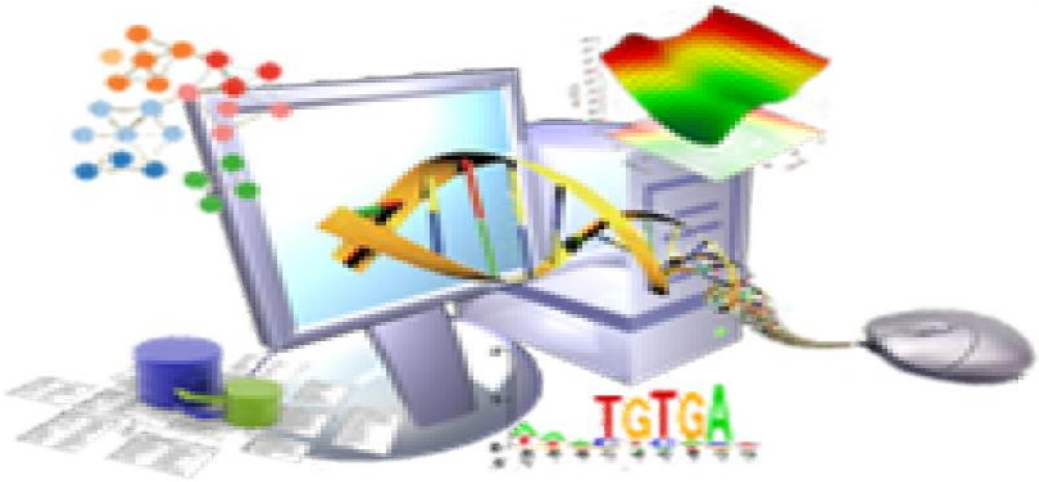
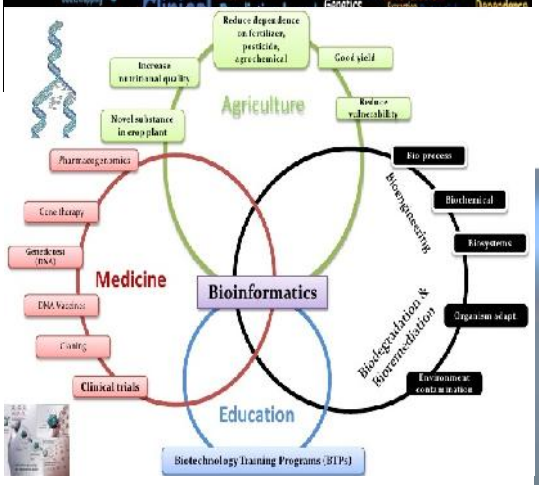
Introductory Bioinformatics

-Student Hand Book

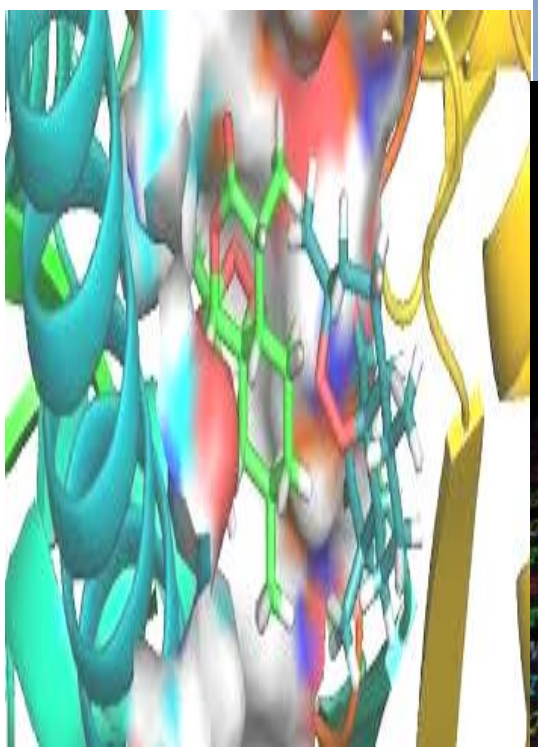
For Life Sciences Students in UG & PG

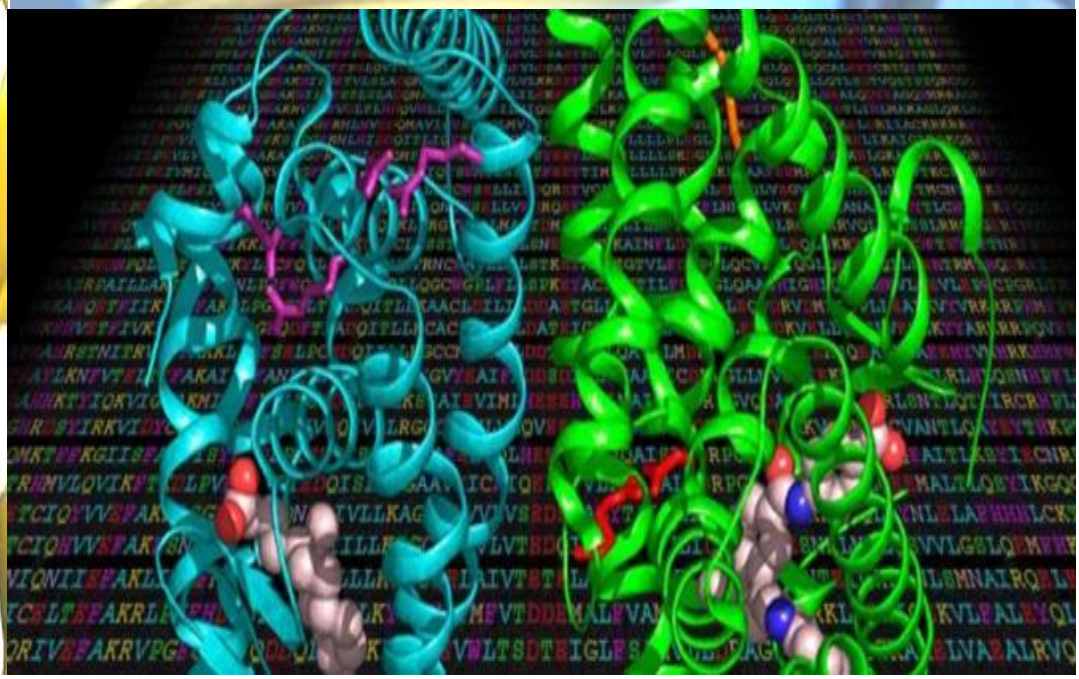




Authors

Prof. D.M. Mamatha
Dr. K. Swetha kumari
Ms. S. Kalpana





Introductory Bioinformatics

- Student Hand Book

(For Life Sciences Students in UG & PG)

By

Dr. D.M. MAMATHA PhD, FISEC, FISCA,
Professor & Head

Dr. K. SWETHA KUMARI MSc (Bioinfo), PhD.
&
Ms. S. KALPANA MSc (Bioinfo)

Dept. of Sericulture

Sri Padmavati Mahila Visvavidyalayam

(Women's University)

Tirupati, A.P. India

2016

International E - Publication

www.isca.co.in

International E - Publication
427, Palhar Nagar, RAPTC, VIP-Road, Indore-452005 (MP) INDIA
Phone: +91-731-2616100, Mobile: +91-80570-83382
E-mail: **contact@isca.co.in** , Website: **www.isca.me** , **www.isca.co.in**

ISBN: 978-93-84659-62-2

Foreword

The past two decades has witnessed not only a flood of protein sequence and structure data generated by large-scale genomic sequencing and structural genomics projects, but also an ensuing growth of size and number of databases and computer programs designed to manage and process these data. The crowd of bioinformatics tools accessible to molecular biologists offers several solutions to various steps of process sequence–structure–function analyses. Often the choice of which tool to apply depends more on its esteem, sometimes stemming from the availability of an intuitive web-server interface, rather than on an understanding of the underlying principles or on the user's ability to utilize all the information given by the program, including the assessment of confidence of the results. Being well-informed and trained in molecular biology and biotechnology and self-taught in bioinformatics, the authors are interested in both the development of computational tools and their optimal application in the realm of experimental biology, especially in the studies of protein–nucleic acid interactions. Regardless of the loads of literature on bioinformatics and on molecular biology of proteins that network with nucleic acids, only little timely volumes are available to the synthesis of these two research areas. This book is mainly for a molecular biologist who wants to analyze, search or manipulate protein structure or sequence data and to integrate these analyses with their experimental investigations to interpret the obtained results or to plan further studies better. Thus, the book comprises of useful strategies for studying protein function with the aid of bioinformatics, described in step by step form. In the spirit of this series, all case studies involve analyses of proteins involved in interactions with nucleic acids – from ribosome assembly and structure, to posttranscriptional RNA modification, to DNA restriction and repair. The bioinformatics field is a very fast-moving one, and every effort was made to produce this volume as rapidly as possible so the methods would be timely. I hope that *Introductory Bioinformatics* will serve as a useful compendium of methods both to tenderfoot in the field of bioinformatics-aided experimental molecular biology and biochemistry as well as to Research scholars actively affianced in this research area.

- ***Authors***

CONTENTS **Pg. No**

I. Basics

1.1 Nucleic Acids:

1.1.1 2′-Deoxyribonucleic acid (DNA):

DNA (2′-deoxyribonucleic acid) is the molecular store of genetic information in nearly all living systems. It is a large polymeric molecule composed of monomers known as *nucleotides*. Each nucleotide consists of a heterocyclic base, a pentose sugar (2′-deoxy-D-ribofuranose), and a phosphate group. There are four heterocyclic bases in DNA: adenine (A), guanine (G), cytosine (C) and thymine (T).

Their structures and numbering system are shown below:

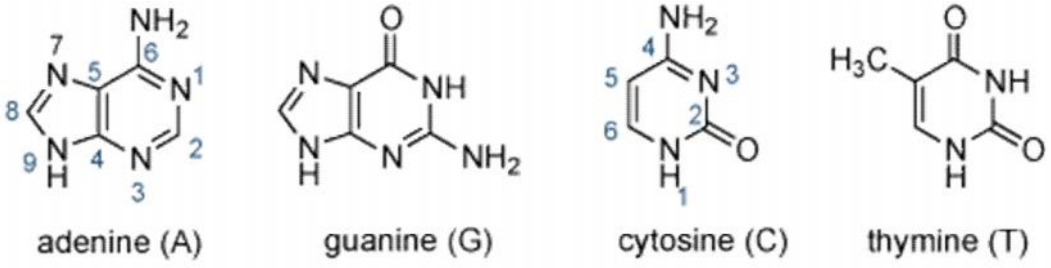


Fig: 1 Chemical structure of the heterocyclic bases of DNA

Adenine and guanine are purines and cytosine and thymine are pyrimidines.

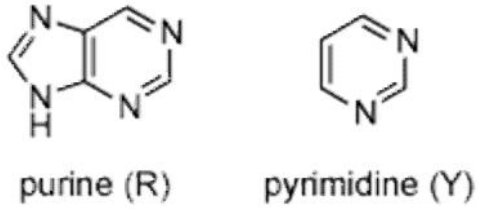


Fig: 2 Structures of the purine and pyrimidine heterocyclic ring systems

The deoxyribose sugar is shown in Figure 3. As a free sugar it can *mutarotate* under certain conditions, adopting furanose, acyclic and pyranose forms; but in DNA it is fixed as a furanoside.

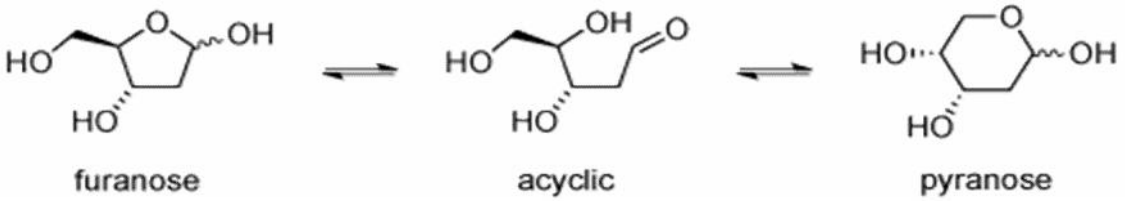


Fig: 3 Mutarotation of 2′-deoxy-D-ribose allows interconversion between furanose, acyclic and pyranose forms

The phosphate group can be found at the 5′- or 3′-position of the sugar depending on the method used to break down DNA to produce the nucleotides. Removal of the phosphate gives rise to a nucleoside. The heterocyclic base is linked to the 1′-position of the sugar. The chemical structures of a deoxynucleoside and a deoxynucleotide are shown in Figure 4.

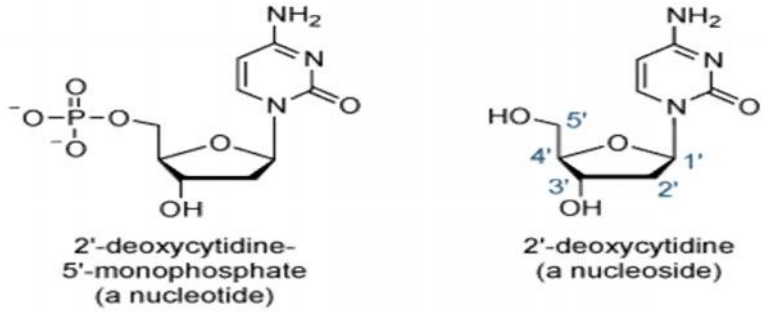


Fig: 4 Structures of a deoxynucleotide and a deoxynucleoside

The bond joining the 1′-carbon of the deoxyribose sugar to the heterocyclic base is the *N*-glycosidic bond. Rotation about this bond gives rise to *syn* and *anti*-conformations. Rotation about this bond is restricted and the *anti*-conformation is generally favoured, partly on steric grounds.

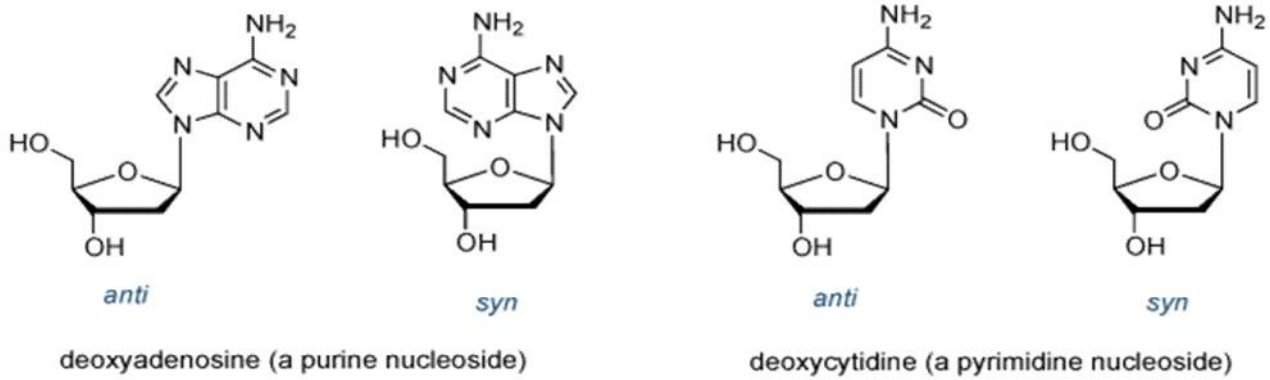


Fig: 5 Structures of *syn* and *anti*-nucleoside conformations

The nucleosides in the DNA duplex adopt the *anti*-conformation (there are very few exceptions to this rule, one of which is guanosine monophosphate, in which the guanine base adopts the *syn* conformation about the glycosidic bond).

The 1′-position of the deoxyribose sugar is the *anomeric centre*. If a substituent attached to the 1′-carbon lies on the same face of the sugar ring as the 5′-hydroxyl group, it is known as the β-anomer; if the substituent is on the opposite side of the sugar ring it is the α-anomer (Figure 6). All of the nucleosides in DNA are in the β-configuration.

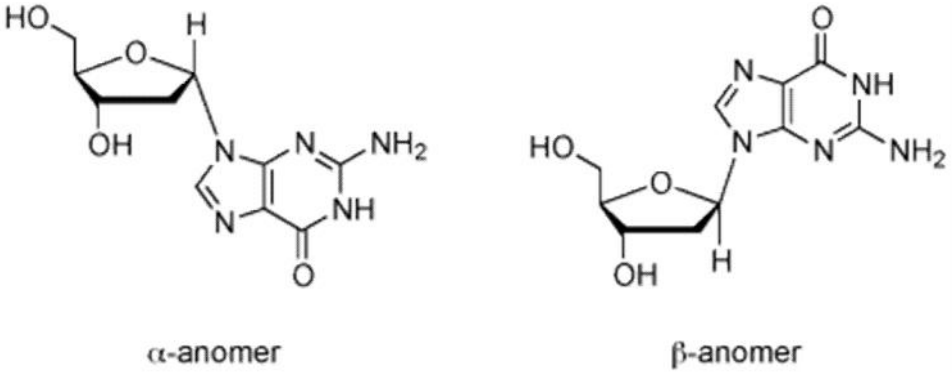


Fig: 6 Structures of nucleoside α- and β-anomers

1.1.2 RIBONUCLEIC ACID (RNA):

Some organisms, for example retroviruses, use ribonucleic acid (RNA) instead of DNA as their store of genetic information. RNA is chemically very similar to DNA but there are two important differences:

- RNA has a hydroxyl group attached to the 2'-position of the sugar, and
- The pyrimidine uracil (U; Figure 7) replaces thymine in RNA.

The key biological role of RNA is as a messenger: it reads the genetic code in DNA (transcription) and transports it to the ribosome, where it is decoded into the sequence of a protein (translation).

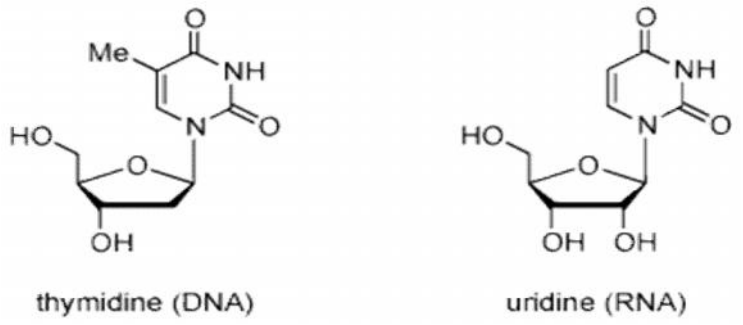
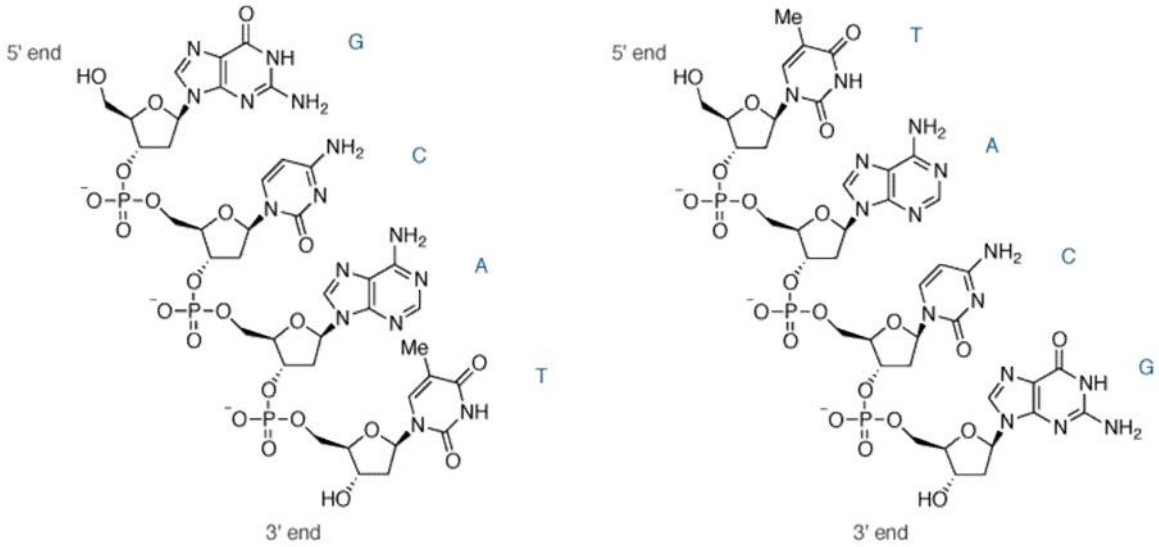


Fig: 7 Structures of uridine and thymidine nucleosides

(Thymidine occurs in DNA; Uridine in RNA;)

1.1.3 OLIGONUCLEOTIDES:

A dinucleotide (dimer) of DNA or RNA is formed by covalently linking the 5′-phosphate group of one nucleotide to the 3′-hydroxyl group of another to form a *phosphodiester bond*. An oligonucleotide (oligomer) is formed when several such bonds are made, and naturally-occurring nucleic acids are linear, high molecular weight molecules of this kind. At physiological pH (7.4) each phosphodiester group exists as an anion (hence the term *nucleic acid*), and nucleic acids are therefore highly charged polyanionic molecules (Figure 8).



a b

Fig: 8 Chemical structures of the oligonucleotides dGCAT (left) and dTACG (right)

One end of a nucleic acid strand has a 5′-hydroxyl group (primary hydroxyl) and the other end has a 3′-hydroxyl group (secondary hydroxyl). The nucleic acid chain therefore has directionality. The tetranucleotide in Figure 8a has the sequence 5′-GCAT-3' and the tetranucleotide in Figure 8b has the sequence 5′-TACG-3'. By convention, the prefixes 5′- and 3′- are not written, and nucleic acid sequences are written in the 5' to 3' direction. The oligonucleotides GCAT and TACG are distinct molecules with different chemical and biophysical properties.

1.1.4 NUCLEIC ACID DUPLEXES:

The chemical structure of a single strand of DNA gives little insight into its biological function as a carrier of genetic information. However, when James Watson and Francis Crick showed in 1953 that DNA adopts a double-stranded structure (duplex), the mechanism of DNA replication (copying) became apparent. The double-helical structure was principally elucidated from X-ray fibre diffraction data (acquired by Rosalind Franklin and Maurice Wilkins) and Chargaff's rules. Erwin Chargaff discovered that the molar amount of adenine in DNA was always equal to that of thymine and the same was true for guanine and cytosine (i.e. number of moles of G = number of moles of C). Watson and Crick were able to explain this by building models to show that the two strands of DNA are held together by hydrogen bonds between individual bases on opposite strands. The purine base A always pairs with the pyrimidine T and the purine G always pairs with the pyrimidine C (Figure 9).

