_\theta(x)$ of the function T (Dixon 2006). For example, if this function is the mean, then we average all estimates $\hat{\theta}_b$ as shown in (1) to obtain the bootstrapped estimate of the mean, $\hat{\theta}^{boot}$.

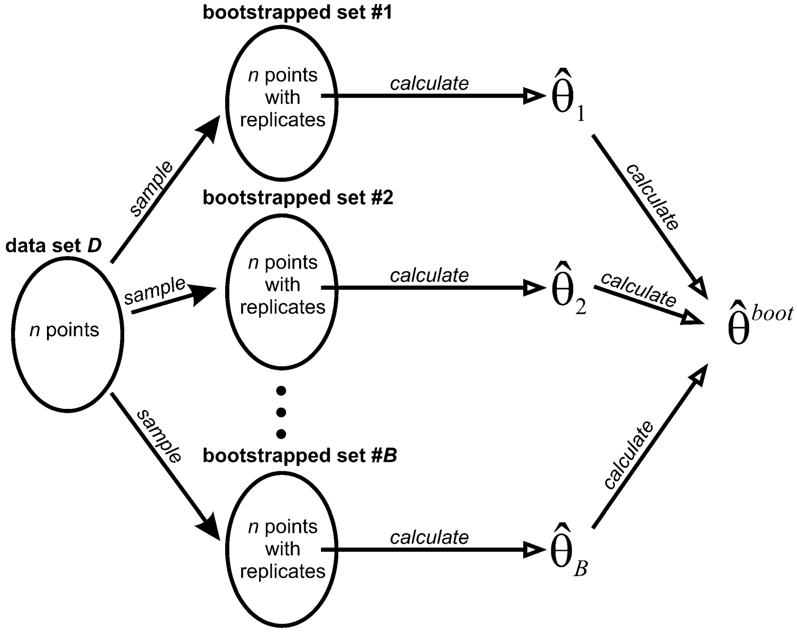
$$\hat{\theta}^{boot} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b \tag{1}$$

Characteristics

Bootstrapping is a general data resampling strategy for estimating any population parameter. For example, from the empirical distribution of the B statistics $\hat{\theta}_b$, we can calculate the bootstrapped estimates for bias and variance of the sample statistic $\hat{\theta}$ as follows:

$$bias\{\hat{\theta}\} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b - \hat{\theta}) \tag{2}$$

Bootstrapping, Fig. 1 Simplified schematic of the standard bootstrap process. From the data set D comprising n points, B bootstrap samples are generated by repeated uniform sampling with replacement. From each bootstrap sample, an estimate of the parameter θ is calculated. The B estimates are used to derive the bootstrapped estimate $\hat{\theta}^{boot}$



$$\text{variance}\{\hat{\theta}\} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \frac{\sum_{b=1}^{B}\hat{\theta}_b}{B}\right)^2 \qquad (3)$$

For $B \rightarrow \infty$, the bootstrapped estimate approaches the true parameter of interest. Therefore, by increasing the number of bootstrap samples, we can improve the accuracy of our estimate.

To create one bootstrap sample, the data set D with n cases is sampled n times. In the standard bootstrap, this sampling is uniform; hence, each of the n data points in D has the same probability $1/n$ of being selected for the bootstrap sample. Thus, the probability that a point is *not* selected for the bootstrap sample is $(1 - 1/n)^n$, which is approximately $e^{-1} \approx 0.368$ for large n. Therefore, the expected number of distinct points in a bootstrap sample is about 0.632 times n.

The standard non-parametric resampling is arguably the most widely used type of bootstrap. Alternative methods include the *balanced bootstrap*, which warrants that each case occurs exactly B times in B bootstrap samples (Dixon 2006). The *smoothed bootstrap* adds a small amount of noise to each sampled case (Silverman and Young 1987). Both balanced and smoothed bootstrap stabilize the variance estimation. Whereas the standard bootstrap assumes that the data points are independent and identically distributed random variables, the *moving blocks bootstrap* is applicable to correlated observations, for example, time series data (Künsch 1989; Dixon 2006).

