Chapter-2

Medical Sciences in Information Technology age

2.1 Information Technology in Medical Sciences

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. Recent improvements of computer technology have resulted in a new approach to medical illustration and the scientific research process (Corl F.M et al, 2000). 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. Technical issues in medical informatics are addressed, with emphasis on tools for medical practitioners to willingly and effectively use computers to capture data and to access information; the conversion of paper records to electronic data to facilitate automation; and system and application integration based on patient medical documents and information (Chang Ifay F, et al, 1990). The introduction of computers into the clinical laboratory raises issues that are difficult to resolve by the methods of information science or medical science applied in isolation. The melding of these two disciplines, together with the contributions of other disciplines, has created a new field of study called medical information science (TL Lincoln and RA Korpman, 1980). Information technology provides an
opportunity for the health care industry to fulfill its mandate for delivering high-quality care in a cost-effective manner. Technical developments, such as the Intranet and computer-based medical records; regulatory mandates; and employer cooperative groups are pushing the industry toward paying increased attention to the quality and cost of care (Suri S et al, 2002). After the advent of Information Technology, Communications Backup and the data available on-line, the doctors are able to make accurate diagnosis and prescribe the patients a suitable treatment.

Information and communications technologies (ICT) range from the simple (telephone, audio-conferencing) to the sophisticated (virtual environments, learning repositories) and can increase access to medical education and enhance learning and collaboration for learners at all levels and for institutions. While ICTs are being used and offer further potential for medical education enhancement, challenges exist, especially for rural areas. Finally, there is need for more rigorous research to more clearly identify advantages and disadvantages of specific uses of ICTs in medical education (Sargeant J.M et al, 2005). Information Technology and communication technology has made inroads into all diverse branches of medicine. Research and Development made in the field of information technology has brought a great revolution in medical imaging, telecooperation, education and training. The rapid growth of diagnostic-imaging technologies over the past two decades has dramatically increased the amount of non textual data generated in clinical medicine. The architecture of traditional, text-oriented, clinical information systems has made the integration of digitized clinical images with the patient record problematic. Systems
for the classification, retrieval, and integration of clinical images are in their infancy. Recent advances in high-performance computing, imaging, and networking technology now make it technologically and economically feasible to develop an integrated, multimedia, electronic patient record (Lowe H.J et al, 1995). As a result of the information Technology and communication technology, telemedicine came to the force. These technologies made it possible to gather and disseminate the medical information for the best medical practices. Most hospital policies prohibiting the use of wireless devices cite reports of disruption of medical equipment by cellular telephones. The past few years have seen rapid advances in communication and information technology (C&IT), and the pervasion of the worldwide web into everyday life has important implications for education. Most medical schools provide extensive computer networks for their students, and these are increasingly becoming a central component of the learning and teaching environment (Ward J.P.T. et al, 2001).

One of the most significant developments in healthcare over the past 25 years has been the widespread deployment of information and communication technologies. These technologies have had a wide-ranging impact on the organization of healthcare, on professional practice and on patients' experience of illness and its management (Heath C et al, 2003).

2.2 Drug Design

Drug design is the approach of finding drugs by design, based on their biological targets. Typically a drug target is a key molecule involved in a particular
metabolic or signaling pathway that is specific to a disease condition or pathology. Computer-aided drug design methods are effective tools for drug discoveries. It is only a small number of compounds that are required to be synthesized in practice, if the computer-aided drug design methods are used to predict the activities of the compounds. The methods are based on the interaction theory between drugs and their target proteins. Steric, electrostatic and hydrophobic complementarities are important in the interaction between the drug and the target protein (Hirono S et al, 2002).

Some approaches attempt to stop the functioning of the pathway in the diseased state by causing a key molecule to stop functioning. Drugs may be designed that bind to the active region and inhibit this key molecule. Modern strategies of computer-aided drug design (CADD) are reviewed. The task of CADD in the pipeline of drug discovery is accelerating of finding the new lead compounds and their structure optimization for the following pharmacological tests. The main directions in CADD are based on the availability of the experimentally determined three-dimensional structure of the target macromolecule (Veselovsky A.V. Ivanov A.S.et al, 2003). However these drugs would also have to be designed in such a way as not to affect any other important molecules that may be similar in appearance to the key molecules.

The structure of the drug molecule that can specifically interact with the biomolecules can be modeled using computational tools. These tools can allow a drug molecule to be constructed within the biomolecule using knowledge of its structure and the nature of its active site. Nowadays the in silico scenario for drug design is
totally dependent on structural biology and structural bioinformatics. A myriad of free bioinformatics applications and services have been posted on the web. The information is given in a logical manner, following the drug design process i.e. characterization of a protein target, modelling the protein using sequence homology, optimization of the protein structure and finally docking of small ligands into the active site (Carpy A.J.M., Marchand-Geneste N et al, 2006)

2.2.1 Information Technology in Drug Discovery

New drug discovery from early on involved a trial-and-error approach on naturally derived materials and substances until the end of the nineteenth century. The first half of the twentieth century witnessed systematic pharmacological evaluations of both natural and synthetic compounds. However, most new drugs until the 1970s were discovered by serendipity. With the exponential development of molecular biology on one hand and computer technology on the other, it became possible from 1980 onwards to place drug discovery on a rational basis. Cloning of genes has led to the development of methodologies for specific receptor-directed and enzyme-directed drug discoveries. Advances in recombinant DNA and transgenic technologies have enabled the production of human hormonal and other endogenous biomolecules as new drugs (Kaul P.N et al, 1998). As we have entered the 21st century, the convergence of pharmaceutical science and informational technology is providing unprecedented opportunities for prevention and management of diseases. The application of software technology has penetrated into all spheres of drug discovery, development and manufacturing process. The advent of combinatorial chemistry has
revolutionized the new drug discovery processes. The information technology driven approach involving the automated High Throughput Screening process now permits up to 50,000-100,000 samples to be studied per day (Mukherjee A.K, Ghosh A.C et al, 2002). Drug industry is one of the major players guiding the development of the bioinformatics field. Many of the large pharmaceutical companies have established internal bioinformatics groups whose purpose is to beat the competition to solutions of a problem that may give their company that crucial edge in producing the next major drug. Information technologies for chemical structure prediction, heterogeneous database access, pattern discovery, and systems and molecular modeling have evolved to become core components of the modern drug discovery process. As this evolution continues, the balance between in silico modeling and 'wet' chemistry will continue to shift and it might eventually be possible to step through the discovery pipeline without the aid of traditional laboratory techniques (Augen J et al, 2002). Systems biology promises to impact significantly on the drug discovery process. One of its ultimate goals is to provide an understanding of the complete set of molecular mechanisms describing an organism. Although this goal is a long way off, many useful insights can already come from currently available information and technology (Apic, G, et al, 2005). As the pharmaceutical industry changes in response to unprecedented challenges, information technology supporting the drug discovery process must evolve and be redefined in order to improve research productivity and success rates. Improvements to traditional computational chemistry applications will continue to provide incremental benefits, but drug discovery's productivity imperative can only be
met from the more widespread use of vital decision support tools and through discovery informatics (Koontz B.S et al, 2005).

