A Review On Role of Bio-Computing In Analysis of Drug Targeted Proteins

Narendra Maddu
Assistant Professor
Department of Biochemistry
Sri Krishnadevaraya University,
Anantapur-515 003. AP, India

Abstract: This explains in very general terms how a biologist goes about using computer programs for sequence analysis. The fundamental building blocks of life are proteins. In addition, much of the structure of a cell is made up of proteins. Proteins are variable length linear, mixed polymers of 20 different amino acids. Other terms used more or less inter changeably for amino acid polymers are peptides and polypeptides. These topologically linear polymers fold upon themselves to generate a shape characteristic of each different protein, and this shape along with the different chemical properties of the 20 amino acids determine the function of the protein.

Index Terms:*Bio-computing, Drug Design, Genomics, Bioinformatics.*

Introduction

One of the most important concepts in modern biology is that the functional properties of proteins is determined largely by the sequence of the 20 amino acids in the linear polypeptide chain; that in many cases proteins are largely self-folding. Thus, in theory, knowing the sequence of a protein (the order with which the amino acids occurred) one could infer its function. Four different nucleotides taken three at a time can result in 64 different possible triplet codes; more than enough to encode 20 amino acids. The way that these 64 codes are mapped onto 20 amino acids is first, that one amino acid may be encoded by 1 to 6 different triplet codes, and second, that 3 of the 64 codes, called stop codons, specify "end of peptide sequence". Where multiple codons specify the same amino acid, the different codons are used with unequal frequency and this distribution of frequency is referred to as "codon usage". Codon usage varies between species.

The fact that DNA nucleotides need to be read three at a time to specify a protein sequence implies that a DNA sequence has three different reading frames determined by whether you start at nucleotide one, two, or three. (Nucleotide four will be in the same frame as nucleotide one and so on). Both strands of DNA can be copied into RNA (for translation into protein). Thus, a DNA sequence with its (inferred) complementary strand can specify six different reading frames. It is possible to chemically determine the sequence of amino acids in a protein and of nucleotides in RNA or DNA. However, it is vastly easier at present to determine the sequence of DNA than that of RNA or protein. Since the sequence of a protein can be determined from the DNA sequence which encodes it, most protein sequences are in fact inferred from DNA sequences. Although it is possibly true in theory that given a protein sequence one can infer its properties, current state of the art in biology falls far short of being able to implement this in practice. Current sequence analysis is a painful compromise between what is desired and what is possible. Some of the many factors which make sequence analysis difficult are discussed in this section.

As noted above, the difficulty of sequencing proteins means that most protein sequences are determined from the DNA sequences encoding them. Unfortunately, the cellular pathway from DNA to RNA to Protein includes some features that complicates inference of a protein sequence from a DNA sequence. Many proteins are encoded on each piece of DNA, and, so when confronted with a DNA sequence, a biologist needs to figure out where the code for a protein starts and stops. This problem is even more difficult because the human genome contains much more DNA than is needed to encode proteins; the sequence of a random piece of DNA is likely to encode no protein whatsoever. The DNA which encodes proteins is not continuous, but rather is frequently scattered in separate blocks called exons. Many of these problems can be reduced by sequencing of RNA (via cDNA) rather than DNA itself, because the cDNA contains much less extraneous material, and because the separate exons have been joined in one continuous stretch in the RNA (cDNA). There are situations, however, where analysis of RNA is not possible and the DNA itself needs to be analyzed.

Although a much greater fraction of RNA encodes protein than does DNA, it is certainly not the case that all RNA encodes protein. In the first case, there can be RNA up- and down-stream of the coding region. These non-coding regions can be quite large, in some cases dwarfing the coding region. Further, not all RNAs encode proteins. Ribosomal RNA (rRNA), transfer RNA (tRNA), and the structural RNA of small nuclear ribonucleoproteins (snRNA) are all examples of non-coding RNA.

However, by combining a variety of computational approaches with some laboratory biology, people have been fairly successful at accomplishing this in many specific cases. Nonetheless, this problem is currently considered one of the most important in computational biology. Once you have obtained a protein sequence, inferring structure and function represent vastly greater problems. As is noted above, the structure of a protein is produced by the folding of a peptide chain back on itself, and in some cases, the association of multiple peptide chains. This folding can occur as rotation can occur around both bonds within the constituent amino acids as well as the bonds that join the amino acids one to another. Unfortunately (or fortunately, as life depends on this fact), the number of possible folding patterns is effectively infinite. To help cope with this daunting problem, biologists have divided the structural features of proteins into levels. The first level of structure, termed primary structure, refers just to the sequence of amino acids in the protein; this is what we know. Decades ago, it was found that polypeptide chains can sometimes fold into regular structures; that is, structures which are the same in shape for different polypeptides. One such shape is helical, and is referred to as an alpha helix. In another such shape, the polypeptide chain folds back and forth, producing a sheet-like surface. This structure is referred to as a beta sheet. There are additional examples of secondary structural types into which a polypeptide might fold, and some peptides do not fold into one of these regular structures at all. In fact, most long polypeptide chains (e.g. virtually all real biological proteins) fold into different secondary structures along different portions of their length.

