

# **Computational methods in Bioinformatics:**

## **Introduction, Review, and Challenges**

**CCSE Technical Report**

**Moustafa Elshafei**  
**Department of Systems Engineering**  
May, 2004

### **Contents:**

#### **Abstract**

- 1- Introduction**
- 2- Introduction to Molecular Biology.**
- 3- Gene Banks**
- 4- Gene Identification**
- 5- Sequence Alignment**
- 6- Multiple Sequence alignment and classification**
- 7- Summary and future directions**
- 8- Conclusion**

#### **References**

## **Abstract**

Biotechnology is emerging as a new driving force for the global economy in the 21<sup>st</sup> century. An important engine for biotechnology is Bioinformatics. Bioinformatics has revolutionized biology research and drug discovery. Bioinformatics is an amalgamation of biological sciences, computer science, applied math, and systems science. The report provides a brief introduction to molecular biology for non-biologists, with focus on understanding the basic biological problems which triggered the exponentially growing research activities in the bioinformatics fields. The report provides as well a comprehensive literature review of the main challenging problems, and the current tools and algorithms. In particular, the problems of gene modeling, and gene prediction, similarity search, multiple alignments of proteins, and the protein folding problems are highlighted. The report discusses as well how such tools as dynamic programming, hidden Markov models, statistical analysis, clustering, decision trees, fuzzy theory, and neural networks have been applied in solving these problems.

### **1- Introduction**

Biotechnology is expected to be the new engine of the global economy during the 21<sup>st</sup> century. Biotechnology is creating new products and markets in many areas from agriculture to chemicals and manufacturing processes, from drug discovery to bio-computing and nanotechnology. The growing biotechnology industry and its sectors, like agriculture, marine sciences, human therapeutics, and the environment are considered the new directions for long-term economic growth.

An important engine in Biotechnology development is Bioinformatics. Bioinformatics technology has the potential to revolutionize biology research and drug target discovery. By reducing drug discovery and development costs, bioinformatics facilitates the creation and commercialization of agricultural, pharmaceutical, environmental, and industrial products that might otherwise be cost prohibitive.

The forecast value for the worldwide informatics market in the life science sector was estimated in 2002 to be approximately \$12 billion, and is expected to grow at rate of over 24% per year to almost \$38 billion by 2006 [1]. Advances in genomics in general, including the mapping of genomes from bacteria, viruses, and humans, have provided an enormous amount of data to be mined. The information encrypted within these data promises advances in areas that can dramatically improve quality of life, including personalized medicine, the use of genes to treat diseases, the development of new energy sources, obtaining better matches for organ transplants, and protection from biological and chemical warfare [2]. For example, in the pharmaceutical industry, traditional drug discovery technologies are reaching the limits of their ability to yield innovative new drugs. Consequently, pharmaceutical firms and researchers are increasingly relying on bioinformatics technologies to use genetic information to identify and develop rational, targeted drugs. The expansion of bioinformatics research is expected to accelerate drug development for a wide range of illnesses, from cancer to Alzheimer's disease. The application of bioinformatics has the potential to drive growth in the worldwide pharmaceuticals drug market from the \$240 billion today to \$3 trillion by 2020<sup>1</sup>.

The potential for significant advances in biological and medical science is enormous but is currently hindered by a shortage of trained Bioinformatics professionals. There is an increasing

---

<sup>1</sup> [http://www.bizintelagents.com/reports/kt12412\\_Bioinformatics.html](http://www.bizintelagents.com/reports/kt12412_Bioinformatics.html)

demand from industry and from academia for individuals with training in both biology and computer science. To fill this need, many universities around the world have started new programs in Bioinformatics and related fields that trains students in both the biological and computer sciences [3, 4]. According to the International Society of computational Biology (ISCB) [5], by 2004 over 18 universities in Europe and over 70 universities in North America started undergraduate and/or graduate programs in Bioinformatics.

Bioinformatics is a merge of molecular biology science and “*informatics techniques*” (derived from disciplines such as applied mathematics, systems science, computer science, statistics, Artificial Intelligence and Pattern recognition) to *understand* and *organize* the *information* associated with these molecules, on a *large scale*. In short, bioinformatics is concerned with:

- 1- Organizing data in a way that allows researchers to access existing information and to submit new entries as they are produced, e.g. gene banks, and protein banks
- 2- Development of data mining and analysis tools, e.g., to identify, qualify, and quantify genes and gene products and proteins.
- 3- Modeling, interpreting and predicting biological activities, and how genes and proteins interact in complex biological systems and regulatory networks.

The international human genome project, which starts in 1989 and finished in 2003, created a research fever for sequencing and annotating DNA sequences [ 6,7]. By 2003, more than 180 genomes from different organisms were completed, and another 900 projects are still undergoing [8]. By 2004, the gene banks databases contain more than 35 billion nucleotides of sequences from a wide spectrum of organisms and species. The exponential growth of gene banks entries is clearly illustrated in Fig.1 [9]. According to [10], the number of submitted papers to

Bioinformatics, a well-known journal in the field, has been increasing at rate of almost 40% annually, which reflects the exponential increase in the research activities in this growing field. Despite the increase in data available each year, less than one percent of microbes are known, many genes remain to be found, most of the functions of the “discovered genes” are still unknown, and functions of noncoding DNA remain unidentified [7].

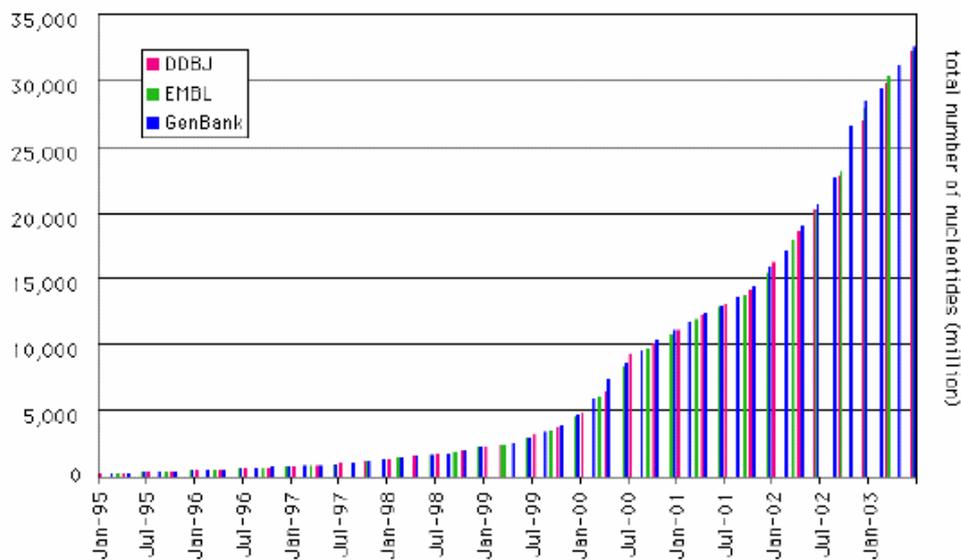


Fig. 1 Exponential growth of gene banks entries [9].

The rapid growth of biological data and the value mined from these data have attracted researchers from many disciplines, e.g. engineering, signal processing, mathematics, physics, operations research, mathematics, and computer science, which has in turn revolutionized the field of bioinformatics.

Statistical methods and mathematical analyses have contributed to the development of new algorithms for DNA and protein sequence analysis and modeling [11-21]. Efficient algorithms based on dynamic programming and Hidden Markov Model (HMM) have been used to discover and assess similarity between sequences, and in gene modeling and prediction [22-31]. More

recent work contributed algorithms using modern Artificial Intelligence tools such as clustering, fuzzy theory, and decision trees [ 32-43], and neural networks and self-organized maps [44-55]. There is also a great need and interest in developing better methods and tools for large scale data mining, visualization, and information integration and management [ 56-64 ]. Robotics and image processing have recently contributed to the discovery of the Microarrays technology. Microarrays allow scientists to analyze expression of many genes in a single experiment quickly and efficiently. They represent a major methodological advance and illustrate how the advent of new technologies provides powerful tools for researchers [65,66 ]

The impact of bioinformatics technology not only has lead to discovery of new concepts in fighting disease [67,68 ], but also lead to a reciprocal impact on such fields as nano-technology and biocomputing [70, 76 ].

## **2. Molecular Biology (gentle introduction)**

This section provides a brief introduction to the science of molecular biology. The objective is to introduce only the basic principles and background that would be needed by non-biologists to understand the molecular biology problems and challenges to be possibly investigated by the researchers and scientists from other fields as computer science, systems science, mathematics, and physics.

### **2.1 Chromosomes**

The classical *chromosome* theory of inheritance holds that chromosomes are the cellular components that physically contain *genes*. [77]. Genes are the functional units of inheritance, and control cell structure and function. Chromosomes consist of a long sequence of molecules

called DNA. A structured gene is a segment of the DNA that code for specific proteins. Non coding genes provide regulatory functions for other genes, or act as templates for molecular acids which control protein synthesis.

The chromosomes in all the cells of the human body are the same (except in sperm, egg and some cells of the immune system). This is because all the cells are derived from the same fertilized egg by cell division. However, the information that does not pertain to the cell's identity is inactive. The number of chromosomes varies from organism to another. In the human genome, there are 46 chromosomes, 2 of which are sex chromosomes, Fig. 2. The number of chromosome of an organism bears no relationship to the organism's complexity. For example, the number of chromosomes in chicken is 78, mouse 40, wheat 42, corn 20, fruit fly 8, and scorpion is 4.

Two types of chromosome pairs occur. *Autosomes* resemble each other in size and structure (one from each parent). For example pairs of chromosome 21 are the same size, while pairs of chromosome 9 are of a different size from pair 21. Sex chromosomes may differ in their size, depending on the species they are from. Cells with two of each type of chromosome are said to be *diploid* whereas cells with only one of each type of chromosome, like sperm cells or egg cells, are said to be *haploid*. But some other organisms such as fungi can be haploid for much of their life cycle.

In humans , males have a smaller sex chromosome, termed the Y, and a larger one, termed the X. Males are thus XY, and are termed *heterogametic*. Females are XX, and are termed *homogametic*.



Fig. 2 The 46 chromosomes of the human<sup>2</sup>.