2.2.2 Analysis in Drug Discovery

The completion of the first draft of the human genome has provided an unprecedented opportunity to understand the genetic and molecular basis of disease. Parallel developments of new biological technologies, such as transcript profiling, allow scientists to examine almost any biological system in high molecular resolution. Contemporary drug discovery research is now focusing on the identification and validation of pharmaceutical targets in the molecular pathways/systems embedded in this information (Bumol, T.F, Watanabe, A.M et al, 2001). Traditionally, pharmaceutical analysis referred to the chemical analysis of drug molecules. However, over the years, modern pharmaceutical analysis has evolved beyond this to encompass combination techniques, high-throughput technologies, chemometrics, microdosing studies, miniaturization and nanotechnology. These analytical advances are now being employed in all stages of drug discovery and the focus of this review will be on how these technologies are being employed within this process (Koh, H.-L et al, 2003). It is becoming increasingly difficult to find new compounds that will lead to new drugs. The current method of identifying new drugs focuses on finding biologically active candidates. Once a potential drug is identified through research, it is patented, and research is then completed. The next step is formulating the drug. Here the focus is on obtaining a method of delivering the drug in a means most comfortable for the patient while considering cost and logistic issues. Advances in the field of substructure
analysis have expanded the applicability of sub structural analysis in multiple fronts in early lead discovery and optimization. It can be applied beyond the management of information, including compound library design and virtual screening to structure activity relationships (Villar H.O, et al, 2007). These issues include the required release rate, costs of formulating the mechanism (powder, liquid, etc), and the form of the crystal structure (if the molecule indeed crystallizes). Once the final formulation of the drug is developed, clinical trials can begin.

2.2.3 Clinical Trials on Discovered Drug

Once a drug is discovered and formulated, clinical trials begin. Dysregulation of lipid metabolism plays a crucial role in the progression of many diseases and in adverse drug response. Changes in lipid metabolism leading to disease are often not detected using biomarker analyses because these changes are defined by a subtle, long-term shift in the concentrations of many metabolites simultaneously (Watkins, S.M et al, 2004). In 1962 the Food Drug and Cosmetic Act were amended which required evidence of effectiveness before a drug could be marketed. All prescription drugs on the market today are categorized as members of existing drug classes. The stakes of drug selection within a class are especially high for chronic interventions that aim to prevent development of disease and its complications, since the evidence of efficacy or questions of safety may not be apparent for many years (MD Curt D Furberg et al, 1999). Currently, emphasis has moved to the efficiency and timeliness of the drug review process as both the public and industry demand prompt reviews and access to experimental drugs. Clinical trials in children have improved outcomes in
areas such as neonatology and HIV Trials in pediatric oncology are certainly notable for achieving high degrees of participation, yet trials in children infected with HIV have also been successful despite great obstacles (Esse N Menson et al, 2004). A pharmaceutical company would like to test on as broad a population as possible while still being able to detect any treatment effect. There are two main reasons why this may occur: the clinical response takes a long time to occur on average, or the response of main interest is difficult to observe or measure. This information is essential as a resource for development of medicines, but is also needed to satisfy licensing requirements, protect patents, promote sales, and advice patients, prescribers, and dispensers. Such information is of great commercial value, and most of it is confidential, protected by regulations about intellectual property rights (Prof Joe Collier FRCP, 2002). With the development of regulatory guidelines and experience with cholinergic drugs in recent years there has been considerable progress in this field, especially with respect to the testing of symptomatic agents. As future research turns to examine compounds which may alter the progression of the disease, new methods need to be developed to allow this question to be studied. As well as washout designs using clinical instruments, there is the need to develop practical biological markers of the disease, such as magnetic resonance imaging scans and biochemical markers in the cerebrospinal fluid or plasma. Further considerations are also underway to design studies which could demonstrate a preventive effect of a compound in Alzheimer's disease, an important future treatment strategy (Gray, J.A., 1999). Incorporating biomarkers into clinical drug trials for Alzheimer's disease (AD)
could 1) increase the homogeneity of patients through improved diagnosis, 2) establish surrogate outcome measures for drug efficacy, 3) test pharmacogenetic bases of drug response, and 4) verify proposed mechanisms of drug action (Growdon J.H, 2001). The pathophysiologic process leading to neurodegeneration in AD is thought to begin long before clinical symptoms develop. Existing therapeutics for AD improve symptoms, but increasing efforts are being directed toward the development of therapies to impede the pathologic progression of the disease. Although these medications must ultimately demonstrate efficacy in slowing clinical decline, there is a critical need for biomarkers that will indicate whether a candidate disease-modifying therapeutic agent is actually altering the underlying degenerative process. A number of in vivo neuroimaging techniques, which can reliably and noninvasively assess aspects of neuroanatomy, chemistry, physiology, and pathology, hold promise as biomarkers (Dickerson, B.C, Sperling R.A et al, 2005).

2.3 Future Technologies for Medical Applications

Information technology has the potential to revolutionise the way medicine is learned by students and healthcare professionals. This potential was recognised by the General Medical Council in their 1993 report Tomorrow's doctors in which the need for future generations of doctors to be familiar with the application and scope of information technology is described. (Mooney G.A., Bligh J.G., 1997). The medical impact of new languages and technologies in wireless Internet, broadband/high speed Internet and virtual reality Internet applications; and analyses their implications for the
relationship between patients and physicians in the 21st century. The progressive increase in the speed of transmission of data on the Internet will allow an audiovisual explosion, which will greatly benefit health professionals. Multimedia is extremely important in medicine and, therefore, the Internet will be promoted as an indispensable tool in continuing medical education, telemedicine and healthcare management (Pareras L.G., 2002). During the last decades laser technology has continuously developed. New types of lasers as ultra-short pulsed lasers in the femtosecond regime entered medical applications in ophthalmology. Diode lasers became more powerful and smaller with a broader range in wavelengths. In future new sources will also be used in medicine, fibre lasers, LEDs and organic LEDs (OLED). (Steiner R et al, 2006) Risks associated with modern drugs expand and develop beyond our existing control and perception mechanisms. New medicines have the ability to transform side effects from the traditional individual physical level to a societal level in the form of economic, political and ethical consequences. Existing assessment methods cannot capture or cope with the side effects we will experience with tomorrow's drugs. We interpret this as a sign of the need for on-going methodological developments within Medical Technology Assessment (MTA) to capture and include the as yet 'unknown' and 'unforeseen' economic, political and ethical risks of modern drugs (Møldrup C et al, 2003). Handheld computers, or personal digital assistants (PDAs), have been used to assist clinicians in medical nutrition since the early 1980s. The term PDA was originally applied to programmable calculators; over time, the capabilities of these devices were expanded to allow for the use of more complicated programs such as
databases, spreadsheets, and electronic books. Slowly, the device evolved into what is more commonly thought of as a PDA, that is, a device such as a PalmOS (PalmSource, Inc, Tokyo, Japan) or PocketPC (Microsoft, Redmond, WA) unit. We present a review of the literature about the use of PDAs in medical nutrition, followed by a discussion of the different types of PDAs and mobile technologies that are commercially available. (Holubar S, Harvey-Banchik L. et al, 2007). Genomic/bioinformatic methods are used to identify genomic polymorphisms characteristic of particular patient response profiles. They could be applied for individual administration in optimizing the doses and development of new therapies. The possibility of using information technology makes the drug discovery and development no longer a labour-intensive, trial-and-error process. Bioinformatics builds the foundations of future medical progress based on the intelligent use of structural databases to understand and predict functional relations, as well as to create modern tools for diagnosis and treatment (Sarafian V, 2006)