The Bootstrap in Systems Biology

The bootstrap and ▶ cross-validation are the two most widely used data resampling strategies in systems biology. Specifically for evaluating classification performance in small sample scenarios such as microarray data classification, the bootstrap is a popular choice.

In ▶ classification tasks, the bootstrap can be used to estimate the prediction error ε of a classifier C. Here, the classifier is built B times using the data in the bootstrap samples, which serve as training sets. The resulting B models are then applied to the original data set D, which serves as test set.

Let y_i denote the true class label of a case $\mathbf{x}_i$. C_b is the classifier resulting from the application of a learning algorithm to the data of the b^{th} bootstrap set, and $C_b(\mathbf{x}_i)$ is the prediction of this classifier for case $\mathbf{x}_i$. $L(y_i, C_b(\mathbf{x}_i))$ is the loss function of the classification. For example, the 0–1 loss function gives

0 if $y_i = C_b(\mathbf{x}_i)$ and 1 otherwise. Equation (4) defines the bootstrapped estimate of the prediction error (Hastie et al. 2008). Here, B is the number of bootstrap samples, and n is the total number of cases.

$$\hat{\varepsilon}^{boot} = \frac{1}{B}\frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n} L(y_i, C_b(x_i)) \tag{4}$$

The expected number of cases that are identical in the data set D and the b^{th} bootstrap sample is 0.368 times n. As the training sets overlap with the test set, the bootstrapped estimate of the prediction error, $\hat{\varepsilon}^{boot}$, is biased downward (i.e., optimistically biased). For ▶ overfitted classifiers, the bootstrapped estimate of the prediction error would then be smaller than their true prediction error ε.

To alleviate the optimistic bias of (4), we can exclude those cases that are contained in *both* the bootstrap samples *and* the original data set. The resulting estimate is called the *leave-one-out bootstrap estimate of the prediction error* (Hastie et al. 2008).

$$\hat{\varepsilon}^{loob} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|S_{-i}|}\sum_{b \in S_{-i}} L(y_i, C_b(x_i)), \tag{5}$$

with S_{-i} being the set of indices of bootstrap samples that do *not* contain the case i, and $|S_{-i}|$ is the number of those samples. Because of the random sampling, it is possible that all bootstrap samples contain the case i, leading to $|S_{-i}| = 0$. To avoid a division by zero, the number of bootstrap samples needs to be sufficiently large so that at least one bootstrap sample does not contain the case i. Alternatively, we could omit those cases that are included in all bootstrap samples (Hastie et al. 2008). Note that, although the estimate $\hat{\varepsilon}^{loob}$ is called the leave-one-out bootstrapped estimate, on average, 0.368 times n cases are left out per bootstrap sample. The leave-one-out bootstrap estimate of the prediction error is not to be confused with the out-of-bag estimate of the error rate resulting from bagged classifiers.

For each bootstrap sample, the expected number of distinct training cases is 0.632 times n. So each classifier is trained on about 63% of the cases only. As a result of the small effective training set, the leave-one-out bootstrapped estimate $\hat{\varepsilon}^{loob}$ tends to overestimate the true prediction error (i.e., $\hat{\varepsilon}^{loob}$ is biased upward). The 0.632 bootstrap (▶ Bootstrapping, 0.632 Bootstrap) addresses this bias by weighing the leave-one-out bootstrapped estimate and the *bootstrapped resubstitution error*, $\hat{\varepsilon}^{resub}$, which is defined in (6).

