International Research Journal of Biotechnology (ISSN: 2141-5153) Vol. 3(9) pp. 141-151, November, 2012
Available online http://www.interesjournals.org/IRJOB

Review

Bioinformatics and scientific research

[1]Ibiam O. F. A. and [2]Ekwe A.

[1]Department of Applied Biology, Faculty of Biological Sciences, Ebonyi State University, Abakaliki. Ebonyi State.
[2]Department of Computer Science, Faculty of Physical Sciences, Ebonyi State University, Abakaliki. Ebonyi State.
E-mail: drakanuibiamjr@yahoo.com

Accepted November 25, 2012

Bioinformatics is the application of information technology to the field of molecular biology. The primary goal of bioinformatics is to increase our understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques (e.g., data mining, machine learning algorithms, and visualization) to achieve this goal. Bioinformatics derives knowledge from computer analysis of biological data. These can consist of the information stored in the genetic code, but also experimental results from various sources, patient statistics, and scientific literature. Research in bioinformatics includes method development for storage, retrieval, and analysis of the data. Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine.

Keywords: Bioinformatics, Analysis, Scientific, Research, Data.

INTRODUCTION

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize, and index the data and for specialized tools to view and analyze the data.

History

The history of computing in biology goes back to the 1920s when scientists were already thinking of establishing biological laws solely from data analysis by induction (e.g. A.J. Lotka, Elements of Physical Biology, 1925). However, only the development of powerful computers, and the availability of experimental data that can be readily treated by computation (for example, DNA or amino acid sequences and three–dimensional structures of proteins) launched bioinformatics as an independent field. Today, practical applications of bioinformatics are readily available through the World Wide Web, and are widely used in biological and medical research. As the field is rapidly evolving, the very definition of bioinformatics is still the matter of some debate.

Bioinformatics is the application of information technology to the field of molecular biology. The term *bioinformatics* was coined by Paulien Hogeweg in 1978 for the study of informatics processes in biotic systems. Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. It is the name given to these mathematical and computing approaches used to glean understanding of biological processes. Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures.

The primary goal of bioinformatics is to increase our understanding of biological processes. What sets it apart

from other approaches, however, is its focus on developing and applying computationally intensive techniques (e.g., data mining, machine learning algorithms, and visualization) to achieve this goal. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expressionand protein-protein interactions, genome-wide association studies and the modeling of evolution. For researchers to benefit from the data stored in a database, these additional requirements must be met:

i. easy access to the information

ii. a method for extracting only that information needed to answer a specific biological question(A reliable and easy query mechanism).

iii. A record generating system.

iv. The data must non-bulky and flexible with consistent data free of redundant data.

At the beginning of the "genomic revolution", a bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues but the development of multifaceted or comprehensive interfaces whereby researchers could both access existing data as well as submit new or revised data.The field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as computational biology. Development of this type of database involved not only design issues but the development of comprehensive interfaces whereby researchers could both access existing data as well as submit new or revised data.

Important sub-disciplines within bioinformatics and computational biology include:

i. the development and implementation of tools that enable efficient access to, and use and management of, various types of information

ii. the development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences

The relationship between computer science and biology is a natural one for several reasons.

i. The phenomenal rate of biological data being produced provides challenges: massive amounts of data have to be stored, analysed, and made accessible.

ii.The nature of the data is often such that a statistical method, and hence computation, is necessary. This applies in particular to the information on the building plans of proteins and of the temporal and spatial organisation of their expression in the cell encoded by the DNA.

iii. There is a strong analogy between the DNA sequence and a computer program (it can be shown that the DNA represents a Turing Machine).

Analyses in bioinformatics focus on three types of datasets: i genome sequences. ii macromolecular structures and iii. functional genomics experiments (e.g. expression data, yeast two–hybrid screens).

However, bioinformatic analysis is also applied to various other data, e.g. taxonomy trees, relationship data from metabolic pathways, the text of scientific papers, and patient statistics. A large range of techniques are used, including primary sequence alignment, protein 3D structure alignment, phylogenetic tree construction, prediction and classification of protein structure, prediction of RNA structure, prediction of protein function, and expression data clustering. Algorithmic development is an important part of bioinformatics, and techniques and algorithms were specifically developed for the analysis of biological data (e.g., the dynamic programming algorithm for sequence alignment).

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

A biological database is an organized body of data or rather logically related biological data which may be large or quite small .This is usually manipulated or managed using a Database management System(DBMS) such as Foxpro, Oracle, Sybase, Informix etc. designed to update, sort, query, delete and retrieve the components of the date stored in a given database. A simple database might be a single file containing many records, each of which includes the same set of information. For example, a record associated with a nucleotide sequence database typically contains fields such as :- Nucleotide ID; Input Sequence: Type Molecule: Scientific name of source organism: Molecular Description; Method of Isolation; and Literature citations associated with the sequence.