The importance of understanding protein structure comes from two factors working together. The first of these is that the *function* of the protein is absolutely dependent on its structure. In fact, one of the most common ways for proteins to loose their function is to have their structure disrupted; for example by heat or mechanical stress (e.g. beating an egg white); only completely and properly folded proteins "work". The second factor is that it is extremely difficult to determine the structure of a protein experimentally. To date, the primary structure of many sequences has been determined (about 30,000). In contrast, the tertiary structure of many fewer (about 500) has been determined. Obviously, then, it would be of great value if tertiary structure could be determined from primary structure. It is not an exaggeration to state that the ability to exactly predict protein structures and, from that, protein function would revolutionize medicine, pharmacology, chemistry and ecology.

Current research on tertiary structure prediction has used two basic approaches; homology based and *ab initio*. Homology-based approaches attempt to determine the tertiary structure of a protein by comparing its primary sequence to that of related proteins whose structure is known. This is a laborious but fairly successful approach. Unfortunately, it requires the existence of similar protein(s) with known structure(s); something not always available. Ab initio approaches try to determine the structure which minimizes free energy. This is done using either Monte-Carlo methods or Neural Net software. Finally, even if/when you determine the tertiary structure of a protein, techniques have not yet been developed for inferring the functional properties of this protein from its structure.

Given the pessimistic view of sequence analysis presented in the previous section, why do we even bother with it? In the first place, the attempt to find methods for successful sequence analysis is a research goal in its own right, one whose potential rewards are so vast as to make it of the first importance. In the second place, although there are many things that sequence analysis *cannot* yet do, there are many very worthwhile things that can currently be done with sequence analysis, and these will be summarized in this section.

Identification of Protein Primary Sequence from DNA Sequence.

The computer programs which are used to infer protein sequence from DNA sequence provide information which can be used to help approach a solution. For example, if you are trying to find out where in a DNA sequence a protein is encoded, it is very useful to know what peptides would be encoded by all six reading frames (Fig.1). A stretch containing many stop codons is a poor candidate for encoding a protein. This will not absolutely tell you where the protein sequence starts and stops, but it will help you guess where that might occur. Programs exist for doing this. In fact, there are many factors you can use to guess where in a

DNA sequence a protein sequence might reside; use of the expected codon bias, presence of characteristic sequences representing regulatory signals in the DNA, and so forth. One family of programs integrates a variety of these approaches, and, using either explicit algorithms or trained neural nets makes a prediction.

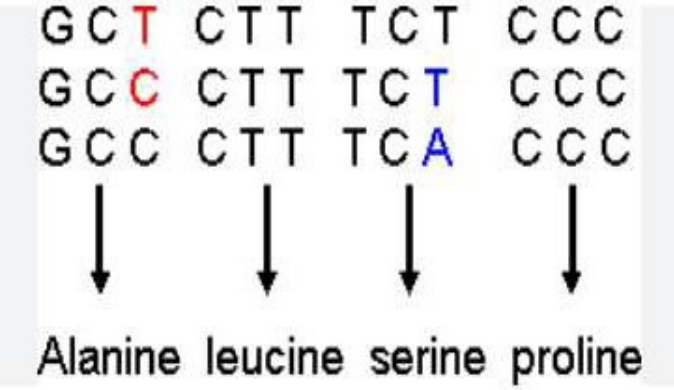

Fig.1 Nucleotide triplet codon to amino acids

Searching of databases for sequences similar to a new sequence

If you have just determined a sequence of an interesting bit of DNA, one of the first questions you are likely to ask yourself is "has anybody else seen anything like this?" Fortunately, there has been a very successful international effort to collect all the sequences people have determined in one place so they can be searched. For DNA sequences, three groups have cooperated in this effort, one in Japan, one in Europe, and one in the United States to produce DDBJ, EMBL and GenBank, respectively. These databases are frequently reconciled with each other, so that searching any one is virtually the same as searching all three. The problem is that these databases are HUGE and, as a result, you must compare your sequence with this vast number of other sequences efficiently. A number of programs have been written to rapidly search a database for a query sequence, two of which, BLAST and FASTA, will be discussed in this course. The techniques used by these programs to make searching rapid result in some loss of rigor of comparison. It is possible (although, as it turns out, unlikely) that a weak but relevant similarity could be missed by these programs. In addition, many times these programs will flag a sequence as being similar to your query sequence when this similarity is not significant. Thus, these programs should be seen as tools for identifying a small subset of sequences from the database for retrieval and further analysis rather than ends in themselves.

Calculation of sequence alignments for evolutionary inferences and to aid in structural and functional analysis

Although it is not possible to *completely* predict the function or shape (structure) of a protein from its sequence *de novo*, *some* useful inferences about structure and function can be drawn, especially by comparing the sequence of a protein of unknown structure and function to sequences of proteins with known structure and function. Second, if the goal of structure/function prediction is to be reached in the future, it will be because of partial analyses done in the present. Third, by comparing the sequence of equivalent proteins from different species of animals (such equivalent proteins are called "homologues"), one can draw inferences about the evolution of these species from their common ancestors.

One of the most useful things people do with sequences is to compare them to other sequences. However, such comparisons are not as easy to make as one might first think. One factor that complicates analysis is that the sequences biologists need to compare are usually not identical, but only similar. In addition to having a small number of substitutions (e.g. a Guanine for an Adenine at one position in a DNA sequence) there will be insertions and deletions in one sequence relative to the other. Also, depending what you are comparing and what you want to learn from the comparison, how you do the comparison will be different. For these reasons, there have been many different kinds of programs written to compare sequences.

