Journal of Engineering
Research and Education
Vol. 5 (2008) 75-88

Bioinformatics: prospect, application and the way forward

Harbant Singh[1], Lim Eng Aik[2]

[1]School of Bioproses, Universiti Malaysia Perlis (UniMAP), Kompleks Pusat Jejawi 3, 02600 Jejawi, Perlis , E-mail: harbant@unimap.edu.my
[2]Insitute of Engineering Mathematics, Universiti Malaysia Perlis (UniMAP), 02600 Jejawi, Perlis , E-mail: ealim@unimap.edu.my

Abstract

The world of bioinformatics is fast expanding into the mainstream of both molecular and computational biology. Bioinformatics is a new discipline that can be described as the science of collecting, modelling, storing, searching, annotating, and analysing biological information. It contains data for complex protein structures such as DNA or RNA that carry out catalysis, sense metabolites and synthesize proteins. The dynamics and structural nature of proteins present a whole new set of informatics challenges to computational community. The ability to relate the traits of protein biomolecules to function in molecular medicine, agriculture, and energy will provide a key to understand this complex world of biotechnology. Advances in nanofabrication of biosensing interfaces in nanotechnology have dramatically impacted on biosensor research in the last few years. Biosensing interfaces has contributed to the rise of self assembles monolayers (SAMs) that give molecular level control over the fabrication of biosensing interface. Nano-wire sensors that range from 1 nm to 100 nm are being used to detect biochemical species at very low concentrations, single viruses and bacteria, DNA sequence variations, small molecular protein interactions and in environmental and food analysis. This paper outlines the current trends and future challenges in bioinformatics.

Keywords: *Bioinformatics, biomolecules, modelling, biosensing, nano-wires*.

INTRODUCTION

Bioinformatics or computational biology is the use of information technology in the field of molecular biology, or the application of computer technology to the management and analysis of biological data. Here, computers are used to gather, analyze and merge biological data for reference of future research in the fields of biology, agriculture, engineering and computer science. It is an emerging interdisciplinary research area and is increasingly being used to improve the quality of life. The ultimate goals of bioinformatics is to uncover the wealth of biological information hidden in the mass of sequence, structure, literature and other biological data in order to obtain a clearer insight into the fundamental biology of organisms and to use this information to enhance the standard of life for mankind.

Bioinformatics could have profound impact on human health, agriculture, the environment, energy and biotechnology, besides enhancing research and development in these areas. It is now being used in the discipline of molecular medicine, to help produce better and more customized medicines to prevent or cure diseases. In addition, it has environmental benefit in identifying waste cleanup bacteria; and in agriculture, it can be used for producing high yielding low maintenance crops such as in the arid regions [1].

WHAT IS BIOINFORMATICS?

Bioinformatics is a new area that can be described as the science of collecting, modelling, storing, searching, annotating, and analysing biological information such as sequence information, evolutionary patterns, prediction of gene function and “silicon-based biology”. It involves a range of activities from data handling, publication, to data mining and analysis. The essential part of bioinformatics is to create new algorithms for the analysis of complex and/or large data sets.

Bioinformatics deals with the issues created by the massive amounts of new types of data obtained through novel biological experiments. The well-known example is the determination of the complete nucleotide sequence of human genome, which has been accomplished [2].

EARLY WORK

Early Work in bioinformatics, as discussed in “*Current Trends in bioinformatics*” [3] focussed mainly in micro arrays. One of the purposes of micro array experiments is to find genes that are either up- or down regulated under specific circumstances. The usual method for detecting such variations within micro array data is to look for genes that have expression levels more than a certain number of standard deviations from the mean or show a several-fold difference between experiment and control measurements.

One of the challenges is to separate relevant signals from the background noise without obtaining too many false positives. Besides, the fact that biological systems are inherently noisy (unnecessary signals data), further errors are introduced by various technical factors. After obtaining a list of up- and down-regulated genes the next step usually involves trying to extract some biological meaning from this data, for example by looking for similar genes in public databases to determine properties such as chromosomal location and functional categories. Micro array data processing involves several steps ranging from image acquisition, image processing, filtering and normalization of raw data to tasks such as data analysis and visualization. Reliable results could be achieved with hierarchical clustering if the process is repeated several times with different starting points and then the most common solution is chosen. Various machine learning approaches, such as neural networks and support vector machines, are also being adopted for the classification of micro array data.

