Advanced Modeling and Optimization, Volume 10, Number 1, 2008

Evolution to Revolution

A review on Bioinformatics

Dipankar Mazumdar and Swapan Raha

Department of Mathematics
Visva-Bharati (A Central University)
Santiniketan 731235, INDIA

dmazumda@science.uva.nl, raha_swapan@rediffmail.com

***Abstract**: This review is an introduction to the field of bioinformatics. Bioinformatics is a multidisciplinary subject devoted to the in-silico analysis of biological information. It requires the help of other natural sciences. In this review we study the developments of bioinformatics. A new definition to the subject is given. The centres of the subject are described. The aims and the future directions are studied.*

Received: January 2007; Revised and Accepted: January 2008.

AMO- ADVANCED MODELING AND OPTIMIZATION ISSN: 1841-4311

Pages 51–62

Introduction

In the previous century it was thought that biology and mathematics can't go together. Biology and mathematics are subjects with different domains. It is true. But the fact is that mathematics does the general formulation which can also be implemented in the biological problems. The latest developments in molecular biology lead mathematics to solve the newly created problems. After the completion of the Human Genome Project, the application of other subjects in molecular biology has taken a start. That new field is named as computational molecular biology or bioinformatics.

The term, Bioinformatics, was coined by Hwa Lim (at The Supercomputer Research Institute at the Florida State University, USA) in the late 1980's, to encompass all forms of computational study and analysis of biological problems [1]. The Central dogma, DNA makes RNA makes Protein, has long been a staple of biology text books. More recently, this paradigm has been extended from individual genes to whole genomes by advances in genomic technologies [2]. In the 21st century Bioinformatics has transformed the discipline of Biology from a purely lab-based science to an information science as well [3]. Many had the vision of establishing bioinformatics in a leadership role over experimental biology, similar to that theoretical physics enjoys over experimental physics. Somewhere along the line, it seems that bioinformatics lost this ambition and became sidetracked onto what physicists would call a 'phenomenological' pathway [4].

Bioinformatics and computational molecular biology have emerged as new branches of science. These are the results of the inevitable marriage of mathematics, information science and biosciences (especially molecular biology). Bioinformatics and Computational Biology are rooted into life sciences as well as computer and information sciences and technologies. Both of these interdisciplinary approaches draw from specific disciplines such as mathematics, physics, computer science, engineering, biology and behavioral science. Bioinformatics applies principles of information sciences and technologies to make the vast, diverse and complex life science data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology.

Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at there interface [5]. Bioinformatics is the technological aspect and Computational (molecular) biology is the scientific aspect [6]. The relationship between computer science and biology is a natural one for several reasons. First, the phenomenal rate of biological data being produced provides challenges: massive amount of data have to be stored, analyzed, and made accessible. Second, the nature of the data is often such that a statistical method, and hence computation, is necessary. This applies in particular to the information on the building plans of proteins and of the temporal and spatial organization of their expression in the cell encoded by the DNA. Third, there is a strong analogy between the DNA sequence and a computer program [7].

Bioinformatics, the development and application of computational tools and approaches for expanding the use of biological data including those to acquire, store, organize, archive, analyze or visualize; has now firmly established itself as an independent discipline within the domain of science [5]. This interdisciplinary approach draws from specific disciplines such as mathematics, physics, computer science, biology, behavioral science and information science. It has applications in various fields such as molecular medicine, personalized medicine, gene therapy, evolutionary studies and drug discovery [8]. Given a nucleotide sequence, the probable amino acid sequence of the encoded protein can be determined using some translation software. Sequence search technique is used to find homologues in model organism. Based on sequence similarity, we can predict the structure of the human protein. Docking algorithm is then used to design molecules that can bind the model structure, leading the way for biochemical assays to test their biological activity on the actual protein. In the process, mathematical concepts like similarity measures (in sequence alignment), differential geometry (in protein structure prediction), design and analysis of efficient algorithm are being extensibly used [9].

Bioinformatics is relatively inexpensive, research and development can be done with modest equipment and plenty of public resources. It can generate predictions that should be tested in the lab. This is a novel and exciting technology platform for the future [10]. The rules of theoretical physics have a predictive power; this part of theoretical endeavor is entirely missing in current bioinformatics. As a consequence, we are still not able to agree on the number of human genes despite having the complete sequence of the human genome at hand [4].

Defining the terms bioinformatics and computational biology is not necessarily an easy task. In the past few years, as the areas have grown, a greater confusion into these two terms has prevailed. For some, the terms bioinformatics and computational biology have become completely interchangeable terms, while for others, there is a great distinction. Computational biology and bioinformatics are multidisciplinary fields, involving researchers from different areas of specialty, including (but in no means limited to) statistics, computer science, physics, biochemistry, genetics, molecular biology and mathematics [11].