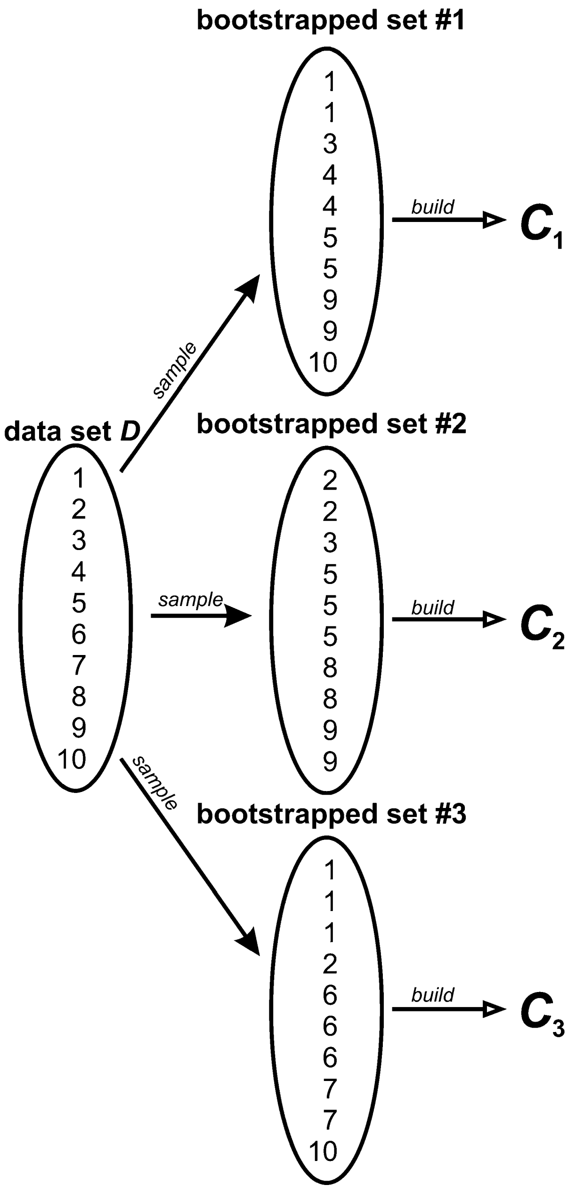
$$\hat{\varepsilon}^{resub} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|S_{+i}|}\sum_{b \in S_{+i}} L(y_i, C_b(x_i)), \tag{6}$$

with S_{+i} being the set of indices of bootstrap samples that contain the case i, and $|S_{+i}|$ is the number of those samples. Like any resubstitution error estimate, $\hat{\varepsilon}^{resub}$ is biased downward, i.e., it is smaller than the true prediction error. To correct the downward bias of the 0.632 bootstrap (▶ Bootstrapping, 0.632 Bootstrap), Efron and Tibshirani (1997) developed the 0.632 +bootstrap (▶ Bootstrapping, 0.632+ Bootstrap).

Figure 2 illustrates the relation between the leave-one-out bootstrapped estimate and the bootstrapped resubstitution error in a simplified example. To calculate the contribution of case 1 to the leave-one-out bootstrapped estimate, the prediction of only model C_2 is relevant because it is derived from a bootstrap sample that does *not* contain case 1. In contrast, to calculate the contribution of case 1 to the bootstrapped resubstitution error, the predictions of only models C_1 and C_3 are used because only the bootstrap samples #1 and #3 contain case 1.

Advantages and Disadvantages of the Bootstrap

The bootstrap is one of the most commonly used resampling strategies for assessing the reliability of performance measures in classification tasks, specifically when the data sets are relatively small. This is generally the case for genomic data sets where the number of features (e.g., probe sets on a microarray) is orders of magnitude smaller than the number of specimens.