Cells of organisms are broadly classified into two main types; *Eukaryotes* and *Prokaryotes*.

**Eukaryote** is a type of cell found in many organisms including single-celled *protists* (microbes, molds, and primitive algae), multi-cellular fungi, plants, and animals, characterized by a membrane-bounded nucleus and other membraneous organelles. The first eukaryotes are encountered in rocks approximately 1.2-1.5 billion years old. **Prokaryote** is a more primitive type of cell, which lacks a membrane-bound nucleus, has no membrane organelles, and have a single circular chromosome. Prokaryotes were the first forms of life on earth, evolving over 3.5 billion years ago.

*Phenotypes* are the observed properties or outward appearance of a trait ( height, shape, color, etc). A phenotype is contributed by one or more gene. A gene can have alternate forms called *alleles*. Many genes have more than two alleles (even though any one diploid individual can only have at most two alleles for any gene), such as the ABO blood groups in humans. Human ABO blood types are determined by alleles A, B, and O. A and B are co-dominants, which are

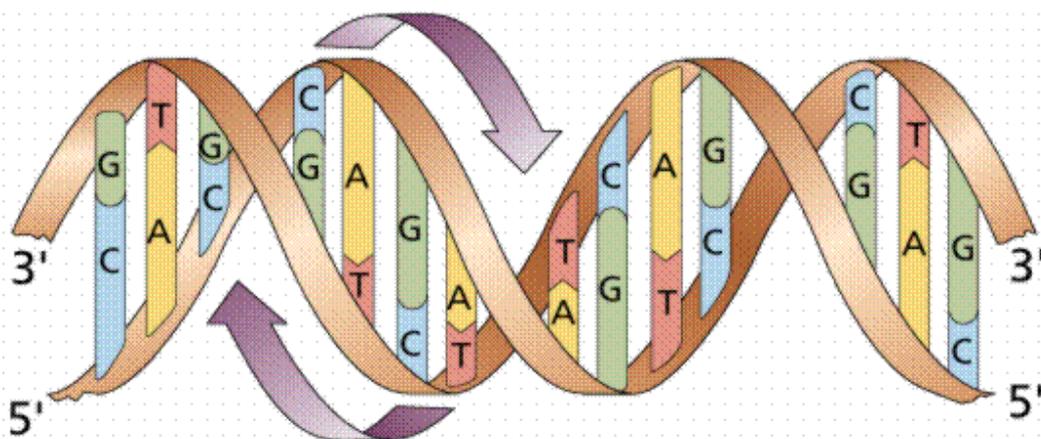
---

<sup>2</sup> [http://www.emc.maricopa.edu/faculty/farabee/BIOBK/Human\\_46,XY.gif](http://www.emc.maricopa.edu/faculty/farabee/BIOBK/Human_46,XY.gif)

both dominant over O. Many traits such as height, shape, weight, color, and metabolic rate are governed by the cumulative effects of many genes. Polygenic traits are not expressed as absolute or discrete characters. Instead, polygenic traits are recognizable by their expression as a gradation of small differences (a continuous variation), which usually follow a normal distribution. Phenotypes are always affected by their environment. Expression of phenotype is a result of interaction between genes and environment.

## 2.1 Deoxyribonucleic acid (DNA) Structure

All information necessary to maintain cell life cycle is embedded in the DNA, a sequence order of four nucleotides: A (*Adenine*), C (*Cytosine*), G (*Guanine*), T (*Thymine*) in the long DNA molecule. DNA is a double helix, with bases to the center (like rungs on a ladder) and sugar-phosphate units along the sides of the helix (like the sides of a twisted ladder). A pairs with T, and C pairs with G. The pairs held together by hydrogen bonds, as depicted in Fig 3.



### Fig. 3. DNA Double Helix<sup>3</sup>

Receiving amino acids from outside and using double DNA helix as a template, a cell produces all materials necessary for its life. Physically DNA is a long molecule intricately packed in space and its structure is determined by the forces of two kinds; covalent bonds and hydrogen bonds.

Covalent bonds provide binding force for the *polynucleotides* chain. Molecule of each nucleotide A, C, G, T is built out of the sugar-phosphate group and the base attached to it. Fig. 4 shows the molecular structure of the Adenine (A) base attached to its Sugar-Phosphate group. Sugar-phosphate groups are naturally polarized. They can bound with each other, forming molecules with hundreds of thousands nucleotides.

On the other hand, Hydrogen bonds are weaker in the order of magnitude, and they provide DNA complementarities. In other words, the two DNA ( equal length) strands are bound by hydrogen bonds. In one of the two strands every A letter is substituted by T in another, C replaced by G, and vice versa. GC-bond is a strong bond provided by three hydrogen bonds, while the AT-bond is weaker, provided with two hydrogen bonds.

The 5' refers to the 5<sup>th</sup> bond of the sugar molecule, see Fig. 3., which in the DNA series is attached to the phosphate group, the 3' refers to the 3<sup>rd</sup> arm of the sugar molecule which is attached to HO the hydroxyle group. Since DNA contains Phosphorous (P) but no Sulphur (S), they tagged the DNA with radioactive Phosphorous-32. Conversely, protein lacks P but does have S, thus it could be tagged with radioactive Sulfur-35.

---

<sup>3</sup> <http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookDNAMOLGEN.html>

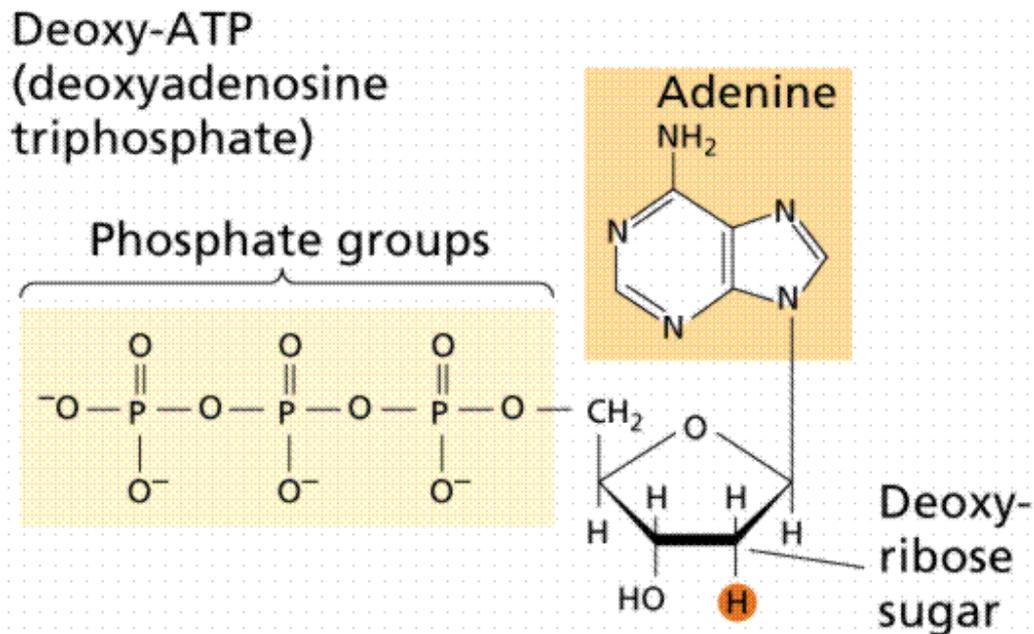
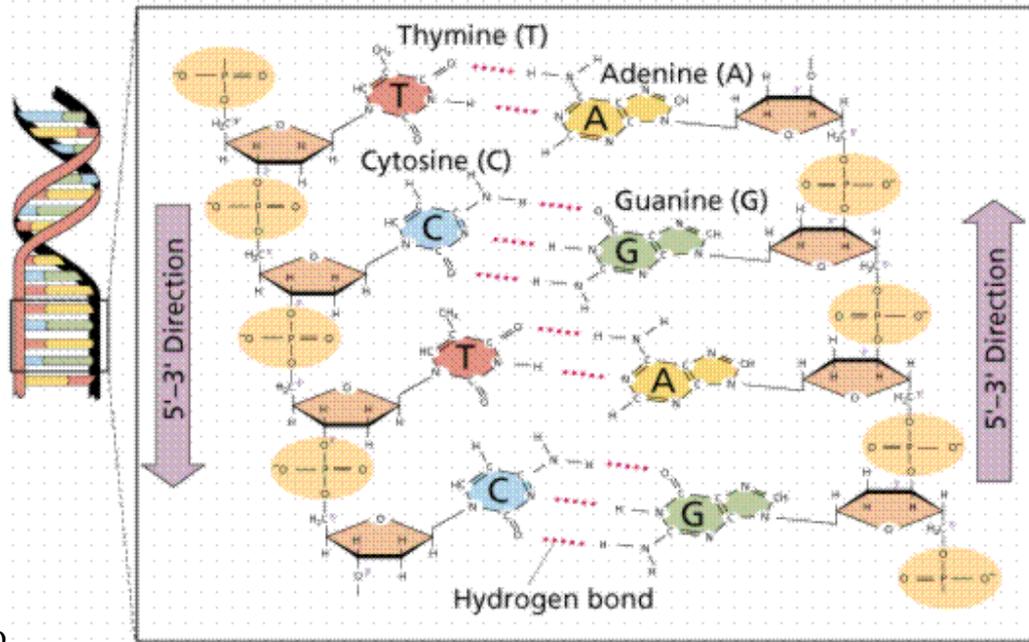


Fig. 4 Chemical structure of the double helix and example of Adenine (A) base<sup>4</sup>.

DNA helix ( 2 nm wide) are rounded on histone fiber of diameter 11 nm, then compacted in 30 nm cromation fiber, then coiled in 700 nm diameter then formed as chromosoms 1400 nm diameter. If the DNA strand of the human genome has 1 mm diameter, it would have stretched to 25km. It would be winded and twisted, and coiled until it becomes a chromosome of 2 ft diameter and 16 ft length

<sup>4</sup> <http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookDNAMOLGEN.html>



0

Fig. 5 Molecular Structure of the double strands DNA.

