Minireview

Genomics & Informatics Vol. 1(1) 1-6, September 2003

Challenges and New Approaches in Genomics and Bioinformatics

Jong Hwa Park*[1] and Kyung Sook Han*[2]

[1]MRC-DUNN Human Nutrition Unit, Hills Road, Cambridge, CB2 2XY, England, UK and Object Interaction Technologies Inc. (OITEK), Seoul, Korea

[2]School of Computer Science and Engineering, Inha University, Inchon, Korea

Introduction

The science of biology aims to answer the question 'What is life?' in the most systematic manner ever developed. Modern biologists are, in fact, as philosophers were in the past. Bioinformatics, which will likely to be the future name of biology, employs information processing technology for biology to interpret the whole life process as a complex system with many computable layers of different kinds of elements. The layers can be encapsulated as classes or components of abstract objects for analysis and simulation. Eventually, the layers can form a recursive and self-similar pattern of information processing, providing a commonality in all the levels of life. As in fractal geometry, these patterns are suggested to be present naturally and universally in biology. That is why computable enzyme circuits of metabolic pathways for cancer can be applied to the simulation of bacterial interactions and even the socio-economical behaviors of human beings (such as the Internet). Remarkably, because of the extent of the challenges and the amount of data it produces, biology has proved itself as the richest information field in science. This new field is focused on genomes as they are the most central data source in life on Earth. We will look at the multi-layered problems of genomics and bioinformatics, as mentioned above, with a new paradigm of research called 'network biology' from complex systems analysis approach of computer science.

Keywords: network biology, interaction among biological entities, protein interaction network

* Corresponding authors:
[1]E-mail j@bio.cc http://bio.cc
[2]E-mail khan@inha.ac.kr, Tel +82-32-860-7388, Fax +82-32-863-4386

Abbreviations: SNP, single nueleotide polymorphism; KEGG, Kyoto Encyclopedia of Genes and Genomes

Accepted 17 February 2003

Historical Background

In 1920, the term "genome" was proposed to denote the totality of all genes on all chromosomes in the nucleus of a cell. However, only when Sanger in MRC, Cambridge published the first genomes of virus and mitochondria in 1970s and 1980s, bioinformatists could make the list of protein inventory in cells. Instead of the traditional reductionist' s approach to analyzing small aspects of genes and proteins, biologists started analyzing whole systems of life. Inevitably, this meant more theoretical and computational paradigms were required. In 1995, the first free-living bacterial genome *Haemophilus influenzae* was published (Fleischmann *et al.,* 1995). Bioinformatists could analyse how genes were duplicated within single genome. It can be called 'individual genomics'. Soon, more complete genomes such as *Mycoplasma genitalium* and yeast were published. So, a new field called 'comparative genomics' became possible to study how homologous genes evolved in different organisms. Around this time, based on molecular hybridization technique, very large scale mRNA expression data became available. This let biologists analyse thousands of gene expression within a short time and computationally analyse the significance of molecular interactions between them. This was probably the most important development for another field called 'functional genomics'. Functional genomics focuses on the interactions and subsequent functions of genes. However, before these relatively recent genomics fields, structural biology now known as structural genomics has been the main research in bioinformatics. Since the first 3D structure of proteins solved by Max Perutz and his colleagues in 1950s and 1960s, bioinformatists have been studying the physical mechanisms of protein folding. This involves topics that deal with sequence search and alignment, structure predictions, molecular interactions and docking and functional analysis. One of the aims of such biophysical informatics was to eventually engineer and design proteins for medical usage. In doing so, bioinformatists employed many computational algorithms, which are essentially the founding tools of general bioinformatics we know of. In summary, modern genomics have the following subfields, namely, individual genomics, comparative genomics, functional genomics and structural genomics. All these fields are informatics research by nature (Fig. 1.)

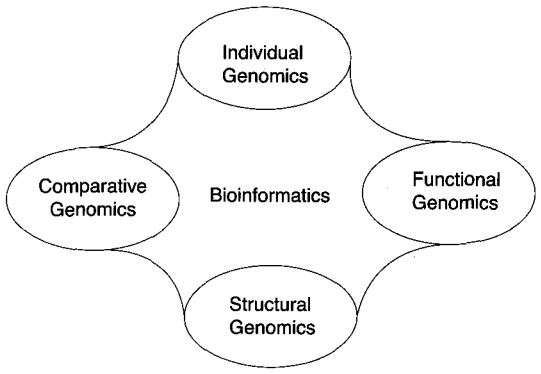

Fig. 1. Genomics fields linked by bioinformatic methods.

