Protein prediction methods – steps of analysis
(version 1.0)

Kornelia Polok,
Department of Genetics, University of Warmia and Mazury
Plac Lodzki 3. 10-967 Olsztyn, kpolok@moskit.uwm.edu.pl
Within EU Marie Curie Transfer of Knowledge Project, GenCrop, Greece, University of Crete

❶BASIC INFORMATION

The three dimensional structure of a protein determines its function. Genomes sequencing projects generate predictions of protein-coding sequences. One fundamental aspect of each predicted protein is its structure and function. Function is often assigned based upon homology to another protein whose function is perhaps already known or inferred. Various types of BLAST are employed to identify such relationships of homology. However, for many proteins sequence identity is extremely limited. On the other hand structure and function are preserved over evolutionary time more than is sequence identity. When a genome is sequenced and a new protein is predicted, we can use the information about other proteins of known structure to:

- predict structure,
- predict structural and functional consequences of mutations,
- predict ligand.

❷PROTEIN STRUCTURE

Primary structure

Linear sequence of amino acids residues in a polypeptide chain. The amino acids are joined by peptide bonds on each side of the Cα carbon atom.

Secondary structure

Secondary structure refers to the arrangements of the primary amino acid sequence into motifs such as α helices, β-sheets and coils. It is determined by the amino acid side chains. The α helices typically are formed from contiguous stretches of 4-10 amino acids residues in length. The β sheets are formed from adjacent β strands composed of 5-10 residues. They are arranged in parallel or antiparallel orientation. In general, proteins tend to be arranged with hydrophobic amino acids in the interior and hydrophilic residues exposed to the surface. This hydrophobic core is produced in spite of the highly polar nature of the peptide backbone. This problem a protein solves by organizing the interior amino residues into secondary structures consisting of α helices and β sheets. Predictions are based on the frequencies of residues found in a helices, b sheets and turns. For example, a proline is extremely unlikely to occur in an a helix. Some approaches are based on individual sequences. As multiple alignments have become increasingly available, the accuracy of related secondary-structure prediction programs has increased. Typically, the most recently developed algoritms have about 70-75% accuracy.

Tertiary structure (3D)

It is a three-dimensional arrangement formed by packing secondary structure elements into globular domains. The structure may depend upon some post-translational modifications, such as the addition of sugars and disulfide bridges. In nature proteins fold spontaneously almost after protein synthesis.

Quaternary structure

It involves arrangements of tertiary structures into several polypeptide chains. Functionally important areas such as ligand-binding sites or enzymatic active sites are formed at the levels of tertiary or quaternary structures.

❸APPROACHES TO DETERMINING PROTEIN STRUCTURE

1. Experimental

   The experimentally by X-ray crystallography and nuclear magnetic resonance (NMR)

2. Comparative homology modelling by comparison to one or more homologous, known structures. It is successful when the percent of amino acid identity between the target and template is >50%.

3. Ab inito methods that use physical principles alone to predict the 3-D structure of a target.

❹PROTEIN DATA BASES

Protein Data Bank (PDB - http://www.rcsb.org/pdb)

It is one principal repository in which the structure is deposited. A broad range of primary structural data are collected, such as atomic coordinates, chemical structures of cofactors, and description of the crystal structure. PDB structures can be searched using NCBI Website through Entrez (structure data base) and BLAST with PDB. A search yields a list of proteins with four-character PDB identifiers

- Related structures - multiple protein structures can be compared simultaneously.
- Literatures
- Domains – shows conserved domains and 3D domains, clicking 3D domains shows the entries for different individual chains that are deposited in PDB for the same protein. Protein structures can be viewed by clicking “the picture”. This requires that the Cn3D software be downloaded. **The Cn3D viewer** shows the structure in seven different formats, and it can be rotated. The corresponding One-D Viewer shows amino acid sequence including helices and sheets. Highlighting any individual amino acid residue or group of residues causes the corresponding region of the protein to be highlighted. **Swiss-PDB viewer offers a large array of options including analysis of mutations**.
- Ligands – the binding pocket in each subunits deposited. A broad range of primary structural data are collected, such as atomic coordinates, chemical str

Taxonomic System for Protein Structures (SCOP, http://scop.mrc-lmb.cam.ac.uk/scop)

It provides a comprehensive description of protein structures and evolutionary relationships based upon a hierarchical classification scheme. At the level of superfamilies proteins probably do share an evolutionary relationships, even if they share relatively low amino acid sequence identity in pairwise alignments. Member of families have a clear evolutionary relationship and sequence identity is about 30%.

CATH Database (http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html)

It is a hierarchical system that describes all known protein domain structures. It clusters proteins at class (C), architecture (A), topology (T) and homologous (H) superfamilies. It provides a deep and broad set of data on the structure of individual proteins, placing them in the context of taxonomy.

- Class (C) describes three main protein folds based on secondary structure prediction.
- Architecture (A) describes the shape of the domain structure as determined by the orientation of the secondary structures.
- Topology (T) describes fold families. Proteins sharing topologies in common are not necessarily homologous.
- Homology (H) clusters proteins that are likely to share homology i.e., descent from a common ancestor.

5 COMPUTATIONAL APPROACHES TO PROTEIN STRUCTURE PREDICTIONS – COMPARATIVE HOMOLOGY MODELLING

Before you start – search the NCBI data bases or/and others to find out if you protein has already been put into the data bases and what is known about it. Make so called "global alignment" with BLAST. If your sequence has already been deposited and annotated, you have probably lost time and money isolating it. However, the most often your sequence is similar to others already deposited. This similarity can be used to predict structure and function of your protein. However, it is not unusual to find that half predicted proteins in a completed genome have no identifiable homologs. Even in that case, a protein may have features such as transmembrane domain, sites for phosphorylations etc and some predicted secondary structure. Such features may give clues to the structure and function of the protein. It is advisable to start from simple protein characterisation and then employ more advanced options. Proteins should be characterised by:

- physical properties,
- cellular localisation and protein topology,
- primary structure,
- post-translational modifications,
- domains and motifs through patterns and profile searches and similarity searches,
- function.

It is advisable to use as many prediction as possible. All programs you can find at the ExPASy serwer (http://ca.expasy.org/tools/).

