Bio-Informatics: A New Life to Computer Sciences

[1]A. KHURSHID, [2]D. C. HOESSLI, [3]M. I. CHOUDHARY,
[1]M. OVAIS, [1]A. QAZI,
[3]ATTA-UR-RAHMAN AND [1, 3]NASIR-UD-DIN*
[1]Institute of Molecular Medicine and Biomedical Sciences, The University of Lahore, Lahore, Pakistan,
[2]Department of Pathology, CMU, University of Geneva, Geneva, Switzerland
[3]HEJ Research Institute of Chemistry, University of Karachi, Karachi, Pakistan

(Received 7th November, 201, revised 27th December, 2001)

Summary: The refinement and precision of biochemical techniques coupled with enhanced computer technology, speed and capacity, have resulted in a converged system of bioinformatics providing us with means to discover and predict biological functions on molecular basis. This rapidly emerging field is in line with explosive growth of life sciences. There is an ever increasing need to manage, integrate and interrogate vast amounts of information data particularly in genomics. The new sciences of genomics, proteomics and bioinformatics are providing us with a sense of continuously emerging revolution in the area of biochemical, and biomedical sciences. The progress in new technologies involving computer sciences, information technology and biological and biochemical systems including that of microarray are likely to translate the biological data into a healthy future. The search for important patterns from vast amounts of fragmentary data of genetics sequence may give biological clues to biologically inspired models and ultimately providing systems for artificial neural networks, genetics algorithms or analog & digital optical pattern recognition.

Introduction

Bioinformatics is a field of science in which biological sciences, computer science and information technology merge into a single discipline. The goal of the field is to discover and predict new biological functions on molecular basis, and design experiments based on available state of the art technology to validate the structural and functional predictions. The major systems of methods involved in these studies are as follows:

1. Development of new algorithms and statistics where we are able to assess the relationships among members of large data sets.
2. Analysis and interpretation of various types of data including nucleotides and amino acid sequences, protein domains, and protein structures.
3. Development and implementation of tools that enable efficient access and management of different types of information.

Bioinformatics is in place since 1980. It originated as a field of science and engineering in the mid 1970s. It has also been termed biocomputing and its development was considerably accelerated by the need for biosequence data analysis. It is allied with the older field of computational biology, and these terms are often used to cover a range of related concepts. These share many similarities with fields of computational chemistry and cheminformatics, and medical informatics, all of which combine principles of informatics with basic sciences. Concepts central to bioinformatics revolve around solving needs for scientific information analysis and management for biosciences, focused on molecular biology and genetics data based on biochemistry.

Data refers to facts concerning people, objects, or events and information is data processed and presented in a form suitable for human interpretation. Databases growth in biological & biochemical fields is exponential, and with the help of bioinformatics representative protein sequences databases are growing rapidly with high information contents [1]. Databases are organized to contain protein sequences and structural data, to integrate

*To whom all correspondence should be addressed.

existing genomics data sources into one database, to create databases to support high-throughput production of genome sequencing, and to automatically construct models for interpreting micro-array results

In the future practice of life sciences, bioinformatics may be defined as a pluridisciplinary approach of computer science, information technology and genetics applied at determining and analyzing massive genetic information, thus providing a comprehensive mean to analyze protein expression, three dimensional structure and function leading to multiple benefits.

Background

Initially, bioinformatic databases were constructed a few years after the first protein sequences became available. The first protein sequence reported was that of bovine insulin in 1956, consisting of 51 residues. Nearly a decade later, the first nucleic acid sequence was reported, that of yeast alanine tRNA with 77 bases. Just a year later, all the available biosequence data were pooled to create the first bioinformatic database. The protein data bank followed in 1972 with a collection of ten X-ray crystallographic protein structures, and the SWISS-PROT protein sequence database bank began in 1987.

Variety of divergent data resources of different types and sizes are now available either in the public domain or more recently from commercial third parties. All of the original databases were organized in a very simple way with data entries being stored in flat files, either one per entry, or as a single large text file. Later on, lookup indexes were added to allow convenient keyword searching of header information.

Following the formation of the databases, tools became available to search sequence databases — at first in a very simple way, looking for keyword matches and short sequence words, and then more sophisticated pattern matching and alignment-based methods. New algorithms like BLAST [2] and FASTA [3] algorithms are now the mainstay of sequence database searching.