What is Bioinformatics?
The latest definition [12]

Bioinformatics is the application of basic and applied sciences for the development of new technologies (computer) to the management of biological information.

The term *biological information* in the given definition means the information encoded in the nucleotide sequence or amino acid sequence. The term *management* in the given definition means to acquire, store, organize, archive, analyze and visualize.
In other words this can be defined as Bioinformatics is the multidisciplinary approach comprising mathematics, biology, computer science, pharmacy, medicine, chemistry,

statistics, biochemistry, linguistics, informatics, physics and health science. It develops mathematical and computational tools for the analysis and organization of biological (molecular biology, medicine, health etc.) data.

Other important definitions

Today, practical applications of bioinformatics are readily available through the World Wide Web and are widely used in biological and medical research. As the field is rapidly evolving, the very definition of bioinformatics is still the matter of some debate [7]. To get the clear picture of the definitions of bioinformatics some are given below.

Definition: as described in ([9, 13]).
(Molecular) **bio**- informatics: bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical chemistry) and applying "**informatics techniques**" (derived from disciplines such as applied mathematics, computer science and statistics) to **understand** and **organize** the **information** associated with these molecules, on a **large scale**. In short, bioinformatics is a management information system for molecular biology and has many **practical applications**.

Definition: NIH [5] Bioinformatics is the research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data including those to acquire, store, organize, archive, analyze or visualize such data.

Definition: NCBI [3] Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. At the beginning of the 'genomic revolution', a bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues but the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data.

Definition: Stanford University [14] The definition of bioinformatics is not universally agreed upon. Generally speaking, we define it as the creation and development of advanced information and computational technologies for problems in biology, most commonly molecular biology (but increasingly in other areas of biology). As such, it deals with methods for storing, retrieving and analyzing biological data, such as nucleic acid (DNA/RNA) and protein sequences, structures, functions, pathways and genetic interactions. Some people construe bioinformatics more narrowly, and include only those issues dealing with the management of genome project sequencing data. Others construe bioinformatics more broadly and include all areas of computational biology, including population modeling and numerical simulations.

Definition: UCLA [15] Bioinformatics is the study of the inherent structure of biological information and biological systems. It brings together the avalanche of systematic biological data (e.g. genomes) with the analytic theory and practical tools of computer science and mathematics.

This formulation of the 'bioinformatics problem' opens the way for biologists to collaborate with mathematicians and computer scientists, because it aims to translate 'biology problems' into new challenges that are interesting to theoreticians–problems of information content, structure and encoding, which inherently interest theorists.

Definition: Pasteur Institute [16] Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information stored in the genetic code, but also experimental results from various sources, patient statistics, and scientific literature. Research in bioinformatics includes method development for storage, retrieval, and analysis of the data. Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine.

Definition: Georgia Tech [17] Bioinformatics is an integration of mathematical, statistical and computer methods to analyze biological, biochemical and biophysical data.

Definition: Weizmann Institute [6] The body of tools, algorithms and know-how needed to handle complex biological information the technological aspect.

Aims and scope

Informatics is the creation, development, and operation of databases and other computing tools to collect, organize, and interpret data [23]. Two Principal approaches underpin all studies in bioinformatics. First is that of comparing and grouping the data according to biologically meaningful similarities and second, that of analyzing one type of data to infer and understand the observations for another type of data [9]. Three key areas of this field are the organization of knowledge in databases, sequence analysis and structural bioinformatics [7]. The ultimate goal of the field is to enable the discovery of new Biological insights as well as to create a global perspective from which unifying principles in biology can be discerned [23].

The rationale for applying computational approaches to facilitate the understanding of various biological processes includes: a more global perspective in experimental design and the ability to capitalize on the emerging technology of database-mining - the process by which testable hypotheses are generated regarding the function or structure of a gene or protein of interest by identifying similar sequences in better characterized organisms [3].

The core data comprises of the genomes and proteomes of human and other organisms, 3-D structures and functions of proteins, microarray data, metabolic pathways, cell lines and hybridoma, biodiversity etc. The sudden growth in the quantitative data in biology has rendered data capture, data warehousing and data mining as major issues for biotechnologists and biologists. Availability of genomic and other data has resulted in the realization of the inherent biocomplexity issues which call for innovative tools for synthesis of knowledge [8].