5.1. THE 1ST STEP – TRANSLATION

The first thing to do with the nucleotide sequence is to translate it into a protein. When the sequence is derived from any cDNA that indeed represents a mRNA, the translation can be made directly from a sequence using any of listed programs. If a sequence has been derived from genomic DNA, it is important to determine which nucleotide to start translation, and then to stop, i.e., to determine **Open Reading Frame (ORF)**. Every region of DNA has six possible reading frames, three in each direction. Typically only one reading frame is used in translating a gene, and **this is often the longest ORF**. In most species an open reading frame starts with **ATG** (methionine) and ends with a stop codon TAA, TAG, TGA (UAA, UAG, UGA in mRNA).

- **TRANSLATE** – a tool, which allows the translation of a nucleotide sequence to a protein sequence. All six frames are translated. After choosing a frame one of methionine entries should be selected. This will create a virtual Swiss-Prot entry.
- **TRANSEQUE**– nucleotide to protein translation from EMBOSS.
- **ORF Finder (NCBI)** – the potentiall ORFs are presented as coloured bars with the coordinates. By clicking on a bar, the alignment of amino acid and DNA sequence is available. The program offers automatic link to BLAST Searches.
- **Codon Usage Database** – http://www.justbio.com/translator/index.php
- **JCat** – Codon Adapter Tool – enables to express a gene in an organism having different codon usage.

In a case of DNA sequence it is good to try first to use introns and exons prediction tools at the NCBI server. Or try following programs or websites:

- **GENSCAN** (http://genes.mit.edu/GENSCAN.html)
- **Bioinformatics**, Genomics, Proteomics, Biotechnology and Molecular Biology Directory (http://www.bioinformatics.vg/biolinks/bioinformatics/verbose/Gene%2520Identification.shtml)

- **GenomeThreader** Gene Prediction Software
- **Bioinformatics** Links Directory

5.2. The 2ND step – physical properties of proteins

Proteins are characterized by a variety of physical properties that derive both from their nature as an amino acid polymer and from a variety of post-translational modifications. Several tools can be used to characterized general features of proteins:

- **AACompItool** – the results are send by an e-mail **(not very useful)**
- **AACompSimtool** – the results are send by an e-mail **(not very useful)**
- **Compute pI/Mw** - Molecular weight and isoelectric point **(very useful)**
- **IsotopIdent** – Predicts a theoretical isotopic distribution of a peptide, nucleotide
- **ProtParam** - Analysis of different physic-and chemical parameters including amino acids content and hydropathicity. For example, aliphatic protein with high lysine and/or arginine content can be a protein related with DNA **(very useful)**
- **Peptide cutter** - Predicts potential protease and cleavage sites **(very useful)**
- **TagId** – generates list of proteins close to a given pI and Mw (**Moderately useful)**

5.3 The 3RD step – cellular localisation and protein topology

Proteins can have a variety of functions and thus different cellular localisation. The proteins can be also classified upon their relationships to phospholipid bilayers i.e., those that are soluble and exist in the cytoplasm, in the lumen of an organelle or in the extracellular environments and those that are membrane attached associated wit a lipid bilayer. They can be integral membrane proteins or they may be peripherally attached to membranes. Proteins are also targeted to their cellular localisation because of intrinsic signals embedded in their primary structure sequence. Topology prediction, locating transmembrane segments can give important information about the structure and function of a protein as well as help in locating domains. If a protein has about 500 amino acids or more, it is rather certain, that this protein has more than a single domain.

5.3.1. Cellular localisation

A protein's subcellular localization is influenced by several features present within the protein's primary structure, such as the presence of a signal peptide or membrane-spanning α-helices.

- **PSORT** – subcellular localisation of proteins. Program analyses several features at once to generate an overall prediction of localisation site. Originally developed for prediction of protein localization in Gram-negative bacteria, PSORT was expanded into a suite of programs (PSORT, PSORT II, iPSORT) capable of handling proteins from all classes of organisms.
- **SOSOUI** – several parameters including solubility
- **TargetP** – localisation and signal peptides

5.3.2. Hydrophobic segments

In principal, hydrophobic segments correspond to transmembrane segments although it is not a rule.

- **ProtScale** – It gives amino acid representation – compute the profile produce by any amino acid. The hydropathicity and hydrophilicity, transmembrane tendency, secondary structure, conformational parameters and mutability can be searched.
- **SAPS** – (Statistical Analysis of Protein Sequence) calculates the amino acid content, charges, segments with high and low charges, hydrophobic and transmembrane segments.

5.3.3. Transmembrane regions

- **DAS** – transmembrane prediction server. Predictions based on pairwise comparisons. Two cutoffs are used, a "strict" one at 2.2. DAS score, and a "loose" one at 1.7 DAS score. The former indicates the number of matching segments while the latter indicates the exact segments.
- **PHDtopology** – predicts transmembrane segments, retrieves data informing about probability, 9 = maximum, 1 – minimum, defines region orientation. **Program is in PredictProtein software.**
- **TMHMM** – predicts transmembrane segments
- **TMPRED** – based on the transmembrane protein data basis. The program retrieves the number of transmembrane helices, table of correspondence, suggested topology models with start and stop positions of each segment, and graphical interpretation.

5.3.4. Signal peptides

- **NetNes** – prediction of protein rich export signals
- **ProP** – signal peptide cleavage site
- **Secretome** – non-classical and leaderless secretion of protein (there is no option for plants)
- **SignalP**– predicts signal peptides and their cleavage sites
- **TatP** – twin arginine signal peptides

5.3.5. Other non-globular regions

- **GLOBE**– predicts non-globular regions. **Program is in PredictProtein software.**
- **SEG**– predicts low complexity regions. **Program is in PredictProtein software.**
- **SMART** - predicts intrinsic disorders and low complexity regions

5.4. The 4TH step – Primary structure analysis

5.4.1. Repeat searches

Regions of low complexity, long stretches of repeated residues (Proline, Glutamine, Serine and Threonive) often indicate linker sequences and can be good indicators where to split proteins into domains. There are several useful programs to detect repeated sequences.

- **RADAR (RAPID)** – Rapid Automation Detection of Repeats and Alignment. Many large proteins have evolved by internal duplication. Any internal sequence repeats correspond to functional and structural units. RADAR is searching for segmenting a query sequence into repeats. It identifies short composition biased as well as gapped approximate repeats and complex architecture repeats. The program retrieves alignments score, Z-score and alignment:

  e.g., 9-137 (118.41/121.39) LPNL...aap indicates
  start and stop (alignment score of this repeat unit/ Z-score) alignment... not aligned amino acids.