Bioinformatics Rationale

Bioinformatics has profound utility as it provides with means to utilize cataloging and retrieval of incoming vast biological information, "mainly from human genome project" data on computers, offering a more global perspective in experimental design, and data-mining. As we move from the "one scientist-one gene/protein/disease" paradigm of the past to a consideration of whole organisms, we gain opportunities for new insights into health and disease, and the process by which testable hypotheses are generated regarding the function or structure of a gene or protein of interest by identifying similar sequences in better characterized organisms are only possible using computer-assisted methods. Furthermore, it is important to be able to link the system that drug research needs, that is to compress the time frames for development, reduce the cost of discovering new drugs and bringing them to patients, to identification of genetic disease targets, and manage the rapidly expanding information available in genomic and other databases. Application of the breakthrough tools developed within disciplines such as combinatorial chemistry, screening of immense libraries of natural and chemically synthesized compounds for desirable biological and therapeutic activity utilizing bioinformatics in combination with model studies to rationales for drug discoveries.

Role of Genomics

The complete set of instructions for making a functional organism reside in the genome. It contains the master blueprints for all cellular structure and activities for the lifetime of the cell or organism. Human genome project was started in 1988 with the goal to validate the information that makes up the genetic blueprint of human beings. The genetic information fostered by the human genome project (HGP) with advances in computer technology resulted in the relatively young field of bioinformatics [4], leading to new gene discovery with a view of understanding the genetic information that is encoded within the genome of an organism. The science is essentially divided into two main activities, one concerned with "structural genomics", the other with "functional genomics" [5]. Within genomics there are many newly emerging specialist disciplines, such as comparative genomics, epigenomics, structural genomics and the most important pharmacogenomics. "Pharmacogenomics" generally refers to a particular application of genomic technologies in drug discovery and development [6].

Structural genomics is predominantly concerned with identifying gene sequences and the

three dimensional structure of the proteins encoded by genes, and functional genomics is more specifically focused upon understanding and interpreting gene activity. Predicting the three-dimensional structure of a protein from its amino acid sequence is one of the most important current problems of modern biology [7]. Both rely upon bioinformatics for interpretation in a system-wide context and gene expression and protein analysis for comprehensive measurements and validation of specific a gene.

The human genome project's success in sequencing the bases of DNA has revolutionized molecular biology. It created new knowledge about fundamental biological processes. The genome community was an early adopter of the Web, finding in it a way to publish its vast accumulation of data and to manipulate and analyze the data to facilitate access by use of bioinformatics resources.

Genome sequencing has progressed very rapidly due to improvements in speed and capacity of computers, opening up the possibility that many of the genome centers will be able to sequence specific genomes. In early years of genomics, human genome scientists used automatic gene sequencing technology to compile one of the largest database of human and microbial genes. The major biological challenge in functional genomics has been to understand the effects genes have through their relationship with RNA and associated active proteins. As a result of this, genomics research will lead to the development of techniques beyond the current ability to define functional differences between biological states. Stimulating biological understanding and improved technology analysis tools, together with the impact of interdisciplinary sciences, will provide genomics with the ability to accurately predict biological mechanisms and more precisely infer functions. Each successive generation of genomic databases have become more comprehensive, they have moved from offering raw sequenced data only, to gene expression profiles and interest-stimulating protein-protein interaction profiles. Efficient, accurate and automatic clustering of large protein sequence datasets with new algorithm has been developed [8].

With the advent of new techniques, genomics will continue to rapidly grow as it already has over the last several years. A successful integration of bioinformatics methods into drug discovery programs can improve target discovery and validation and accelerated drug development by focusing research efforts on novel genomics targets [9]. The evolution of this model is likely to be in the form of alliances with larger biotech and pharmaceutical companies, since the barriers to successful entry into this side of the market are still limited by time frames for clinical trials and regulatory approvals.