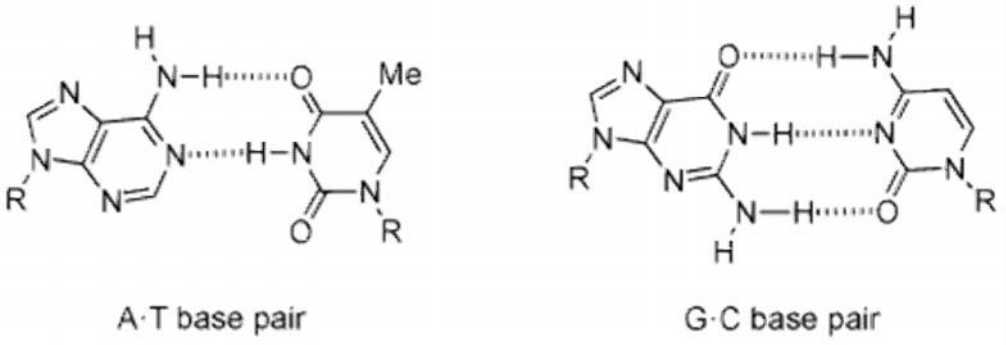


Fig: 9 Hydrogen bonding in the A·T and C·G Watson-Crick DNA base pairs

Not surprisingly A·T and G·C are known as Watson-Crick base pairs. They are pseudo symmetric and if an A·T base pair is laid over any other base pair (T·A, G·C or C·G) the

phosphodiester backbones fall on top of each other (Figure 10). Thus all four base pairs fit neatly into the double helix.

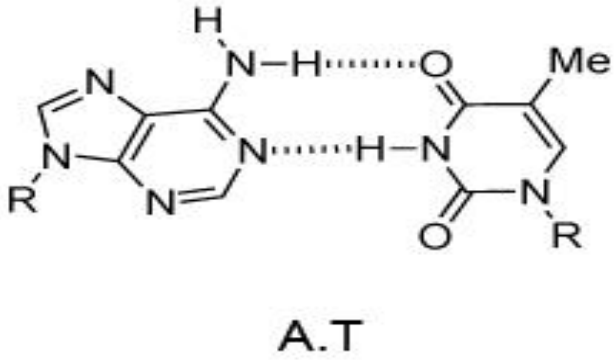


Fig 10: Animated overlay of G·C and A·T base pairs

The sequence of one strand of DNA precisely defines the sequence of the other; the two strands are said to be *complementary*, and are sometimes called *reverse complements* of each other. The two strands are antiparallel, with the 5′-end of one strand adjacent to the 3′-end of the other. The two strands coil around each other to form a right-handed double helix, with the hydrophobic base pairs in the centre and the sugars and negatively charged phosphates forming the external hydrophilic backbone. The term "right-handed" indicates that the backbone at the front of the molecule facing the observer slopes down from top right to bottom left. The planar heterocyclic bases stack one on another and the separation between successive base pairs along the helix axis is around 0.34 nm. One helical turn (a full 360° turn of the double helix) is repeated every 10 to 11 base pairs. The stability of the duplex is derived from both base stacking and hydrogen bonding. The principal form of double helical DNA, B-DNA (Figure 11, middle), has a wide major groove and a narrow minor groove running around the helix along the entire length of the molecule. Proteins interact with the DNA in these grooves (principally in the major groove) and some small drug molecules (e.g. netropsin, distamycin) bind in the minor groove (see Nucleic acid-drug interactions).

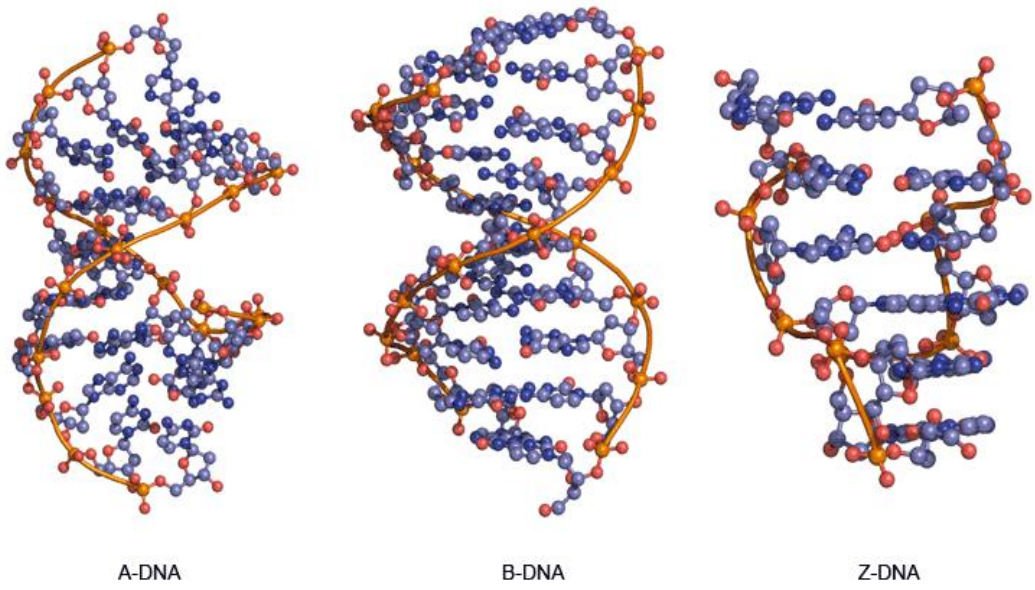


Fig: 11 Three-dimensional structures of A-DNA, B-DNA and Z-DNA

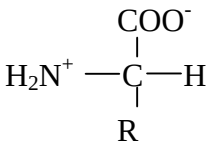
RNA can also form a right-handed duplex using the same base pairing rules (A·U and G·C), but the RNA duplex has a distinctive shape (the A-form) in which the major groove is deep and the minor groove very shallow. Under conditions of low humidity DNA can also adopt the A-form (Figure 11, left), and there are a number of other conformations of DNA and RNA, most of which are subtle variations on the A- and B-forms.

One drastically different DNA conformation, Z-DNA, (Figure 12, right) was determined from an X-ray crystal structure of a chemically synthesized DNA strand of the alternating sequence dCGCGCG, which spontaneously forms a duplex in aqueous buffer. Z-DNA is left-handed and has a dinucleotide repeat unit, so the backbone is not smooth but appears to "zig-zag". It is not clear whether Z-DNA has biological relevance.

1.2 Proteins:

Proteins are polymers of amino acids and constitute the largest fraction (besides water) of cells. Amino acids are compounds containing carbon, hydrogen, oxygen and nitrogen. They serve as monomers (building blocks) of proteins and composed of an amino group, carboxyl group, a hydrogen atom, and a distinctive side chain, all bonded to a carbon atom, the α-carbon.

In an α-amino acid, the amino and carboxylic groups are attached to the same carbon atom, which is called the α-carbon atom. The various α-amino acids differ with respect to the side chain attached to the α-carbon atom. The general structure of amino acid is



Amino acids can act as both acids and bases. When an amino acid is dissolved in water, its solution exists in dipolar ion or zwitter ion. Hence, it is an amphoteric molecule. The different amino acids and their structures were as shown below:

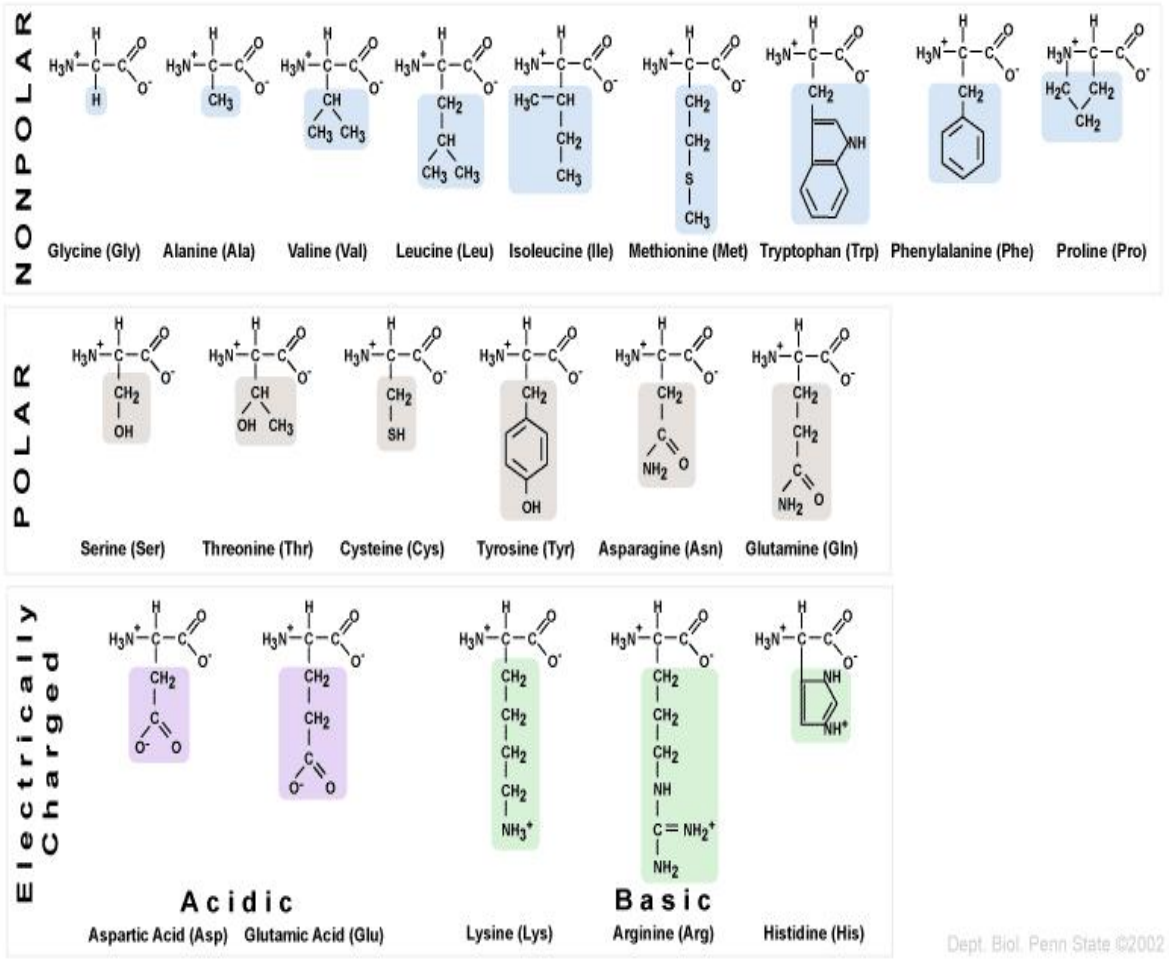
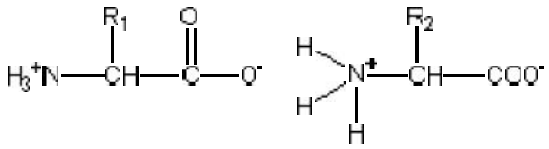


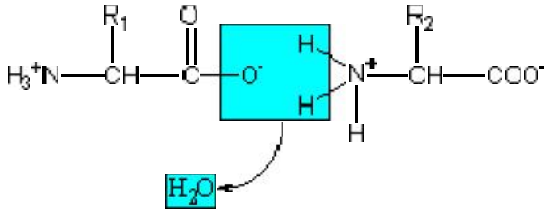
Fig: 12 Classification of amino acids

1.2.1 Peptide Bond Formation:

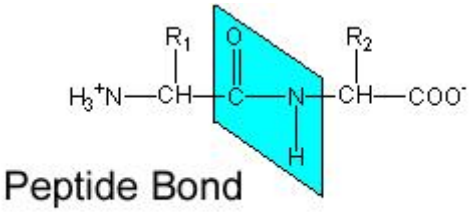
First, two amino acids are brought together. The acid group of the first is close to the amine group of the second.



Next, a water molecule is eliminated, leaving a bond between the acid carbon of the first amino acid and the amine nitrogen of the second



The peptide bond is formed between the two amino acids.



1.2.2 Protein Structure:

Proteins are unbranched polymers constructed from 22 standard α- amino acids. They have four levels of the structural organization.

Primary Structure:

The primary structure (1^0structure) of a polypeptide is its amino acid sequence. The amino acids are connected by peptide bonds. Primary structure of polypeptides determines the higher levels of structural organization.

Secondary Structure:

The most common types of secondary structure (2^0 structure) are the α-helix and β-pleated sheet. Both α-helix and β-pleated sheet patterns are stabilized by hydrogen bonds between the carbonyl and N-H groups in the polypeptide's backbone.

Tertiary structure:

The term tertiary structure (3^0 structure) refers to the unique three dimensional conformations that globular proteins assumes as a consequence of interactions between the side chains in their primary structure. The following types of covalent and non-covalent interactions stabilize tertiary structure:

1. Hydrophobic interactions
2. Electrostatic interactions
3. Hydrogen bonds
4. Vanderwall forces of interaction
5. Covalent bonds (Disulphide bonds)

Quaternary Structure:

Many proteins like hemoglobin, are composed of two or more polypeptide chains and each polypeptide is called a sub-unit. Polypeptide sub-units assemble to form quaternary structure and are held together by non-covalent interactions. The schematic view of all structures of proteins is shown below:

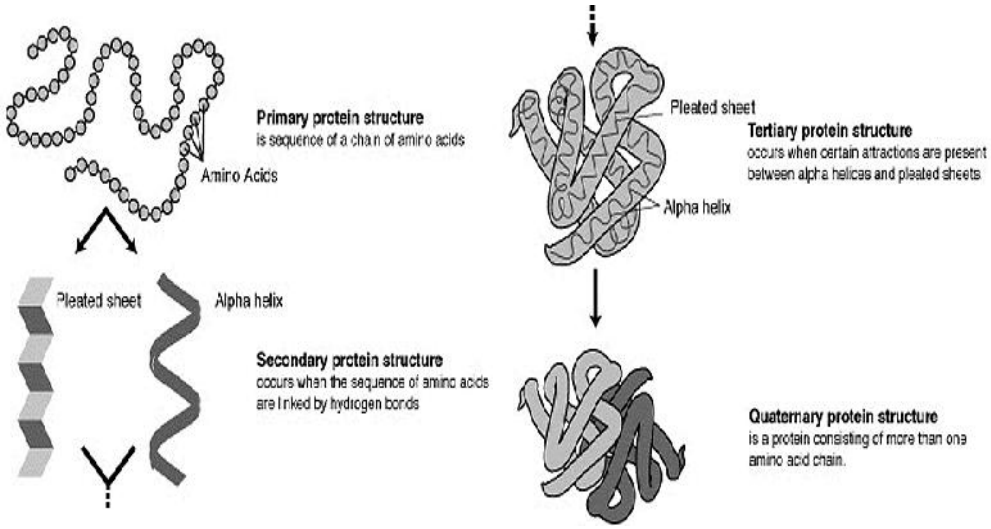


Fig: 13 Structural Organization in proteins

Source: image from the National Human Genome Research Institute (NHGRI)

1.3 The Genetic Code:

The Genetic Code uses three bases to specify each amino acid.

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe UUC Phe	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr	UGU Cys UGC Cys	U C A G
	UUA Leu UUG Leu		UAA Stop UAG Stop	UGA Stop UGG Trp	
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
			CAA Gln CAG Gln		
A	AUU Ile AUC Ile AUA Ile	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn	AGU Ser AGC Ser	U C A G
	AUG Met		AAA Lys AAG Lys	AGA Arg AGG Arg	
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G
			GAA Glu GAG Glu		

Fig: 14 Genetic code

Source:www.cbs.dtu.uk

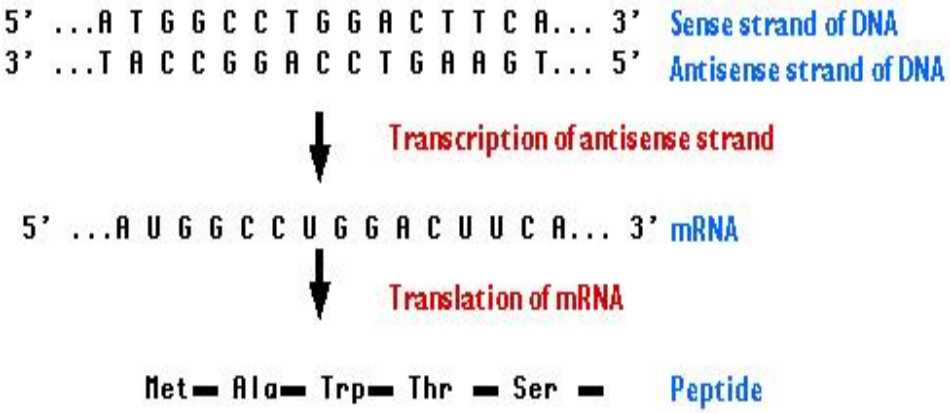
1.4 Central Dogma

The flow of Genetic Information:

This is known as The Central Dogma of Molecular Biology

The Relationship between Genes and Proteins is

- Most genes encode the information for the synthesis of a protein
- The sequence of bases in DNA codes for the sequence of amino acids in proteins



Shown below is an Illustration of the transcription of DNA to RNA to protein which forms the backbone of molecular biology.

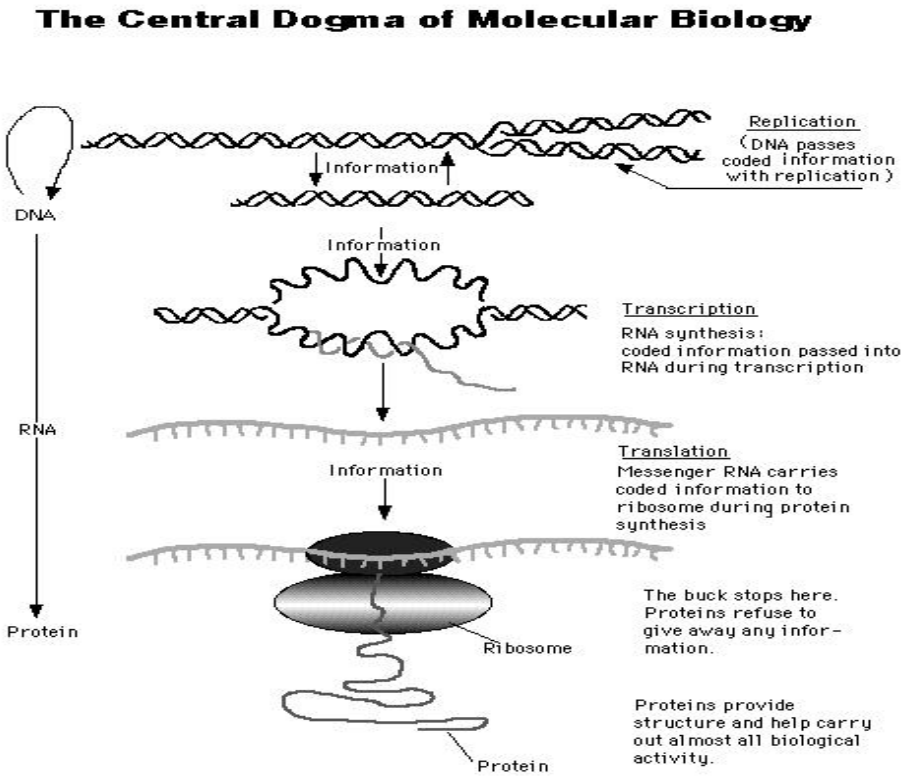


Fig: 15 Central Dogma
Source:http://nuweb.neu.edu/

Or in the words of Francis Crick: *Once information has passed into protein, it cannot get out again.*

Diagrammatic representation of Central dogma inside the cell:

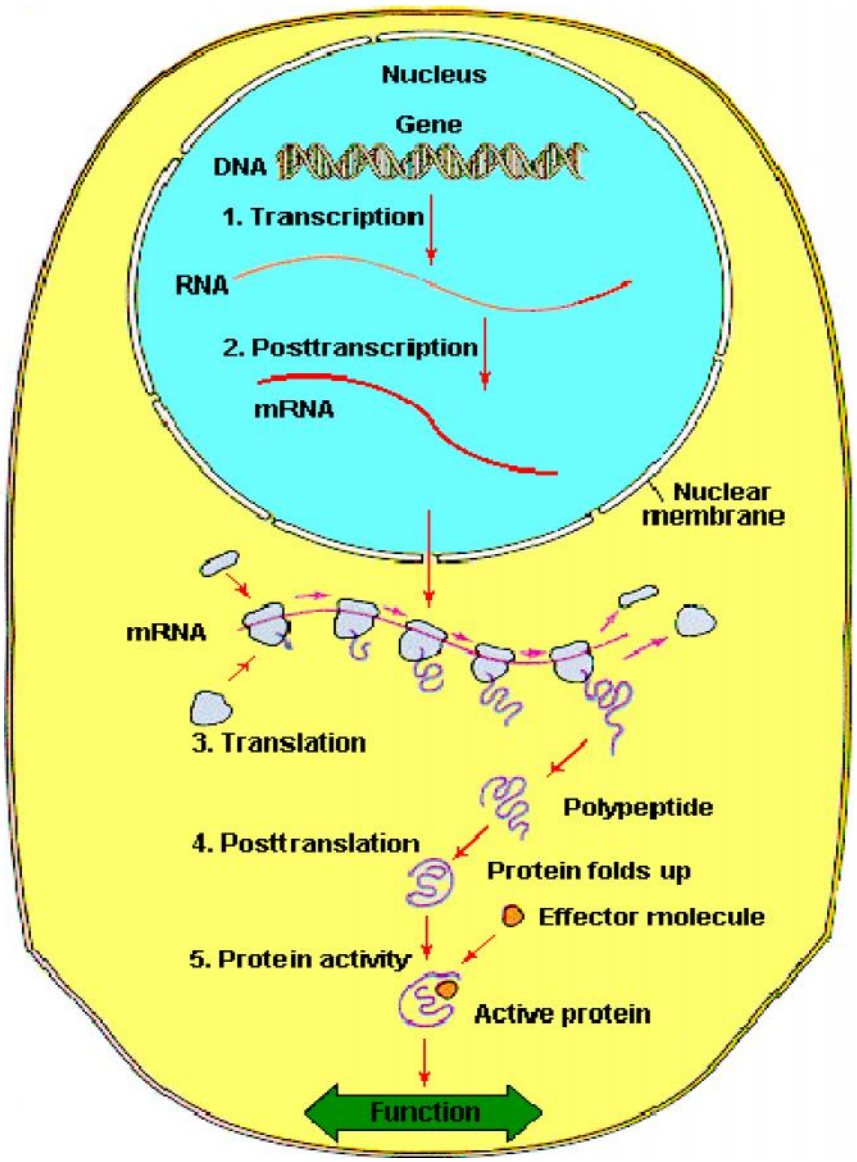
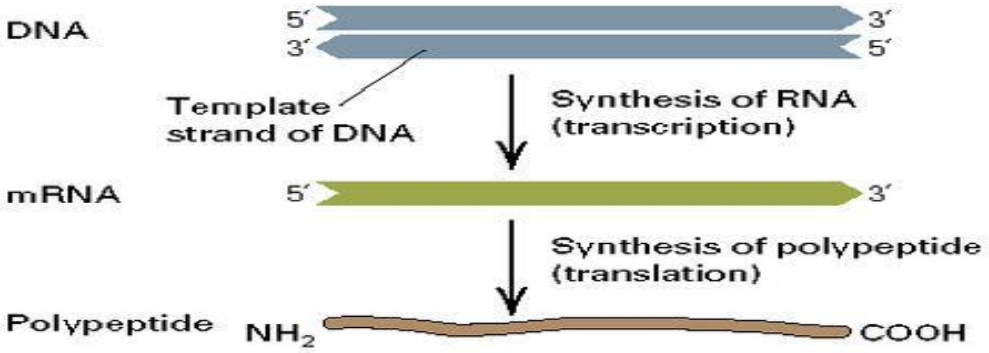


Fig: 16 Central dogma inside the cell

Source: http://nuweb.neu.edu/



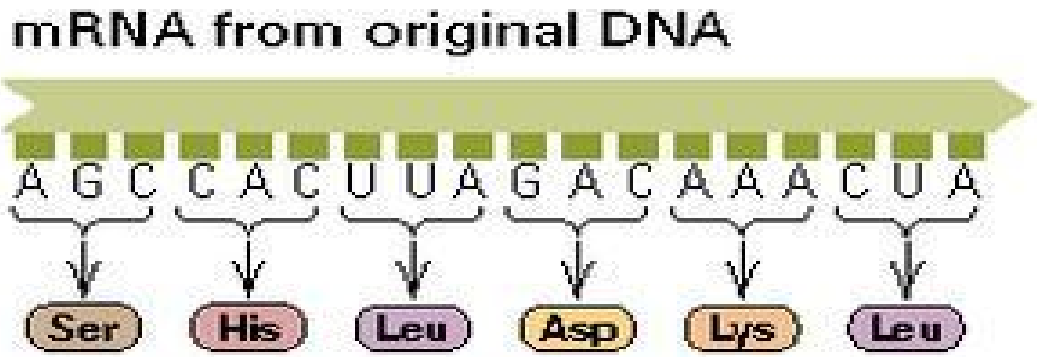


Fig: 17 Central dogma

Source: http://nuweb.neu.edu/

II. Introduction to Bioinformatics

Definition:

Application of computational techniques to understand and organize the information associated with biomolecular molecules.

