African Journal of Biotechnology Vol. 7 (5), pp. 495-503, 4 March, 2008
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB07.899
ISSN 1684–5315 © 2008 Academic Journals

Review

Recent developments in life sciences research: Role of bioinformatics

Vaddi Srinivasa Rao[1], Sussant Kumar Das[1], Vaddi Janardhana Rao[1] and Gedela Srinubabu[2]*

[1]Department of Computer Science, Berhampur University, Orissa-760007, India.
[2]International center for Bioinformatics and Center for Biotechnology, Andhra University College of Engineering (A), Visakhapatnam-530003, India.

Accepted 22 January, 2008

Life sciences research and development has opened up new challenges and opportunities for bioinformatics. The contribution of bioinformatics advances made possible the mapping of the entire human genome and genomes of many other organisms in just over a decade. These discoveries, along with current efforts to determine gene and protein functions, have improved our ability to understand the root causes of human, animal and plant diseases and find new cures. Furthermore, many future Bioinformatic innovations will likely be spurred by the data and analysis demands of the life sciences. This review briefly describes the role of bioinformatics in biotechnology, drug discovery, biomarker discovery, biological databases, bioinformatic tools, bioinformatic tasks and its application in life sciences research.

Key words: Bioinformatics, drug discovery, biological databases, information technology.

INTRODUCTION

Bioinformatics and computational biology are rooted in 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 and engineering, biology, and behavioral science. Bioinformatics and computational biology each maintain close interactions with life sciences to realize their full potential. 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 their interface.

The growth of the biotechnogy industry in recent years is unprecedented, and advancements in molecular modeling, disease characterization, pharmaceutical discovery, clinical healthcare, forensics, and agriculture fundamentally impact economic and social issues worldwide. Research and discovery activities in the life sciences were once limited to a single gene or protein, but the development of computational and information systems (integrated with biotechnology) has facilitated a shift to high-throughput screening (thousands of samples per day) and high-content detection systems (thousands of data points per sample), and the supporting information system represents the enabling factor in these endeavors. The National Center for Biotechnology Information (NCBI) maintains a growing collection of databases housing genetic sequence and protein structure/activity data (among other biological data sets), which are currently growing at an exponential rate. On February 15, 2005, the NCBI published its 146th release of GenBank, a flat-file database of gene sequences, which contains 42,734,478 gene sequences (www.ncbi.nim.gov). In addition, genomic research enjoyed a landmark accomplishment with the completion and publication of the human genome in February, 2001 (Venter, 2001; Lander, 2001). This represents one of two fundamental advancements in the last 10 years that have moved genomic research almost completely into the computational domain, and dedicated information systems now represent the enabling factor for life sciences research.

As mentioned, the human genome has been sequenced and draft version was published (Venter, 2001).

*Corresponding author. E-mail: srinubabuau6@gmail.com.Tel: +91-891-2844204, +91- 9394073150, +91- 8941- 216037. Fax: +91-891-2755547.

Several updates are publishing to support this draft version to understand the human genome clearly. Yet the boon of high-throughput gene sequencing is not limited to the human species, many other genomes have been completely sequenced representing a breadth of mammalian, agricultural, viral and bacterial organisms, and this fundamental advancement in genomic data availability offers scientists the informational basis of living systems. The other fundamental breakthrough in biotechnology can best be described as "high-content" genomic screening, where the integrity (single nucleotide polymorphisms) and activity (gene expression profiling) of every gene in a known genome can be detected. This is accomplished using the DNA microarray technology platform, which can detect tens of thousands of genes using a small functionalized system the size of a postage stamp (Brown and Botstein, 1999), (Debouck and Goodfellow, 1999) and (Heller, 2002). The information flowing from genomic laboratories using DNA microarray technology constitutes hundreds of thousands of gene-specific measurements each day. The overall impact of this revolutionary technology depends upon an integrated information system to analyze and store the data, as well as computational systems to design each of these gene-specific detections (hybridizations). The Food and Drug Administration (FDA) has recently published guidelines for the development of biotechnology methods, such as the DNA microarray platform, for use in genome-based prognostics and diagnostics in humans (www.fda.gov/cdrh) which marks the beginning of a new era in healthcare that utilizes a patient's genome sequence to enable "personalized medicine."

With this explosion of molecular data and biotechnology capabilities, the pharmaceutical, biotechnology, and healthcare industries are dependent on professionals with information systems and technology skills to fully exploit these resources and translate molecular and cellular data into new genetic and therapeutic discoveries, as well as develop new biotechnologies to impact human health. Similarly, information technology professionals must be trained in the utilization of biomedical data structures and capable of developing and integrating information systems with biotechnology systems to meet this challenge.

