BIOINFORMATICS

A SEMINAR REPORT

Submitted by

NITHYA K. PILLAI

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF ENGINEERING

COCHIN UNIVERSITY OF SCIENCE &TECHNOLOGY,

KOCHI-682022

AUGUST 2008

DIVISION OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF ENGINEERING

COCHIN UNIVERSITY OF SCIENCE & TECHNOLOGY,

COCHIN-682 022

Certificate

Certified that this is a bonafide record of the Seminar Entitled

"BIOINFORMATICS"

Done by the following Student

Nithya K Pillai

Of the VIIth semester, Computer Science and Engineering in the year 2008 in partial fulfillment of the requirements to the award of Degree Of Bachelor Of Technology in Computer Science and Engineering of Cochin University of Science and Technology.

Mrs.Rahna P. Muhammed — *Seminar Guide*

Dr. David Peter — *Head of the Department*

Date:

ACKNOWLEDGEMENT

At the outset, I thank the Lord Almighty for the grace, strength and hope to make my endeavor a success.

I also express my gratitude to **Dr. David Peter, Head of the Department** and my Seminar Guide for providing me with adequate facilities, ways and means by which I was able to complete this seminar. I express my sincere gratitude to him for his constant support and valuable suggestions without which the successful completion of this seminar would not have been possible.

I thank **Mrs.Rahna P. Muhammed,** my seminar guide for her boundless cooperation and helps extended for this seminar. I express my immense pleasure and thankfulness to all the teachers and staff of the Department of Computer Science and Engineering, CUSAT for their cooperation and support.

Last but not the least, I thank all others, and especially my classmates and my family members who in one way or another helped me in the successful completion of this work.

NITHYA K.PILLAI

ABSTRACT

Rapid advances in bioinformatics are providing new hopes to patients of life threatening diseases. Gene chips will be able to screen heart attack and diabetics years before patients develop symptoms. In near future, patients will go to a doctor's clinic with lab- on- a- chip devices. The device will inform the doctor in real time if the patient's ailment will respond to a drug based on his DNA. These will help doctors diagnose life-threatening illness faster, eliminating expensive, time-consuming ordeals like biopsies and sigmoidoscopies. Gene chips reclassify diseases based on their underlying molecular signals, rather than misleading surface symptoms. The chip would also confirm the patient's identity and even establish paternity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.INTRODUCTION

Bioinformatics is an inter disciplinary research area. It is a fusion of computing, biotechnology and biological sciences. Bioinformatics is poised to one of the most prodigious growth areas in the next to decades. Being the interface between the most rapidly advancing fields of biological and computational sciences, it is immense in scope and vast in applications.

Bioinformatics is the study of biological information as it passes from its storage site in the genome to the various gene products in the cell. Bioinformatics involves the creation and computational technologies for problems in molecular biology. As such, it deals with methods for storing, retrieving and analyzing biological data, such as nuclei acid (DNA/RNA) and protein sequence, structures, functions, path ways and interactions. The science of Bioinformatics, which is the melding of molecular biology with computer science, is essential to the use of genomic information in understanding human diseases and in the identification of new molecular targets of drug discovery. New discoveries are being made in the field of genomics, an area of study which looks at the DNA sequence of an organism in order to determine which genes code for beneficial traits and which genes are involved in inherited diseases.

If you are not tall enough, the stature could be altered accordingly. If you are weak and not strong enough, your physique could be improved. If you think this is the script for a science fiction movie, you are mistaken. It is the future reality.

2. EVOLUTION OF BIOINFORMATICS

DNA is the genetic material of organism. It contains all the information needed for the development and existence of an organism. The DNA molecule is formed of two long polynucleotide chains which are spirally coiled on each other forming a double helix. Thus it has the form of spirally twisted ladder. DNA is a molecule made from sugar, phosphate and bases. The bases are guanine(G), cytosine(C), adenine(A) and thiamine(T).Adenine pairs only with Thiamine and Guanine pairs only with Cytosine. The various combinations of these bases make up with DNA. That is; AAGCT, CCAGT, TACGGT etc. An infinite number of combinations of these bases is possible. And then the gene is a sequence of DNA that represents a fundamental unit of heredity. Human genome consists of approximately 30,000 genes, containing approximately 3 billion base pairs.

Currently, scientists are trying to determine the entire DNA sequence of various living organisms. DNA sequence analysis could identify genes, regulatory sequences and other functions. Molecular biology, algorithms, and computing have helped in sequencing larger portions of genomics of several species. Sequence is the determination of the order of nucleotides in a DNA as also the order of amino acids in a protein. Sequence analysis, which is at the core of bioinformatics, enables function identification of genes.

The human found in every cell of a human being consists of 23pairs of chromosomes. These chromosomes constitute the 3 billion letters of chemical code that specify the blue print for a human being. Human Genome Project, one of the best known projects in the world. The world Human Genome Project, a vast endeavor aimed at reading this entire DNA code will completely transform biology, medicine and biotechnology. Using this entire code all 30,000 human genes will be identified; all 5000 inherited diseases will become diagnosable and potentially curable; and drug design will be completely transformed. The Genome Project focuses on two main objective:

mapping-pinpointing the genomic location of all genes and markers; and DNA sequencing-reading the chemical "text" of all the genes and their intervening sequences. DNA sequences are entered in to large data bases, where they can be compared with the known genes, including inter-species comparisons. The explosion of publicly available genomic information resulting from the Human Genome Project has precipitated the need for bioinformatics capabilities.

Determination of genome organization and gene regulation will promote the understanding of how humans develop from single cells to adults, why this process some times goes wrong, and the changes that take place as people age. Bioinformatics finds applications in medicine for recommending individually tailored drugs based on an individual's profile. It helps to identify a specific genetic sequence that is responsible for a particular disease, its associated protein, and protein function. For curing the disease a new drugs can be developed.

3. HUMAN ELECTRONICS

The nucleus is the most obvious organelle in the human cell. Within the nucleus is the DNA responsible for providing the cell with its unique characteristics. The DNA is similar in every cell of the body, but depending on the specific cell type; some genes may be turned on or off-that is why a liver cell is different from a muscle cell, and a muscle cell is different from a fat cell. About 99.9% of the sequence is identical between any two people. But because the small percentage of DNA that differs can relate to an individual's disease. Scientists are comparing sequence using DNA chips from healthy people and those from patients with a specific disease to help identify genetic targets for drug discovery information about genetic variation can help to predict which patients are likely to benefit from specific drugs

The most significant and the biggest application of DNA chips is the use of DNA micro arrays for expression profiling. In expressions profiling the chip controls how different parts of the genes turned on or off to create certain types of cells. If the gene is expressed in one way, it may result in normal muscle, for instance. If it is expressed in another way, it may result in a tumor. By comparing these different expressions, researchers hope to discover ways to predict and perhaps to prevent diseases.