- **REPRO** – de novo repeat prediction – it is able to recognise distant repeats in a single query sequence. This technique relies on a local alignment strategy to find non-overlapping top scoring fragments followed by a graph-based interactive clustering procedure to delineate the repeats. The analysis is done in two steps:
  1. Calculation a list of N-top scoring, non-overlapping local alignment. N should be specified by the user.
  2. Then, prediction is made. Gap opening and extension penalties can be chosen. Default values are 10 and 1, respectively, however, values can be changed to make more specific searches.
- **XSTREAM** – de novo tandem repeat and architecture modelling but only for chosen organisms.
- **REP** – search for a collection of repeats in known protein families.
- **SAPS** – Useful program that calculates several parameters:
  1. Number of residues, molecular weight

2. Usage of all 20 amino acids
3. Cleavage distribution
♦a significant cumulative positive score indicates a region of high charge concentration.
♦positive, negative and mixed charge clusters are recognised.

4. Transmebrane segments
5. Repeated structure – calculates separated, tandem and periodic repeats
6. Multiplets
7. Periodicity analysis
8. Spacing

5.4.2. Coiled regions

Coiled coils regions sometimes indicate where proteins can be divided into domains.

•**AGADIR**– an algorithm to predict the helical content of peptides. It predicts the helical behaviour of monomeric peptides. It accepts two modifications at N-terminus (accetylation or succymilation) and one at the C-terminus (amidation). Some parameters can be chosen.

•**COILS** – It compares a sequence into a database of known parallel two-stranded coiled coils and delivers a similarity score. By comparing this score to the distribution of scores in globular and coiled-coil proteins the program calculates the probability that the sequence will adopt a coil-coil confirmation.

•**MarCoil1** – It retrieves predicted coiled domains and probability.

•**PARACOIL2** – predicts parallel coil fold from sequence using pairwise residue probabilities with the paracoil algorithm.

•**ProtScale** – It gives amino acid representation – compute the profile produce by any amino acid. The hydropathicity and hydrophilicity, transmembrane tendency, secondary structure, conformational parameters and mutability can be searched.

•**SMART** - predicts coiled-coils regions

•**2ZIP**– predicts Leucine Zippers.

5.4.3. Other motifs

•**PESTfind** – It searches for PEST motifs i.e., with high concentration of proline (P), glutamic acid (E), serine (S), threonine (T) and to a lesser extent aspartic acid (D). PEST motifs reduce the half lifes of proteins dramatically so they are target proteins for proteolitic degradation.

5.5. The 5H step – Post-translational modifications

5.5.1. Organella transit peptides

•**ChloroP** – chloroplast transit peptides.

•**LipoP** – lipoproteins and signal peptides in Gram positive bacteria.

•**MITOPROT** – mitochondrial targeting sequence.

•**Predator** – mitochondrial and plastid targeting sequence.

•**PTS1** – peroxisomal targeting signal containing proteins.

5.5.2. Prediction of glycosylation sites

The majority of programs are for mammalian proteins. For all Eucaryotes two programs are available: OSPET and YinOYang.

5.5.3. GPI modification site

•**big P1Predator** – C-terminal modifications. No option for plants.

•**GPI-SOM –** GPI anchor and cleavage sites.

5.5.4. N-terminal sites – fatty acid attachment

Numerous proteins could be covalently modified by a variety of lipids including myristate (C14), palmitate (C16), farnesyl (C15), geramylgeranyl (C20), glycosylphospatidyloinositol (GPI). Most of lipid modifications are irreversible. S-palmitoylation – thioacylation (S-acylation) could attach 16-carbon saturated fatty acids to specific cysteine residues. This process enhances the surface hydrophobicity and membrane affinity of protein substrates and play important role in modulating protein trafficking.

•**Myristoylator** – predicts N-terminal myristoylation of proteins by neural networks in eukaryota. A sequence should have N-terminal glycine.

•**CSSPalm** – fatty acids (**does not work with Linux**)

5.5.5. N-terminal sites – phosphorylation

•**NetPhos** – Ser, Thr and Tyr phosphorylation sites in eucaryotic proteins.

•**NetPhosK** – kinase specific phosphorylation sites in eucaryotic proteins.

•**DisulFIND** – predicts cytosine sulfation sites. **Program is in PredictProtein software.**

•**Sulfinator** – predicts tyrosine sulfation sites. Tyrosine sulfation is important post-translational modification of proteins that go through the secretory pathway.

•**Sulfosite** – predicts tyrosine sulfation sites (**not readable in Linux**).

•**SUMOplot** – predicts SUMO protein attachment sites. It is related with experimental data – it explains larger Mw than expected on SDS gels due to attachment of SUMO protein (11 kD) at multiple positions of a protein.

•**TermiNator** – predicts N-terminal methionine excision, N-terminal acetylation, N-terminal myristoylation, and S-palmitoylation of either prokaryotic or eukaryotic proteins. .

•**ProP**– arginine and lisine cleavage sites in eucaryotic protein segments.

5.6. THE 6TH STEP – DOMAINS ASSIGNING

One of the most basic questions about a gene or protein is whether it is related to any other gene or protein. Relatedness of two proteins at the sequence level suggests that they are homologous. The regions of significant amino acid identity between at least two proteins can be identified. Such regions that share significant structural features and sequence identity are defined as signatures. **A signature denotes a protein category, such as domain or family of motifs.** Structurally conserved and variable regions can be identified by patterns and similarity searches. These analyses are accomplished by sequence alignments

Domains

A domain (module)is a region of a protein that can adopt a particular three dimensional structure. It has a distinctive secondary structure and a hydrophobic core. A group of proteins that share a domain is called **a family**. Further, protein domains are classified upon the subcellular localization (intracellular, extracellular) or in terms of function. Domains are evolutionary related. Homologous domains with common functions usually show sequence similarities. The entire protein may consist of one domain, others have more domains. Comparison of two proteins indicate that the domains occupy different regions of each protein. A domain may be repeated several times.

Motifs

Motifs (fingerprints) are short, conserved regions of proteins. A motif consist of a pattern of amino acids that characterises a protein family. Motifs are subsets of protein domains. The size of a defined motif is between 10-20 contiguous amino acids residues. Some common motifs form a transmembrane domain or a consensus phosphorylation sites. Small motifs can provide a signature for a protein. Set of sequence motifs needs no

necessary represent homologs.

Repeats

A repeat is a region that is not expected to fold into a globular domain of its own.

Profile

A profile is table of position-specific scores and gap penalties representing an homologous family

Locating domains

If a protein has more than about 500 amino acids, it is nearly certain that it can be divided into domains. If possible it is advisable to split such a large protein up and consider each domain separately. The location of domains can be predicted in a few different ways. Signs of domains and assigning sites for split.