Proteomics

With the availability of genome sequence, proteomics is becoming increasingly important. Expression information from mRNA and proteins is required to understand gene network [10]. The importance of total proteins expression is motivating and promoting proteomics. Proteomics relies mainly on microchemical characterization of peptides separated by two dimensional gel electrophoresis and monitor synthesis rates, expression level and post translational modifications. As with genomic data derived from microarray analysis, an information framework is necessary to organize proteomics data. An elaborate and effective link has to be established between protein level and nucleic acid level information about genes and gene network. A further challenge is to evolve an information system to define the three-dimensional structure of the proteins and the modified proteins, i.e. glycoproteins, phosphoproteins, and other post-translationally modified proteins, to define the structure-function relationship of the total expressed proteins. The human genome project has provided in depth studies on gene isolation, sequencing, translation-related phenomena and protein structure utilizing cell biology, molecular biology and protein chemistry.

Bioinformatics: Future of life sciences.

The new 'big science' of genomics sets out not only to reveal the DNA sequence of all the 100,000 or so genes which constitute our total genetic makeup, or genome, but further to understand how these genes function in health and disease. From this knowledge, new therapies will emerge to combat those diseases for which effective treatment still remains elusive. To bring sufficient data on the human genome and on the variations in the many phenotypes of disease and health, develop effective and accurate data-mining tools to allow us to take advantage of the data. The focus of genomics is DNA. However, DNA is not a basic building block from which living cells are made but it is, instead, an

information carrier. Proteins are the core structural and functional molecules through which all other biomolecules and organisms are made and are, therefore, the central effectors of health and disease at the molecular level. The DNA sequence determines corresponding sequences of genes and the active sequences of proteins. It also deals with differences in DNA that distinguish the proteins of one species or individual from another.

The science of proteomics sets out to reveal the structure and function of the entire repertoire of proteins expressed in an organelle, cell, tissue or organism at any time. These sciences need an equally potent partner to store, manage, retrieve, analyze, integrate vast amounts of data and design experiment to validate the formulations, findings and predictions based on these data, now being produced globally, hence bioinformatics. Initial interest in bioinformatics was fuelled by the need to create huge databases, such as GenBank, SWISS-PROT and EMBL (European Molecular Biology Laboratory) and DNA Database of Japan to store and compare the DNA sequence data erupting from the human genome and other genome sequencing projects. The GenBank sequences database incorporates publicly available DNA sequences of > 55, 000 different organisms [11]. SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation [12] and the EMBL Nucleotide Sequence Database is maintained at the European Bioinformatics Institute (EBI) in an international collaboration with the DNA Data Bank of Japan [13], and a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance [14].

Bioinformatics, today embraces protein structure analysis, gene and protein functional information, data from patients, pre-clinical and clinical trials, and the metabolic pathways of numerous species. The management and, more importantly, accessibility of this data is directly attributable to the development of the Internet, particularly the world wide web (www), and a data specification for storing and describing biomolecular intra- and interactions, molecular complexes and pathways [15].

Microarrays and DNA Chips

The aim of bioinformatics is to create software solutions such as programs that discover and identify the links between the thousand of genes, which interact to create a disease state or behavioral problems, such as microarray or DNA Chip technology. DNA microarrays technology is produced by *in situ* synthesis of oligonucleotide [16,17], or the immobilization of pre-fabricated molecules [18].

In order to analyze a large number of genotypes at a time microarrays and DNA Chips, obtained by chemical syntheses, should be used for the presentation of large number of predefined probes to a genomic DNA target [19]. The most fascinating prospect for the application of DNA Chips is likely to be search of mutation detection, with the increasing range of genetic disorder, for which unique mutation has been identified.

The synthesis of the probe on the Chip surface utilizes the technology of fabrication the sequential application of thin layer material in different pattern [20]. Analyses on DNA microarrays considerably depend on spot quality and a low background signal of the glass support. Currently, glass slides are mainly used as support medium because of their favorable optical characteristics especially for transcriptional profiling analyses [21,22]. DNA arrays, or biochips, represent blossoming field technically and from a business point of view as well [23], this technology reverses the old fashion approach of screening the thousand of clones manually. Instead of screening probe a cloned gene, PCR product, or synthesized oligonucleotide, a defined DNA fragment occupies each portion or "probe cell" in the array, and the array is probed with the unknown sample [24].

With the development of new ways to fabricate the Chips either on glass or plastic, and availability of arrays of different sizes, researchers now have the option of readymade chips or building their own customized Chips in their lab.