Bioinformatics and its application

The rationale for applying computational approaches to

facilitate the understanding of various biological processes includes:

i. A more global perspective in experimental design
ii. The ability to capitalize on the emerging technology of database-mining-the process by which testable hypotheses are generated regarding the function or structure of a gene or protein of interest by identifying similar sequences in better characterized organisms

Bioinformatics and research

Bioinformatics was applied in the creation and maintenance of a database to store biological information at the beginning of the "genomic revolution", such as nucleotide and amino acid sequences. In order to study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. It has a large impact on biological research. Human genome project would be meaningless without the bioinformatics component. The goal of sequencing projects, for example, is not to corroborate or refute a hypothesis, but to provide raw data for later analysis. Once the raw data are available, hypotheses may be formulated and tested *in silico*. In this manner, computer experiments may answer biological questions which cannot be tackled by traditional approaches.

Three key areas are the organisation of knowledge in i. Databases, ii. Sequence analysis and iii. Structural bioinformatics.

Organizing biological knowledge in databases

Biological raw data are stored in public databanks (such as Genbank or EMBL for primary DNA sequences). The data can be submitted and accessed via the world wide web. Protein sequence databanks like trEMBL provide the most likely translation of all coding sequences in the EMBL databank. Sequence data are prominent, but also other data are stored, e. g. yeast two–hybrid screens,expression arrays, systematic gene–knock–out experiments, and metabolic pathways.The stored data need to be accessed in a meaningful way, and often contents of several databanks or databases have to be accessed simultaneously and correlated with each other. Special languages have been developed to facilitate this task (such as the Sequence Retrieval System (SRS) and the Entrez system). An unsolved problem is the optimal design of inter–operating database systems.

Databases provide additional functionality such as access to sequence homology searches and links to other databases and analysis results, for example, SWISSPROT contains verified protein sequences and more annotations describing the function of a protein. Protein 3D structures are stored in specific databases (for example, the Protein Data Bank now primarily curated and developed by the Research Collaboratory for Structural Bioinformatics). Organism specific databases have been developed (such as ACEDB, the A C. Elegans Data Base for the *C. elegans* genome, FLYBASE for *D. melanogaster* etc). A major problem are errors in databanks and databases (mostly errors in annotation), in particular since errors propagate easily through links. Also databases of scientific literature (such as PUBMED, MEDLINE) provide additional functionality, e.g. they can search for similar articles based on word–usage analysis. Text recognition systems are being developed that extract automatically knowledge about protein function from the abstracts of scientific articles, notably on protein–protein interactions.

Analysing sequence data

The primary data of sequencing projects are DNA sequences. These become only really valuable through their annotation. Several layers of analysis with bioinformatics tools are necessary to arrive from a raw DNA sequence at an annotated protein sequences: establish the correct order of sequence contigs to obtain one continuous sequence; find the translation and transcription initiation sites, find promoter sites, define open reading frames (ORF); find splice sites, introns, exons; translate the DNA sequence into a protein sequence, searching all six frames; compare the DNA sequence to known protein sequences in order to verify exons etc with homologuous sequences. Some completely automated annotation systems have been developed (e.g., GENEQUIZ), which use a multitude of different programs and methods.

The protein sequences are further analysed to predict function. The function can often be inferred if a sequence of a homologous protein with known function can be found. Homology searches are the predominant bioinformatics application, and very efficient search methods have been developed . The often difficult distinction between orthologous sequences and paralogous sequences facilitates the functional annotation in the comparison of whole genomes. Several methods detect glycolysation, myristylation and other sites, and the prediction of signal peptides in the amino acid sequence give valuable information about the subcellular location of a protein.The ultimate goal of sequence annotation is to arrive at a complete functional description of all genes of an organism. However, function is an ill–defined concept. Thus, the simplified idea of “one gene – one protein – one structure – one

function" cannot take into account proteins that have multiple functions depending on context (e.g., subcellar location and the presence of cofactors). Well-known cases of "moonlighting" proteins are lens crystalline and phosphoglucose isomerase. Currently, work on ontologies is under way to explicitly define a vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing.