Through Silicon and Biology Junction

Microchips, constructed with a variety of microfabrication technologies (photolithography, micropatterning, microjet printing, light-directed chemical synthesis, laser stereo-chemical etching, and microcontact printing) are being applied to molecular biology. The new microchip-based analytical devices promise to solve the analytical problems faced by many molecular biologists (eg, contamination, low throughput, and high cost). They may revolutionize molecular biology and its application in clinical medicine, forensic science, and environmental monitoring
(Cheng J et al, 1996). In conventional medicine, diagnosis remains mostly the art of neglecting remote dangers in favor of likelier ones. Diagnostic tools are often too expensive or too inaccurate to be deployed widely. But in the near future, diagnostic gene chips will rely not on spying crude symptoms but detecting the underlying molecular processes that trigger disease in the weeks, months, or years before the patient feels a twinge. 'Molecular Electronics' is a state-of-the-art technology which is emerging as one of the augmentations to today's silicon technology. Specifically it will attempt to meet the same dynamics of current-day computing. However, these devices will consist of proteins and other large molecules rather than silicon circuits. The theory of a 'molecular' electronic computing device differs from that of digital computing in that the fields of biology, chemistry, computer science and physics will come together to engineer or create new architectures for 21st century computing devices (Saia Paul, 1998). Diagnostic tools are often too expensive or too inaccurate to be deployed widely. But in the near future, diagnostic gene chips will rely not on spying crude symptoms but detecting the underlying molecular processes that trigger disease in the weeks, months, or years before the patient feels a twinge. DNA chips are elegantly simple in concept: thin wafers of glass or plastic embedded with strips of DNA rather than, like silicon chips, tiny transistors. DNA chips are small, solid supports such as microscope slides onto which thousands of cDNAs or oligonucleotides are arrayed, representing known genes or simply EST clones, or covering the entire sequence of a gene with all its possible mutations. Fluorescently labeled DNA or RNA extracted from tissues is hybridized to the array. Laser scanning
of the chip permits quantitative evaluation of each individual complementary sequence present in the sample. DNA chip technology is currently being proposed for qualitative and quantitative applications, firstly for the detection of point mutations, small deletions and insertions in genes involved in human diseases or affected during cancer progression; secondly, to determine on a genome-wide basis the pattern of gene expression in tumors, as well as in a number of experimental situations. The extraordinary power of DNA chips will have a strong impact on medicine in the near future, both in the molecular characterization of tumors and genetic diseases and in drug discovery and evaluation (De Benedetti, et al, 2000). DNA chips are miniaturized « dot-blot » system on which thousands of DNA fragments (PCR products, oligos) spotted onto cm² Nylon membrane or glass slide support are hybridized with complementary DNA that have been previously labeled with radioactive $^{33}$P or with specific fluorophores. The hybridization is then revealed using dedicated apparatus (phosphoimager for radioactive imaging; CDD camera or Scanner Laser for fluorescence) and recorded for further image analysis and data treatment. The strength of DNA chips technology rests on the capacity to provide parallel information's at the level of a whole genome or to identify mutations and polymorphisms within large DNA sequences (le Berre-Anton V, François J et al, 2001). DNA chips are miniaturized Microsystems based on the ability of DNA to spontaneously find and bind its complementary sequence in a highly specific and reversible manner, known as hybridization. Labeled DNA molecules in a sample are analyzed by DNA probes tethered at distinct sites on a solid support. The composition
of the DNA sample is then deduced by analyzing the signal generated by labels present at each probe site. Applications are widespread: fundamental research, cancer or microbiology diagnostics, genotyping, gene expression, pharmacogenomics, and environmental control. Medical application consists, for example, in the identification and detection of mutations in genes responsible for cancers, or DNA chip analysis of individual polymorphisms which may provide a guide towards the most efficient treatment (Cuzin M, 2001). They exploit the natural tendency of double-stranded DNA molecules to bind with their complementary partners, in a process called hybridization. Once researchers have identified a particular strip of DNA within a virus or bacteria or genetic disease, that strip can be used to track down a matching strand from a patient's blood sample or biopsy specimen. Dozens, even hundreds of potentially offending pathogens, genetic diseases, or other ailments can be diagnosed on the surface of a single chip.

Doctors are currently treating most cancers in a uniform fashion. For example, everyone with pancreatic cancer or liver cancer ends up getting similar treatments, even though genetic markers are beginning to tell us that these cancers really come in a wide variety, each requiring its own unique approach. This is perhaps one reason why most cancer treatments usually leave a significant number of people behind, often with more than half (and in the case of pancreatic cancer, 90 percent) failing standard treatments. The recent development of advanced analytical and bio separation methodologies based on micro arrays and biosensor is one of the strategic objectives of the so-called post-genomic. In this field, the development of micro fabricated
devices could bring new opportunities in several application fields, such as predictive oncology, diagnostics and anti-tumor drug research. The so called "Laboratory-on-a-chip technology", involving miniaturization of analytical procedures, is expected to enable highly complex laboratory testing to move from the central laboratory into non-laboratory settings. The main advantages of Lab-on-a-chip devices are integration of multiple steps of different analytical procedures, large variety of applications, sub-micro liter consumption of reagents and samples, and portability (Gambari R et al, 2003)

The merger of medicine and microchip is in one sense only natural. Efficient drug delivery remains an important challenge in medicine: continuous release of therapeutic agents over extended time periods in accordance with a predetermined temporal profile; local delivery at a constant rate to the tumor microenvironment to overcome much of the systemic toxicity and to improve antitumour efficacy; improved ease of administration, and increasing patient compliance required are some of the unmet needs of the present drug delivery technology (Sharma S et al, 2006). These new technologies are being adopted throughout the drug industry, but they are being most effectively implemented in the R&D infrastructure of some of the bioinformatics companies for development of embracing new tools.

2.4 Information Technology in Biology

Biology (from Greek: βίος, bio, "life"; and λόγος, logos, "speech" lit. "to talk about life"), also referred to as the biological sciences, is the scientific study of life.
Biology examines the structure, function, growth, origin, evolution, and distribution of living things (http://en.wikipedia.org/wiki/Biology).

2.5 Information Technology + Biology = Biocomputing

James Watson and Francis Crick described a simple arrangement of simple components following simple rules that captured a mechanism, a biological computer that can generate incredible diversity and complexity. A biological systems having an enormous capability as control systems for agility or regulation, for pattern recognition, information storage, sensor fusion is of great interest to computer scientists. The ability to manipulate systems on the molecular scale naturally leads to speculation about the rational design of molecular-scale machines. Cells might be the ultimate molecular-scale machines and our ability to engineer them is relatively advanced when compared with our ability to control the synthesis and direct the assembly of man-made materials. Indeed, engineered whole cells deployed in biosensors can be considered one of the practical successes of molecular-scale devices. However, these devices explore only a small portion of cellular functionality (Simpson, M.L., et al, 2001). The idea in exploring the multidimensional characteristics of biocomputing systems involves multilevel protein logic interaction with hybrid digital systems (bio/digital/analog systems). In the protein biocomputing world, all models are multilevel and nonlinear exponential in nature. These biocomputing systems are linearized by boundary limited frequency, time response and dimensional ranges. The protein model and its correct development is probably the most difficult aspect for biocomputing issues (Ewing, R. L, et al, 2005).