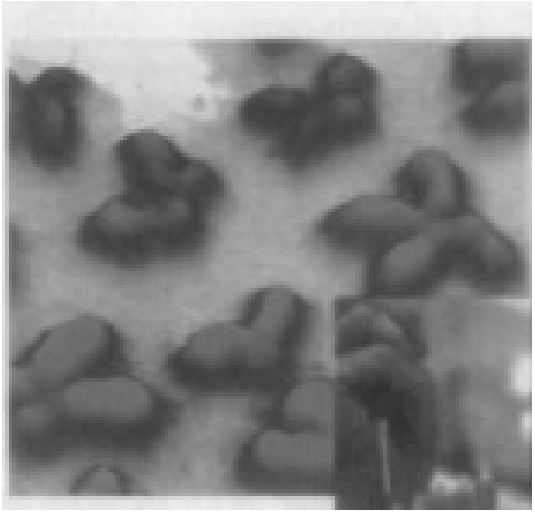


Fig 1.DNA

Electronic circuit can be incorporated in the chip to detect various states of DNA. DNA carries an electric charge. That charge can be read on the chip, just like cells on a memory array. This DNA chip would like to diagnose life-threatening bacterial infections.

In DNA the medium is a chain of two units (phosphate & ribose), and the most easily recognizable message is provided by a sequence of letters (bases) attached to the chain. The DNA has two sequences of letters wrapped in the form of a double helix. The DNA has two sequences of letters wrapped around each other in the form of a double helix. One is the complement of other, so that the sequence of one string (strand) can be inferred from the sequence of other. The DNA sequence of bases encodes 20 amino acids. Under instructions received from DNA, amino acids join together in the same order as they are encoded in DNA to form proteins. Chains of amino acids, which fold in complicated ways, play a major role in determining how we interact with the environment.

Genomic information is revolutionizing life sciences. The quest for under standing how genetic factors contribute to human disease is gathering speed. The 46 human chromosomes house almost three billion base pairs of DNA that contain 30,000 to 40,000 protein-coding genes. Using bioinformatics find out how genes contribute to diseases that have a complex pattern of inheritance, such as diabetics, asthma, and mental illness. No one gene can tell whether a person has a disease or not. A number of genes may make a subtle contribution to a person's susceptibility to a disease. Gene may also affect how a person reacts to the environment. As the entire human genome is too big a sequence on its own, sequencing and reading a genome demand heavy computational resources.

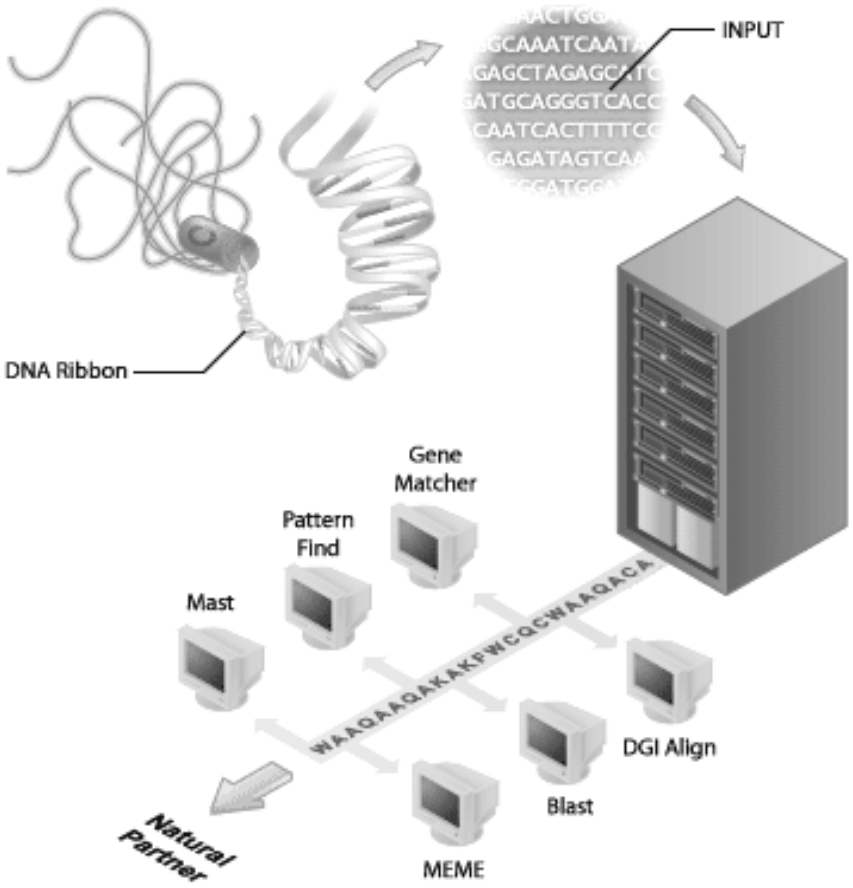
Bioinformatics is largely, although not exclusively, a computer-based discipline. Computers are important in bioinformatics for two reasons:

First, many bioinformatics problems require the same task to be repeated millions of times. For example, comparing a new sequence to every other sequence stored in a database or comparing a group of sequences systematically to determine evolutionary

relationships. In such cases, the ability of computers to process information and test alternative solutions rapidly is indispensable.

Second, computers are required for their problem-solving power.

Typical problems that might be addressed using bioinformatics could include solving the folding pathways of protein given its amino acid sequence, or deducing a biochemical pathway given a collection of RNA expression profiles. Computers can help with such problems, but it is important to note that expert input and robust original data are also required.



.

Fig 2. Involvement of computers

We start with an overview of the sources of information: these may be divided into raw DNA sequences, protein sequences, macromolecular structures, genome sequences, and other whole genome data. Raw DNA sequences are strings of the four baseletters comprising genes, each typically 1,000 bases long. The GenBank repository of nucleic

acid sequences currently holds a total of 9.5 billion bases in 8.2 million entries (all database figures as of August 2000). At the next level are protein sequences comprising strings of 20 amino acid-letters. At present there are about 300,000 known protein sequences.

