An Introduction to Bioinformatics for Biological Sciences Students

Department of Microbiology and Immunology, McGill University

Version 2.5 (For the BIOC-300 lab), March 2006

Contributors

The first edition of the Introduction to Bioinformatics for biological sciences students was written during the summer of 2004 at McGill University for the Bioinformatics Project (BIP) as part of the U2 undergraduate laboratory in Microbiology and Immunology (MIMM-386).

What you are holding in your hands is the second edition of the manual put together by a new group of students during the summer of 2005. From the first edition, it contains only the section on biological databases and the main institutes that develop and maintain them (other parts were included in an "extended" version of the manual instead). The biggest change from the first edition is that the manual now includes the exercise sheets and tutorial written during the course of the BIP's first year of existence, making this volume the comprehensive resource students need to understand the material covered in the BIP, but also to perform the exercises.

This version has been adapted for the BIOC-300D Laboratory in Biochemistry course lab on bioinformatics.

Main contributors to the abridged version

- Cédric Sam (cedric.sam@elf.mcgill.ca)
- Oksana Kapoustina
- Abrar Khan

Contributors to all sections including those not in the abridged version

- Belinda Befort (PROSITE, Phylip)
- Scott Bunnell (Editing)
- Mansoureh Hakimi (Exercises review)
- Oksana Kapoustina (BLAST, ClustalW)
- Abrar Khan (Editing)
- François Pepin (Editing)
- Cédric Sam (Institutes, Databases, Editing, Exercises, original layout)
- Sean Wiltshire (Introduction, editing)

Faculty members who contributed to this manual

- Dr Silvia Vidal (silvia.vidal@mcgill.ca)
- Dr Nicholas Acheson
- Dr Malcolm Baines

Copyright © 2004, 2005 Department of Microbiology and Immunology of McGill University. All rights reserved.

Table of Contents

Chapter 1: Bioinformatics Institutes

This section will be an overview of the major actors in the field of bioinformatics, what are the services they offer and what sort of databases they each manage. These research institutes were all established in different countries, but their reach, their funding sources, their staff are now well worldwide.

1.1 NCBI: The National Center for Biotechnology Information (USA)

The NCBI is a unit of the National Library of Medicine (NLM), which is in turn a branch of the National Institutes of Health (NIH). The NCBI is located in Bethesda, MD, in the outskirts of Washington DC.

1.1.1 Database resources at the NCBI

Here's an overview of a few of the databases hosted by NCBI and the services which come with them[1].

- Database Retrieval Tools - **Entrez** is an integrated retrieval system for the databases hosted by NCBI. **Taxonomy** indexes over 150 000 organisms that are represented by at least one nucleotide or protein sequence. **LocusLink** provides a single query interface to curated sequence and descriptive information about genetic loci.
- BLAST family of sequence-similarity search programs.
- Resource for gene-level sequences - **UniGene** is a system which partitions GenBank sequences into non-redundant set of gene-oriented clusters. There are many other specialized databases for single nucleotide polymorphisms (**dbSNP)**, and for information on Major Histocompatibility Complex (**dbMHC**).
- Resources for genome-scale analysis - **Entrez Genomes** provides access to genomic data and includes genomes spanning from microbes to multicellular organisms.
- Eukaryotic Genomic Resources - **Map Viewer** displays genome assemblies using sets of aligned chromosomal maps.
- Structural databases - The NBCI **MMDB** is built by processing entries from the Protein Data Bank.

1.1.2 PubMed: The ultimate biomedical literature database

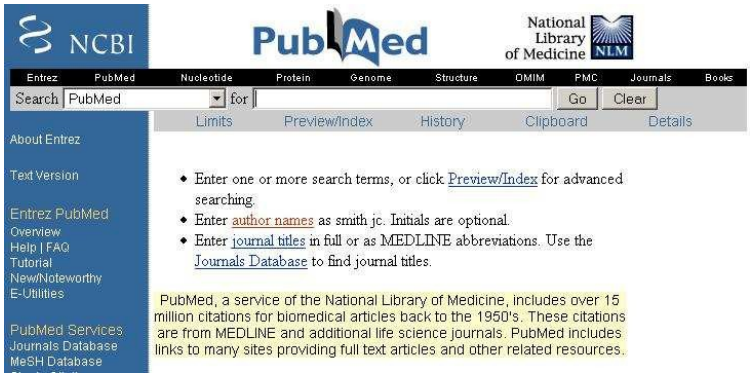
MEDLINE is the NLM's premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. MEDLINE contains bibliographic citations and author abstracts from more than 4,800 biomedical journals published in the United States and 70 other countries. The database contains over 12 million citations dating back to the mid-1960. Coverage is worldwide, but most records are from English-language sources or have English abstracts.

PubMed is NCBI's biomedical literature database giving access to citations compiled in databases such as MEDLINE. To the average user, Pubmed just equals Medline, although a website describes the difference between both concepts: (http://www.nlm.nih.gov/pubs/factsheets/dif med pub.html).

.What you need to know is that PubMed is a biomedical literature giving access to the MEDLINE database, but also to certain non-medical article featured in MEDLINE journals. What you read in a textbook today has almost always been

1 Source: Database resources of the National Center for Biotechnology Information, Nucleic Acids Res.2004 Jan 1; 32 Database issue: D35-40

published and debated through peer-reviewed journals. Reading "review" articles in prominent journals like **Science** or **Nature** is a good way to start familiarizing yourself with peer-reviewed journals.



NCBI'S PUBMED - HTTP://WWW.NCBI.NIH.GOV/PUBMED

1.2 EBI: The European Bioinformatics Institute

"The European Bioinformatics Institute (EBI) is a non-profit academic organization that forms part of the European Molecular Biology Laboratory (EMBL)." (EBI website, http://www.ebi.ac.uk/) The EBI is located in Cambridgeshire, United Kingdom, and was established in 1992. It is the European equivalent of the NCBI. In 2004, EBI was funded primarily by the EMBL (45%) and the European Union (25%), but also by the National Institutes of Health (NIH) in the USA (accounting for about 10%). Many applications are available from EBI through a web interface. Here are some examples[2]:

- Homology & Similarity - the **BLAST** or **Fasta** programs can be used to look for sequence similarity. (Note: The BLAST provided by EBI is different from the one provided by NCBI (it's "WU-BLAST", by Washington U in St.Louis, rather than "NCBI-BLAST").
- Protein Functional Analysis **- InterProScan** can be used to search for motifs in your protein sequence.
- Sequence Analysis - **ClustalW** a sequence alignment tool.
- Structural Analysis **- MSDfold** or **DALI** can be used to query your protein structure and compare it to those in the Protein Data Bank (PDB).
- Tools Miscellaneous - **Expression Profiler** a set of tools for clustering, analysis and visualization of gene expression and other genomic data.

As well as applications, the following are databases maintained by EBI[3]:

- **EMBL Nucleotide Database** - Europe's primary collection of nucleotide sequences is maintained in collaboration with Genbank (USA) and DDBJ (Japan). (Note: These are the three partners of The International Nucleotide Sequence Database Collaboration (INSD). See Science, Brunak et al 298 (5597):1333)
- **UniProt Knowledgebase** - a complete annotated protein sequence database. It is a central repository of protein sequence and function created in 2002 by joining the information contained in Swiss- Prot/TrEMBL (Switzerland-Europe), and PIR (USA). See Curr Opin Chem Biol 2004 Feb 8(1):76-80, a recent article on Uniprot and Protein sequence databases at large.
- **Macromolecular Structure Database** - European Project for the management and distribution of data on macromolecular structures.
- **ArrayExpress** - for gene expression data.

2 (Source: EBI Services - http://www.ebi.ac.uk/services/)

3 (Source: EBI Databases - http://www.ebi.ac.uk/databases/)

- **Ensembl** - Providing up to date completed metazoic genomes and the best possible automatic annotation.

THE EUROPEAN BIOINFORMATICS INSTITUTE WEBSITE - HTTP://WWW.EBI.AC.UK/

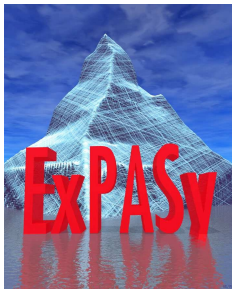
1.3 SIB: The Swiss Institute of Bioinformatics

"The SIB is an academic not-for-profit foundation established on March 30, 1998 whose mission is to promote research, the development of databanks and computer technologies, teaching and service activities in the field of bioinformatics, in Switzerland with international collaborations." (SIB website, http://www.isb-sib.ch/)

The SIB maintains a number of important databases such as the **Swiss-Prot/TrEMBL** protein databases, the **PROSITE** protein families and domains database and the **SWISS-2DPAGE** database of 2D-gels, plus many other specialized databases.

The SIB is also active in developing software tools like **Melanie** for the analysis of 2-D gels, **Swiss- PdbViewer** for the visualization of 3-D structures (such as those found in the Protein Data Bank, or PDB, database), and **SWISS-MODEL**, a fully-automated server which takes protein sequences and tries to model their 3-D structure from known 3-D structures found in the PDB.