Research at the interface of biology and information technology may lead to important new information systems (algorithms and software) and computer technologies; biology performs at levels many orders of magnitude better than silicon-
based systems. Bimolecular computing is an emerging field at the interface of computer science, biological science and engineering. It uses DNA and other biological materials as the building blocks for construction of living computational machines to solve difficult combinatorial problems (Fu, P, et al, 2007).

**Enable Biology through Information Technology**

The life science research of the 21st Century – is rich, and will become exceptionally rich, in diverse, complex data, while focusing on a system-level of understanding, enabling predictive capacity. The language for understanding biology at a systems level will be bioinformatics/IT as calculus/math has been the language for understanding the physical sciences

**2.6 Bioinformatics**

Bioinformatics is the application of computer technology to the management of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development. Research and development in bioinformatics and computational biology require the cooperation of specialists from the fields of biology, computer science, mathematics, statistics, physics, and such related sciences. It is the comprehensive application of mathematics (e.g., probability and graph theory), statistics, science (e.g., biochemistry), and computer science (e.g., computer algorithms and machine learning) to the understanding of living
systems Bioinformatics can be broadly defined as the creation and development of advanced information and computational techniques for problems in biology. More narrowly, bioinformatics is the set of computing techniques used to manage and extract useful information from the DNA/RNA/protein sequence data being generated (at high volumes) by automated techniques (e.g. DNA sequencers, DNA micro arrays) and stored in large public databases (e.g. GenBank, Protein DataBank). Certain method for analyzing genetic/protein data has been found to be extremely computationally intensive, providing motivation for the use of powerful computers (Abd-Elsalam, K.A, 2003)

What is Bioinformatics?

Bioinformatics is the field that studies the applications of computers to molecular biology and biochemistry. It encompasses:

- the information processing needs of biological data;
- the acquisition of knowledge from the data;
- the mathematical modeling and computer modeling of phenomena in a cell;
- and
- Visualization of models and information.

Bioinformatics is the driver for many of the discoveries from genomics. Researchers in genomics need collaborators with a broad range of expertise in computer systems, software, and bioinformatics.
As shown in the diagram 2.1 when we look bioinformatics at abstract level it mainly contains 3 layers.

**Figure 2.1: Abstract view of Bioinformatics**

### Purpose of Data Manager

1. Store large data sets of many different kinds
2. Access them efficiently and cross-reference them
3. Data integrity and secured Databases. Everything can be stored in a relational database or flat file or XML document. A modeling notation tailored to data type and tailored to scientists Intuitive ways to query the data Support for
efficient answering of queries query optimization indexes compact physical
storage

Purpose of Task Manager

1. Define a dataset, computation on a dataset
2. Describe combinations of computations and computational resources
3. Examine intermediate results

Purpose of Management Layer

Browsing, querying more important than transactions Modeling of user’s
interests, preferences, data quality, quantifying uncertainty, need confidence values
and alternative results (by priority/confidence) databases

2.6.1 Bioinformatics Data Mining

Introduction:

Knowledge/information can be seen as the patterns or characteristics of the
data. It is much more valuable than data. Thus, a new technology field has emerged in
the mid 1990's to deal with the discovery of knowledge/information from data. It is
called knowledge discovery in databases (KDD) or simply data mining (DM) (Chen et
al., 1996) Diabetes mellitus is a major health problem for every corner of the world for
many reasons of life styles. There is a long history of diabetic registries and databases
with systematically collected patient information. We examine one such diabetic data
warehouse, showing a method of applying data mining techniques, and some of the
data issues, analysis problems, and results. The diabetic data warehouse is from a large
integrated health care system in the New Orleans area with 30,383 diabetic patients. Methods for translating a complex relational database with time series and sequencing information to a flat file suitable for data mining are challenging (Breault J.L et al, 2002).

Data mining is ready to deal with this challenge because recent developments in data mining have shown an increasing interest on mining of complex data (as exemplified by graph mining, text mining, etc.). By incorporating the relationships of the data along with the data itself (rather than focusing on the data alone), complex data injects semantics into the mining process, thus enhancing the potential of making better contribution to knowledge economy (Chen, Z, et al, 2007). Database technology manages domain data. Machine learning, learns patterns from data by using its intelligence, and statistics finds statistical parameters of the data.

From the day Human Genome Project has become a reality huge amounts of data are being produced and collected. The biologist needs information technology to help manage and analyze such large and complex data sets. To access the data, online data banks for data storage are available. Most of the data collected have been put on the World Wide Web and can be shared and accessed online. For example, GenBank (http://www.ncbi.nlm.nih.gov/Genbank/), EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/), the mystery of life hidden in the biological data might be decoded much faster and more accurately with the data mining technologies with the help of bioinformatics.
Goal of data mining should be, new, mined information needs to be carefully verified and correct; it should be meaningful and can be easily understood. And applicable to certain problem domain what search is meant for. To meet the goal of data mining, data mining uses different technologies; three of them are databases, machine learning, and statistics. As shown in the diagram figure2.2 it contains many iterative steps.

Figure 2.2: Data mining process model
2.6.2 The Data mining process

Data mining is an iterative process that typically involves the following phases:

Stair Case for Mining

Step 1: Problem Identification

Step 1.1: Starts with the understanding of the business problem.

Step 1.2: Data mining experts, business experts, and domain experts work closely together to define the project objectives from a business perspective.

Step 1.3: The project objective is then translated into data mining problem identification.

Step 2: Data exploration

Step 2.1: Domain experts collect, describe, and explore the data.

Step 2.2: They identify quality problems of the data.

Step 2.3: Domain experts understand the meaning of the metadata.

Step 3: Data preprocessing

Step 3.1: Domain experts collect, cleanse, and format the data because some of the mining functions accept data only in a certain format

Step 3.2: Build the data model for the modeling process

Step 4: Modeling

Data mining experts select and apply various mining functions and use different mining functions for the same type of data mining problem. Some of
the mining functions require specific data types. The modeling phase and the evaluation phase are coupled.

Step 5: Information interpretation

Data mining experts evaluate the model. If the model does not satisfy their requirements, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved and finally satisfied.

Step 6: Visualization

The mining results are exported into database tables or into other applications, for example, spreadsheets.

It should be noticed that data mining is costly. Therefore, besides going through the above mentioned steps with care, an estimate of the data mining project in advance is necessary. Although certain degree of uncertainty is always involved in a data mining project, with the cooperation of the data mining expert and the domain expert, better understanding of the problem with a clearer objective can be achieved. This increases the possibility of success of the data mining project.