Public Molecular Databases

It is very important to learn how to use and understand the databases that store the wealth of information that is so useful to the molecular biologist. Even it refer DNA database searching, but the principles are the same for proteins. Let's imagine that we have made a cDNA library, and that we have sequenced some of the cDNA clones. We don't know anything about these sequences, and looking at the string of letters representing nucleotides (A, C, G, T) doesn't give us many clues. We certainly cannot tell what the *in vivo* function is. Now let's consider an analogy: a word is another sequence of letters that can either mean a lot (if you know which word the letters spell), or nothing at all (if you don't know the word). If I were to present you with an English word, that you had never seen or heard before, you would be able to find out what it meant by looking it up in a **DICTIONARY**. This is what we're going to do with our molecular sequences. We are going to look up our unknown sequence with vast molecular DATAbase, and try to find out more about it. We will search for the sequence in the vast molecular databases available on the Internet.

The Main Databases

These are currently main public databases:

1. **EMBL**: at the European Molecular Biology Laboratory, Cambridge, UK.
2. **GenBank**: at NCBI, a division of NLM at the NIH campus, USA.
3. **DDBJ**: the DNA Databank of Japan.
4. **GSDB:** Genome Sequence Database. Uses Super computer with algorithms.

Each database collects and processes new sequence data and relevant biological information from scientists in their region e.g. EMBL collects from Europe, GenBank from the USA. These databases automatically update each other with the new sequences collected from each region, every 24 hours. The result is that they contain exactly the same information, except for any sequences that have been added in the last 24 hours. This is an important consideration in your choice of database. If you need accurate and up to date information (e.g. if you plan to publish), you must search an up to date database. The present manual will frequently make use of the EMBL & GenBank databases. When you are doing the assignments during the manual you may like to search all other databases. You will be able to see for yourself that the results are the same, even if reported in different fashions. You may find that you prefer the presentation of one database. You can safely continue to use your favorite database, knowing that the contents are identical (to within 24 hours).

In 1998, there were more than 1200 million base pairs from over 1.6 million sequences in the EMBL database, and the doubling size was estimated at only around one year! To organize this huge amount of information, the database has been split into numerous divisions (17 in 1998). Each entry (sequence and relevant information) is allocated to exactly one division. The division is indicated by a three letter code, specified when you receive information about a sequence.

Data coding in EMBL :

Identifier (ID)

This number is the unique identifier. It is allocated when a sequence submitted to the EMBL database is accepted for publishing. It will never change. It will be quoted in an EMBL report and may be quoted in the description line of a FASTA report.

Example: In an EMBL Report - **HS498971**

Nucleic Acid Identifier (NI)

This number is assigned to each *version* of an entry. While the identifier (EMBL) and accession number (GenBank) never changes, a new NI number will be allocated each time the sequence is modified, however minor the change might be.

Example: In an EMBL Report - **NI g2462721**

Data coding in GenBank:

Accession Number (AC)

The accession number is allocated when the record is first entered into the database, and will *never* be changed. It consists of one letter followed by 5 digits (X12345), or (more recently) two letters followed by six digits (XY123456). This number is also reported in EMBL reports.

Examples: In a GenBank Report - **ACCESSION: U49897**
In a FASTA Report - **gb|U49897**

GI Number (GI, NID)

This number is referred to as the NI number in EMBL --they are identical. The GI number tracks *versions* of an entry, and was until recently, only quoted on the NID line in a GenBank report. Database collaboration efforts have prompted a change: the GI number is now quoted in a new line called VERSION. Although now redundant, the NID line will remain for quite some time.

Examples: In a GenBank Report - **NID: g2462721**
[NID: Now redundant.]
In a GenBank Report - **VERSION:** U49897.1 **GI: 2462721**
In a FASTA Report - **gi|2462721**

Accession version

This is a new field, and the one to pay most attention to. The first number is the never changing accession number, followed by a period and a version number. The version number starts at one, and increases by one each time the sequence changes. The second number is the GI number (see example under GI Number, above).

Example: In a GenBank Report - **VERSION: U49897.1**
GI: 2462721

It is essential to remember that the EMBL ID and GenBank Accession numbers are unique. It is possible to search these databases, and others (e.g. protein databases) quoting just these unique identifiers. Other information is stored along with the sequence. Each piece of information is written on it's own line, with a code defining the line. For example, DE, description; OS, organism species; AC, accession number *etcetera*. Most are self explanatory from the content. Sometimes, a user friendly graphical report is made available, making it even easier to read the results of a search. Relevant biological information is usually described in the feature table (FT).

The Protein Sequence Databases

PIR-International Protein Sequence Database

Previously called just PIR, this is the oldest molecular sequence database available (established 1984). The entries arise from international collaborative efforts and are organized biologically e.g. by structural, functional or evolutionary relationships. The entries include amino acid sequences, and in many cases further annotation including: citations (linked to Medline for abstracts); nucleotide database references; current genetic information (including map position and the start codon if not AUG). *PIR is, in part, a redundant database*. Sequences are made public as soon as the database curators receive them, even before annotation or classification is verified. Redundancy has it's disadvantages, most notably the repetition of sequences in different entries may include discrepancies. The redundancy at PIR can be advantages, as sequences are made public very quickly. The database is updated weekly.