Or

Application of computer science techniques to solve biological problems

History:

Building on the recognition of the importance of information transmission, accumulation and processing in biological systems, in 1970 Paulien Hogeweg, coined the term "Bioinformatics" to refer to the study of information processes in biotic systems. This definition placed bioinformatics as a field parallel to biophysics or biochemistry (biochemistry is the study of chemical processes in biological systems). Examples of relevant biological information processes studied in the early days of bioinformatics are the formation of complex social interaction structures by simple behavioral rules, and the information accumulation and maintenance in models of prebiotic evolution.

One early contributor to bioinformatics was Elvin A. Kabat, who pioneered biological sequence analysis in 1970 with his comprehensive volumes of antibody sequences released with Tai Te Wu between 1980 and 1991. Another significant pioneer in the field was Margaret Oakley Dayhoff, who has been hailed by David Lipman, Director of the National Center for Biotechnology Information, as the "mother and father of bioinformatics."

At the beginning of the "genomic revolution", the term bioinformatics was re-discovered to refer to the creation and maintenance of a database to store biological information such as nucleotide sequences and amino acid sequences. Development of this type of database involved not only design issues but the development of complex interfaces whereby researchers could access existing data as well as submit new or revised data.

Goals: In order to study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these

activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data. This includes nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analyzing and interpreting data is referred to as computational biology. Important sub-disciplines within bioinformatics and computational biology include:

- The development and implementation of tools that enable efficient access to, use and management of, various types of information
- The development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets. For example, methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences.

The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques to achieve this goal. Examples include: pattern recognition, data mining, machine learning algorithms, and visualization. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, and the modeling of evolution.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.

Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. Bioinformatics is the name given to these mathematical and computing approaches used to glean understanding of biological processes.

Approaches: Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning different DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures.

There are two fundamental ways of modeling a Biological system (e.g., living cell) both coming under Bioinformatics approaches.

Static:

- Sequences – Proteins, Nucleic acids and Peptides
- Interaction data among the above entities including microarray data and Networks of proteins, metabolites

Dynamic:

- Structures – Proteins, Nucleic acids, Ligands (including metabolites and drugs) and Peptides (structures studied with bioinformatics tools are not considered static anymore and their dynamics is often the core of the structural studies)
- Systems Biology comes under this category including reaction fluxes and variable concentrations of metabolites
- Multi-Agent Based modeling approaches capturing cellular events such as signaling, transcription and reaction dynamics

A broad sub-category under bioinformatics is structural bioinformatics.

The Major research areas or the applications of bioinformatics involve the following:

- Biological databases
- Sequence analysis & Genome annotation
- Computational evolutionary biology
- Literature analysis
- Analysis of gene expression & Analysis of regulation
- Analysis of protein expression
- Analysis of mutations in cancer
- Comparative genomics
- Network and systems biology
- High-throughput image analysis
- Structural bioinformatic approaches like Prediction of protein structure, Molecular Interaction, Docking algorithms.

III. Biological Databases – Networking for the Biologist

As biology has increasingly turned into a data-rich science, the need for storing and communicating large datasets has grown tremendously. The obvious examples are the nucleotide sequences, the protein sequences, and the 3D structural data produced by X-ray crystallography and macromolecular NMR. A new field of science dealing with issues, challenges and new possibilities created by these databases has emerged: bioinformatics.

Bioinformatics is the application of Information technology to store, organize and analyze the vast amount of biological data which is available in the form of sequences and structures of proteins (the building blocks of organisms) and nucleic acids (the information carrier). The biological information of nucleic acids is available as sequences while the data of proteins is available as sequences and structures. Sequences are represented in single dimension where as the structure contains the three dimensional data of sequences.

Sequences and structures are only among the several different types of data required in the practice of the modern molecular biology. Other important data types includes metabolic pathways and molecular interactions, mutations and polymorphism in molecular sequences and structures as well as organelle structures and tissue types, genetic maps, physiochemical data, gene expression profiles, two dimensional DNA chip images of mRNA expression, two dimensional gel electrophoresis images of protein expression data. A biological database is a collection of data that is organized so that it contents can easily be accessed, managed, and updated. There are two main functions of biological databases:

Make biological data available to scientists.

As much as possible of a particular type of information should be available in one single place (book, site, and database). Published data may be difficult to find or access and collecting it from the literature is very time-consuming. And not all data is actually published explicitly in an article (genome sequences!).

To make biological data available in computer-readable form

Since analysis of biological data almost always involves computers, having the data in computer-readable form (rather than printed on paper) is a necessary first step.

Data Domains

- Types of data generated by molecular biology research:
  - – Nucleotide sequences (DNA and mRNA)
  - – Protein sequences
  - – 3-D protein structures
  - – Complete genomes and maps
- Also now have:
  - – Gene expression
  - – Genetic variation (polymorphisms)

3.1 Biological Databases

When Sanger first discovered the method to sequence proteins, there was a lot of excitement in the field of Molecular Biology. Initial interest in Bioinformatics was propelled by the necessity to create databases of biological sequences. Biological databases can be broadly classified into sequence and structure databases. Sequence databases are applicable to both nucleic acid sequences and protein sequences, whereas structure database is applicable to only Proteins. The first database was created within a short period after the Insulin protein sequence was made available in 1956. Incidentally, Insulin is the first protein to be sequenced. The sequence of Insulin consisted of just 51 residues (analogous to alphabets in a sentence) which characterize the sequence. Around mid nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases (individual units of nucleic acids) was found out. During this period, three dimensional structures of proteins were studied and the well known Protein Data Bank was developed as the first protein structure database with only 10 entries in 1972. This has now grown in to a large database with over 10,000 entries. While the initial databases of protein sequences were maintained at the individual laboratories, the

development of a consolidated formal database known as SWISS-PROT protein sequence database was initiated in 1986 which now has about 70,000 protein sequences from more than 5000 model organisms, a small fraction of all known organisms. These huge varieties of divergent data resources are now available for study and research by both academic institutions and industries. These are made available as public domain information in the larger interest of research community through Internet (www.ncbi.nlm.nih.gov) and CDROMs (on request from www.rcsb.org). These databases are constantly updated with additional entries.

Databases in general can be classified in to **primary**, **secondary** and **composite** databases. A **primary** database contains information of the sequence or structure alone. Examples of these include Swiss-Prot & PIR for protein sequences, GenBank & DDBJ for Genome sequences and the Protein Databank for protein structures.

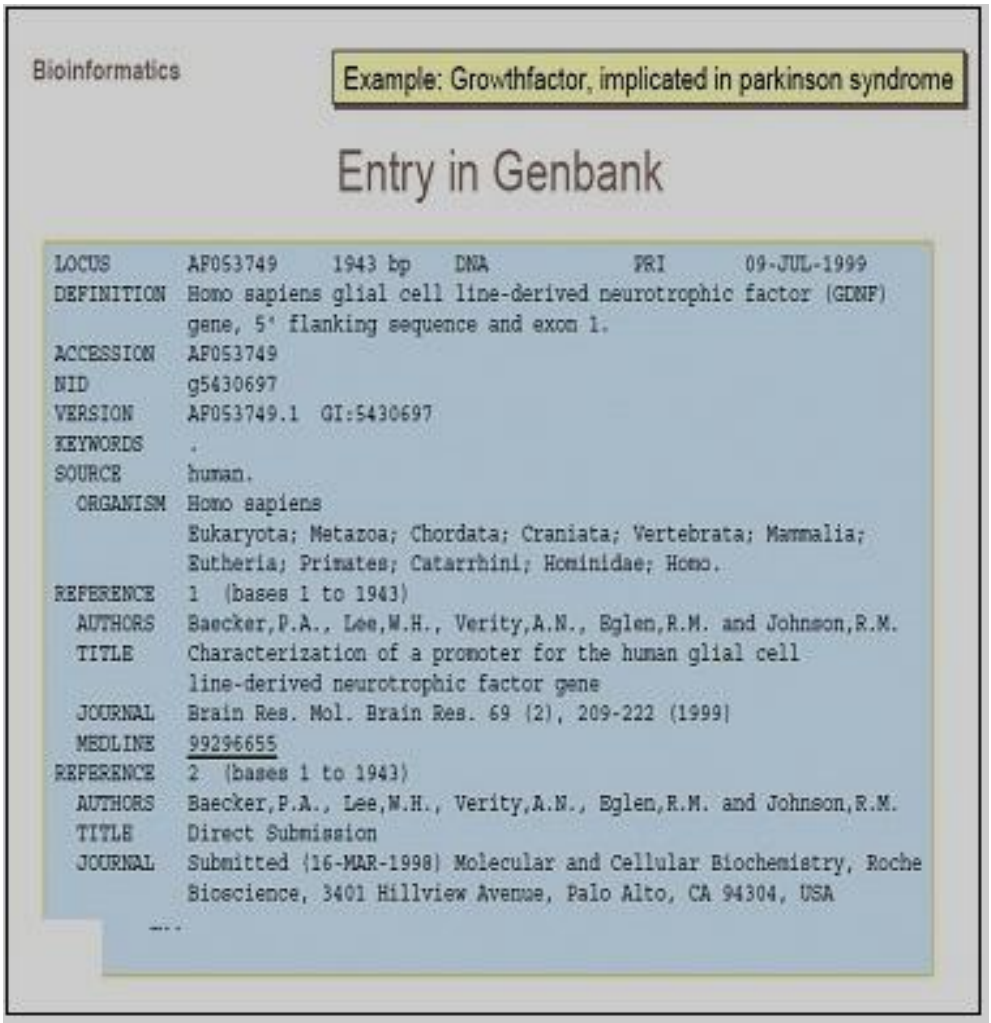
A **secondary** database contains derived information from the primary database. A secondary sequence database contains information like the conserved sequence, signature sequence and active site residues of the protein families arrived by multiple sequence alignment of a set of related proteins. A secondary structure database contains entries of the PDB in an organized way. These contain entries that are classified according to their structure like all alpha proteins, all beta proteins, etc. These also contain information on conserved secondary structure motifs of a particular protein. Some of the secondary database created and hosted by various researchers at their individual laboratories includes SCOP, developed at Cambridge University; CATH developed at University College of London, PROSITE of Swiss Institute of Bioinformatics, eMOTIF at Stanford.

Composite database amalgamates a variety of different primary database sources, which obviates the need to search multiple resources. Different composite database use different primary database and different criteria in their search algorithm. Various options for search have also been incorporated in the composite database. The National Center for Biotechnology Information (NCBI) which hosts these nucleotide and protein databases in their large high available redundant array of computer servers, provides free access to the various persons involved in research. This also has link to OMIM (Online

Mendelian Inheritance in Man) which contains information about the proteins involved in genetic diseases.

Primary Nucleotide Sequence Repository – GenBank, EMBL, DDBJ:

These are three chief databases that store and make available raw nucleic acid sequences. GenBank is physically located in the USA and is accessible through NCBI portal over internet. EMBL (European Molecular Biology Laboratory) is in UK and DDBJ (DNA databank of Japan) is in Japan. They have uniform data formats (but not identical) and exchange data on daily basis. Here we will describe one of the database formats, GenBank, in detail. The access to GenBank, as to all databases at NCBI is through the Entrez search program. This front end search interface allows a great variety of search options.



Bioinformatics

Example: Growthfactor, implicated in parkinson syndrome

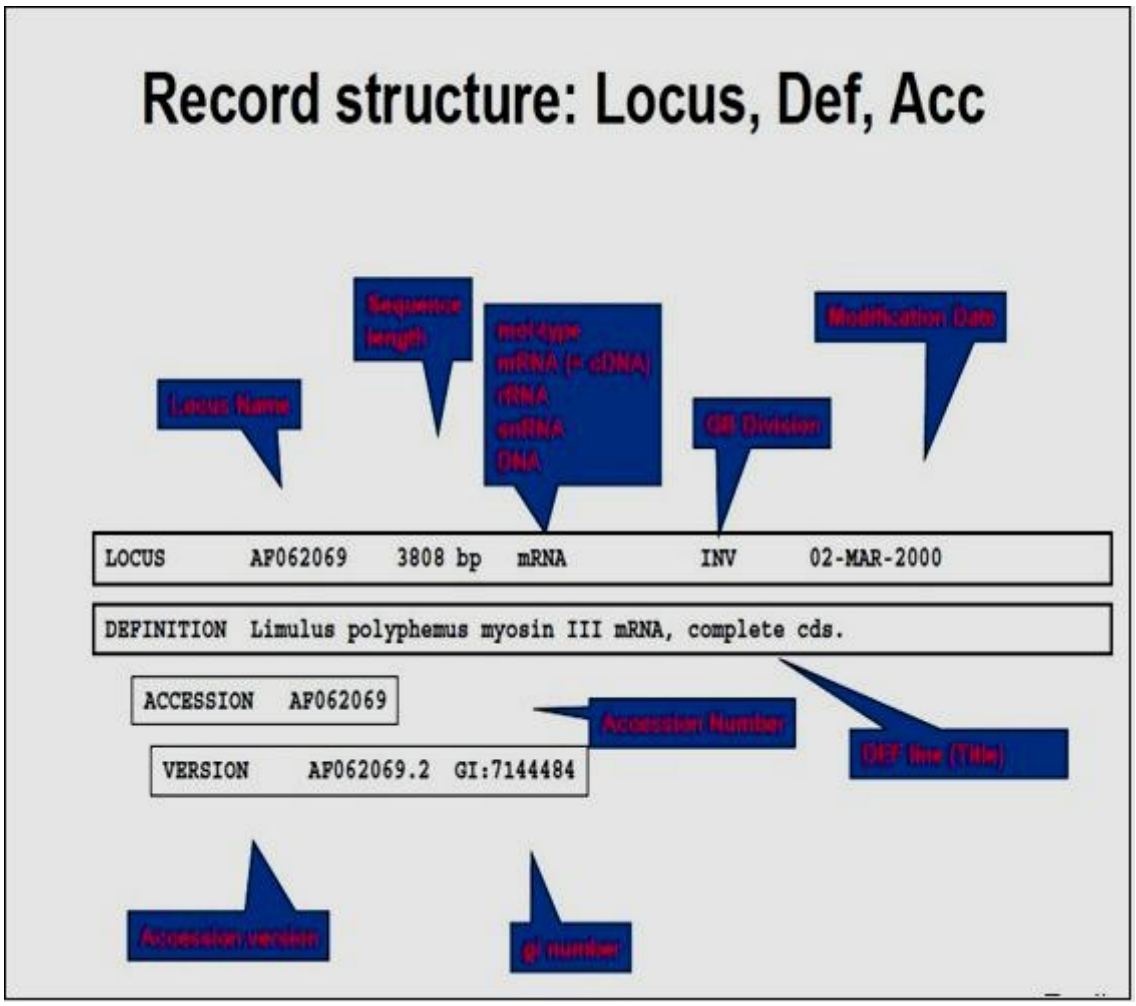
Entry in Genbank

```
LOCUS       AF053749     1943 bp    DNA             PRI       09-JUL-1999
DEFINITION  Homo sapiens glial cell line-derived neurotrophic factor (GDNF)
            gene, 5' flanking sequence and exon 1.
ACCESSION   AF053749
NID         g5430697
VERSION     AF053749.1  GI:5430697
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;
            Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1943)
  AUTHORS   Baecker,P.A., Lee,W.H., Verity,A.N., Eglen,R.M. and Johnson,R.M.
  TITLE     Characterization of a promoter for the human glial cell
            line-derived neurotrophic factor gene
  JOURNAL   Brain Res. Mol. Brain Res. 69 (2), 209-222 (1999)
  MEDLINE   99296655
REFERENCE   2  (bases 1 to 1943)
  AUTHORS   Baecker,P.A., Lee,W.H., Verity,A.N., Eglen,R.M. and Johnson,R.M.
  TITLE     Direct Submission
  JOURNAL   Submitted (16-MAR-1998) Molecular and Cellular Biochemistry, Roche
            Bioscience, 3401 Hillview Avenue, Palo Alto, CA 94304, USA
            ...
```

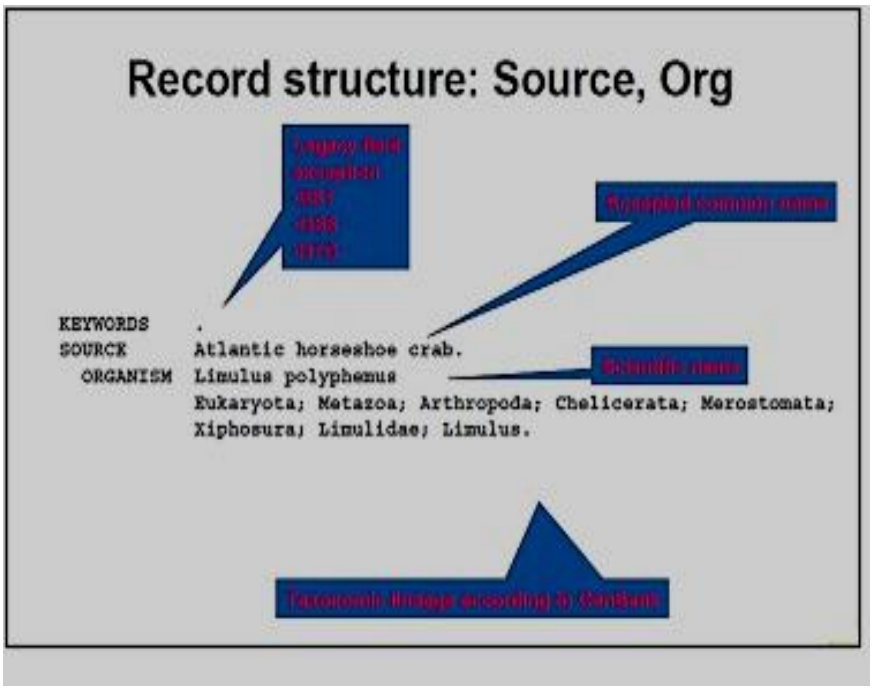
Bioinformatics

Example: Growthfactor, implicated in parkinson syndrome

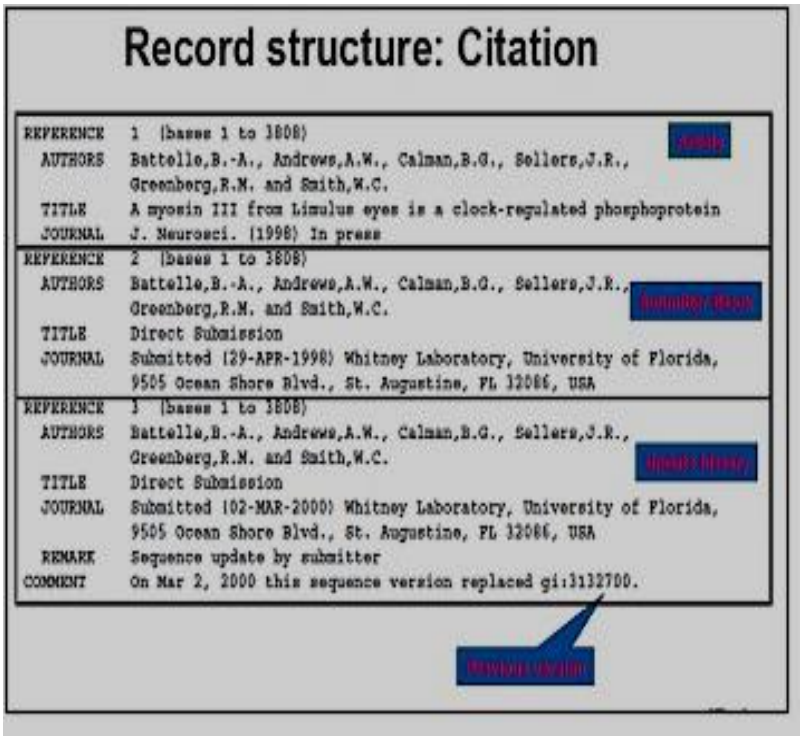
```
FEATURES             Location/Qualifiers
     source          1. .1943
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="5"
                     /map="5p12-p13.1"
     gene            1. .>1943
                     /gene="GDNF"
     misc_feature    1. .1643
                     /gene="GDNF"
                     /note="5' flanking region"
     mRNA            1644. .>1817
                     /gene="GDNF"
                     /product="glial cell line-derived neurotrophic factor"
     5'UTR           1644. .>1817
                     /gene="GDNF"
     exon            1644. .1817
                     /gene="GDNF"
                     /number=1
BASE COUNT      356 a    662 c    576 g    349 t
ORIGIN
     GAATTCAGGT CCAATGGCTT CCGGAAAACA GGTTTCTGCT TAGCAAAGAC ATGCCCTATT        60
     TAGTACATTA TTTTAGAGGT ACAGCCAATT CCATGCCCCA TGTGAATGAA ATGTATTTAT       120
     GGTTATAGCC ATGCACAGGG TGTGTAAGGA CTTGCCCTCC TCCTGTCCTC TACAAAAGAA       180
     GGCTCAGGCA GCTTCTGGTG GTGAACTAAC CAACAAAAGG AATGCCCAGA AGGTCTCACC       240
     TCTCCCATCC ACAGAGCTCT GGAATGGGGG CCGGGCCCCT GATCGCTGGA AACTCAGCAT       300
     CCAAGTGGGC GCTTGCTGAA GTTTCCCATC TGCATTTTCG AAAATCTGGA TAAAAGCAGG       360
     TTTAGCTCAA CCTCCCCTAA CCCGTTCCTG ATAAAGTGAT CTTACGCCTC TGGAATTGGG       420
     ...
```