Arrays Application

Arrays provide a method for rapid genotyping thus facilitating the diagnosis of diseases for which a gene mutation has been identified. They also assist in the identification of sentinel genes that demonstrate altered expression in a given cell or tissue type in response to drug exposure, and help in the selection of custom and rational drug therapy, identification of signature genes indicative of a disease process.

Arrays can also identify candidate targets for therapeutic intervention.

Dynamic Controls

Artificial neural networks (ANN) are collections of mathematical models that emulate some of the observed properties of biological nervous systems and draw on the analogies of adaptive biological learning. The key element of the ANN paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements that are analogous to neurons and are tied together with weighted connections that are analogous to synapses.

Today ANNs are being applied to an increasing number of real world problems of considerable complexity such as developing bioinformatic tools. Optical pattern recognition is the process by which objects or patterns are imaged optically and then analyzed by computer algorithms. They are good pattern recognition engines and robust classifiers, with the ability to generalize in making decisions about imprecise input data. They offer ideal solutions to a variety of classification problems such as speech, character and signal recognition patterns, as well as functional prediction and system modeling where the physical processes are not understood or are highly complex.

Cellular Function

Functional Genomics aims at the systematic understanding of gene networks. It uses computational algorithms to identify co-regulated genes from gene expression (e.g. microarray) and sequence data (e.g. binding sites, motifs). In addition, it applies text data-mining techniques to identify protein-protein and protein-DNA interaction from literature. Using biological knowledge of functional genomics, we canpoint to potential targets for further simulation and experimental verification. Currently, the focus is on genes involved in metabolic pathways and intracellular signal transduction to elucidate the qualitative and quantitative dynamics of pathway and signaling processes in the cell. It has a scope in more complex targets, for example apoptosis and intercellular signaling networks such as cell growth control and differentiation by hematopoeitic growth factors (Fig 1).

Cell functions through information analysis and computer simulation

Complex behavior of living cells is achieved by networks of interactions between proteins and other biomolecules. Bioinformatics conducts information analysis and computer simulation to elucidate the cellular mechanisms on the basis of information on protein structure and function. Using

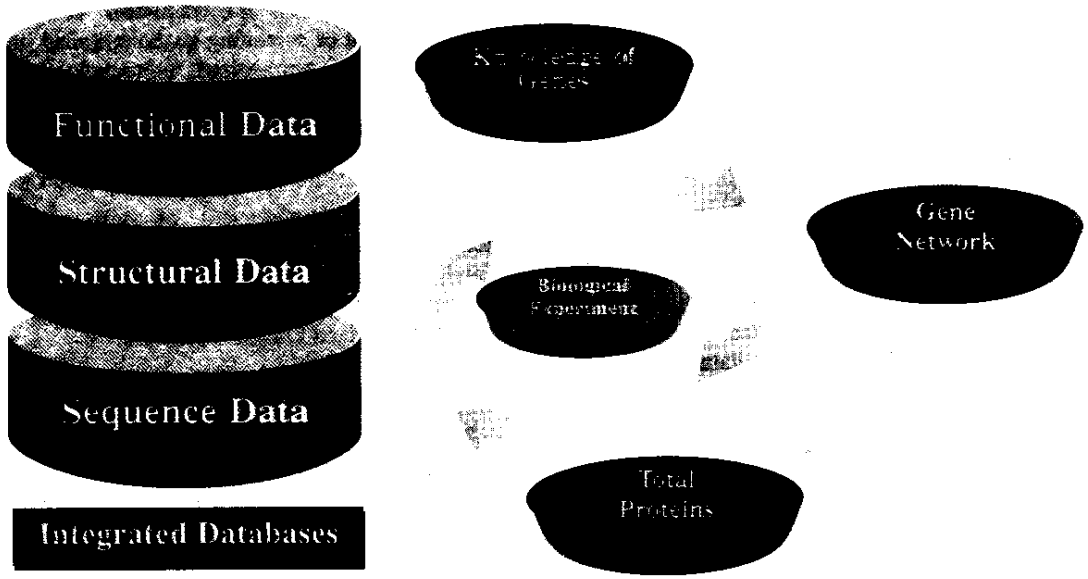

Fig. 1: Knowledge of Genes: Gene knowledge depicting motives providing different basis for the combination of accumulated data in a gene and its operation in the genetic network resulting in total protein synthesis and protein function.

