한국산학기술학회논문지
Vol. 11, No. 3, pp. 971-977, 2010

Bioinformatics : Latest Application and Interdisciplinary Field of Computer Science

Ki-Bong Kim[1*]
[1]**Department of Biomedical Technology, Sangmyung University**

전산학의 최신 응용 및 학제 분야인 생명정보학

김기봉[1*]
[1]상명대학교 공과대학 의생명공학과

요 약 A flood of biological data has caused many challenges in computing. Bioinformatics, the application of computational techniques to analyze the information associated with biomolecules on a large-scale, has now firmly established itself as an interdisciplinary subject in molecular biology, and encompasses a wide range of subject areas from structural biology, genomics, proteomics, systems biology, biostatistics to computer science. In this review, I provide an introduction and overview of the current state of bioinformatics. Looking at the types of biological information and databases that are commonly used, I also deals with some of bioinformatics application domains which are closely related to areas of computer science.

Abstract 본 바이오데이터들이 홍수처럼 쏟아지면서 컴퓨터를 활용해 해결해야 할 많은 문제들이 야기되고 있다. 생체분자들과 연관된 정보들을 대량으로 분석하는 전산기술 응용분야인 생명정보학은 학제간 학문분야로서 현재 확고히 자리매김하고 있으며. 구조생물학, 유전체학, 단백질체학, 시스템 생물학, 생물통계학 및 전산학 등 광범위한 학문영역을 포괄한다. 본 총설에서는 생명정보학에 대한 최근 상황과 전반적인 응용분야에 대해 소개하고자 한다. 일반적으로 널리 사용되는 생명정보 및 바이오데이터베이스들의 유형을 살펴보고, 전산학 분야와 긴밀한 관계가 있는 몇 가지 생명정보학 응용 분야들을 다루고자 한다.

Key Words : Bioinformatics, Structural Biology, Genomics, Proteomics, Systems Biology

1. Introduction

Biological data has been increasing at an exponential rate. For example as of December 2009, the GenBank repository of nucleic acid sequences contained 112,910,950 entries [1] and the PDB database of biological macromolecules structures contained 62,926 entries [2]. On average, these databases are doubling in size every 15 months. In addition, since the publication of the H. influenza genome [3], complete sequences for over 1,000 organisms have been released [4]. As a result of this surge in biological data, computers have become indispensable to biological research. Such an approach is desirable because of the ease with which computers can handle large quantities of data and probe the complex dynamics observed in nature.

Bioinformatics, the subject of the current review, is often defined as the application of computational techniques to understand and organize the information associated with biological macromolecules [5]. Now it cannot be denied that bioinformatics is an emerging interdisciplinary field bringing together researchers from different academic backgrounds, most prominently molecular biology and computer science. According to the NCBI (National Center for Biotechnology Information) in USA, the main goal of bioinformatics is to “enable the

*Corresponding Author : Ki-Bong Kim(kbkim@smu.ac.kr)
Received October 26, 2009 Revised (1st January 12, 2010, 2nd March 15, 2010) Accepted March 18, 2010

discovery of new biological insights", accomplished with the use of various computational tools. Molecular biologists are generally interested in the "functional aspect" of bioinformatics, such as analyzing data, discovering patterns, and developing methods of prediction for functions of organisms [6]. The computer scientists involved in this field mainly develop algorithms and information systems for storage, retrieval and analysis of biological data [7].

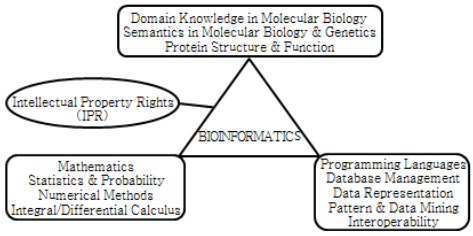
Newer technologies, such as the Internet, are quickly changing the way scientists share their data, results, ideas and research in general [6]. In the bioinformatics field, the collection of high-throughput experimental data in a central 'warehouse' where it can be shared is saving researchers many years of lab benchwork and helping them narrow down research focuses [6].

As to knowledge in biology, I hope the reader will be able to recall some of the rudiments of biology. The reader should keep in mind the pervasive role of evolution in biology. It is evolution that allows us to infer the cell behavior of one species, say the human, from existing information about the cell behavior of other species like the mouse, the worm, the fruit fly, and even yeast. In the past decades, biologists have gathered information about the cell characteristics of many species. With the help of evolutionary principles, that information can be extrapolated to other species. However, most available data is fragmented, incomplete, and noisy. So if one had to characterize bioinformatics in logical terms, it would be reasoning with incomplete information. That includes providing ancillary tools allowing researchers to compare carefully the relationship between new data and data that had been validated by experiments.