2.7 Data Mining Techniques

Many concepts and techniques are useful for the same goal of mining hidden information from data. Among them algorithms, databases, statistics, machine learning, and information retrieval are essential. Algorithms and parallel processing techniques are mainly to accelerate the mining process.
2.7.1 Bioinformatics and Machine Learning

Bioinformatics applies data mining, i.e., modern computer-based statistics, to biomedical data. It leverages on machine learning approaches, such as artificial neural networks, decision trees and clustering algorithms, and is ideally suited for handling huge data amounts (Wiemer J.C, Prokudin, A et al, 2004). Machine learning is a long-developed field in artificial intelligence (AI). Machine learning and discovery for bioinformatics copes with various kinds of problems raised by researchers with the help of AI technologies (Miyano, Satoru, Shinohara, Ayumi, 1995). It focuses on automatic learning from a data set. A suitable model with many parameters is built first for a certain domain problem and an error measure is defined. A learning (training) procedure is then used to adjust the parameters according to the predefined error measure. The purpose is to fit the data into the model. There are different theories for the learning procedure, including gradient decent, expectation maximization (EM) algorithms, simulated annealing, and evolutionary algorithms. The learning procedure is repeated until the error measure reaches zero or is minimized. After the learning procedure is completed with the training data, the parameters are set and kept unchanged and the model can be used to predict or classify new data samples. Machine learning is used in a large number of bioinformatics applications and studies. The application of machine learning techniques in other areas such as pattern recognition has resulted in accumulated experience as to correct and principled approaches for their use. The aim of this paper is to give an account of issues affecting the application of machine learning tools, focusing primarily on general aspects of
Different learning schemes have been developed and discussed in the machine learning literature. Important issues include the learning speed, the guarantee of convergence, and how the data can be learned incrementally. There are two categories of learning schemes: (1) supervised learning and (2) unsupervised learning. Supervised learning learns the data with an answer. Meaning, the parameters are modified according to the difference of the real output and the desired output (the expected answer). The classification problem falls into this category. On the other hand, unsupervised learning learns without any knowledge of the outcome. Clustering belongs to this category. It finds data with similar attributes and put them in the same cluster.

Various models like neural networks (NN), decision trees (DT), genetic algorithms (GA), fuzzy systems, and support vector machines (SVM) have proved very useful in classification and clustering problems. But machine learning techniques usually handles relatively small data sets because the learning procedure is normally very time-consuming. To apply the techniques to data mining tasks, the problem with handling large data sets must be overcome.

2.8 Data Mining for Bioinformatics

Certain bioinformatics problems have solutions via data mining technologies
2.8.1 Protein Structure Prediction

With so many known proteins, the functions of most proteins are still waiting for further investigation. A protein's function depends upon its structure. If the structure of a protein is known, it would be easier for the biologist to infer the function of the protein. The prediction of protein structure and the precise understanding of protein folding and unfolding processes remains one of the greatest challenges in structural biology and bioinformatics. Computer simulations based on molecular dynamics (MD) are at the forefront of the effort to gain a deeper understanding of these complex processes. Currently, these MD simulations are usually on the order of tens of nanoseconds, generate a large amount of conformational data and are computationally expensive (Berrar D et al, 2005).

Many sequences, and in some cases structures, of proteins that induce an allergic response in a topic individuals have been determined in recent years. This data indicates that allergens, regardless of source, fall into discreet protein families. Similarities in the sequence may explain clinically observed cross-reactivates between different biological triggers (Ivanciuc O et al, 2002). Disordered regions in proteins are relatively frequent and important for our understanding of molecular recognition and assembly, and protein structure and function. From an algorithmic standpoint, flagging large disordered regions is also important for ab initio protein structure prediction methods. Here we first extract a curedt, non-redundant, data set of protein
disordered regions from the Protein Data Bank and compute relevant statistics on the length and location of these regions (Cheng J et al, 2005).

It is desirable that a protein's structure can be decided from its sequence through computational methods. A protein sequence is called the primary structure of the protein. Hydrogen bonding of the molecules results in certain substructures called the secondary structure. Interactions between secondary structures assemble them into the tertiary structure. If a protein is composed by more than one unit, it is a quaternary structure. Although an ordered 3D structure is generally considered to be a necessary precondition for protein functionality, there are disordered counter examples found to have biological activity. The objectives of our data mining project are: (1) to generalize from the limited set of counter examples and then apply this knowledge to large data bases of amino acid sequence in order to estimate commonness of disordered protein regions in nature, and (2) to determine whether there are different types of protein disorder. For general disorder estimation, a neural network based predictor was designed and tested on data built from several public domain data banks through a nontrivial search, statistical analysis and data dimensionality reduction (Romero P et al, 2000).

2.8.2 Finding Gene expression

Gene expression is the process by which the inheritable information in a gene, such as the DNA sequence, is made into a functional gene product, such as protein or RNA (http://en.wikipedia.org/wiki/Gene_expression). These expressions are stored in
central databases. Gene expression databases contain a wealth of information, but current data mining tools are limited in their speed and effectiveness in extracting meaningful biological knowledge from them. Online analytical processing (OLAP) can be used as a supplement to cluster analysis for fast and effective data mining of gene expression databases (Alkharouf N.W et al, 2005). Although many genetically inherited diseases currently recorded in different databases like LocusLink, OMIM are already linked to a region of the human genome, about 450 have no known associated gene and Gene expression. In most hereditary syndromes like cancer, finding a relation between various genetic mutations within a gene (genotype) and a patient's clinical cancer history (phenotype) is challenging; these challenges can be addressed with the help of Data mining techniques.

2.8.3 Genetic Analysis

The simulated sequence data for the Genetic Analysis Workshop 12 were analyzed using data mining techniques provided by SAS ENTERPRISE MINER™ Release 4.0 in addition to traditional statistical tests for linkage and association of genetic markers with disease status. (Czika W.A et al, 2001).

2.8.4 Primary structures analysis

In molecular biology, biological macromolecules, like desoxyribonucleic acids (DNA) and proteins are coded by strings, called 'primary structures'. For a long time, biologists gathered these primary structures in large databases. Now, they focus on
analyzing these primary structures in order to extract useful knowledge. Data mining approaches can be helpful to reach this goal. (Maddouri M, Elloumi, M et al, 2002)

2.8.5 DNA sequence Analysis

DNA sequence is one of the basic and important data among biological data. Researching DNA sequence data and then comprehending life essential is a necessary task in post-genomic era. At present, data mining technique is one of the most efficient data analysis means, which finds out information hidden in data. It has also become main data analysis technique adopted in Bioinformatics. It has been applied in DNA sequence analysis, which has got wide attention and rapid development. And considerable research achievements have emerged that provides an overview of research progress in DNA sequence data mining field. In more detail, it proposes three research phases including statistics-based data mining methods application, general data mining methods application, and specialized DNA sequence-oriented data mining methods design, and then elaborates that sequence similarity is foundation of DNA sequence data mining technique. It also analyzes and comments some key techniques in this field by combining with biological background, such as DNA sequential pattern, association, clustering, classification and outlier mining. (Zhu, Y.-Y., Xiong, Y et al, 2007)
2.9 Dynamic Programming

The Smith-Waterman algorithm is a dynamic programming method for determining similarity between nucleotide or protein sequences. The algorithm was first proposed in 1981 by Smith and Waterman and is identifying homologous regions between sequences by searching for optimal local alignments. To find the optimal local alignment, a scoring system including a set of specified gap penalties is used (Smith and Waterman et al, 1981). Dynamic programming algorithms solve optimization problems, problems in which there are a large number of possible solutions, but only one (or a small number of) best solutions are given. Dynamic programming methods are a general class of algorithms that are often seen both in sequence alignment and other computational problems. Dynamic programming has provided a powerful approach to optimization problems, but its applicability has been somewhat limited because of the large computational requirements of the standard computational algorithm (Francelin, Roseli Ap et al, 1993). A dynamic programming algorithm finds the best solution by first breaking the original problem into smaller sub problems and then solving. These pieces of the larger problem have a sequential dependency; that is, the fourth piece can be solved only with the answer to the third piece, the third can be solved only with the answer to the second, and so on.
2.9.1 Applying Dynamic Programming on Genetic Sequence

A DNA sequence or genetic sequence is a succession of letters representing the primary structure of a real or hypothetical DNA molecule or strand, with the capacity to carry information. (http://en.wikipedia.org/wiki/DNA_sequence)

![Sample Electropherogram printout from automated sequencer showing part of a DNA sequence](image)

To compare the nucleotides or amino acids that appear at corresponding positions in two or more sequences, we must first assign those correspondences. Sequence alignment is the identification of residue-residue correspondences. It is the basic tool of bioinformatics.