Challenging Aspects in Genomics

In each genomics subfields, there are frontier projects where new techniques and insights are needed. For individual genomics the following problems are important.

Individual genomics:

1) Complete mapping of evolutionary relationship between non-redundant gene sets in genomes, such as the human genome. It is estimated that there are about 30,000 genes in human genome. However, most of them have homologues within the genome. Therefore distinctive genes or proteins may not exceed 5,000. In other words, there is a very significant degree of redundancy. Mapping all the redundant genes and subtle differences between them is also very important.
2) Genes are composed of modules. The most useful unit in the study of proteins using bioinformatics is often called domains and motifs. The functional diversity of human body comes from the combinatorial assembly of such modular units. Surprisingly, bioinformatists have found there are only 1,000 kinds of different protein domain structures in nature, which would not exceed 2,000 when proteins are completely catalogued So, elucidating the relationships of these combined domains and functions will result in a kind of biological periodic table.
3) Genome structure (DNA level) and subsequent regulation mechanism. It is not clear why the present organization has been selected for the human genome. This must be related to its control mechanism unless there are some physical restraints. For the understanding of genetic regulation, detailed information on gene organization in relation to expression and metabolic pathway is necessary.

Comparative genomics

1) Reliable and complete set of genome based taxonomy is required. Many proteins share structures and widely spread in many different taxonomic branches of life. By comparing the distribution of proteins, it is possible to map their evolution and to predict their functions . This is a part of the bigger project that is named as 'mapping protein universe' .
2) Comparative structural genomics. Even though protein structures often have a canonical fold, they differ subtly depending on their functions. This is why one protein fold may have dramatically different functions (such as gamma crystalline structure which also works as an enzyme). By categorizing the subtle differences on the surface of proteins, we can make databases of protein functional surfaces for protein domains. This will enhance drug discovery dramatically.
3) Application of comparative genomics to detection and classification of species. The presently known microorganisms are a biased set of culturable species in nature. By developing diagnostic tools using the difference between genomes, it is possible to expand the study of genomes toward organisms that are difficult to grow. Diagnostic genome chips can be an example of this. This may lead us to discovering new pathogens.
4) Small variations in gene sequences and their effect on chemicals. Single nucleotide polimorphism (SNP) is a high-resolution comparative genomics between very closely related genomes such as different human races and primates. This is often regarded as an important research field for possibly different drug response to slightly different genes.

Functional genomics

DNA chips received a lot of attention when first widely introduced. However, the quality of the data they produced was not often reliable for analysis. Also, mRNA expression level that the DNA chips measure is not directly co-related to the expression levels and subsequent effect levels of protein function. Now, protein chips are available and the functional genomics using such DNA with protein chips data will result in much more reliable data.

1) Database management for large scale chip data is not trivial. Standardization is necessary with format conversion. Also, automatic data uploading and downloading servers are necessary. The task should be coordinated by an international organization. Many different conditions used in large scale expression experiment are critical to analysis. Object oriented or relational database systems are suitable for this.
2) Efficient algorithms for clustering, large scale automatic data mining, and visualization are required. Due to the

size of variables involved, it is an absolute requirement to develop a set of very efficient and highly integrated bioinfrastructure.

3) Fully automatic and yet reliable annotation system. The literature information, sequence, structure and other less automated information are to be combined into an automatic annotation system. Recently, text parsing algorithms are actively developed for this purpose. However, the reliability is an important problem yet to be solved. In addition to devising more powerful automatic annotation system such as natural language processing, it is necessary to establish a standard form for storing data.

Structural genomics

Structural genomics is perhaps the core of genome study. This is because structure is the most definite way of representing genetic entity. The major issues in structural genomics are:

1) Completing the list of all the protein folds in nature and classifying them in a biological meaningful scheme. There are numerous protein structure classification systems and they need to be coordinated to be accessed efficiently.
2) Protein structural interface classification. Protein folds interact with other proteins through their surfaces. A major new task of structural genomics is to map all such surfaces according to interactions and functions.
3) RNA structures. RNA is an important part of genomics. However, due to the difficulty in solving the structure of RNA, the problem has not been tackled effectively yet. A major challenge in structural genomics is to develop a very fast and large scale determination method of RNA structures.