Given that bioinformatics has evolved to meet the objectives of life sciences research, courses and curricula that offer training in bioinformatics, or biomedical informatics (Friedman, 2004) have primarily originated in life science programs where traditional life sciences training in genetics, genomics and proteomics is augmented with training in sequence alignments, genomic data analysis and protein modeling. This training largely involves the use of existing software applications dedicated to queries in gene and protein sequence databases. Yet this training is not appropriate for students enrolled in information technology programs that lack what is considered prerequisite knowledge in biology and genetics. However, students enrolled in information technology programs are developing skills that are directly applicable to biomedical informatics such as data formats, database structure and development, applications development, and systems design. Given that both fields of study are important to training in biomedical informatics, interdisciplinary training must be developed to improve the skills relevant for both the groups.

It is important to mention that the terminology utilized for training information technology students is described herein as "biomedical informatics" since the learning objective of this training is to apply information sciences and technology skills to all sub disciplines within this domain, and aspects from all defined sub disciplines are included in the courses described. It is recognized that "bioinformatics" is a sub discipline of "biomedical informatics" (Friedman, 2004) and the primary emphasis of subject matter in the course series falls within the realm of bioinformatics. This larger emphasis on bioinformatics is intentional simply to provide the necessary background in areas most lacking in existing information technology education, since students have little or no background in life sciences.

DRUG DISCOVERY

Developing a new drug is a tedious and expensive undertaking, despite promising discoveries and multi-billion dollar investments for new drug development is quietly undergoing crisis. The recently developed high-throughput experimental technologies, summarized by the terms genomics, transcriptomics, proteomics and metabolomics provide for the first time ever the means to comprehensively monitor the molecular level of disease processes. The "-omics" technologies facilitate the systematic characterization of a drug target's physiology, thereby helping to reduce the typically high attrition rates in discovery projects, and improving the overall efficiency of pharmaceutical research processes. Currently, huge amount of data is producing day by day from the new experimental technologies automatically bottleneck occurs for analysis of this data. A lack of scalable database systems and computational tools for target discovery has been recognized as a major hurdle. Hans Peter Fischer (2005) reviewed on novel *in silico* methods to reconstruct regulatory networks, signaling cascades, and metabolic pathways, with an emphasis on comparative genomics and microarray-based approaches. Promising methods, such as the mathematical simulation of pathway dynamics are discussed in the context of applications in discovery projects.

Drug discovery can be categorized as a series of processes that can be measured by the number of candidates identified in a given period with a defined level of resources. Productivity and speed become critical discovery performance metrics. The discovery process is typically defined as being composed of four distinct, yet related, processes: (1) target identification/validation, (2) lead identification, (3) lead optimization, and (4) discovery /development interface, with early drug discovery encompassing the first three of these processes (Shayne, 2005). The successes in Target Identification due to the application of DNA sequencing and genomics databases have created serious bottlenecks downstream in the drug discovery pipeline. The pharmaceutical industry has eased the bottleneck at the Lead Identification step by employing bioinformatic tools to test large libraries and databases against a growing list of targets. Computational systems biology is an emerging field in biological simulation that attempts to model or simulate intra- and intercellular events using data gathered from genomic, proteomic or metabolomic experiments. The need to model complex temporal and spatiotemporal processes at many different scales has led to the emergence of numerous techniques, including systems of differential equations, Petri nets, cellular automata simulators, agent-based models and pi calculus (Wayne and Wishart, 2007). Unprecedented growth in the interdisciplinary domain of biomedical informatics reflects the recent advancements in genomic sequence availability, high-content biotechnology screening systems, as well as the expectations of computational biology to command a leading role in drug discovery and disease characterization. These forces have moved much of life sciences research almost completely into the computational domain. Importantly, educational training in biomedical informatics has been limited to students enrolled in the life sciences curricula (Thesseling, 2003), yet much of the skills needed to succeed in biomedical informatics involve or augment training in information technology curricula. Kane and Brewer (2007) described the methods and rationale for training students enrolled in information technology curricula in the field of biomedical informatics, which augments the existing information technology curriculum and provides training on specific subjects in biomedical informatics not emphasized in bioinformatics courses offered in life science programs, and does not require prerequisite courses in the life sciences.