Families of similar sequences contain information on sequence evolution in the form of specific conservation patters at all sequence positions. Multiple sequence alignments are useful for building sequence profiles or Hidden Markov Models to perform more sensitive homology searches. A sequence profile contains information about the variability of every sequence position. improving structure prediction methods (secondary structure prediction). Sequence profile searches have become readily available through the introduction of PsiBLAST [3]; studying evolutionary aspects, by the construction of phylogenetic trees from the pair-wise differences between sequences: for example, the classification with 70S, 30S RNAs established the separate kingdom of Archeae; determining active site residues, and residues specifc for subfamilies; predicting protein–protein interactions; analysing single nucleotide polymorphisms to hunt for genetic sources of diseases.

Many complete genomes of microorganisms and a few of eukaryotes are available . By analysis of entire genome sequences a wealth of additional information can be obtained. The complete genomic sequence contains not only all protein sequences but also sequences regulating gene expression. A comparison of the genomes of genetically close organisms reveals genes responsible for specific properties of the organisms (for example infectivity). Protein interactions can be predicted from conservation of gene order or operon organisation in different genomes. Also the detection of gene fusion and gene fission (that is, one protein is split into two in another genome) events helps to deduce protein interactions.

Structural bioinformatics

This branch of bioinformatics is concerned with computational approaches to predict and analyse the spatial structure of proteins and nucleic acids. Whereas in many cases the primary sequence uniquely specifies the three–dimensional (3D) structure, the specific rules are not well understood, and the protein folding problem remains largely unsolved. Some aspects of protein structure can already be predicted from amino acid content. Secondary structure can be deduced from the primary sequence with statistics or neural networks. When using a multiple sequence alignment, secondary structure can be predicted with an accuracy above 70 %.3D models can be obtained most easily if the 3D structure of a homologous protein is known (homology modelling, comparative modelling). A homology model can only be as good as the sequence alignment: whereas protein relationships can be detected at the 20% identity level and below, a correct sequence alignment becomes very difficult, and the homology model will be doubtful. From 40 to 50% identity, the models are usually mostly correct; however, it is possible to have 50% identity between two carefully designed protein sequences with different topology (the so–called JANUS protein). Remote relationships that are undetectable by sequence comparisons may be detected by sequence–to–structure–fitness (or threading) approaches: the search sequence is systematically compared to all known protein structures. Predictions of protein 3D structure remain the major challenge *ab initio*, and some progress has been made recently by combining statistical with force–field based approaches.

Membrane proteins are drug targets. It is estimated that membrane receptors form 50% of all drug targets in pharmacological research. However, membrane proteins are underrepresented in the PDB structure database. Since membrane proteins are usually excluded from structural genomics initiatives due to technical problems, the prediction of transmembrane helices and solvent accessibility is very important. Modern methods can predict transmembrane helices with a reliability greater than 70 %.Understanding the 3D structure of a macromolecule is crucial for understanding its function. Many properties of the 3D structure cannot be deduced directly from the primary sequence. Obtaining better understanding of protein function is the driving force behind structural genomics efforts, which can be thus understood as part of functional genomics. Similar structure can imply similar function and general structure–to–function relationships can be obtained by statistical approaches, for example, by relating secondary structure to known protein function or surface properties to cell location.

The increased speed of structure determination necessary for the structural genomics projects makes an independent validation of the structures (by comparison to expected properties) important, particularly. Structure validation helps to correct obvious, for example, in the covalent structure, and leads to a more standardized representation of structural data, by agreeing on a common atom name nomenclature. The knowledge of the structure quality is a prerequisite for further use of the structure, as found in molecular modelling or drug design. In order to make as much data on the structure and its determination available in the databases, approaches for

automated data harvesting are being developed. Structure classification schemes, as implemented in the SCOP, CATH, and FSSP databases, elucidate the relationship between protein folds and function, and shed light on the evolution of protein domains.

Protein folds can be analysed for whole genomes, protein–protein interactions predicted on the sequence level can be studied in more detail on the structure level and single Nucleotide polymorphisms can be mapped on 3D structures of proteins in order to elucidate specific structural causes of disease. More detailed aspects of protein function can be obtained also by force–field based approaches. Whereas protein function requires protein dynamics, no experimental technique can observe it directly on an atomic scale, and motions have to be simulated by molecular dynamics (MD) simulations. Free energy differences between binding energies of different protein ligands could be characterized by MD simulations. Molecular mechanics or molecular dynamics based approaches are also necessary for homology modelling and for structure refinement in X–ray crystallography and NMR structure determination.