PIR1: Classified, annotated, verified and non-redundant with respect to other PIR1 entries.
PIR2: Essentially indistinguishable from PIR1. Classification may not be quite so extensive as in PIR1.
PIR3: Not classified, annotated or verified. No attempts have been made to reduce redundancy.
PIR4: Un-encoded or untranslated

SWISS-PROT Protein Sequence Database

SWISS-PROT (established 1986) is a protein sequence database, accessible from the Swiss EMBL Outstation, EXPASY. SWISS-PROT excels in annotation, exhibits *very little redundancy* and is thoroughly integrated with other databases (Fig.2).

The extensive annotation and exhaustive to reduce redundancy mean that entries can take time before they are made available to public. But when they are available they are complete and thorough resource. Annotation is updated with information from published review articles, and by external expert referees. The entries are similar in layout to EMBL entries, with similar two letter codes defining the contents of each line. These include *CC* (comment), *FT* (feature table) and *KW* (keywords). Annotation includes information about the protein's function, post-translational modifications, disease associated deficiency, domains, structure and more. Where applicable, SWISS-PROT entries are cross referenced with PDB, a database of experimentally determined protein structure. Three dimensional (3D) models can be viewed with most web browsers, or files can be downloaded for local viewing.

Fig.2 SWISS-PROT home page

TrEMBL

TrEMBL is a supplement to SWISS-PROT that contains computer annotated translations of EMBL. When entry annotation and verification is complete, it is moved from TrEMBL to SWISS-PROT (assuming the entry does not already exist, in which case they will be merged). Since preparing entries for SWISS-PROT is so time consuming, TrEMBL basically attempts to bridge the gap, and provide a redundant database of (less extensively) annotated translations of coding sequences (CDS) that are *not* listed in SWISS-PROT. TrEMBL has two main sections. SW-TrEMBL (SWISS-PROT TrEMBL), which contains sequences that are *en route* to SWISS-PROT. In contrast, REM-TrEMBL stores the remaining entries. This includes entries specifically excluded from SWISS-PROT, such as the many variations of immunoglobulins and T-cell receptors, synthetics sequences, fragments of less than eight amino acids, CDS from patent applications and EMBL CDS translations where the curators have strong evidence that the nucleotide does not code for real proteins.

Sequence Comparison Tools

Suppose you are a Molecular Biologist, who has discovered an unknown fragment of DNA deduced from a gel, which you have had sequenced. You will try to find out as much as you can about the sequence. Here the sequence alignments provide a powerful way to compare novel sequences with previously characterized genes. Both functional and evolutionary information can be inferred from well designed queries and alignments made using comparison tools with the data records available at various public databases.

Basic Local Alignment Search Tool – **BLAST- www.ncbi.nlm.nih.gov/BLAST**

BLAST is the algorithm used by a family of five programs that will align your query sequence against sequences in a molecular database. Statistical methods are applied to judge the significance of matches. Reported alignments (i.e. sequences in the database that could be identical to your query sequence) are reported in order of significance, as estimated by the applied statistics. The BLAST algorithm has been optimised for sequence alignment, but not for motif-searching. Listed below are the definitions for each BLAST flavour, as described by NCBI.

BLASTN: Compares a nucleotide query sequence against a nucleotide sequence database.
BLASTP: Compares an amino acid query sequence against a protein sequence database.
BLASTX: Compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
T***BLAST***N: Compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).
T***BLAST***X : Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

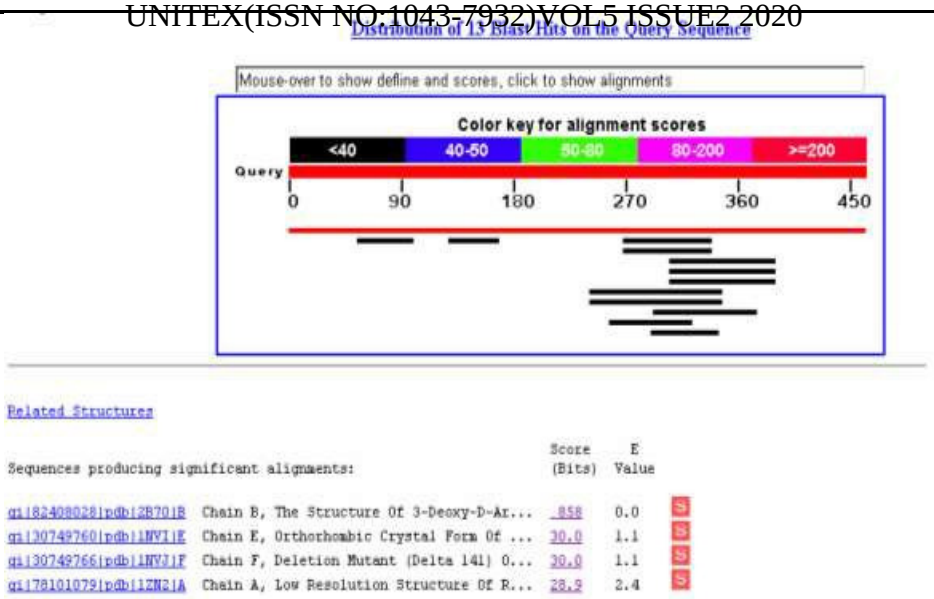

Fig.3. BLAST-N results with query sequence.