The year 2003 marks two major milestones in genomics: the completion of the sequencing of the human genome [7], and the 50th anniversary of the discovery of the DNA double helix. The human genome project reveals the sequence of the entire human genome of 3 billion nucleotide pairs, constituting the human 46 chromosomes. Table 1 compares the length of the human genome with other organisms. Genes are segments of DNA which code for specific protein. The number of predicted genes in the human genome is estimated between 30,000 to 40,000 genes, compared to 13,600 for the fruit fly, and over 14,000 in mosquitoes [8]

Organism	Genome length in thousands of nucleotide pairs
Virus	5
E.Coli	4700
Corn	4,500,000

Salamander	72,500,000
Human being	3,000,000

Table 1. Comparison of Genome length in some organisms.

A gene consists of coding and non coding segments, called *exons*, and *introns* respectively. Exon is a section of a gene which codes biological information. Exons can be classified in four classes: "starting" exon, "inner" exon, "terminal" exon and "single" exon (in case when the gene has no introns). Replacement of one nucleotide in an exon for another one may change properties of coded protein radically. So, so exon compositions are practically identical for genes of organisms of the same species. Moreover, genomes of higher species contain many genes which are almost the same base sets as their distant primitive ancestors. A more detailed structure of genes will be discussed in Section 4.

Sections of DNA, that do not code information, may be junk or introns. Junk DNA fills areas between genes. Junk DNA forms the skeleton of DNA, that is its secondary space structure. It seems that small changes in junk composition don't lead to considerable modifications in DNA properties. The major part of Eukaryotes DNA is believed to be a junk DNA or of unknown functions. Eukaryotes have only 10% of their DNA coding for proteins. Humans may have as little as 1% coding for proteins. Viruses and prokaryotes use a great deal more of their DNA. Almost half the DNA in eukaryotic cells is repeated nucleotide sequences. Introns are areas dividing exons in a gene. In translation process introns are cut out and the information coded in them, if any, is not present in the resulting protein.

## 2.2 Proteins

Every function in a cell is controlled by some kind of proteins. Every protein has a specific cell function. Proteins are formed by concatenation (strands) from 20 amino acids. Typical length is several hundreds amino acids, while DNA length is millions to hundred of millions of base pairs. Protein is a single dimension chain, but tends to fold into complex structures. A chain of amino acid is called Polypeptide. Protein are generated based on a code in genes. Protein synthesis is also governed by a genetic code. A segment of the DNA that codes for a specific polypeptide is known as a structural gene.

Every 3 base pairs in DNA can be mapped into 64 possible combinations. The three are called *codons*. The 64 possible codons are mapped into, Start, Stop, and one of the 20 amino acids. For example ATG: START ( the start of a protein synthesis region).

TAA: STOP ( end of a protein synthesis region)

AAA: *Lysine* amino acid, etc.

A stop codon marks the end of a coding region. A section in DNA extending from one stop to a next stop (TAA) could likely contain a gene, and is called Open Reading Frame (ORF).

Complex protein structures like Haemoglobin are made up of one or more polypeptide molecules. During protein synthesis, the DNA coding sequence acts as the blue prints from which a template, called RNA, is constructed and used in the actual protein synthesis.

The following table gives the mapping of codons to the 20 amino acids, start, and the stop codons. The mapping is not one-to-one. While the mapping from a coding DNA sequence to the amino acid sequence is straight forward, the inverse mapping, to identify a section of DNA which code for a specific protein is a more tricky problem.

Second Letter										
		T		C		A		G		
<b>First Letter</b>	<b>T</b>	<b>TTT</b>	<b>Phenylalanine</b>	<b>TCT</b>	<b>Serine</b>	<b>TAT</b>	<b>Tyrosine</b>	<b>TGT</b>	<b>Cysteine</b>	<b>T</b>
		<b>TTC</b>	<b>(Phe)</b>	<b>TCC</b>	<b>(Ser)</b>	<b>TAC</b>	<b>(Tyr)</b>	<b>TGC</b>		<b>C</b>
		<b>TTA</b>	<b>Leucine</b>	<b>TCA</b>		<b>TAA</b>	<b>Stop</b>	<b>TGA</b>	<b>Stop</b>	<b>A</b>
		<b>TTG</b>		<b>TCG</b>		<b>TAG</b>	<b>Stop</b>	<b>TGG</b>	<b>Tryptophan</b>	<b>G</b>
	<b>C</b>	<b>CTT</b>	<b>Leucine</b>	<b>CCT</b>	<b>Proline</b>	<b>CAT</b>	<b>Histidine</b>	<b>CGT</b>	<b>Arginine</b>	<b>T</b>
		<b>CTC</b>	<b>(Ieu)</b>	<b>CCC</b>	<b>(pro)</b>	<b>CAC</b>	<b>(His)</b>	<b>CGC</b>		<b>C</b>
		<b>CTA</b>		<b>CCA</b>		<b>CAA</b>	<b>Glutamine</b>	<b>CGA</b>		<b>A</b>
		<b>CTG</b>		<b>CCG</b>		<b>CAG</b>	<b>(Gln)</b>	<b>CGG</b>		<b>G</b>
	<b>A</b>	<b>ATT</b>	<b>Isoleucine (Ile)</b>	<b>ACT</b>	<b>Threonine</b>	<b>AAT</b>	<b>Asparagine</b>	<b>AGT</b>	<b>Serine</b>	<b>T</b>
		<b>ATC</b>		<b>ACC</b>	<b>(Thr)</b>	<b>AAC</b>	<b>(Asn)</b>	<b>AGC</b>		<b>C</b>
		<b>ATA</b>		<b>ACA</b>		<b>AAA</b>	<b>Lysine</b>	<b>AGA</b>	<b>Arginine</b>	<b>A</b>
		<b>ATG</b>	<b>Metionnine</b> <b>(Met)</b> <b>Start codon</b>	<b>ACG</b>		<b>AAG</b>	<b>(Lys)</b>	<b>AGG</b>		<b>G</b>
	<b>G</b>	<b>GTT</b>	<b>Valine</b>	<b>GCT</b>	<b>Alanine</b>	<b>GAT</b>	<b>Aspartic</b>	<b>GGT</b>	<b>Glycine</b>	<b>T</b>
		<b>GTC</b>	<b>(Val)</b>	<b>GCC</b>	<b>(Ala)</b>	<b>GAC</b>	<b>Acid (Asp)</b>	<b>GGC</b>		<b>C</b>
		<b>GTA</b>		<b>GCA</b>		<b>GAA</b>	<b>Glutamic</b>	<b>GGA</b>		<b>A</b>
		<b>GTG</b>		<b>GCG</b>		<b>GAG</b>	<b>Acid (Glu)</b>	<b>GGG</b>		<b>G</b>

Table 2 Mapping of DNA codons to amino acids.

Protein-coding sequences are interrupted by non-coding regions. Non-coding interruptions are known as intervening sequences or introns. Coding sequences that are expressed are exons.

The Genes length vary between 30k-250k pb, exon regions can be between 69 to 3106 bp, with mean value of about 150 bp. Introns can be as large as 32k bp.[78].

## 2.3 Ribonucleic acid (RNA)

RNA is a single stranded nucleic acid consisting of 4 types of nucleotides similar to the DNA. However, there are two chemical differences distinguish RNA from DNA. The first difference is in the sugar component. RNA contains ribose, while DNA contains deoxyribose. The second difference is that the thymine (T) in DNA is replaced by uracil (U) in RNA. In other words the RNA sequence consists of the 4 bases ( A,U,C,G).

RNA play central role in protein synthesis It was observed that although DNA was located in the eukaryotic nucleus, proteins were being synthesized in the cell in the presence of abundant RNA [77]. Most of this cellular RNA could be found in the site of protein synthesis and called ribosomes. There are three types of RNA that participate in the synthesis of protein: messenger RNA (mRNA), which carries the genetic information from the DNA and used as a template for protein synthesis. Ribosomal RNA (rRNA), which is a major constituent of the cellular particles called ribosomes on which protein synthesis actually takes place. A set of transfer RNA (tRNA), each of which incorporates a particular amino acid subunit into the growing protein when it recognizes a specific group of three adjacent basis in the mRNA. In simpler language, mRNA is the template of the protein product, tRNA is a general purpose protein generation machine, while rRNA is the factory floor.

The sequence of amino acids in a polypeptide is dictated by the codons in the messenger RNA (mRNA) molecules from which the polypeptide is translated. The sequence of codons in the mRNA is, in turn, dictated by the sequence of codons in the DNA from which the mRNA is

transcribed. The mRNA is constructed from the protein coding genes in the DNA after removing the noncoding introns from the DNA sequence as shown in Fig. 6.

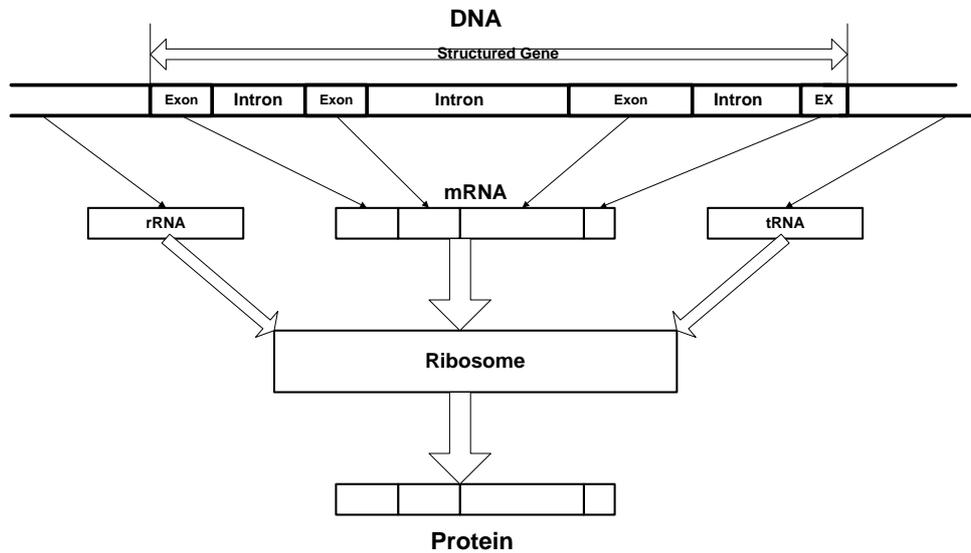


Fig. 6 Construction of protein from DNA.