In this review I provide an introduction and overview of the current state of bioinformatics. In addition, I also discuss the main principles that underpin bioinformatics analyses, look at the types of biological information and databases that are commonly used, and finally examine some of bioinformatics application domains that are areas of computer science.

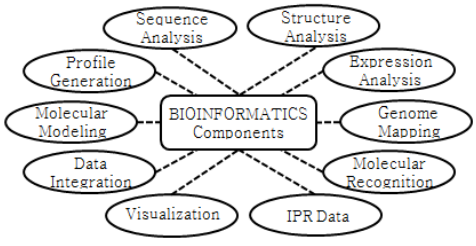
2. Skills for Bioinformatics and Its Components

The skills required to do bioinformatics experiments are: (a) domain knowledge in molecular biology; (b) computational and programming skills; and (c) mathematics (Fig. 1) [8]. Domain knowledge in molecular biology, genetics and protein structures is highly appreciated. Computational and programming skills include UNIX or LINUX, PERL, Phython, HTML, database management, data representation and storage, patterns and data mining and biological data interoperability. Mathematical skills required are statistics and probability, numerical methods and integral and differential calculus. Furthermore, knowledge on IPR (intellectual property rights) is welcomed.



[Fig. 1] Skills for bioinformatics

The components of bioinformatics include sequence analysis, profile generation, structure analysis, molecular modeling, expression analysis, genome mapping, molecular recognition (docking), IPR data, graphics in biology, data integration and data management (Fig. 2) [8]. Each of these components is a subject by themselves. It is much appreciated to refer standard books in bioinformatics for further information on these aspects.



[Fig. 2] Bioinformatics components

3. Bioinformatics Data

Bioinformatics data consists of different views of biological information. Bioinformatics databases are diverse in their data formats, and are highly redundant. The bioinformatics data views include biological sequences (DNA, RNA, and proteins), gene or protein expression, functional properties, molecular interactions, clinical data, system descriptions, and related publications. The data appears as sequences, sequence annotations, structural models, physical maps, clinical records, interaction pathways, gene and protein expressions, protein-protein interactions, and other forms in data sources such as databases, private data collections, and related publications.

There is substantial diversity and variation in bioinformatics data, even among databases containing data of the same type. Each database has its own infrastructure and proprietary data format. Common data standards and data exchange formats are not established in this field. For example, sequence entries are described in different formats in GenBank [9], Swiss-Prot [10], and EMBL [11]. GenBank developed the ASN.1 (Abstract Syntax Notation One) format while Swiss-Prot designed its own format. The Swiss-Prot data format differs slightly from that of the EMBL database. Recent introduction of XML (eXtensible Markup Language) as the generic data exchange format has also given rise to several variants of the XML representations of bioinformatics data. Despite a variety of available data formats, a universal protocol for data exchange has not been established, contributing to the complication of bioinformatics data management.

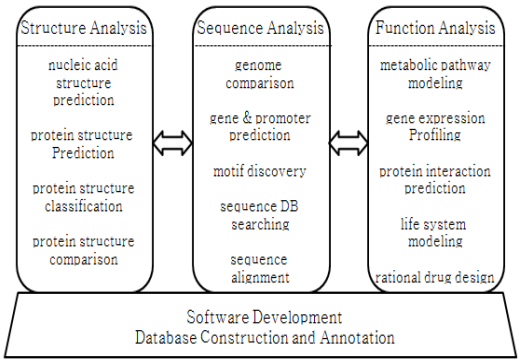
[Table 1] Major institutions worldwide for storing genetic and biological data

Database	Institution	Description	Country
GenBank	NCBI	National Center for Biotechnology Information	USA
EMBL	EBI	European Bioinformatics Institute	Europe
DDBJ	CIB	Center for Information Biology	Japan

In recent years, biological data has increased to multiple folds. This is exemplified by the amount of data available at public funded institutions (NCBI, EBI and CIB). GenBank, EMBL and DDBJ [12] are the main databases for genetic data available in the internet for public use. They are hosted by USA, Europe and Japan (Table 1). These databases exchange data on a daily basis despite being hosted at different geographical locations.