- Homology occurs only other a portion of the probe sequence and the other sequences are whole.
- Regions of low complexity, long stretches of repeated residues (proline, glutamine, serine or threonine) often separate domains in multidomain proteins.
- Transmembrane segments can be dividing points since they easily separate extracellular from intracellular domains.
- Coiled- coils region may indicate where proteins can be divided.
- Secondary structure prediction methods often predict regions of proteins to have different structural classes.

If a protein is separated into domains, then it is necessary to repeat all the database searches and alignments using the domains separately.

5.6.1. Patterns of profile searches

The programs predicting domains usually describe the key residues that are conserved and define the family. For example:

[IV] – G – x – G – T – [LIVMF] – x(2) – [GS] means:

isoleucine or valine at the 1st position, glicine at the 2nd, any amino acid at the 3rd, glicine (4th), trypsine (5th), lisine, isoleucine, valine, methionine or fenylalanine at 6th, any amino acid at 7th and 8th and glicine or serine at 9th. Protein signature databases have become vital tools for identifying distant relationships in novel sequences and hence are used for the classification of protein sequences and for inferring their function. . Searches before applying BLAST greatly help in inspecting BLAST results.

- **3of5** – searches for user defined patterns
- **9aa TAD** – predicts of nine amino acid transactivation domain – for mammalian, yeast, viral and plant proteins.
- **ELM** - Eukaryotic Linear Motif resource for functional sites in proteins. It is a resource for predicting functional sites in eukaryotic proteins. Putative functional sites are identified by patterns (regular expression) **(Difficult to connect)**
- **FingerPRINTScan** – scans a protein sequence against the PRINS Protein Fingerprint database.
- **HAMAPScan** – scans a sequence against HAMAP families. The HAMAP families are a collection of orthologous **microbial** families manually created by expert curators.
- **HITS (MotifScan)** – relationships between protein sequences and motifs. **(very long processing)**
- **InterProScan** – integrated motif search in PROSITE, Pfam, PRINTS and other family and domain databases.
- **PATTINPROF** – scans a protein sequence or a protein database for one or several patterns. Patterns should be defined by a user.

•PPSearch – scans a sequence against PROSITE and determines the function of uncharacterised proteins.
•PROSITEScan (ProScan) - scans a sequence against PROSITE.
•ScanProsite – scans a sequence against PROSITE and UNIPROT, predicts intra-domain features, gives alignment of all members and enzyme descriptions.
•SMART – Simple Modular architecture Research Tool – prediction of transmembrane, coiled coil regions, segments of low compositional complexity, signal peptides, regions containing repeats. If a domain is predicted, BLAST can be automatically run out. SMART can also do the alignments.

5.6.2. Similarity searches

Given the choice of aligning a DNA sequence or the sequence of the protein it encodes, it is usually more informative to compare protein sequence because not all changes in a DNA sequence resulted in the change of the amino acid (3rd position of a codon), many amino acids share related biophysical properties (e.g., lysine and arginine are basic amino acids), and protein sequences can identify homologues from organisms that last shared a common ancestor over 1 billion years ago while DNA sequences allow lookback times up to about 600 MYA. Therefore when a nucleotide coding sequence is analysed, it is often preferable to study its translated protein. However**, nucleotide comparisons are appropriate while confirming the identity of a DNA sequence, searching for polymorphism, analysing the identity of a cloned DNA fragments** and many others.

Homology
Two sequences are homologous if they share a common evolutionary ancestry. Sequences are either homologous or not, there are not degrees of homology. Homologous proteins almost always share a significantly related three-dimensional structure. When two sequences are homologous, their amino acid or nucleotide sequences share significant similarity.

Identity or similarity
Quantities that describe the relatedness of sequences. Notably, two molecules may be homologous without sharing statistically significant amino acid or nucleotide identity. In general, three-dimensional structures diverge much more slowly than amino acid sequence identity. The percent similarity of two protein sequences is the sum of both identical and similar matches. The percent of identity is the sum only of identical matches. Percent identity is not an exact indicator of the number of mutations that have occurred across a protein sequence. Any position is subject to multiple hits. For example, when a protein sustains about 250 hits per 100 amino acids, it may have about 20% identity with the original protein.
•ORTHOLOGS – homologous sequences in different species that arose from a common ancestral gene during speciation. Orthologs are presumed to have similar biological functions.
•PARALOGS – are homologous sequences that arose by a mechanism such as gene duplication. For example α and β hemoglobin have descended side by side during the history of an organism. The α hemoglobin in man and mouse have resulted from speciation, so that the history of the gene can reflects the history of the species.

The relatedness of any two proteins (or nucleotide sequences) can be assessed by **performing a pairwise alignment**.

Pairwise alignment
The relatedness of any two proteins (or nucleotide sequences) can be assessed by **performing a pairwise alignment. It is the process of lining up two sequences to achieve maximal levels of identity**. The purpose is to assess the possibility of homology nevertheless, the strongest evidences always come from structural studies in combination with evolutionary analyses. The study of homologous protein or DNA

sequences by pairwise alignment involves an investigation of the evolutionary history of that sequence. As we examine a variety of homologous proteins, we can observe a wide range of conservation between family members. Some positions are perfectly conserved while the others represent conservative substitutions i.e., one amino acid is replaced by another with similar properties. There are many algorithms used for pairwise alignment. **A heuristic algorithm** is one that makes approximations of the best solutions without exhaustively considering every possible outcome.

- **Substitutions** – results in the alignment of two non-identical amino acids.
- **INSERTIONS, DELETIONS** – results in gaps in the alignments. The effects of gaps is to make the overall length of each alignment exactly the same. Gaps can allow the full alignment of two proteins or nucleotide sequences.

Quantitative scoring system for pairwise alignments

Rules by which evolutionary changes occur in protein sequences are described by several models. These systems accounts for scores between any proteins.

- **PAM (Accepted Point Mutations)** – an accepted point mutation is a replacement of one amino acid in a protein by another that has been accepted by natural selection. Such mutation occurs when:
  - a gene undergoes a DNA mutation such that it encodes a different amino acid (sense mutation);
  - the entire species adopts that change as the predominant form of the protein.

Accepted mutations are based on empirically observed amino acid changes. Based on observed frequencies, the relatives **mutabilities** of each amino acid are calculated (number of times each amino acid mutated by the overall frequency of its occurrence). The less mutable residuals have important structural features, the most mutable can easily be replaced by the others. The most common substitutions are:

- glutamic acid for aspartic acid (both are acidic)
- serine for alanine (hydroxylated)
- serine for threonine (hydroxylated)
- isoleucine for valine (hydrophobic)

The least mutable amino acids are generally specified by one or two codons.