Genomics as Network Biology

The layers of biology, such as genomic layer, protein layer and metabolic pathway layer, can be best represented and analyzed as computable networks. All biological entities can be represented as networks, and we posit that networks are the ultimate representation of all life processes. For example, a protein is a network or graph in which nodes represent amino acids and edges represent chemical forces. Each of the 20 amino acids, in turn, can be represented as a network of atoms of carbon, nitrogen, oxygen and so on. This can go down as far as the boundary of matter and non-matter or go up as far as or beyond two humans having a conversation. The conversation can be regarded as information processing with a relatively precise syntax and a highly context dependent grammar, which is essentially the same process as two proteins interacting probabilistically to produce some biological functions. DNA or genome has already been suggested as a dynamic storage of a language system as early as 1980s (Searls, D.B, 1993) with precise computable finite states. Recent research in complex systems (Bianconi and Barabasi, 2001) has also suggested some far reaching commonality in the organization of information in problems from biology, computer science and physics such as the Bose-Einstein condensate, which is a special state of matter.

However, only in the last 5 or so years, has bioinformatics truly shifted its focus from individual genes, proteins, structures and search algorithms to the viewpoint of large-scale networks. Suddenly, biologists find the links between the internet and metabolic pathways, structural interactions of proteins via a network topology (scale-free network (Jeong, *et al.*, 2000). We are becoming more certain that the future of biology lies in the networks of biological entities. Then, what are the challenges and future trends for the network biology? Three main challenges lie ahead in network biology. They are 1) representing biological entities and making databases, 2) mapping the networks efficiently and 3) modeling, simulating, analysing and predicting the networks. The critical problems of all the networks boil down to physical and informational 'interactions' among biological entities. Hence, the above challenges are best tackled by mapping the entities with interaction maps or networks. In building such interaction networks, proteins are the most useful for us. However, only recently have we observed new research outcomes on large scale identification of proteins and mapping of their interactions even though proteomics started as early as 1980s when Sanger group in Cambridge tried to verify the expressed genes from complete genomes. Due to its difficulty and importance, mapping protein interaction is perhaps comparable to the human genome project. Until new technologies for more sophisticated and large-scale detection methods for many different kinds of chemicals in cells, the protein entities and their networks will occupy the core of future bioinformatics research. The following paragraphs show where important information on protein networks come from.

Protein Interaction Networks

The interactions of protein entities (commonly the protein domains and their complex) can be represented differently and found (the first challenge) in many different forms. Four major forms are explained here.

Biological data from literature

The first and most obvious one is the literature of all kinds

of biological fields. Many biological articles provide some degree of protein interaction information. The main problem of this form is that the signal to noise ratio is poor due to massively irrelevant and confusing text bodies. Therefore, they need to be parsed logically, predicted and verified. Artificial intelligence techniques, including natural language processing, are often employed with reasonable success. In the future, with the data mining processes included, the whole literature itself will form a giant biological entity which, in essence, is not much different from a whole genome (a textome, Tsoka and Ouzounis, 2000, or archiome).

Metabolic data source

The second form comes from the metabolic pathways information. Interactions in this case are often linked by biological substrates within directed graphs or circuits of enzymes. A good example of this representation is the KEGG. Practically, this representation is close to the electronic circuits of switches. When, seemingly distributed biological entities interact with each other with a switching mechanism, certain emergent properties occur and the whole circuit becomes alive or starts to control and process the information flow (Walhout, 2000). The physical material for the processing is often associated with energy in chemical forms. Finding regulatory principles and rules is critical to correctly analyzing this form of interaction data.

Genetic interaction data source

The third form of interaction is found in molecular genetics methods such as yeast two hybrid system (Y2H, Walhout, 2000). This method in a massively large scale produces genetically predicted or verified protein interactions. Whole genome scale interaction experiments are now possible (Uetz *et al.,* 2000). Due to the volume of data, a graphical representation of protein-protein interactions has proven to be much easier to understand than a long list of interacting proteins. However, visualizing protein-protein interactions is not easy, even for relatively simple organisms such as yeast. Fig. 2, for example, shows the largest connected component of the Y2H data, visualized by a 3D layout program.