Bootstrapping, Fig. 2 Illustration of the leave-one-out bootstrapped estimate and the bootstrapped resubstitution error. The data set contains $n = 10$ cases; here, only the case indices are shown. From the original data set D, three bootstrap samples are drawn, and three classifiers (C_1, C_2, and C_3) are built

The non-parametric bootstrap method does not rely on any assumptions about the data distribution. Therefore, bootstrapping can be applied to real-world data sets for which specific distributional assumptions are questionable and classic statistical procedures thus problematic. However, it is also possible to include distributional assumptions into the sampling, for example, we may assume that the data follow approximately a normal distribution. Or, on the basis of our prior beliefs, we may assign different prior probabilities to the individual cases for being selected for a bootstrap sample. The bootstrap is then a parametric method. Thus, the bootstrap can be performed either with or without a predefined probability model (Davison and Hinkley 1997). A particular advantage of the bootstrap is that its implementation is straightforward. The language and environment ► R, for example, provides several functions for parametric and non-parametric bootstrapping (see, e.g., the R library boot). Specifically, R provides functions for computing a variety of bootstrapped confidence intervals (see, e.g., the R function boot.ci). For instance, using the percentile bootstrap, we derive a confidence interval for our statistic of interest as follows: (1) we sample several hundred bootstrap samples with replacement; (2) we calculate the statistic for each sample; and (3) we use the $\alpha/2$ and $(1 - \alpha/2)$ percentiles of the distribution to obtain an empirical $(1 - \alpha/2)$-level confidence interval. For many real-world data sets, the assumptions underlying conventional confidence intervals are violated. Specifically for small data sets, the assumption of an asymptotic distribution may not hold. Bootstrapped confidence intervals are then an interesting alternative.

The bootstrap requires multiple samplings of the data set. The computational costs may be considered a disadvantage, specifically if all possible subsamples are generated. The exhaustive subsampling, however, is generally not necessary. In the context of performance assessment, bootstrapped estimates of error rates tend to be optimistically biased, specifically for the leave-one-out bootstrap and the 0.632 bootstrap (Molinaro et al. 2005). For small data sets, the 0.632+ bootstrap performs competitively with leave-one-out cross-validation and ten-fold cross-validation (Simon 2007). Although bootstrapping is often recommended for small sample scenarios, it is no panacea for a lack of data (Jiang and Simon 2007; Isaksson et al. 2008).

Cross-References

- ▶ Bootstrapping, 0.632 Bootstrap
- ▶ Bootstrapping, 0.632+ Bootstrap
- ▶ Classification
- ▶ Cross-Validation
- ▶ Learning, Attribute-Value
- ▶ Model Testing, Machine Learning
- ▶ Overfitting
- ▶ R, Programming Language

References

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, New York

Dixon PM (2006) Bootstrap resampling. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of environmetrics. Wiley, New York, pp 212–220

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

Efron B (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc 78 (382):316–331.

Efron B, Tibshirani R (1997) Improvement on cross-validation: the 0.632+ bootstrap method. J Am Stat Assoc 92:548–560.

Efron B, Tibshirani R (1998) An introduction to the bootstrap. Chapman & Hall, London

Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning. Springer Series in Statistics, New York/ Berlin/Heidelberg

Isaksson A, Wallman M, Goeransson H, Gustafsson MG (2008) Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recogn Lett 29: 1960–1965

Jiang W, Simon R (2007) A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. Stat Med 26:5320–5334

Künsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann Stat 17:1217–1241

Molinaro AM, Simon R, Pfeiffer RM (2005) Prediction error estimation: a comparison of resampling methods. Bioinformatics 21(15):3301–3307

Silverman BW, Young GA (1987) The bootstrap: to smooth or not to smooth? Biometrika 74:469–479

Simon R (2007) Resampling strategies for model assessment and selection. In: Dubitzky W, Granzow M, Berrar D (eds) Fundamentals of data mining in genomics and proteomics. Springer, Heidelberg

Bootstrapping, 0.632 Bootstrap

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

Definition

The 0.632 bootstrap weighs the bootstrapped resubstitution error (▶ Bootstrapping), $\hat{\varepsilon}^{resub}$, and the leave-one-out bootstrapped estimate of the prediction error (▶ Bootstrapping), $\hat{\varepsilon}^{loob}$, as follows:

$$\hat{\varepsilon}^{.632} = 0.368 \;\; \hat{\varepsilon}^{resub} + 0.632 \;\; \hat{\varepsilon}^{loob} \tag{1}$$