1.3.1 How to access SIB's resources?



EXPASY: THE SWISS PROTEOMICS SERVER - HTTP://CA.EXPASY.ORG/ (CANADIAN MIRROR)

ExPASy (Expert Protein Analysis System, http://ca.expasy.org/) is the SIB's proteomics web server. ExPASy is the website to use to access to all of the aforementioned SIB databases and analytical tools (and Swiss-Jokes http://www.expasy.org/cgi-bin/sw-jokes.pl).

1.4 Bioinformatics in Canada

The website of the Canadian Bioinformatics Resource in Ottawa hosts well-known bioinformatics applications, such as BLAST, ClustalW and a web version of the popular phylogenetics program Phylip. (http://cbr-rbc.nrc-cnrc.gc.ca/)

Chapter 2: Molecular Biology Databases

Databases are large collections of data arranged for ease of search and retrieval. Common databases such as GenBank, PDB or Swiss-Prot exist as (extremely) large files which can be downloaded from public sites for various data manipulations on local private computers, or more practically, consulted on-line by molecular biologists at large using search tools such as BLAST.

2.1 Introduction

This section will cover nucleotide sequence, protein structure and protein sequence databases. Some of the main databases are found below[4]:

- **Biomedical literature**: PubMed.
- **Species-specific:** SGD, FlyBase, WormBase, MGI.
- **Nucleotide sequences:** GenBank, EMBL, DDBJ.
- **Genome sequences:** Entrez Genome, TIGR databases.
- **Protein sequences**: GenPept, Swiss-Prot/TrEMBL, PIR.
- **Macromolecular 3-D Structures:** Protein Data Bank, MMDB.
- **Protein and peptide mass spectroscopy:** PROWL.
- **Post-translational modifications:** RESID.
- **Biochemical and biophysical information:** ENZYME, BIND.
- **Biochemical pathways:** PathDB, KEGG, WIT.
- **Microarray chips data:** ArrayExpress, SMD.
- **2D-PAGE:** SWISS-2DPAGE.
- **Protein families and domains:** PROSITE, Pfam, InterPro, ProDom.

2.2 Nucleotide Sequence Databases

Nucleotide sequences (DNA and RNA) are essential pieces of information. Researchers might use protein- coding nucleotide sequences to produce large quantities of protein for various experiments in the wet lab.

4 This listing is vaguely based on the one found in “Developing bioinformatics computer skills” by Cynthia Gibas and Per Jambeck, O’Reilly & Associates, 2001.

2.2.1 The Big Three: GenBank, DDBJ and EMBL

NCBI's GenBank (USA), EBI's EMBL Nucleotide Sequence Database (Europe), and the DNA Databank of Japan (DDBJ, Japan) are the three biggest nucleotide sequence databases in the world. Their main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

The NCBI hosts the most well-known database, GenBank. As a result of the International Nucleotide Sequence Database (INSD) Collaboration between the NCBI, EMBL and DDBJ, new submissions are shared between databases, leading them to have similar content (although the annotations can differ). This collaboration between the three institutes has existed for 16 years[5].

2.2.2 Entrez: NCBI's multi-purpose search engine

Entrez can be used to search any of the NCBI-hosted databases. Pubmed is one of NCBI's databases; it is the scientific publications database.

The NCBI website is not easy to navigate and takes a lot of fooling-around before one can safely sail from place to place[6]. You can use Entrez directly from NCBI's homepage http://www.ncbi.nih.gov/, but you will be missing out on many of the search options. If you want to search PubMed or another database, just click on the upper bar link, with Entrez being the cross-database search (useful when you want all the information about a specific gene).

Refine your search

Some parameters can be used to refine your search. In general, you might want to start by limiting your searches. So these options are only accessible through each specific Entrez flavor. Using the "Limits" link at the bottom of the Entrez search bar/box, you specify many parameters, such as the fields you want to limit your search to. There are also limits specific to the type of Entrez you are using: for example, with Entrez Nucleotide (GenBank), you can decide the type of nucleotide (genomic DNA/RNA, mRNA or rRNA), date of modification, the subset of GenBank it belongs to, etc. In Entrez PubMed, you can conveniently specify a range for the date of publication or choose the publication type, among other things.

NCBI UniGene website

To search Genbank, use Entrez: http://www.ncbi.nih.gov/Entrez/ (case-sensitive).

2.2.3 NCBI's UniGene

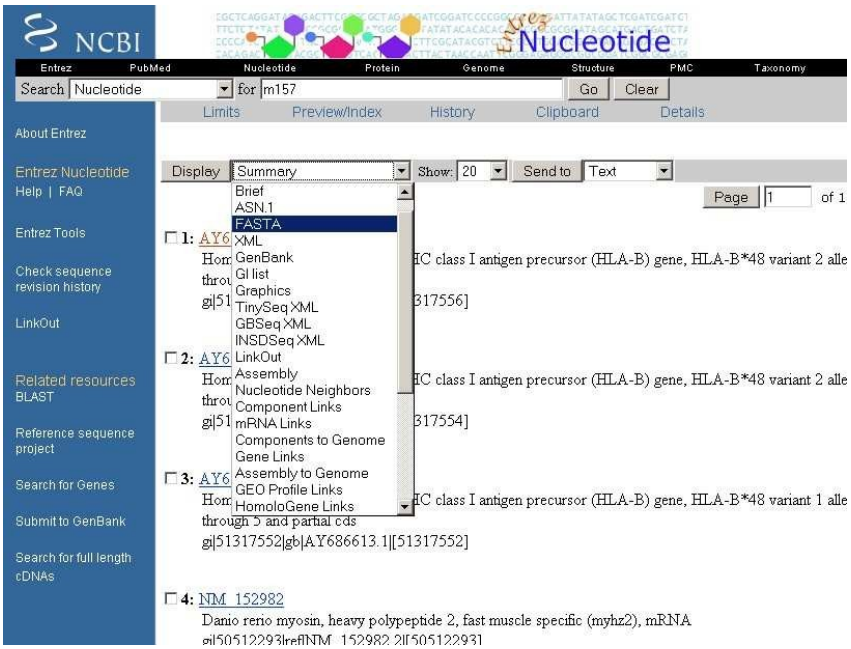
"Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location." (NCBI website)

UniGene is not a database per se, it is rather a system for "automatically partitioning GenBank sequences, including expressed sequence tags (ESTs), into a non-redundant set of gene-oriented clusters". (Wheeler DL, et al. Database Resources of the National Center for Biotechnology. Nucl Acids Res 31:28-33;2003)

The importance of UniGene is its role in organizing the numerous sequences contained in public databases that can relate to a single gene. Means of organizing the information have been overwhelmed by the deluge of sequences coming from the various genome projects, and UniGene is an effort to group nucleotide sequences (ESTs and mRNA) of selected organisms by genes they are related to.

5 See: Nucleotide Sequence Database Policies, Science 298 (5597): 1333 15 Nov 2002

6 The Entrez Help Document is a useful resource (http://web.ncbi.nih.gov/entrez/query/static/help/helpdoc.html)



By default, the results of an Entrez search in GenBank are formatted for the "Sequence" view. A pull-down menu allows you to select the view fitting your needs (such as FASTA, one of the popular formats accepted by sequence analysis programs).

dbEST and UniGene

Complementary DNA (cDNA) is single-stranded DNA synthesized from a mature mRNA template. Now, what are "Expressed Sequence Tags" (ESTs)? They are short sequences generated by sequencing the ends of these cDNA molecules.

ESTs are important gene mapping and discovery tools because they can be used as primers to amplify genomic DNA spanning a region presumably bounded on one side by the EST. The EST database (dbEST) is one of the many divisions of GenBank, the NCBI nucleotide database. As of September 2004, there were some 5.7 million Homo sapiens ESTs in dbEST.

By design, ESTs in dbEST may be redundant, as several different ESTs might be derived from mRNA expressed by the same gene. This is where UniGene comes into play.

What does UniGene contain exactly?

UniGene regroups ESTs, mRNAs, high-throughput cDNAs (HTC), etc., representing a unique gene into clusters. UniGene is an automated system, and has so far reduced 4.6 million Homo sapiens sequences to some 107,014 gene clusters. Clusters are never stable, they can be merged together at any point based on certain criteria as new sequences are added to GenBank/UniGene.

Every cluster has its own webpage through UniGene's web interface. From that page, related information about the cluster can be found: tissue types in which the gene is expressed, protein similarities with clusters in model organisms (say if you wanted to express a human gene in mice using the murine counterpart), LocusLink report for the gene and its location in the genome.

How do you search UniGene?

Each species has its own summary page with all kinds of nice statistics about the number and size of the clusters as sequences get added to the system by the thousand every week.

Typically, you might know the name of the gene, or could be looking for the cluster to which your nucleotide sequence belongs to. UniGene is the database, and Entrez is NCBI's all-purpose search engine, so Entrez UniGene is naturally the way to go.