Any assignment of correspondences that preserves the order of the residues within the sequences is an alignment. Gaps may be introduced.

Given two text strings:

First string = FTFTALELLAVAV

Second string = FTALLLAAV
Global match: Align all of one sequence with all of the other. Bioinformatic searches are designed to detect global sequence similarity and short contiguous amino acid sequence identity (Silvanovich A, et al, 2007)

Local match: Find a region in one sequence that matches a region of the other. The Smith-Waterman algorithm for local sequence alignment is one of the most important techniques in computational molecular biology (Arslan A.N., et al, 2001)

Global FTFTALILLAVAV
F—TAL-LLA-AV

Local FTFTALILL-AVAV
—FTAL-LLAAV—

**Figure 2.4: This illustrates mismatches, insertions and deletions**

The technique of dynamic programming can be applied to produce global alignments via the Needleman-Wunsch algorithm, and local alignments via the Smith-Waterman algorithm. Dynamic programming works by first solving all these sub problems, storing each intermediate solution in a table along with a score, and finally choosing the sequence of solutions that yields the highest score. Dynamic programming is one of the major problem solving methodologies in a number of disciplines such as operations research and computer science. It is also a very important and powerful tool of thought (Sniedovich, M., Lew, A., et al, 2006) The goal of the dynamic programming algorithm is to maximize the total score for the alignment. This problem is broken down into sub problems of aligning each residue
from one sequence with each residue from the other. The solution is a decision as to whether the residues should be aligned with each other, a gap should be introduced in the first sequence, or a gap should be introduced in the second sequence. Each high-scoring choice rules out the other two low-scoring possibilities, so that if information about the accumulated scores is stored at each step.

The types of analysis that can be done with sequence data are:

- Knowledge-based single sequence analysis for sequence characteristics
- Pairwise sequence comparison and sequence-based searching
- Multiple sequence alignment
- Sequence motif discovery in multiple alignments
- Phylogenetic inference

2.10 Bioinformatics role in Sequence Analysis

2.10.1 Sequence analysis

1. Since the sequenced in 1977, the DNA sequences of hundreds of organisms have been decoded and stored in databases, the development of powerful techniques for DNA sequencing has enabled sequencing of large amounts of gene fragments and even complete genomes. Important new techniques for physical mapping, DNA sequencing and sequence analysis have been developed (Sterky F, Lundeberg J et al, 2000). Sequence analysis in
bioinformatics is an automated, computer-based examination of characteristical fragments

2.10.2 Analysis of gene expression

Gene expression is the process by which the inheritable information in a gene, such as the DNA sequence, is made into a functional gene product, such as protein or RNA (http://en.wikipedia.org/wiki/Gene_expression). This expression yields many results for eg. Patterns of gene expression across tissues revealed a precise distinction between normal and tumor samples, and revealed a striking group of about 400 genes that were overexpressed in tumor tissue (Welsh JB et al, 2001). Expression arrays facilitate the monitoring of changes in the expression patterns of large collections of genes. The analysis of expression array data has become a computationally-intensive task that requires the development of bioinformatics technology for a number of key stages in the process, such as image analysis, database storage, gene clustering and information extraction (Tamames, J et al, 2002).

2.10.3 Analysis of protein expression

Protein expression is a subcomponent of gene expression. It consists of the stages after DNA has been translated into amino acid chains, which are ultimately folded into proteins. Protein expression is commonly used by proteomics researchers to denote the measurement of the presence and abundance of one or more proteins in a particular cell or tissue. (http://en.wikipedia.org/wiki/Protein_expression) To identify the putative functions and structural features of these new genes, integrated genomic
approaches of data mining, expression profiling, and bioinformatic predictions were used. (Breton G et al, 2003) New tools for the biologists to compare their datasets with others, as well as examples of future bioinformatic tools that can be used for developing gene networks and pathways for a given set of data (Harris, S.E., Harris, M.A et al, 2002)

2.11 Bioinformatics role in Sequence Alignment

One of the most central methods in bioinformatics is the alignment of two protein or DNA sequences. However, so far large-scale benchmarks examining the quality of these alignments are scarce. On the other hand, recently several large-scale studies of the capacity of different methods to identify related sequences has led to new insights about the performance of fold recognition methods (Elofsson A et al, 2002). Sequence alignment is the identification of residue-residue correspondences. Where the residues in one position are deemed to have a common evolutionary origin. If the same letter occurs in both sequences then this position has been conserved in evolution. If the letters differ it is assumed that the two derive from an ancestral letter

2.11.1 Global Alignment

This alignment attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size A general global alignment technique is called the Needleman-Wunsch algorithm and is based on dynamic programming.
2.11.2 Local Alignment

These alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The Smith-Waterman algorithm is a general local alignment.

2.11.3 Global vs. Local Alignment

All the routines for the local alignment are exactly the same as the routines for the global alignment except that during the construction of matrix D the alignment is restarted each time the score becomes higher than a cutoff. The second difference is that the backtracking starts from the lowest element in the matrix, wherever it is.

Pairwise alignment

Pairwise sequence alignment methods are used to find the best-matching piecewise (local) or global alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time. Pairwise sequence alignments aim to decide whether two sequences are related and, if so, to exhibit their related domains. Recent works have pointed out that a significant number of true homologous sequences are missed when using classical comparison algorithms. This is the case when two homologous sequences share several little blocks of homology, too small to lead to a significant score (Nédélec E et al, 2005). The characteristics of pairwise alignment, critical to bioinformatics applications, make it uniquely suitable for the Cell processor. We have implemented two different strategies for pairwise alignment on Cell. We analyze the bottlenecks for each strategy, a diagonal approach which uses...
the vector processing capabilities of the SPE and a row wise approach in which cells are dependant and has expensive branches. (Sachdeva V, Kistler, M, Speight, E et al, 2006)

2.11.4 Tools for Global Alignment

There are different tools for global alignment out which few of them are very frequently used for better results and performance

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Sequence Type</th>
<th>Author</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNADot</td>
<td>Web-based dot-plot tool</td>
<td>Nucleotide</td>
<td>R. Bowen</td>
<td>1998</td>
</tr>
<tr>
<td>DOTLET</td>
<td>Java-based dot-plot tool</td>
<td>Both</td>
<td>M. Pagni and T. Junier</td>
<td>1998</td>
</tr>
<tr>
<td>MCALIGN2</td>
<td>explicit models of indel evolution</td>
<td>DNA</td>
<td>J. Wang et al.</td>
<td>2006</td>
</tr>
<tr>
<td>MUMmer</td>
<td>Suffix-Tree based</td>
<td>Nucleotide</td>
<td>S. Kurtz et al.</td>
<td>2004</td>
</tr>
<tr>
<td>Needle</td>
<td>Needleman-Wunsch dynamic programming</td>
<td>Both</td>
<td>A. Bleasby</td>
<td>1999</td>
</tr>
<tr>
<td>Ngila</td>
<td>logarithmic and affine gap costs and explicit models of indel evolution</td>
<td>Both</td>
<td>R. Cartwright</td>
<td>2007</td>
</tr>
<tr>
<td>ProbA (also propA)</td>
<td>Stochastic partition function sampling via dynamic programming</td>
<td>Both</td>
<td>U. Mückstein</td>
<td>2002</td>
</tr>
<tr>
<td>stretcher</td>
<td>Memory-optimized but slow dynamic programming</td>
<td>Both</td>
<td>I. Longden (modified from G. Myers and W. Miller)</td>
<td>1999</td>
</tr>
</tbody>
</table>