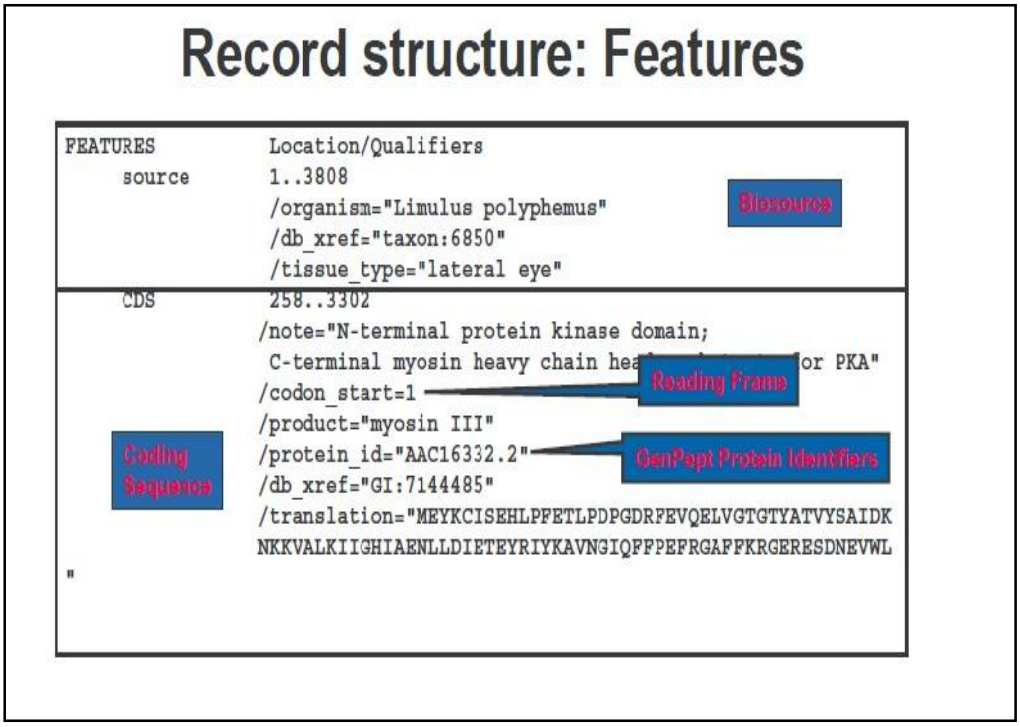
The word accession number defines a field containing unique identification numbers. The sequence and the other information may be retrieved from the database simple by searching for a given accession number. Taking the field names in order, we have first all the word 'LOCUS'. This is a GenBank title that names the sequence entry. Apart for accession number, it also specifies the number of bases in the entry, a nucleic acid type, a code word PRI that indicates the sequence is from primate, and the date on which the entry was made. PRI is one of the 17 keyword search that are used to classify the data. The next line of the file contains the definition of the entry, giving the name of the sequence. The unique accession number came next, followed by a version number in case the entries have gone through more than one version.

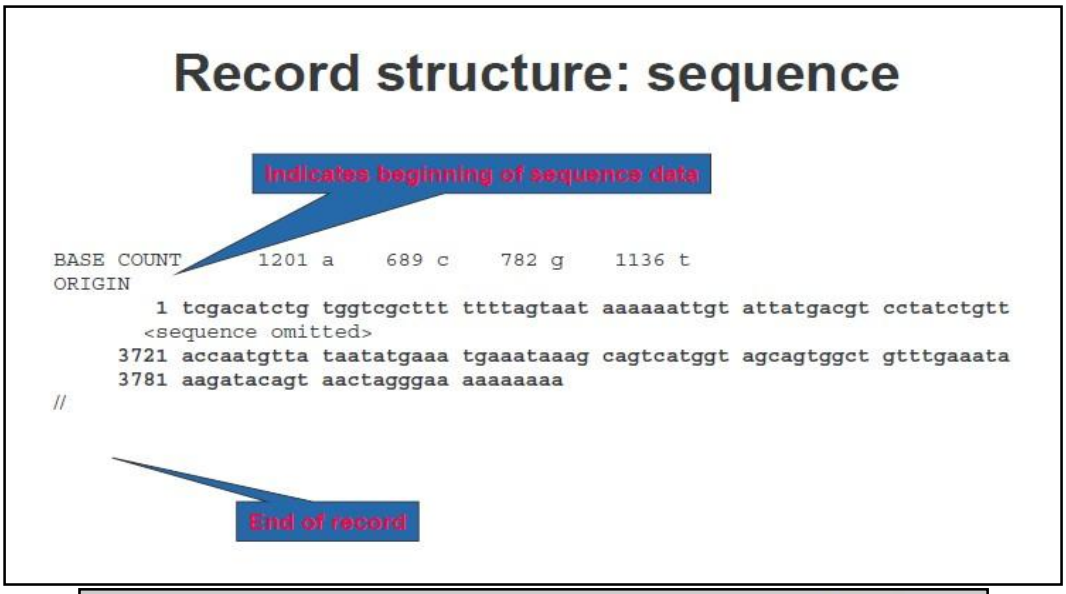


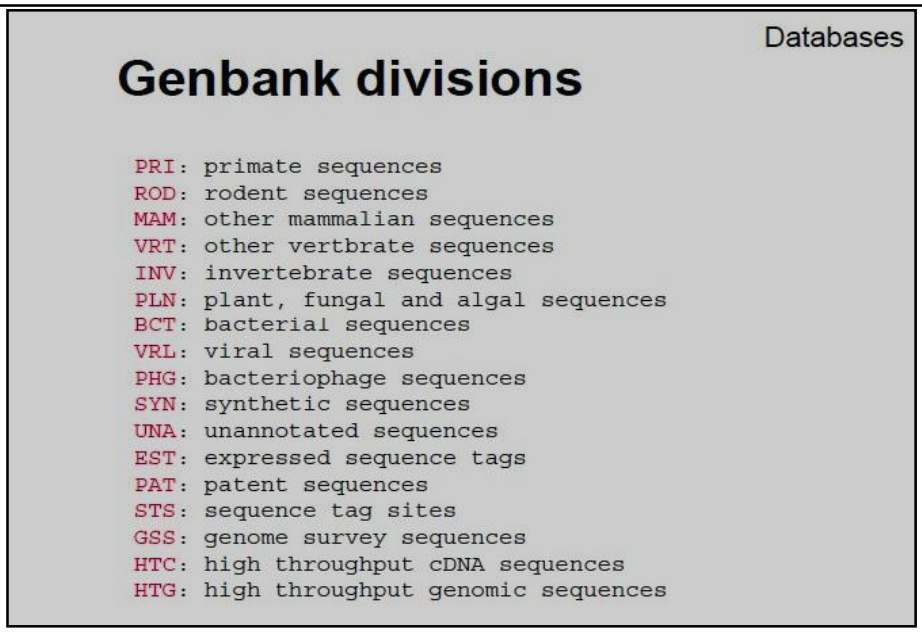
The next item is a list of specially defined keywords that used to index the entries. Next come a set of SOURCE records which describe the organism from which sequence was extracted. The complete scientific classification is given. This is followed by publication details.



In the beginning, sequences were extracted from the published literature and painstaking entered in the database. Each entry was therefore associated with a publication. The features table includes coding region, exons, introns, promoters, alternate splice patterns, mutation, variations and a translation into protein sequence, if it code for one. Each feature may be accompanied by a cross-reference to another database. After the feature table, a single line gives the base count statistics for the sequence. Finally comes the sequence itself. The sequence is typed in lower cases, and for ease of reading, each line is divided into six columns of ten bases each. A single number on the left numbers the bases.







The above description does not cover all the fields used in GenBank, but only the more important ones.

Primary Protein Sequence Repositories

PIR-PSD or protein information resource – protein sequence database, at the NBRF (National Biomedical Research Foundation, USA), and SWISS-PROT at the SBI (Swiss Biotechnology Institute), Switzerland are protein sequence databases.

The PIR-PSD is a collaborative endeavor between the PIR, the MIPS (Munich Information Centre for Protein Sequences, Germany) and the JIPID (Japan International

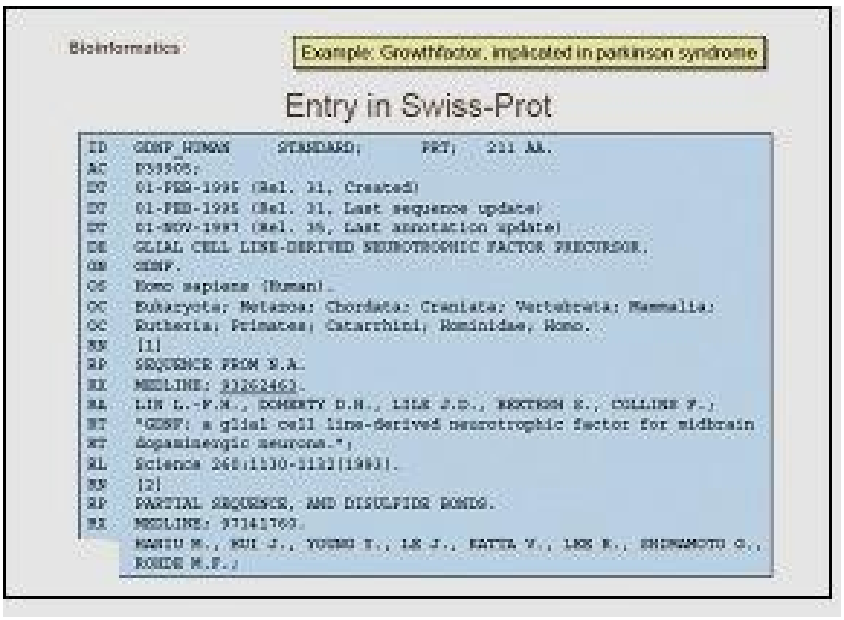
Protein Information Database, Japan). The PIR-PSD is now a comprehensive, non-redundant, expertly annotated, object relational DBMS. It is available at http://pir.georgetown.edu/pirww. A unique characteristic of the PIR-PSD is its classification of protein sequences based on the super family concept. Sequence in PIR-PSD is also classified based on homology domain and sequence motifs. Homology domains may correspond to evolutionary building blocks, while sequence motifs represent functional sites or conserved regions. The classification approach allows a more complete understanding of sequence function structure relationship.

The other well known and extensively used protein database is SWISS- PROT (http://www.expasy.ch/sprot). Like the PIR-PSD, this curated proteins sequence database also provides a high level of annotation. The data in each entry can considered separately as core data and annotation. The core data consists of the sequences entered in common single letter amino acid code, and the related references and bibliography. The taxonomy of the organism from which the sequence was obtained also forms part of this core information. The annotation contains information on the function or functions of the protein, post-translational modification such as phosphorylation, acetylation, etc., functional and structural domains and sites, such as calcium binding regions, ATP-binding sites, zinc fingers, etc., known secondary structural features as for examples alpha helix, beta sheet, etc., the quaternary structure of the protein, similarities to other protein if any, and diseases that may rise due to different authors publishing different sequences for the same protein, or due to mutations in different strains of an described as part of the annotation.

Lines of code in SWISS-PROT database:

Code	Expansion	Remarks
ID	Identification	Occurs at the beginning of the entry. Contains a unique name for the entry, plus information on the status of the entry. If it has been checked and conforms to SWISS-PROT standards, it is called STANDARD.

AC	Accession numbers	This is a stable way of identifying the entry. The name may change but not the AC. If the line has more than one number, it means that the entry was constituted by merging other entries.
DT	Date	There are three dates corresponding to the creation date of the entry and modification dates of the sequence and the annotation respectively
DE	Description	Lines that start with the identifier contain general description about the sequence.
GN	Gene name	The name of the gene (or genes) that codes for the protein
OS, OG,OC	Organism name, Organelle, Organism classification	The name and taxonomy of the organism, and information regarding the organelle containing the gene e.g. mitochondria or chloroplast, etc.
RN, RP,RX,RART,RL	Reference number, Position, comments, cross-reference, authors, title and location.	Bibliographic reference to the sequence. This includes information (following the code RP) on the extent of work carried out by the authors.
CC	Comments	These are free text comments that provide any relevant information pertaining to the entry.
DR	Database cross-reference	This line gives cross-references to other databases where information regarding this entry is also found. As for example to structural information for the protein in the PDB.
KW	Keywords	This line gives a list of keywords that can be used in indexes. Search programs very often simply go through such indices to identify required information
FT	Features Table	These lines describe regions or sites of interest in the sequence, e.g. post-transitional modifications, binding sites, enzyme active sites and local secondary structures
SQ	Sequence Header	This line indicates the beginning of the sequence data and gives a brief summary of its contents.



Both PIR-PSD and SWISS-PROT have software that enables the user to easily search through the database to obtain only the required information. SWISS-PROT has the SRS or the sequence retrieval system that searches also through the other relevant databases on the site, such as TrEMBL.

TrEMBL (for Translated EMBL) is a computer-annotated protein sequence database that is released as a supplement to SWISS-PROT. It contains the translation of all coding sequences present in the EMBL Nucleotide database, which have not been fully annotated. Thus it may contain the sequence of proteins that are never expressed and never actually identified in the organisms.

Derived or Secondary databases of nucleotide sequences

Many of the secondary databases are simply sub-collection of sequences culled from one or the other of the primary databases such as GenBank or EMBL. There is also usually a great deal of value addition in terms of annotation, software, presentation of the information and the cross-references. There are other secondary databases that do not present sequences at all, but only information gathered from sequences databases.

An example of the former type of database is the FlyBase or The Bereley Drosophila Genome Project (http://www.fruitfly.org). A consortium sequenced the entire genome of the fruit fly *D. Melanogaster* to a high degree of completeness and quality.

Another database that focuses on a single organism is ACeDB. More than a database, this is a database management system that was originally developed for the *C. Elegans* (a nematode worm) genome project. It is a repository of not only the sequence, but also the genetic map as well as phenotypic information about the *C. Elegans* nematode worm.

The comprehensive Microbial Resource maintained by TIGR (The Institute for Genomic Research) at http://www.tigr.org allows access to a database called Omniome. This contains all the focus on one organism. Omniome has not only the sequence and annotation of each of the completed genomes, but also has associated information about the organisms (such as taxon and gram stain pattern), the structure and composition of their DNA molecules, and many other attributes of the protein sequences predicted from the DNA sequences. The presence of all microbial genomes in a single database facilitated meaningful multi-genome searches and analysis, for instance, alignment of entire genomes, and comparison of the physical proper of proteins and genes from different genomes etc.

A database of the genomes of mitochondria and other such organelles is available at the Organelle Genome Database at the University of Montreal, Canada, and is called GOBASE (http://megasun.bch.umontreal.ca/gobase).

Derived or Secondary databases of amino acid sequences - Subcollections

Another family of a database focused on a particular family protein is GPCRGB (http://rose.man.pozen.pl/aars/). These are transmembrane protein used by cells to communicate with the outside world. They are involved in vision, smell, hearing, taste and feeling. GPCRGB is in fact more than a collection of sequences of the protein family. It includes additional data on multiple sequences alignments. Ligands and ligands binding data, 3D models, mutation data, literature reference, disease patterns, cell lines, protocols, vectors etc. It is fully integrated information system with data, and browsing and query tools.

MHCPep (http://wehih.wehi.edu.au/mhcpep/) is a database comprising over 13000 peptide sequences known to bind the Major Histocompatibilty Complex of the immune system. Each entry in the database contains not only the peptide sequence, which may be 8 to 10 amino acid long, but in addition has information on the specific MHC molecules

to which it binds, the experimental method used to assay the peptide, the degree of activity and the binding affinity observed , the source protein that, when broken down gave rise to this peptide along with other, the positions along the peptide where it anchors on the MHC molecules and references and cross links to other information.

The CluSTr (Cluster of SWISS-PROT and TrEMBL proteins at http://ebi.ac.uk.clustr) database offers an automatic classification of the entries in the SWISS-PROT and TrEMBL databases into groups of related proteins. The clustering is based on the analysis of all pair wise comparisons between protein sequences.

Similar to CluSTRr is the COGS or Cluster of Orthologous Groups of database that is accessible at htp://ncbi.nlm.nih.gov/COG. An orthologous group of proteins is one in which the members are related to each other by evolutionary descent. Such orthology may not be just from one protein to another, and then to another and so on down the line. It may involve one-to-many and many-to-many evolutionary relationships, and hence the term 'groups'. COGS is thus a database of phylogenetic relationships. The approximately 2500 groups have been divided into 17 broad categories. The utility of COGS, as of CluSTr, is that it helps assign function to new protein sequences without going through tedious biochemical discovery processes.

Derived or Secondary databases of amino acid sequences – Patterns and Signature

A set of databases collects together patterns found in protein sequences rather than the complete sequences. The patterns are identified with particular functional and/or structural domains in the protein, such as for example, ATP binding site or the recognition site of a particular substrate. The patterns are usually obtained by first aligning a multitude of sequences through multiple alignment techniques. This is followed by further processing by different methods, depending on the particular database.

PROSITE is one such pattern database, which is accessible at http://www.expasy.ch/prosite. The protein motif and pattern are encoded as "regular expressions". The information corresponding to each entry in PROSITE is of the two forms – the patterns and the related descriptive text. The regular expression is placed in a format reminiscent of the SWISS-PROT entries, with a two letter identifier at beginning

of the each line specifying the type of information the line contains. The expression itself is placed on line identified by "PA". The entry also contains references and links to all the proteins sequences that contains that pattern. The related descriptive text is placed in a documentation file with the accession number making the connection to the expression data.

In the PRINTS database (http://www.bioinfo.man.ac.uk/dbbrowser/PRINTS), the protein sequence patterns are stored as 'fingerprints'. A finger print is a set of motifs or patterns rather than a single one. The information contained in the PRINT entry may be divided into three sections. In addition to entry name, accession number and number of motifs, the first section contains cross links to other databases that have more information about the characterized family. The second section provides a table showing how many of the motifs that make up the finger print occurs in the how many of the sequences in that family. The last section of the entry contains the actual finger prints that are stored as multiply aligned set of sequences, the alignment being made without gaps. There is therefore one set of aligned sequences for each motif.

The ProDom protein domain database (http://www.toulouse.inrs.fr/prodom.html) is a compilation of homologous domains that have been automatically identified sequence comparison and clustering methods using the program PSI-BLAST. No identification of patterns is made.. The focus is here to look for complete and self-contained structural domains and the search methods includes signals for such features. A graphical user interface allows easy interactive analysis of structural and therefore functional homology relationships among protein sequences.

A database called Pfam contains the profiles used using Hidden markov models (http://www.sanger.ac.uk/Software/Pfam). HMMs build the model of the pattern as a series of match, substitute, insert or delete states, with scores assigned for alignment to go from one state to another. Each family or pattern defined in the Pfam consists of the four elements. The first is the annotation, which has the information on the source to make the entry, the method used and some numbers that serve as figures of merit. The second is the seed alignment that is used to bootstrap the rest of the sequences into the multiple alignments and then the family. The third is the HMM profile. The fourth element is complete alignment of all the sequences identified in that family.

Structure Databases:

Structure databases like sequence databases comes in two varieties, primary and secondary. Strictly speaking there is only one database that stores primary structural data of biological molecules, namely the PDB. In the context of this database, term macromolecule stretches to cover three orders of magnitude of molecular weight from 1000 Daltons to 1000 kilo Daltons Small biological and organic molecules have their structures stored in another primary structure database the CSD, which is also widely used in biological studies. This contains the three dimensional structure of drugs, inhibitors and fragments or monomers of the macromolecule.

The primary structure database - PDB and CSD

PDB stands for Protein Databank. In spite of the name, PDB archive the three-dimensional structures of not only proteins but also all biologically important molecules, such as nucleic acid fragments, RNA molecules, large peptides such as antibiotic gramicidin and complexes of protein and nucleic acids. The database holds data derived from mainly three sources. Structure determined by X-ray crystallography form the large majority of the entries. This is followed by structures arrived at by NMR experiments. There are also structures obtained by molecular modeling. The data in the PDB is organized as flat files, one to a structure, which usually means that each file contain one molecule, or one molecular complex.

The Cambridge Structural Database (CSD) was originally a project of the University of Cambridge, which is set up to collect together the published three-dimensional structure of small organic molecules. This excludes proteins and medium sized nucleic acid fragments, but small peptides such as neuropeptides, and monomer and dimers of nucleic acid finds a place in the CSD. Currently CSD holds crystal structures information for about 2.5 lakhs organic and metal organic compounds. All these crystal structures have been obtained using X-ray or neuron diffraction technique. For each entry in the CSD there are three distinct types of information stored. These are categorized as bibliographic information, chemical connectivity information and the three- dimensional coordinates. The annotation data field incorporates all of the bibliographic material for

the particular entry and summarized the structural and experimental information for the crystal structure.