Algorithms and computational challenges

Most of the work on algorithms has been associated with protein databases. To eliminate redundancy and speed up homology, clustering by choosing representative sequences is an efficient simplistic strategy within large protein databases. Corresponding references could be attached to the processed data using certain algorithms with the analysis software.

Protein interactions and pathways

Experimental protein interaction data is obtained on a large scale from yeast two-hybrid experiments. Experiments might vary not only in the proteins they test for interactions but also in how detailed information such as place and time of interactions are recorded and assuming that the interactions should be classified in a binary way or with probabilities. Promoter analysis is another way to detect protein interactions. This technique is based on the fact that co-regulated genes usually produce gene products that lie within the same pathway. Protein interaction maps can be used to establish potential pathways, which are essential for the understanding of gene functions. The function of a gene can often be predicted more reliably by its functional context than its sequence.

Protein function

Structural similarities between proteins are considered to be more relevant than pattern or sequence similarity because structures are known to be more conserved than the underlying sequences. Therefore protein structures rather than sequences should be used for predicting functions.

Clinical and research applications

Micro arrays are widely applied in cancer research. Gene expression data can be linked to clinical data, such as, survival rates and used to determine high-risk patient groups that should receive more therapy, now that the human genome has been sequenced. The mechanism by which repressors control gene expression and insight into the **tumour**-suppressor protein (**pRb**) pathway has been incorporated in micro arrays research. For drug discovery purposes the genome should be seen to consist not only of the basic sequence but also of individual variations.

IMPORTANT TRENDS IN BIOLOGICAL RESEARCH

The most important trend in modern biology is the increasing availability of high-throughput (HT) data. The earliest forms of HT were genome sequences, and to a lesser degree, protein sequences, however now many forms of biological data are available via automated or semi-automated experimental systems. This data includes gene expression data, protein expression, metabolomics (study of metabolism in biological organism), mass spec data, imaging of all sorts, protein structures and the results of mutagenesis and screening experiments conducted in parallel. So an increasing quantity and diversity of data are major trends. To acquire a sound understanding of this biological data, it is necessary to integrate (finding and constructing correspondences between elements) and curate (checked for errors, linked to the literature and previous results and organized) the vast information gathered. The challenges in producing high-quality, integrated data sets are immense and long term.

The second trend is the general acceleration of the pace of asking those questions that can be answered by computation and by HT experiments. Using the computer, a researcher can be 10 or 100 times more efficient than by using wet lab experiments alone. Bioinformatics can identify the critical experiments necessary to address a specific question of interest. Thus, the biologist that is able to leverage bioinformatics is in a fundamentally different performance regime that those that can't.

The third trend is the beginnings of simulation and modelling technologies that will eventually lead to predictive biological theory. Today, simulation and modelling applied at the whole cell level is suggestive of what is to come, the ability to predict an organism's phenotype computationally from just a genome and environmental conditions. This capability is probably five years away for microbial organisms and 10 to 20 years away for complex eukaryotes (such as the mouse and human).

The role of cyber-infrastructure in biological research

As we noted above, modern biology will become increasingly coupled to modern computing environments. This means that rates of progress of some (but not all) biological investigations will become rate limited by the pace of cyber-infrastructure (CI) development. Certainly, it will make it much easier for the biologist to gain access to both data and computing resources (perhaps without them knowing it) once cyber-infrastructure is more developed and in place. Today, we have early signs of how some groups will use access to large-scale computing to support communities by developing gateways or portals that provide access to integrated databases and computing capabilities behind a web-based user interface. But, that is just the beginning. It is possible to imagine that, in the future, laboratories will be directly linked to data archives and to each other, so that experimental results will flow from HT instruments directly to databases which will be coupled to computational tools for automatically integrating the new data and performing quality control checks in real-time (not that dissimilar from how high-energy physics and astronomy work today). In field research, cyber-infrastructure can not only connect researchers to their databases and tools while they are in the field, but it will enable the development of automated instruments that will continue working in the field after the scientists and graduate students have returned home.