Drug design, exploits the knowledge of the 3D structure of the binding site (or the structure of the complex with a ligand) to construct potential drugs, for example, inhibitors of viral proteins or RNA. In addition to the 3D structure, a force field is necessary to evaluate the interaction between the protein and a ligand,to predict binding energies. In virtual screening, a library of molecules is tested on the computer for their capacities to bind to the macromolecule. Patient databases with genetic profiles as in cardiovascular diseases, diabetes, cancer, etc. may play an important role in the future for individual health care, by integrating personal genetic profile into diagnosis, despite obvious ethical problems. The goal is to analyse a patient's individual genetic profile and compare it with a collection of reference profiles and other related information. This may improve individual diagnosis, prophylaxis, and therapy.

Major research areas of bioinformatics

Sequence analysis

This has to do with Sequence alignment and Sequence database. Since the Phage Φ-X174 was sequenced in 1977, the DNA sequences of hundreds of organisms have been decoded and stored in databases. The information is analyzed to determine genes that encode polypeptides, as well as regulatory sequences. A comparison of genes within species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematic to construct phylogenetic trees). Computer programs are now used to search the genome of thousands of organisms, containing billions of nucleotides. These programs would compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, in order to identify sequences that are related, but not identical. A variant of this sequence alignment is used in the sequencing process itself. The so-called shotgun sequencing technique (which was used, for example, by The Institute for Genomic Research to sequence the first bacterial genome, *Haemophilus influenzae*) does not give a sequential list of nucleotides, but instead the sequences of thousands of small DNA fragments (each about 600-800 nucleotides long). The ends of these fragments overlap and, when aligned in the right way, make up the complete genome. Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes.

Another aspect of bioinformatics in sequence analysis is the automatic search for genes and regulatory sequences within a genome. Not all of the nucleotides within a genome are genes. Within the genome of higher organisms, large parts of the DNA do not serve any obvious purpose. This so-called junk DNA may, however, contain unrecognized functional elements. Bioinformatics helps to bridge the gap between genome and proteome projects--for example, in the use of DNA sequences for protein identification.

Genomics

Estimating the number of genes in an organism basing on the number of nucleotide base pairs was not reliable, due to the presence of high numbers of redundant copies of many genes. Genomics has corrected this situation. Useful genes can be selected from a gene library thus constructed and inserted into other organisms for improvement or harmful genes can be silenced. In the areas of structural genomics, functional genomics and nutritional genomics, bioinformatics plays a vital role.

Comparative genomics

The core of comparative genome analysis is the establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms. It is these intergenomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. A multitude of evolutionary events acting at various organizational levels shape genome evolution. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral

transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation. The complexity of genome evolution poses many exciting challenges to developers of mathematical models and algorithms, who have recourse to a spectra of algorithmic, statistical and mathematical techniques, ranging from exact, heuristics, fixed parameter and approximation algorithms for problems based on parsimony models to Markov Chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models. Many of these studies are based on the homology detection and protein families computation.

Analysis of gene expression

The expression of many genes can be determined by measuring mRNA levels with multiple techniques including microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed *in-situ* hybridization. All of these techniques are extremely noise-prone and/or subject tobias in the biological measurement, and a major research area in computational biology involves developing statistical tools to separate signal from noise in high-throughput gene expression studies. Such studies are often used to determine the genes implicated in a disorder: one might compare microarray data from cancerous epithelial cells to data from non-cancerous cells to determine the transcripts that are up-regulated and down-regulated in a particular population of cancer cells.

Genome annotation

It has to do with Gene finding. In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation software system was designed in 1995 by Dr. Owen White, who was part of the team that sequenced and analyzed the first genome of a free-living organism to be decoded, the bacterium *Haemophilus influenzae*. Dr. White built a software system to find the genes (places in the DNA sequence that encode a protein), the transfer RNA, and other features, and to make initial assignments of function to those genes. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA are constantly changing and improving.

Genome Mapping

Genomic maps serve as a scaffold for orienting sequence information. A few years ago, a researcher wanting to localize a gene, or nucleotide sequence, was forced to manually map the genomic region of interest, a time-consuming and often painstaking process. Today, thanks to new technologies and the influx of sequence data, a number of high-quality, genome-wide maps are available to the scientific community for use in their research. Computerized maps make gene hunting faster, cheaper, and more practical for almost any scientist. In a nutshell, scientists would first use a genetic map to assign a gene to a relatively small area of a chromosome. They would then use a physical map to examine the region of interest close up, to determine a gene's precise location. In light of these advances, a researcher's burden has shifted from mapping a genome or genomic region of interest to navigating a vast number of Web sites and databases.

Map Viewer

A Tool for Visualizing Whole Genomes or Single Chromosomes

Map Viewer is a tool that allows a user to view an organism's complete genome, integrated maps for each chromosome (when available), and/or sequence data for a genomic region of interest. When using Map Viewer, a researcher has the option of selecting either a "Whole-Genome View" or a "Chromosome or Map View". Genome View displays a schematic for all of an organism’s chromosomes, whereas the Map View shows one or more detailed maps for a single chromosome. If more than one map exists for a chromosome, Map Viewer allows a display of these maps simultaneously.