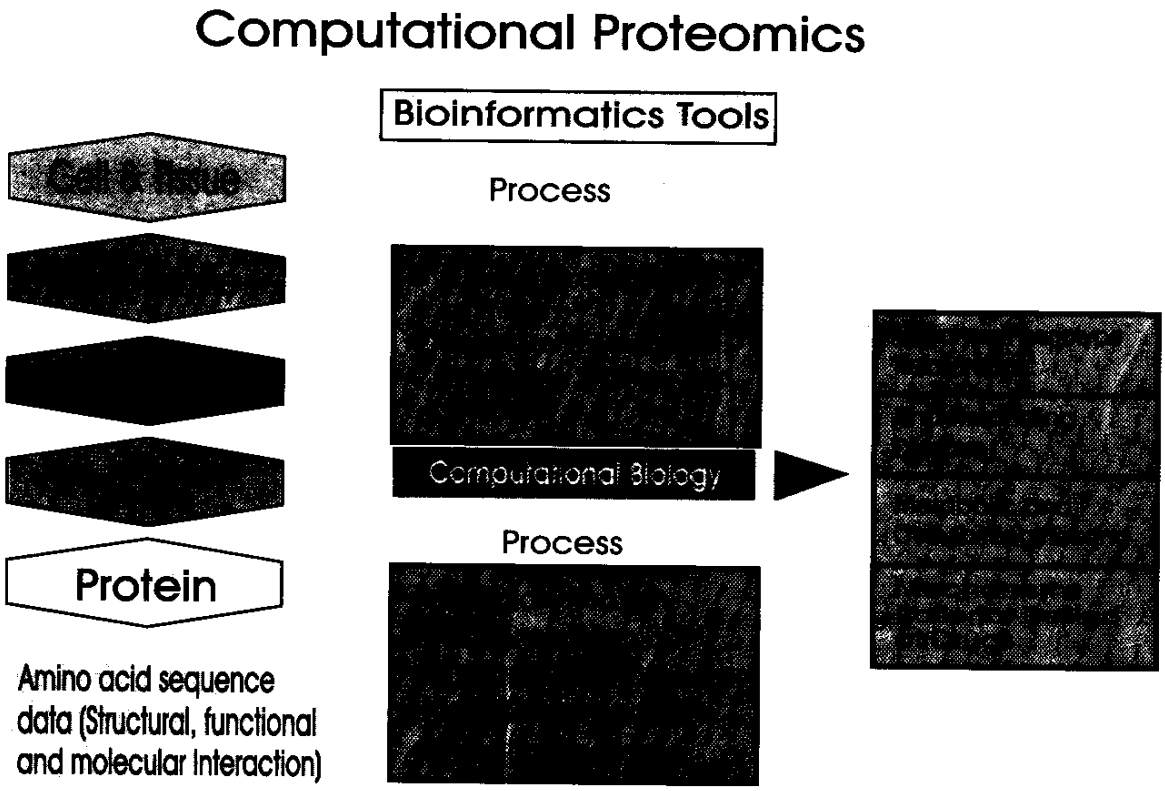

Fig. 2: Computational Proteomics: Process of research simulation to elucidate the cellular mechanism on the basis of amino acid sequences data information. Bioinformatic tools provide classification of protein sequences, and functional relationship, and to predict the protein function using 3D structure. Also to elucidate molecular mechanism to design novel drugs for diseases.

amino acid sequence and structure data, proteins are systematically classified. This classification provides useful knowledge that is acquired concerning relationships among sequence, structure, function, and evolution. A large number of examples of interactions of proteins with other proteins and modified molecules are being collected and organized into a database. In this way, on the basis of information on structures and functions of individual molecules, the functions and behavior of a system are theoretically modeled. Using computer simulations, behavior of cells is predicted in response to such external and internal factors as environmental conditions, drugs, and genetic mutations and polymorphisms. The current research efforts are expected to make a major contribution to biomedical science. For example, it will help to elucidate molecular mechanisms of various diseases and to rationally design novel drugs. It will also be useful in order to predict the effects of differences in genotypes such as Single Nucleotide Polymorphism (SNPs) on phenotypes, e.g., disease susceptibility and drug responses, and achieving evidence-based personalized medicine by prescribing appropriate drug, for the right patient at the right time, in the right dose for proper treatment and medication according to the individual's characteristics (Fig. 2).