Derived or Secondary databases of bimolecular structures

NDB stands for Nucleic acid data bases. It is a relational database of three-dimensional structures containing nucleic acid. This encompasses DNA and RNA fragments, including those with unusual chemistry such as NDB, and collections of patterns and motifs such as SCOP, PALI etc. The structures are the same as those found in the PDB and therefore the NDB qualifies to be called a specialized sub collection. However a substantial amount, unlike the PDB, the NDB is much more than just a collection of files. The structure of DNA has been classified into A, B and Z polymorphic forms, based on the information specified by authors. Other classes include RNA structures, unusual structures and protein-nucleic acid complexes. These classes of structures are arranged in the form of an ATLAS of Nucleic acid containing structures, which can be browse and searched to obtain the structure or structures required. Each entry in the ATLAS has information on the sequence, crystallization condition, references and details of the parameters and the figures of the merit used in structure solution. The entry has links not only to the coordinated but also to automatically generated graphical views of the molecule. NDB also have archives of structural geometries calculated for all the structures or for a subset of them. And finally, the database stores average geometrical parameters for nucleic acids, obtained by statistical analysis of the structures. These parameters are widely used in computer simulations of nucleic acids and their interactions. The NDB may be accessed at http://ndbserve.rutgers.edu/NDB/.

The SCOP database (Structural Classification of Proteins: http://scop.mrc-lmb.cam.ac.uk/scop/) is a manual classification of protein structures in a hierarchical scheme with many levels. The principal classes are the family, the super family and the fold. SCOP is a searchable and browsable database. In other words, one may either enter SCOP at the top of the hierarchy or examine different folds and families as one pleases, or one may supply a keyword or a phrase to be search the database and retrieve corresponding entries. Once a structure, or a set of structures, has been selected, they may be obtained or viewed wither as graphical images. Each entry also has other annotation

regarding function, etc., and links to other databases, including other structural classification such as CATH.

CATH stands for Class, Architecture, Topology and Homologous super family. The name reflects the classification hierarchy used in the database. The structures chosen for classification are a subset of PDB, consisting of those that have been determined to a high degree of accuracy.

Conclusion

The present challenge is to handle a huge volume of data, such as the ones generated by the human genome project, to improve database design, develop software for database access and manipulation, and device data-entry procedures to compensate for the varied computer procedures and systems used in different laboratories. There is no doubt that Bioinformatics tools for efficient research will have significant impact in biological sciences and betterment of human lives.

IV. Sequence Submission tools

Not only does the research worker want to query, retrieve and analyse sequences, occasionally they also want to submit their own sequences to the databanks be it GenBank or EMBL. The three major organizations that collect sequence information work in collaboration with each other so that sequences entered into GenBank are transferred daily by FTP to both EBI and DDBJ (and vice versa) in an attempt to keep the major databases synchronized. At any given time the three institutes are continually swapping data so it is a false idea to believe that any one database is more current than the other. All three institutes have online methods of submitting sequence data through the Web. The NCBI were the first to come online with BANKIT. The EBI then followed with WEB1N and the Japanese at DDBJ have Sakuara, It should also be noted that the NCBI developed a stand-alone program for MAC'S, PC's, and Unix called SEQUIN that allows the end-user to enter their data from a personal computer and to send the submission via e-mail or to simply post the disk to the appropriate institute where it is then uploaded into the database. Sequin is strongly recommended if you have bulk submissions to make,

4.1 Bankit at NCBI

Bankit is convenient for quick submission of sequence data to the NCBI. Bankit allows you to enter sequence information into a form, edit as necessary, and add biological annotation (e.g. coding regions, mRNA features). Bankit transforms your data into GenBank format for you to review and when your record is completed, it can be submitted directly to GenBank. You have the option of adding information by using text boxes to describe in your own words the source of the sequence and its biological features. The entry screen from Bankit is shown in *Figure* 4.1. The GenBank annotation staff reviews the submitted textual information, incorporates it into the appropriate structured fields, and returns the record by e-mail for your review.

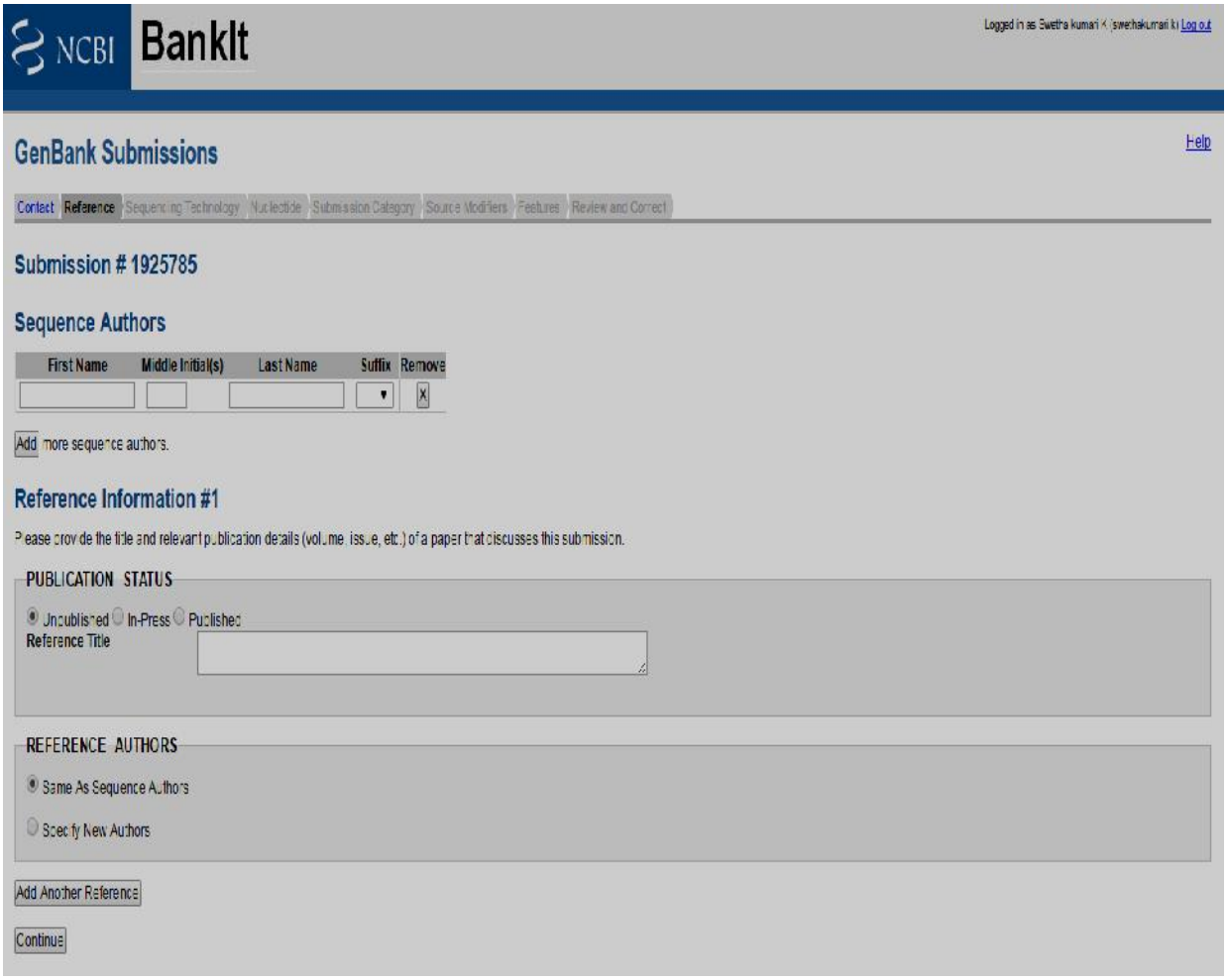


Fig: 4.1 Bankit Submission page

4.2 Sequin from NCBI

Sequin is stand-alone program for the MAC, PC/Windows and UNIX. Sequin is an interactive, graphically oriented program based on screen forms and controlled vocabularies that guide you through the process of entering your sequence and providing biological and bibliographic annotation. Sequin is designed to simplify the sequence submission process and to provide graphical viewing and editing options. This program is optimal for submitting multiple sequences, mutation studies, phylogenetic sets, population sets, and segmented sets. It incorporates robust error checking and accommodates very long sequences and complex annotations. Although Sequin has been implemented by the NCBI, the opening screen allows you to select which database you would like to submit you sequence to be it GenBank, EMBL, or DDBJ. Usually when a sequence is submitted there may be a process whereby the submitter has to be in contact

with the annotators of the sequence by telephone to clarify certain details. Therefore it is wise to choose a submission centre in your geographical region if you want to avoid long distance telephone calls. A screen capture of Sequin is shown in Figure 4.2, Once you have completed the submission depending on which database you have selected at the beginning you will be prompted to send an e-mail to *gbsub@ ndri.nlm.nih.gov* for NCBI, *datasubs@ebi.ac.uk* for EMBL or *ddbjsub@ddbj.nig.ac.jp* for DDBJ. Sequin runs on Macintosh, PC/Windows, and UNIX computers. The program itself, along with its on-line help documentation, is available by anonymous FTP from the *EBI (UK)* at ftp://ftp.ebi.ac.uk/pub/software/sequin/ or from the NCBI *(USA)* at ftp://ncbi.nlm.nih.gov/sequin/ a useful FAQ to help you if you run into problems during submission can be found at. http://www.ebi.ac.uk/~sterk/sqndocs/index.html

Fig: 4.2 Sequin Submission page

4.3 Webin from EBI

The EBI WWW tool (Webin) guides the user through a sequence of WWW forms allowing the user to submit sequence data and descriptive information in an interactive and easy way. All the information required to create a database entry will be collected during this process

(a) Submitter information.

(b) Release date information,

(c) Sequence data, description, and source information.

(d) Reference citation information.

(e) Feature information (e.g. coding regions, regulatory signals etc.).

Data submissions are usually processed within two working days of receipt and the authors are sent notification of their accession number(s). Authors will be asked whether their submitted data can be made available to the public immediately or whether they should be withheld until an author-specified date. Data are never withheld after publication. Once a database entry has been created from a submission, a copy is sent to the submitter for their reference and for comments or corrections. However, it often happens that the entry is correct when it is created but, with the passage of time, becomes out of date. The authors may make corrections to the sequence itself, or may discover new features of the sequence. Since such findings are often not published, the only way to keep entries correct and up to date is if the authors communicate their new findings to the database. At the EBI this can be done by completing an update form available from the Anonymous FTP, site FTP.EBI.AC.UK in the file: pub/databases/embl/release/ update.doc or via the WWW at the URL http://www.ebi.ac.uk/ebi_docs/update.html. A new service that has been instituted at EBI is scanning for vectors before submission for potential vector contamination by running a BLASTN search against EMVEC, a vector database containing information on more than 2000 vectors from the EMBL/GenBank/DDBJ Database SYN(thetic) division. The results will list sequences producing significant alignments and associated information like vector name, score, alignment, etc. The EB1 suggests that you remove vector contamination from your sequence data before submitting to the database. Screen capture of Webin is shown in Figure 4.3.

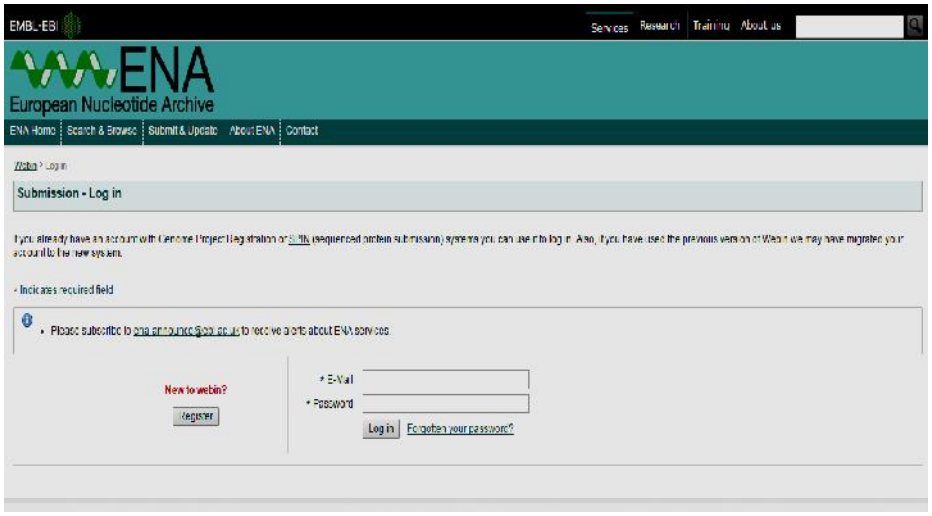


Fig: 4.3 Webin Home page

4.4 Nucleotide Sequence Submission System (NSSS) from DDBJ:

SAKURA is a web-based DNA data submission system from DDBJ. The name SAKURA has been changed to NSSS Nucleotide Sequence Submission System. The URL for NSSS http://www.ddbj.nig.ac.jp/sub/websub-e.html which can be accessed from the DDBJ Home Page (http://www.ddbj.nig.ac.jp)

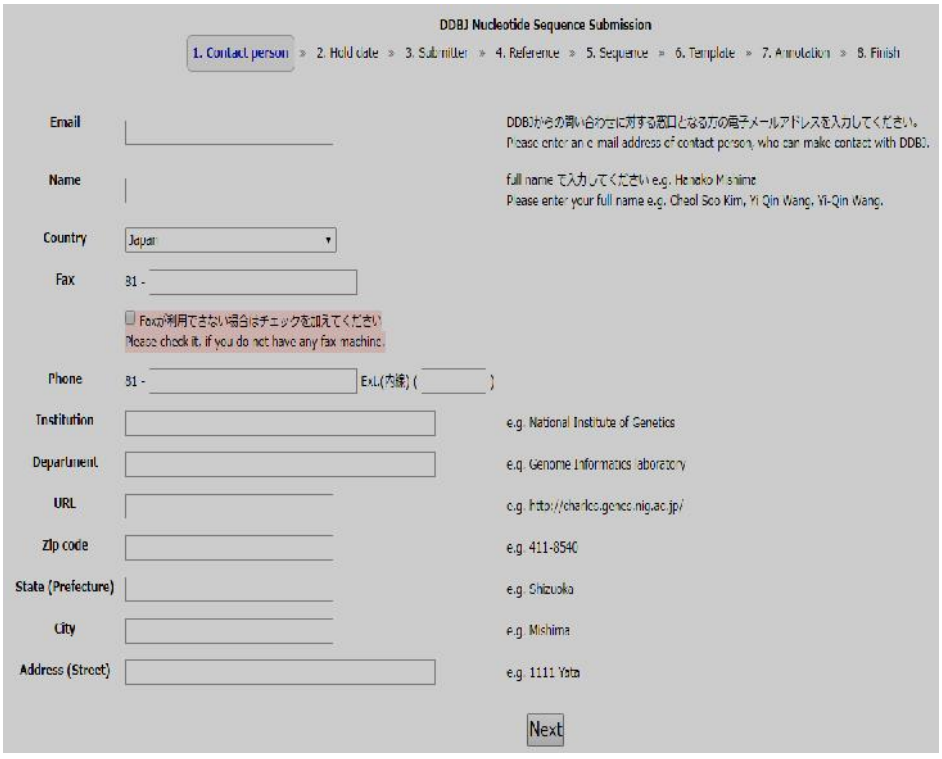


Fig: 4.4 NSSS Home page

You can select either the English or Japanese version. However, data input must be done in English only, regardless of language version selected. NSSS allows you to save your document before completion and submit multiple sequences sequentially.

Conclusion:

Historically there has been collaboration between EBI, NCBI, and DDBJ, These three sites are still the only places that have the infrastructure set up to handle the submission of nucleotide sequences to the databases, be they EMBL or Genbank or DDBJ. For this reason they are also looked upon as the only places where you can do queries and retrieval, or perform homology searches, or multiple sequence alignments. This is no longer true and with the advent of EMBnet, many of the national nodes are able to supply services that are not offered by the major centers. These three major centers have a policy of making all of their databases publicly available, and when distributed network of databases exists in many different parts of the globe then it can only be for the benefit of molecular biologists worldwide.

V. Sequence Analysis

1. Aim: To perform pairwise and multiple sequence alignment of the given query sequence (Nucleotide/Protein) using BLAST and CLUSTALW respectively.

2. Introduction:

Sequence analysis in molecular biology includes a very wide range of relevant topics:

- The comparison of sequences in order to find similarity, often to infer if they are related (homologous)
- Identification of intrinsic features of the sequence such as active sites, post translational modification sites, gene-structures, reading frames, distributions of introns and exons.
- Identification of sequence differences and variations such as point mutations and single nucleotide polymorphism (SNP) in order to get the genetic marker.
- Revealing the evolution and genetic diversity of sequences and organisms
- Identification of molecular structure from sequence alone

Pairwise Sequence alignment:

BLAST is the tool used for homology search. BLAST stands for Basic Local Alignment Search Tool. The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. There are different types of BLAST programs depending upon type of query sequence.

1. Nucleotide Blast Search : a nucleotide database using nucleotide query
2. Protein blast Search : protein database using a protein query
3. blastx Search : protein database using a translated nucleotide query
4. tblastn Search : translated nucleotide database using a protein query
5. tblastx Search : translated nucleotide database using a nucleotide query

Multiple Sequence Alignment:

ClustalW2:

ClustalW2 is a widely used multiple sequence alignment computer program. ClustalW2 is now retired and Clustal omega has been introduced. Clustal omega is a general purpose multiple sequence alignment program for DNA or proteins. It attempts to calculate the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen.

3. Requirements:

System Configuration: Windows 7 operating system with Internet

Input requirements: Nucleotide/Protein Sequence

4. Methodology:

BLAST:

- ✓ Retrieve query nucleotide/protein sequence from NCBI
- ✓ Paste the retrieved sequence in the input form of Blast home page
- ✓ The blastn/blastp will search for similar DNA/proteins to the query against non-redundant protein sequences.
- ✓ The obtained results are interpreted

CLUSTAL Omega:

- ✓ Retrieve five JHEH sequences obtained from BLAST output
- ✓ Open the clustal omega home page
- ✓ Paste the set of JHEH DNAs from different insects
- ✓ Enter your sequences option and click on ‘Run’.
- ✓ The whole set of sequences were aligned by default program parameters in clustal omega, and the alignment file is displayed.

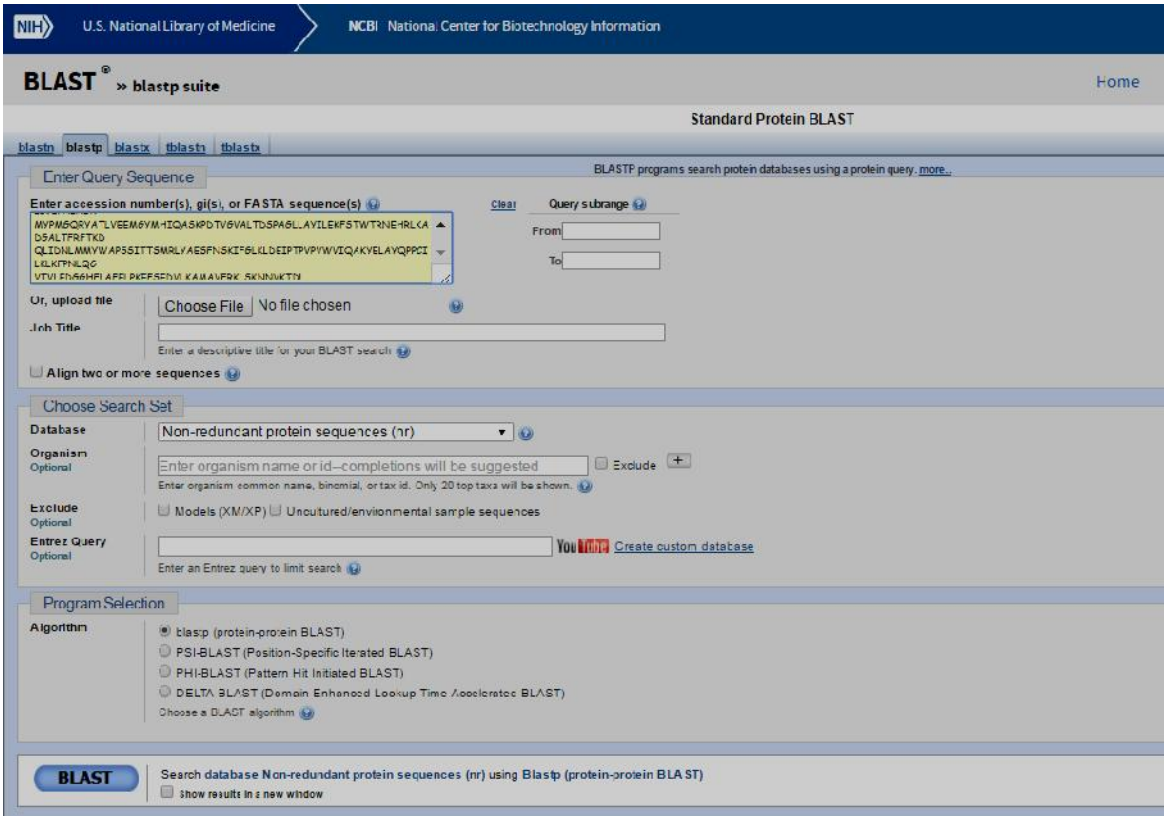


Fig: 5.1 Blast Home page

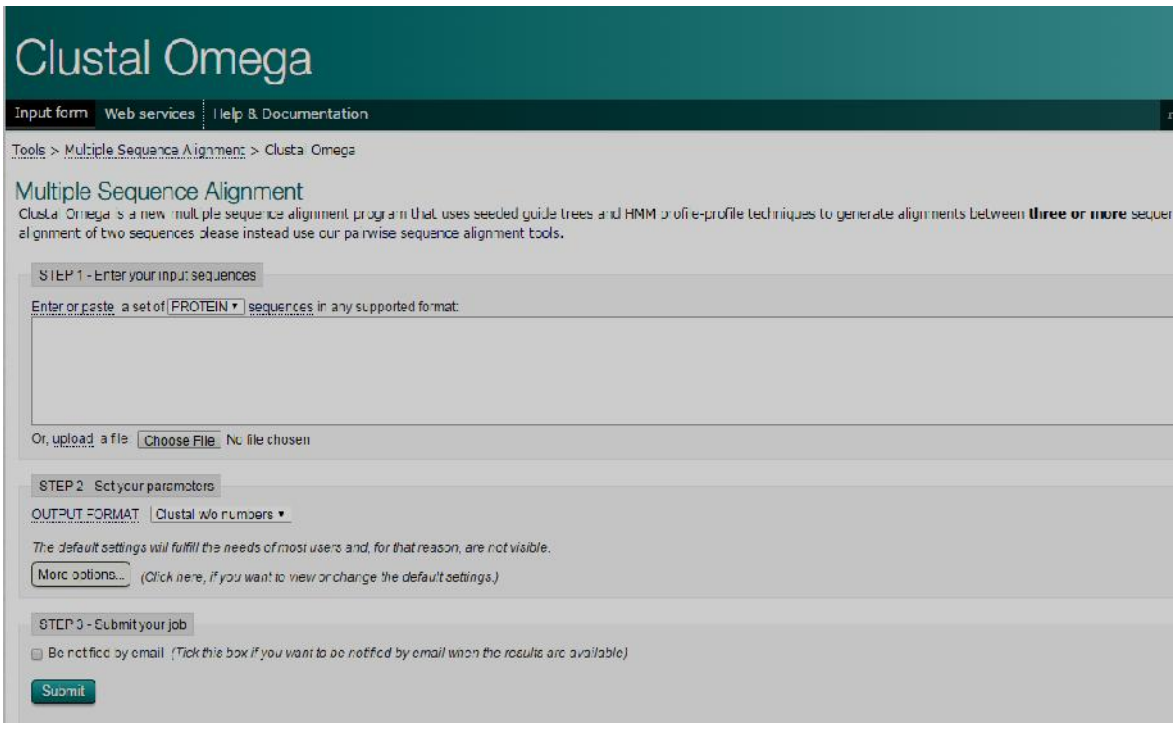


Fig: 2 Clustal W Home page

5. Results and Discussion:

In BLAST output, results are ranked within their alignment score. Score (result´s punctuation), query coverage (how much of the query is aligned), e-value (result´s significance) and %identity are features to be taken into account. Check if a result is good or not in their accordance to evolution.