NCBI UniGene website

http://www.ncbi.nih.gov/UniGene

2.3 Protein Sequence Databases

Protein databases are the natural extension of nucleotide sequences. They come in two varieties: “Sequence Repositories” and “Universal Curated Databases”. Sequence repositories are generally just places where protein sequences are compiled with minimal attention given to provide non-redundant entries. GenBank is a well-known example of sequence repository. In contrast, universal curated databases are manually organized and looked after by experts.

Among protein databases, NCBI’s GenPept is an example of a sequence repository. It contains the translation of nucleotide sequences contained in the GenBank-EMBL-DDBJ triumvirat mentioned earlier. Curated databases contain information validated by expert biologists and thus considered highly reliable. Swiss-Prot/TrEMBL and PIR are examples of such databases, and we will look more closely at their history and modes of functioning before talking about UniProt, an effort by Apweiler, Bairoch and Wu’s groups to establish networks for sharing protein information around the world.

2.3.1 What can you find in a curated protein database?

Information on each protein is very specific and presumably highly reliable. Anything essential such as the accession number, the source organism and (sometimes very many) references can be found. Cross- references with other databases can be useful for researchers who might be interested to learn more about a given structural domain contained in the protein, or the related nucleotide sequences if one wishes to express the protein for various assays.

2.3.2 Swiss-Prot and TrEMBL

“Swiss-Prot is a curated biological database of protein sequences created in 1986 by Amos Bairoch during his PhD and developed by the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. It strives to provide a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. As of July 5, 2004, Swiss-Prot (release 44.0) contains 153,871 entries.” (Wikipedia)

THE SWISS-PROT LOGO.

Each entry of Swiss-Prot (http://ca.expasy.org/sprot/) is carefully inspected by specialists from around the world to ensure a high quality of the information contained. This is a long process, and more and more sequences are added to the database every day. That’s where TrEMBL comes to the rescue.

TrEMBL: Translation of EMBL nucleotide sequence database

TrEMBL is a computer-annotated supplement to Swiss-Prot introduced in 1996 as a solution to preserve the high editorial standards of Swiss-Prot while making new sequences available to the public.

TrEMBL contains translations of all coding regions in the DDBJ-EMBL-GenBank nucleotide databases, and protein sequences extracted from the literature or submitted to UniProt, which are not yet integrated into Swiss-Prot. TrEMBL allows these sequences to made publicly available quickly without diluting the high quality annotation found in Swiss-Prot.

Searching Swiss-Prot/TrEMBL

In an effort to create a single source of protein information, the UniProt consortium was established. Searching and using Swiss-Prot/TrEMBL is similar to searching and using the UniProt databases, so this section will actually be covered below.

2.3.3 PIR-PSD: The Protein Sequence Database

"The Protein Information Resource (PIR), located at Georgetown University Medical Center, is an integrated public bioinformatics resource that supports genomic and proteomic research and scientific studies." (PIR website, http://pir.georgetown.edu/) The PIR maintains the Protein Sequence Database (PSD), an annotated protein database similar to Swiss-Prot. The PSD grew out of the Atlas of Protein Sequence and Structure (1965-1978) edited by the late Margaret Dayhoff[7].

2.3.4 UniProt

Nucleotide sequence databases were united under the International Nucleotide Sequence Database (INSD) Collaboration, but curated protein databases didn't have their equivalent body until 2002, when the UniProt consortium was established between the developers of the main existing annotated protein databases: the EBI/SIB (Swiss-Prot & TrEMBL) and the PIR (Protein Sequence Database (PSD)).

UniProt is a very recent addition that aims to replicate the efforts of UniGene in the amino acid sequence world. The first version (1.0) of UniProt was officially launched 15-Dec-2003, and its second version (2.0) on 5-Jul-2004. Both of these were in fact the most current versions of Swiss-Prot/TrEMBL and PSD merged together.

Databases making up UniProt are:

- The **UniProt Archive (UniParc)** is the most comprehensive publicly accessible non-redundant protein sequence database available. It includes sequences from databases hosted by the founding members of UniProt but also sequences derived from other public databases such as PDB, RefSeq or EMBL. As its name implies, UniParc is an archive, so every time a change is made to an entry on the native database, UniParc takes note of it, updates the sequence and keeps the old version.
- Initially, the **UniProt Knowledgebase (UniProt)** consists of the merging of the Swiss-Prot, TrEMBL and PSD entries, but will later be derived from the UniParc database[8]. UniProt will retain the organization of the Swiss-Prot/TrEMBL duo (Swiss-Prot as a manually-curated database and TrEMBL as a computer-annotated database) and integrate data from PIR-PSD that's not already in Swiss- Prot/TrEMBL.
- The **UniProt Non-redundant Reference (UniRef)** is, as its name implies, a collection of non- redundant protein sequences. UniRef sequences come from the UniProt knowledgebase, and the non- redundancy is generated automatically. Sequences are compared with each other and if there is sequence homology, they are merged together and added as a single entry in UniRef.

Searching UniProt

The search interface is slightly different among the three UniProt associates, even if the same tools are essentially offered. Because all computers connecting from North America are redirected to the PIR's UniProt site, we will only consider that version of the search interface. (PIR's UniProt website: http://www.pir.uniprot.org/)

The two main search tools are:

1. **Text Search**, which allows you to search in one field in particular or all of them at once for plain query strings. One of UniProt's layers/databases (as discussed in the previous section) must be selected. Query boxes can be added as you go by clicking the add input box or + box button, with the corresponding boolean operator (and, or, not) which are used to concatenate the query terms.
2. **BLAST**, which is a sequence-alignment program to search a protein sequence against a database, UniProt in our case.

---

7 Dr Dayhoff (1925-1983) was considered a pioneer of bioinformatics. She developed a number of algorithms for alignment and comparison, as well as protein and DNA databases. A footnote (!) in her biography would probably be the single-letter code for amino acids she came up with...

8 The difference between database and knowledgebase is subtle. The web defines knowledgebase as " A collection of in- formation used to answer questions", while a database is " A collection of data arranged for ease and speed of search and retrieval"

But that's not it... "There are various tools and analyses available from the individual UniProt consortium member web sites and other sites that complement the UniProt Databases. These are categorized as Similarity Search, Multiple Sequence Alignment, Batch Retrieval, Proteomics, and Bibliography. There is also a section for Comprehensive Tools/Links Lists." (http://www.uniprot.org/search/tools.shtml on the UniProt website)

Reading a UniProt entry

A UniProt entry is just text organized in a consistent format. Every entry contains information about the following items:

1. **Entry Information**: Entry name, Accession Nb, etc.
2. **Name and origin of the protein**: The protein's full name, a description, the species.
3. **References:** Articles referring to this protein.
4. **Comments:** Combination of various fields concerning that protein, like a description of the protein's function, etc.
5. **Database Cross-References**: Links to other databases concerning the protein of interest (such as domains it contains).
6. **Features:** A description of the domains, disulfide bonds, transmembrane regions, etc., with begin/end position and length.
7. **Sequence:** The peptide sequence in plain text.

The default view from UniProtPIR is of course the PIR view. Probably because UniProt is still in its infant stages, the EBI format (SRS) and the SIB format (Niceprot) are also offered as alternatives. All views show the same information, with fields ordered slightly differently.

Uniprot website

http://www.uniprot.org/

2.4 Protein Families and Domains Databases

Before talking about Protein families and domains databases, it is important to outline some of the concepts of molecular evolution (itself a major field of study in bioinformatics and biomathematics). A protein family is a group of evolutionarily related proteins (Wikipedia).

Evolution is an expensive process, in the sense that if an enzyme doesn't "work", you die. Most of the mutations will appear as neutral (there's a nucleotide change, but it's either in non-coding regions, or it doesn't change the amino acid the codon ultimately coded for) or as having deleterious effect (the aa sequence is changed, modifying the structure and function of the protein, and you die!), and a few mutations will slightly change the structure of a protein, in the long run conferring a selective advantage to the organism carrying it.

Within a protein, a structural domain ("domain") is an element of overall structure that is self-stabilizing and often folds independently of the rest of the protein chain. Many domains are not unique to the protein products of one gene or one gene family but instead appear in a variety of proteins. Domains often are named and singled out because they figure prominently in the biological function of the protein they belong to; for example, the "calcium-binding" domain of calmodulin. Because they are self-stabilizing, domains can be "swapped" by genetic engineering between one protein and another to make chimeras. A domain may be composed of one, more than one or not any structural motifs. (Wikipedia)

But such change is slow. If evolution depended only on point mutations, we wouldn't be here today. Instead, we must see proteins as collections of domains. Protein domains themselves are made of sequences of the simplest secondary structures, α-helices, β-sheets and turns (segments between helices and sheets). Therefore, in an oversimplified conclusion, the swapping, deletion, or duplication of these building blocks (entire genes, domains, secondary structures) are at the origin of most significant evolutionary changes.