Genomic functions at the organism level

DNA changes, mutations, are reflected not only in hereditary diseases, but also in individual constitutional differences, such as fatigability and response to foods, medicine, and environment. The objectives of the recent projects are to identify what genetic difference (genotype) is responsible for what constitutional variance (phenotype) at the organism level. At present, genotypes can be determined

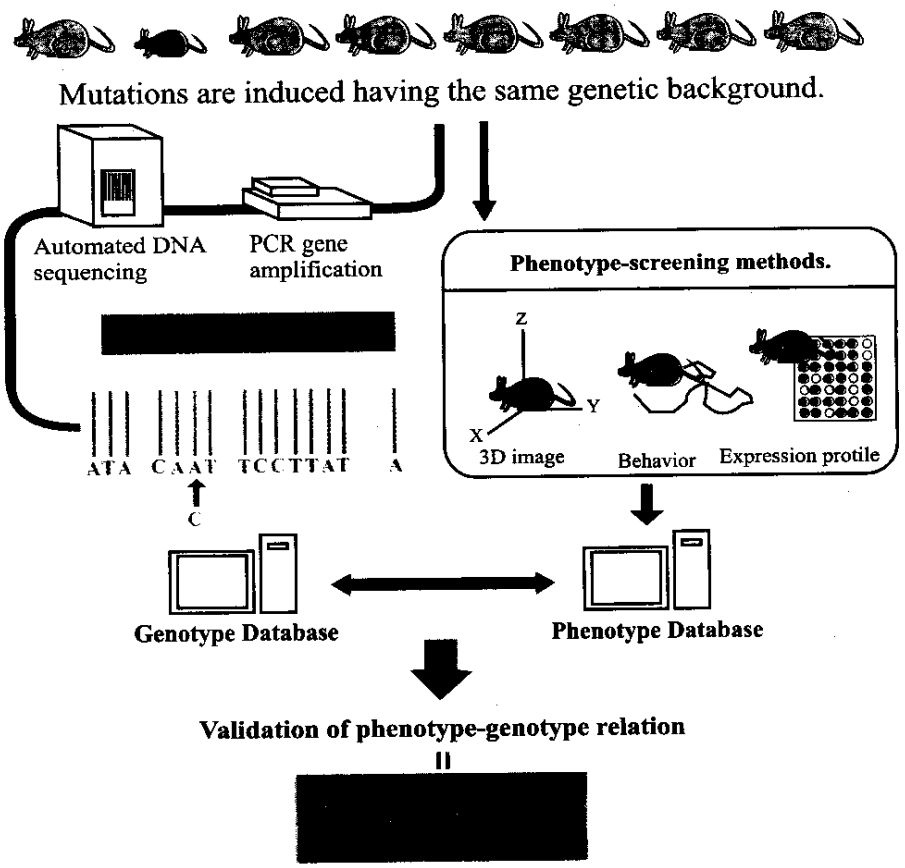

Fig. 3: Understanding the Genomic Function: Processes to find the genetic difference in genotype – phenotype at the organism level to recognize the mutation with genotype using automated DNA sequence to analyses genomic functions, and for phenotype to find the difference using screening method having 3D images, behavior and expression profile to understand the genomic functions.

precisely at the DNA sequence level. However, measurement of phenotypes is still a problem because of environmental effects. It is necessary to use animal models to effectively collect highly reliable information to detect genetic factors. The animal models will provide information for phenotype assessment, affected phenotype(s) based on each identified phenotype, change(s) in DNA due to each recognized phenotype, information on phenotypes and genotypes, and to construct an integrated functional genomics database on a large scale.

This work will make "order-made therapy" feasible and lead to the improvement of "quality of life". Mutant mice with clear genotypes and phenotypes will contribute to medicine and drug discovery as animal models for human disease and development of new treatment methods. By using plant model systems such as Arabidopsis, a similar approach will also contribute to the development of novel inbreeding systems (Fig 3).

