

## Data Mining, a Tool for Systems Biology or a Systems Biology Tool

Nicolas Turenne

INRA, Unité Mathématique Informatique et Génome UR1077, F-78350 Jouy-en-Josas

Received June 01, 2009; Accepted June 15, 2009; Published July 05, 2009

**Citation:** Nicolas T (2009) Data Mining, a Tool for Systems Biology or a Systems Biology Tool. J Comput Sci Syst Biol 2: 216-218. doi:10.4172/jcsb.1000034e

**Copyright:** © 2009 Nicolas T. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

In the past ten years, we have witnessed revolutionary changes in biomedical research and biotechnology. There has also been an explosive growth of biomedical data, ranging from those collected in pharmaceutical studies and life-science investigations, to those identified in “omics” research by discovering sequential patterns, gene functions, and protein-protein interactions. The rapid progress of biotechnology and biological data analysis methods has led to the emergence and fast growth of a promising field, namely, systems biology. Systems biology is based on the understanding that the behavior of the whole is greater than would be expected from the sum of its parts. The field is not new, but is regaining some interest for network analysis as a reachable but mysterious, large-scale genomic structure. Thus, the ultimate goal of systems biology is to predict the behavior of the whole system on the basis of the list of components involved. On the other hand, recent progress in data mining research has led to the development of numerous efficient and scalable methods for mining interesting patterns and knowledge in large databases, ranging from efficient classification methods to clustering, outlier analysis, frequent, sequential, and structured pattern analysis methods, and visualization and spatial/temporal data analysis tools.

Data mining is a branch of computing that aims to explore databases, with a view to exploiting useful similarities and links inside contexts. It can be applied to biological data in three ways.

1. Experimental high-throughput data (as screening, microscopy images, micro-arrays) exploited by inference methods for network reconstruction.

2. Since, at present, no unique experiment is able to catch all interactions at the same time, and thousands of publications containing biological facts are available, analysis of scientific literature, coupled with gene ontology, can help genome annotation or network reconstruction, too.

3. Many relational, biological public databases are now

available, hence access and navigation are common user issues addressed in information systems, to which visualization methods become a possible way for user-friendly visual exploration.

Let us illustrate with a concrete example of data mining for network understanding. In 2004 a team from the University of Colorado developed an algorithm, PathMiner, based on heuristic search, to extract, or infer, biotransformation rules from the Kyoto Encyclopedia of Genes and Genomes (KEGG), a web-accessible database of pathways, genes and gene expressions. Using KEGG, the team inferred 110 biotransformation rules about what happens when certain compounds interact. They used these rules, as well as mathematical algorithms, to predict how detoxification pathways would metabolize ethyl and furfuryl alcohol. The model's prediction is correlated with known patterns of alcohol metabolism.

Automated data mining tools are well on their way to development, as searching literature databases shows. We tried to make a text collection from the Web of Science database with a double set of keywords crossing the fields of data mining and systems biology. Using keywords about systems biology (such as “pathway”, “regulatory network”, “protein interaction”, “regulatory network”, “systems biology” or “biological network”), and about data mining (such as “large database”, “amount of data”, “high throughput”, “knowledge discovery”, “mining”, “knowledge extraction”, “information extraction” or “representation”) we retrieved, without difficulty, more than 5,300 papers between 1994 and 2009. The growth is noticeable after 2002, in particular, where the words “systems biology”, “networks” and “gene ontology” emerge in the top ten most-used keywords. Almost half of publications have been published in the past three years. If genetic issues are widespread in papers, we can work out a few species, only 37, catching attention. The major species are as follows:

- For plants, arabidopsis, rice, lotus and cacao;

- For fungi: yeast;
- For prokaryotes: *B. subtilis*, *T. brucei*, *M. aurum*, *M. grisea*, *C. glutanicum*, *M. tuberculosis*, *E. coli*, *P. falciparum*, *P. syringae*, *S. choleraesuis*, *P. gingivalis* and *D. vulgaris*;
- For eukaryotes: human (cancer and hiv diseases), nematode, fly, pig, mouse, rat and zebrafish.

Habits in large-scale studies remain the same as in traditional molecular studies. It is due essentially to lack of massive production of data. But, as will be seen below, new technologies have been developed and shall become a challenge for non-model species investigations.

Different techniques of data mining are applied to network analyses, even if some approaches are not traditional to the knowledge discovery field, which absorbs easily and continuously any statistical computational method for database and knowledge extraction. Two cases can be distinguished, namely, network structure is known or is not known. In the first case, visualization and network inference approaches are generally used. To understand the structure and regularities of the network, a complex systems approach (social networks and network motifs) is typical and largely used for graphical analysis, as signal-transduction networks. Large-scale graphical visualization (spectral, Boolean, and sparse representation) has been used recently to permit important clues in identification, in terms, for instance of dense graph components such as protein hubs, of transcription factors. More recently, low-complexity approaches, such as decision trees, have been studied for visual drug delivery. The main approaches are aimed at extracting functional parts of the network and its topological and statistical properties. If a network structure is not known, there are several methods that capture knowledge from high-throughput data.