Data source	**Data size**	**Bioinformatics topics**
Raw DNA sequence	8.2 million sequences (9.5 billion bases)	Separating coding and non-coding regions Identification of introns and exons Gene product prediction Forensic analysis
Protein sequence	300,000 sequences (~300 amino acids each)	Sequence comparison algorithms Multiple sequence alignments algorithms Identification of conserved sequence motifs
Macromolecular structure	13,000 structures (~1,000 atomic coordinates each)	Secondary, tertiary structure prediction 3D structural alignment algorithms Protein geometry measurements Surface and volume shape calculationsIntermolecular interactions
Genomes	40 complete genomes (1.6 million – 3 billion bases each)	Characterisation of repeats Structural assignments to genes Phylogenetic analysis Genomic-scale censuses (characterisation of protein content, metabolic pathways) Linkage analysis relating specific genes to diseases
Gene expression	largest: ~20 time point measurements for ~6,000 genes	Correlating expression patterns Mapping expression data to sequence, structural and biochemical data

Table 1. Sources of data used in bioinformatics, the quantity of each type of data that is currently available, and bioinformatics subject areas that utilise this data.

4.GENE EXPRESSION

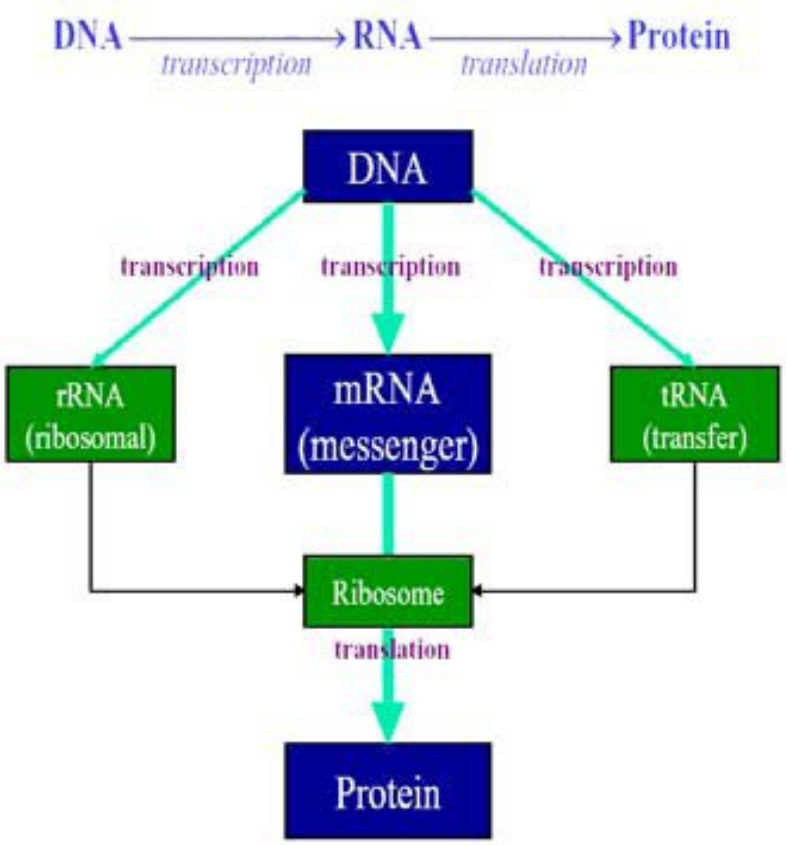


Fig.3 Gene Expression

mRNA encodes the genetic information as copied from the DNA molecules. Transcription is the process in which DNA is copied into an RNA molecule. The resulting linear molecule is an mRNA transcript. tRNA molecules develop a well-defined three-dimensional structure which is critical in the creation of proteins. Translation is the process in which the nucleotide base sequence of the processed mRNA is used to order and join the amino acids into a protein with the help of ribosomes and tRNA.

The 3D structure of proteins is mainly determined by X-ray crystallography and by nuclear magnetic resonance (NMR). It is time consuming and costly.

A powerful new tool available in biology is microarrays. They allow determining simultaneously the amount of mRNA production of thousands of genes. Microarray experiments require threephases. In the first phase one places thousands of different one-stranded chunks of RNA in minuscule wells on the surface of a small glass chip. (This task is not unlike that done by a jet printer using thousands of different colors and placing each of them in different spots of a surface.) The chunks correspond to the RNA known to have been generated by a given gene. The 2D coordinates of each of the wells are of course known.

The second phase consists of spreading—on the surface of the glass— genetic material (again one-stranded RNA) obtained by a cell experiment one wishes to perform. Those could be the RNAs produced by a diseased cell, or by a cell being subjected to starvation, high temperature, etc. The RNA already in the glass chip combines with the RNA produced by the cell one wishes to study. The degree of combined material obtained by complementing nucleotides is an indicator of how much RNA is being expressed by each one of the genes of the cell being studied.

The third phase consists of using a laser scanner connected to a computer. The apparatus measures the amount of combined material in each chip well and determines the degree of gene expression—a real number—for each of the genes originally placed on the chip. Microarray data is becoming available in huge amounts. A problem with this data is that it is noisy and its interpretation is difficult. Microarrays are becoming invaluable for biologists studying how genes interact with each other. This is crucial in understanding disease mechanisms.

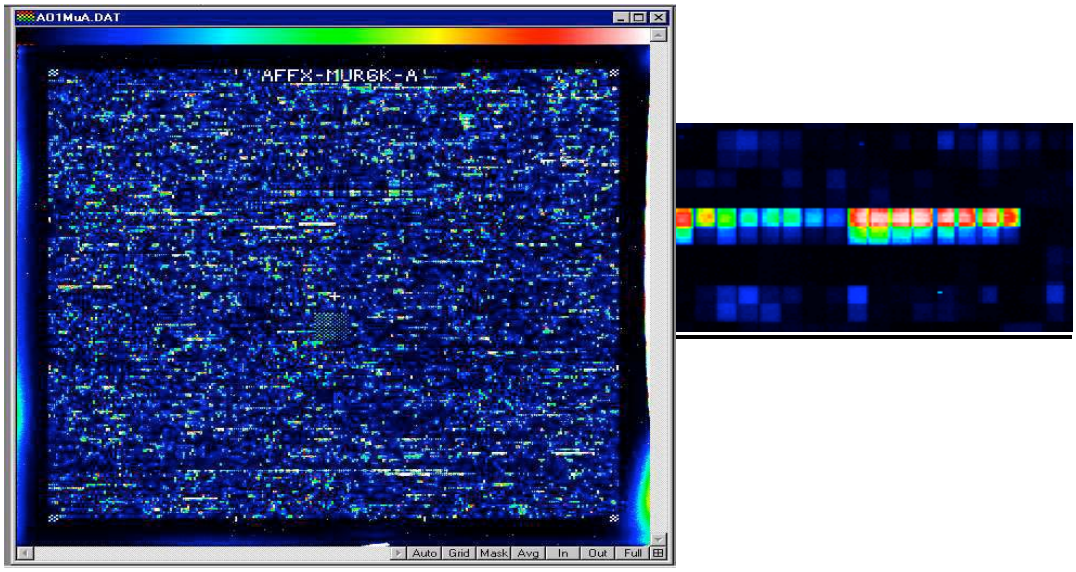


FIG.4 MICROARRAY

The most significant and the biggest application of DNA chips is the use of DNA micro arrays for expression profiling. In expressions profiling the chip controls how different parts of the genes turned on or off to create certain types of cells. If the gene is expressed in one way, it may result in normal muscle, for instance. If it is expressed in another way, it may result in a tumor. By comparing these different expressions, researchers hope to discover ways to predict and perhaps to prevent diseases.