4. Major Applications of Bioinformatics

Collecting, analyzing and processing the diverse set of biological data call for an even more diverse group of applications and algorithms. Some of these are more prominent than others: Genomics, the study of the entire genome of a given organism, is probably the most familiar bioinformatics application domain for many. Genomics itself covers a variety of different topics [13]. While functional genomics involves the identification of specific functions of a gene in the genome, comparative genomics deals with the analysis of similarities and differences in the genomes of different species. Genomics applications are probably the most commonly used bioinformatics applications, because genomic data constitutes the majority of biological data in existence. A close second is proteomics, which studies the entire set of proteins of an organism. While proteomics and genomics applications account for a large number of well-known bioinformatics application in use, there are many other lesser-known application domains in bioinformatics. The applications fall into three areas: sequence analysis, structure analysis, and function analysis (Fig. 3). Biocomputing tool development is at the foundation of all bioinformatics analysis domains. There are intrinsic connections between different areas of analyses [14]. Bioinformatics is still relatively new emerging field with a very diverse variety of problems and subfields; and providing a comprehensive listing of all bioinformatics applications is outside the scope of this article. In this section, I identify and briefly describe some of the more important bioinformatics application domains.



[Fig. 3] Overview of various subfields of bioinformatics

4.1 Sequence Alignment

Modern genomics relies on the paradigm of evolution which suggests that all organisms are related to each other as a result of the evolutionary process. This implies that similar proteins should exist in related species, and hence the sequences that code these proteins should also be similar. This similarity is called homology in molecular biology. The most important practical use of homology is in the determination and confirmation of coding regions in DNA sequences. Comparing a sequence region with regions known to encode certain proteins in other organisms can be useful for confirming whether the region might be instrumental in coding a similar protein, or not. A high degree of homology between two sequences from different organisms might also provide clues on the evolutionary links between the organisms. Sequence alignment algorithms are used to align two sequences in a way to maximize the similarity between them. In many cases, commonly used pattern matching techniques are not readily applicable to biological sequences due to the fuzzy nature of the data. To illustrate this point, I consider an alignment of the DNA sequence "AGTAGC" with "GAG". An ordinary string matching algorithm will not be able to find "GAG" within the larger string; but in case of biological sequence data we can use a gap to accommodate the non-matching character "T":

AGTAGC
-G-AG-

The number of matching residues (or symbols) contributes positively to the similarity score and gap/non-matching residues contribute negatively. Sequence alignment algorithms aim to maximize the similarity score by finding the best alignment possible, which generally means that gaps should be avoided wherever possible. In the specific case of amino acid sequences, the contribution of each residue match or the penality associated with gaps or non-matching residues is generally read from scoring matrices that were computed with evolutionary relationships of different amino acids taken into account. These matrices are also called substitution matrices, and the most important substitution matrices in use are BLOSUM (Blocks Substitution Matrix) and PAM (Percent Accepted Matation) [13]. These matrices are provided in several different varieties based on the degree of identity between the sequence to be aligned. The most widely used heuristic-based sequence alignment algorithms are BLAST [15] and FASTA [16]; both of which can align DNA, RNA or protein sequences.

4.2 Multiple Sequence Alignment

Another widely used and effective technique is multiple alignment, which helps align several sequences of symbols, so identical symbols are properly lined up vertically, with gaps allowed within symbols. The sequences may represent variants of the same proteins in various species; the goal is to find conserved parts of the proteins that are unchanged during evolution. Finding conserved parts of proteins also provides hints about a possible function of proteins. Methods for multiple alignments are based on dynamic programming techniques developed for pairwise alignment.

After aligning multiple genomic or protein sequences, biologists usually depict trees representing the degree of similarity among the sequences being studied. Depicting evolutionary trees is in itself a domain within bioinformatics called phylogenetic trees. The problem of matching spatial structures can be viewed as a combination of computational geometry and computer graphics.

Probably the most widely used multiple sequence alignment tool is CLUSTAL version [17], which uses a progressive algorithm. This application generates the phylogenetic tree by calculating a distance matrix for every pair of sequences in the input set, and completes progressive alignment by using this tree as a guide.

4.3 Protein Structure Prediction

The capability to deduce the three-dimensional physical structure of a protein from its amino acid sequence is highly valued because of its potential applications in the drug discovery process. Since the biochemical function of a protein depends on its physical shape, pharmacology researchers need to know the precise shape of the protein to be able to design a compound that can attach to it. Currently, the only way to determine the structural layouts of proteins is through experimental methods such as NMR and X-ray crystallography. Among all of the diverse application domains in bioinformatics, protein structure prediction probably has one of the highest requirements for computing performance. In theory, inferring the tertiary structure of a protein from its amino acid sequence data should be possible through the use of molecular dynamics formulations, because the structure of the protein is dictated by the attraction and repulsion forces between the structural elements forming it. In practice, a solution by molecular modeling seems elusive, primarily because of the overwhelming computational cost and inaccuracies introduced by the experimental processes needed to determine the parameters needed for the computation. As a result, current research in this field has focused on statistical and empirical methods which exploit the existing repository of protein structure data.