Proteins are all somehow evolutionarily related and the information obtained from protein families and domains databases is crucial to understand the relationships between proteins, to infer function for newly discovered proteins and the biological importance of certain protein domains.

2.4.1 PROSITE

Prosite is both a database and a collection of tools. As a database, it serves to collect the amino acid sequence patterns for different peptide motifs. The collection of motifs is drawn from analysis of the amino acid sequences in the SWISS-PROT/Tremble database. The main tool of interest to the user is the peptide scan function of ScanProsite, which detects the presence motifs from the database in any amino acid sequence of interest. Other tools available but not covered in this manual include the motif scan function of ScanProsite, tools which scan against other motif databases and tools which allow the user to scan various databases in search of as yet unnoted motifs and create profile for them.

Prosite was written by L. Falquet, M. Pagni, P. Bucher, N. Hulo, C.J. Sigrist, K. Hoffmann and A. Bairoch, was produced by a collaboration between the Swiss Institute of Bioinformatics(SIB) and the European Bioinformatics Institute (EBI) and is hosted on ExPASy (Expert Protein Analysis System) the proteomics server of SIB. It is available in Canada via the mirror site at http://ca.expasy.org/prosite/.

THE PROSITE LOGO.

What does the Prosite database contain?

The Prosite database consists of only two files: the data file PROSITE.DAT and the documentation file PROSITE.DOC. Both are text files and both contain exactly one entry for each motif which has been identified by Prosite.

The format of each data file entry depends on whether it represents a pattern or profile described motif. While both give a data file identification name, data file accession number and pointer to the motif's documentation entry, a motif pattern entry give a one-line pattern description and a motif profile entry will give a multiple-line weight matrix. A pattern description defines the exact amino acid sequence expected for the motif whereas a profile weight matrix defines gives score values for all the different amino acids for each site.

The entries of the documentation file all conform to a single format and each contain the documentation entry accession number, the corresponding data entry accession number and identification name and any important documentation information regarding the entry in free-format text (ex. Biochemical, taxonomic, anatomical and source information).

Search PROSITE using ScanProsite

The aim of ScanProsite is to identify the occurrences of any motifs from the Prosite database in the sequence indicated by the user. To do this, the tool scans through the entire amino acid sequence once with each motif entry in the PROSITE.DAT file. The scanning process consists of progressively aligning the pattern or profile with different positions in the sequence.

The first alignment matches the first position of the pattern or profile to the first position of the sequence and compares all the now aligned sites. If the pattern finds a match or the profile score is high enough (both situations are called hits), the positions are marked as being the pattern or profile's motif. The pattern or profile's alignment head then moves forward one site in the sequence to begin comparison all over again. Thus, identified specific motifs may overlap. As this scanning process is done with all the different patterns and profiles, different identified motifs may also overlap. Because too many or too extensive overlaps are likely meaningless, once all possible motifs have been identified, Prosite implements an algorithm to select among them.

Operation and interpretation of ScanProsite

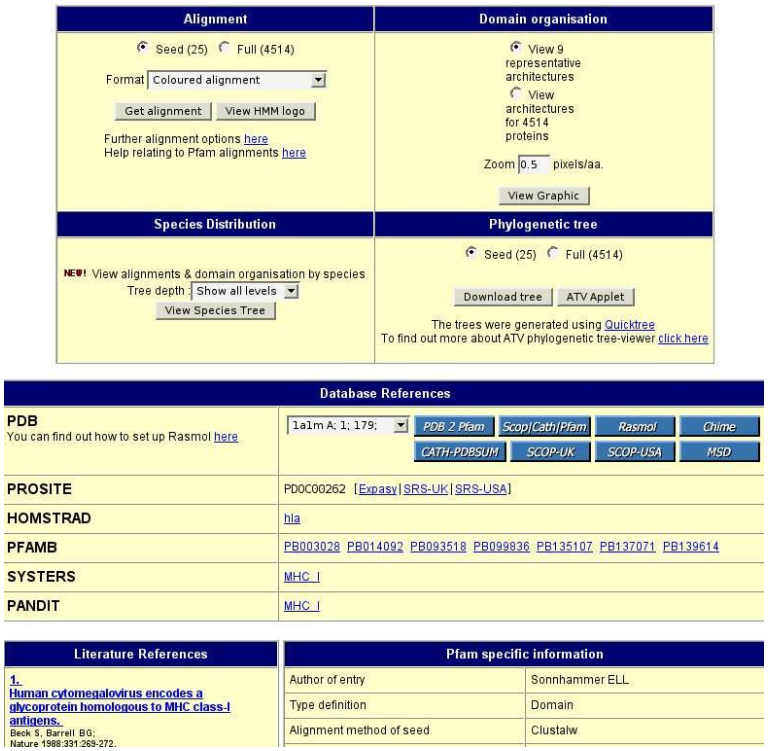
The Quick Scan tool on the main page of Prosite performs exactly the same function as ScanProsite with all but one default option set. The input can be a sequence (in raw, FASTA or Swiss-Prot format), an accession number for a protein sequence in the Swiss-Prot/TrEMBL database or an ID for a protein sequence in the Protein Data Bank (PDB).

There is one option whose setting must be considered by the user: Exclude patters (and profiles) with a high probability of occurrence.

2.4.2 Pfam: Protein families database of alignments and HMMs

Pfam is a large collection of multiple sequence alignments and hidden Markov models[9] (HMMs) covering many common protein domains and families.

While there is only one Pfam database in circulation, there are many websites from which it is accessible from (same db, different interfaces). These sites aren't mirrors of each other, but the services offered on them are equivalent. There are Pfam sites in Sweden, South Korea and France, but the main ones are Sanger Institute's in the UK[10] and Washington University's in St.Louis, MI[11].



ALIGNMENTS, PHYLOGENETIC TREES, STRUCTURES AND OTHER RELEVANT INFORMATION CAN BE DOWNLOADED FROM A PFAM ENTRY PAGE.

What's in Pfam?

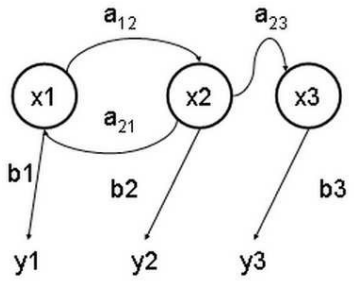
Pfam is divided in two sets of protein families

- Pfam-**A** families are based on curated multiple alignments. A certain number of proteins (ranging from around 10 to a few thousands) are chosen to form the "seed" group representing a protein family. An example of protein family can be the "Class I Histocompatibility antigen, domains alpha 1 and 2" (Code: MHC I), which regroups MHC-I-like proteins based on HMMs .
- Pfam-**B** is based on an automated clustering of the proteins in UniProt not already in Pfam-A from a database called ProDom.

9 A hidden Markov model is essentially a statistical model, which has found interesting applications in describing protein families, as well as in computerized speech recognition.

10 http://www.sanger.ac.uk/Software/Pfam/

11 http://pfam.wustl.edu/



How to visualize a Markov model. . . (x: States of the Markov model, a: Transition probabilities, b: Output probabilities, y: Observable outputs.) (Picture: Wikipedia)

The data in a Pfam entry will include the following:

- A **seed alignment** which is a hand edited multiple alignments representing the family.
- **Hidden Markov Models (HMM)** derived from the seed alignment which can be used to find new members of the domain and also take a set of sequences to realign them to the model. One HMM is in ls mode (global) the other is an fs mode (local) model.
- A **full alignment** which is an automatic alignment of all the examples of the domain using the two HMMs to find and then align the sequences.
- **Annotation** which contains a brief description of the domain links to other databases and some Pfam specific data. To record how the family was constructed.

What can I do with Pfam?

With full alignments and hidden Markov models, one has a lot of raw information representing a family of protein. Keeping up with the MHC-I family of proteins, some viruses, like HCMV, encode MHC-I-like proteins in order to evade killing from Natural Killer cells. Using the data contained in the MHC I Pfam entry, one could then build programs to scan viral protein databases for novel candidates in viral host resistance. This is a first step in an experiment that will necessarily include a wet lab component.