5. CHIP ELECTRONICS

Chip electronics can be divided into two.

5.1. Biochips

5.2. Clinical chip

5.1.BIOCHIP

Biochip is an IC who's electrical and logical functions are performed by protein molecules appropriately manipulated. Advances in molecular biology and semiconductor fabrication have resulted in new formats for hybridization arrays. Instead of these being based on a membrane or a glass slide platforms these arrays several electrodes covered by a thin layer of agarose coupled with affinity moiety. Each micro electrode is capable of generating a controllable electric current that can be used to draw biological samples, reagents and probes to specify locations on the chip surface. The number of genes covered by these arrays depends on the number of electrodes made within the area of that array.

5.2. CLINICAL CHIPS

A decade ago, an eight-year old kid jumped from his swing set and landed flat, shattering a leg bone where most kids would have sprained an ankle. An X-ray revealed this problem. Where there should have been hard bone, a soft tumour was present. The kid needed a precise diagnosis. If the cancer was aggressive, it needed immediate treatment with the powerful but toxic drug 'adriamycin'. If the tumour was growing slowly, doctors had the time to try out weaker but safer drugs.

A biopsy was inconclusive. Like many paediatric bone tumours, the kid's tumour was a small, round blue- cell tumour. The doctor had a problem treating the kid. As "adriamycin" could cause serious heart damage, doctors weren't willing to give it to the kid. Of all blue cell tumours spreads aggressively enough to require this potentially deadly medicine. Doctors hoped that a less toxic medicine will do and gave the same to

the kid, resulting in the death of the kid just after six months. Today, rapid advances in bioinformatics are providing new hopes to such patients. The new technology enables doctors to proceed straight to genetic codes that instruct tumours to grow, finding invisible molecular signals that differentiate cancers as well as a host of other deadly diseases.

The key to this life saving, cost effective diagnostic power is a tiny glass chip peppered with DNA strips, called the gene chip. Today, 60% of gene chips are used for are research purposes, where these are speeding up drug design and helping researchers to mine genomic data bases.

6. PRESENT GOALS OF BIOINFORMATICS

The present role of bioinformatics is to aid biologists in gathering and processing genomic data to study protein function. Another important role is to aid researchers at pharmaceutical companies in making detailed studies of protein structures to facilitate drug design. Typical tasks done in bioinformatics include:

—*Inferring a protein's shape and function from a given a sequence of amino acids*.

—*Finding all the genes and proteins in a given genome.*

—*Determining sites in the protein structure where drug molecules can be attached.*

To perform these tasks, one usually has to investigate *homologous* sequences or proteins for which genes have been determined and structures are available. Homology between two sequences (or structures) suggests that they have a common ancestor. Since those ancestors may well be extinct, one hopes that similarity at the sequence or structural level is a good indicator of homology.

7. ALGORITHMS FREQUENTLY USED IN BIOINFORMATICS

Based on the availability of the data now present the various algorithms that lead to a better understanding of gene function. They can be summarized as follows:

(1) COMPARING SEQUENCES.

Given the huge number of sequences available, there is an urgent need to develop algorithms capable of comparing large numbers of long sequences. These algorithms should allow the deletion, insertion, and replacements of symbols representing nucleotidesor amino acids, for such transmutations
occur in nature.

(2) CONSRUCTING EVOLUTIONARY TREES

These trees are often constructed after comparing sequences belonging to different organisms. Trees group the sequences according to their degree of similarity. They serve as a guide to reasoning about how these sequences have been transformed through evolution. For example, they infer homology from similarity, and may rule out erroneous assumptions that contradict known evolutionary
processes.

(3) DETECTING PATTERNS IN SEQUENCE

There are certain parts of DNA and amino acid sequences that need to be detected. Two prime examples are the search for genes in DNA and the determining of subcomponents of a sequence of amino acids
(secondary structure). There are several ways to perform these tasks. Many of them are based on machine learning and include probabilistic grammars, or neural networks.

(4) DETERMINING 3D STRUCTURES FROM PATTERNS

The problems in bioinformatics that relate sequences to 3D structures are computationally difficult. The determination of RNA shape from sequences requires

algorithms of cubic complexity. The inference of shapes of proteins from amino acid sequences remains an unsolved problem.

7.1. Comparing Sequences

From the biological point of view sequence comparison is motivated by the fact that all living organisms are related by evolution. That implies that the genes of species that are closer to each other should exhibits similarities at the DNA level; one hopes that those similarities also extend to gene function.

The following definitions are useful in understanding what is meant by the comparison of two or more sequences. An *alignment* is the process of lining up sequences to achieve a maximal level of identity. That level expresses the degree of similarity between sequences. Two sequences are *homologous* if they share a common ancestor, which is not always easy to determine. The degree of similarity obtained by alignment can be useful in determining the possibility of homology between two sequences.
In biology, the sequences to be compared are either nucleotides (DNA, RNA) or amino acids (proteins). In the case of nucleotides, one usually aligns identical nucleotide symbols. When dealing with amino acids the alignment of two amino acids occurs if they are identical or if one can be derived from the other by substitutions that are likely to occur in nature. An alignment can be either *local* or *global*. In the former, only portions of the sequences are aligned, whereas in the latter one aligns over the entire length of the sequences. Usually, one uses gaps, represented by the symbol "-", to indicate that it is preferable not to align two symbols because in so doing, many other pairs can be aligned. In local alignments there are
larger regions of gaps. In global alignments, gaps are scattered throughout the alignment. A measure of likeness between two sequences is *percent identity*: once an alignment is performed we count the number of columns containing identical symbols. The percent identity is the ratio between that number and the number of symbols in the (longest) sequence. A possible measure or *score* of an alignment is calculated by summing up the matches of identical (or similar) symbols and counting gaps as negative.

With these preliminary definitions in mind, we are ready to describe the algorithms that are often used in sequence comparison.