*Table: 2.1 Table showing the Tools for Global Alignment*

(http://en.wikipedia.org/wiki/Sequence_alignment_software)
2.11.5 Tools for local alignment

There are different tools for local alignment out which few of them are very frequently used for better results and performance.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Sequence Type</th>
<th>Author</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASTZ</td>
<td>Seeded pattern-matching</td>
<td>Nucleotide</td>
<td>Schwartz et al.</td>
<td>2003</td>
</tr>
<tr>
<td>JAligner</td>
<td>Open source Java implementation of Smith-Waterman</td>
<td>Both</td>
<td>A. Moustafa</td>
<td>2005</td>
</tr>
<tr>
<td>Matcher</td>
<td>Memory-optimized needleman but slow dynamic programming (based on LALIGN)</td>
<td>Both</td>
<td>I. Longden (modified from W. Pearson)</td>
<td>1999</td>
</tr>
<tr>
<td>REPuter</td>
<td>Suffix-Tree based</td>
<td>Nucleotide</td>
<td>S. Kurtz et al.</td>
<td>2001</td>
</tr>
<tr>
<td>SIM</td>
<td>Local similarity</td>
<td>Both</td>
<td>X. Huang and W. Miller</td>
<td>1991</td>
</tr>
<tr>
<td>SSEARCH</td>
<td>Local (Smith-Waterman) alignment with statistics</td>
<td>Protein</td>
<td>W. Pearson</td>
<td>1981 (Algorithm)</td>
</tr>
<tr>
<td>Water</td>
<td>Smith-Waterman dynamic programming</td>
<td>Both</td>
<td>A. Bleasby</td>
<td>1999</td>
</tr>
</tbody>
</table>

*Table: 2.2 Table showing the Tools for Local Alignment* *(http://en.wikipedia.org/wiki/Sequence_alignment_software)*
2.12 Bioinformatics Role in Evolutionary studies

Multiple sequence alignment

Multiple sequence alignment is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set. Molecular Biologists frequently compute multiple sequence alignments (MSA) to identify similar regions in protein families. Progressive alignment is a widely used approach to compute MSA. However, aligning a few hundred sequences by popular progressive alignment tools requires several hours on sequential computers. Due to the rapid growth of biological sequence databases biologists have to compute MSA in a far shorter time (Oliver T et al, 2005). Optimization method that unifies sequence and structure information. The alignment score is based on standard amino acid substitution probabilities combined with newly computed three-dimensional structure alignment probabilities (Shatsky, M et al, 2006). The Clustal series of programs are widely used in molecular biology for the multiple alignment of both nucleic acid and protein sequences and for preparing phylogenetic trees. The popularity of the programs depends on a number of factors, including not only the accuracy of the results, but also the robustness, portability and user-friendliness of the programs (Chenna R et al, 2003).

2.12.1 Phylogenetic Trees

A phylogenetic tree, also called an evolutionary tree, is a tree showing the evolutionary relationships among various biological species or other entities that are
believed to have a common ancestor. (http://en.wikipedia.org/wiki/Phylogenetic_tree)
A molecular phylogenetic tree is a tree-structured graph that represents the evolutionary process of genes, and is constructed from sequence data (such as DNA sequences) obtained from several organisms. Although molecular phylogenetic trees are fundamental data structures in evolutionary analysis, no database system is available that can match trees in the database against a user-supplied tree by comparing tree structures (Yoshikawa, Takanobu, 1999). Phylogenetic tree estimation plays a critical role in a wide variety of molecular studies, including molecular systematics, phylogenetics, and comparative genomics. Finding the optimal tree relating a set of sequences using score-based (optimality criterion) methods, such as maximum likelihood and maximum parsimony, may require all possible trees to be considered, which is not feasible even for modest numbers of sequences (Whelan S, 2007). Since evolution plays a key role in biology, it is natural to attempt to depict it using trees, leaves represent various organisms, species, or genomic sequences; an internal node P stands for an abstract organism (species, sequence)

2.12.2 Phylogenetic Analysis

The construction of phylogenetic trees is an NP-hard problem as the number of possible trees increase exponentially with the number of DNA or protein sequences included in the analysis. Most standard methods for inferring phylogenetic trees require an optimality criterion and a tree-search algorithm (Hoef-Emden, K., 2005). This tree-based representation of the relationships among species is a phylogenetic tree; it has since been adopted as a convenient schematic for depicting evolutionary
relatedness based on sequence similarity. The quantitative nature of sequence relationships has allowed the development of more rigorous methods and rules for tree drawing. While hand-drawn trees of life may branch fancifully according to what is essentially an artist's conception of evolutionary relationships, modern phylogenetic trees are strictly binary; that is, at any branch point, a parent branch splits into only two daughter branches.

Only by analysis of much larger sets of data can theories of whole-organism phylogeny be suggested. There is a variety of phylogenetic analysis software available for many operating systems. One of them is PHYLIP package, and is accessible from the PHYLIP web page (http://evolution.genetics.washington.edu/phylip.html).

2.13 Biomarkers

A Biomarker is a substance used as an indicator of a biologic state. It is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (http://en.wikipedia.org/wiki/Biomarker)

Biomarkers of disease play an important role in medicine and have begun to assume a greater role in drug discovery and development. The challenge for biomarkers is to allow earlier, more robust drug safety and efficacy measurements. Their role in drug development will continue to grow for the foreseeable future. For biomarkers to assume their rightful role, greater understanding of the mechanism of
disease progression and therapeutic intervention is needed. In addition, greater understanding of the requirements for biomarker selection and validation, biomarker assay method validation and application, and clinical endpoint validation and application is needed (Colburn, W.A., Keefe, D.L., et al., 2003). Most emphasis was placed on biomarkers of deleterious effects, since these are of greatest relevance to the subject of this review. The area of greatest activity was found to be that relating to biomarkers of mutagenic, genotoxic and carcinogenic effects. This is also one of the major areas of concern in considerations of the beneficial and deleterious effects of dietary components, and also the area in which regulatory testing requires studies of the longest duration (Bottrill, K. et al, 1998). Molecular biomarkers for breast cancer are of several types. Risk biomarkers are those associated with increased cancer risk and include mammographic abnormalities, proliferative breast disease with or without atypia, family clustering and inherited germ-line abnormalities. Surrogate endpoint biomarkers are tissue, cellular or molecular alterations that occur between cancer initiation and progression (Beenken, S.W., Bland, K.I et al, 2002). Biomarkers include a vast array of measurements that reflect exposure, effect, and/or susceptibility. The development and validation of potential biomarkers is a long-term endeavor that proceeds from basic research to pilot human studies to full-scale epidemiological investigations. The past decade has seen extensive research investigation of biomarkers and the beginnings of their practical application for risk assessment and environmental health management (Decaprio A.P et al, 1997). Biomarkers for Alzheimer's disease have many diagnostic and therapeutic uses. They can be used to
facilitate diagnosis, particularly early diagnosis and possibly presymptomatic diagnosis. In evaluating treatment aimed at modifying the course of AD, biomarkers can help to assess whether the drug is hitting a target or influencing disease mechanisms in the brain, and can help with dose-finding. In definitive clinical trials, biomarkers can be used as surrogate outcome measures (Galasko D et al, 2005). In surveillance, biomarkers can be used as indicators of hazard, exposure, disease, and population risk. However, to obtain rates for these measures, the population at risk will need to be assessed. In medical screening, biomarkers can serve as early indicators of disease in asymptomatic people. This allows for the identification of those who should receive diagnostic confirmation and early treatment (Schulte P.A et al, 2005).