Systems biology is one such approach to analyze the published DNA sequence of human genome, and has been increasingly recognized as a very important area of research, as it places specific molecular targets within a context of overall biochemical action. Understanding the complex interactions between the components within a given biological system that lead to modifications in output, such as changes in behavior or development, may be important avenues of discovery to identify new therapies. Central nervous system (CNS) drug discovery in the post-genomic era is rapidly evolving (Shaikh and Kerwin, 2002). Older empirical methods are giving way to newer technologies that include bioinformatics, structural biology, genetics, and modern computational approaches. In the search for new medical therapies, and in particular treatments for disorders of the central nervous system, there has been increasing recognition that identification of a single biological target is unlikely to be a recipe for success; a broad perspective is required (Nichols, 2006). Computational models of cells, tissues and organisms are necessary for increased understanding of biological systems. In particular, modeling approaches will be crucial for moving biology from a descriptive to a predictive science (Tramontano, 2006). Pharmaceutical companies identify molecular intervenetions that they predict will lead to therapies at the organism level, suggesting that computational biology can play a key role in the pharmaceutical industry. Kumar (2006) discussed the pharmaceutically-relevant computational modeling approaches currently used as predictive tools.

BIOINFORMATICS IS A KEY LIFE SCIENCE R AND D ACTIVITY

Rapid advances in technologies like genomic as well as bioinformatics coupled with a unique collaboration between industry and academia are beginning to show the true potential for the human genome project to affect patient healthcare. By knowing the sequence of the human genome and beginning to unravel the location and sequence of all genes and their variants, scientists can establish a better understanding of the mechanisms for diseases, with subsequent availability of new treatments. Because of the vast amount of data coming out of the Human Genome Project, bioinformatics tools and databases have become an integral part of pharmacogenomic and disease susceptibility gene research. They play an important role in candidate gene identification, gene finding, SNP detection, genotyping and genetic analysis. Public sources of databases and tools abound, although it is sometimes difficult to determine the quality, consistency and sustainability of these sources. The data-management challenges arising from this heady sampling of the genome were making a strong impression, in both the public and private sectors, and the as-yet-unresolved (and highly charged) question of the patent ability of genes led to a land rush on intellectual property (Kiley, 1992).

Bioinformatics data integration and tool standardization are critical to the success of association and linkage studies. The underlying data models accommodate the variability inherent in subject collections, the ability to trace the data source, and the automation and archival storage of analysis results. A fully traceable data source is important, as we are often faced with anomalies in data at a late stage that can be very time consuming to resolve in an infrastructure that does not facilitate data integration. The polymorphism database component includes data from public and proprietary sources. The subject phenotypes (a relevant measure of disease severity, disease progression and/or disease sub classification for disease genetics or a relevant measure of drug response for pharmacogenetics) and genotype components are fully integrated with the source databases.

The subject database component also includes reference collections and allele frequency information needed for analysis. This model has proven useful in analyzing reasonably large datasets. The model is scalable to variations in volumes and expandable to accommodate a variety of markers. The performance for very high volumes (e.g. genome-wide scans of a large population) is currently being investigated. SNPs are the most common markers for disease-gene and drug-response associations (McCarthy and Hilfiker, 2000). However, to detect association at a SNP near a complex disease gene, the appropriate SNPs must be chosen for analysis. In addition, the order and relationship of SNP markers is extremely important.

The cost of doing high-density genome-wide association scans is still quite high, so, using a haplotype-based SNP map would maximize the information content and reduce the resource needs. The use of haplotypes has been discussed in great detail, including their benefits and limitations (Stephens, 1999; Marth, 2001). One limitation of haplotypes that needs to be considered is the fact that frequencies of most clinically significant AEs are low (< 5–10%) so the use of commonly occurring haplotypes (those with frequencies of at least 10%) may overlook important genetic associations (Lai, 2002). Another approach that has been advocated to reduce the cost of genotyping is DNA pooling. Instead of analyzing SNPs from individual subjects, DNA from responders is pooled and compared with pooled DNA from control subjects. The advantages and disadvantages of this approach are reviewed in detail elsewhere (Chanock, 2001).

DISEASE GENETICS AND PHARMACOGENETICS

Genotypic data can be combined with accurate phenotypic data and analyzed to determine the SNPs and/or haplotypes associated with disease susceptibility and/or drug response. A high-density genome association scan can be used to thoroughly evaluate the genes that modify a patient's response to medications (i. e. pharmacogenetics) and to push the limits of disease gene identification in appropriate populations (i.e. disease genetics). Examples of the use of the candidate gene approach and/or the whole genome scan approach are described below as they relate to disease genetics and pharmacogenetics.

Disease genetics

In the past, disease genetics has focused on monogenic diseases such as Huntington's disease in which the expression of a particular variant of a single gene will, in the vast majority of cases, lead to disease. There are innumerable monogenic diseases, each of which affects only a small number of patients. In contrast, disease genetics research is now focused on identification of genes associated with common diseases (diseases affecting thousands or millions of people). These common diseases are multi-factorial [i.e. dependent on complex interactions between numerous environmental factors and a number of alternative forms (alleles) of genes called disease susceptibility genes] and polygenic (involving more than one gene in their multi-factorial pathogenesis) Middleton, 2000). The overall goal of disease genetics is to identify how genetic variation can influence disease susceptibility and to improve our understanding of the molecular processes resulting in clinically overt disease. New treatments can then be designed to target these molecular processes to prevent and/or treat the disease.