The BLAST programs are the fastest currently available. The speed was originally achieved (version 1.4) in part by forbidding gaps in the sequence. As mentioned before, gaps affect the quality of your results (Fig.3). Most servers now offer BLAST version 2.0 (BLAST2.0), which can perform gapped alignments. The search is accelerated by the heuristic nature of the statistical algorithm used: BLAST performs fast, local rather than slower, global alignments. BLAST does not try to match the whole sequence.

FASTA (pronounced FAST-Aye) stands for **FAST-A**ll, reflecting the fact that it can be used for a fast protein comparison or a fast nucleotide comparison. This program achieves a high level of sensitivity for similarity searching at high speed. This is achieved by performing optimized searches for local alignments using a substitution matrix, in this case a DNA identity matrix. The FASTA algorithm and family of programs are similar to BLAST in that they both align a query sequence against all of the sequences in a database and return the most significant matches. Whereas BLAST relies on the sum match probability for each local alignment for the sequence, FASTA scores only *exact* matches. FASTA allows gapped searches to be made. Like BLAST, FASTA is heuristic, sacrificing some speed for sensitivity. FASTA comes in several flavours, and you should choose the most appropriate program when searching.

fasta3: A DNA query sequence is aligned against a DNA sequence database. A protein query sequence will be aligned against a protein database.

tfasta3: Align a protein query sequence against a DNA sequence database, translating the DNA sequences 'on-the-fly'.

***fastx3*:** Align a DNA query sequence against a protein sequence database, comparing the translated DNA sequence in three frames.

***tfastx3*:** Align a protein sequence to a DNA sequence database. Align with forward and reverse frame shift mutations.

Clustal W: Clustal W is a general purpose global multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. The basic information they provide is identification of conserved sequence regions (Fig.4). This is very useful in designing experiments to test and modify the function of specific proteins, in predicting the function and structure of proteins, and in identifying new members of protein families. Sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment). This is true for pairwise and multiple alignments.

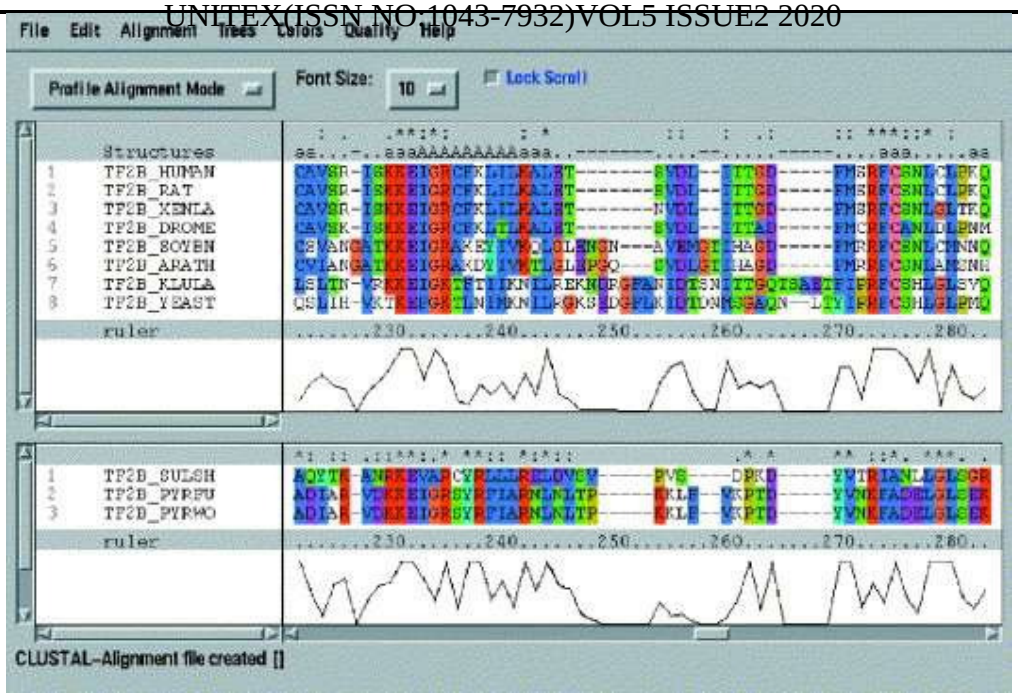

Fig.4. Model Alignment of sequences in multiple way

Global alignments need to use gaps (representing insertions/deletions) while local alignments can avoid them, aligning regions between gaps. ClustalW is a fully automatic program for global multiple alignment of DNA and protein sequences. The alignment is progressive and considers the sequence redundancy. Trees can also be calculated from multiple alignments. The program has some adjustable parameters with reasonable defaults.

Clustal X : Clustal X is a new windows interface for the ClustalW multiple sequence alignment program. It provides an integrated environment for performing multiple sequence and profile alignments and analysing the results. The sequence alignment is displayed in a window on the screen. A versatile coloring scheme has been incorporated allowing you to highlight conserved features in the alignment. All sequences must be in 1 file, one after another. 7 formats are automatically recognised: NBRF/PIR, EMBL/SWISSPROT, Pearson (Fasta), Evolutionary relationships can be seen via viewing Cladograms or Phylograms. Multiple alignments of protein sequences are important tools in studying sequences (Fig.5).