An RNA gene is any gene that encodes RNA that functions without being translated into a protein. Commonly-used synonyms of "RNA gene" are noncoding RNA or non-coding RNA (ncRNA), and functional RNA (fRNA). Non-coding RNA (ncRNA) genes produce functional RNA molecules rather than encoding proteins

tRNA and rRNA are also coded in the DNA in RNA genes. However, since the late 1990s, many new RNA genes have been found, and thus RNA genes may play a much more significant role than previously thought. Even so, they are probably not as significant or numerous as the protein-coding genes. Several abundant, small non-mRNAs, other than rRNA and tRNA, were detected and isolated biochemically, New RNAs continue to appear [79]. However, almost all

means of gene identification assume that genes encode proteins, so even in the era of complete genome sequences, ncRNA genes have been effectively invisible [80]. Recently, several different systematic screens have identified a surprisingly large number of new ncRNA genes. Non-coding RNAs seem to be particularly abundant in roles that require highly specific nucleic acid recognition without complex catalysis, such as in directing post-transcriptional regulation of gene expression or in guiding RNA modifications.

### **3- Gene Banks & Web Resources**

There is an enormous amount of resources available free on the internet, including gene and protein sequence banks, software, and literatures. A summary of the key resources and banks is given below and in table III.

#### **Primary Web Resources**

- European Molecular Biology Laboratory, Germany  
[http:// www.embl-heidelberg.de](http://www.embl-heidelberg.de)
- ExPASy Molecular Biology Server, Swiss Institute of Bioinformatics, Switzerland  
<http://ca.expasy.org/>
- National Center for Biotechnology Information, USA  
[http:// www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)
- San Diego Supercomputer Center, USA  
[http:// www.sdsc.edu](http://www.sdsc.edu)
- Entrez  
<http://www3.ncbi.nlm.nih.gov/Entrez/>

- Human genome project:  
<http://www3.ncbi.nlm.nih.gov/genome/guide/http://www.ornl.gov/TechResours/>
- Whole genome analysis:  
<http://www.ncbi.nlm.nih.gov/COG/>
- Protein Data Bank (PDB)  
<http://www.rcsb.org/pdb/>
- Structural Classification of Proteins (SCOP)  
<http://scop.mrc-lmb.cam.ac.uk/scop/index.html>
- CATH: Protein Structure Classification  
[http://www.biochem.ucl.ac.uk/bsm/cath\\_new/index.html](http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html)

### **New Frontiers**

- Target identification in drug design, agriculture, biocatalysis:  
<http://www.labmed.umn.edu/umbbd/index.html>
- Differential digital display (Cancer genome anatomy project):  
<http://www.ncbi.nlm.nih.gov/ncicgap/>
- Array technologies:  
<http://cmgm.stanford.edu/pbrown/>
- Metabolic pathways:  
<http://www.ecocyc.org/>; <http://www.genome.ad.jp/kegg/>

TABLE III  
 TOOLS AND SERVICES AVAILABLE ON THE WORLD WIDE WEB FOR  
 BIOINFORMATICS

Tool/Service	Description
GenBank	A repository for scientists to submit new data
Bookshelf	Search information with freely accessible online biomedical texts
PubMed	Access MEDLINE's 11 million biomedical journal citations
PubMed Central	A digital archive for the full text of over 100 life sciences journals
OMIM	Comprehensive catalog of inherited diseases, maintained by Dr. McKusick at Johns Hopkins University
Entrez	A text-based search engine used for major databases such as PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and more
dbSNP	Database of single nucleotide polymorphisms (SNPs)
BLAST	The Basic Local Alignment Search Tool used for comparing gene and protein sequences with public databases.
Map Viewer	Shows integrated views of chromosome maps for several organisms
LocusLink	Covers information on official nomenclature, aliases, sequence accessions, and related web sites
UniGene	Assists in gene discovery, gene mapping projects, and large-scale expression analysis
ORF Finder	Identifies all possible ORFs in a DNA sequence
Electronic PCR	Allows users to search their DNA sequence for sequence tagged sites (STSs)
VAST Search	A structure/structure similarity search service, compares 3D coordinates of new protein structures to existing database
CCAP	Compiles information on chromosome aberrations associated with different cancers
Human-Mouse Morphology Maps	Compares genes in homologous segments of DNA from human and mouse sources
VecScreen	A tool for identifying segments of nucleic acid sequence
dbHMC	An open, publicly accessible platform for DNA and clinical data related to the human Major Histocompatibility Complex
CGAP	Aims to decipher molecular anatomy of cancer cells
Spidey	Aligns one or more mRNA sequences to a single genomic sequence

Entrez<sup>5</sup> is a quick entry point for people who want to investigate known proteins or structures. The Entrez interface lets you search for a protein sequence or a 3D molecular structure using instead of a specific sequence, a name (organism, protein, or gene), identification number, author name, etc. Entrez integrates the scientific literature, DNA and protein sequence databases, 3D protein structure and protein domain data, population study datasets, expression data, and assemblies of complete genomes into a tightly interlinked system. Help using the literature component of Entrez, known as PubMed, is also available. The Entrez help contains a description of the database and its features, basic search techniques and advanced search techniques, and explains the various display formats, how to save results.

For example, to get a nucleotide sequence from the genome of say E.Coli bacteria,

- 1- go to the Entrez web page
- 2- select search for “nucleotide”
- 3- in the query field type: *E.Coli AND 100:500[SLEN]* this will search for nucleotide sequences of Sequence Length [SLEN] between 100 and 500 bp.
- 4- Check one or more of the query results, select the format output from format list box, and choose *send to* text.
- 5- The next web page contains the desired sequence. You can then copy and past in your document. You may also select to save directly the results to a file of your choice.

You can identify proteins of interest by searching a nucleotide string against GenBank using BLASTX or TBLASTX. This will return protein sequences that are identical or similar to the

---

<sup>5</sup> <http://www.ncbi.nlm.nih.gov/Entrez/>

translation product of your gene of interest. These sequences can then be copied and used as queries for further studies.

A number of free standing programs and web based programs are available in order to help researchers find potential coding regions and deduce gene structures for long DNA stretches.

For example, GeneMachine is freely-available for download at <http://genome.nhgr.nih.gov/genemachine>. A public web interface to the GeneMachine server for researchers may be found at <http://genome.nhgr.nih.gov/genemachine/supplement>.

The program allows the user to query multiple exon and gene prediction programs in an automated fashion [ 81].

#### 4- Gene Identification

The problem of automated genes identification may be formulated as following: a sequence of letters A, C, G, T, corresponding to the order of DNA nucleotides in genome, is given at the input of computer program [82,83]. At the output we need to have a list of identified genes with indicated start, end and gene structure, and its division into *exons* and *introns* segments. The accuracy of a given method for identification or classification can be evaluated in terms of the following parameters:

**TP** (true positive) : the frequency of correct patterns being correctly accepted (known and predicted).

**TN** ( True Negative): the frequency of wrong patterns being correctly rejected.

**FP** : the frequency of a wrong pattern being falsely accepted (predicted).

**FN** : the frequency of a correct pattern (known) being rejected.

Based on counts of TP, TN, FP, FN we can define various measures [15, 84], for example:

Sensitivity (SN), also called coverage, is defined as

$$SN = TP / (TP + FN) \quad (1)$$

and Specificity (SP) is defined as

$$SP = TN / (TN + FP) \quad (2)$$

A pattern has maximum sensitivity, if it occurs in all patterns in the family and maximum specificity, if it does not occur in any sequence outside the family. If we want to combine these two measures to one score, we may use Correlation Coefficient (CC)

$$CC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FP)}} \quad (3)$$

This expression has a value 1 when there are no false positive or false negatives, and decreases towards zero as the number of false positives and false negatives grows.

Three different approaches can be distinguished in gene identification methods. They could be called *similarity search*, *content search* and *signal search*.

*Similarity search* is one of the first group of methods that were applied to identify genes in new genomes. It is based on the fact that the function of a gene defines to some extent its nucleotide composition, and if two genes code similar products or functions then the corresponding sites of DNA will be similar. One of the early attempts to evaluate the possibilities of similarity search in a new genome using already known analogs in a database was made by [85]. Rather big collection of genetic sequences in Genbank was arbitrarily divided into two halves. Then genes from one part of the collection were searched with use of the other part as a database. The result was almost 75% correctly identified genes. But when applied to the real new experimentally annotated genomes the method gave only 20-25% of identified genes. Due to the large variability between species, similarity search can at most identify up to 50% of all genes in new genomes.

*Content search* is based on the fact that statistical characteristics, calculated in DNA analysis, differ considerably in coding and non-coding regions. Many features based on observation of structure of nucleotide compositions in genes and junk DNA have been proposed. The earliest features were the frequencies of codon (triplets) usage. Some types of Fourier-transform were investigated and their ability for gene identification was systematically tested [86].

Content search methods based on discriminant functions in multidimensional space of the features were proposed [ 87]. This approach yielded quite good results and some methods proposed were included in computer programs (for example, HEXON, GRAIL) that became real

instruments for primary investigation of new decoded sequences. These programs usually use discriminating rule that is trained on the known analogous samples.

The methods of content search and similarity search share a common concept which can be called "comparison with sample". In case of similarity search such comparison is made at the level of alphabet, while in case of content search the comparison is based on statistical characteristics.

*Signal search* is the third principle of genes identification. Signal search is based on the hypotheses about physical and chemical processes initiating transcription. The molecule that initiates the start of transcription "recognizes" it by the presence of active sites - signals, that are short sequences with a definite structure. There is no clear concept of what are the factors that cause some sites of DNA to serve as signals. Signals as promoters, initiators and terminators of transcription are known, but all these sequences may occur in DNA without initiating any process.

At the early stages of using signal search there were hopes that it would be possible to construct one or more consensus signal sequences and to measure the distance from DNA site to the consensus (using alignment). In these early approaches, the first letter of consensus sequence is the most frequent first letter in all already known signals, the second is the most frequent second letter and so on [88]. Though this approach turned out to be too primitive, at present one of its generalization is successfully applied (when all four letters are used rather than one with calculated probabilities, and resulting consensus is a probability matrix [89]).