Typically, new disease susceptibility genes have been identified using a combination of linkage and association studies. The linkage studies involve collection of DNA samples and extensive clinical phenotypic data from multiple members of affected families. Markers are typed throughout the genome, and, using linkage analysis algorithms, chromosomal regions harboring disease genes are identified (Stoll, 2000). The regions are identified using highly informative markers on the basis of their chromosomal location by taking advantage of the meiotic process of recombination as apparent in families segregating for the disease (Kruglyak, 1999).

Markers closest to the disease gene show the strongest correlation with disease patterns in families. These linkage studies allow identification of a region on a chromosome and large portions (1–20 cM) of the DNA (which may include 10–1000 genes) that may be linked to a specific disease. Candidate genes within the region can sometimes be inferred from the genome-wide databases that are currently available. Unfortunately, most of the few validated disease genes were not obvious candidates. Association studies are then conducted to identify the causative mutation responsible for the disease either using family-based association studies or unrelated case-control association studies. The key to success for linkage and association studies is the availability of high quality clinical information, available appropriate genotypic data and the ability to link such data (see above). Linkage and/or association studies have been reported to identify susceptibility genes for many therapeutic areas.

The potential benefits of the human genome project are beginning to be realized with the availability of technology advances and bioinformatics tools. The identification of disease susceptibility genes and the development of many new treatments are the longer-term benefits. In the shorter term, the benefits will be the ability to predict those patients at risk for experiencing adverse reactions or patients with a high probability of experiencing improved efficacy (i.e. pharmacogenetics). As progress is made in the area of disease genetics and pharmacogenetics, our understanding of disease susceptibility and its interrelationship with drug response will improve, making targeted therapy (i. e. the right drug to the right patient) a reality.

BIOINFORMATICS FOR CLINICAL DECISION SUPPORT SYSTEMS

One of the most promising areas in bioinformatics is computer-aided diagnosis, where a computer system is capable of imitating human reasoning ability and provides diagnoses with an accuracy approaching that of expert professionals. This type of system could be an alternative tool for assisting dental students to overcome the difficulties of the oral pathology learning process. Borra et al. (2007) developed an open decision-support system based on Bayes' theorem connected to a relational database using the C++ programming language, developed software was tested in the computerization of a surgical pathology service and in simulating the diagnosis of 43 known cases of oral bone disease. The simulation was performed after the system was initially filled with data from 401 cases of oral bone disease. The system allowed the authors to construct and to manage a pathology database, and to simulate diagnoses using the variables from the database. The integration of patient-specific genomic information into the electronic medical record (EMR) will create many opportunities to improve patient care. Key to the successful incorporation of genomic information into the EMR will be the development of laboratory information systems capable of appropriately formatting molecular diagnostic and cytogenetic findings in the EMR. Due to the lack of granular genomics-related content in existing medical vocabularies, the adoption of new standards for describing clinically significant genomic information will be an important step toward recognizing the genome-enabled EMR (Hoffman, 2007). Appropriate capture of patient-specific genomic results in the EMR will generate new opportunities to utilize this information in clinical decision support, including automated response to pharmacogenomic-based risks

DATABASES AND TOOLS USED IN BIOINFORMATICS

The functional aspect of bioinformatics is the representation, storage, and sharing of data. The design of databases, design and development of tools to retrieve data from the databases and creation of user interfaces are considered as the infrastructure of bioinformatics. Biomedical researchers are using the computerized databases since 1960s (Neufeld and Cornog, 1999). After bioinformatics came into existence during mid 1980s, the US government has established the National Center for Biotechnology Information through a legislation in 1988. This is a division of National Library of Medicine. NCBI is the official agency for creation of information and Help, Control, Health and Disease.

The growing usage of information technology in biological sciences paved way to a large number of websites, databases, tools and software available on the world wide Web for open access. In addition, huge amounts of literature is also available for ready reference. The internet has changed the way in which the data in a Central Data Warehouse is shared by the researchers across the globe.