Clustal (*.aln), GCG/MSF (Pileup), GCG9 RSF and GDE flat file.

The line above the ruler is used to mark strongly conserved positions. Three characters ('*', ':' and '.') are used:
'*' indicates positions which have a single, fully conserved residue
':' indicates that one of the following 'strong' groups is fully conserved:-
'.' indicates that one of the following 'weaker' groups is fully conserved.

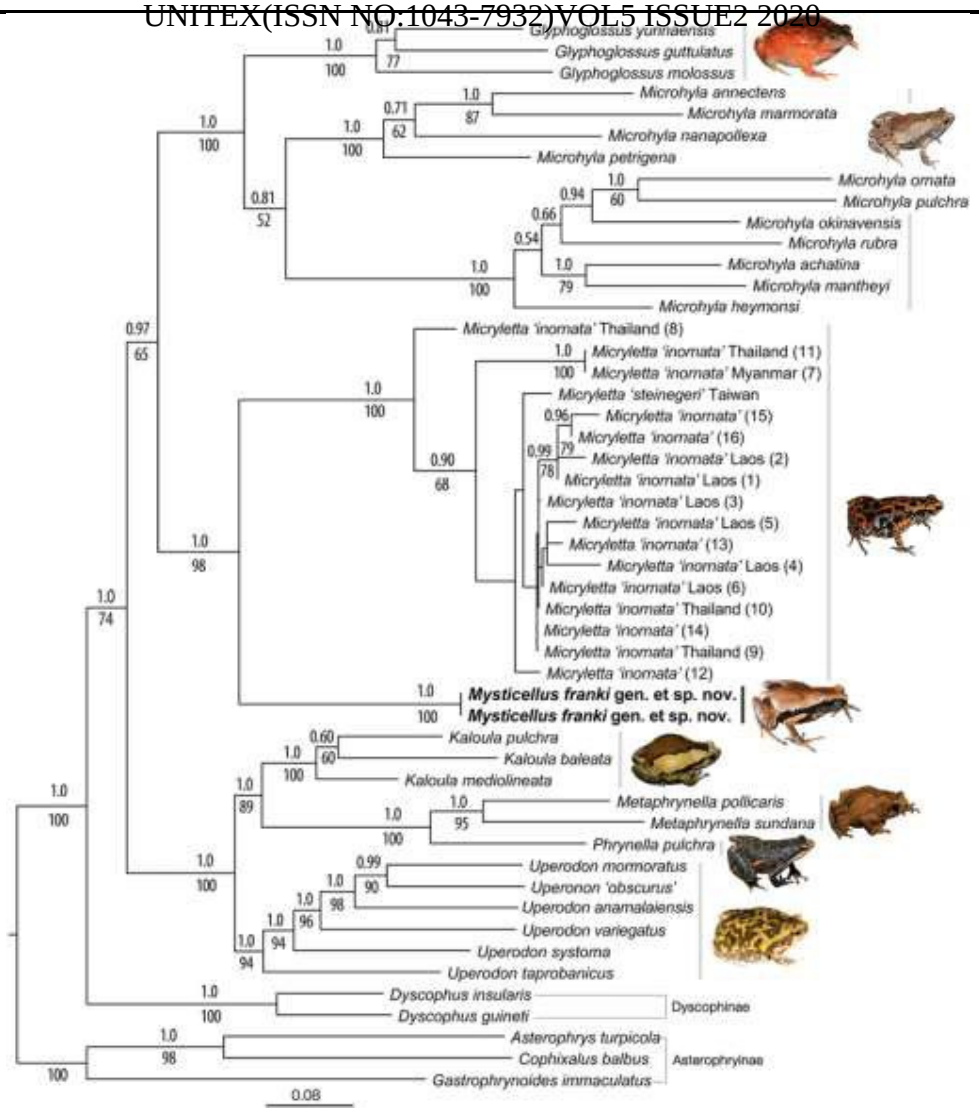

Fig.5. Phylogram of m multiple sequence alignment

References:

Academic Press Dictionary of Science and Technology http://www.harcourt.com/dictionary/ A nice online dictionary. Easy to use, comprehensive.

ALIGN 5-1-1-DUPUWE/France http://www2.igh.cnrs.fr/bin/align-guess.cgi or http://genome.eerie.fr/fasta/align-query.html Applies the BLOSUM50 matrix to deduce the optimal alignment between two sequences.

Analysis and Annotation Tool for Finding Genes in Genomic Sequences 5-3-2-DcDPDFWE/USA

Australian National Genomic Information Service (ANGIS) http://morgan.angis.su.oz.au

BioComputing Hypertext Coursebook http://www.techfak.uni-bielefeld.de/bcd/Curric/ A very thorough source of learning. Especially good for readers who want to get to grips with the "nuts and bolts" of bioinformatics**Bioinformatics and Biology Resources on the Internet** http://aeiveos.wa.com/biology/index.html An excellent site, well worth a visit!

Berkeley Fly Database 5-5-2-DDW (UI)/USA http://www.fruitfly.org/bfd/ Search for sequences by name or map location, and optionally view a clickable image of sequenced contigs aligned alongside the fly chromosomes. Searches can be limited to available sequences only. Retrieve P1, BAC or cosmid genomic clones, P element insertion lines, YAC, STS and more.