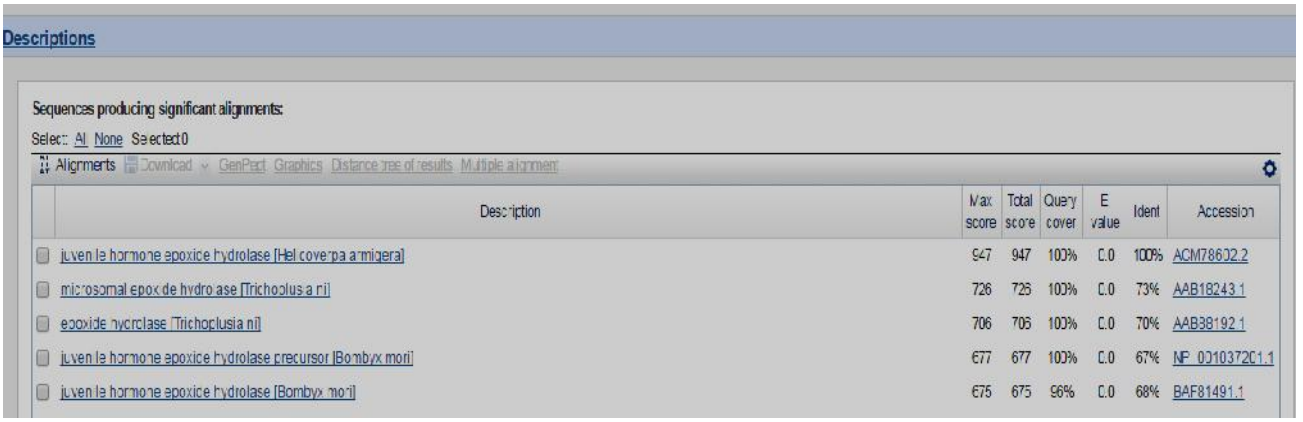


Fig: 5.3 Blast output showing homologues

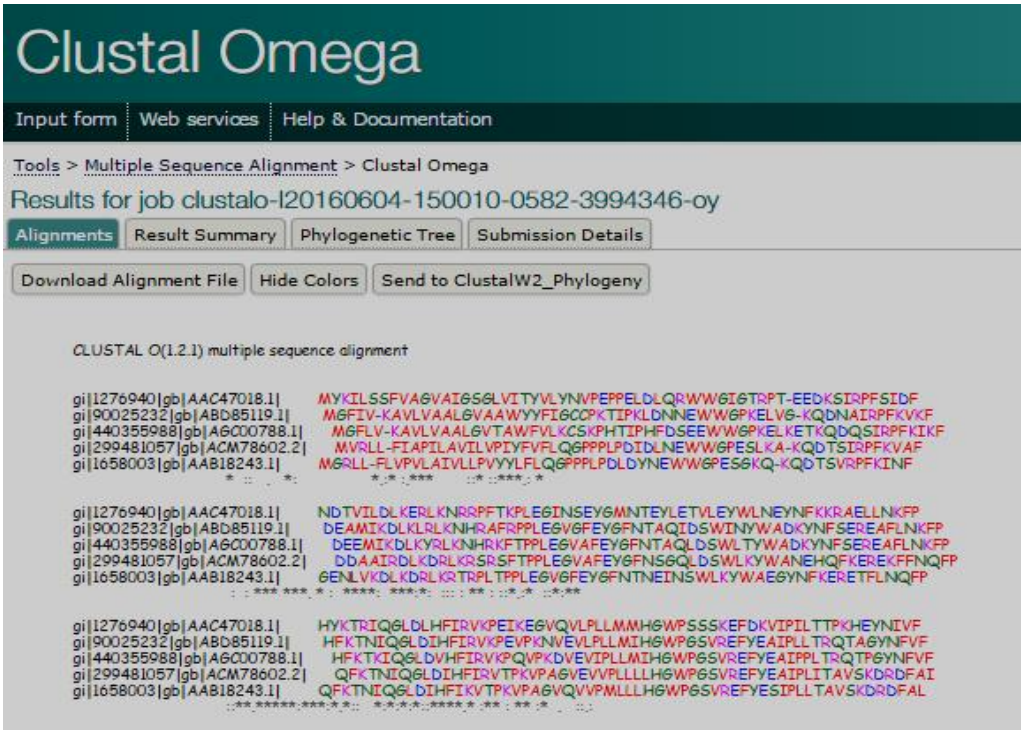


Fig: 5.4 Multiple sequence alignment

6. Inference:

Blast results were interpreted and the homologous sequences for *Helicoverpa armigera* JHEH were identified.

Multiple sequence alignment enabled us to identify the conserved regions of the JHEH sequence among other five species.

7. Reference:

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
- Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega (2011) Molecular systems biology 7 :539
- The EMBL-EBI bioinformatics web and programmatic tools framework (2015 July 01) Nucleic acids research 43 (W1) :W580-4
- Analysis Tool Web Services from the EMBL-EBI(2013 July) Nucleic acids research 41 (Web Server issue) :W597-600

VI. Restriction Digestion analysis

1. Aim: To perform restriction digestion analysis of the given query Nucleotide sequence using "Restriction enzyme digest of DNA".

2. Introduction:

Restriction Digestion is the process of cutting DNA molecules into smaller pieces with special enzymes called Restriction Endonucleases. These special enzymes recognize specific sequences in the DNA molecule wherever that sequence occurs in the DNA. The cleavage method makes use of an important class of DNA-cleaving enzymes isolated primarily from bacteria. These enzymes are called restriction endonucleases or restriction enzymes, and they are able to cleave DNA molecules at the positions at which particular short sequences of bases are present. The resulting digested DNA is very often selectively amplified using PCR, making it more suitable for analytical techniques such as agarose gel electrophoresis, and chromatography. It is used in genetic fingerprinting analysis.

There are numerous types of restriction enzymes, each of which will cut DNA differently. Most commonly-used restriction enzymes are Type II restriction endonuclease. There are some that cut a three base pair sequence while others can cut four, six, and even eight. Each enzyme has distinct properties that determine how efficiently it can cut and under what conditions. Different restriction enzymes may also have different optimal temperatures under which they function.

Restriction digest is most commonly used as part of the process of the molecular cloning of DNA fragment into a vector (such as a cloning vector or an expression vector). The vector typically contains a multiple cloning site where many restriction sites may be found, and a foreign piece of DNA may be inserted into the vector by first cutting the restriction sites in the vector as well the DNA fragment, followed by ligation of the DNA fragment into the vector.

3. Requirements:

System Configuration: Windows 7 operating system with Internet

Input requirements: Nucleotide Sequence

4. Methodology:

- ✓ Retrieve query nucleotide sequence from NCBI
- ✓ Open the home page of Restriction enzyme digestion of DNA tool from http://insilico. ehu.es /restriction/main/index2.php
- ✓ Paste the sequence in the textbox and get the list of restriction enzymes
- ✓ The obtained results are interpreted

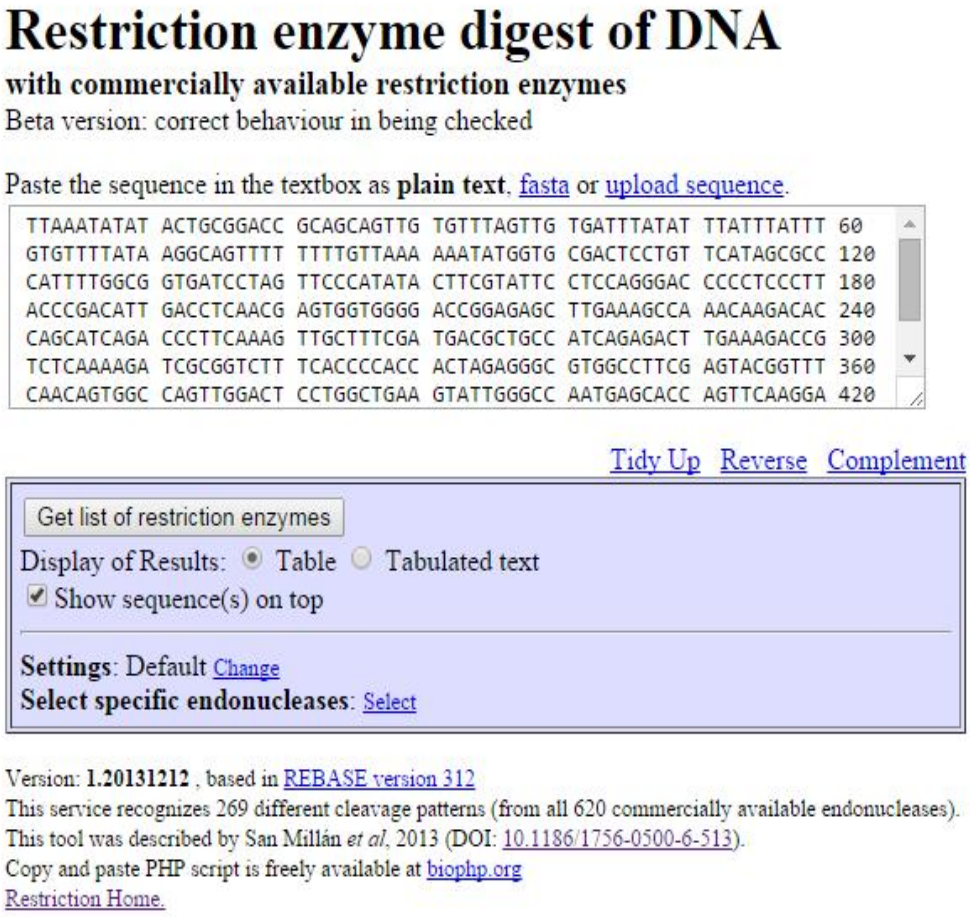


Fig: 6.1 Restriction Digestion analysis of DNA sequence

5. Results & Discussion:

Restriction enzyme digestion analysis of query JHEH nucleotide sequence of *Helicoverpa armigera* DNA sequence was interpreted and the specific cuts and positions in the sequence were identified.

```
Lengh of code: 490
G+C=46%

TTAAATATAT ACTGCGGACC GCAGCAGTTG TGTTTAGTTG TGATTTATAT TTATTTATTT GTGTTTTATA AGGCAGTTTT TTTTGTTAAA AAATATGGTG  100
CGACTCCTGT TCATAGCGCC CATTTTGGCG GTGATCCTAG TTCCCATATA CTTCGTATTC CTCCAGGGAC CCCCTCCCTT ACCCGACATT GACCTCAACG  200
AGTGGTGGGG ACCGGAGAGC TTGAAAGCCA AACAAGACAC CAGCATCAGA CCCTTCAAAG TTGCTTTCGA TGACGCTGCC ATCAGAGACT TGAAAGACCG  300
TCTCAAAAGA TCGCGGTCTT TCACCCCACC ACTAGAGGGC GTGGCCTTCG AGTACGGTTT CAACAGTGGC CAGTTGGACT CCTGGCTGAA GTATTGGGCC  400
AATGAGCACC AGTTCAAGGA ACGGGAGAAG TTCTTCAACC AGTTCCCGCA GTTCAAGACT AACATTCAAG GGCTTGATAT ACATTTCATT
```

AccII,Bsh1236I,BspFNI,BstFNI,BstUI,MvnI	CG^CG	1	313	
AciI,BspACI,SsiI	C^CG_C or G^CG_G	5	14,19,128,313,446	
AcoI,CfrI,EaeI	Y^GGCC_R	1	367	
AfaI,RsaI	GT^AC	1	353	
AfiI,Bsc4I,BseLI,BsiYI,BslI	CCNN_NNN^NNGG	2	335,415	
AjnI,EcoRII,Psp6I,PspGI	^CCWGG_	2	162,380	
AluI	AG^CT	1	219	
Alw21I,BsiHKAI,Bbv12I	G_WGCW^C	1	408	
ApeKI,TseI	G^CWG_C	2	21,275	
Asp700I,MroXI,PdmI,XmnI	GAANN^NNTTC	1	431	
AspLEI,BstHHI,CfoI,HhaI	G_CG^C	1	118	
AspS9I,BmgT120I,Cfr13I,Sau96I	G^GNC_C	4	16,167,209,396	
AvaII,Bme18I,Eco47I,SinI,VpaK11BI	G^GWC_C	3	16,167,209	
BalI,MlsI,MluNI,MscI,Msp20I	TGG^CCA	1	369	
BfaI,FspBI,MaeI,XspI	C^TA_G	2	137,332	
BfuCI,Bsp143I,BssMI,BstMBI,DpnII,Kzo9I,MboI,NdeII,Sau3AI	^GATC_	2	132,308	
BisI,BlsI,Fnu4HI,Fsp4HI,GluI,ItaI,SatI	GC^N_GC	2	22,276	
Bme1390I,MspR9I,ScrFI	CC^N_GG	2	164,382	
BmiI,BspLI,NlaIV,PspN4I	GGN^NCC	3	168,169,210	
BmrFI,BssKI,BstSCI,StyD4I	^CCNGG_	2	162,380	
BsaJI,BseDI,BssECI	C^CNNG_G	1	163	
BsaWI	W^CCGG_W	1	211	
BseBI,BstNI,BstOI,Bst2UI,MvaI	CC^W_GG	2	164,382	
BshFI,BsnI,BspANI,BsuRI,HaeIII,PhoI	GG^CC	3	344,369,398	
BsiSI,HapII,HpaII,MspI	C^CG_G	1	212	
Bsp1286I,MhlI,SduI	G_DGCH^C	1	408	
Bst4CI,HpyCH4III,TaaI	AC_N^GT	3	299,356,365	
BstH2I,HaeII	R_GCGC^Y	1	119	
BstKTI	G_AT^C	2	135,311	
CpoI,CspI,RsrII,Rsr2I	CG^GWC_CG	1	16	

Fig: 6.2 Restriction digestion of DNA Output

6.Inference:

In molecular biology and genetic Engineering experiments, restriction digestion analysis is the most important task to be performed before designing of primers, cloning vectors. Restriction digests are also necessary for performing the following analytical techniques like RFLP - Restriction Fragment Length Polymorphism, AFLP - Amplified Fragment Length Polymorphism and STRP - Short Tandem Repeat Polymorphism.

7.Reference:

- Richard J. Roberts* and Dana Macelis 1999 REBASE—restriction enzymes and methylases Nucleic Acids Research, Vol. 27, No. 1 312–313.
- https://en.wikipedia.org/wiki/Restriction_digest

VII. Restriction Mapping

1. Aim: To perform Restriction Mapping of the given query Nucleotide sequence using "Webcutter 2.0".

2. Introduction:

Restriction mapping is a process of obtaining structural information on a piece of DNA by the use of restriction enzymes.

Restriction Mapping:

Restriction mapping involves digesting DNA with a series of restriction enzymes and then separating the resultant DNA fragments by agarose gel electrophoresis. The distance between restriction enzyme sites can be determined by the patterns of fragments that are produced by the restriction enzyme digestion. In this way, information about the structure of an unknown piece of DNA can be obtained.

Uses of Restriction Mapping:

Restriction map information is important for many techniques used to manipulate DNA. One application is to cut a large piece of DNA into smaller fragments to allow it to be sequenced. Genes and cDNAs can be thousands of kilobases long (megabases - Mb); however, they can only be sequenced 400 bases at a time. DNA must be chopped up into smaller pieces and subcloned to perform the sequencing. Also, restriction mapping is an easy way to compare DNA fragments without having any information of their nucleotide sequence.

3. Requirements:

System Configuration: Windows 7 operating system with Internet

Input requirements: Nucleotide Sequence

4. Methodology:

- ✓ Retrieve query nucleotide sequence from NCBI
- ✓ Open the home page of Webcutter tool from http://rna.lundberg.gu.se/cutter2/
- ✓ Paste the sequence in the textbox and click on "Analyze sequence" by keeping the parameters default
- ✓ The obtained results are interpreted

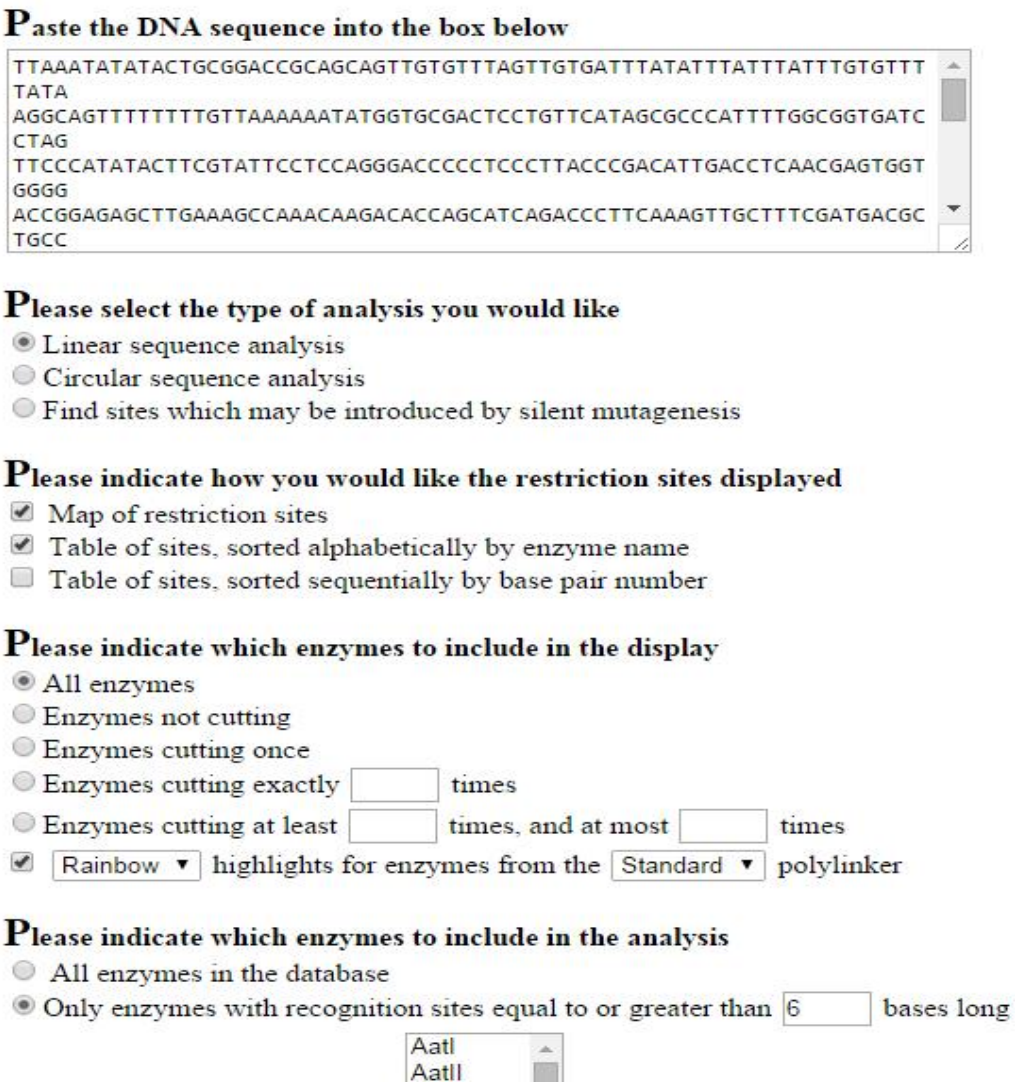


Fig: 7.1 Input of DNA sequence for Restriction mapping

5. Results and Discussion:

Restriction mapping of JHEH DNA sequence by using webcutter results were interpreted and the different types of restriction enzymes and the position of their cuts in JHEH nucleotide sequence of *Helicoverpa armigera* were identified.