Data on accepted mutations and frequencies of amino acids are used to generate **a mutation probability matrix,** describing the probability that original amino acid will be replaced by another amino acid over a defined evolutionary interval. **The unit of evolutionary divergence is defined the interval in which 1% of the amino acids have been changed between two sequences**.

- PAM1 – based upon an alignment of proteins, all of which are at least 85% identical within a protein family. In this matrix some substitutions are very rare (tryptophan to threonine).
- PAM100 – reflects the amino acid substitutions that occur in distantly related proteins. It is derived by multiplying the PAM1 by itself, up to hundreds of times.
- PAM250 – reflects the amino acid substitutions that occur in distantly related proteins. It is derived by multiplying the PAM1 by itself, up to 250 of times. It is one of the most common matrices used for BLAST searches. It applies to an evolutionary distance where proteins share about 20% amino acid identity. The PAM250 assumes the occurrence of 250 point mutations per 100 amino acids.

To derive a scoring system, PAM matrix is converted into a log-odds matrix (relatedness odds matrix). An “odd ratio” is the probability that some amino acid a will change to amino acid b in some PAM interval.

$S(a, b) = 10 \log (M_{ab}/p_b)$ where M_{ab} is the probability that the aligned pair of amino acids a, b, represents and authentic alignment, p_b i the normalized frequency of some amino acid. For example:

S(a, tryptophan) = 10 log (0.55/0.01) = 17.4, a score indicates that authentic alignment is 50 times more frequent that the chance alignment. When two sequences are aligned, the score is simple the sum of the scores for all aligned residues. Using PAM1 matrix, amino acid substitutions would earn large penalties and would generally not be tolerated. But using PAM250 many amino acid substitutions would receive relatively minor penalties, and a significant score is possible (Figure 1).

EXAMPLE

Comparison using PAM250 matrix an overlap of 20 identical residues over a span of 81 amino acids is detected with 24.7% identity. Using PAM40 the best aligned sequence is only 10 amino acids in length (60% identity) that is biologically meaningless.

•**BLOSUM** (**Blocks Substitution Matrix)** – based on BLOCK database, which consisted of over 500 group of local multiple alignments of distantly related protein sequences.

- ♦BLOSUM62 – based upon an alignment of proteins, all of which are at least 62% identical within a protein. It is the default scoring matrix used by most BLASTs (Figure 1).
- ♦BLOSUM80 – based upon an alignment of proteins, all of which are at least 80% identical within a protein. It is the default scoring matrix used by most BLASTs.
- ♦BLOSUM95 – based upon an alignment of proteins, all of which are at least 95% identical within a protein. It is the default scoring matrix used by most BLASTs.

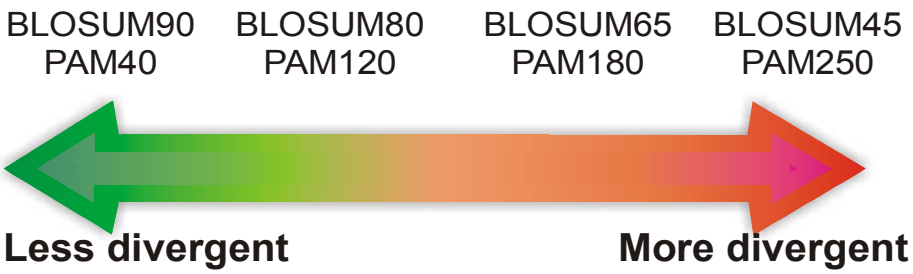

Figure 1. Relationships of PAM and BLOSUM matrices

Alignment algorithms

When two proteins are aligned, there is an enormous number of possible alignments. There are two types of alignments, global and local.

•**GLOBAL SEQUENCE ALIGNMENT (the Needleman-Wunsch algorithm)** – focuses on the entire sequences. It produces an optimal alignment of two protein or DNA sequences allowing the introduction of gaps. The approach is based on three steps, setting up a matrix, scoring the matrix and identifying the optimal alignment. The optimal matrix is identified by an optimal path while moving along the diagonal of the matrix.

- ♦Setting up a matrix (Figure 2)

  Two sequences are compared in a two-dimensional matrix. A perfect alignment would be represented by a diagonal line, any mismatches would be still represented by a diagonal line, gaps are represented by horizontal (gaps in the vertical sequence) or vertical line (gaps in the horizontal sequence).
- ♦Scoring the matrix (Figure 3)

  The goal is to find the best alignment. The amino acid identity gains a score of +1. The assignment of scores begins at the bottom right of the matrix, corresponding to the carboxy termini of the protein. The scores must increase and this means that to calculate the score we must look down and to the right of the matrix.

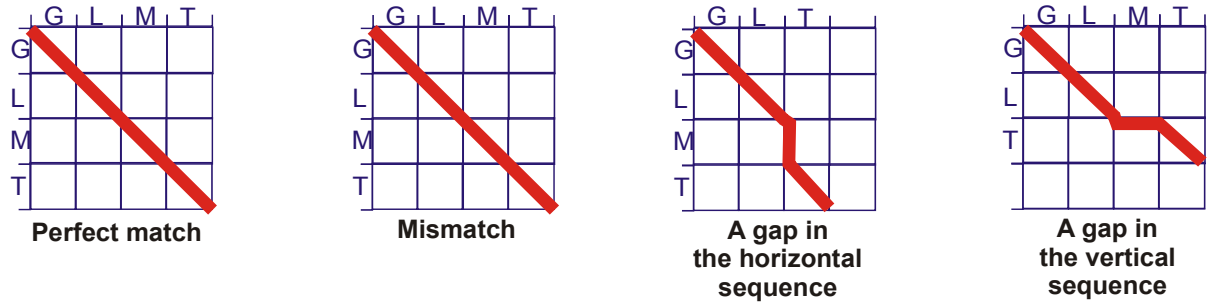

Figure 2. Alignment of two sequences using a global alignment algorithm

	A	B	C	N	J	R	Q	C	L	C	R	P	M
A	1												
J					1								
C			1					1		1			
J					1								
N				1									
R											1		
C			1					1		1			
K													
C			1					1		1			
R						1					1		
B		1											
P												1	

Assigning scores for identical amino acids

	A	B	C	N	J	R	Q	C	L	C	R	P	M
A	1												
J					1								
C			1					1		1			
J					1								
N				1									
R											1		
C			1					1		1			
K													
C			1					1		1			
R			1	1	1	1	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

	A	B	C	N	J	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
J	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
J	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	1	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Matrix filling taking into account values on the right and bottom

	A	B	C	N	J	R	Q	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
J	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
J	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	1	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

Designing a path

1. ABCNJ - RQCLCR - PM
   AJC - JNR - CKCRBP -

2. ABC - NJRQCLCR - PM
   AJ CJ N- R -CKCRBP -

Figure 3. Alignment of two sequences using a global alignment algorithm – basis of scoring

•**LOCAL SEQUENCE ALIGNMENT**– focuses on the regions of greatest similarity between two proteins. BLAST represents a simplified form of local alignment. A local sequence alignment algorithm resembles the global one, there is no penalty for some internal positions and the alignment does not extend to the ends of sequences.