Berkeley *Drosophila* **Genome Project BLAST Searches 5-4-3-DDFWE/USA** http://www.fruitfly.org/blast/ Search for your sequence using the WU-BLAST 2.0 algorithm for *D. melanogaster* sequence data, including EST's, genomic sequences, STS's or sequences derived from them, P element insertion sites and transposons.

Blitz http://www.ebi.ac.uk/searches/blitz.html

Beauty http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html Beauty is an enhanced BLAST search, returning output which predicts the function of the protein **Brunel University Online Teaching Programme** http://www.brunel.ac.uk/depts/bl/project/front.htm

BLAST 2 Similarity Search (EMBNet) 5-3-2-PUWE/Switzerland http://www.ch.embnet.org/software/frameBLAST.html WU-BLAST 2.0 similarity searches.

BLAST2 Search with Post-processing (EMBL) 4-3-2-PUW/Germany http://dove.embl-heidelberg.de/Blast2/ WU-BLAST 2.0 search with post-processing.

Center for Biotechnology (NCBI) of The National Library of Medicine (NLM) at The National Institutes for Health (NIH) campus, USA. **Colour INteractive Editor for Multiple Alignments (CINEMA)** http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA2.1/ A comprehensive and popular site. Allows the user to visualise and manipulate aligned protein sequences. *Makes use of Java. I recommend you access this site from a fast workstation!*

DNA Databank of Japan (DDBJ) http://www.ddbj.nig.ac.jp

European Drosophila Genome Project http://edgp.ebi.ac.uk/

European Molecular Biology Laboratory (EMBL) http://www2.ebi.ac.uk/Help/General/general.html Cambridge, UK.

European *Drosophila* Genome Project BLAST server 5-4-3-DDPDFWE/UK http://edgp.ebi.ac.uk/www-blast.html Search using the WU-BLAST 2.0 (gapped aligned) or the original BLAST (which does not allow gaps). The database includes *Drosophila* genomic data, EST's, STS's, P element sites, transposons, repeats, and proteins.

ExPASY (Swiss Institute of Bioinformatics) http://www.expasy.ch/ ExPASY is not the Swiss EMBnet node!

Entrez http://www.ncbi.nlm.nih.gov/Entrez/ Start here for pretty much anything!

Forward and Reverse Translation 5-5-2-DUPUFWE/UK http://www.sanger.ac.uk/Software/Wise2/genewiseform.shtml Translate a protein sequence into a genomic DNA sequence, and vice versa. *This is a WWW interface to the pgwise software application. Those who are proficient with this package may like to take advantage of it's extra capabilities by adding execution critera.*

FASTA 3 (EMBL) http://www2.ebi.ac.uk/fasta3/ FASTA 3 similarity search.

FASTA http://www2.igh.cnrs.fr/bin/fasta-guess.cgi FASTA similarity search and a clean, basic and simple interface.

being tested.

http://genome.cs.mtu.edu/aat.html Identifies genes in a DNA sequence by comparing it to cDNA and protein sequence databases (including those at HGI, TIGR, dbEST, Swiss-Prot and nr).

GenBank http://www.ncbi.nlm.nih.gov/Genbank/ GenBank at the National Center for Biotechnology (NCBI) of The National Library of Medicine (NLM) at The National Institutes for Health (NIH) campus, USA.

Gene Cards(*) http://bioinfo.weizmann.ac.il/cards/ A very useful site providing comprehensive information and links. Direct links to GenBank, SWISS-PROT and MedLine. Includes synonyms, similar genes in other organisms, gene products and details about disorders.

Gene Cards: Diseases with a Genetic Association (*) http://bioinfo.weizmann.ac.il/cards-bin/listdiseasecards http://bioinfo.weizmann.ac.il/cards/cards-bin/listdiseasecards View the complete (comprehensive**Genome Sequence DataBase (GSDB)** http://seqsim.ncgr.org/ The National Center for Genome Resources, Genome Sequence Database. The server is a supercomputer with genomic algorithm accelleration.

GenomeNet (Japan) http://www.genome.ad.jp/

Genome Sequence DataBase (GSDB) http://seqsim.ncgr.org/ The National Center for Genome Resources, Genome Sequence Database. The server is a supercomputer with **Online Mendelian Inheritance in Man (OMIM)** http://www3.ncbi.nlm.nih.gov/Omim/ Database of human genes and their disorders, with textual information, images and references. Links to Entrez and MedLine.

Genome Database (GDB)http://www.hgmp.mrc.ac.uk/gdb or http://gdbwww.gdb.org/ *Funding for this project has been withdrawn. This valuable database will remain online, but it should be noted that no new entries will be recorded after 31st July 1998.*

GENATLAS (*) **5-5-2-DUWI/France** http://bisance.citi2.fr/GENATLAS/ A comprehensive, easy to use site. Search gene, marker or phenotype or linkage databases. Useful, relevant links provided in the results. The user can also locate the desired gene from a graphical clickable map of disease related or other mapped genes on a chromosome.

Gene Cards(*) http://bioinfo.weizmann.ac.il/cards/ A very useful site providing comprehensive information and links. Direct links to GenBank, SWISS-PROT and MedLine. Includes synonyms, similar genes in other organisms, gene products and details about disorders.