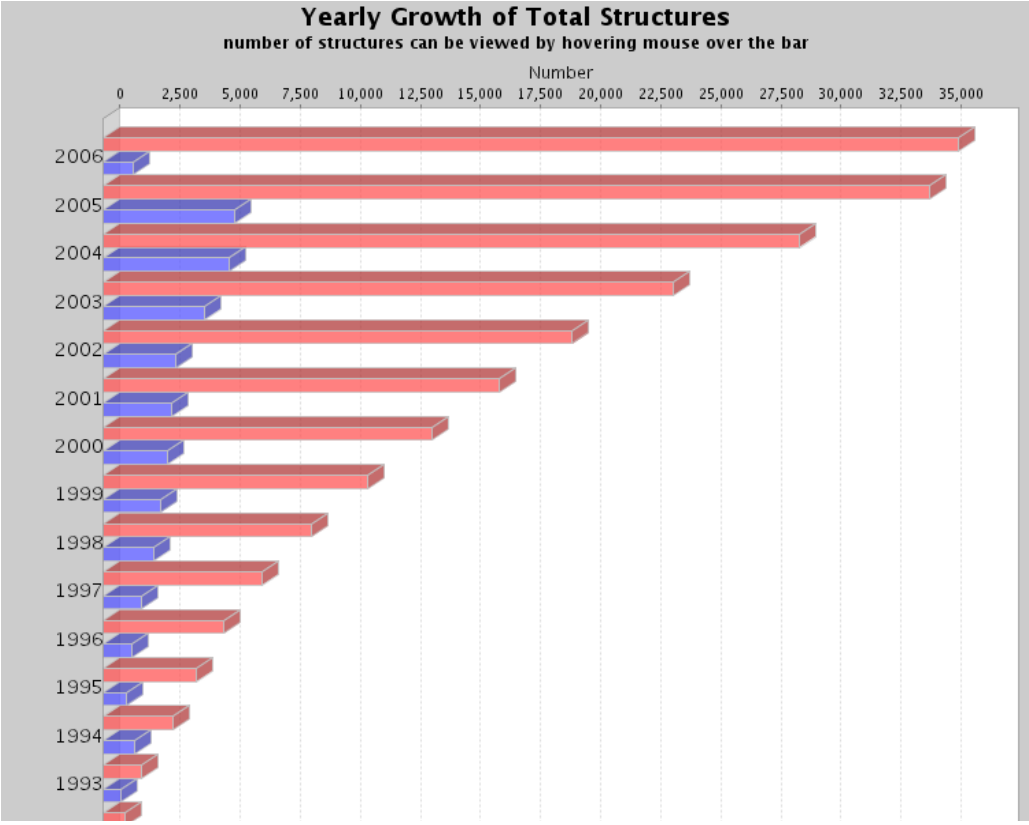
2.5 3-D Structure Databases

From the protein sequence, the ultimate goal would be to decipher the function based on the sequence alone. While sequence comparisons are somewhat useful in this manner, knowing the three-dimensional structure can get us a step closer to this goal. This is done, in part, by elucidating the interaction between macromolecules and by comparing the spatial arrangement of the polypeptide chain. As well, 3-D structures have been of prime importance in the rational development of new drugs (versus the good old trial and error approach). PDB is the most comprehensive structural database, and we will now go more in depth about it.

2.5.1 PDB: The Protein Data Bank

The Protein Data Bank (PDB, http://www.rcsb.org/pdb/) is the single worldwide repository of (despite its name) protein, nucleic acid and other biologically-relevant 3-D structures. “These data, typically obtained by X-ray crystallography or NMR spectroscopy, are submitted by biologists and biochemists from around the world, are released into the public domain, and can be accessed for free. The database is the central repository for biological structural data.” (Wikipedia)

PDB is hosted and managed by a three research centers based in the United States that are also part of the Research Collaboratory for Structural Bioinformatics (RCSB) consortium. For reference, they are Rutgers University, the San Diego Supercomputer Center (SDSC), and the Center for Advanced Research in Biotechnology (near Washington DC)



As of March 2006, the PDB is holding a total of 35579 3-D structures (27204 in Sept 2004), among which 32519 are proteins, peptides or viruses, 1448 are proteins/nucleic acids complexes, 1510 are nucleic acids only, and 102 are other compounds.

Determining the 3-dimensional structure of macromolecules, in particular of proteins, is a daunting task involving X-ray crystallography or NMR spectroscopy. The success of these experimental techniques is difficult to predict, and structure determination is often likened to an art. X-ray crystallography, for instance, requires the growth of a protein crystal up to 1 mm of size from a highly purified protein source[12].

Contents

What information is contained in each entry of PDB? "A variety of information associated with each structure is available, including sequence details, atomic coordinates, crystallization conditions, 3-D structure neighbors computed using various methods, derived geometric data, structure factors, 3-D images, and a variety of links to other resources." (PDB website)

Two file formats are available to represent the structural data contained in a PDB entry (and other information such as name of molecule, references, etc˙). They are namely the "PDB" and the "macromolecular Crystallographic Information File" (or "mmCIF") formats, which consist essentially of plain text specifying spatial coordinates of the atoms and bonds of the molecule in question. You could read all of this code if you opened the files from a text editor like Notepad, but that's not generally what you'd want. Instead, visualization programs are used to convert the text information into molecules you can twist and turn in space.

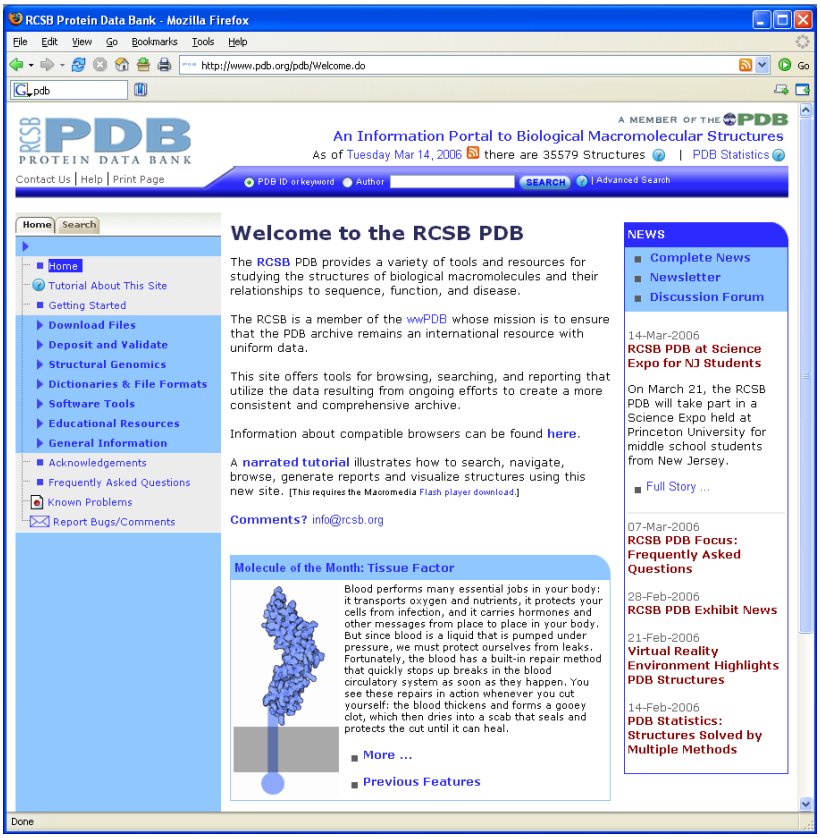
Searching PDB

This section will now cover methods for searching the PDB database for information that's relevant to us.

1. The simplest search tool is SearchLite, which is available directly on PDB's homepage (as the text input box right in the middle of the webpage). You may enter a query as a PDB ID (a unique 4-character alphanumerical string, scientific papers sometimes use these), the authors of the structure or the full text search, which is basically any text that's found associated with an entry.

12http://www.whatislife.com/reader/techniques/techniques.html

2. Other search tools are linked from the homepage:
    - **QuickSearch**: Searches the entries like SearchLite, but also all of the support pages making up PDB.
    - **SearchFields**: Searches against specific fields of information - for example, deposition date or author.
    - **Status Search**: Searches on the status of an entry, on hold or released.
3. Interactive Search: Among structures obtained through one of the types of search previously mentioned, you can choose a subset of structures to perform additional searches. Structures to search within can be selected through a pull-down menu on a query results page, or manually by checking the box corresponding to a target entry.



THE PROTEIN DATA BANK HOMEPAGE (HTTP://WWW.PDB.ORG/)

2.5.2 Viewing Structures with RasMol and derivatives

RasMol is considered the grandfather of many molecular visualization tools out there. Its first version was released by Roger Sayle at the University of Massachusetts in 1992. Molecular visualization tools back then had to be run on graphics workstations, but RasMol, being an extremely well optimized program, could run on then moderately powerful computers.

It is still being used nowadays, although technology has since evolved and several convenient web-based tools have been developed. One piece of software adopted by the PDB is **Jmol**, which does not need to be installed and can be run from the web on any computer equipped with a browser and Java. Jmol borrows its scripting language from RasMol, and users can interact with the program using a command-line interface, which allows the user to perform several tasks that would otherwise be too complex in a point-and-click fashion (such as selecting a range of amino acids and to "highlight" them).

Another nice program to view and customize structures is **DeepView** (aka Swiss Pdb-Viewer) developed by the Swiss Bioinformatics Institute and GlaxoSmithKline. Another fairly popular downloadable program is **PyMol**. Both of these programs are used in a production environment and have advanced view and customization features you will not find in RasMol or web-based applications.

2.6 Other databases

There's bound to be a database suited to your needs. For instance, the McGill Center for Bioinformatics hosts a database called HERA, which compiles all human proteins known to reside in the endoplasmic reticulum.