Significance of pairwise alignments

Significance of pairwise alignment is not easy to assess however, a rule is that if two protein sequences share at least 25% of similarity over a span of at least 150 amino acids, they are probably significantly related. Generally, it is accepted that:

- two proteins are unrelated if they share 10-20% amino acid identity when aligned.
- two proteins are regarded as homologous if they share 40% amino acid identity over a reasonable long stretch (70-100 amino acids),
- two proteins are probably significantly related if they share 20-25% amino acid identity over a reasonable long stretch (70-100 amino acids). In that case when proteins share limited identity, statistical tests should be used to determine if they are related. The most common is calculation of Z score

Z-Score – indicates how far and in what direction an item deviates from its distribution's mean. The Z-score expresses the divergence of the experimental result x, from the most probable mean value. The larger the value of Z-score the less probably the experimental result is due to a chance. Probability calculated from Z-score indicates the likelihood that results are not due to a chance. The lower probability the more probable results are not due to a chance.

For sequence alignments Z score is calculated using a real comparison of a sequence, then a sequence is scrabbled (shuffled) 100 times, 100 alignments are performed and the authentic score is compared to the mean randomized score. If 100 alignments of a shuffled sequence have a score less than the authentic score then the probability that real one occurred by a chance is less than 0.01. For global alignments a Z score can not be converted into a probability value. For local alignments the expected value (E) can be calculated , which describes the number of hits that can be obtained by chance.

Similarity searches by pairwise alignments - programs

Rules by which evolutionary changes occur in protein sequences are described by several models. These

•**MPSearch** – it utilises an exhausting algorithm which is recognised as the most sensitive sequence comparison method available. It is capable of identifying hits in cases where BLAST fails.

•**ScanPS** – it is a program for comparing a protein sequence against a database of protein sequences. It is capable of identifying multiple domain matches.

•**BLAST (Basic Local Alignment Search Tool)** - it is a family of programs that allows a DNA or protein query sequence to be compared to a database. A DNA sequence can be converted into six potential proteins. Applications of BLAST include:

- ♦ identifying orthologs and paralogs,
- ♦ determining the identity of a DNA or protein sequence,
- ♦ discovering new genes,
- ♦ investigating expressed sequence tags,
- ♦ predicting protein structure and function.

BLAST search steps

❶Specifying a sequence of interest

❷Selecting BLAST program

- ♦**Nucleotide blast** (blastn, megablast, discontiguous blast)– for a nucleotides query
- ♦**Protein blast** (blastp, psi-blast, phi-blast) – for a protein query
- ♦**Blastx** – protein database against a translated anucleotide query
- ♦**tblastn** – translated nucleotide database against a protein querry
- ♦**tblastx** – translated nucleotide database against a translated nucleotide query

In addition more specialized searches can be perform. This include:

- ♦searches of archives
- ♦finding conserved domains in a sequence
- ♦finding sequences that have similar domain architecture (cdart)
- ♦search sequences that have gene expression profiles (GEO)
- ♦search immunoglobulines (IgBLAST)
- ♦Search for SNPs (snp)
- ♦screen for vector contamination (vecscreen)
- ♦Align two sequences using BLAST (bl2seq)

❸Selecting a database

For proteins:

- ♦nr database (nonredundant) – combined protein records from GenBank, the Protein Data Bank (PDB), SwissProt, PIR, PFF **(the most frequent choice).**

For DNA:

- ♦nucleotide nr database (nonredundant) – nucleotide sequences fromGenBank, EMBL, DDBJ, PDB. It is derived by merging several main protein and DNA databases. **(the most frequent choice)**
- ♦EST database

❹Program options (search paramethers)

❖**CD-Search**

Conserved Domain Database

❖**Limit by Entrez Query**

By entering the user defined term (e.g., protease) and performing a BLAST search with a sequence.

❖**E value**

The E value is a number of different alignments with scores equal or greater than some scores S that are expected to occur by chance. The E value equal to 1e-4 i.e., 1 x 10^{-4} or 0.0001 for a score 45 bits indicates that only in 1 per 10 000 alignments a score 45 is expected to occur by chance. **Increasing the E value returns more hits.** The default number 10 means that 10 hits with scores equal or better than the alignment score S are expected to occur by chance. Although hits with E values much higher than 0.1 are unlikely to reflect true sequence relatives, it is useful to examine hits with lower significance (E between 0.1-10) for short regions of similarity. In the absence of longer similarities , these short regions may allow the tentative protein assignment of biochemical activities or ORFs. The significance of any such region must be assessed on a case by case basis.

❖**Word size**

The BLAST first divides the querry into a series of smaller sequences (words) of a particular length (word size). For proteins the larger word size the more accurate search. The default is 3 or 2 and

there is rarely a need to change it. For nucleotide, the default word size is 11. Lowering the word size yields more accurate search.

❖**Matrix**

PAM and BLOSUM matrices are available for protein searches. The default matrix is BLOSUM65.

❖**Filter low complexity**

It is appropriate to filter low complexity region. They appear as X's in the alignment. Low complexity regions are those having commonly found stretches of amino acids or nucleotides with limited information content (dinucleotide regions, rich in one or two amino acids, hydrophobic amino acids from a transmembrane domain).

❖**Mask lower case letter**

Some types of low complexity sequences may not be detected by the filtering option, for example coiled-coil and transmembrane regions. They are detected by the programs listed under point 1. Since these regions may lead to mismatches it is worthy to mask them manually by lower case letters (or X in nucleotide databases).

❖**Ungapped**

The default **gapped** setting of BLAST 2.0 reports the best local alignments and is suitable for most applications. An **ungapped** search, on the other hand, may be desirable when hits that align to the entire length of the query are most interesting. An ungapped search can be specified by checking the ungapped option or by increasing the gap existence penalty

❖**Advanced options**

The complete list of Advanced options for use with blastp are as follows:

∗G - Cost to open a gap; default = 11 (typically 10-15)

∗L - Cost to extend a gap; default = 1 (typically 1-2)

∗E - Expectation value (E); default = 10.0

∗W - Word size; default is 11 for blastn, 3 for other programs.