Structural interaction data source

The last and probably the most precise one is coming from the physical and structural interactions between proteins. Proteomics data from mass spectrometry can provide relatively reliable physical interactions of proteins with

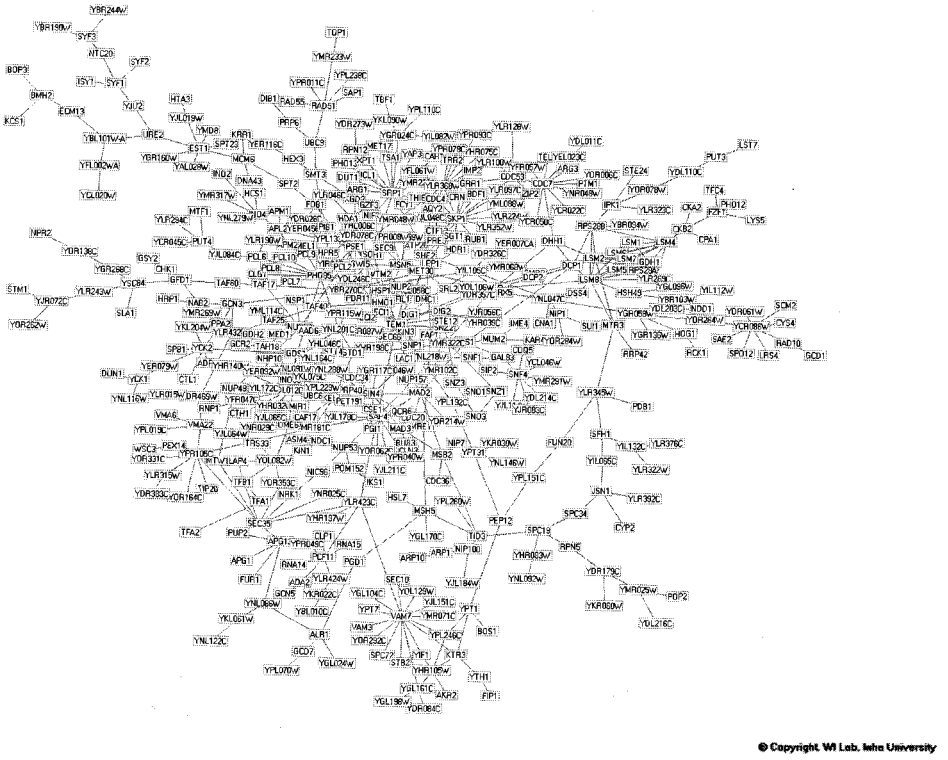

Fig. 2. Network of the largest connected component of the Y2H data, containing 473 nodes and 543 edges.

identification of new proteins. Another valuable source is the PDB (Protein Data Bank), in which 3D coordinate values of protein structures are stored. Using the precise 3D structure, protein interactions can be generalized and drawn in a map, which encompasses all the known protein topologies and their interactions (Uetz *et al.*, 2000). The advantage of this interaction map is that it can reveal the evolutionary paths of interactions as it lies at the protein family level rather than at the individual protein level (Fig. 3).

Not only the above four different sources of interaction networks can form different layers of infrastructure in bioinformatics, but they also overlap and interact with each other, resulting in a super-network of information. This pattern will recurse, eventually forming a tightly yet probabilistically controlled network called life as modeled by human beings (as humans can only model it), whether it is called Gaia, Galaxia or something else.

Summary

In conclusion, the seemingly fuzzy and disorganized data of biology with thousands of different layers ranging from molecule to the Internet have refused so far to be mapped precisely and predicted successfully by mathematicians, physicists or computer scientists. Genomics and bioinformatics are the fields that process such complex data. The insights on the nature of biological entities as complex interaction networks are opening a door toward a generalization of the representation of biological entities. The main challenge of genomics and bioinformatics now lies in 1) how to data mine the networks of the domains of bioinformatics, namely, the literature, metabolic pathways, and proteome and structures, in terms of interaction; and 2) how to generalize the networks in order to integrate the information into computable genomic data for computers regardless of the levels of layer. Once bioinformatists succeed to find a general principle on the way components interact each other to form any organic interaction network at genomic scale, true simulation and prediction of life in silico will be possible.