Human Genome Map Database (HuGeMap) http://www.infobiogen.fr/services/Hugemap Genetic and physical maps of the human genome. Connected with the gene radiation hybrid mapping database RHdb.

HGMP-RC Primers Database http://www.hgmp.mrc.ac.uk/local-data/Primers.html

List of other Genome Sites http://www.hgmp.mrc.ac.uk/GenomeWeb/ List of other genome web sites. Concise and clearly presented.

Multiple Sequence Alignemnt with MAP 4-3-2-DcUPUFWE/USA http://genome.cs.mtu.edu/map/map.html Calculates the global alignment of DNA or protein sequences using an algorithm which computes the best overlapping alignment without penalising terminal gaps. Long internal gaps in short sequences are not penalised.

Multiple Sequence Alignment with Hierarchial Clustering 5-5-3-PUW/France http://www.toulouse.inra.fr/multalin.html Sequence alignment with a colour output where differing or similar amino acids in the alignment can be highlighted.

Molecular Probe Data Base (MPDB or MOLPROBE) http://www.biotech.ist.unige.it/interlab/mpdb.html

Network Protein Sequence Analysis 5-4-1-PUW/France VSNS BioComputing Division Multiple Alignment Resource Page http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/ An excellent, comprehensive resource for multiple sequence alignment, software and tutorials.

Nucleotide to Protein (ExPASY) 5-2-1-DUW/Switzerland http://www.expasy.ch/tools/dna.html Translates a nucleotide sequence (DNA/RNA) into a protein sequence (amino acids).

Nucleotide to Protein (EMBL) 5-4-1-DUW/UK http://www.ebi.ac.uk/contrib/tommaso/translate.html Translates a nucleotide sequence into a protein sequence.

ORF Finder http://www.ncbi.nlm.nih.gov/gorf/gorf.html Finds likely open reading frames in a sequence.

Primer 3 5-5-2-DUW/Norway http://www2.no.embnet.uio.org/primer/primer3.cgi? Select PCR primers for your nucleotide sequence **Protein and cDNA Translation 5-4-2-DcUPU/UK** http://www.sanger.ac.uk/Software/Wise2/protein2cdna.shtml

Protein Colourer 5-2-1-PUW/UK http://www.ebi.ac.uk/htbin/visprot.pl Colour a protein sequence (raw text or SWISS-PROT Acc. No.) by properties e.g. hydrophobicity.

.**Pairwise Sequence Alignment 3-4-2-DcUPUFWE/USA** http://genome.cs.mtu.edu/align/align.html Computes the global alignment between two sequences. Compare DNA with DNA, cDNA or protein. For DNA and cDNA, settings (gap open penalty, gap extension etc.) can be defined.

Random Protein Sequence Generator 5-5-2-PUW/Switzerland http://www.expasy.ch/sprot/randseq.html Random protein sequence generator! Output in FASTA format (the format most commonly required by bioinformatics search sites).

Residue Periodicity Review this site http://o2.dbuoa.gr/FT/ Study the periodicity of residues in a protein sequence.

Sequence Retrieval System (SRS) http://srs.hgmp.mrc.ac.uk/

SeqNet: UK Node of European Molecular Biology Network (EMBNet) http://www.seqnet.dl.ac.uk/About/SEQNET/

TIGR HGI Gene Expression Data 4-2-1-DDcW/USA http://www.tigr.org/tdb/hgi/searching/hgi_xpress_search.html Search for tissue specific transcripts e.g. "lung". cDNA libraries can also be searched.

The Institute for Genomic Research (TIGR) http://www.tigr.org/

The Sanger Centre http://www.sanger.ac.uk/

The Sanger Centre Database Search Services 5-2-5-DDPDFWE/UK -Clean, simple design. http://www.sanger.ac.uk/DataSearch/ BLAST and WU-BLAST 2.0 searches which can be refined to finished and/or unfinished genomic sequences.

The Bio-Web (**) http://www.cellbiol.com/ Resources for Molecular and Cell Biologists. Superb site; well maintained. Links, news etc etc. Very comprehensive. genomic algorithm accelleration.

The International Immunogenetics Database (IMGT) http://imgt.cnusc.fr:8104 Contains expertly annotated sequences and alignment tables for Ig, TCR and MHC sequences.

The Institute of Genomic Research Databases http://www.tigr.org/tdb/tdb.html Many databases including microbial, parasites, human, human cDNA, mouse, rat, *Arabidopsis*, zebrafish, and others.

UK Human Genome Mapping Project - Resource Center (HGMP-RC) http://www.hgmp.mrc.ac.uk/

UTR Home Page http://bigarea.area.ba.cnr.it:8000/EmbIT/UTRHome/ Internet resources for sequence analysis of 5' and 3' untranslated regions of eukaryotic mRNAs. Includes specialised UTR databases and tools for analyses of UTR regions.

Via OMIM You can search for a 'disease gene' at OMIM. Click the "DNA" button in the results display and follow the link to the mRNA sequence. Note the absence of U (uracil): this sequence is referred to in GenBank reports as mRNA, but the sequence is a cDNA sequence. http://www.nih.gov

Web Cutter - Restriction Enzyme Mapping Utility http://rna.lundberg.gu.se/cutter2/ Map restriction enzyme sites on your sequence. Easy to use and comprehensive options.