Using Map Viewer, researchers can find answers to questions such as

i. Where does a particular gene exist within an organism's genome?
ii. Which genes are located on a particular chromosome and in what order?
iii. What is the corresponding sequence data for a gene that exists in a particular chromosomal region?
iv. What is the distance between two genes?

Proteomics

Proteomics involves the sequencing of amino acids in a protein, determining its three dimensional structure and

relating it to the function of the protein. Before computer processing comes into the picture, extensive data, particularly through crystallography and NMR, are required for this kind of a study. With such data on known proteins, the structure and its relationship to function of newly discovered proteins can be understood in a very short time. In such areas, bioinformatics has an enormous analytical and predictive potential. Protein folding alone of the most significant and fundamental problem in biological science realizing this, IBM in Dec 1999, had built a supercomputer, which is 2 million times faster than the today's fastest desktop PC. This new computer nicknamed "Blue Gene" by IBM researchers will be capable of performing more than one quadrillion operations per second. Better understanding of how proteins fold will give scientists and doctors better insight into diseases and ways to combat them.

Prediction of protein structure

It has to do with Protein structure prediction Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment.(Of course, there are exceptions, such as the bovine spongiform encephalopathy-aka Mad Cow Disease-prion.) Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as one of *secondary*, *tertiary* and *quaternary* structure. A viable general solution to such predictions remains an open problem. As of now, most efforts have been directed towards heuristics that work most of the time.

One of the key ideas in bioinformatics is the notion of homology. In the genomic branch of bioinformatics, homology is used to predict the function of a gene: if the sequence of gene *A*, whose function is known, is homologous to the sequence of gene *B,* whose function is unknown, one could infer that B may share A's function. In the structural branch of bioinformatics, homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably. One example of this is the similar protein homology between hemoglobin in humans and the hemoglobin in legumes (leghemoglobin). Both serve the same purpose of transporting oxygen in the organism. Though both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes? Other techniques for predicting protein structure include protein threading and *de novo* (from scratch) physics-based modeling.

Analysis of protein expression

Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is very much involved in making sense of protein microarray and HT MS data; the former approach faces similar problems as with microarrays targeted at mRNA, the latter involves the problem of matching large amounts of mass data against predicted masses from protein sequence databases, and the complicated statistical analysis of samples where multiple, but incomplete peptides from each protein are detected.

Analysis of regulation

Regulation is the complex orchestration of events starting with an extracellular signal such as a hormone and leading to an increase or decrease in the activity of one or more proteins. Bioinformatics techniques have been applied to explore various steps in this process. For example, promoter analysis involves the identification and study of sequence motifs in the DNA surrounding the coding region of a gene. These motifs influence the extent to which that region is transcribed into mRNA. Expression data can be used to infer gene regulation: one might compare microarray data from a wide variety of states of an organism to form hypotheses about the genes involved in each state. In a single-cell organism, one might compare stages of the cell cycle, along with various stress conditions (heat shock, starvation, etc.). One can then apply clustering algorithms to that expression data to determine which genes are co-expressed. For example, the upstream regions (promoters) of co-expressed genes can be searched for over-represented regulatory elements.

Analysis of mutations in cancer

In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways. Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancer. Bioinformaticians continue to produce specialized automated systems to manage the sheer volume of

sequence data produced, and they create new algorithms and software to compare the sequencing results to the growing collection of human genome sequences and germline polymorphisms. New physical detection technology are employed, such as oligonucleotide microarrays to identify chromosomal gains and losses (called comparative genomic hybridization), and single nucleotide polymorphism arrays to detect known *point mutations*. These detection methods simultaneously measure several hundred thousand sites throughout the genome, and when used in high-throughput to measure thousands of samples, generate terabytes of data per experiment. Again the massive amounts and new types of data generate new opportunities for bioinformaticians. The data is often found to contain considerable variability, or noise, and thus Hidden Markov model and change-point analysis methods are being developed to infer real copy number changes.

Another type of data that requires novel informatics development is the analysis of lesions found to be recurrent among many tumors.

Protein Modeling

The process of evolution has resulted in the production of DNA sequences that encode proteins with specific functions. In the absence of a protein structure that has been determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, researchers can try to predict the three-dimensional structure using protein or molecular modeling. This method uses experimentally determined protein structures (templates) to predict the structure of another protein that has a similar amino acid sequence (target).