Aims

To improve the content and utility of databases, develop better tools for data generation, capture, and annotation, develop and improve tools and databases for comprehensive functional studies, develop and improve tools for representing and analyzing sequence similarity and variation, create mechanisms to support effective approaches for producing robust, exportable software that can be widely shared. Organization of data in such a way, that it allows researchers to access existing information and to submit new entries as they are produced. While data-creation is an essential task, the information stored in these databases is useless unless analyzed. Thus the purpose of bioinformatics extends well beyond mere volume control.

To develop tools and resources that help in the analysis of data. For example, having sequenced a particular protein, it is with previously characterized sequences. This requires more than just a straightforward database search. As such, programs such as FASTA and PSI-BLAST much consider what constitutes a biologically significant resemblance. Development of such resources needs extensive knowledge of computational theory, as well as a thorough understanding of biology. Use of these tools to analyze the individual systems in detail and frequently compared them with few that are related.

The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned [23].

In 1909, Danish botanist Wilhelm Johanssen coined the word gene for the hereditary unit found on a chromosome. Nearly 50 years earlier, Gregor Mendel had characterized hereditary units as factors- observable differences that were passed from parent to offspring.

The completion of a *working draft* of the human genome, an important milestone in the Human Genome Project, was announced in June 2000 at a press conference at the White House and was published in the February 15, 2001 issue of the journal Nature [3].

Biological Sciences modeled by Mathematical Sciences		
Genetics	Statistics	*Mendel*, 1866
Population Genetics	Differential and Difference Equations	*Fisher,Wright, Sewall*, 1920s
Epidemiology	Differential Equations and Statistics	1950s
Neurology – Networks	Partial Differential Equations	*Mc. Culloch , Pitts*, 1943 and *Hogdkin – Huxley*, 1952
DNA segments and cloning	Graph Theory	*Benzer*, 1959

The future

Bioinformatic tools are software programs that are designed for extracting the meaningful information from the mass of data and to carry out this analysis step. Factors that must be taken into consideration when designing these tools are (i) the end user (the biologist) may not be a frequent user of computer technology, (ii) these software tools must be made available over the internet given the global distribution of the scientific research community. Bioinformatics techniques are used for whole genome and sequence analysis and experimental analysis involving thousands of genes simultaneously, DNA chips and array analysis -expression arrays, comparative analysis between species and strains, analysis of the proteome of an organism, medical applications for genetic disease-pharmaceutical and biotechnology industry, forensic applications, agricultural applications.

The potential of bioinformatics in the identification of useful genes leading to the development of new gene products, drug discovery and drug development has led to a paradigm shift in biology and biotechnology. These fields are becoming more computationally intensive. The new paradigm, now emerging, is that all the genes will be known and in the sense of being resident in database available electronically and the starting point of biological investigation will be theoretical and a scientist will begin with a theoretical conjecture and only then turning to experiment to follow or test the hypothesis. The shift from gene biology has resulted in the development of strategies-from lab techniques to computer programs to analyze whole batch of genes at once. Genomics is revolutionizing drug development, gene therapy, and our entire approach to health care and human medicine.

The genomic discoveries are getting translated in to practical biomedical results through bioinformatic applications. Work on proteomics and genomics will continue using highly sophisticated software tools and data networks that can carry multimedia databases. Thus, the research will be in the development of multimedia databases in various areas of life sciences and biotechnology. There will be an urgent need for development of software tools for data mining, analysis and modeling, and downstream processing. Security of data, data transfer and data compression, auto checks on data accuracy and correctness will also be major research areas of bioinformatics. The use of virtual reality in drug design, metabolic pathway design, and unicellular organism design, paving the way to design and modification of multi-cellular organisms, will be the challenges which bioinformatics scientist and specialist have to tackle.

Information technology, particularly the internet, is utilized to collect, distribute and access ever-increasing data which are later analyzed with mathematics and statistics based tools. Bioinformatics has a key role to play in the cutting edge research and development areas such as functional genomics, proteomics, protein engineering, pharmacogenomics, discovery of new drugs and vaccines, molecular diagnostic kits, agro-biotechnology etc [8].

Conclusion

The key to bioinformatics is integration, integration; integration- says Jim Golden [20]. Bioinformatics is used largely in the field of human genome research, by the Human Genome Project that has been determining the sequence of the entire human genome (about 3 billion base pairs). It is essential in using genomic information to understand diseases. It is also used largely for the identification of new molecular targets for drug discovery [21, 22]. Bioinformatics is emerging as the most powerful technique in the coming age. Mathematics will be used more and more in solving the biological problems. There is equal scope for mathematicians as well as for biologists. There is a vast area to study. Many fields have already been studies but many are untouched, too.