2.13.1 Bioinformatics for Biomarkers

In today's information-driven culture, there is virtually no walk of life that is not impacted on by computing. As a bridging discipline in the health sciences with activities that span both basic science and clinical interests, modern pharmacology is no exception. As the plethora of data and databases spawned by the 'omics' generation expand in number and complexity, bioinformatics is necessary to manage, integrate and exploit this cohort of data so that the appropriate links to molecular pathology and therapeutic response can be made (Whittaker, P.A., et al, 2007). The DNA microarray has revolutionized cancer research. Now, scientists can obtain a genome-wide perspective of cancer gene expression. One potential application of this technology is the discovery of novel cancer biomarkers for more accurate diagnosis and prognosis,
and potentially for the earlier detection of disease or the monitoring of treatment effectiveness. Because microarray experiments generate a tremendous amount of data and because the number of laboratories generating microarray data is rapidly growing, new bioinformatics strategies that promote the maximum utilization of such data are necessary (Rhodes, D.R., Chinnaiyan, A.M., et al., 2004).

Figure 2.5: Advantages of Biomarker at different stages of treatment

Biomarkers discovery is often carried out by comparing physiological changes between normal and disease states. As shown in the figure 2.5 it has many advantages. After examining up-regulated and down-regulated genes, proteins and metabolites,
researchers will identify new biomarker associated with particular diseases that can serve as biomarkers. The development and validation of biomarkers can often follow well-established principles of scientific investigation, but implementation of biomarkers into drug development requires careful consideration of many other practical issues. During the study planning process, careful consideration must be given to the suitability of the biomarker for the specific needs and objectives of the drug development team, performance of the biomarker at the specific sites, standardization and data transmittal, potential use of the biomarker at other stages in the life cycle of the compound, and a clear understanding of how the results will be used to make decisions about the compound. A close collaboration between the biomarker expert and the drug development team is needed maximize the likelihood that the full value of the biomarker will be realized (Peterson, B.T., McCarthy, T.J., Murphy, P, et al, 2007)

2.14 Goals of Bioinformatics

Bioinformatics, in its broad sense, involves application of computer processes to solve biological problems. A wide range of computational tools are needed to effectively and efficiently process large amounts of data being generated as a result of recent technological innovations in biology and medicine. A number of computational tools have been developed or adapted to deal with the experimental riches of complex and multivariate data and transition from data collection to information or knowledge (Kapetanovic, I.M., et al 2004). The goals are as lofty as the development of rational
drugs and antimicrobial agents, development of new enhanced bacterial strains for bioremediation and pollution control, development of better and easy to administer vaccines, the development of protein biomarkers for various bacterial diseases, and better understanding of host-bacteria interaction to prevent bacterial infections (Bansal, A.K., et al., 2005). Bioinformatics, in its broad sense, involves application of computer processes to solve biological problems. A wide range of computational tools are needed to effectively and efficiently process large amounts of data being generated as a result of recent technological innovations in biology and medicine. A number of computational tools have been developed or adapted to deal with the experimental riches of complex and multivariate data and transition from data collection to information or knowledge (Kapetanovic, I.M., et al., 2004).

Bioinformatics research can be classified under three major approaches: (1) analysis based upon the available experimental wet-lab data, (2) the use of mathematical modeling to derive new information, and (3) an integrated approach that integrates search techniques with mathematical modeling (Bansal, A.K., et al., 2005).

Since protein complexes play a crucial role in biological cells, one of the major goals in bioinformatics is the elucidation of protein complexes. A general approach is to build a prediction rule based on multiple data sources, e.g., gene expression data and protein interaction data, to assess the likelihood of two proteins having complex association (Van Berlo, R.J.P et al., et al., 2007).
Application of Goals

Application of Bioinformatics in various Fields. Bioinformatics is the use of IT in biotechnology for the data storage, data warehousing and analyzing the DNA sequences.

**Bioinformatics is being used in following fields:**

- Molecular medicine
- Personalized medicine
- Preventative medicine
- Gene therapy drug development
- Microbial genome applications
- Waste cleanup
- Climate change Studies
- Alternative energy sources
- Biotechnology
- Antibiotic resistance
- Forensic analysis of microbes
- Bio-weapon creation
- Evolutionary studies
- Crop improvement
- Insect resistance
- Improve nutritional quality
- Development of Drought resistance varieties
- Veterinary Science

The developed bioinformatics techniques have potential to facilitate (i) the discovery of causes of diseases, (ii) vaccine and rational drug design, and (iii) improved cost effective agents for bioremediation by pruning out the dead ends (Bansal, A.K., et al., 2005).
References


18. Iain Chalmers; Otolaryngology - Head and Neck Surgery Better information systems are needed to help patients and clinicians integrate clinical research within everyday clinical practice, Volume 137, Issue 4, Supplement 1, October 2007, Pages S69-S71


28. Hyuk-Sang Kwon, Jae-Hyoung Cho, Hee-Soo Kim, Jin-Hee Lee, Bok-Re Song, Jung-Ah Oh, Je-Ho Han, Hee-Seung Kim, Bong-Yun Cha, Kwang-Woo Lee, Development of web-based diabetic patient management system using
29. Lee T, Yu-Ting Yeh, Chien-Tsai Liu and Ping-Ling Chen
Development and evaluation of a patient-oriented education system for
diabetes management; Int. J. Med. Informatics, Volume 76, Issue 9, September
2007, Pages 655-663

30. Finkelstein, J., Khare, R., Vora, D., Arora, M., Design and implementation of
home automated telemanagement in chronic obstructive pulmonary disease,
Proceedings of the IEEE Symposium on Computer-Based Medical Systems,

calls with nurse follow-up on diabetes treatment outcomes in a Department of
Veterans Affairs health care system: A randomized controlled trial, Diabetes

32. Piette, J.D., Weinberger, M., McPhee, S.J., The effect of automated calls with
television nurse follow-up on patient-centered outcomes of diabetes care: A

33. Lobach Df: Hammond We: Computerized Decision Support Based On A
Clinical Practice Guideline Improves Compliance With Care Standards. Am J
Med 102:89-98, 1997


45. Clark Cm Jr, Snyder Jw, Meek RI, Et Al; A Systematic Approach To Risk Stratification And Intervention Within A Managed Care Environment Improves Diabetes Outcomes And Patient Satisfaction. Diabetes Care 24:1079-1086, 01

46. Graham G, Nugent, L, Stroucse K, Information Everywhere; How The Her Transformed Care At Vha, J Ahima


66. Lapane, K.L., Dubé, C.E., Schneider, K.L., Quilliam, B.J., Misperceptions of patients vs providers regarding medication-related communication issues, American Journal of Managed Care 13 (11), pp. 613-618, 2007