Biological databases

The rapid development of genome technologies, especially automatic sequencing techniques, has produced a huge amount of data consisting essentially of nucleotide and protein sequences. For instance, the number of sequences in GenBank increases exponentially. To store, characterize, and mine such a large amount of data requires many databases and programs hosted in high-performance computers. Until now, there has been several databases, for example GenBank (Benson, 2004) Uniprot (Apweiler et al., 2004), PDB (Berman, 2000) KEGG (Kanehisa et al., 2000), PubMed Medline, etc., covering not only nucleotide and protein sequences but also their annotations and related research publications. The programs include those for sequence alignment, prediction of genes, protein structures, and regulatory elements, etc., some of which are organized into packages such as EMBOSS (Rice et al., 2000) PHYLIP (felsenstein, 1989) and GCG Wisconsin. In general, these databases are built independently by various academic or commercial organizations and their input and output data formats follow their own standards (e.g., Fasta, Genbank, EMBL, SRS, etc.), most of which are incompatible. The programs themselves are even more complex in that they are implemented using a variety of programming languages and on different operating systems, are operated in different ways using input and output data in a wide range of formats. Biologists try to discover biological functions from sequences using informatics techniques but are frequently frustrated by the processes of searching for suitable tools, learning how to use these tools, and translating data formats between them.

To facilitate biologists' research, an integrative informatics platform is needed in which many kinds of databases and programs are integrated with a common input–output data format and uniform graphical user interface (GUI). To build such an integrative informatics platform, workflow is recognized as a potential solution. Some existing efforts include Biopipe (Hoon et al., 2003) BioWBI (Leo, 2004) Taverna et al. (2004) Wildfire et al. (2005) etc. All of them provide mechanism to integrate bioinformatics programs into workflows. Biopipe is based on programming language perl. It looks lack of user-friendly interface for building workflow so far. BioWBI and

Tarverna use Web-Services for components to construct workflows. However, to convert a 3rd-party program into Web-Services, they lack of integrative GUI environment. Wildfire aims at using workflow to provide huge computing capability to bioinformatics application. However, there is no integrative environment provided for multiple users to collaborate in the same large-scale bioinformatics project. (Ambiguity sentence deleted)

A biological Database is a huge collection of persistent data supported by software meant for update, retrieve and maintain data. There are different types of databases depending on the nature and type of data stored in the database. The data in a biological database may be of type sequences or structures 2D gel or 3D structure images. In the case of protein sequence analysis, primary, composite and secondary databases are needed. These databases store different levels of protein sequence information.

Primary databases

The growing demands for the sequence information during 1980s, a lot of primary database projects were taken up and resulted in the creation of nucleic acid and protein sequence databases. Some of the important DNA sequence databases are Gene Bank (USA), EMBL (Europe) and DDBJ (Japan). These databases exchange data on regular basis to ensure consistency of data.

The early 1960s witnessed the development of the Protein Sequence database at the National Biomedical Research Foundation (NBRF). Currently this database is split into four distinct sections designated as PIR1 thru PIR4. They differ in terms of the quality of data and the level of annotation. Some of the important protein sequence databases are MIPS, SWISS – PORT etc. The primary databases suffer from the problem of proliferation which gives rise to a variety of problems. These problems are alleviated by the development of Composite Protein sequence databases that provide sequence searching more effective. NRDGB is such a database that contains comprehensive and up-to-date information.

Secondary databases

Secondary databases contain the results of analysis of the sequences in the primary databases. Secondary databases contain pattern data. As an example SWISS – PORT has emerged as a popular primary source for a number of secondary databases like PROSITE, profiles, Pfam etc. A Tertiary database is derived from the information stored in secondary databases. In addition to these databases composite protein pattern databases, structure classification databases are also available. A bibliographic database is a database that is used for collecting published articles, abstracts and full research papers with links to individual records. Researchers and scientists extensively use Pubmed and Agricola for their studies.

BIOINFORMATICS TOOLS

Earlier generation of Bioinformatics used tools and applications with text based interface. BLAST is a most popular tool that is widely used by biologists (Madden, 1996). This is an algorithm for searching large databases of Protein or DNA sequences. The NCBI provides web based implementation that searches the massive sequences and annotated data. Programming languages like Perl and Python are used to interface with biological databases and parse output from programs written in routine languages like C, C++ etc., to implement bioinformatics algorithms. Bioinformatic meta search engines like sequence profiling tools are available to find relevant information from several databases.

The recent development is, Simple Object Access Protocol (SOAP) (http://www.w3.org/TR/soap) based interfaces, developed for a variety of bioinformatics applications that allow using programs running on one computer in part of the world to use algorithm, data and computer resources on servers in other parts of the world. The large availability of SOAP based bioinformatics and web services along with the open source bioinformatics collections lead to the next generation bioinformatics tools called integrated bioinformatics platform. These tools range from a web based interface to an extensible bioinformatics work flow development environment.