7.1.1. Pairwise Alignment. Many of the methods of pattern matching used in computer science assume that matches contain no gaps. Thus there is no match for the pattern *bd* in the text *abcd*. In biological sequences, gaps are allowed and an alignment *abcd* with *bd* yields the representation:

a b c d

− b − d.

Similarly, an alignment of *abcd* with *buc* yields:

a b − c d

− b u c −.

The above implies that gaps can appear both in the text and in the pattern. Therefore there is no point in distinguishing texts from patterns. Both are called sequences. Notice that, in the above examples, the alignments maximize matches of identical symbols in both sequences. Therefore, sequence alignment is an optimization problem. A similar problem exists when we attempt to automatically correct typing errors like character replacements, insertions, and deletions. Google and Word, for example, are able to handle some typing errors and display suggestions for possible corrections. That implies searching a dictionary for best matches.

7.1.2. Aligning Amino Acids Sequences. The DP algorithm is applicable to any sequence provided the weights for comparisons and gaps are properly chosen. When aligning nucleotide sequences the previously mentioned weights yield good results. A more careful assessment of the weights has to be done when aligning sequences of amino acids. This is because the comparison between any two amino acids should take evolution into consideration. Biologists have developed 20×20 triangular matrices that provide the weights for comparing identical and different amino acids as well as the weight that should be attributed to gaps. The two more frequently used matrices are known as PAM (Percent Accepted Mutation) and BLOSUM (Blocks Substitution Matrix). These matrices reflect the weights obtained by comparing the amino acids substitutions that have occurred through evolution. They are often called substitution matrices.

7.1.3. Complexity Considerations and BLAST. The quadratic complexity of the Dpbased algorithms renders their usage prohibitive for very large sequences. Recall that the present genomic database contains about 30 billion base pairs (nucleotides) and thousands of users accessing that database simultaneously would like to determine if a sequence being studied and made up of thousands of symbols can be aligned with existing data. That is a formidable problem! The program called BLAST (Basic Local Alignment Search Tool) developed by the National Center for Biotechnology Information (NCBI) has been designed to meet that challenge. The best way to explain the workings of BLAST is to recall the approach using dot matrices. In BLAST the sequence, whose presence one wishes to investigate in a huge database, is split into smaller subsequences. The presence of those subsequences in the database can be determined efficiently (say by hashing and indexing).

7.2. Phylogenetic Trees

Since evolution plays a key role in biology, it is natural to attempt to depict it using trees. These are referred to as phylogenetic trees: their leaves represent various organisms, species, or genomic sequences; an internal node *P*stands for an abstract organism (species, sequence) whose existence is presumed and whose evolution led to the organisms whose direct descendants are the branches emanating from *P*. A motivation for depicting trees is to express—in graphical form—the outcome of multiple alignments by the relationships that exist between pairs or groups of sequences. These trees may reveal evolutionary inconsistencies that have to be resolved. In that sense the construction of phylogenetic validates or invalidates conjectures made about possible ancestors of a group of organisms.

There are several types of trees used in bioinformatics. Among them, we mention the following:

(1) ***Unrooted*** trees are those that specify distances (differences) between species. The length of a path between any two leaves represents the accumulated differences.

(2) ***Cladograms*** are ***rooted trees*** in which the branches' lengths have no meaning; the initial example in this section is a cladogram.

(3) ***Phylograms*** are extended cladograms in which the length of a branch quantifies the number of genetic transformations that occurred between a given node and its immediate ancestor.

(4) ***Ultrametric*** trees are phylograms in which the accumulated distances from the root to each of the leaves is quantified by the same number; ultrametric trees are therefore the ones that provide most information about evolutionary changes. They are also the most difficult to construct. The above definitions suggest establishing some sort of molecular clock in which mutations occur at some predictable rate and that there exists a linear relationship between time and number of changes.

7.3. Finding Patterns in Sequences

It is frequently the case in bioinformatics that one wishes to delimit parts of sequences that have a biological meaning. Typical examples are determining the locations of promoters, exons, and introns in RNA, that is, gene finding, or detecting the boundaries of α-helices, β-sheets, and coils in sequences of amino acids. There are several approaches for performing those tasks. They include neural nets, machine learning, and grammars, especially variants of grammars called probabilistic In this subsection, we will deal with two of such approaches. One is using grammars and parsing. The other, called Hidden Markov Models or HMMs, is a probabilistic variant of parsing using finite-state grammars. It should be remarked that the recent capabilities of aligning entire genomes also provides means for gene finding in new genomes: assuming that all the genes of a genome $G1$ have been determined, then a comparison with the genome $G2$ should reveal likely positions for the genes in $G2$.

7.3.1. Grammars and Parsing. Chomsky's language theory is based on grammar rules used to generate sentences. In that theory, a nonterminal is an identifier naming groups of contiguous words that may have subgroups identified by other nonterminals. In the Chomsky hierarchy of grammars and languages, the finite-state (FS) model is the lowest. In that case, a nonterminal corresponds to a state in a finite-state automaton. In context-free grammars one can specify a potentially infinite number of states. Context-free grammars (CFG) allow the description of palindromes or matching parentheses, which

cannot be described or generated by finite-state models. Higher than the context-free languages are the so-called contex sensitive ones (CSL). Those can specify repetitions of sequence of words like *ww*, where *w* is any sequence using a vocabulary. These repetitions cannot be described by CFGs. Parsing is the technique of retracing the generation of a sentence using the given grammar rules. The complexity of parsing depends on the language or grammar being considered. Deterministic finite-state models can be parsed in linear time. The worst case parsing complexity of CF languages is cubic. Little is known about the complexity of general CS languages but parsing of its strings can be done in finite time. The parse of sentences in a finite-state language can be represented by the sequence of states taken by the corresponding finite-state automaton when it scans the input string. A tree conveniently represents the parse of a sentence in a context-free language. Finally, one can represent the parse of sentence in a CSL by a graph. Essentially, an edge of the graph denotes the symbols (or nonterminals) that are grouped together.

In what follows, we briefly describe the types of patterns that are necessary to detect genes in DNA.nonterminal *G*, defined by the rules below, can roughly describe the syntax of genes:

$G \rightarrow PR$

$P \rightarrow N$

$R \rightarrow EIR|E$

$E \rightarrow N$

$I \rightarrow gtNag$,

where *N* denotes a sequence of nucleotides *a , c , g , t* ; *E* is an exon, *I* an intron, *R* a sequence of alternating exons and introns and *P* is a promoter region, that is, a heading announcing the presence of the gene. In this simplified grammar, the markers *gt* and *ag* are delimiters for introns. Notice that it is possible to transform the above CFG into an equivalent FSG since there is a regular expression that defines the above language. But the important remark is that the grammar is highly ambiguous since the markers *gt* or *ag* could appear anywhere within an exon an intron or in a promoter region. Therefore, the grammar is descriptive but not usable in constructing a parser. The alternation exons-

introns can be interpreted in many different ways, thus accounting for the fact that a given gene may generate alternate proteins depending on contexts.