Traditional statistical data analysis methods, in addition to artificial intelligence techniques, have helped, for some time, to network reconstruction such as probabilistic Bayesian inference (and its Naive Bayes variant), artificial neural networks, clustering (using correlation or mutual information metrics) and logistic regression. For the textual analysis, categorization tasks exploit mainly discrete Markov models and dictionary-based methods (for syntactic parsing), inductive-based methods (for multi-relational data) and adjacency matrix approaches (for protein names co-occurrences such as kernel-based or maximum entropy classifiers). In such reconstruction methods, evaluation often drives quality of performance, counted as a noise ratio with the help of false positives, false negatives, expert knowledge and, when available, gold standards. Biological data are noisy and principal component analysis is used for dimensionality

reduction. Inductive-based (supervised learning, making evaluation easily computable) and matrix formulation (unsupervised learning) methods are quite different variants, but semi-supervised learning becomes an alternative.

Apart from network considerations, data mining can be implemented to supply differential equations models, achieving benefits, for instance, from genetic algorithm and fuzzy logic in the case of a multi-objective evolutionary-simplex approach. Representation is also a key concept for modeling, especially for ontology conception. Emphasis on data sharing and interoperability gives impulse to ontological representation of cells or spatio-temporal saliency to anticipate the nature of knowledge to grab from texts or to share in databases. One particularly important research question in the bio-text mining area is how terminological resources, such as ontologies, can best support information retrieval (IR) and information extraction (IE) solutions and vice versa. In theory, we can expect that large, terminological resources cover well the domain knowledge and efficiently contribute to one basic information extraction step, i.e. to named entity recognition, in both IR and IE. In reality, conceptual resources, such as ontologies, form poor terminological resources, since they have never been designed to serve this purpose. From a text mining perspective, they fall short of covering a significant part of the domain knowledge, i.e. they are still sparsely populated, and do not incorporate morphological and syntactical variability; again, this is not the purpose of an ontological resource. Ontologies are not designed to support text mining but, rather, to improve the annotation of database content. Although text mining solutions intend to fill databases with content, it is not the case that a text mining solution finds ontological concepts easily in the literature. Furthermore, ontological resources are not designed to support text mining solutions, in the sense that ontological terms fit the demands of a natural language processing system. However, the text mining community exploits ontological resources to link generated evidence from the literature to the ontological concepts, and biological researchers put significant effort into the development of increasingly complete ontological resources. Text mining makes use of standalone techniques, domain-independent machine learning and natural language processing. A drawback, however, is that many current systems in the life sciences use very little linguistic information, i.e., typically, only word stems or part-of-speech tags. This may lead to misinterpretations of generated evidence, since, for instance, negations and subject-object relationships are ignored. Using more linguistic information is, therefore, an obvious possibility to improve systems, especially as tools for generating such information, in principle, are available in the computational linguistics (CL) community. If such attempts seem

promising, they report disappointing results. The CL community suffers from a lack of data standards and ontology updating. Terminological normalization and systematic integration of a systems biology markup language should provide some helpful orientation.

Data mining is data-dependent, not only for text mining, but also for biological data. Global Sequencing is a new high-throughput sequencing technologies open challenge for data-mining applications. Since 2004, massively parallel DNA sequencing technologies (MPS) have exploded onto the scene, offering dramatically higher throughput and lower per-base costs than had previously been possible with electrophoretic sequencing. Application of this generation and next-generation sequencing will allow for sequencing 1,000 human genomes, characterizing thousands of transcriptomes and microbial diversities within a few years with unprecedented depth and resolution. Tens of millions of sequencing tags can now be obtained at a cost similar to what tens of thousands used to cost. Next-generation technologies are coming.

Over the past year, implementations of MPS have been applied to profile protein-DNA interactions, cytosine methylation, genetic variation, genomic rearrangements, transcriptomes and biodiversity studies. Such platforms as the Roche (454) GS FLX sequencer, Illumina genome analyser and the Applied Biosystems SOLiD sequencer, are able to produce millions of short-length sequence reads. The first type of output is suitable for genome resequencing, the whole transcriptome acquisition, microRNA discovery, methylation inference, ChIPSeq experiments and SNP discovery. The second type of output would be useful for the whole genome sequencing, assessing of structural re-arrangements and DNA copy number alterations, as well as for SNP dis-

covery. Such data should give a good impulse to data-mining methods usage for a large set of species, since, for instance, more than 200 hundred bacterial species populate the human body and need to be sequenced and studied as a whole. The challenge here can be robust, and optimal methods are needed to enhance low training, good computational complexity and working on the fly with a flow of data. Perspectives can also be network comparison from multi-conditional sequencing experiments.

As mentioned in the introduction, the goal of systems biology relies on capabilities of prediction. State-of-the-art methods are related to formal methods of dynamics systems area, such as Monte-Carlo stochastic simulation or ordinary differential equations. Poincaré, at beginning of the twentieth century, showed that such equation systems are not able to produce exact predictions with interactions between components of a modeled system. Von Bertalanffy, in the 1950s, discovered that life organisms can be seen as open systems, being non-chaotic and less dependent on initial conditions because of internal regulations. Some questions, however, remain hard to answer, for instance, which genes and interactions are required to be included in a model, how to estimate kinetic parameters and how to select a particular solution of the model hypothesis space. A combination of hypotheses from data-mining, *in silico* experimentation from simulations, and wet-laboratory validation will make the systematic identification of useful genes in pathways.

## Resources

Bioconductor Project (data mining and micro-array processing) <http://www.bioconductor.org/>

BioCreative Project (bio text mining) <http://www.biocreative.org/>