Notable accomplishments in applying CI to biology research

There are a handful of systems that have fundamentally changed how biologists work. The most important has been the system developed by the National Center for Biotechnology Information [4] including Entrez, which is a search engine (google like) that supports searching across many types of biological data. There are similar systems like this in Europe [5] and Japan [6]. These systems and systems like them have provided the global community access to sequence data (starting out as outgrowths from genome and protein sequence databases) and more recently to publications, annotations, linkage maps, expression data, phylogeny data, metabolic pathways, regulatory and signally data, compounds and molecular structures. Search techniques have expanded from keywords to computed properties (sequence similarity) that enable one to find connections between biological or chemical entities. While these systems have enormous user bases and require considerable computing capabilities for indexing and integration, they are essentially client/server in nature, and the computing that an end user can request is closely controlled.

Approximately a decade ago a number of groups began to produce more flexible tools that support a more unstructured workflow, enabling the user to construct their own mini-environment to pursue computational approaches to problems. One of the first such systems was the Biology Workbench developed at the University of Illinois and now hosted at the University of California, San Diego [7]. Other systems were developed to provide access to a specific type of data (e.g. microbial genomes) in well engineering data integrations. These systems are often associated with teams of curators. Three are particularly important: the Institute for Genomic Research's Comprehensive Microbial Resource [8]; the SEED, an annotation system developed by the Fellowship for the Interpretation of Genomes at the University of Chicago [9]; and the DOE's Joint Genome Institute's Integrated Microbial Genomes resource [10]. These systems provide the user with an integrated view of hundreds of genomes and provide a rich environment for discovery.

Biosensing in Biological Research

Biosensors commonly comprise a biological recognition molecule immobilised onto the surface of a signal transducer to give a solid state analytical devices. Advances in nanofabrication of biosensing interfaces are one of the two major areas where nanotechnology has dramatically impacted on biosensor research in the last few years. Nanofabrication of biosensing interfaces is on of the two major areas where nanotechnology has dramatically impacted on biosensors research in the last few years. Biosensing interfaces has contributed to the rise of self assembles monolayers (SAMs) that give molecular level control over the fabrication of biosensing interface. The application of new nanomaterials (nanoparticles) or nanotubes (nanoporous materials) to biosensing is currently influencing biological research [19].

Nano-wires in Biological Research

Nano-wire sensors are a new dawn in technology. Nano-wires have been defined as wires with at least one dimension in the range of 1-100 nm. This technology is gaining importance in biosensor research. Single crystalline silicon nano-wire sensor has been used to detect biochemical species at very low concentrations, single viruses and bacteria, DNA and DNA sequence variations, and small molecular protein interactions. The sensor allows real-time and online detection with a quick respond time (collecting real-time interaction data). Nano-wire sensors are also engage in environmental and food analysis [20].

PETASCALE COMPUTING – EARLY WORK AND NOW

Supercomputing has come a long way in the past half-century. Far from CDC's single operation scalar processors in the 1960s, present day terascale computers in development by companies like Intel boast up to 100 processor cores and the ability to perform one trillion operations per second.

Now, academics have turned their attention to petascale computers that are said to be capable of performing one quadrillion - that's *one million billion* - operations per second as discussed in "*Petascale computers: the next supercomputing wave*" [17]. Running at nearly ten times the speed of today's fastest supercomputers, petascale computing is expected to open the doors to solving global challenges such as environmental, sustainability, disease prevention, and disaster recovery.

Petascale Computing is the present state-of-the-art in High Performance Computing that leverages the most cutting edge large-scale resources to solve grand challenge problems in science and engineering. Nowadays, Science has withstood centuries of challenges by building upon the community's collective wisdom and knowledge through theory and experiment. However, in the past half-century, the research community has implicitly accepted a fundamental change to the scientific method. In addition to theory and experiment, computation is often cited as the third pillar as a means for scientific discovery.