7.3.2. Hidden Markov Models *(HMMs).* HMMs are widely used in biological sequence analysis. HMMs can be viewed as variants of *probabilistic or stochastic* finite-state transducers (FSTs). In an FST, the automaton changes states according to the input symbols being examined. On a given state, the automaton also outputs a symbol. Therefore, FSTs are defined by sets of states, transitions, and input and output vocabularies. There is as usual an initial state and one or more final states. The automata that we are dealing with *can be and usually are nondeterministic.* Therefore, upon examining a given input symbol, the transition depends on the specified probabilities. An HMMis a probabilistic FST in which there is also a set of pairs [*p*, *s*] associated to each state; *p* is a probability and *s* is a symbol of the output vocabulary. The sum of the *p*'s in each set of pairs within a given state also has to equal 1. One can assume that the input vocabulary for an

HMM consists of a unique dummy symbol (say, the equivalent of an empty symbol). Actually, in the HMM paradigm, we are solely interested in state transitions and output symbols. As in the case of finitestate automata, there is an initial state and a final state. Upon reaching a given state, the HMM automaton produces the output symbol *s* with a probability *p*. The *p*'s are called emission probabilities. As we described so far, the HMM behaves as a string generator.

The main usage of HMMs is in the reverse problem: recognition or parsing. Given a sequence of *H*'s and *T*'s, attempt to determine *the most likely* corresponding state sequence of *F*'s and *L*'s.

7.4. Determining Structure

From the beginning of this article, we reminded the reader of the importance of structure in biology and its relation to function. In this section, we review some of the approaches that have been used to determine 3D structure from linear sequences. A particular case of structure determination is that of RNA, whose structure can be approximated in two dimensions. Nevertheless, it is known that 3D knot-like structures exist in RNA. This

section has two subsections. In the first, we cover some approaches available to infer 2D representations from RNA sequences. In the second, we describe one of the most challenging problems in biology: the determination of the 3D structure of proteins from sequences of amino acids. Both problems deal with minimizing energy functions.

7.4.1. RNA Structure. It is very convenient to describe the RNA structure problem in terms of parsing strings generated by context-free-grammars (CFG). As in the case of finite-state automata used in HMMs we have to deal with highly ambiguous grammars. The generated strings can be parsed in multiple ways and one has to choose an optimal parse based on energy considerations. RNA structure is determined by the attractions among its nucleotides: A (adenine) attracts U (uracil) and C (cytosine) attracts G (guanine). These nucleotides will be represented using small case letters. The CFG rules:

$S \rightarrow aSu/uSa/\varepsilon$

generate palindrome-like sequences of *u*'s and *a*'s of even length. One could map this palindrome to a 2D representation in which each *a* in the left of the generated string matches the corresponding *u* in the right part of the string and viceversa. In this particular case, the number of matches is maximal. This grammar is nondeterministic since a parser would not normally know where lies the middle of the string to be parsed. The grammar becomes highly ambiguous if we introduce a new nonterminal *N* generating any sequence of *a*'s and *u*'s. $S \rightarrow aSu/uSa/N\ N \rightarrow aN/uN/\varepsilon$. Now the problem becomes much harder since any string admits a very a large number of parses and we have to chose among all those parses the one that matches most *a*'s with *u*'s and vice versa. The corresponding 2D representation of that parse is what is called a *hairpin loop.* An actual grammar describing RNA should also include the rules specifying the attractions among *c*'s and *g*'s:

$S \rightarrow cSg/gSc/.$

7.4.2. Protein Structure. The largest repository of 3D protein structures is the PDB (Protein Data Base): it records the actual *x*, *y*, *z* coordinates of each atom making up each of its proteins. That information has been gathered mostly by X-ray crystallography and NMR techniques. There are very valuable graphical packages (e.g., Rasmol) that can present the dense information in the PDB in a visually attractive and useful form allowing the user to observe a protein by rotating it to inspect its details viewed from

different angles. The outer surface of a protein consists of the amino acids that are *hydrophilic* (tolerate well the water media that surrounds the protein). In contrast, the *hydrophobic* amino acids usually occupy the protein's core. The configuration taken by the protein is one that minimizes the energy of the various attractions and repulsions among the constituent atoms.

A *domain* is a portion of the protein that has its own function. Domains are capable of independently folding into a stable structure. The combination of domains determines the protein's function. Protein folding, the determination of protein structure from a given sequence of amino acids, is one of the most difficult problems in present-day science. The approaches that have been used to solve it can only handle short sequences and require the capabilities of the fastest parallel computers available.

8.APPLICATIONS

Biochips can be mainly classified into two based on the applications:

1. Internal biochips
2. External biochips

Applications of internal biochips are

1. Glucose measurement
2. Brain surgery for Parkinson's disease
3. Cochlear implant
4. Eye implant
5. Personal identification

Applications of external biochips are

1. lab on a chip
2. mass spectrometry

8.1 APPLICATION OF INTERNAL BIOCHIPS

1. GLUCOSE MEASUREMENT

Nowadays diabetics measure the level of sugar glucose in their blood by using a skin prick and a hand held blood test and medicate them with insulin. The disadvantage of this simple system is that the need to draw blood makes the diabetics not to test the sugar levels themselves as often as they could.

By using Biochips the measurement can be done in a much simpler way. The chips are of size less than an uncooked grain of rice can be injected under the skin. It sense the glucose level and send the result back out by radio frequency communication.

2. BRAIN SURGERY FOR PARKINSON'S DISEASE

Parkinson's disease is caused by a brain messenger dopamine, which is a product of dying brain cells. This disease causes uncontrolled movements or tremors on body parts.

Drug therapy for Parkinson's disease aims to replace dopamine but the drugs effect wear off after some time. This causes the erratic movements coming back to the patients.

Activa implant is a biochip which uses high frequency electrical pulses to reversibly shut off the thalamus. These chips turn off brain signals that cause the uncontrolled movements or tremors. The implantation surgery is far simpler. Electrodes will be entered in to the thalamus region whose extension is connected to the pulse generator placed near the chest. The pulse generator generates pulses according to the heart beat of the patient.

If there are any post operative problems the simulator (pulse generator) can be simply turned off.