At present tens of programs and algorithms realize automated gene identification. A recent excellent overview of the performance of some of them is given in [90]. The most effective

programs in fact use several approaches simultaneously. Unfortunately different algorithms show different results on different databases of annotated genomes. Second, so far there is no single opinion how to compare one program with another (especially it concerns comparing predicted gene structures).

### Gene model

The gene model used in Genescan [78] is depicted in Fig.3, the model consists of 13 forward states and 13 reverse states. The Start state generates one of the two initiation codons used by prokaryotes (ATG or GTG); the Terminate state generates one of the three stop codons (TAA, TAG or TGA);

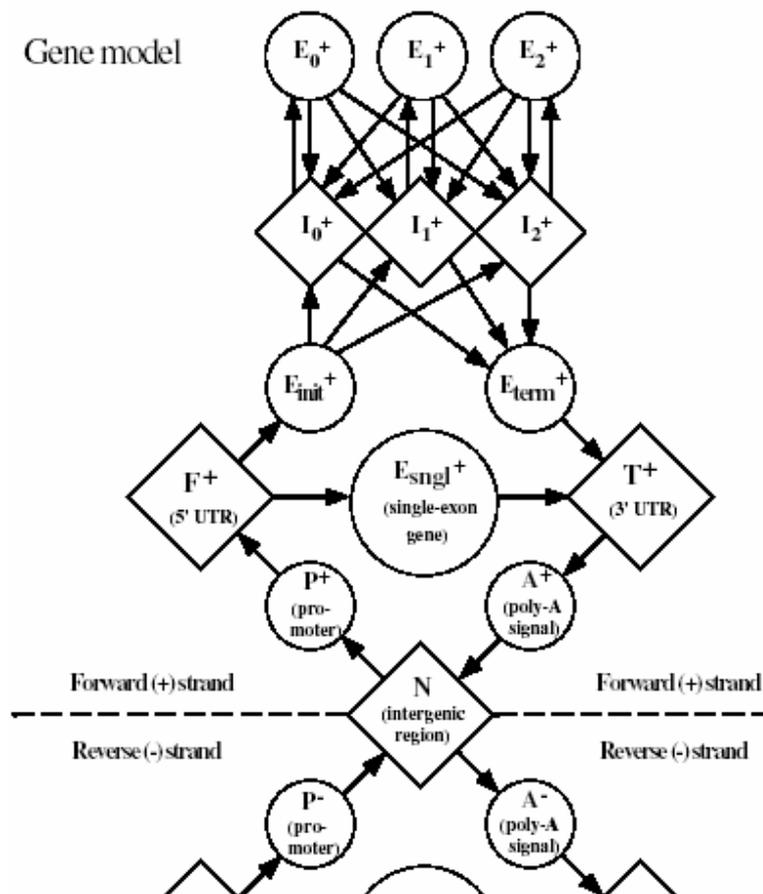


Fig. 7. Gene state model [78].

Starting from an intergenic region and moving in the forward direction, the program expects to find first a “promotor site”. This upstream promoter site is (T,A) rich called TATA box (25-30 base-pairs(bps) . Following the promotor site (if any), the program allocates the starting region of the gene, known as the 5’ UTR (untranslated region), that is the program F+ state. The F+ state extends from the start of transcription to just before the translation initiation signal. The  $E_{init}$  state is the initial exon. If this exon is not the only exon ( $E_{single}$ ), the program tries to identify an intron region. With a few exception, virtually all introns begin with (GT), called donor splice signal, and end with (AG), called acceptor signal. Since exons must be multiple of three nucleotides, while introns do not follow this rule, there could be phase shift from exon region to another exon region. This 3 possible phase shifts are accounted for by including three internal exon phases  $\{ E_0, E_1, E_2 \}$ , and three internal intron phases  $\{ I_0, I_1, I_2 \}$ .  $E_{term}$  is the terminal exon. The 3’ UTR region is characterized by a signal of the form (AATAA + A-rich-sequence 20-30 bps away). In the model described here, the reading frame is kept track of by dividing introns and internal exons according to their “phase”. Thus, an intron which falls between codons is considered phase 0; after the first base of a codon, phase1; and after the second base of a codon, phase 2. Internal exons are similarly divided according to the phase of the previous intron, which determines the codon position of the first base pair of the exon, hence the reading frame. For example, if the number of complete codons generated for an initial exon is  $c$  and the phase of the subsequent intron is  $k$ , then the total length of the exon is  $d=3c+k$ ; The components of an  $E_k^+$  (forward-strand internal exon) state will be encountered in the order: acceptor site, coding region, donor site, while the components of an  $E_k^-$  (reverse-strand internal exon) state will be encountered in the order: inverted complement of donor site, inverted complement of coding region, inverted complement of acceptor site

The GeneScan algorithm is based on a generalized Hidden Markov Model GHMM. The GHMM model consists of four main components: a vector of initial probabilities  $\Pi$ , a matrix of state transition probabilities  $\mathbf{T}=\{t_{ij}\}$ , a set of length distributions  $\{f_i\}$ , and a set of sequence generating models  $P_i$ ;  $i=0-26$ ;

The program takes a DNA sequence  $S$  of length  $L$ , and generates a “parse”  $\phi$  consisting of set of a state sequence states  $Q=\{q_1, q_2, \dots, q_n\}$ , with associated lengths  $D=\{d_1, d_2, \dots, d_n\}$ , and sequence segmentation  $S = \{s_1, s_2, \dots, s_n\}$ .

The joint probability of a parse  $\phi$  and a sequence  $S$  is given by

$$P(\phi, S) = \pi_1 f_{q_1}(d_1) P_{q_1}(s_1) \prod_{k=2}^n t_{q_{k-1}, q_k} f_{q_k}(d_k) P_{q_k}(s_k) \quad (5)$$

Where,  $\pi_1$  is the probability of the first state. The objective then is to find the optimal parse  $\phi$  which maximizes the conditional probability of  $\phi$  given the DNA sequence  $S$ .

$$P(\phi | S) = \frac{P(\phi, S)}{P(S)}$$

With a few assumptions, the above problem can be solved efficiently using the Viterbi algorithm [91]. Other programs exist for gene finding, for example GRAIL (Gene Recognition and Analysis Link) based on neural network [92], HMMGene based on a different HMM model, some sensors or mile stones, e.g, start and stop codons, frequency of codons, frequency of repeats [31]; MORGAN is based on decision trees [93], FGENEH/FGENES Predicts exons by known splice site features [94], and MZEF uses quadratic discrimination function analysis [95].

## 5- Sequence alignments

Sequence alignment is a tool to compare 2 sequences. Needleman-Wunsch [96] is one of the earliest *global alignment* algorithms to find the optimum alignment (including gaps) of two sequences when considering their entire length. The method uses dynamic programming to search for the optimal global alignment. A tool was developed based on this algorithm known as “Needle”. Needle finds an alignment with the maximum possible score where the score of an alignment is equal to the sum of the matches taken from the scoring matrix.

On the other hand, *local alignment* algorithms search for regions of local similarity between two sequences and need not include the entire length of the sequences. Local alignment methods are very useful for scanning databases when it is desired to find matches between small regions of sequences, for example between protein domains. A popular algorithm known as “Water”, based on Smith-Waterman algorithm [97]. Water is a member of the class of algorithms that can calculate the best score and local alignment in the order of  $(m \times n)$  steps, (where 'n' and 'm' are the lengths of the two sequences).

FASTA and BLAST are also popular tools for similarity search. Both methods rely on identification of brief sub-sequences (k-tuples), which serve as the core of an alignment. Multiple k-tuples can be combined and extended as seeded for more extended alignment, allowing also deletion, insertion, or changes between two sequences. BLAST (Basic Local Alignment Search Tool) [98] is the most popular sequence comparison algorithm optimized for speed to search sequence databases for optimal local alignments to a query. The BLAST algorithm, developed by the National Center for Biotechnology Information (NCBI) at the

National Library of Medicine<sup>6</sup>, is a heuristic for finding locally optimal sequence alignments. There are several versions of BLAST. The BLAST family of programs can be used to compare an amino acid, query sequence against a protein sequence database, or a nucleotide query sequence against a nucleotide sequence database, as well as other combinations of protein and nucleic acid. The initial search is done for a word of length "W" that scores at least "T" when compared to the query using a substitution matrix. Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S". The "T" parameter dictates the speed and sensitivity of the search.

FASTA [99], a sort for "Fast All" or "FastA", is the first widely used algorithm for database similarity searching. Similar to BLAST, the program looks for optimal local alignments by scanning the sequence for small matches called "words". Initially, the scores of segments in which there are multiple word hits are calculated. Later the scores of several segments may be summed to generate a combined score. The sensitivity and speed of the search are inversely related and controlled by the "k-tuples" variable which specifies the size of a "word".

## **6- Classification and Multiple Sequence Alignment**

DNA and protein sequence classification is an important problem in computational biology [89]. Discovering closely related homologues, i.e. members of the same family of proteins or the corresponding genes in different related species has been a major task in computational biology. When organisms are remote relatives, the homology signal begins to submerge in noise, and the problem becomes increasingly challenging.

---

<sup>6</sup> [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)

There are two different, but related classification problems. The first is how to find a classifier function for a family of bio-sequences. This is a function which takes a sequence as argument, returning TRUE for members of the family and FALSE for non members. Both positive examples (members of the family) and negative examples (sequences not in the family) are given as a training set. In the second problem only positive examples (family members) are given, and the goal is to extract a description of features conserved in (characterizing) the family. In many cases it is desired to discover what is called a conservation function, and the evolutionary relations. This class of problems is known as the **Multiple Sequence Alignment (MSA)** problem.

The techniques for solving the first problem can be categorized into the following three classes:

A) Sequence Alignment This approach aligns the unlabelled sequence  $S$  with members of a set  $C$  using an existing tool, such as FASTA and BLAST, and assigns  $S$  to  $C$  if the best alignment score for  $S$  is sufficiently high.

B) Consensus search: this approach takes a collection of sequences of the class  $C$  and generates composite subsequences by taking the majority base at each position in multiple alignment of sequences in  $C$ . The consensus sequence is then used to identify sequences in uncharacterized biosequence [100, 89].