Since the emerging field of protein structure prediction is still very active and many important questions remain unanswered, it is somewhat difficult to name applications in wide use. An example application THREADER [18] which evaluates the compatibility of known protein folds and amino acid sequences. Rost et al. [19] describe the important issues in protein structure prediction while providing a good introduction to this application domain.

4.4 Systems Biology

While the insights provided by genomics, proteomics and functional genomics have allowed scientists to understand the origins of life better and formulate drugs for diseases, they do not afford us a complete understanding of the complex events that take place in even the simplest organisms. An emerging paradigm called systems biology aims to combine the numerous computational biology techniques and data in order to model entire biological systems such as tissues, organs, even complete life forms. The ambitious goal of being able to model extremely complex biological systems, if realized, could allow scientists with a very useful tool to observe important biochemical pathways in silico and even test the effects of new medications using computer simulations.

Since systems biology is a very recent concept, the kind of applications and data formats it will require are not fully determined yet. The breadth of the effort implies that algorithms, workloads and data formats from all domains of bioinformatics are likely to be utilized. A good and concise introduction to the concepts and challenges of systems biology is presented by Morel et al. in [20]. Finkelstein et al. [21] present a useful discussion of the computational requirements of the emerging field of systems biology, focusing on the modeling challenge.

Needless to say, computational requirements of systems biology applications will be enormous and the field is already being recognized as a ”grand challenge” in high performance computing. Freddolino et al. have published the results of the first complete simulation of an entire life form [22]. They used the NAMD molecular dynamics framework [23] to simulate the modeled virus for 13ns. of simulated time. Even though the simulated life form (the satellite tobacco mosaic virus) was a very simple and primitive organism containing a total of two proteins, their full simulation ran on a 256-node (512 processors) Intel Itanium 2-based SGI Altix NUMA SMP system at a speed of 1.1ns of simulated time per day.

4.5 Rational Drug Discovery

In order to maintain growth and meet their targets, drug companies need to develop at least several new drugs with high sales potential every year. An important reason behind this pressure on pharmaceutical companies is the relatively short patent protection periods for drugs. Once a company loses patent protection on one of its flagship drugs, revenues from that product fall precipitously as competitors rush to produce generic versions. Faced with the difficulty of coming up with suitable candidates for such drugs using traditional techniques, large pharmaceutical companies started exploring the use of bioinformatics in drug design during the mid-1990’s [24].

Traditional drug discovery heavily relied on animal models of diseases and chemicals whose therapeutic values have been determined to some extent. In contrast, bioinformatics-centered drug discovery starts with the identification of genes associated with a certain disease or desired drug response. The detection of disease-related genes can be accomplished through the use of microarrays that detect the presence of related mRNA. Using comparative genomics techniques, similar sequences in other organisms can be identified and a human protein structure can be modeled. Armed with this information, researchers can then target this protein and design a compound to bind to this molecular target.

Commonly referred to as the rational drug design, this approach to drug discovery requires the use of many different tools in the bioinformaticists' arsenal [25]. Genomics applications such as sequence alignment are used to find homology between human genes and their counterparts in other species. Protein profile searching applications might be used to find similar patterns in human and animal proteins, and protein structure prediction techniques play an important role in finding the chemicals to bind to the physical structures of target proteins. The interplay of different applications and algorithms in the rational drug discovery process suggests that the performance of these diverse applications is crucial for success in rapid and efficient drug discovery.

5. Conclusion

With the current deluge of data, computational methods have become indispensable to biological investigations. Originally developed for the analysis of biological sequences, bioinformatics now encompasses a wide range of subject areas including structural biology, genomics and gene expression studies. In this review, I provided an introduction and overview of the current state of field. In particular, we discussed the types of biological information and databases that are commonly used, examined some of the studies that are being conducted, and finally looked at several practical applications of the field. Two principal approaches underpin all studies in bioinformatics [25]. First is that of comparing and grouping the data according to biologically meaningful similarities and second, that of analysing one type of data to infer and understand the observations for another type of data. These approaches are reflected in the main aims of the field, which are to understand and organize the information associated with biological molecules on a large scale. As a result, bioinformatics has not only provided greater depth to biological investigations, but also added the dimension of breadth. In this way, we are able to examine individual systems in detail and also compare them with those that are related in order to uncover common principles which apply across many systems and highlight unusual features which are unique to some.