Some of the bioinformatics programmers have set up free open source bioinformatics projects to develop and distribute the tools and modules they produce (Vallabhajosyula and Sauro, 2007) described the development of a useful graphical user interface for stochastic simulation of biochemical networks which allows model builders to run stochastic simulations of their models and perform statistical analysis on the results. These include the construction of correlations, power-spectral densities and transfer functions between selected inputs and outputs. The software is licensed under the BSD open source license and is available at http://sourceforge.net/projects/jdesigner. In addition, a more detailed account of the algorithms employed in the tool can be found at the Wiki at http://www.sys-bio.org/sbwWiki. Stajich and Lapp (2006) reviewed the important work in open-source bioinformatics software that has occurred over the past couple of years. The survey is intended to illustrate how programs and toolkits whose source code has been developed or released under an Open Source license has changed informatics-heavy areas of life science research. Rather than creating a comprehensive list of all tools developed over the last 2-3 years, they used a few selected projects encompassing toolkit libraries, analysis tools, data analysis environments and interoperability standards to show how freely available and modifiable open-source software can serve as the foundation for building important applications, analysis workflows and resources. Argraves et al. (2005) presented ArrayQuest, a web-based DNA microarray analysis process controller. Key features of ArrayQuest are that (1) it is capable of executing numerous analysis programs such as those written in R, BioPerl and C++; (2) new analysis programs can be added to ArrayQuest Methods Library at the request of users or developers; (3) input DNA microarray data can be selected from public databases (i.e., the Medical University of South Carolina (MUSC) DNA Microarray Database or Gene Expression Omnibus (GEO)) or it can be uploaded to the ArrayQuest center-point web server into a password-protected area; and (4) analysis jobs are distributed across computers configured in a backend cluster. To demonstrate the utility of ArrayQuest they have populated the methods library with methods for analysis of Affymetrix DNA microarray data.

TASKS OF BIOINFORMATICS

The tasks of Bioinformatics involve the analysis of sequence information. This involves the following activities.

Identifying the genes in the DNA sequences from various organisms.
Identifying families of related sequences and the development of models.
Aligning similar sequences and generating Phylogenetic trees to examine evolutionary relationships.
Finding all the genes and proteins of a genome from a given sequence of amino acids.
Predicting active sites in the protein structures to attach drug molecules.

Gene ontology, a semantic framework could be used to underpin a range of important bioinformatics tasks, such as the querying of heterogeneous bioinformatics sources or the systematic annotation of experimental results (Baker et al., 1999).

Applications of bioinformatics

Computational biology has found its applications in many areas. It helps in providing practical tools to explore Proteins and DNA in number of other ways. Bio-computing is useful in recognition techniques to detect similarity between sequences and hence to interrelate structures and functions. Another important application of Bioinformatics is the direct prediction of protein 3-Dimensional structure from the linear amino acid sequence. It also simplifies the problem of understanding complex genomes by analyzing simple organisms and then applying the same principles to more complicated ones. This would result in identifying potential drug targets by checking homologies of essential microbial proteins. Bioinformatics is useful in designing drugs.

Aims of bioinformatics

The aims of Bioinformatics are:

1. To organize data in a way that allows researchers to access existing information and to submit new entries as they are produced.
2. To develop tools and resources that aid in the analysis and management of data.
3. To use this data to analyze and interpret the results in a biologically meaningful manner.
4. To help researchers in the Pharmaceutical industry in understanding the protein structures to make the drug design easy.

Backofen and Gilbert (2004) discussed the definition of bioinformatics, give a classification of the problem areas which bioinformatics addresses, and illustrate these in detail with examples.

Information technology and medical sciences integration

If we closely examine, the evolution of the Computer Applications in the field of Medical Diagnostics for the last two and a half decades, it is evident that it has evolved as a good usher rather than an assistant or a supporter. Presently the computers are extensively used by experts of various fields in solving complex tasks that are beyond the reach of human capabilities that are limited in nature. The field of medicine, particularly the medical diagnostics falls in this ambit. The modern medical diagnosis, calls for a multi-dimensional knowledge unlike the older practices and usage. Such knowledge can not be found in one expert. As the complexity of the diseases is in rise day in and day out, the medical diagnosis has become an intricate job and a difficult task. The quest for accuracy in diagnosis calls for computer applications in this context. Figure 1 represents the role of bioinformatics for the target therapeutic discovery in connection with proteomics, information technology and life sciences research.