C) Inductive learning/ Neural networks: This approach takes a set of sequences of the class  $C$  and a set of sequences not in  $C$  and then, based on these sequences and using learning techniques, AN artificial Neural Network (ANN) determines whether or not the unlabelled sequence  $S$  belongs to  $C$  [50,51,89,101]

**Multiple Sequence Alignment** MSAs are essential bioinformatics tools. MSA will continue to be a central to the sequence-based biological analysis for many years to come.

MSAs are required for phylogenetic analysis, to scan databases for remote members of a protein family and structure prediction. No perfect method exists for assembling a multiple sequence alignment and all the available methods are heuristic approximations.

The most commonly used methods for doing multiple sequence alignments use a progressive alignment algorithm, called ClustalW, [101]. Progressive alignments algorithms [102, 103] depend on a progressive assembly of the multiple alignments, where sequences or alignments are added one by one so that never more than two sequences (or multiple alignments) are simultaneously aligned using dynamic programming. This approach has the great advantage of speed and simplicity combined with reasonable sensitivity, even if it is by nature a heuristic that does not guarantee any level of optimization.

Recent techniques have focused on the design of iterative methods [104], for example iterative dynamic programming [105], and Genetic Algorithm, SAGA [106]. In consistency based methods, DiAlign [107], T-Coffee [08], the optimal MSA is the one which optimize all pair-wise alignment. For example, DiAlign [107] assembles the alignment in a sequence-independent manner by combining segment pairs in an order dictated by their score, until every residue of every sequence has been incorporated in the multiple alignment. Iterative alignment methods depend on algorithms able to produce an alignment and to refine it through a series of cycles (iterations) until no more improvements can be made. Iterative methods can be deterministic or stochastic, depending on the strategy used to improve the alignment.

Benchmarking on a collection of reference alignments [109] indicates that ClustalW performs reasonably well on a wide range of situations, while DiAlign is more appropriate for sequences

with long insertions/deletions. Future methods should be able to integrate structural information within the multiple alignments and to allow some estimation of their local reliability.

## 6. Protein Structure Analysis

While sequence analysis focuses on the one dimensional characteristics of the nucleic acids and proteins, it is fact that their three dimensional structure that underlines their structural and functional properties. Much computational biology research is devoted to the prediction of the precise three-dimensional structure of proteins given their amino acid sequence, and to further discover their resulting function [110].

Structural biologists classify protein structure at four levels. A protein's primary structure is the sequence of amino acids in a polypeptide chain. Local runs of amino acids often assume one of two sequence structures: a closely packed helical spiral ("alpha" helix), or a relatively flat structure where successive runs of peptides fold on one another ("beta" sheet). Secondary structure is also called a "coiled" region. The complete, detailed conformation of the molecule, describing how these helices, sheets, coils, and intervening sequences are precisely positioned in three dimensions, is referred to as the protein's tertiary structure (3D structure). There are two approaches to this problem [111]. In the first approach is based on homology with sequences whose tertiary structure is known. In the second approach is derived from first principles based on fundamental atomic interactions. The protein folding problem can be considered as a search for a folding function  $F$ , where  $V=F(S)$ , and  $S$  is the amino acid chain  $S = \{s_1, s_2, \dots, s_n\}$ , where  $s_i$  is a member of the set of 20 amino acids. The vector  $V$  of dimension  $3n$  represents the relative or the absolute positions of each amino acid in a 3D structure. Conceptually, the protein structure would be the one which minimizes the protein chain free energy. The problem can be posed a

search problem for a vector  $V$  which minimizes an Energy function  $E(V,S)$ . The energy function employs a set of information theoretic potential of mean force [115,116]. The first step in this approach is to determine a potential function  $E$ , then selection of a suitable search algorithm. For a protein chain of length  $N$ , the search space would be of order  $10^N$  states. [112] argued that each protein can basically have only 7 states, and accordingly the complexity of the search algorithm would be  $7^N$ .

In fact, the general problems of protein folding, and protein structure are all known to be NP-hard problem [113]. Other investigators observed that there are recurrence patterns in protein folds, and proposed to limit the search to say, 1000 possible protein folds [114]. In this case the problem becomes a “Fold Recognition”, by selecting the most appropriate one. The candidate set is constructed by first searching for closely related proteins in known families of proteins. Then we construct the set of the candidate folding structures from those closely related to the given protein and of known folding structures. The third step is to identify the structure which minimizes an energy function. Another approach is based on limiting the folding recognition to the core part of the protein [113]. It is argued that long chains fold first on a stable core, which has a relatively limited number of 3D patterns. However, determining the core part of a given protein chain is by itself can be a complex and challenging problem.

## **7- Summary and Future Directions**

- 1- Sequence Alignment algorithms locate a region of interest. Raw sequencing is performed on pieces of random lengths between 500 to 5000 pbs. With possible large overlapping parts at both ends. Algorithms align the fragments, and find the pair wise alignments in

the pieces, discover similar sequences in the databases. There a need for much faster and more effective third generation algorithms. This new generation should be built on the knowledge gained about the known genomes and how they are structured.

- 2- Gene finding algorithms try to identify a potential gene region in DNA. However, only 1-3% of human genome is translated into proteins. It is not clear until now what is the purpose (if any) of the large quantities of "junk DNA" , that does not appear to code for any proteins. Characterization of the features of the regulatory RNA genes still to be determined, and development of effective methods for discover and predicting these noncoding genes still an open question. The DNA in the vicinity of genes has several structure features, e.g., promoter region and other binding sites. The stochastic and deterministic properties of these region, and how they can be used to identify genes need further studies. More work still to be conducted to understand the mutation mechanism in genes, and the cell techniques for fault tolerance and error recovery.
- 3- Protein structure prediction: given the linear primary structure of a protein sequence, how it would fold itself into a specific 3D complex shape. The problem involves a vary large search space for the optimal shape based on thermodynamics principles, and possibly covalent interaction and modifications. Once the 3-dimensional structure of a protein is known, it becomes possible to design drugs that inhibit or enhance a protein's activity by fitting into niches in the surface of the protein. It may also be possible to design new proteins with useful properties. Perhaps the more difficult is to determine sequences that give rise to desired structures.
- 4- Homology search: we discovered a new gene, and its function is still to be determined. We then search for members of the same family of proteins or the corresponding genes in

different related species. Local alignment and similarity search algorithms can be used to find the closest matches. However, statistical grouping, clustering, statistical similarity measures are first needed for course classification.

- 5- Multiple Alignment and phylogeny construction: the comparison of DNA and protein sequences in different species is an increasingly important tool for understanding the evolutionary relationships among species. These are typically depicted by phylogenetic trees that indicate how species branched off from ancestral species. There is a great need for developing better probabilistic models for the evolutionary process and metrics for comparing trees or quantifying the robustness of the information deducible from them.
- 6- Modeling Cell Activities: The rate at which proteins are produced and activated is different in different cells and at different times, depending on factors such as the ambient environment of the cell and chemical signals from other cells. Protein expressions, regulation, and interaction can be better understood if new mathematical models are developed. The models can help us to understand the cell activities and reaction to outside stimulus. The results may lead to production of better drugs or to improving the immune system.
- 7- Many processes that go on in living cells can be viewed in computational terms. DNA strands can in a sense be viewed as the tapes of multi-headed Turing machines, from which the designs for proteins (the genes) are read and the proteins themselves then produced. The rate at which proteins are produced and activated is different in different cells and at different times, depending on factors such as the ambient environment of the cell and chemical signals from other cells

- 8- The proliferation of biological data and the need for its systematic and flexible storage, retrieval, and manipulation is creating opportunities in the database field. Current genomic databases are heterogeneous, distributed, and semistructured or with schemas that are in flux, thus offering novel challenges in database design, including its more fundamental aspects.
- 9- DNAmicroarrays: In DNA microarrays, also known as DNA Chips, an unknown fragment of DNA is tested against a large number of DNA fragments arranged in a grid. The DNA chips produce patterns of light which varies in light and intensity depending on the degree of similarity between the unknown DNA specimen and the members of the grid. How can we provide quantitative, consistent, and standardized interpretations from the test results ? and how should arrays be designed so as to maximize the accuracy of readings obtained from it?

## **8- Conclusion:**

Bioinformatics is an emerging field which is expected to be an important contributor to the global economy. Research in this field has already made a major impact on the pharmaceutical industry and drug discovery, agriculture, health care, environment, and protection from biological warfare. The report acts as a single starting point for new comers in this field. It provides an overview of the research activities, and how knowledge from applied math, operations research, artificial intelligence, computer science, and other fields merge to create this field.

## **Acknowledgement**

The author would like to acknowledge KFUPM for its support in conducting this research.

## References

- [1] A. Jacobson, "Bioinformatics booming," IEEE Comput. Sci. Eng. Mag., vol. 4, p. 11, July–Aug. 2002.
- [2] Barbara A. Oakley, and Darrin M. Hanna, "A Review of Nanobioscience and Bioinformatics Initiatives in North America", IEEE Transactions on NanoBioscience, Vol. 3, No. 1, March 2004
- [3] T. Raymer, M.D. Krane, and O.Garcia," Crossing the interdisciplinary barrier: a baccalaureate computer science option in bioinformatics Doom", IEEE Transactions on Education,, Volume: 46 , Issue: 3 , pp. 387 – 393, Aug. 2003
- [4] R. Hughey, and K. Karplus, "Bioinformatics: a new field in engineering education", 31st Annual Frontiers in Education Conference, 2001., Volume: 2 , pp.10-13, Oct. 2001 .
- [5] The International Society of computational Biology (ISCB)  
<http://www.iscb.org/univ.shtml>
- [6] "Genomics and its impact on science and society: the human genome project", U.S. Department of Energy, Washington DC, 2003.
- [7] Human Genome Project (HGP) Information,  
[http://www.ornl.gov/sci/techresources/Human\\_Genome/project/50yr.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/project/50yr.shtml)  
Oak Ridge National Laboratory, US Department of Energy.
- [8] Genome On-Line Database (GOLD), <http://wit.integratedgenomics.com/GOLD/>
- [9] S.A. De Carvalho Jr., Sequence Alignment Algorithms, MSc., King's College, University of London, 2003.
- [10] C. Sander, "The journal Bioinformatics, key medium for computational biology," Bioinformatics, vol. 18, pp. 1–2, 2002.
- [11] E. Jain, "Current trends in bioinformatics," Trends Biotechnol., vol. 20, pp. 317–319, 2002.