Although molecular modeling may not be as accurate at determining a protein's structure as experimental methods, it is still extremely helpful in proposing and testing various biological hypotheses. Molecular modeling also provides a starting point for researchers wishing to confirm a structure through X-ray crystallography and NMR spectroscopy. Because the different genome projects are producing more sequences and because novel protein folds and families are being determined, protein modeling will become an increasingly important tool for scientists working to understand normal and disease-related processes in living organisms.

The Four Steps of Protein Modeling

i. Identify the proteins with known three-dimensional structures that are related to the target sequence
ii. Align the related three-dimensional structures with the target sequence and deter mine those structures that will be used as templates.
iii. Construct a model for the target sequence based on its alignment with the template structure(s).
iv. Evaluate the model against a variety of criteria to determine if it is satisfactory

Protein-protein docking

In the last two decades, tens of thousands of protein three-dimensional structures have been determined by X-ray crystallography and Protein nuclear magnetic resonance spectroscopy (protein NMR). One central question for the biological scientist is whether it is practical to predict possible protein-protein interactions only based on these 3D shapes, without doing protein-protein interaction experiments. A variety of methods have been developed to tackle the Protein-protein docking problem, though it seems that there is still much work to be done in this field.

Pharmaco and cheminformatics

Many aspects of bioinformatics are relevant in pharmacology. Drug targets in infectious organisms can be revealed by whole genome comparisons of infectious and non–infectious organisms. The analysis of single nucleotide polymorphisms reveals genes potentially responsible for genetic diseases. Prediction and analysis of protein 3D structure is used to develop drugs and understand drug resistance.

Drug Design

It is now possible, through computer algorithm based bioinformatic procedures, to identify and structurally modify a natural product, to design a drug with the desired properties and to assess its therapeutic effects, theoretically. Such procedures, similar to an architect's on board plan before construction, are described as *in silico* (in the computer, based on silicon chip technology), as opposed to the earlier *in vitro* (in experimental models) and *in vivo* (in clinical trials) methods. The risk involved in the earlier random processes of drug discovery methods is largely removed by bioinformatics. Cheminformatics involves organization of chemical data in a logical form to facilitate the process of understanding chemical properties, their relationship to structures and making inferences.

Drug Modification

Several synthetic products are quite useful but cannot be

used by one and all for certain side effects in some people. For example, aspartame (marketed under different trade names) is a dipeptide of aspartic acid and phenylalanine, and is 300 times sweeter than cane sugar. Aspartame is widely used as an alternate sweetener by diabetics and others who cannot take sweeteners loaded with calories. Unfortunately, pregnant women and people suffering from phenylketonuria, a disorder due to an impaired metabolism of phenylalanine, should not use aspartame. It would be useful if phenylalanine were substituted by some other amino acid without affecting its sweetness, to remove the restriction on its use.

Molecular phylogenies

Phylogeny is the origin and evolution of organisms. Biologists have constructed very elegant systems of classifications for the known organisms, though problems persist. Extensive work was carried out this way, comparing a very large number of organisms of plants and animals. Amino acid sequences and characteristics of proteins are also used in systematic.

Computational evolutionary biology

Evolutionary biology is the study of the origin and descent of species, as well as their change over time. New insight into the molecular basis of a disease may come from investigating the function of homologs of a disease gene in model organisms. In this case, homology refers to two genes sharing a common evolutionary history. Scientists also use the term homology, or homologous, to simply mean similar, regardless of the evolutionary relationship. Equally exciting is the potential for uncovering evolutionary relationships and patterns between different forms of life. With the aid of nucleotide and protein sequences, it should be possible to find the ancestral ties between different organisms. Thus far, experience has taught us that closely related organisms have similar sequences and that more distantly related organisms have more dissimilar sequences. Proteins that show a significant sequence conservation, indicating a clear evolutionary relationship, are said to be from the same protein family. By studying protein folds (distinct protein building blocks) and families, scientists are able to reconstruct the evolutionary relationship between two species and to estimate the time of divergence between two organisms since they last shared a common ancestor. This is connected with phylogenetics, which is the field of biology that deals with identifying and understanding the relationships between the different kinds of life on earth.

Hence, informatics has assisted evolutionary biologists in several key ways; it has enabled researchers to:

i. trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone,

ii. more recently, compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, horizontal gene transfer, and the prediction of factors important in bacterial speciation,

iii. build complex computational models of populations to predict the outcome of the system over time

iv. track and share information on an increasingly large number of species and organisms.

The area of research within computer science that uses genetic algorithms is sometimes confused with computational evolutionary biology, but the two areas are unrelated.