This weighted average corrects the upward bias of the leave-one-out bootstrapped estimate. However, $\hat{\varepsilon}^{.632}$ can now be biased downward (Hastie et al. 2008). Consider the example of a classification problem with two balanced classes where the class labels are independent of the data set attributes. For this data set, the true prediction error of any classifier cannot be smaller than 0.5. However, consider the 0.632 bootstrap estimate for the 1-nearest neighbor classifier (Hastie et al. 2008). Its resubstitution error is zero because 1-NN uses the class label of the duplicated training case to predict the class label of the corresponding test case in D. The leave-one-out bootstrapped estimate is 0.5 because for those cases that are not included in the bootstrapped samples, 1-NN performs like a random guesser. Thus, the 0.632 bootstrapped estimate of the prediction error is $\hat{\varepsilon}^{.632} = 0 + 0.632 \cdot 0.5 = 0.316$, which underestimates the true error rate of 0.5 in this example. To correct the downward bias of the 0.632 bootstrap, Efron and Tibshirani (1997) developed the 0.632+ bootstrap (▶ Bootstrapping, 0.632+ Bootstrap).

Cross-References

▶ Bootstrapping, 0.632+ Bootstrap
▶ Bootstrapping

References

Efron B, Tibshirani R (1997) Improvement on cross-validation: the 0.632+ bootstrap method. J Am Stat Assoc 92:548–560

Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning. Springer Series in Statistics, New York/Berlin/Heidelberg

Bootstrapping, 0.632+ Bootstrap

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

Definition

The 0.632+ bootstrap adjusts the weights of the 0.632 bootstrap (▶ Bootstrapping, 0.632 Bootstrap) as follows (Hastie et al. 2008).

$$\hat{\varepsilon}^{.632+} = (1-w)\ \hat{\varepsilon}^{resub} + w\ \hat{\varepsilon}^{loob},\ \text{with}$$
$$w = \frac{0.632}{1-0.368R}\ \text{and}\ R = \frac{\hat{\varepsilon}^{loob} - \hat{\varepsilon}^{resub}}{\hat{\varepsilon}^{random} - \hat{\varepsilon}^{resub}} \quad (1)$$

Here, $\hat{\varepsilon}^{resub}$ is the bootstrapped resubstitution error (▶ Bootstrapping), and $\hat{\varepsilon}^{loob}$ is the leave-one-out bootstrapped estimate of the prediction error (▶ Bootstrapping). $\hat{\varepsilon}^{random}$ is the estimate of the prediction error under the assumption that the class labels are independent of the data set attributes (▶ Learning, Attribute-Value). R is an estimate of the degree of ▶ overfitting. Hence, the weights for the 0.632+ bootstrap are not fixed as in the 0.632 bootstrap but determined based on the degree of ▶ overfitting. Efron and Tibshirani (1997) developed the 0.632+ bootstrap to address the downward bias of the 0.632 bootstrap (▶ Bootstrapping, 0.632 Bootstrap).

Cross-References

▶ Bootstrapping
▶ Bootstrapping, 0.632 Bootstrap
▶ Learning, Attribute-Value
▶ Overfitting

References

Efron B, Tibshirani R (1997) Improvement on cross-validation: the 0.632+ bootstrap method. J Am Stat Assoc 92:548–560

Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning. Springer Series in Statistics, New York/Berlin/Heidelberg