Bioinformatics integrates the advances in different subject areas to solve complex problems in Life Sciences [8]. Two principal approaches underpin all studies in bioinformatics. First is that of comparing and grouping the data according to biologically meaningful similarities and second, that of analyzing one type of data to infer and understand the observations for another type of data [9]. Three key areas are the organization of knowledge in databases, sequence analysis and structural bioinformatics [7]. The paradigm has been shifted. Now a biologist can forget test tubes, petri dishes and pipettes. One of the few pieces of equipment that can be honestly labeled ubiquitous in biology today is the computer. Many unsolved problems of biology may now be solved applying other sciences and computer technologies. That is why Haussler says Bioinformatics is an enormous frontier [20].

Acknowledgement

The research has been partially supported by the UGC-SAP DRS Phase-I project under the Department of Mathematics, Visva-Bharati.

References

[1] http://www.thermo.com/com/cda/resources/resources detail/1,,37,00.html

[2] Henikoff, S, (2002), *Editorial- Beyond the Central Dogma*, Bioinformatics, 18 (2): 223-225.

[3] http://www.ncbi.nlm.nih.gov/

[4] Claverie, J.M., *From bioinformatics to Computational Biology*, Genome Research, 10:1277-1279.

[5] NIH Working definition of Bioinformatics and Computational Biology, July 17, 2000.

[6] http://bioinfo.weizmann.ac.il/courses/BCG

[7] Nilges, M., Linge, J.P., Unit de Bio-Informatique Structurale, Institut Pasteur, France.

[8] www.geocities.com/bioinformaticsweb

[9] Luscombe, N.M., Greenbaum, D., Gerstein, M., (2001), *What is bioinformatics? An introduction and overview*, Yearbook of Medical Informatics: 83-99.

[10] Ouzounis, C., (2000), *Two or three myths about bioinformatics*, Bioinformatics, 16 (3): 187-189.

[11] http://www.biochem.northwestern.edu

[12] Mazumdar, D., (2006)*, A sequence alignment model for bioinformatics drug design*, Abstract Proceedings of the 13th West Bengal State Science and Technology Congress, 2006: 35.

[13] Reichhardt, T., (1999), *It's sink or swim as a tidal wave of data approaches*, Nature, 399 (6736): 517-520

[14] http://smi-web.stanford.edu/people/altman/bioinformatics.html#one

[15] http://www.doe-mbi.ucla.edu/˜leec/UCLA-bioinf/bioinf-RFC.html

[16] http://www.pasteur.fr/recherche/unites/Binfs/definition/

[17] http://www.biology.gatech.edu/bioinformatics/whatis.html)

[18] http://www.cite-sciences.fr/english/ala cite/expo/tempo/defis/home/home.html

[19] http://cancerweb.ncl.ac.uk/cgibin/

[20] Chicurel, M., (2002), Bioinformatics: Bringing It All Together, Nature, 419: 751-757.

[21] http://80211-planet.webopedia.com/TERM/B/bioinformatics.html

[22] http://www.internet.com

[23] http://www.DOEGenomes.org

[24] Burge, C., (2002), Genome Technology, 17.

[25] Grenberg, H.S. and Briggs, W.L., (2004), *Mathematical Models in Molecular Biology*, University of Colorado at Denver, Lecture Note.

[26] L. Allison, http://www.csse.monash.edu.au/ lloyd/

[27] Pevzner, P.A., (1998), *Computational molecular biology- An Algorithmic approach*, The MIT Press.

[28] http://arbl.cvmbs.colostate.edu/molkit/dnadot/bkg.html

[29] Kanehisa, M., Bork, P., (2003), *Bioinformatics in the post sequence era*, Nature Genetics, 33: 305-310.

[30] http://www.icp.ucl.ac.be/bio links.html

[31] Viklund, H., (2003), *Development of a modular HMM package for biological sequence analysis*, Thesis.

[32] Mazumdar, D., (2005), *A definition of Bioinformatics*. http://www.geocities.com/itsmedip
[33] http://www.biochem.northwestern.edu
[34] Russell, P.J., iGenetics, Benjamin Cummings, USA, 2002.
[35] Brown, T.A., Genomes, John Wiley & Sons (Asia) Pte Ltd.
[36] Gwynne, P., (1999), Science and Technology, Science, Special advertising supplement.
[37] Levenshtein, V.I., (1965), *Binary codes capable of correcting deletions, insertions and reversals*, Doklady Akademii Nauk SSSR, 163(4): 845-848.
[38] Bowen, R. A., (2004), Molecular Toolkit, Lecture Note.
[39] Mount, D.W., (1999), http://www.blc.arizona.edu/courses/bioinformatics/mult-seqalign.html
[40] http://www.blc.arizona.edu/courses/bioinformatics/distance.html
[41] http://ibc.wustl.edu/msa.html
[42] http://www.blc.arizona.edu/courses/bioinformatics/phylogenetic.html#anchor424246
[43] Baldi, P., Chauvin,Y., Hunkapillar, T., and McClure, M.A., (1994), *Hidden markov models of biological primary sequence information*, Proc. Nat. Acad. Sci. (USA) ,91: 1059-1063.
[44] Richon, A.B., (2005), *A Short History of Bioinformatics*, Network Science, http://www.netsci.org/Science/Bioinform/feature06.html