Doctors, medical societies, and associations could critically appraise internet information and act as decentralized "label services" to rate the value and trustworthiness of information by putting electronic evaluative and descriptive "tags" on it (Eysenbach et al., 1998). After the advent of Information Technology, Communications Backup and the data available on-line, the doctors are

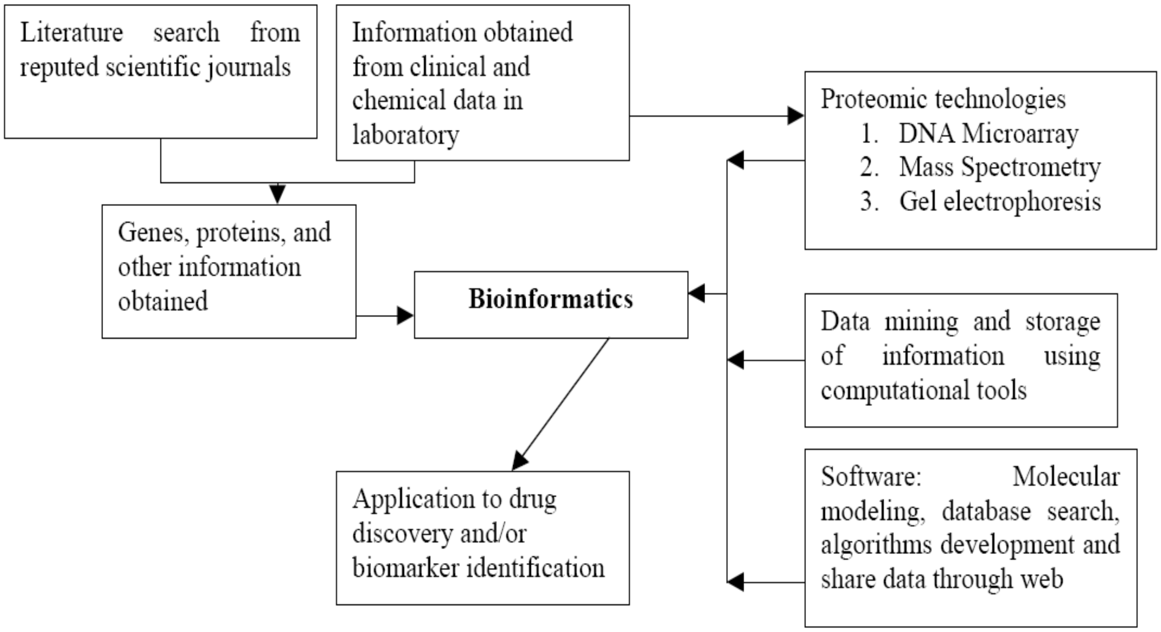
Literature search from reputed scientific journals
Information obtained from clinical and chemical data in laboratory
Proteomic technologies
1. DNA Microarray
2. Mass Spectrometry
3. Gel electrophoresis
Genes, proteins, and other information obtained
Bioinformatics
Data mining and storage of information using computational tools
Software: Molecular modeling, database search, algorithms development and share data through web
Application to drug discovery and/or biomarker identification

Figure 1. Bioinformatics relationship to life sciences, proteomics and information technology.

able to make accurate diagnosis and prescribe the patients a suitable treatment.

Information and communication technology has made in roads into all diverse branches of medicine. Research and Development made in the field of information technology has brought a great revolution in medical imaging, tele co-operation, education and training. As a result of the information and communication technology telemedicine came. These technologies made it possible to gather and disseminate the medical information for the best medical practices. Ongoing digitization of patient information will greatly facilitate the assessment of treatment outcomes. The challenge in this area is to distribute this information efficiently and promptly. This has been met, to some extent, by the moves toward so-called evidence-based medicine (Sackett et al., 1996)

The said technologies are to be developed further and fine tuned in the days to come so as to make it affordable and accessible. Another important development of the technology is to see the real time video consultation or multimedia consultation could happen. Medical Informatics is a fast emerging field that can be defined as the study, invention, and implementation of structures and algorithms to improve communication, understanding and management of medical information.

REFERENCES

Apweiler R (2004). UniProt: the universal protein knowledgebase, Nucleic Acids Res. 32: 115119.

Argraves GL, Jani S, Barth JL, Argraves WS (2005). ArrayQuest: a web resource for the analysis of DNA microarray data, BMC Bioinformatics 6: 287.

Backofen R, Gilbert D (2001). Bioinformatics and Constraints, Constraints 6: 141-156.

Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R, Brass A (1999). An ontology for bioinformatics applications., Bioinformatics, 15: 510-520.

Benson DA (2004). GenBank: update. Nucleic Acids Res., 32: 23-26.

Berman DA (2000). The protein data bank, Nucleic Acids Res., 28: 235-242.

Borra RC, Andrade PM, Corrêa L, Novelli MD (2007). Development of an open case-based decision-support system for diagnosis in oral pathology., Eur. J. Dent. Educ. 1: 87-92.