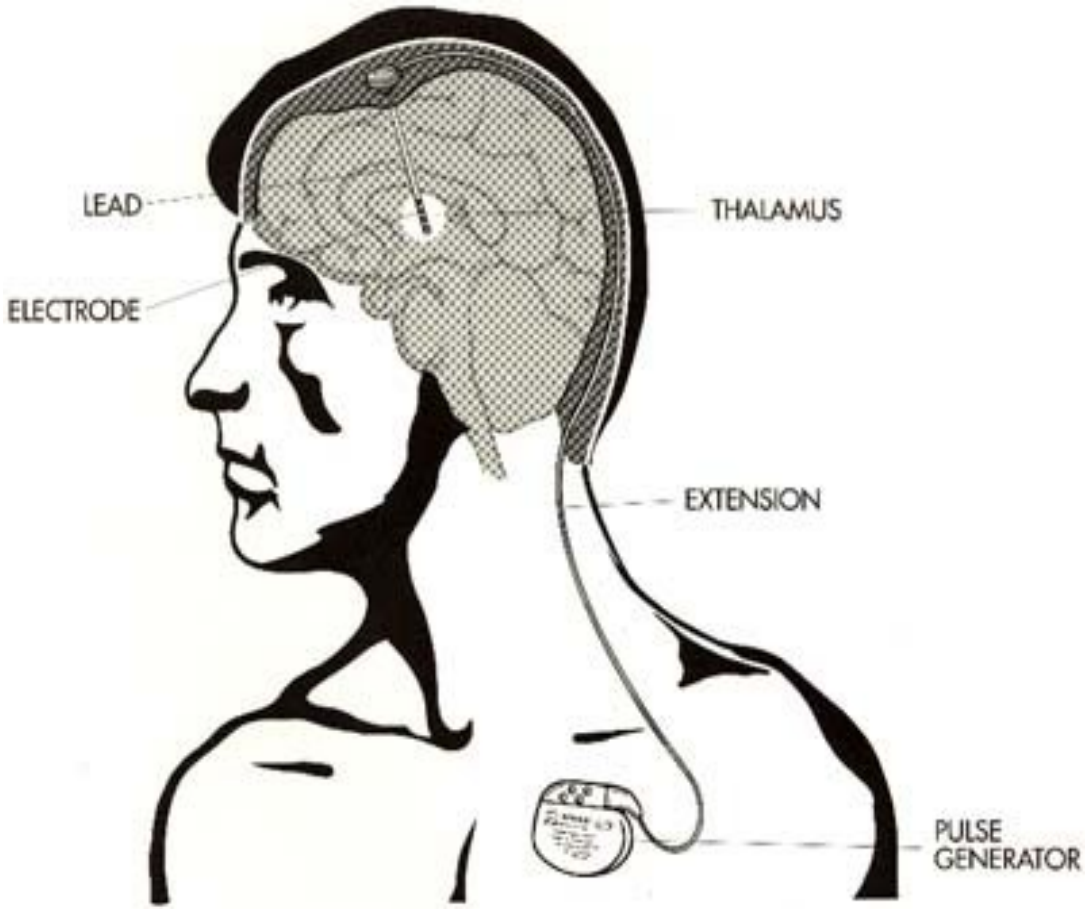


FIG.2. ACTIVA IMPLANT

3. COCHLEAR IMPLANT

Hearing aids used in present days are glorified amplifiers, but the cochlear implant is for patients who have lost the hair cells that detect sound waves. For these individuals no amount of amplification is enough.

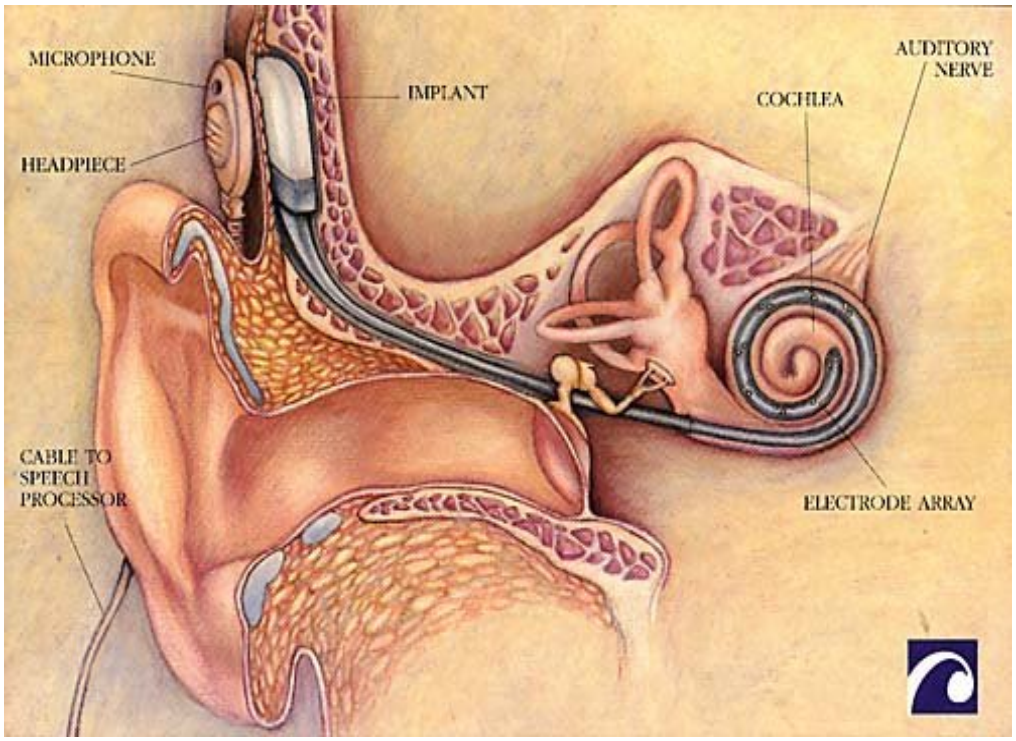


Fig.3. cochlear implant

The cochlear implant delivers electrical pulses directly to the nerve cells in the cochlea, the spiral shaped structure that translates sound into nerve pulses. In normal hearing individuals, sound wave set up vibrations in the walls of the cochlea, and hair cells detect these vibrations.

High frequency noises vibrate the base of the cochlea, while low frequency notes vibrate near the top of the spiral.

The cochlear implant does the job of the hair cells. It splits the frequencies of incoming noises into a number of channels and then stimulates the appropriate part of cochlea.

Increasing the number of channels will improve sound perception. But speech is perceived in an area of the cochlea only 14 mm long and spacing the electrodes to close to each other causes signals to bleed from one channel to another. This causes a broad version of hearing.