- [12] G. Singh, "Statistical modeling of DNA sequences and patterns," in *An Introduction to Bioinformatics*, S. Krawtz, S. Krawtz, and D. Womble, Eds. Totowa, NJ: Humana, 2002.
- [13] L. R. Cardon and G. D. Stormo, "Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments," *J. Mol. Biol.*, vol. 223, pp. 159–170, 1992.
- [14] R. Arratia, E. S. Lander, S. Tavare, and M. S. Waterman, "Genomic mapping by anchoring random clones: A mathematical analysis," *Genomics*, vol. 11, pp. 806–827, 1991.
- [15] G. A. Churchill and M. S. Waterman, "The accuracy of DNA sequences: Estimating sequence quality," *Genomics*, vol. 89, pp. 89–98, 1992.
- [16] J. Felsenstein, "Evolutionary trees from DNA sequences: A maximum likelihood approach," *J. Mol. Evolut.*, vol. 17, pp. 368–376, 1981.
- [17] A. Rzhetsky and M. Nei, "Statistical properties of the ordinary least squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference," *J. Mol. Evolut.*, vol. 35, pp. 367–375, 1992.
- [18] E. M. Crowley, K. Roeder, and M. Bina, "A statistical model for locating regulatory regions in genomic DNA," *J. Mol. Biol.*, vol. 268, pp. 8–14, 1997.
- [19] P. Baldi, S. Brunak, P. Frasconi, G. Pollastri, and G. Soda, "Exploiting the past and the future in protein secondary structure prediction," *Bioinformatics*, vol. 15, pp. 937–946, 1999.
- [20] R. Sanchez and A. Sali, "Large-scale protein structure modeling of the *saccharomyces cerevisiae* genome," in *Proc. Nat. Acad. Sci.*, vol. 954, 1998, pp. 13 597–13 602.
- [21] T. D. Moloshok, R. R. Klevecz, J. D. Grant, F. J. Manion, W. F. T. Speier, and M. F. Ochs, "Application of bayesian decomposition for analyzing microarray data," *Bioinformatics*, vol. 18, pp. 566–575, 2002.

- [22] Z. Galil and R. Giancarlo, “Speeding up dynamic programming with applications to molecular biology,” *Theor. Comput. Sci.*, vol. 64, pp. 107–118, 1989.
- [23] D. Gusfield, “Efficient algorithms for inferring evolutionary trees,” *Networks*, vol. 21, pp. 19–28, 1991.
- [24] T. Hunkapiller, R. J. Kaiser, B. F. Koop, and L. Hood, “Large-scale and automated DNA sequence determination,” *Science*, vol. 254, pp. 59–67, 1991.
- [25] R. Idury and M. S. Waterman, “A new algorithm for shotgun sequencing,” *J. Comput. Biol.*, 1995.
- [26] M. S. Waterman, “Efficient sequence alignment algorithms,” *J. Theor. Biol.*, vol. 108, pp. 333–337, 1984.
- [27] , “Rapid dynamic programming algorithms for RNA secondary structure,” *Adv. Appl. Math.*, vol. 7, pp. 455–464, 1986.
- [28] H. Carillo and D. Lipman, “The multiple sequence alignment problem in biology,” *SIAM J. Appl. Math.*, vol. 48, pp. 1073–1082, 1988.
- [29] D. Baker and A. Sali, “Protein structure prediction and structural genomics,” *Science*, vol. 294, pp. 93–96, 2001.
- [30] A. G. Pedersen, P. Baldi, S. Brunak, and Y. Chauvin, “Characterization of prokaryotic and eukaryotic promoters using hidden Markov models,” in *Proc. 4th Int. Conf. Intelligent Systems Molecular Biology*, 1996, pp. 182–191.
- [31] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler, “Hidden Markov models in computational biology: Applications to protein modeling,” *J. Mol. Biol.*, vol. 235, pp. 1501–1531, 1994.

- [32] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, 2nd ed. Cambridge, MA: MIT Press, 2001.
- [33] D. J. Galas, M. Eggert, and M. S. Waterman, "Rigorous pattern recognition methods for DNA sequences: analysis of promoter sequences from *E. coli*," *J. Mol. Biol.*, vol. 186, pp. 117–128, 1985.
- [34] L. Pickert, I. Reuter, F. Klawonn, and E. Wingender, "Transcription regulatory region analysis using signal detection and fuzzy clustering," *Bioinform.*, vol. 14, pp. 244–251, 1998.
- [35] J. T. L. Wang, Q. Ma, D. Shasha, and C. H. Wu, "New techniques for extracting features from protein sequences," *IBM Syst. J. (Special Issue on Deep Computing for the Life Sciences)*, vol. 40, pp. 426–441, 2001.
- [36] J. T. L. Wang, B. A. Shapiro, and D. Shasha, *Pattern Discovery in Biomolecular Data: Tools, Techniques and Applications*. London, U.K.: Oxford Univ. Press, 1999.
- [37] V. Faramarz, "Pattern recognition techniques in microarray data analysis," *Ann. NY Acad. Sci.*, vol. 980, pp. 41–64, 2002.
- [38] D. Dembele and P. Kastner, "Fuzzy C-means method for clustering microarray data," *Bioinformatics*, vol. 19, pp. 973–980, 2003.
- [39] J. Tamames, D. Clark, J. Herrero, J. Dopazo, C. Blaschke, J. M. Fernandez, J. C. Oliveros, and A. Valencia, "Bioinformatics methods for the analysis of expression arrays: Data clustering and information extraction," *J. Biotechnol.*, vol. 25, pp. 269–283, 2002.
- [40] W. Schmitt and W. S. Waterman, "Linear trees and RNA secondary structure," *Disc. Appl. Math.*, vol. 51, pp. 317–323, 1994.
- [41] J. Herrero and J. Dopazo, "Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns," *J. Proteome Res.*, vol. 1, pp. 467–470, 2002.

- [42] H. Resson, R. Reynolds, and R. S. Varghese, "Increasing the efficiency of fuzzy logic-based gene expression data analysis," *Physiol. Genomics*, vol. 13, pp. 107–117, 2003.
- [43] A. Sturn, J. Quackenbush, and Z. Trajanoski, "Genesis: Cluster analysis of microarray data," *Bioinformatics*, vol. 18, pp. 207–208, 2002.
- [44] B. Rost and C. Sander, "Combining evolutionary information and neural networks to predict protein secondary structure," *Proteins*, vol. 19, pp. 55–72, 1994.
- [45] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins*, vol. 47, pp. 228–235, 2002.
- [46] T. L. Bailey and C. P. Elkan, "Unsupervised learning of multiple motifs in biopolymers using expectation maximization," *Mach. Learn.*, vol. 21, pp. 51–83, 1995.
- [47] C. M. Bishop, *Neural Networks for Pattern Recognition*. London, U.K.: Oxford Univ. Press, 1995.
- [48] I. Mahadevan and I. Ghosh, "Analysis of *E. coli* promoter structures using neural networks," *Nucleic Acids Res.*, vol. 22, pp. 2158–2165, 1994.
- [49] A. G. Pedersen and J. Engelbrecht, "Investigations of *E. coli* promoter sequences with artificial neural networks: New signals discovered upstream of the transcriptional start point," in *Proc. 3rd Int. Conf. Intelligent Systems Molecular Biology*, 1995, pp. 292–299.
- [50] C. H. Wu, "Artificial neural networks for molecular sequence analysis," *Comput. Chem.*, vol. 21, pp. 237–256, 1997.
- [51] C. H. Wu and J. McLarty, *Neural Networks and Genome Informatics*. Amsterdam, The Netherlands: Elsevier, 2000.

- [52] Q. Ma, J. T. L. Wang, D. Shasha, and C. H. Wu, "DNA sequence classification via an expectation maximization algorithm and neural networks: A case study," *IEEE Trans. Syst., Man, Cybern. C*, vol. 31, pp. 468–475, Nov. 2001.
- [53] T. Sawa and L. Ohno-Machado, "A neural network-based similarity index for clustering DNA microarray data," *Comput. Biol. Med.*, vol. 33, pp. 1–15, 2003.
- [54] A. Mateos, J. Dopazo, R. Jansen, Y. Tu, M. Gerstein, and G. Stolovitzky, "Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons," *Genome Res.*, vol. 12, pp. 1703–1715, 2002.
- [55] Y. Xu, F. M. Selaru, J. Yin, T. T. Zou, V. Shustova, Y. Mori, F. Sato, T. C. Liu, A. Olaru, S. Wang, M. C. Kimos, K. Perry, K. Desai, B. D. Greenwald, M. J. Krasna, D. Shibata, J. M. Abraham, and S. J. Meltzer, "Artificial neural networks and gene filtering distinguish between global gene expression profiles of barrett's esophagus and esophageal cancer," *Cancer Res.*, vol. 62, pp. 3493–3497, 2002.
- [56] P. A. Pevzner and M. S. Waterman, "A fast filtration for the substring matching problem," *Lecture Notes in Computer Science, Combinatorial Pattern Matching*, vol. 684, pp. 197–214, 1993.
- [57] B. Prum, F. Rodolphe, and E. Tuckerheim, "Finding words with unexpected frequencies in DNA sequences," *J. R. Stat. Soc. Ser. B.*, vol. 55, pp. 205–220, 1995.
- [58] U. Ukkonen, "Finding approximate patterns in strings," *J. Algorithms*, vol. 6, pp. 132–137, 1985.
- [59] P. Bertone and M. Gerstein, "Integrative data mining: The new direction in bioinformatics," *IEEE Eng. Med. Biol. Mag.*, vol. 20, pp. 33–40, Jul.–Aug. 2001.
- [60] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena,

E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezcimen, P. Rocca-Serra, and S. A. Sansone, "ArrayExpress—A public repository for microarray gene expression data at the EBI," *Nucleic Acids Res.*, vol. 31, pp. 68–71, 2003.

[61] P. Riikonen, J. Boberg, T. Salakoski, and M. Vihinen, "Mobile access to biological databases on the Internet," *IEEE Trans. Biomed. Eng.*, vol. 49, pp. 1477–1479, Dec. 2002.

