



Bioinformatics in the post-genomics era: Towards Biology of the Systems

Ali Masoudi-Nejad

Laboratory of Bioinformatics & Bioknowledge Systems

Institute of Biochemistry & Biophysics. Tehran University, Iran

E-mail: amasoudin@ibb.ut.ac.ir

Abstract

A grand challenge in genome science of the 21st century is to computationally predict systemic functional behaviors of the cell, organism, and the ecosystem from genomics and molecular information, especially for the purpose of medical, industrial, and other practical applications. This will require new bioinformatics, not simply for screening the large-scale data, but rather for reconstructing the system and computing its interactions with environment. While traditional genomics and other types of -omics approaches have contributed to our understanding of the genomics space of possible genes and proteins that make up the biological systems, new chemical genomics initiatives will give us glimpse of the chemical space of possible compounds and reactions and pathways that exist as an interface between the biological systems and the natural environment. The goal of this approach is to develop a ‘Life Simulator’, which will be attained, step by step, hierarchically from subsystem simulators of subcellular mechanisms, whole cell simulators, cell development simulators, organ simulators, physiological simulators, pathological simulators to body simulators.

An ultimate goal of bioinformatics in next decade is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity, such as molecular interaction networks involving various cellular processes and Phenotypes (morphological, physiological, and behavioral aspects) of entire organisms.

To increase our understanding of cellular processes from genome information, pathway database, for example KEGG and EGENES have been created in the past decade. Whereas most databases concentrate on molecular properties (for example, sequences, 3D structures, motifs and gene expressions), these databases tackle complex cellular properties, such as metabolism, signal transduction and cell cycle, by storing the corresponding networks of interacting molecules in computerized forms, often as graphical pathway diagrams. In this review we will introduce KEGG and EGENES as an example for knowledge based system for efficient analysis of genes and ESTs, linking genomic information with higher order functional information in a single



پنجمین همایش ملی بیوتکنولوژی جمهوری اسلامی ایران
3-5 آذر ماه 1386، سالن اجلاس سران
The 5th National Biotechnology Congress of Iran
24-26 Nov, 2007, Summit Meeting Conference Hall, Tehran- Iran



database. The higher order functional information in these databases are stored in the PATHWAY database, which contains graphical representations of cellular processes, such as metabolism, genetic information processing, environmental information processing, and cellular process.

Bioinformatics in the post-genomics era: Towards Biology of the Systems

Ali Masoudi-Nejad

Laboratory of Bioinformatics & Bioknowledge Systems
Institute of Biochemistry & Biophysics. Tehran University, Iran

In the past decade, bioinformatics has become an integral part of research and development in the biological sciences. Bioinformatics now has an essential role both in deciphering genomic, transcriptomic and proteomic data generated by high-throughput experimental technologies and in organizing information gathered from traditional biology. Sequence-based protocols for analyzing individual genes or proteins have been elaborated and expanded, and different methods have been developed for analyzing large numbers of genes or proteins simultaneously. Thanks to innovations in high-throughput measurement technologies and information technologies, genome-wide analysis is becoming available in a broad range of research fields from DNA sequences, gene and protein expressions, protein structures and interactions, to pathways or networks analysis. **To date, the genome sequence for over 660 different species have been finished and sequencing of 1500 other prokaryotic and 854 eukaryotic genomes is under development.**

Until recently, there was no common terminology for the different aspects of function. The first steps towards a common vocabulary for protein function have been taken by the Gene Ontology Consortium so that functional features can be compared and described better. The Gene Ontology Consortium categorizes currently accumulated and dynamically changing knowledge into three systematic terminologies or ‘ontologies’: the ‘molecular function’ of an individual protein; the ‘biological process’ in which a protein is involved; and the ‘cellular component’ in which a protein functions. Identification of gene functionality has started a new level of bioinformatics research in post-genome era: automated reconstruction and comparison of pathways of newly sequenced organisms by analyzing of global network of reactions catalyzed by enzymes. This approach uses the knowledge of known biochemical pathways and enzymes, identifies the enzyme function of new genes in a newly sequenced genome using BLAST based search or using pair-wise genome comparison of evolutionary close genomes, and matches the product and substrate of chemical reactions catalyzed by enzymes to build the network of reactions.

In contrast to the sequence, which is a simple one-dimensional object, the network of interacting molecules is represented as a complex graph object. Mathematically, a graph is a set of nodes and edges and, depending on what is to be taken as a node, different types of graphic objects can be defined. For example, the



protein sequence is a graph object consisting of amino acids (nodes) connected by peptide bonds (edges), whereas the protein 3D structure is a graph object consisting of atoms (nodes) and atomic interactions (edges). To understand higher functions, it is necessary to consider 'higher' graph objects: the KEGG/EGENES database consists of three such objects called 'protein network', 'gene universe' and 'chemical universe', for which the nodes are proteins, genes and chemical compounds, respectively. These databases of higher graph objects have paved the way for developing graph algorithms, such as those for detecting local graph similarities among pathways, expression profiles and genomic contexts. The complexity of network topology arises from complex patterns of connections (interactions) and not simply from the size of the network (measured by the number of nodes). This may have biological implications, especially in view of the surprisingly few genes found in the human genome. Graphs and patterns of node connections are static in nature. Predicting network dynamics is far more difficult than simply predicting connection patterns, as has been accomplished in metabolic reconstruction. Here again, by designing high-throughput experiments that systematically perturb dynamic environments and collecting enough experimental data, network dynamics may become computable, at least for dynamic changes in response to small environmental perturbations.