∗V - Number of one-line descriptions (V) [Integer]; default = 100

∗b - Number of alignments to show (B) [Integer]; default = 100

A gap compensates for insertions and deletions. Since a single mutation may cause insertion or deletion of more than one residue, the presence of a gap is penalized heavily, whereas a lesser penalty is ascribed to extensions of a gap.

Limited values for gap existence and extension are supported for these three programs. Supported and suggested values (Existence, Extension) include: 10,1; 10,2; 11,1; 8,2; and 9,2.

❺BLAST output

❖**Top**

Details of the search

❖**Middle part**

A. A list of database sequences that match the query sequence

B. A graphical overview. Each bar represents a database sequence that matches the query sequence. The matches are sorted out according to decreasing E values, so the most similar hits are shown at the top in red then pink, greed, blue, black. Hatched areas (if any) represents the nonsimilar sequence between distinct regions of similarity found within the same database entry.

❖**Alignments – one line descriptions**

A. Descriptions beginning with the common sequence identifiers (RefSeq, pdb)

B. Brief description of the sequence

C. The bit score

D. The expect value E

E. Link to the full GenBank entry

❖**Alignments**

A series of pairwise alignments with the percent identity and the percent similarity (positives)

❻Interpretation of results

❖**E value**

As E approaches to zero, the probability that the alignment occurred by chance approaches to zero.

❖**Raw and bit score**

Raw scores are calculated from the substitution matrix and gap penalty parameters that are chosen. The bit score is calculated from the raw score by normalizing and takes into account the size of the database being queried.

❖**P value**

The P value is the probability of a chance alignment. The P and E values are different ways of representing the significance of alignments. The most highly significant P values are those close to zero (traditionally less than 0.05).

❖**Identification of homologues**

When a BLAST search is performed, a true positive should first be defined as a database match that is homologous to the query sequence. Homology is inferred based on sequence similarity and from statistical evaluation of the search results. The first way is to inspect the E value (e-105 = 1 x 10^{-105}, 3e-13 = 3 x 10^{-13}). However, the search results should be supplemented with evaluation of protein structure and function. Proteins with the non significant E value (e.g., 0.54) can be homologous. To decide whether two sequences are homologous or not some additional points should be taken into account.

A. The size of proteins (can be the same for homologous).
B. The presence of a common motif or signature.
C. Sequences are a part of a reasonable multiple alignment.
D. Similar biological function
E. Similar a three-dimensional structure
F. If a BLAST search results in a marginal match to another protein, perform a new BLAST using that distantly related protein.

❖**Handling with too many results**

A. Limit by “Entrez querry”.
B. Limit by organism.
C. Use a portion of the query sequence especially for multidomain proteins.
D. Adjust the scoring matrix and E value

❖**Handling with multidomain protein**

A. Repeat searches with each domains.
B. Limit by organism.
C. Searches can be also perform in the EST database.
D. Adjust the scoring matrix and E value

❼Advanced BLAST

There are three types of specialized BLAST, organism-specific BLAST, molecules-specific and specialized database algorithms.

❖**Organism-specific**

There are many databases with sequences from different organisms, plants are not the major component of these databases.

❖Molecule-specific

The programs allow searches of particular molecules e.g., immunoglobulines.

❖Position Specific BLAST (Psi-BLAST)

It is especially useful for distantly related proteins that share only limited sequence identity. They can have the same 3D structure but no apparent similarity in alignments. The search process is continued iteratively up to the moment no new results are found. Typically, after the second or third iteration, the bit score continues to rise and the E value is dropping, the number of gaps decreasing. However, it is reasonable to mask all disturbing regions and to use only a single domain in multidomain proteins.

❖Pattern HIT Initiated BLAST (PHI-BLAST)

It is especially useful for identifying patterns or signatures and defining the protein family.

❖BLAST for gene discovery

A. The database (best choice is DNA or EST database) is search with a known sequence.

B. Some hits match the query exactly or nearly exactly. They are not novel genes.

C. Some hits match the query significantly but proteins are not annotated. They are candidate novel genes.

D. Some hits match the query non-significantly. They can be novel genes.

5.7. THE 7TH STEP – MULTIPLE ALIGNMENT

A multiple sequence alignment is a collection of three or more protein or nucleic acid sequences that are partially or completely aligned. Homologous residues are aligned in columns across the length of the sequences. These residues are homologous in an evolutionary sense i.e., they are probably derived from a common ancestor. Aligned residues also tend to occupy corresponding positions in the three-dimensional structures of aligned proteins. Multiple sequence alignments are easy to generate for closely related sequences but as soon as the sequences exhibit some divergence, the multiple alignments become extremely difficult. This is because protein sequences evolves more rapidly than structures. Sequences can share only about 20-30% of similarity but the three-dimensional structures can be nearly identical. A multiple sequence alignment can be generate because of some conservative features:

- highly conserved residues such as cysteines,
- conserved motifs such as transmembrane regions,
- conserved features of the secondary structure (α-helices, β-sheets),
- regions that show consistent pattern of insertions and deletions.

Multiple alignments provide information about:

- protein domain structure,
- location of residues likely to be involved in protein function,
- residues likely to be buried in the protein core or exposed to solvent,
- more data for homology modelling and secondary structure predictions.

While doing multiple alignment, it is advisable to look through the output carefully and throw things that do not appear to be a member of a sequence family.

5.7.1. Approaches to perform multiple sequence alignments

The most popular approaches involved the progressive and hierarchical alignment. Many databases of multiple sequence alignment are available.

- **PFAM** – Pfam-A is a manually curated collection of protein families in the form of multiple sequence alignments and profile HMMs. Pfam-B is a helpful supplement.
- **SMART** – database of protein families implicated in cellular signalling, extracellular domains and chromatic function.
- **CONSEVED DOMAIN DATABASE** – database at NCBI enabling to identify conserved domains.
- **BLOCKS** – ungapped sequence alignment of the most highly conserved proteins.
- **PROSISTE** – contains description of proteins and protein families.

Multiple sequence alignment can be generated using programs such as ClustalW and ClustaX. The ClustalX is a windows interface for ClustalW.