Amos Bairoch, the founder of the Swiss-Prot database, lists many databases of biological relevance, which takes more than a half of his already extensive links page! (http://www.expasy.org/alinks.html)

2.7 References

- General references
  1. Introduction to Molecular Biology Databases (by Rolf Apweiler, EBI's Swiss-Prot coordinator)

     http://www.ebi.ac.uk/swissprot/Publications/mbd1.html
  2. The NCBI Handbook

     http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook
- 3-D structure databases
  1. A reference article used when citing PDB: H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. Nucleic Acids Research, 28 pp. 235-242 (2000)
  2. More publications from PDB are available on the PDB Info webpage http://www.rcsb.org/pdb/info.html. It's information you can skim through during your spare time.
  3. Examples of mmCIF, the file format used in PDB:

     http://ndbserver.rutgers.edu/mmcif/examples/
- Protein sequence databases
  1. UniProt User Manual

     http://ca.expasy.org/sprot/userman.html
  2. Protein Sequence Databases (by Apweiler, Bairoch and Wu): A short overview of the existing protein sequence databases, and what differ between them (Curr Opin Chem Biol Feb 2004 8(1):76-80)

Chapter 3: Tutorials

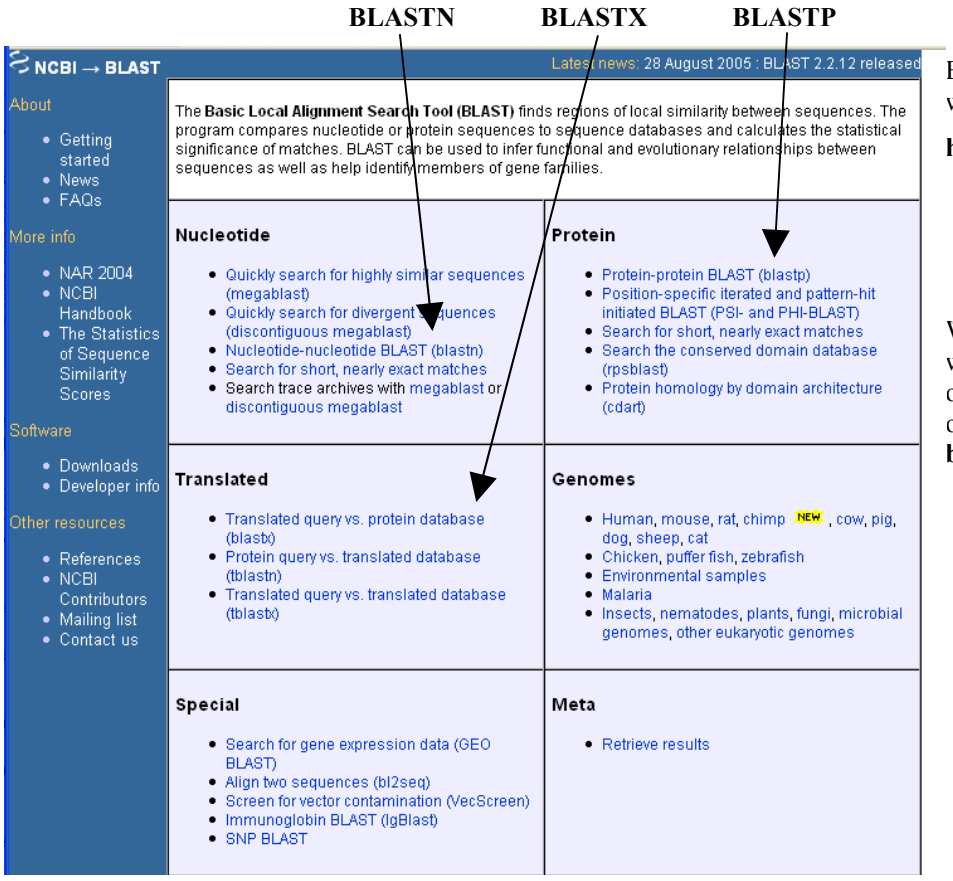
Tutorial: How to use BLAST to search for homologous sequences? (and using NCBI ORF finder)

By Oksana Kapoustina <oksana.k@gmail.com> and Abrar Khan <abrar.khan@mail.mcgill.ca> (layout: Cedric Sam)

Version 2.0, August 2005.

*BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool) is a bioinformatics tool that is used to compare an unknown sequence (from now on we will call this sequence a query sequence) to millions of known sequences in a database. Therefore the choice of, the completeness and the integrity of the database are essential to a BLAST search. BLAST hosted by NCBI works by comparing a query sequence to all the sequences in the NCBI databases. It does so by looking for "regions of similarity" between the query sequence and sequences contained in the database.*

Part 1: Using the web-based version of BLAST



BLAST can be found on the NCBI website:

http://www.ncbi.nih.gov/blast/

When you enter the BLAST website you will be given a choice of different BLAST programs. For our purposes we will be using **blastn**, **blastp** and **blastx**.

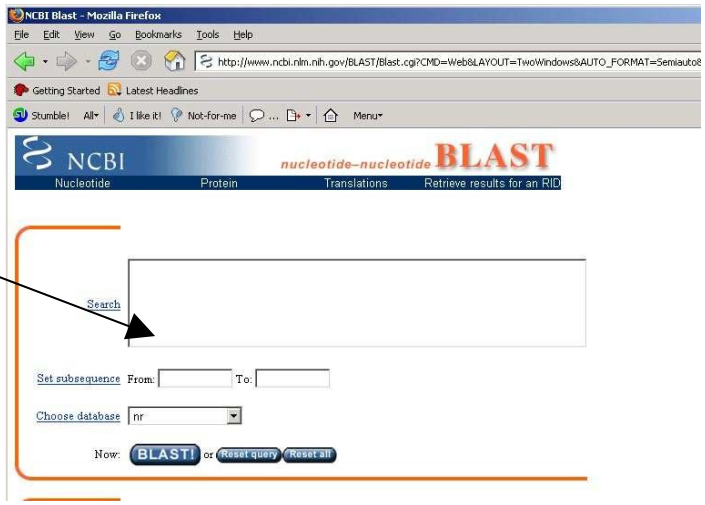
- **blastn** is used to compare an unknown nucleotide sequence to the NCBI nucleotide database
- **blastp** compares a protein sequence to a protein database
- and **blastx** takes a nucleotide sequence, converts it into its protein complement and compares it to the protein database

Note: Just remember that you have to compare nucleotides with nucleotides and proteins with proteins, unless you are using blastx.

Part 2: The BLAST form

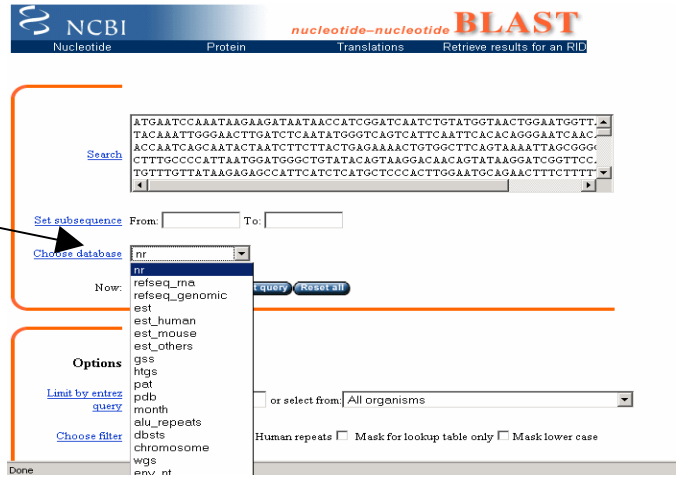
Once you have chosen the appropriate BLAST program you will see the BLAST input window. You can enter the query sequence into the search window. The sequence can be in plain text format. **FASTA** (.fa) format will also be accepted.

Insert sequence



Here we inserted a query sequence into the search window. In this case we are using blastn and the sequence inserted is a nucleotide sequence.

Choose a database

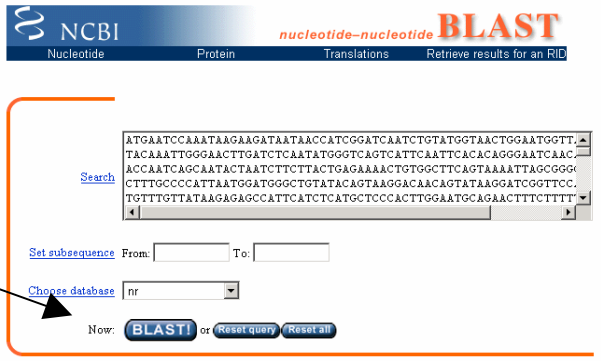


Choosing a database:

It is now time to choose the database that BLAST will use to search matches for the query sequence. As it was already mentioned there is a number of databases that BLAST can use. All of them contain sequences that have already been identified by the researchers. If you click on "choose database" and scroll down you will notice that some databases are specific for an organism and some can only be used for nucleotides or proteins.