Acknowledgments

Authors were supported by the Ministry of Information and Communication of South Korea under grant number IMT

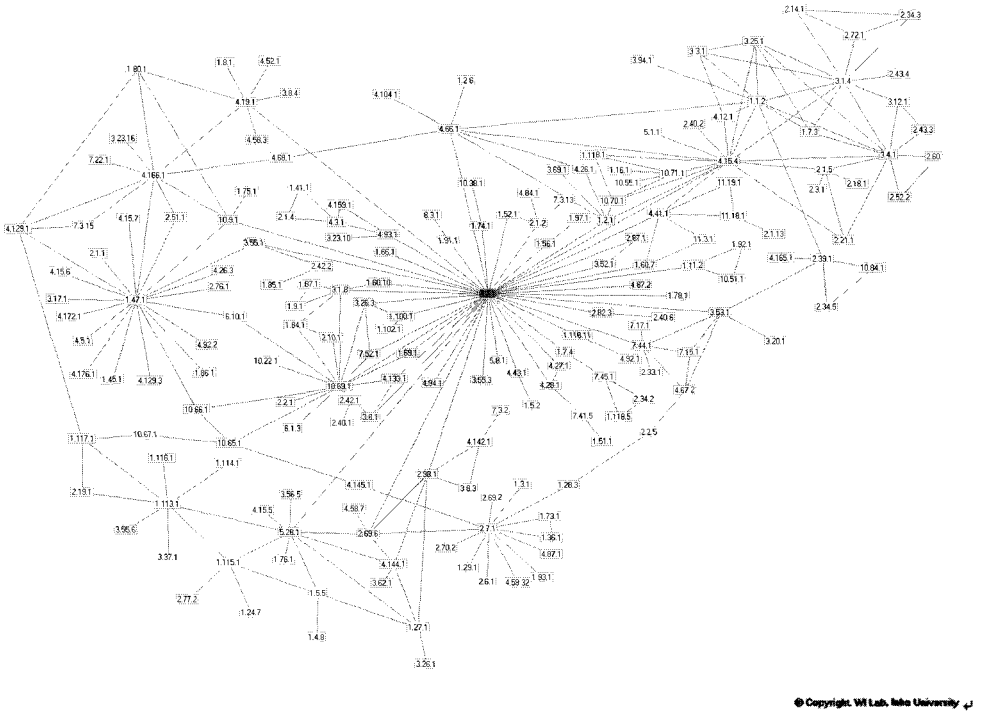

Fig. 3. Part of PSIMAP (Protein Structural Interaction Map, Park et al., 2001), which shows all the known protein fold interactions. The interactions are phylogenetic, i.e., it is based on evolutionarily determined family-family interactions. It works as the basic skeleton of more specific protein-protein interactions. Protein families directly interacting with protein family 3.2.1 (shown in blue color) are highlighted by yellow color.

2000-C3-4.

References

Bianconi, G. and Barabasi, A. (2001) Bose-Einstein condensation in complex networks. Physical Review Letters, 86, 5632-5635.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269, 496-512.

Jeong, H.,Tombor, B., Albert, R., Oltvai, Z.N., and Barabasi, A.L., (2000) The large-scale organization of metabolic networks. Nature, 407, 651-654.

Park, J., Lappe, M., and Teichmann, S.A. (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol., 307. 929-938.

Searls, D.B., (1993) The computational linguistics of biological sequences. In Artificial Intelligence and Molecular Biology (L. Hunter ed.), AAAI Press, The MIT Press, 47-120.

Thieffry, D. and Thomas, R. (1998) Qualitative analysis of gene networks, Pac Symp. Biocomput. 77-88.

Tsoka, S. and Ouzounis, C.A. (2000) Recent developments and future directions in computational genomics. FEBS Letters, 480, 42-48.

Uetz, P., Giot, L., Cagney, G., Mansfield, TA., Judson, RS., Knight, JR., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili A., Li Y., Godwin B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, JM. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403, 623-627.

Walhout, AJ., Boulton. SJ. and Vidal, M. (2000) Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. Yeast, 17, 88-94.