Computational science enables us to investigate phenomena where economics or constraints preclude experimentation, evaluate complex models and manage massive data volumes, model processes across interdisciplinary boundaries, and transform business and engineering practices.

Increasingly, cyber-infrastructure is required to address our national and global priorities, such as sustainability of our natural environment by reducing our carbon footprint and by decreasing our dependencies on fossil fuels, improving human health and living conditions, understanding the mechanisms of life from molecules and systems to organisms and populations, preventing the spread of disease, predicting and tracking severe weather, recovering from natural and human-caused disasters, maintaining national security, and mastering nanotechnologies.

Computing biological data often face several fundamental intellectual challenges, such as, highlighted in Table 1 (please refer Appendix). These problems can be associated with formation of the universe, the evolution of life, the properties of matter, genome and population modelling.

WHAT'S NEXT?

The information presented in this paper illustrates the fact that bioinformatics have proven to be a fruitful application domain for protein functional characterization. Bioinformatics systems have started from systems aimed at computational methods to modelling. They have evolved into the modelling systems for research practice, drug design and clinical forecast e.g. glucose level. It is expected that the interest in bioinformatics in the health sciences (e.g. classification of liver disease), agriculture (e.g. precision farming) and pharmaceuticals (e.g. vaccine development) is increasing, as longer life expectancy focuses on health care and human biology. Ultimately, realizing this vision is one of the main goals of bioinformatics. With this in mind, we here, attempt to show a roadmap for the future of bioinformatics and point out some pitfalls that may arise along the way.

Future challenges for bioinformatics

The next couple of years will bring a shift toward producing "smarter" tools for biologists that will increase their efficiency in mining data. We hope to see in the next five years, the development of tools that will take advantage of integrated genomics, proteomics and metabolic pathways data to unravel the regulatory network that controls cell development. High-impact problems as mentioned in Table 2 (please refer Appendix) could be addressed in the next couple of years for further investigations.

CONCLUSION

This paper has described the current state of bioinformatics in research and projected future opportunities and challenges in this discipline. Current research is marked by its richness, with its wide range of application domains and its integration with numerous complementary approaches and technologies. Opportunities abound, include new research focussing, on data mining, development of modelling systems better designed to account for the complexity of bioinformatics and integrating multi-disciplines into clinical and pharmaceutical settings.

We will find a more fruitful niche in bioinformatics, once, while minimising pitfalls, researchers join together to exploit the vast opportunities in this field.

REFERENCES

[1] Anon, Retrieved from www.winentrance.com/career_course/

[2] Kraulis P., (2000) Stockholm Bioinformatic center, 9 Sept 2000 Retrieved from bioinfo.se/kurser/swell/per/bioinfo-general.html

[3] Jain E., (2002, August) Current Trends in bioinformatics, *TRENDS in Biotechnology, Elsevier science, Vol.20 No.8*

[4] National Center of Biotechnology Information Retrieved from http://www.ncbi.nlm.nih.gov/

[5] European Bioinformatics Institute Retrieved from http://www.ebi.ac.uk/

[6] GenomeNet Database Resource Retrieved from http://www.genome.jp/

[7] San Diego Supercomputer Center Biology WorkbenchRetrieved from http://workbench.sdsc.edu/

[8] Comprehensive Microbial Resource Retrieved from http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi

[9] The Seed: an Annotation / Analysis Tool Provided by The Fellowship for Interpretation of Genomes Retrieved from http://theseed.uchicago.edu/FIG/index.cgi

[10] Integrated Microbial Genomes Retrieved from http://img.jgi.doe.gov/cgi-bin/pub/main.cgi

[11] Building A Cyberinfrastructure For The Biological Sciences (CIBIO), *A Bio*

Advisory Committee Workshop, July 14-15, 2003 Retrieved from http://research.calit2.net/cibio/archived/CIBIO_Overview_Report.pdf

[12] Minakshi D., (2005) Recent Trends in Biotechnology, *News, Current Science, Vol. 88, No. 7.*