Bioinformatics is a developing interdisciplinary science. The involvement of other sciences such as computer science holds great promise. That is, this century's major research and development efforts will probably be in the biological and health sciences. Computer science departments planning to diversify their offerings can thus only gain through early entry into bioinformatics. Still unclear is whether bioinformatics will eventually become an integral part of computer science or will develop into an independent application. Regardless of the outcome, computer scientists are sure to benefit from being active and assertive partners with biologists.

As to the future of the relationship between computer science and biology, it is worth mentioning an interview given by Knuth [26]. He argues that major discoveries in computer science are unlikely to occur as frequently as they did in the past few decades. On the other hand, he states that "Biology easily has 500 years of exciting problems to work on...".

참고문헌

[1] GenBank release 175.0 notes (December 15 2009).
[2] http://www.rcsb.org/pdb/home/home.do
[3] Fleischmann, R. D., et al. "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd", Science, 269(5223):496-512, 1995.
[4] http://www.integratedgenomics.com
[5] Luscombe, N. M., et al. "What is bioinformatics? An introduction and overview", Yearbook of Medical Informatics, 83-99, 2001.

[6] Gibas, C. J. and Jambeck, P. "Developing bioinformatics computer skills", Sebastopol, Calif.: O'Reilly, 2001.

[7] MacMullen, W. J. and Denn, S. O. "Information problems in molecular biology and bioinformatics", Journal of the American Society for Information Science and Technology, 56(5):447-456, 2005.

[8] Pandjassarame Kangueane "Bioinformatics Discovery", Springer Press. 2009.

[9] Benson et al. "GenBank", Nucleic Acids Research, 36(1):D25-D30, 2007.

[10] Bairoch, A. and R. Apweiler. "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000", Nucleic Acids Research, 28(1):45-48, 2000.

[11] Cochrane et al. "Petabyte-scale innovations at the European Nucleotide Archive", Nucleic Acids Research, 37:D19-D25, 2009.

[12] Sugawara, H., et al. "DDBJ with new system and face", Nucleic Acids Research, 36(1):D22-24, 2008.

[13] Mount, D. "Bioinformatics: Sequence and Genome Analysis", 2nd Ed. Cold Spring Harbor Press, Cold Spring Harbor, NY, 2004.

[14] Xing, J. "Essential Bioinformatics", Cambridge University Press, 2006.

[15] Altschul, S. F., et al. "Basic local alignment search tool", Journal of Molecular Biology, 215(3):403-410, 1990.

[16] Pearson, W. R. and D. J. Lipman. "Improved tools for biological sequence comparison", Proc. Natl. Acad. Sci. USA, 85(8):2444-2448, 1988.

[17] Larkin, M. A., et al. "Clustal W and Clustal X version 2.0", Bioinformatics, 23(21):2947-8, 2007.

[18] Jones, D. T., R. T. Miller, and J. M. Thornton. "Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing", Proteins, 23:387-397, 1995.

[19] Rost, B. and S. O'Donoghue. "Sisyphus and prediction of protein structure", Computer Applications in the Biosciences, 13:345-356, 1997.

[20] Morel, N. M., et al. "Primer on medical genomics part xiv: Introduction to systems biology-a new approach to understanding disease and treatment", Mayo Clinic Proceedings, 79:651-658, May 2004.

[21] Finkelstein, A., et al. "Computational challenges of systems biology", IEEE Computer, 37(5):26-33, 2004.

[22] Freddolino, P. L., et al. "Molecular dynamics simulations of the complete satellite tobacco mosaic virus", Structure, 14:437-449, 2006.

[23] Phillips, J. C., et al. "Scalable molecular dynamics with NAMD", Journal of Computational Chemistry, 26:1781-1802, 2005.

[24] Jones, C. "The commercialization of bioinformatics", In Electronic Journal of Biotechnology, volume 3, 2000.

[25] Luscombe, N. M., D. Greenbaum, and M. Gerstein. "What is bioinformatics? A Proposed Definition and Overview of the Field", Method Inform. Med. 4:346-358, 2001.

[26] Knuth, D. Computer literacy interview (Dec. 7, 1993); www.literateprogramming.com/clb93.pdf.

Kibong Kim [Regular member]

- Feb. 1992 : Kyungpook National Univ., Microbiology, BS
- Feb. 1997 : Kyungpook National Univ., Microbiology, MS
- Aug. 2003 : Chungnam National Univ., Computer Engineering, PhD
- Apr. 1999 ~ Aug. 2003 : SmallSoft Co. LtD., Director /CTO
- Sep. 2003 ~ Current : Sangmyung Univ., Dept. of Medical Biotechnology, Associate Professor

<Research Interests>
Biodata Mining, Machine Learning, Genome Informatics