Bottleneck Enzyme

▶ Rate-limiting step

Bottom-Up Determination

▶ Interlevel Causation

Bottom-Up Hierarchical Clustering

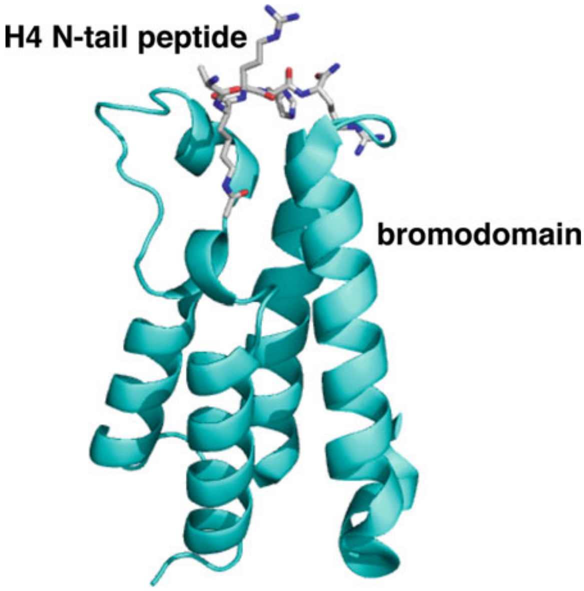
▶ Hierarchical Agglomerative Clustering

Bromodomain

Toshiya Senda[1] and Naruhiko Adachi[2]
[1]Biomedicinal Information Research Centre (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
[2]Structure-guided Drug Development Project, JBIC Research Institute, Japan Biological Informatics Consortium, Tokyo, Japan

Definition

Bromodomains, initially identified in the *Drosophila* protein Brahma, is an acetylated histone recognition domain and found in various chromatin factors. ATP-dependent nucleosome-remodeling factors and histone acetyltransferases frequently contain bromodomains (Allis et al. 2006). Typical bromodomains are composed of approximately 110 amino acid residues. Tertiary structure analyses showed that the bromodomain adopts a four-α-helix-bundle structure and has a cavity at the top of the molecule to accommodate an acetylated lysine residue (Fig. 1).



Bromodomain, Fig. 1 Crystal structure of bromodomain in complex with an N-tail peptide of histone H4 (PDB code: 1E6I). Bromodomain and the N-tail peptide of histone H4 are shown in cartoon (*cyan*) and stick models (carbon atoms in *white*), respectively

A tandem repeat type bromodomain, which is designated as a double bromodomain, was first found in the TAF1 (also called as CCG1 or TAF(II)250) subunit of TFIID. This double bromodomain interacts with histone chaperone CIA/Asf1. The Rsc4 subunit of the RSC complex, which is an ATP-dependent nucleosome-remodeling factor, also contains tandem bromodomains. The physical interactions between the bromodomain and factors for the nucleosome structural change suggest that histone acetylation affects nucleosome structural change.

Cross-References

- ▶ Histone Post-translational Modification to Nucleosome Structural Change

References

Allis CD, Jenuwein T, Reinberg D, Caparros M-L (2006) Epigenetics, 1st edn. Cold Spring Harbor Laboratory Press, New York

Brownian Ratchet

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

Definition

Brownian Ratchet is a model mechanism for translocation of RNA polymerases. In this model, the transcript molecule, complexed with the template DNA, is supposed to thermally move forward and backward rapidly in the catalytic site of RNA polymerase. Thus, an eventual big backward movement makes the backtracked complex, with the 3′-end of the transcript molecule protruding from RNA polymerase. A big forward movement dissociated the transcript in abortive synthesis. In normal elongation, a big backward movement is hampered by the ratchet of the collision with the bridge helix, and a big forward movement by the reaction with a substrate nucleoside triphosphate (NTP), which is inserted into a space occasionally formed between the helix and the 3′-end of the transcript. In this model, a transcription complex with the transcript molecule at any given position is merely a conformation but no longer a chemical species. Thus, there must be neither a chemical transition state nor an activation energy of translocation, and the system should be described in dynamics rather than kinetics.

Cross-References

▶ Transcription in Bacteria

Building Blocks of Gene Regulatory Network

▶ Network Motifs of Gene Regulatory Networks