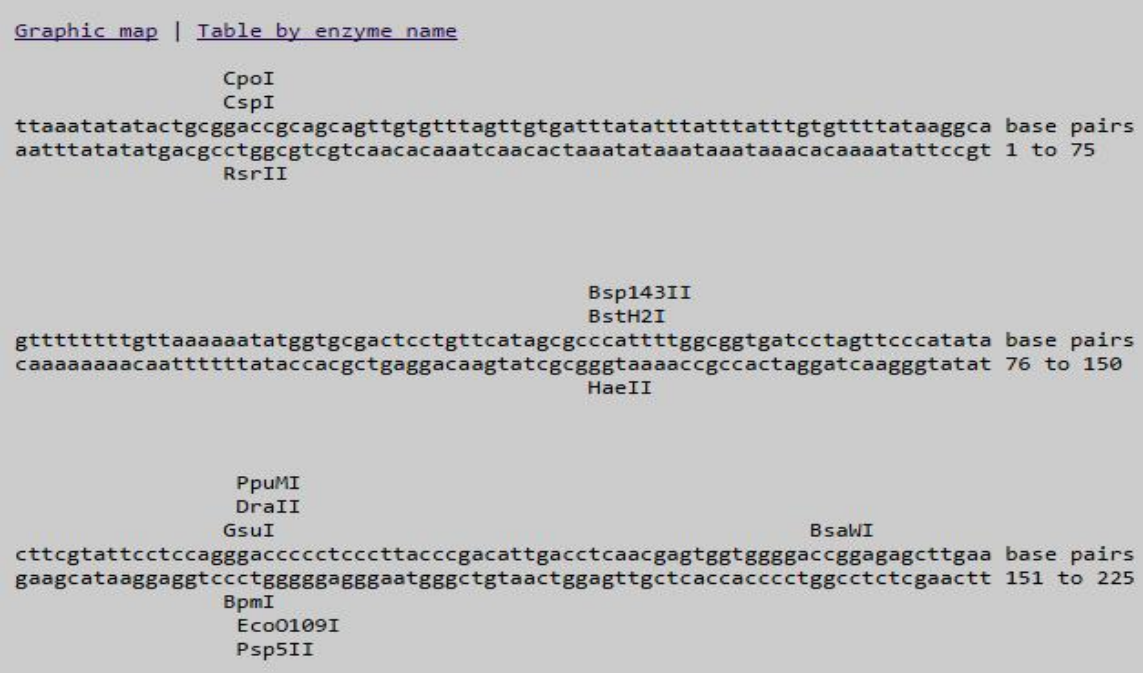


Fig: 7.2 Restriction mapping by using webcutter tool

6. Inference:

Restriction mapping is digesting the DNA with a series of restriction enzymes and separating the resultant DNA fragments by agarose gel electrophoresis. The distance between restriction enzyme sites can be determined by the patterns of fragments that are produced by the restriction enzyme digestion. In this way, information about the structure of an unknown piece of DNA can be obtained.

7.References:

1. Webcutter 2.0, copyright 1997 Max Heiman.
2. http://rna.lundberg.gu.se/cutter2/

VIII. Primer Designing

1. Aim: To design primers for cloning of a query Nucleotide sequence using "Primer 3" tool.

2. Introduction:

Primer optimization has two goals: efficiency and selectivity. Efficiency involves taking into account such factors as GC-content, efficiency of binding, complementarity, secondary structure, annealing and melting point (Tm). Primer selectivity requires that the primer pairs not fortuitously bind to random sites other than the target of interest, nor should the primer pairs bind to conserved regions of a gene family. If the selectivity is poor, a set of primers will amplify multiple products besides the target of interest.

The design of appropriate short or long primer pairs is only one goal of PCR product prediction. The important design considerations described below are a key to specific amplification with high yield. The preferred values indicated are built into all our products by default.

Primer Length: It is generally accepted that the optimal length of PCR primers is 18-22 bp. This length is long enough for adequate specificity and short enough for primers to bind easily to the template at the annealing temperature.

Primer Melting Temperature: Primer Melting Temperature (T_m) by definition is the temperature at which one half of the DNA duplex will dissociate to become single stranded and indicates the duplex stability. Primers with melting temperatures in the range of 52-58 ^oC generally produce the best results. Primers with melting temperatures above 65^oC have a tendency for secondary annealing. The GC content of the sequence gives a fair indication of the primer T_m.

Primer Annealing Temperature: The primer melting temperature is the estimate of the DNA-DNA hybrid stability and critical in determining the annealing temperature. Too high T_a will produce insufficient primer-template hybridization resulting in low PCR product yield. Too low T_a may possibly lead to non-specific products caused by a high number of base pair mismatches. Mismatch tolerance is found to have the strongest influence on PCR specificity.

GC Content: The GC content (the number of G's and C's in the primer as a percentage of the total bases) of primer should be 40-60%.

GC Clamp: The presence of G or C bases within the last five bases from the 3' end of primers (GC clamp) helps promote specific binding at the 3' end due to the stronger bonding of G and C bases. More than 3 G's or C's should be avoided in the last 5 bases at the 3' end of the primer.

Primer Secondary Structures: Presence of the primer secondary structures produced by intermolecular or intramolecular interactions can lead to poor or no yield of the product. They adversely affect primer template annealing and thus the amplification. They greatly reduce the availability of primers to the reaction.

i) Hairpins: It is formed by intramolecular interaction within the primer and should be avoided.

ii) Self Dimer: A primer self-dimer is formed by intermolecular interactions between the two (same sense) primers, where the primer is homologous to itself. Generally a large amount of primers are used in PCR compared to the amount of target gene. When primers form intermolecular dimers much more readily than hybridizing to target DNA, they reduce the product yield.

iii) Cross Dimer: Primer cross dimers are formed by intermolecular interaction between sense and antisense primers, where they are homologous.

Repeats: A repeat is a di-nucleotide occurring many times consecutively and should be avoided because they can misprime. For example: ATATATAT. A maximum number of di-nucleotide repeats acceptable in an oligo is 4 di-nucleotides.

Runs: Primers with long runs of a single base should generally be avoided as they can misprime.

3' End Stability: It is the maximum ΔG value of the five bases from the 3' end. An unstable 3' end will result in less false priming.

Avoid Template Secondary Structure: A single stranded Nucleic acid sequences is highly unstable and fold into conformations (secondary structures). The stability of these template secondary structures depends largely on their free energy and melting temperatures(T_m). Consideration of template secondary structures is important in designing primers, especially in qPCR. If primers are designed on secondary structures

which are stable even above the annealing temperatures, the primers are unable to bind to the template and the yield of PCR product is significantly affected. Hence, it is important to design primers in the regions of the templates that do not form stable secondary structures during the PCR reaction.

Avoid Cross Homology: To improve specificity of the primers it is necessary to avoid regions of homology. Primers designed for a sequence must not amplify other genes in the mixture. Commonly, primers are designed and then BLASTed to test the specificity.

3. Requirements:

System Configuration: Windows 7 operating system with Internet

Input requirements: Nucleotide Sequence

4. Methodology:

- ✓ Retrieve query nucleotide sequence from NCBI
- ✓ Open the home page of primer3 tool.
- ✓ Paste the sequence and run the program with default parameters.

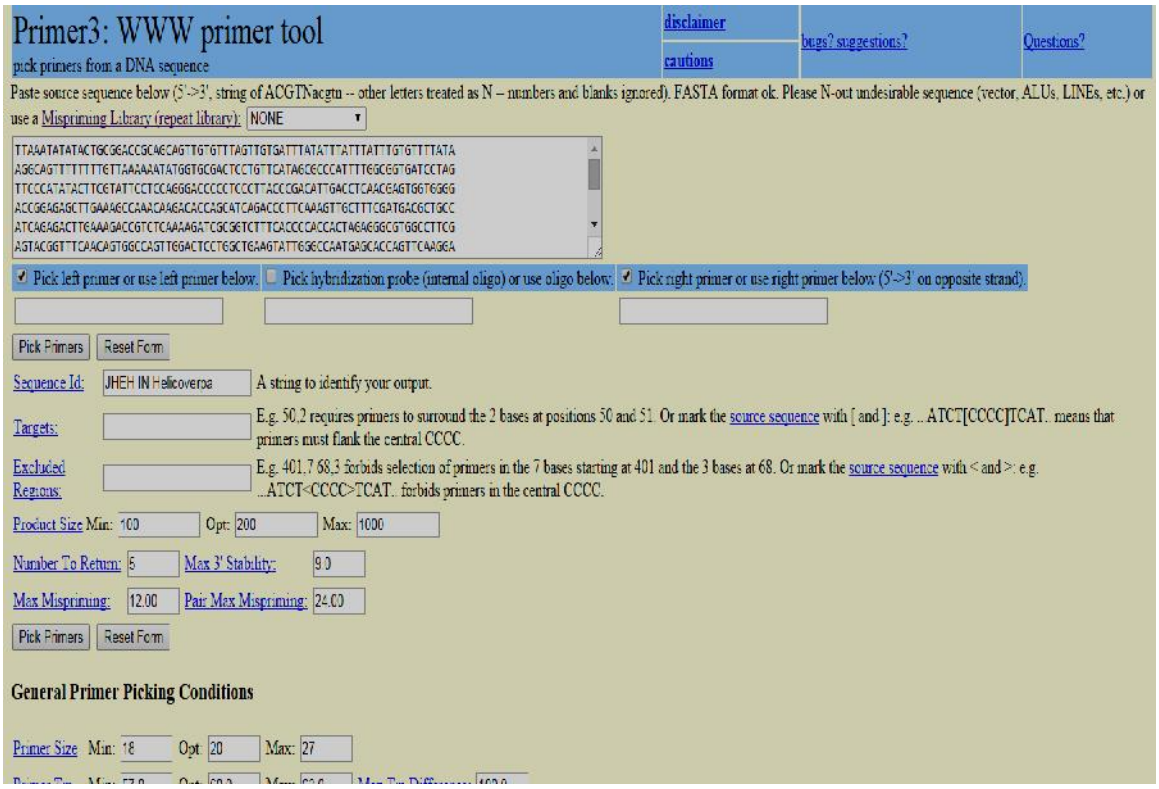
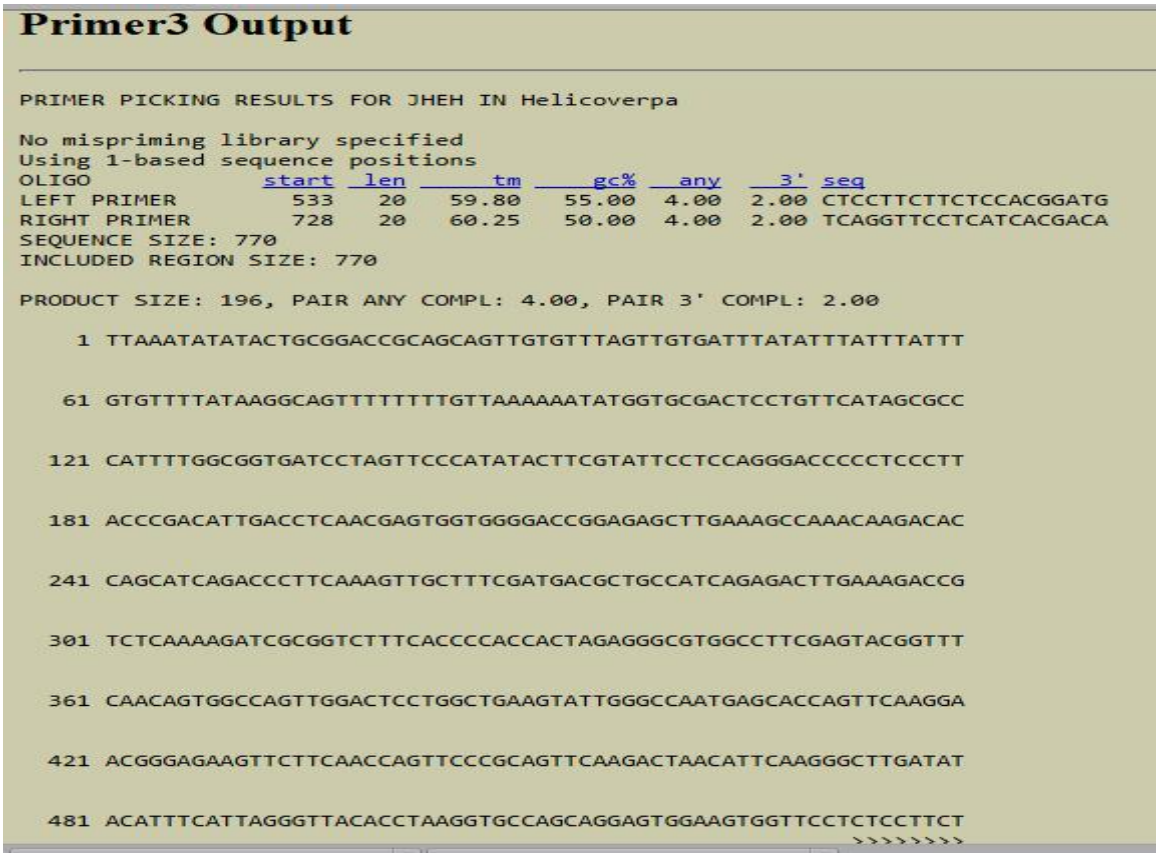


Fig: 8.1 Primer Designing tool Primer3 home page

5. Results and Discussion:

Many primer sequences have been displayed in the output and the best possible forward and reverse primers have to be chosen keeping, 3' and 5' complementarities, hairpin loops in consideration.



Primer3 Output

```
PRIMER PICKING RESULTS FOR JHEH IN Helicoverpa

No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len      tm      gc%   any   3'   seq
LEFT PRIMER        533   20   59.80    55.00  4.00  2.00 CTCCTTCTTCTCCACGGATG
RIGHT PRIMER       728   20   60.25    50.00  4.00  2.00 TCAGGTTCCTCATCACGACA
SEQUENCE SIZE: 770
INCLUDED REGION SIZE: 770

PRODUCT SIZE: 196, PAIR ANY COMPL: 4.00, PAIR 3' COMPL: 2.00

    1 TTAAATATATACTGCGGACCGCAGCAGTTGTGTTTAGTTGTGATTTATATTTATTTATTT

   61 GTGTTTTATAAGGCAGTTTTTTTTGTTAAAAAATATGGTGCGACTCCTGTTCATAGCGCC

  121 CATTTTGGCGGTGATCCTAGTTCCCATATACTTCGTATTCCTCCAGGGACCCCCTCCCTT

  181 ACCCGACATTGACCTCAACGAGTGGTGGGGACCGGAGAGCTTGAAAGCCAAACAAGACAC

  241 CAGCATCAGACCCTTCAAAGTTGCTTTCGATGACGCTGCCATCAGAGACTTGAAAGACCG

  301 TCTCAAAAGATCGCGGTCTTTCACCCCACCACTAGAGGGCGTGGCCTTCGAGTACGGTTT

  361 CAACAGTGGCCAGTTGGACTCCTGGCTGAAGTATTGGGCCAATGAGCACCAGTTCAAGGA

  421 ACGGGAGAAGTTCTTCAACCAGTTCCCGCAGTTCAAGACTAACATTCAAGGGCTTGATAT

  481 ACATTTCATTAGGGTTACACCTAAGGTGCCAGCAGGAGTGGAAGTGGTTCCTCTCCTTCT
                                                    >>>>>>>>
```

Fig: 8.2 Primer3 output

6. Inference:

Primer Designing of *Helicoverpa armigera* JHEH sequence was performed, the results were interpreted and the finest forward and reverse primers in the given DNA sequence are recognized.

7. Reference:

Andreas Untergasser, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C. Faircloth, Maido Remm, and Steven G. Rozen. Primer3—new capabilities and interfaces, Nucleic Acids Res. 2012 Aug; 40(15): e115.

IX. Phylogenetic Analysis

1. Aim: To perform phylogenetic analysis and evolutionary studies on set of Nucleotide sequences using "Clustal W2 phylogeny" tool.

2. Introduction:

A phylogenetic tree or evolutionary tree is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities—their phylogeny—based upon similarities and differences in their physical or genetic characteristics. The taxa joined together in the tree are implied to have descended from a common ancestor. In a *rooted* phylogenetic tree, each node with descendants represents the inferred most recent common ancestor of the descendants and the edge lengths in some trees may be interpreted as time estimates. Each node is called a taxonomic unit. Internal nodes are generally called hypothetical taxonomic units, as they cannot be directly observed. Trees are useful in fields of biology such as bioinformatics, systematics, and comparative phylogenetics.

Types:

Rooted Tree: A rooted phylogenetic tree is a directed tree with a unique node corresponding to the (usually imputed) most recent common ancestor of all the entities at the leaves of the tree. The most common method for rooting trees is the use of uncontroversial outgroup close enough to allow inference from sequence or trait data, but far enough to be a clear outgroup.

Unrooted Tree: Unrooted trees illustrate the relatedness of the leaf nodes without making assumptions about ancestry. They do not require the ancestral root to be known or inferred. Unrooted trees can always be generated from rooted ones by simply omitting the root. By contrast, inferring the root of an unrooted tree requires some means of identifying ancestry. This is normally done by including an outgroup in the input data so that the root is necessarily between the outgroup and the rest of the taxa in the tree, or by introducing additional assumptions about the relative rates of evolution on each branch, such as an application of the molecular clock hypothesis.

Bifurcating tree: Both rooted and unrooted phylogenetic trees can be either bifurcating or multifurcating, and either labeled or unlabeled. A rooted bifurcating

tree has exactly two descendants arising from each interior node , and an unrooted bifurcating tree takes the form of an unrooted binary tree, a free tree with exactly three neighbors at each internal node.

3. Requirements:

System Configuration: Windows 7 operating system with Internet

Input requirements: Nucleotide/Protein Sequence

4. Methodology:

- ✓ Retrieve five query nucleotide sequences from NCBI and save them in a single notepad.
- ✓ Open the Clustalw2 home page and select the Phylogenetic tree.
- ✓ Paste the notepad file of the query sequences and
- ✓ Run the program with default parameters.

Fig: 9.1 Clustal W2_phylogeny home page

5. Results and Discussion:

The results were interpreted and the rooted Phylogenetic tree was obtained. The homolog sequences for *Helicoverpa armigera* JHEH sequence were identified.

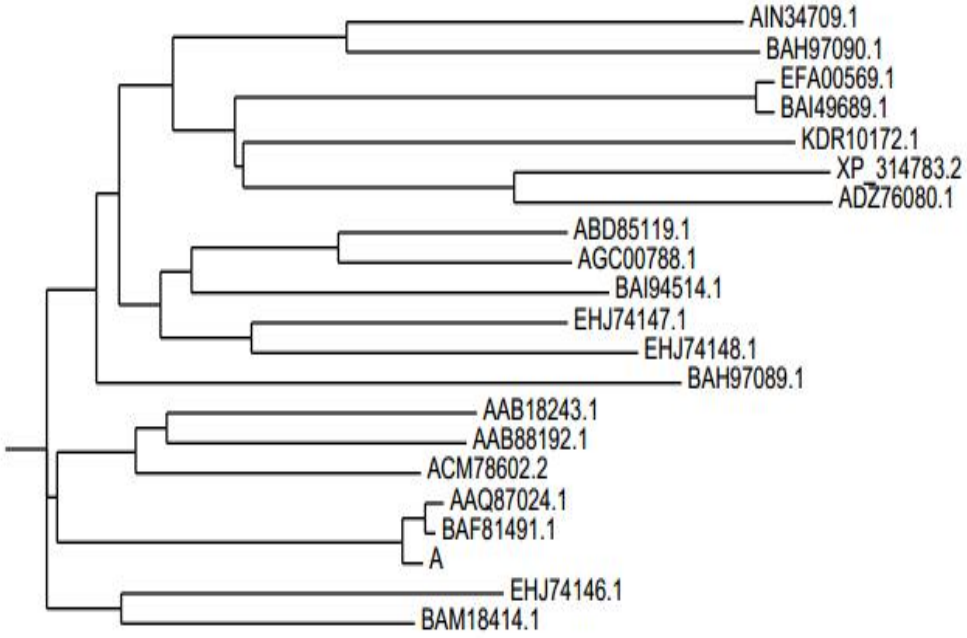


Fig: 9.2 Phylogenetic Tree

6. Inference:

By phylogenetic tree, the distant and close species can be identified. Based on the distances in the tree, the homology has been calculated. Evolutionary relationships can be predicted with these phylogenetic studies.

7. Reference:

- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al., DG ClustalW and ClustalX version 2 (2007), Bioinformatics 2007 23(21): 2947-2948.
- Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R A new bioinformatics analysis tools framework at EMBL-EMBL-EBI (2010), Nucleic acids research 2010 Jul, 38 Suppl: W695-9

X. Primary Structure prediction of proteins

1. Aim: To predict primary structural features of a given protein sequence using PROTPARAM tool.

2. Introduction:

There are 20 different standard L-α-amino acids used by cells for protein construction. Amino acids, as their name indicates, contain both a basic amino group and an acidic carboxyl group. This difunctionality allows the individual amino acids to join together in long chains by forming *peptide bonds:* amide bonds between the -NH2 of one amino acid and the -COOH of another. Sequences with fewer than 50 amino acids are generally referred to as *peptides*, while the terms *protein* or *polypeptide* are used for longer sequences. A protein can be made up of one or more polypeptide molecules. The end of the peptide or protein sequence with a free carboxyl group is called the *carboxy-terminus* or *C-terminus*. The terms *amino-terminus* or *N-terminus* describe the end of the sequence with a free α-amino group.

The amino acids differ in structure by the substituent on their side chains. These side chains confer different chemical, physical and structural properties to the final peptide or protein. The structures of the 20 amino acids commonly found in proteins are shown in Figure 1. Each amino acid has both a one-letter and three-letter abbreviation. These abbreviations are commonly used to simplify the written sequence of a peptide or protein.

Depending on the side-chain substituent, an amino acid can be classified as being acidic, basic or neutral. Although 20 amino acids are required for synthesis of various proteins found in humans, we can synthesize only 10. The remaining 10 are called essential amino acids and must be obtained in the diet.

The amino acid sequence of a protein is encoded in DNA. Proteins are synthesized by a series of steps called transcription (the use of a DNA strand to make a complimentary messenger RNA strand - mRNA) and translation (the mRNA sequence is used as a template to guide the synthesis of the chain of amino acids which make up the protein). Often, post-translational modifications, such as glycosylation or phosphorylation, occur which are necessary for the biological function of the protein. While the amino acid sequence makes up the *primary structure* of the protein, the

chemical/biological properties of the protein are very much dependent on the three-dimensional or tertiary structure.

ProtParam is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered protein sequence.