- The database fit for our purposes is the **nr** database.
- For the complete list and description of BLAST databases you can refer to or the manual.
- After entering the sequence and selecting a database click on the BLAST button

Click BLAST

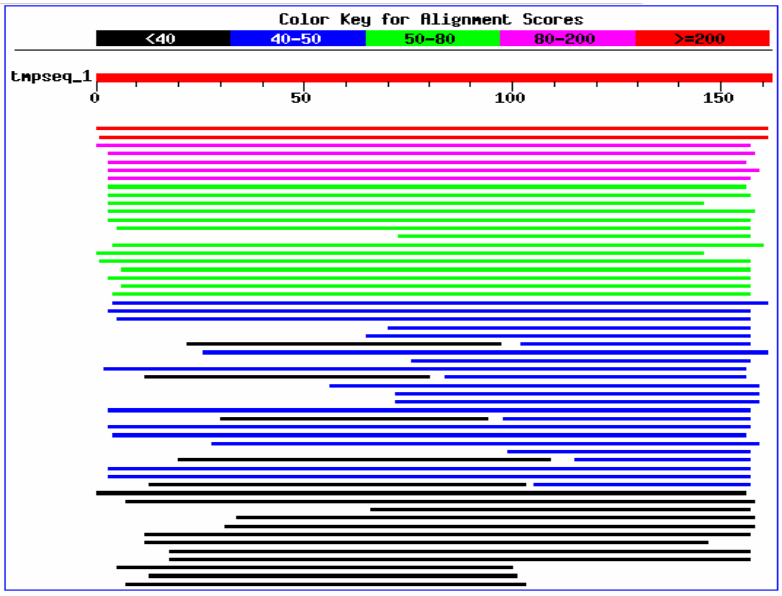


Now you should see the following screen. It will let you know the estimated search time and the ID of your request. To continue click the "Format!" button. This will produce a window where the results of your query will be displayed after the BLAST program will have processed it.

Click Format! To continue

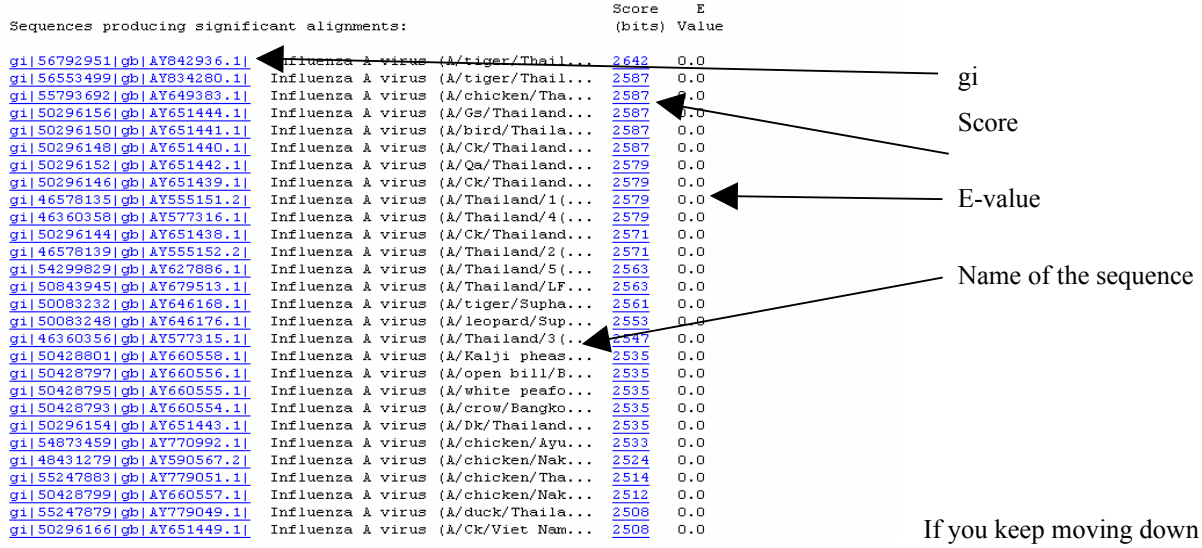
Part 3: The BLAST results page

The wait may be quite long in the case of long query sequences, or during peak hours. Once the search is over you will see the BLAST Results window. If you scroll down you will see a picture representation of your search, which will look something like this:

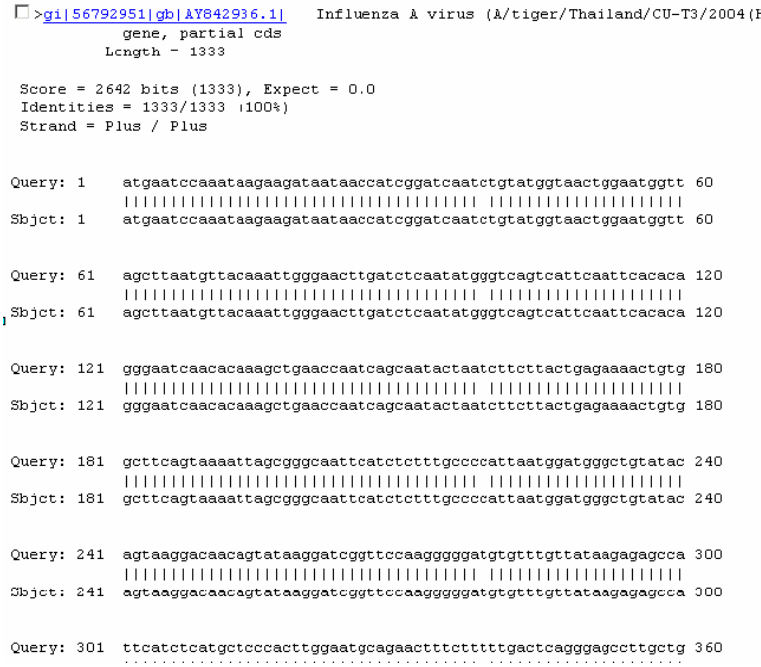


- Each of the lines on the picture represents a match between a database sequence and a query sequence.

- There can me numerous matches for one query sequence. Only the top few matches are shown in the picture representation.
- The picture is colour-coded (you can see the colour map on top of the picture) Red lines represent matches with the highest scores (> 200), green lines are for the lowest scores and so on.
- If you click on any line in the picture you will be taken to a page that shows the alignment of the matching sequence with the query sequence.
- If you scroll down further you will see a list of all the matching sequences in the database
- On the left is the gi number – it is a unique identifier for a sequence within a database.
- Clicking on the gi identifier will summon a new page with a complete description of the sequence provided by GenBank
- On the right you can see the **scores** and **E-values**
- Clicking on the score takes you to the alignment of the database match with the query



If you keep moving down the page you will notice that most of it is taken up by the alignments that look something like the next picture below.



This line lets you know the score and the E-value of this specific alignment

This line shows the number of residues in the query sequence and in the alignment that are identical

- Alignments represent two sequences, the query sequence and the matching sequence, lined up against each other. This will help you determine how many mutations there are and where exactly they are located.
- The query sequence is usually on top and the database match is usually on the bottom.
- The numbers on each side of the sequence represent residue numbers. (eg: the first line of the alignment shows residues from 1 through 60).

Part 3: Interpreting the BLAST results page

Scores: Scores in BLAST represent the extent of similarity between the query sequence and a database sequence. They are based on the percent identity/conservation observed when the sequences are optimally aligned against each other. Naturally, the higher is the score the more similar are the sequences.

E-Values: E-values, also called the Expect values, are the measures of the "background noise" in an alignment. They represent the number of matches with high scores that can occur in a searched database purely by chance. Eg., E-value of 1 that means that there could be at least one sequence in a database that has a high alignment score (i.e. it will be considered a matching sequence for the query) but it is not really a match for the query sequence (it has a high score purely by chance). Therefore, your goal is to find E-values that are closest to "0". An E-value of "0" means that the match is one of a kind and, therefore, it is significant. If you are wondering why some unrelated sequences can have high alignment scores with the query sequence refer to the BLAST section of the manual that you were given but largely, this question is beyond the scope of our course.

Hint: For the purposes of this exercise we will concentrate on the scores rather than the E-values. i.e. As long as your result has a very high score it does not have to have a perfect E-value.