[13] Genomics: GTL Roadmap, Systems Biology for Energy and Environment, (2005, August) U.S. Department of Energy Office of Science Retrieved from http://doegenomestolife.org/roadmap/index.shtml

[14] Retrieved from http://www.tgbioportal.org/

[15] Grant C. B. and Paula E. S., Bioinformatics: Recent Trends in Programs, Placements and Job Opportunities, *Report to the Alfred P. Sloan Foundation, Grant # 2002-5-58 BCMB, June 2004.*

[16] Stevens, R. "Trends in Cyberinfrastructure for Bioinformatics and Computational Biology," *CTWatch Quarterly*, Volume 2, Number 3, August 2006
Retrieved from http://www.ctwatch.org/quarterly/articles/2006/08/trends-in-cyberinfrastructure- for-bioinformatics-and-computational-biology/

[17] Liz, T. “Petascale computers: the next supercomputing wave”, ITnews: Breaking IT news for Australian Business, 29 Nov 2007.
Retrieved from http://www.itnews.com.au/Feature/4081,petascale-computers-the-nextsupercomputing-wave.aspx

[18] Anon, Ministry of Science, Technology and Innovation (MOSTI) Retrieved from http://ernd.mosti.gov.my/escience/

[19] Gooding J.J., (2008) Nanotechnology and biosensors: From detecting small molecules and drugs to the monitoring of the activity of whole cells. *1st Regional Conference on Biosensor and Biodiagnostics 2008. 21-22 May 2008,* Hotel Nikko Kuala Lumpur, Malaysia.

[20] Faris W., (2008) Nano-wires: A breakthrough for biosensors. *1st Regional Conference on Biosensor and Biodiagnostics 2008. 21-22 May 2008,* Hotel Nikko Kuala Lumpur, Malaysia.

APPENDIX

Table 1. Challenges and suggested solutions in computation

Problems/Challenges	Problems Explanation	Possible Method of Solutions
Study of the origins, function, structure, and evolutionary history of genes and genomes	By studying the details of individual gene history and protein families, we can begin to understand the factors that influence molecular evolution, refine our strategies for building large-scale databases of protein structures, and lay the foundation for understanding the role of horizontal gene transfer in evolution.	Large-scale sequence analysis, sequence-based phylogenic analysis
The structure, function, dynamics, and evolution (SFDE) of proteins and protein complexes	Proteins are the building blocks for biological processes. Using modelling and simulation, we can begin to understand how proteins work, how they evolve to optimize their functions, how complexes are formed and function, and how we can modify proteins to alter their functions.	Large-scale molecular dynamics
Predictive protein engineering	Many processes of interest to the biological community are mediated by proteins, ranging from biocatalysis of potential fuel stocks to the production of rare and unique compounds to the detoxification of organic waste products. Large-scale modelling and simulation can be used to attack the problem of rational protein design, whose solution may have long-term impact on our ability to address, in an environmentally sound manner, a wide variety of energy and environmental problems.	Large-scale molecular dynamics and electronic structure

The SFDE of metabolic, regulatory, and signaling networks	Understanding the function of gene regulation is one of the major challenges of 21st century biology. By employing a variety of mathematical techniques coupled with large-scale computing resources, researchers are beginning to understand how to reconstruct regulatory networks, map these networks from one organism to another, and ultimately develop predictive models that will shed light on development and disease.	Graph-theoretic and network analysis methods and stochastic modelling and analysis techniques
The structure, function, dynamics, and evolution (SFDE) of DNA, RNA, and translation and transcription machinery in the cell	The standard dogma of molecular biology relates the transcription of DNA to messenger RNA, which is then translated to produce proteins. This is the foundation of the information-processing operation in all living organisms. The molecular complexes that mediate these processes are some of the most complex nanomachines in existence. Via large-scale modelling and simulation of protein/RNA complexes such as the ribosome and the splisosome, we will improve our understanding of these fundamental processes of life.	Large-scale molecular dynamics and stochastic modelling
The SFDE of membranes, protein and ion channels, cell walls, and internal and external cellular structures	Membranes are the means that nature uses for partitioning biological functions and supporting complexes of proteins that are responsible for supporting the cell's ability to interact with its neighbours and the environment. Large-scale modelling is the means by which we can understand the formation, function, and dynamics of these complex molecular structures.	Large-scale molecular dynamics and mesoscale structural modelling