Measuring biodiversity

Biodiversity of an ecosystem might be defined as the total genomic complement of a particular environment, from all of the species present, whether it is a biofilm in an abandoned mine, a drop of sea water, a scoop of soil, or the entire biosphere of the planet Earth. Databases are used to collect the species names, descriptions, distributions, genetic information, status and size of populations, habitat needs, and how each organism interacts with other species. Specialized software programs are used to find, visualize, and analyze the information, and most importantly, communicate it to other people. Computer simulations model such things as population dynamics, or calculate the cumulative genetic health of a breeding pool (in agriculture) or endangered population (in conservation). One very exciting potential of this field is that entire DNA sequences, or genomes of endangered species can be preserved, allowing the results of Nature's genetic experiment to be remembered in silico, and possibly reused in the future, even if that species is eventually lost.

Modeling biological systems

This has to do with Systems biology .Systems biology involves the use of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes which comprise metabolism, signal transduction pathways and gene regulatory networks) to both analyze and visualize the complex connections of these cellular processes. Artificial life or virtual evolution attempts to understand evolutionary processes via the computer simulation of simple (artificial) life forms.

High-throughput image analysis

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content biomedical imagery. Modern image analysis systems augment an observer's ability to make measurements from a large or complex set of images, by improving accuracy, objectivity, or speed. A fully developed analysis system may completely replace the observer. Although these systems are not unique to biomedical imagery, biomedical imaging is becoming more important for both diagnostics and research. Some examples are:
i. high-throughput and high-fidelity quantification and sub-cellular localization (high-content screening, cytohistopathology).
ii. morphometrics
iii. clinical image analysis and visualization
iv. determining the real-time air-flow patterns in breathing lungs of living animals
v. quantifying occlusion size in real-time imagery from the development of and recovery during arterial injury
vi. making behavioral observations from extended video recordings of laboratory animals
vii. infrared measurements for metabolic activity determination
viii. inferring clone overlaps in DNA mapping, e.g. the Sulston score

Software and tools

Software tools for bioinformatics range from simple command-line tools, to more complex graphical programs and standalone web-services available from various bioinformatics companies or public institutions. The computational biology tool best-known among biologists is probably BLAST,an algorithm for determining the similarity of arbitrary sequences against other sequences, possibly from curated databases of protein or DNA sequences. BLAST is one of a number of generally available programs for doing sequence alignment. The NCBI provides a popular web-based implementation that searches their databases.

Web services in bioinformatics

SOAP and REST-based interfaces have been developed for a wide variety of bioinformatics applications allowing an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of the world. The main advantages lay in the end user not having to deal with software and database maintenance overheads. Basic bioinformatics services are classified by the EBI into three categories: SSS (Sequence Search Services), MSA (Multiple Sequence Alignment) and BSA (Biological Sequence Analysis). The availability of these service-oriented bioinformatics resources demonstrate the applicability of web based bioinformatics solutions, and range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible bioinformatics workflow management systems.

Partnership in bioinformatics

i. Data Gatherers: Enormous amounts of basic data from biomolecular chemistry and relatedareas, very painstakingly gathered over long years by experimental and analytical scientists, are the body and substance of bioinformatics.
ii .Data Processors: The second party use skills of complex software, to serve the needs of the data gathered.
iii. Process Product Users: End users of products.

Role of bioinformatics in biotechnology

The term ‘bioinformatics’ is the short form of ‘biological informatics’, just as biotechnology is the short form of ‘biological technology’. Anthony Kerlavage, of the Celera Genomics, defined bioinformatics as ‘Any application of computation to the field of biology, including data management, algorithm development, and data mining’. Clearly, a number of divergent areas, many of them outside biotechnology, come under bioinformatics. Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. Initial interest in Bioinformatics was propelled by the necessity to create databases of biological sequences.

The first database was created within a short period after the Insulin protein sequence was made available in 1956.The sequence information generated by the human genome research, initiated in 1988 has now been stored as a primary information source for future applications in medicine. The available data is so huge that if compiled in books, the data would run into 200 volumes of 1000 pages each and reading alone (ignoring understanding factor) would require 26 years working around the clock. For the population of about 5 billion human beings with two individuals differing in three million bases, the genomic sequence difference atabase would have about

15,000,000 billion entries. The present challenge to handle such a huge volume of data is to improve database design, develop software for database access and manipulation, and device data-entry procedures to compensate for the varied computer procedures and systems used in different laboratories. A single experiment can now yield data on the transcription level of 100,000 different mRNA species from a given tissue (Winzeler *et al.*, 1998).