4. EYE IMPLANT

Vision occurs as the light reflected from a body is received by photoreceptors, the light sensing cells at the back of the eye. Blindness occurs if the photoreceptors are lost in retinitis pigmentosa, a genetic disease and in related macular degeneration.

The chip used in eye implant does the function of photoreceptors. The chip will be at least ten times smaller than the thickness of the human hair with an area of 1mm2. There will be a camera mounted on a pair of glasses. The camera will detect and encode the scene and then send it into the eye as a laser pulse. The laser will also provide the energy to drive the chip. The energy required for stimulating a nerve cell in the eye is almost 100 times lower than that required in stimulating a nerve cell in an ear.

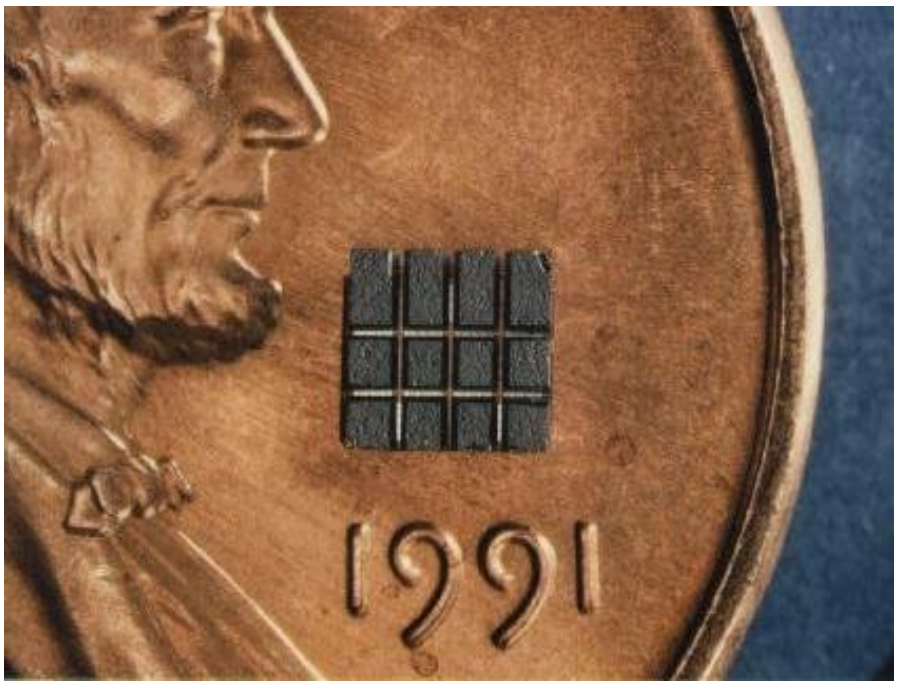


Fig.4. size of a chip implanted in eye (compared with a penny)

5. PERSON IDENTIFICATION

Biochips when implanted into human body can have an identification number, and all the details about that person. This can help agencies to locate lost children, soldiers and Alzheimer's patient. Biochips are widely used in identification of criminals and terrorists in America.

8.2 APPLICATIONS OF EXTERNAL CHIPS

1. LAB ON A CHIP

Biochips scan, process biological data very rapidly. The technology is commonly known as 'lab on a chip'. The idea of a cheap and reliable computer chip look alike that performs thousands of biological reactions is very attractive to drug developers. Because these chips automate highly repetitive laboratory tasks by replacing cumbersome equipment with miniaturized micro fluidic assay chemistries. Biochips are able to provide ultrasonic detection methodologies at significantly lower costs per assay than traditional and also amount of space.

Applications of lab on a chip are basically two

- For the detection of mutations in specific genes as diagnostic "markers" on the onset of a particular disease. E.g.: HIV gene chip
- To detect the differences in gene expression levels in cells those are diseased versus those that are healthy.

  E.g.: cancer studies

2. MASS SPECTROMETRY

Mass spectrometry determines molecular structures from ionized samples of materials. Biochips can be used to perform mass spectrometry and researches are going in that area. This can help in saving much space and time in laboratories

Gene chip will be able to screen diseases like heart attack and diabetics years before patients develop symptoms. These will help doctors diagnose life-threatening illness faster, eliminating expensive, time-consuming ordeals like biopsies and sigmoidoscopies, or simple blood, saliva, stool, or urine tests. Gene chips reclassify diseases based on their underlying molecular signals, rather than misleading surface symptoms.

9. FUTURE DEVELOPMENT

Researchers are working on Bioinformatics that will perform fundamental body changes apart from customizing looks of the people. If you aren't born perfect, any disease and deformity, you need not despair. Because rapid advances in bioinformatics are providing new hopes to such patients. Researchers are going on in the field of` biological computers' hybrid machine like science fiction cyborg which would blend organics and electronics in a single machine. The information processing and storage capabilities of organic molecules are far more superior than the devices that human have been able to create out of silicon.

A few specific areas that fall within the scope of bioinformatics are as follows:

1.Sequence assembly –

The genome of an organisation is assembled from thousands of fragments, which must be correctly 'switched' together. this process requires sophisticated computer- based methods and is carred out by bioinformatics specialists.

2.Sequence (gene) analysis –

Once the DNA sequence of a fragment of the genome is determined, the next step is the understanding of the function of the gene. This involves various analyses, which are carried out by high- powered computing and specialised software. Many would concider this activity as the most important area of focus within bioinformatics.

3.Proteomics –

A relatively new area, proteomics studies not the entire genome, but the portion of the genome that is expressed in particular cells. This involves the collections between

patterns of expression of the genes and a particular disease state to determine likely targets for drug and/or gene therapy. Bioinformatics specialists work closely with scientists to accomplish the same.

4.Pharmacogenomics –

Alterations in the genome at specific positions can be associated with particular disease states, reduced or increased sensitivity to particular drugs, or with side effects to those medications. Such databases are rapidly evolving, and are likely to play an important role in the future drug development efforts and in the design of clinical trails. Bioinformatics experts are at the forefront to collect, analyse and apply this crucial data.

10. CONCLUSION

Days aren't far off when beauty saloons will perform fundamental body changes apart from customizing looks of the people. If you aren't born perfect, free from any diseases and deformity, you need not despair. Rapid advances in bioinformatics are providing new hopes to such patients. At the first sign of physical defect or deformity, people will shop around for a better and stronger organically grown heart, brain, or kidney, as the case may be. With bioinformatics man kind will be able to prolong its life or , even live forever.

11.REFERENCES

Websites

www.electronicsforu.com

www.inbios.org

www.bioinformatics.org

www.biochip.org