Appendix

Landmark events upto HGP

1933-Electrophoresis is introduced by Tiselius for separating proteins in solution
1951-Proposal of structure for the alpha-helix and beta-sheet
1953-Watson and Crick propose the double helix model for DNA based on x-ray data obtained by Franklin and Wilkins
1955-The sequence of the first protein to be analyzed, bovine insulin, is announced by F. Sanger
1964-tRNA 75 bases (old, slow, complicated method)
1970-The details of the Needleman-Wunsch algorithm for sequence comparison are published
1972-The first recombinant DNA molecule is created by Paul Berg and his group
1973-The Brookhaven Protein Data Bank is announced
1977-Sanger, Maxam and Gilbert develop DNA sequencing methods
1977-First complete DNA genome: X174 DNA - 5386 bases
1980-Detailing the use of multi-dimensional NMR for protein structure determination
1981-human mitochondrial DNA - 16,569 bases
1981-The Smith-Waterman algorithm for sequence alignment
1982-GenBank database is established
1983-First human disease gene (Huntington disease) is mapped
1984-First public discussions of the sequencing of the human genome
1985-PCR is introduced as a method
1986-First introduction of automated DNA sequencer from Applied Biosystems
1986-Tobacco chloroplast DNA - 155,844 bases
1986-The term "Genomics" appeared for the first time to describe the scientific discipline of mapping, sequencing, and analyzing genes. The term was coined by Thomas Roderick as a name for the new journal
1986-The SWISS-PROT database is created by the Department of Medical Biochemistry of the University of Geneva and the European Molecular Biology Laboratory (EMBL)
1987-The physical map of e. coli is published (Y. Kohara, et. al., Cell 51: 319-337)
1988-The Human Genome Initiative is started
1988-The FASTA algorithm for sequence comparison is published by Pearson and Lupman
1990-The Human Genome Project was officially launched at USA (The first gene for breast cancer [BRCAI] was also mapped)
1990-The BLAST program (Altschul, et. al.) is implemented
1991-First paper describing ESTs is published
1991-The creation and use of expressed sequence tags (ESTs) is described (J. Craig Venter, et. al., Science, 252: 1651-1656)
1992-First papers describing sequencing of 100,000 kb regions of the human genome
1992-BAC cloning vector is introduced

1993-First five year plan for the HGP in the USA is published
1994-Smallpox genome sequence published (186,000 bp)
1994-HGP's human genetic mapping goals are achieved
1994-Development of the first assembly algorithm for handling tens of thousands of DNA sequences
1995-First bacterial genome sequence (*Haemophilus influenzae*) is published.
1996-First archaeal genome sequence (*Methanococcus janaschii*) is published
1996-Completed Yeast genome
1996-*Saccharomyces cerevisiae* genome sequence is published
1996-Bermuda principles for rapid and open data release are established
1996-The Prosite database is reported by Bairoch et.al.
1996-Affymetrix produces the first commercial DNA chips
1997-*Escherichia coli* (4.7 Mbp) genome sequence is published
Between 1997 to 1998 several pathogenic bacterial genomes sequenced 1.Helicobacter pyloris (ulcers) 2.Treponema pallidium (Syphilis) 3.Borrelia burgdorferi (Lyme disease) 4.Chlamydia trachomatis (trachoma - blindness) 5.Rickettsia prowazekii (epidemic typhus) 6.Mycobacterium tuberculosis (tuberculosis)
1998-*Caenorhabditis elegans* genome sequence is published
1999-Full scale human genome sequencing begins
1999-Sequencing of the first human chromosome (chr.22) completed
2000-Draft versions of the human genome sequence are completed
2000-The genome for *Pseudomonas aeruginosa* (6.3 Mbp) is published
2000-*Drosophila melanogaster* genome sequence is published
2000-*Arabidopsis thaliana* genome sequence is published
2001-Draft versions of the human genome sequence is published
2002-Draft versions of the mouse and rice genomes sequence is published