3. Requirements:

System Configuration: Windows 7 operating system with Internet

Input requirements: Protein Sequence

4. Methodology:

- ✓ Retrieve query JHEH protein sequence from NCBI.
- ✓ Open the home page of Protparam from the URL http://web.expasy.org/protparam/
- ✓ Paste the query protein sequence and click on 'Compute'
- ✓ The results are interpreted

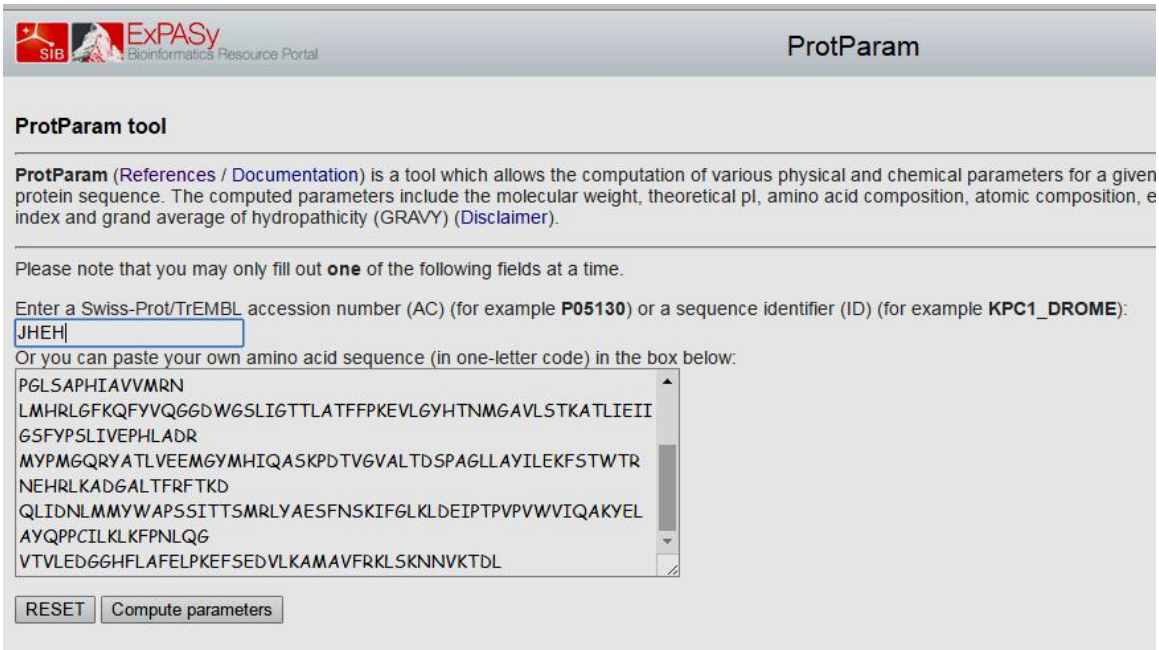


Fig: 10.1 Protparam home page

5. Results and Discussion:

The molecular weight, amino acids ratio, hydropathicity of the given query protein sequence were calculated using Protparam tool (Fig. 10.2)

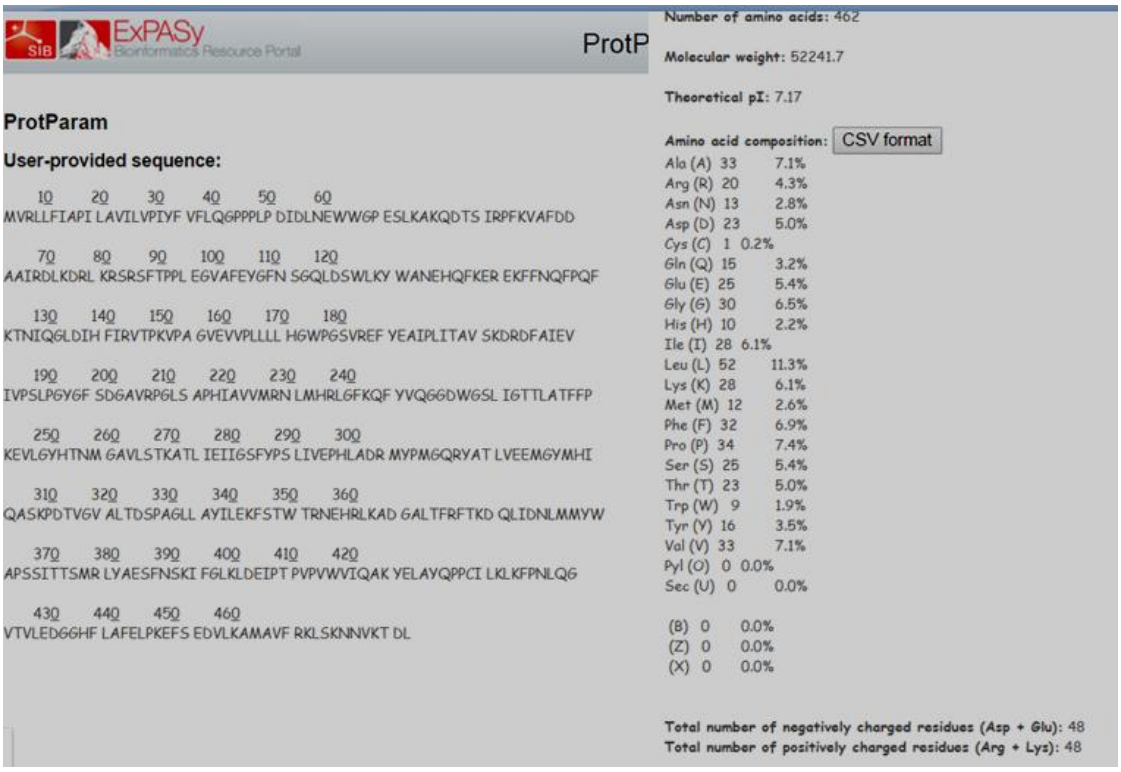


Fig: 10.2 Primary structural details of query sequence

6. Inference:

Protein identification and analysis software performs a central role in the investigation of proteins from two-dimensional (2-D) gels and mass spectrometry. The parameters like the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) can be computed.

7. Reference:

- Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.; *Protein Identification and Analysis Tools on the ExPASy Server;* (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005). pp. 571-607
- http://www.particlesciences.com/news/technical-briefs/2009/protein-structure.html

XI. Secondary Structure Analysis

1. Aim: To predict secondary structural features of a given protein sequence using Chou-Fasman Secondary structure prediction tool.

2. Introduction:

Stretches or strands of proteins or peptides have distinct characteristic local structural conformations or *secondary structure*, dependent on hydrogen bonding. The two main types of secondary structure are the α-helix and the ß-sheet.

The *α-helix* is a right-handed coiled strand. The side-chain substituents of the amino acid groups in an α-helix extend to the outside. Hydrogen bonds form between the oxygen of the C=O of each peptide bond in the strand and the hydrogen of the N-H group of the peptide bond four amino acids below it in the helix. The hydrogen bonds make this structure especially stable. The side-chain substituents of the amino acids fit in beside the N-H groups.

The hydrogen bonding in a *ß-sheet* is between strands (inter-strand) rather than within strands (intra-strand). The sheet conformation consists of pairs of strands lying side-by-side. The carbonyl oxygens in one strand hydrogen bond with the amino hydrogens of the adjacent strand. The two strands can be either parallel or anti-parallel depending on whether the strand directions (N-terminus to C-terminus) are the same or opposite. The anti-parallel ß-sheet is more stable due to the more well-aligned hydrogen bonds.

3. Requirements:

System Configuration: Windows 7 operating system with Internet

Input requirements: Protein Sequence

4. Methodology:

- ✓ Retrieve query JHEH protein sequence from NCBI.
- ✓ Open the home page of CFSSP from the URL http://cho-fas.sourceforge.net/
- ✓ Paste the query protein sequence and click on ‘Predict’
- ✓ The results are interpreted

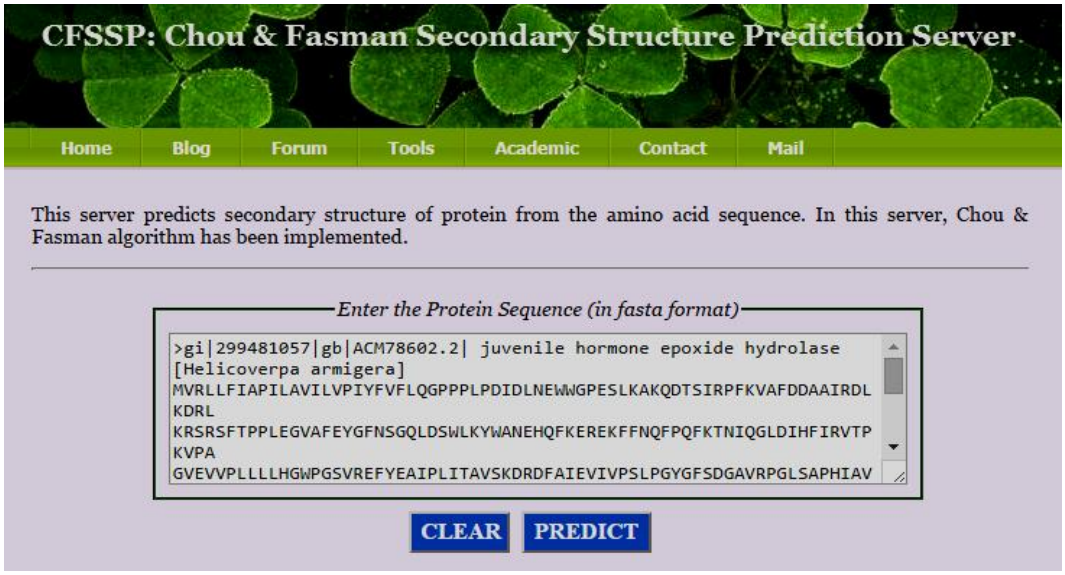


Fig: 11.1 CFSSP prediction page

5. Results and Discussion:

Fig: 11.2 Secondary structure prediction of JHEH protein using CFSSP

6. Inference:

The secondary structure of protein is predicted by CFSSP. Secondary structure features like Helix, sheets, coils and turns are identified in the given query protein sequence.

7. Reference:

- Peter Y. Chou, and Gerald D. Fasman. Prediction of protein conformation. *Biochemistry*. 1974 Jan; 13(2), pp 222-245.
- Peter Y. Chou, and Gerald D. Fasman. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry*. 1974 Jan; 13(2): pp 211-222.
- http://www.particlesciences.com/news/technical-briefs/2009/protein-structure.html

XII. Protein Tertiary structure Visualization

1. **Aim:** To Visualize the protein three dimensional structure of a protein using Rasmol.
2. **Introduction:**

The overall three-dimensional shape of an entire protein molecule is the *tertiary structure*. The protein molecule will bend and twist in such a way as to achieve maximum stability or lowest energy state. Although the three-dimensional shape of a protein may seem irregular and random, it is fashioned by many stabilizing forces due to bonding interactions between the side-chain groups of the amino acids. Under physiologic conditions, the hydrophobic side-chains of neutral, non-polar amino acids such as phenylalanine or isoleucine tend to be buried on the interior of the protein molecule thereby shielding them from the aqueous medium. The alkyl groups of alanine, valine, leucine and isoleucine often form hydrophobic interactions between one-another, while aromatic groups such as those of phenylalanine and tyrosine often stack together. Acidic or basic amino acid side-chains will generally be exposed on the surface of the protein as they are hydrophilic.

The formation of disulfide bridges by oxidation of the sulfhydryl groups on cysteine is an important aspect of the stabilization of protein tertiary structure, allowing different parts of the protein chain to be held together covalently. Additionally, hydrogen bonds may form between different side-chain groups. As with *disulfide bridges*, these hydrogen bonds can bring together two parts of a chain that are some distance away in terms of sequence. *Salt bridges*, ionic interactions between positively and negatively charged sites on amino acid side chains, also help to stabilize the tertiary structure of a protein.

RasMol is a free, interactive molecular-graphics viewer. The program reads in the 3-D coordinates for a molecule using the pdb file format. It displays the molecule in various representations and allows one to rotate the molecule interactively. RasMol-ucb allows simultaneous viewing of multiple molecules. RasMol is a computer program written for molecular graphics visualization intended and used primarily for the depiction and exploration of biological macromolecule structures, such as those found in the Protein Data Bank. It was originally developed by Roger Sayle in the early 90s.

RasMol includes a language (for selecting certain protein chains, or changing colors etc.). Jmol and Sirius have incorporated the RasMol scripting language into its commands. The tertiary structures of proteins can be retrieved from a protein database called "Protein Databank (PDB)" for visualization.

3. Requirements:

System Configuration: Windows 7 operating system with Internet

Input requirements: Protein Structure in PDB format

4. Methodology:

✓ Retrieve query JHEH protein sequence from NCBI.

✓ To open a file: File/Open from within RasMol.

✓ To select a molecule: Click on molecule name in the Molecules window.

✓ Moving the molecule(s):

Action	**PC**	**MAC**
Rotation	Left-mouse button (Click & Hold)	Mouse button (Click & Hold)
Translation	Right-mouse button (Click & Hold)	<OpenApple> and mouse button
Zoom	<alt><SHIFT> and left-mouse button	<SHIFT> and mouse button
Z-Rotation	<alt><SHIFT>and right-mouse button	<option><SHIFT> and mouse button

✓ To change to different representations (i.e., CPK, stick, ribbon, etc.): Display/Stick

✓ To determine distances, angles, dihedral angles:

✓ Click on appropriate icon in the Molecules window (the secondary window)

✓ Click once on the appropriate number of individual atoms. (With angles, clicking on atoms must be in the appropriate order.)

✓ To rotate bond, Click on rotate angle icon in the Molecules window (the secondary window) Click on 2 atoms

✓ Click on rotate angle icon in the Molecules window (the secondary window) again Click on 3rd atom. Use mouse button to rotate

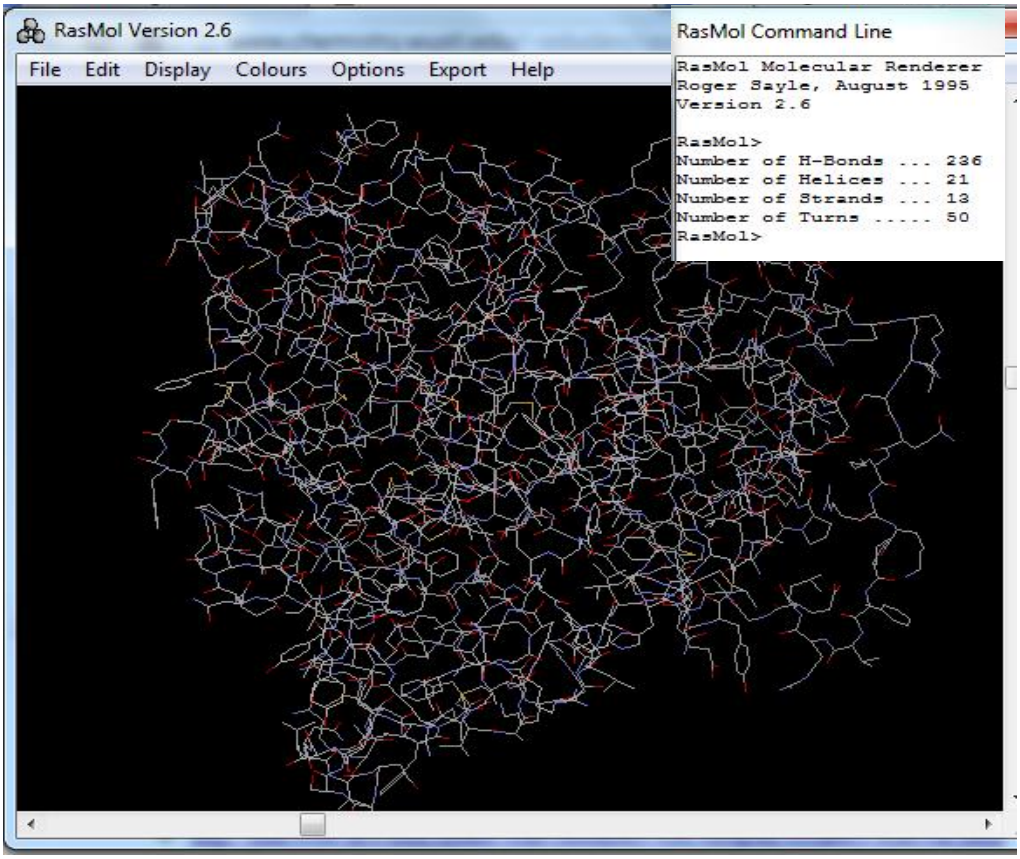


Fig: 12.1 Protein visualisation in Rasmol and Command Window

5. Results and Discussion:

The ribbon format, chemical structure, balls and sticks format of query protein structure has been shown in fig11.2. Different structural features can be visualised using Rasmol.

6. Inference:

Rasmol is a open source software to see protein structure in three dimensions and thereby helping in understanding behavior of proteins. It is easy to use, runs on many platforms, requires meager computational resources, is extremely powerful and is free.

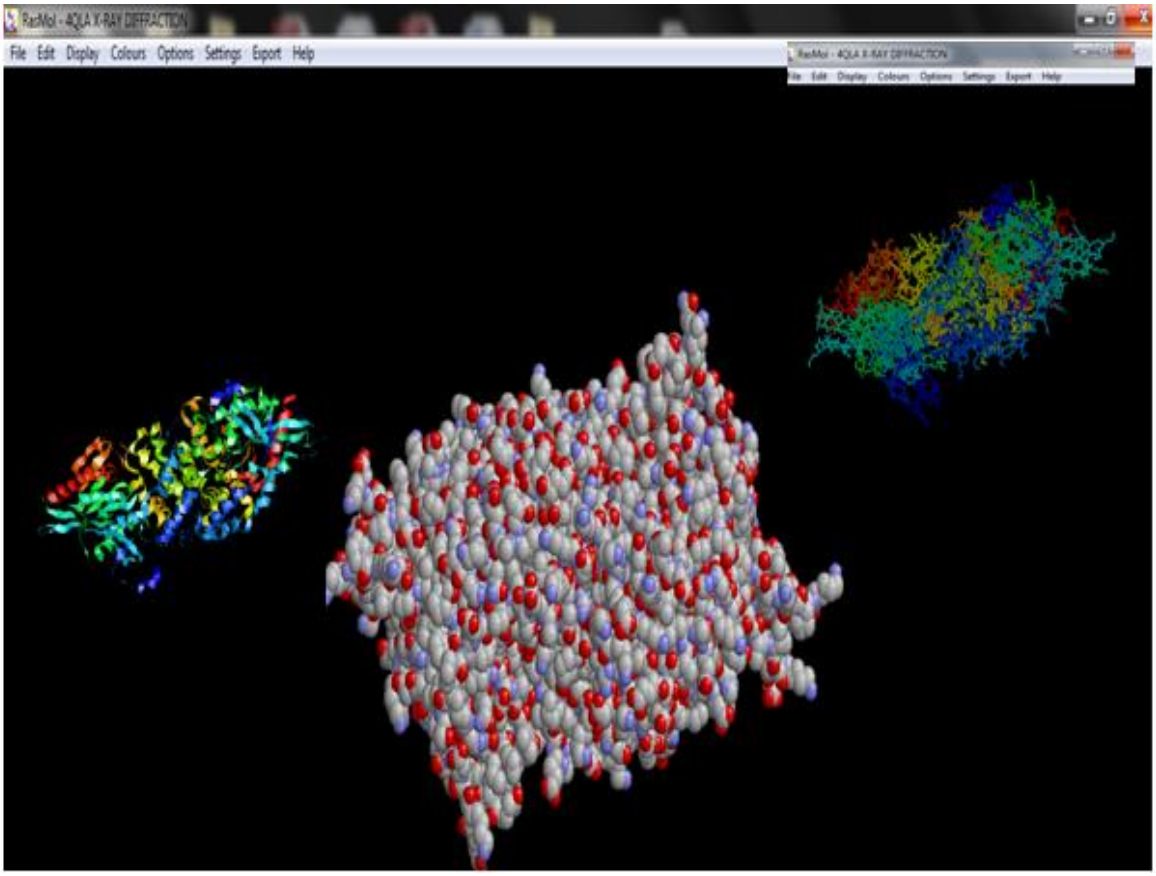


Fig: 12.2 Ribbon and Ball-stick view of a protein using Rasmol

7. Reference:

- RasMol: Biomolecular graphics for all, by Roger A. Sayle and E. James Milner-White, Trends in Biochemical Sciences 20(Sept):374-376, 1995. RasMol was first widely distributed via the Internet in June, 1993, but this is the original paper publication describing RasMol.
- http://www.particlesciences.com/news/technical-briefs/2009/protein-structure.html

About the Authors:

D.M.Mamatha, Professor & Head in the Department of Sericulure, Sri Padmavati Mahila Visva vidyalayam did her PhD in India and two PostDocs at University of California, Davis USA. Specialised in Insect Physiology, Recombinant biopesticide development, Bioinformatics and DNA Barcoding. She has received US-Fulbright award, DBT-Young Scientist & WHO - Achievement awards. She was successful in getting major research grants to the tune of Rs. 100.9 lakhs from DBT, UGC, DBT, Indo-US and Indo-Egypt funding bodies. She has published 24 + International publications in peer reviewed & high impact factored journals. She visited Malaysia, Singapore, USA, Canada,Thailand, Nepal, and Egypt on various academic & research assignments.

Dr. K. Swetha kumari did her PhD in the Department of Sericulture, Sri Padmavati Women's University. She qualified APSET. She did her Masters in Bioinformatics in S.V. University. She underwent Hands on training programs on DNA Barcoding, Baculovirus Expression vector system, Genomics, Transcriptomics & Proteomics in India's top premier labs viz., in PHCDBS-Aurangabad, CDFD-Hyderabad, IBAB-Bangalore respectively. She has seven International publications and has co-authored a book. Her areas of interests are Gene cloning, Molecular Biology, Recombinant protein production and characterization, Transgenics, Biomaterial Science, Homology Modeling and Docking studies, Proteomics and Genomics, DNA Barcoding. Dr. Swetha shows unending quench in updating her research skills and knowledge, which is critical in this rapidly changing scientific world.

Ms. S. Kalpana presently doing Ph.D., in the topic related to Molecular Biology and Bioinformatics. She did Masters in Bioinformatics from S.V.University, Tirupati. Her expertise is on Basic Bioinformatics theory as well as practical. She qualified GATE. She has International publications in *Insilico* studies and Computational Biology. She underwent one month workshop on "Basic Cloning techniques & Gene expression studies" in UGC Networking Resource Centre in the Division of Biological Sciences at IISc, Bangalore. She has completed 15 days Hands on training in Genomics Transcriptomics and Proteomics from Shodaka Life Sciences, IBAB Bangalore. Her areas of research are Gene cloning, Molecular Biology, Recombinant protein production and characterization, Modeling and Docking studies, Proteomics and Genomics, DNA Barcoding.