In the past decade, bioinformatics was characterized by the development of innovative computational methods to help generate and analyze various large-scale data and by the creation of new databases of biological knowledge as a direct result of the large-scale analyses. We consider that this is only the beginning of the path to our ultimate goal of understanding the basic principles underlying the complexity of living cells and organisms. Enumeration in biology is no longer limited to the lists of molecular parts such as genes (genome), mRNAs (transcriptome), proteins (proteome) and metabolic compounds (metabolome). More extensive lists include the interactome, which incorporates sets of protein-protein interactions, and the localizome, which describes the subcellular localizations of proteins. The repertoire of different lists will continue to grow as high-throughput experimental methods are further elaborated and expanded. Of course, on its own the bottom-up approach from large-scale data in genomics and proteomics will not be sufficient for understanding the higher complexity of biological systems. Efforts to computerize our knowledge on cellular functions, at present either by the controlled vocabulary of Gene Ontology or by graph representation in KEGG, will both facilitate the computational mapping of genomic data to complex cellular properties or detect any empirical relationships between genomic and higher properties. Although the field is already looking forward to a 'systems biology' approach and to simulations of whole cells, much of the effort must be devoted to capturing even higher properties, such as ontology for human diseases and the computable representation of cellular networks. In addition, the dependence of functionality on the context (such as experimental conditions, cell status and environment) is currently mostly ignored; in other words,

several other levels of complexity will have to be considered before we can come to a more basic understanding of life as a series of complex information systems.

In parallel with the data-driven research approach that focuses on speedy handling and analyzing of the huge amount of data, a new approach is gradually gaining power. This is a 'model-driven research' approach, which incorporates biological modeling in its research framework. Computational simulations of biological processes play a pivotal role. By modeling and simulating, this approach aims at predicting and even designing the dynamic behaviors of complex biological systems, which is expected to make rapid progress in life science researches and lead to meaningful applications to various fields such as health care, food supply and improvement of environment. Model-driven research takes the approach that sets up a biological model by combining the knowledge of the system with related data and simulates the behavior of the system in order to understand the biological mechanism of the system. It is simply called, "Systems Biology". The living body is composed of numerous subsystems. These include various subsystems, large and small, by which the flows of energy, material and information are controlled. This is a hierarchical system working consistently with many metabolic systems, transcriptional control systems, signal transduction systems, cell cycles, apoptosis systems, various physiological and pathological systems, organ systems, and other systems from the molecular level, cellular level, tissue level, organ level to the body level. Systems biology aims to model and simulate such various systems and visualize the results for the better understanding of life mechanisms.

There are some important features and merits of system biology approach. One of the aims is to take important knowledge in the form of qualitative biological theories and try to express this as explicitly and quantitatively as possible. Thus implicit knowledge can be transferred to become explicit knowledge and disparate human knowledge can accumulate in an integrated way. This approach also tries to model the dynamic behavior of the system. Life systems are inherently dynamic, but papers or books cannot fully express the dynamism. Computer models can handle and visualize such dynamic behavior. Thirdly, this approach will make us recognize the lack of knowledge through model building. There are many unknown pathways or mechanisms and also unknown parameters that govern the mechanisms. Conducting research or measurement of those unknown regions per se is one of the merits. Simulation can identify missing components. Furthermore, we can propose appropriate design of experiments with the prediction of results from such simulations.

Until now, bioinformatics has been a practical discipline through which to meet the needs for informatics technologies in large-scale data production in genomics and other high-throughput areas of biology. But as data are converted to knowledge and empirical rules lead to principles, bioinformatics is bound to become a more fundamental discipline. Bioinformatics in



future will facilitate and quicken the analysis of systemic level behavior of cellular processes, and to understanding the cellular processes in order to treat and control microbial cells as factories. For the last decade, bioinformatics techniques have been developed to identify and analyze various components of cells such as gene and protein function, interactions, and metabolic and regulatory pathways. The next decade will belong to understanding cellular mechanism and cellular manipulation using the integration of bioinformatics, wet lab, and cell simulation techniques. Bioinformatics in the future will encompass not only biology and practical aspects of informatics (computer science), but also mathematics and theoretical foundations to detect the basic architectures of complex biological information systems, and physics and chemistry to integrate physical and chemical principles with biological principles. When we have a complete computer representation of living cells and organisms and know the principles of how they compute, then, in the words of Sydney Brenner, “computational biology will become biological computation”.

To increase our understanding of cellular processes from genome information, pathway database, for example KEGG and EGENES have been created in the past decade. Whereas most databases concentrate on molecular properties (for example, sequences, 3D structures, motifs and gene expressions), these databases tackle complex cellular properties, such as metabolism, signal transduction and cell cycle, by storing the corresponding networks of interacting molecules in computerized forms, often as graphical pathway diagrams. Inevitably, it is necessary to collect data and knowledge from published literature accumulated over many years from traditional studies of biology. At least for metabolic pathways, the past knowledge is relatively well organized in these databases, providing a reference data set for annotating genomes (the ‘metabolic reconstruction’) and for screening microarray and other high-throughput experimental data.

In this review we will introduce KEGG and EGENES as a knowledge based system for efficient analysis of genes and ESTs, linking genomic information with higher order functional information in a single database. The higher order functional information is stored in the PATHWAY database, which contains graphical representations of cellular processes, such as metabolism, genetic information processing, environmental information processing, and cellular process. Functional assignments is a process of linking a set of genes/transcripts in each genome with a network of interacting molecules in the cell, such as a pathway or a complex, representing a higher order biological function.