- **CLUSTALX** – input sequences in the FASTA format should be saved as a text file and then loaded. First no gaps are included, and "complete sequence alignment" is performed. The alignment should be manually inspected and modified as necessary. Do not feed the program with all sequences BLAST find.
  - First, remove all redundant sequences.
  - Chose 10-20 the most similar sequences representing different species and align them using the multiple alignment mode.
  - Add 5-10 more distant sequences using the profile mode. A profile is a position scoring matrix. Profiles are created from existing alignments. New sequences can be optionally aligned in to existing alignments.

5.7.2. Features of a "good alignment"

One general problem is in deciding if a set of sequences are well aligned ("significant alignment") and if they are related to each other. Significance of alignments is a very difficult thing in a statistical sense but it is possible to take some simple steps to check an alignment.

- **OVERALL LOOK** – real alignments of homologous sequences have "neat-looking" blocks of alignments separated by sections of gaps (gaps indicate loop regions with no conserved secondary structure).
- **PATTERN OF CONSERVATION** – an examination of the pattern of conservation in the conserved blocks usually indicates some partially or weakly conserved columns (stars for fully conserved or dots for weakly conserved). If the sequences are not all homologous, there are very few stars or dots. There are gaps everywhere, indicating that there is no pattern of conservation.
- **QUALITY MENU** – guides to columns, residues, or sequences.
- **PARAMETERS** – pairwise parameters do not need to be change. They do not have effects on multiple alignments. Multiple alignments parameters should be changed for DNA sequences. Gap penalties set to 7.5 and gap extensions set to 3.33 usually give better alignments. For sequences showing less than 30% of residues identity it is worthy to use the Delay Divergent Sequences option. The best matching can also be find if the negative matrix option is used.

5.8. THE 8^{H} STEP – COMPARATIVE HOMOLOGY MODELLING

If a protein shows significant homology to another protein of know three-dimensional structure, then a fairy accurate model can be obtained via homology modelling. It is also possible to built models if a suitable fold recognition has been found. Models can be generated using several web service or programs.

- **SWISSMODEL** – it is a fully automated protein structure homology modelling server. The first step is alignment, and then modelling. It is possible to send a protein sequence only. However, it is recommended doing this if the degree of sequence homology is high (50% or above). Sequence alignments, particularly those involving proteins having low percent sequence identities can be inaccurate, and the model will be wrong. Therefore, it is better to look over alignment carefully before building a model.
- **MODELLER** – the program used for homology or comparative modelling of protein three-dimensional structures. The user provides an alignment sequence to be modelled with known related structures and the program automatically calculated a model containing all non-hydrogen atoms. It can also performs de novo modelling of loops in protein structures, optimization of various models of protein structures, multiple alignment, clustering, searching of sequence databases, comparison of protein structures.

Once a three-dimensional model is built, it is useful to look at 3D structures using:

- **MolMol** –
- **RasMol** – t

6 AB INITO STRUCTURE PREDICTIONS

In the absence of detectable homology, protein structure may be assessed de novo. Protein folding is modelled based on global free-energy minimum estimates and there is an overall comparison to known structures. The “Rosetta Stone: method is one of the most successful ab initio strategies.

6.1. SECONDARY STRUCTURE PREDICTIONS

The aim of secondary structure prediction is to provide the location of alpha and beta helices. All methods rely on availability of large families of homologous sequences. In practise it is recommended getting as many methods as possible and combining this with human insights. If it is possible to align all predictions with multiple sequence alignment a consensus picture will be get.

- **APSSP** – Advanced Protein Secondary Structure Prediction (**very long processing**).
- **GOR** – secondary structure prediction method.
- **HNN** – Hierarchical Neural Network method.
- **HTMSRAP** – Helical Transmembrane Segment Rotational Angle Prediction. Transmembrane segments should first be decided.
- **JPRED**– a consensus method for protein secondary structure prediction. First it makes alignments, then predict structure. The analysis usually takes some time, so it is better to use e-mail.
- **JUFO** – predicts secondary structure from sequence, multiple alignment, from sequence and a 3D structure, protein chemical shifts predictions, strain loop strands motifs.
- **PORTER** – (**very long processing**).
- **PredictProtein**– service for sequence analysis, structure and function predictions. The program retrieves smultiple sequence alignment, PROSITE sequence motifs, low complexity regions, nuclear localisation signals, regions lacking regular structure (NORs) and predict a secondary structure, solvent accesibility, helices, coiled-coil regions, disulfide bonds, sub-cellular localisation and functional annotations.
- **PROF**– cascade multiple classifiers for secondary structure prediction.
- **PSIpred**– various protein structure prediction methods.
- **Rosetta**– works by simultaneous optimization of side chain conformation and rigid body position of the two docking partners.
- **SOMPA**–
- **SSPro**– Secondary Structure Prediction using bidirectional recurrent neural networks.
- **DLP-SVM**– domain linker prediction.

6.2. 3D STRUCTURAL SIMILARITIES - TERTIARY STRUCTURE PREDICTIONS

The aim of tertiary structure prediction is to find a suitable fold for a protein among known 3D structures. Proteins often adopt similar folds despite no significant sequence or functional similarity. For many proteins there is a suitable structure in the database from which to built a 3D model. Methods of protein fold recognition attempt to detect similarities between protein 3D structure that are nor accompanied by any sequence similarity. The methods of protein fold recognition always give somehow different results, so each method should be run on as many homologoues as possible. The accuracy of methods should be taken into account and they should not be treated as a white boxes. Moreover, function should also be taken into account. If there is a functional similarity, then a remote homologue can be detected. Alignments done by programs can only be used as starting points.

6PROTEIN FUNCTION

Function is defined as the roll of a protein in a cell. Function can be characterized in several ways.

- A protein has a biochemical function synonymous with its molecular function.
- Functional assignments is based upon homology. If a hypothetical protein is homologous to an enzyme, it is often provisionally assigned that enzymatic function.
- Function may be assigned based upon structure.
- Functional assignments is based upon homology. If a hypothetical protein is homologous to an enzym

The manual is under constant improvements and this is the first but not the last version.

The manual was produced during my fellowship at the University of Crete, Greece (prof. Michael Kokkinidis) within EU Marie Curie Transfer of Knowledge project entitled "**Genomic Approaches for Crop Improvement**" (GenCrop) N° MTKD-CT-2004-509834. co-ordinated by the Department of Genetics, University of Warmia and Mazury. Because the manual is addressed to undergraduate and PhD students of the Department of Genetics, some basic information is included. However, it can be used by anybody who will find it useful. Please cite the WebPage if you use it.

I would like to thank you Prof. Michael Kokkinidis and all his team for help during studying protein prediction methods.