Whole-genome scale metabolic modelling	With the number of completed genome sequences reaching 1,000 in the next few years, we are on the verge of a new class of biological problem; reconstructing the function of entire genomes and building models that enable the prediction of phenotypes from the genotype. With petascale modelling it will become feasible to quickly produce a whole genome scale model for a new sequenced organism and begin to understand the organism’s lifestyle prior to culturing the organism	Linear-programming and optimization
Population, community and ecosystem modeling	Large-scale computing is making it feasible to model ecosystems by aggregating models of individuals. With petascale computing capabilities, this technique can begin to be applied to natural environments such as soils and to artificial environments such as bioreactors, in order to understand the interactions between different types of organisms and their ability to cooperatively metabolize compounds important for carbon cycling.	Numerical solution of partial differential equations, ordinary differential equations, and simple ordinary differential equations

Table 2. Projections for future research and its objectives

Field / Area of Study	Objectives
Determining the detailed evolutionary history of each protein family	*This will enable rational planning for structural biology initiatives and will provide a foundation for assessing protein function and diversity.*
Determining the frequency and detailed nature of horizontal gene transfers in prokaryotes	*This will shed light on the molecular and genetic mechanisms of evolution by means other than direct "Darwinian" descent and will contribute to our understanding of the acquisition of virulence and drug resistance in pathogens and the means by which prokaryotes adapt to the environment*
Automated construction of core metabolic models for all the sequenced Department of Energy (DOE) genomes	*This will enable dramatic acceleration of the promise of the Genomes to Life (GTL) program and the use of microbial systems to address DOE mission needs in energy, environment, and science.*
Predict essential genes for all known sequenced micro-organisms	*This will enable a broader class of genes and gene products to be targeted for potential drugs and to predict culturability conditions for environmental microbes.*
Computational screening all known microbial drug targets against the public and private databases of chemical compounds to identify potential new inhibitors and potential drugs	*The resulting database would be a major national biological research resource that would have a dramatic impact on worldwide health research and fundamental science of microbiology*
Model and simulate the precise cellulose degradation and ethanol and butanol biosynthesis pathways at the protein/ligand level to identify opportunities for molecular optimization	*This would result in a set of model systems to be further developed for optimization of the production of biofuels*
Model and simulate the replication of DNA to understand the origin of and the repair mechanisms of genetic mutations	*This would result in dramatic progress in the fundamental understanding of how nature manages mutations and understanding which molecular factors determine the broad range of organism susceptibility to radiation and other mutagens*

Model and simulate the process of DNA transcription and protein translation and assembly	*This would enable us to move forward on understanding post-transcription and post-translation modification and epi-genetic regulation of protein synthesis.*
Model and simulate the interlinked metabolisms of microbial communities	*This project is relevant to understanding the biogeochemical cycles of extreme, natural and disturbed environments and will lead to the development of strategies for the production of bio-fuels and the development of new bio-engineered processes based on exploiting communities rather than individual organisms.*
In silico	*Prediction of mutations and activity, conformational changes, active site alterations and functional and structural analyses of proteins of importance in healthcare or industrial biotechnology.*
Agricultural Biosystem	*Precision farmer's, modelling for disease, forecasting, fertilizer and application such as climax change using biosensors.*
Bioprocessing*	*Data handling of biomolecules to be used in computer modelling and simulation to identify pathways involved in targeted disease leading to new therapeutic targets or pathways.*
Agriculture genomic and gene discovery*	*Data mining on genome sequence for new genes related to traits listed and also identification and functional analysis of important genes related to yield, plant development, plant resistance to pest and disease, abiotic stress resistance.*
Metabolite engineering*	*Modelling on certain substances which will optimize genetic and regulatory processes within cells to increase the cells production.*
Development of computational algorithms for bioinformatics system*	*Development of algorithms and statistics that enables efficient storage and access of data analyses, data management or data integration.*

* ref. [18]