Conclusions

New methodologies based on precision analytical techniques coupled with "super computers", will radically change the practice of science and its implementation. The Internet will have evolved into the interspaces. The next decade will emerge as a indispensable bioinformatics structure, and the global information infrastructure will support semantics for indexing and analysis of

available information. Observing the gene research areas, gene therapy approach will be focused on treating diseases based on modifying the expression of genes toward a therapeutic goal. Main strategies for organ generation and related cell therapies include stem cell[2] technology, and the xenotransplantation based upon the combination of a number of new technologies such as cloning and germ line genetic engineering, in particular, gene targeting, that permit the manipulation of genes. Stem Cell strategies rely on injection of stem cells and growth factors that may be given to patients to regenerate damaged organs. Organs could be grown outside of the patient's body using either the cells of the patient or a donor alongwith reorganized animal organs and tissue for transplantation into human patients.

With the ease of data information about gene expression at the protein level, the critical data on the genotype–phenotype relationship in variety of settings will be available. Such data is vital and necessary for metabolic and cellular engineering. The above mentioned procedures are likely to increase substantially in future, laying the basis of novel approaches to health and disease.

Acknowledgement

In part bioinformatic studies on glycoproteins has been supported by the TDR-WHO grant No:970604 to Nasir-ud-Din.

Reference

1. J. Park, L. Holm, A. Heger and C. Chothia. *Bioinformatics,* **16** (5), 458 (2000).
2. S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and Lipman. *Nucleic Acid Res:* **25**, 3389 (1997).
3. W.R. Pearson, and D.J. Lipman, *Proc. Natl Acd. Sci USA.* **85,** 2444 (1998).
4. W. T. Durso. *The Scientist,* **11** 13 (1997)
5. P. Hieter, and M. Boguski, *Scinece,* **278**, 601 (1997).
6. K.K. Jain. *Drug Discovery Today* **4**, 50 (1999).
7. N. Siew and D. Fischer. *IBM System Journal,* **40**, 410 (2001).
8. A. J. Enright and C. A. Ouzounis. *Bioinformatics,* **16**, 451 (2000).
9. L.L. Bellavance, M.S. Donlan, and S. Sharp, *Genetic Engineering News,* **19** 32 (1999).
10. Hetimanikatis, and E.K. Lee, *Metab Eng. 1:* 275 (1999).
11. D. A. Benson, I. K. Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp and D. L. Wheeler. *Nucleic Acid Res.*, **28**, 15 (2000).
12. A. Bairoch and R. Apweiler. *Nucleic Acid Res.* **28**, 45 (2000)
13. G. Stoesser, W. Baker, A. Broek, E. Camon, M. G. Pastor, C. Kanz, T. Kulikova, V. Lombard, R. Lopez, H. Parkinson, N. Redaschi, P. Sterk, P. Stoehr and M. A. Tuli. *Nucleic Acid Res.*, **29**, 17 (2001).
14. G. Pesole, S. Liuni and M. D'Souza. *Bioinformatics,* **16**, 439 (2000).
15. G. D. Bader and C. W. V. Hogue. *Bioinformatics,* **16**, 465 (2000).
16. U. Maskos, and E.M. Southem, *Nucleic Acid Res.* 20, 1679 (1992).
17. S.P.A. Fodor, R.P. Rava, X.C. Huang, A.C. Pease, C.P. Holmes, and C.L. Adams, *Nature* **364**, 555 (1993).
18. M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. *Science,* **270**, 467 (1995).
19. E.M Southerm, *Trends Genet.,* **12** (3), 110 (1996).
20. D.T. Burke, M.A. B. C Mastrangelo *Genome Res,* **7**, 189 (1997).
21. M.B., Eisen, and P.O. Brown. *Methods Enzymol.*, 303, 179 (1999).
22. J. G. Hacia, L.C. Brody, M.S Chee, S.P. Fodor, F.S. Collins. *Nat Genet.*, **14**, 367 (1996).
23. B. Sinclair. *The Scientist* **13**: 18 (1999).
24. F. Diehl, S. Grahlmann, M. Beier and J. D. Hoheisel, *Nucleic Acids Res.* **29**, e38 (2001).