The whole area of biology can immensely benefit from the bioinformatic approach. Bioinformatics tools for efficient research will have significant implications in life sciences and betterment of human lives.The rapidly emerging field of bioinformatics promises to lead to advances in understanding basic biological processes and, in turn, advances in the diagnosis, treatment, and prevention of many genetic diseases. Bioinformatics has transformed the discipline of biology from a purely lab-based science to an information science as well. Increasingly, biological studies begin with a scientist conducting vast numbers of database and Web site searches to formulate specific hypotheses or to design large-scale experiments. The implications behind this change, for both science and medicine, are staggering.

REFERENCES

Achuthsankar SN (2007). Computational Biology andBioinformatics-A gentle Overview, Communications of Computer Society of India.

Altschul SF,Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*:3389–3402

Aluru, Srinivas, (2000). *Handbook of Computational Molecular Biology*.Chapman and Hall/Crc, (Chapman and Hall/Crc Computer and Information Science Series).

Anthony K, Bonazzi V, M di T, Charles L, Peter L, Frank M, Richard M, Marc N, Mark Y, Jinghui Z, Paul T (2002).The Celera Discovery System. *Nucleic Acids Res*., *30*(1): 129-136.

Bairoch A, Apweiler R (2000).The SWISS–PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res. 28*:45–48 Barnes, M.R. and Gray, I.C.,(2003).*Bioinformatics for Geneticists*, first edition Wiley.

Baldi P, Brunak S (2001). *Bioinformatics: The Machine Learning Approach*, 2nd edition. MIT Press.

Baxevanis AD, Ouellette BFF (2005). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, third edition. Wiley. Baxevanis, A.D., Petsko, G.A., Stein, L.D., and Stormo, G.D(2007).*Current Protocols in Bioinformatics*.JohnWiley. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000). The Protein Data Bank.*Nucleic Acids Res.8*:235–42.

Cambridge University Press. Pearson WR (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.132:*185–219. Pevzner, P. A.(2000) *Computational Molecular Biology: An Algorithmic Approach* The MIT Press.

Claverie JM, Notredame C (2003). *Bioinformatics for Dummies*. Wiley. Critianini,N. and Hahn,M.(2006).*Introduction to ComputationalGenomics*, Cambridge University Press.

David W (2000). *Bioinformatics:Sequence and Genome Analysis* Spring Harbor Press. Pachter,L and Sturmfels,B (2005)."Algebraic Statistics for Computational Biology"

Dicks J, Anderson M, Cardle L, Cartinhour S, Couchman M, Davenport G, Dickson J, Gale M, Marshall D, May S, McWilliam H, O'Malia A, Ougham H, Trick M, Walsh S, Waugh R (2000). U.K CropNet: a collection of databases and bioinformatics resources for crop plantgenomics.*Nucleic Acid Res.,* 28 (1):104-107. Durbin,R.,Eddy,S.,Krogh,A.andMitchison,G.(1998).*Biological analysis*. Cambridge University Press.

Fleischmann RD (1995). Whole–genome random sequencing and assembly of haemophilus–influenzae.*Science269:*496–51.

Gilbert D (2004).Bioinformatics software resources.*Briefings in Bioinformatics,5:* 300-304.http://www.fbae.org/Channels/bioinformatics/BIOINFORMATICSN.*htmhttp://www.mrc-lmb.cam.ac.uk/genomes/madanm/articles/bioinfo.htm* Important projects*:* Species 2000 project; uBio Project;Partnership for Biodiversity Informatics.

Keedwell E (2005). *Intelligent Bioinformatics.*:*The Application of Artificial IntelligenceTechniques to Bioinformatics Problems*.John Wiley. Kohane, *et al*. (2002)*Microarrays for an Integrative Genomics.* The MIT Press. Lund, O. *et al*. (2005) *Immunological Bioinformatics.*

The Genome International Sequencing Consortium .(2001).Initial sequencing and analysis of the human genome.*Nature 409*:860–921

Tisdall J (2001). "Beginning Perl for Bioinformatics" O'Reilly.

The MIT Press. McCullough MJ, Mc-Cusker JH, Stevens DA, Wodicka L, Lockhart DJ, Davis RW (1998). Direct Allelic variation scanning of the Yeast *Genome.Science, 281:* 1194-1197.

Tramontano, Christian DD, B Alfonso, V David, B David, M Christophe, L Isidore R; Chris, S; Christos, A. O. (2003). Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics* 19(6): 717- 726. Venter,J.C,.(2001).The sequence of the human genome.*Science 291*:1304–1351.

Waterman MS (1995). *Introduction to Computational Biology: Sequences, Maps and Genomes*. CRC Press.