[62] J. P. Lee, D. Carr, G. Crinstein, J. Kinney, and J. Saffer, "The next frontier for bio- and cheminformatics visualization," *IEEE Comput. Graph. Appl.*, vol. 22, pp. 6–11, Sept.–Oct. 2002.

[63] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein, "GoMiner: A resource for biological interpretation of genomic and proteomic data," *Genome Biol.*, vol. 4, p. R28, 2003.

[64] K. J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W. C. Reinhold, B. Zeeberg, W. Ajay, and J. N. Weinstein, "MatchMiner: a tool for batch navigation among gene and gene product identifiers," *Genome Biol.*, vol. 4, p. R27, 2003.

[65] R. Ekins and F. W. Chu, "Microarrays: Their origins and applications," *Trends Biotechnol.*, vol. 17, pp. 217–218, 1999.

[66] T. P. Dooley, E. V. Curto, R. L. Davis, P. Grammatico, E. S. Robinson, and T. W. Wilborn, "DNAmicroarrays and likelihood ratio bioinformatic methods: discovery of human melanocyte biomarkers," *Pigment Cell Res.*, vol. 16, pp. 245–253, 2003.

[67] J. R. Baker Jr, A. Quintana, L. Piehler, M. Banaszak-Holl, D. Tomalia, and E. Raczka, "The synthesis and testing of anti-cancer therapeutic nanodevices. biomedical microdevices," *Biomed. Microdev.*, vol. 3, pp. 59–67, 2001.

- [68] T. Hamouda and J. R. Baker Jr, "A novel surfactant nanoemulsion with a unique nonirritant topical antimicrobial activity against bacteria, enveloped viruses and fungi," *Microbiol. Res.*, vol. 156, pp. 1–7, 2001.
- [69] R. K. Soong, G. D. Bachand, H. P. Neves, A. G. Olkhovets, H. G. Craighead, and C. D. Montemagno, "Powering an inorganic nanodevice with a biomolecular motor," *Science*, vol. 290, pp. 1555–1558, 2000.
- [70] C. D. Montemagno, "Nanomachines: A roadmap for realizing the vision," *J. Nanoparticle Res.*, vol. 3, pp. 1–3, 2001.
- [71] G. Wu, H. Ji, K. Hansen, T. Thundat, R. Datar, R. Cote, M. Hagan, A. K. Chakraborty, and A. Majumdar, "Origin of nanomechanical cantilever motion generated from biomolecular interactions," *Proc. Nat. Acad. Sci.*, vol. 98, pp. 1560–1564, 2001.
- [72] S.-J. Park, T. A. Taton, and C. A. Mirkin, "Array-based electrical detection of DNA using nanoparticle probes," *Science*, vol. 295, pp. 1503–1506, 2002.
- [73] R. Bashir, "Biologically mediated assembly of artificial micro and nanostructures," in *CRC Handbook of Nanoscience, Engineering, and Technology*, W. Goddard, D. Brenner, S. Lyshevski, and G. Iafrate, Eds. Boca Raton, FL: CRC, 2003.
- [74] S. I. Stupp and P. V. Braun, "Molecular manipulation of materials: biomaterials, ceramics, and semiconductors," *Science*, vol. 277, p. 1242, 1997.
- [75] H. Hess, J. Howard, and V. Vogel, "Surface imaging by self-propelled nanoscale probes," *Nanoletters*, vol. 2, pp. 113–116, 2002.
- [76] G. H. Pollack, "Micro-and nano-scale motion in the cell," presented at the Int. MEMS Workshop, Singapore, 2001.
-

- [77] D.L. Hartl and E.W. Jones. Genetics: principles and analysis, Jones and Bartlett publishers, Toronto, Canada, 1998.
- [78] C. Burge, Identification of Genes in Human Genomic DNA, Ph.D. Thesis, Stanford University, 1997.
- [79] Sean R. Eddy , “Non-Coding RNA Genes and the Modern RNA World”, Nature Reviews Genetics, vol. 2, no. 12, pp. 919-929, December, 2001.
- [80] V.A. Erdmann, et al. “The non-coding RNAs as riboregulators”, Nucleic Acids Res. 29, 189-193, 2001.
- [81] I. Makalowska, J.F. Ryan, and A.D. Baxevanis, “GeneMachine: Gene prediction and sequence annotation”, Bioinformatics Application Note, Vol 17, No. 9, 2001, pp. 843-844.
- [82] C. Burge and S. Karlin, “Prediction of Complete Gene Structures in Human Genomic DNA”, J. Mol. Biol. (1997) 268, 78-94.
- [83] R. Guigo, ”Computational Gene Identification: an open problem”, Comp. Chem. Vol. 21, No. 4, pp. 215-222, 1997.
- [84] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen, “ Pattern Discovery in Biosequences”, ICGI, 1998.
- [85] O. Seely .Jr., D.F. Feng, D.W. Smith , D. Sulzbach , R. Doolittle, (1990)Genomics 8,71.
- [86] Fickett J.W..The Gene Identification Problem:An Overview For Developers.Computers Chem.,1996.Vol.20,No.1,pp.103-118.
- [87] G.J. McLachlan, Discriminant Analysis and statistical Pattern Recognition’, John Wiley, New York, (1992)
- [88] R. Staden, “Methods for Calculating the Probabilities of Finding Patterns in Sequences”, CABIOS, Vo. 5, pp. 89-96, 1989.

- [89] Gelfand MS, “ Prediction of function in DNA sequence analysis”, *J Comput Biol* 1995, 2:87-115.
- [90] Rogic S.,Mackworth A.K.,Ouellette F.B.Evaluation of Gene-Finding Programs on Mammalian Sequences.”, *Genome Research*.Vol.11, No. 5, pp.817-832 (2001).
- [91] L. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceedings of the IEEE*, 77(2), 1989.
- [92] Y. Xu, and E.D. Uberbacher, ”Computational gene prediction using neural networks and similarity search”, in S.L. Salzberg, D.B. Searls, and S. Kasif (eds.), *Computational Methods in Molecular Biology*, Elsevier Science, 1998.
- [93] S.L. Salzberg, ”Decision Trees and Markov chains for gene finding”, in S.L. Salzberg, D.B. Searls, and S. Kasif (eds.), *Computational Methods in Molecular Biology*, Elsevier Science, 1998.
- [94] Solovyev, V.V., Salamov, A.A., Lawrence, C.B., “Identification of human gene structure using linear discriminant functions and dynamic programming”, In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology* (eds. C. Rawling et al. ), pp. 367–375. AAAI Press, Menlo Park, CA. 1995.
- [95] M.Q. Zhang, “Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci.* 94: 565–568, 1997.
- [96] Needleman, S. B. and Wunsch, C. D. *J. Mol. Biol.* 48, 443-453., (1970)
- [97] T.F. Smith, and M.S. Waterman , *J. Mol. Biol* 147(1);195-7, (1981)
- [98] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman,” Basic Local Alignment Search Tool (BLAST)”, *J. Mol. Bio.*, 215:403-410, 1990.

- [99] W.R. Pearson, and D.J. Lipman, "Improved tool for biological sequence comparison", Proc. Natl. Acad. Sci. Vol. 85, pp. 2444-2448, 1988.
- [100] R. Staden, "Computer methods to locate signals in nucleic acid sequences", Nucleic Acids Res. 12, 505-519, (1984).
- [101] Haym Hirsh and Michiel Noordewier (1994)., Using Background Knowledge to Improve Inductive Learning of DNA Sequences. Proceedings of the Tenth IEEE Conference on Artificial Intelligence for Applications (CAIA94), pages 351-357.
- [101] J.D. Thompson, D.G. Higgins, T.J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice", Nucleic Acids Res. 22, 4673-4680, 1994.
- [102] C. Notredame, D.G. Higgins, and J. Heringa, "T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment", J. Mol. Biol. 302, 205-217, (2000).
- [103] F. Corpet, "Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res, 25;16(22):10881-10890, 1988.
- [104] C. Notredame, "Recent progress in Multiple Sequence Alignments", Pharmacogenomics, Jan;3(1):131-44 (2002).
- [105] O. Gotoh, "Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinements as Assessed by Reference to Structural Alignments", J. Mol. Biol. 264(4), 823-838, (1996).
- [106] C. Notredame, D.G. Higgins, "SAGA: Sequence Alignment by Genetic Algorithm", Nucleic Acid Research, Vol. 24, 1515-1524, (1996).
- [107] B. Morgenstern, "DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment", [In Process Citation]. Bioinformatics 15(3), 211-8, (1999).

- [108] C. Notredame, DG. Higgins, and J. Heringa, "T-Coffee: A novel algorithm for multiple sequence alignment. *J. Mol. Biol.* 302, 205- 217 (2000).
- [109] J.D. Thompson, F. Plewniak, and O. Poch, "A comprehensive comparison of multiple sequence alignment programs", *Nucleic Acids Res.* 27(13), 2682-2690 (1999).
- [110] Y.J. Edwards, and A. Cottage, "Bioinformatics methods to predict protein structure and function. A practical approach", *Mol Biotechnol.* 2003 Feb;23(2):139-66.
- [111] D. Baker , and A. Sali, " Protein structure prediction and structural genomics", *Science.* 2001 Oct 5;294(5540):93-6. , PubMed : 11588250.
- [112] M.J. Rooman, J.P.A. Kocher, and S.J. Wodak, *J.Mol.Biol.* Vo. 221, pp. 961-979, 1991.
- [113] R.H. Lathrop, et. al., " Analysis and algorithms for protein sequence-structure alignment", in S.L. Salzberg, D.B. Searls, and S. Kasif (eds.), *Computational Methods in Molecular Biology*, Elsevier Science, 1998.
- [114] A.Grant, D. Lee, C. Orengo, " Progress towards mapping the universe of protein folds. PMID: 15128436 [PubMed - in process], *Genome Biol.*5(5):107. Epub, 2004.
- [115] M.Hendlich et.al, *J. Mol. Bio.*, Vol. 216, pp. 167-180., 1990
- [116] M.J. Sippl, *J. Mol. Bio.*, Vol. 213, pp. 859-883, 1990