Brown PO, Botstein D (1999). Exploring the new world of the genome with DNA microarrays, Nat. Genet., 21(Suppl 1): 33-37.

Chanock S (2001). Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease. Dis. Markers 17: 89-98

Debouck C, Goodfellow PN (1999). DNA microarrays in drug discovery and development, Nat. Genet. 21(Suppl 1): 48-49.

Eysenbach G, Thomas LD (1998). Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information, BMJ 317: 1496-1502.

Felsenstein J (1989). Phylip: phylogeny inference package (Version 3.2). Cladistics 5: 164-166.

Fischer HP (2005). Towards quantitative biology: Integration of biological information to elucidate disease pathways and to guide drug discovery; Biotechnol. Annu. Rev., pp. 1-68.

Friedman CP (2004). Training the next generation of informaticians: The impact of “BISTI” and bioinformatics—A report from the American College of Medical Informatics, J. Am. Med. Informatics Assoc. 11(3): 167-172.

Guidance for Industry and FDA Staff: Class II Special Controls Guidance Document: Drug Metabolizing Enzyme Genotyping System. Available from: www.fda.gov/cdra

Heller MJ (2002). DNA microarray technology: devices, systems, and applications, Annu. Rev. Biomed. Eng. 4: 129-153.

Hoffman MA (2007). The genome-enabled electronic medical record, J. Biomed. Inform. 40(1): 44-46.

Hoon S (2003). Biopipe: a flexible framework for protocol-based bioinformatics analysis, Genome Res. 13: 1904-1915.

Kane MD, Jeffrey LB (2007). An information technology emphasis in biomedical informatics education, J. Biomed. Informatics; 40 67-72.

Kanehisa M, Goto S (2000). KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28: 27-30.

Kiley TD (1992). Patents on random complementary DNA fragments? Science. 257: 915-918.

Kruglyak L (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes, Nat. Genet. 22: 139-144.

Kumar N (2006). Applying computational modeling to drug discovery and development, Drug Dis. Today 11: 806-811.

Lai E (2002). Medical applications of haplotype-based SNP maps: shouldn't we learn how to walk before trying to run? Nat. Genet. 32: 353.

Lander ES (2001). Initial sequencing and analysis of the human genome, Nature 409(6822): 860-921.

Leo PM (2004). BioWBI: an Integrated Tool for building and executing Bioinformatic Analysis Workflows, Bioinformatics Italian Society Meeting (BITS 2004) Padova.

Madden TL, Tatusov RL, Zhang J (1996). Applications of network BLAST server, Methods Enzymol. 266: 131-141.

Marth T, (2001). Single-nucleotide polymorphisms in the public domain: how useful are they? Nat. Genet. 27: 371-372.

McCarthy JJ, Hilfiker R (2000). The use of single-nucleotide polymorphism maps in pharmacogenomics, Nat. Biotechnol., 18: 505-508.

Middleton H (2000). From gene-specific tests to pharmacogenetics. Commun. Genet. 3: 198-203.

Neufeld L, Cornog M (1999). Database history: From dinosaurs to compact discs, J. Am. Soc. Inf. Sci. 37: 183-190.

Rice P, Longden I, Bleasby A (2000). EMBOSS: the european molecular biology open software suite, Trends Genet. 16: 276-277.

Sackett D, Rosenberg W, Gray J (1996). Evidence based medicine: what it is and what it isn't. BMJ 312: 71-72.

Shaikh S, Kerwin RW (2002). Receptor pharmacogenetics: relevance to CNS syndromes, Br. J. Clin. Pharmacol. 54(4): 344-348.

Shayne CG (2005), Drug Discovery Handbook, Wiley-Interscience, ISBN 0-471-21384-5.

Stajich JE, Lapp H (2006). Open source tools and toolkits for bioinformatics: significance, and where are we? Brief Bioinform. 7: 287-296.

Stephens JC (1999). Single-nucleotide polymorphisms, haplotypes and their relevance to pharmacogenetics. Mol. Diagn., 4: 309-317.

Stoll M (2000). New Target Regions for Human Hypertension via Comparative Genomics. Genome Res. 10: 473-482.

Thesseling G (2003). Expanding Access to Bioinformatics Training and Proteomics Software Tools Through Strategic collaboration, Am. Biotechnol. Lab., pp. 10-12.

Vallabhajosyula RR, Sauro HM (2007). Stochastic simulation GUI for biochemical networks., Bioinformatics, 23: 1859-1861.

Venter J (2001). The sequence of the human genome, Science 291(5507): 1304-1351.

Wayne M, Wishart DS (2007). Computational systems biology in drug discovery and development: methods and applications Drug. Discov. Rev. 12: 295-303.