Choosing the best matching sequence

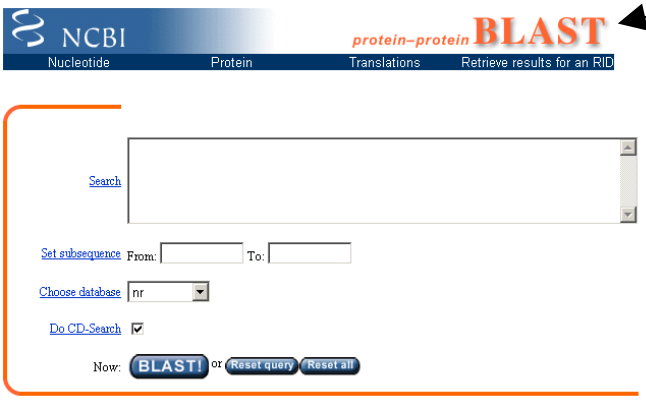
- Once the alignment is complete and you have examined the results you can choose the sequence that matches the query sequence the best. Remember, when choosing the optimal matching sequence you want it to have the lowest E-value and the highest alignment score~
- In the previous example the first sequence in the list of would be a perfect match for the query since it has the highest score and an E-value of "0".

```
                                                                    Score     E-value
gi|56792951|gb|AY842936.1| Influenza A virus (A/tiger/Thail... 2642      0.0
```

- From the name of the match you can infer that the query sequence represents an Influenza A virus gene or a part of it. This gene most likely belongs to the A/tiger/Thailand/CU-T3/2004(H5N1) strain and it codes for the neuraminidase gene.
- (You can obtain all of this information by clicking on the score beside the gene name and examining the header of the alignment.) You can get more information about the Influenza A gene from GenBank by clicking on the gi number beside it.

Part 4: Using Blastp to search the protein databases

As stated previously, the BLAST search pages allow you to select from several different programs (blastn, blastp, blastx). The blastp database takes an amino acid input sequence and compares it with millions of protein sequences within the Blast database. It then provides you with a list of the closest matches found.



Protein-protein means you are in the right database! Depending on what you are looking for, you can modify the database to search within

The following is a list of some important databases used in blastp searches:

nr (default):All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF

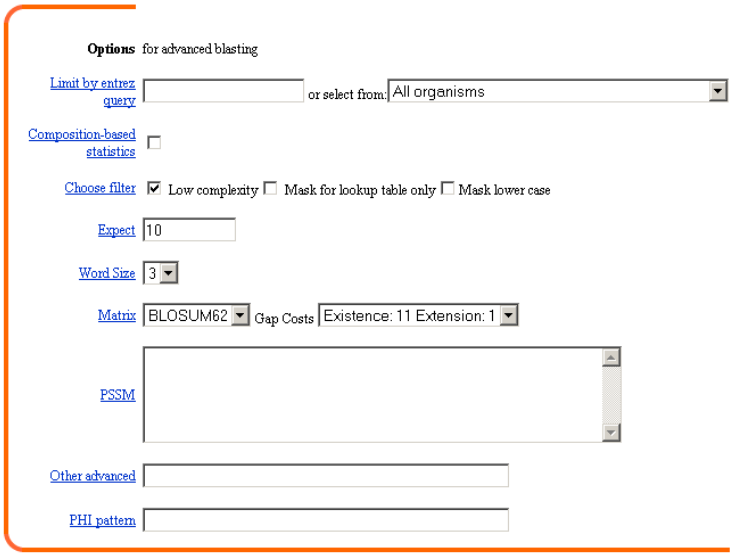
month:All new or revised GenBank CDS translation+PDB+SwissProt+PIR released in the last 30 days.

swissprot: The last major release of the SWISS-PROT protein sequence database.

pat: Protein sequences derived from the Patent division of GenBank.

pdb: Sequences derived from the 3-dimensional structure Protein Data Bank.

Blast Options



Limited to Entrez query: BLAST searches can be limited to the results of an Entrez query against the database chosen. This can be used to limit searches to subsets of the BLAST databases.

Filtering: Mask off segments of the query sequence that have low compositional complexity.

Expect: The statistical significance threshold for reporting matches against database sequences; the default value is 10, meaning that 10 matches are expected to be found merely by chance

Word size: BLAST is a heuristic that works by finding word-matches between the query and database sequences. One may think of this process as finding "hot-spots" that BLAST can then use to initiate extensions that might lead to full-blown alignments.

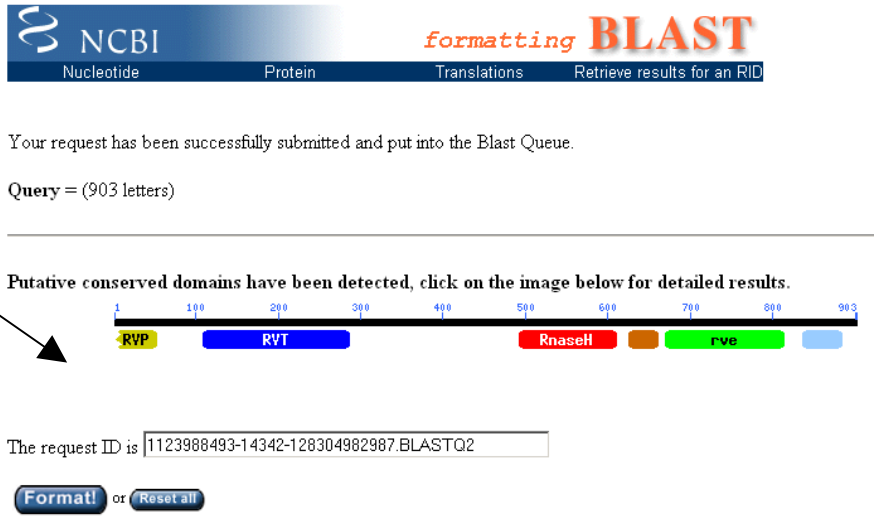
Matrix: A key element in evaluating the quality of a pairwise sequence alignment is the "substitution matrix", which assigns a score for aligning any possible pair of residues. The matrix used in a BLAST search can be changed depending on the type of sequences you are searching**.**

Part 5: Analyzing Conserved Domains using Blastp

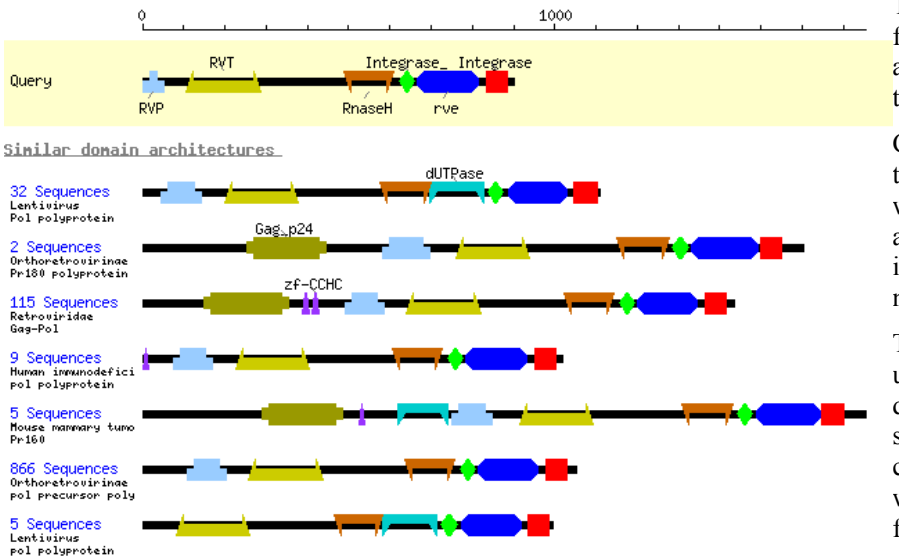
If you are lucky enough to have a sequence that is highly annotated, you may be able to determine the protein function of specific open reading frames through the use of conserved domains using the blastp database.

Conserved domains are a region in a protein sequence that are retained in the 3-D structure of a protein and confer a special function for the protein (i.e. zinc finger domain, Ribonuclease domain)

Conserved Domains



Clicking on the colored conserved domains above will open a more detailed outlook of the various domains and their positions within your ORF.



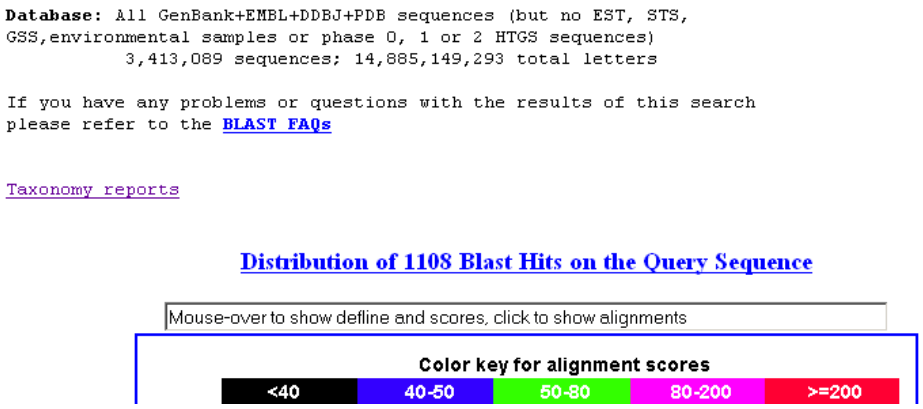
The domain relatives button looks for similarity of domain architecture between different taxonomical groups.

Clicking on the conserved domains in the graphical view or the tabular view will direct you to the **Pfam** website and provide you with more information about the structure and nature of that domain.

The domain relatives page (above) is useful in analyzing homology of the domains between evolutionary species. It also shows domains in close proximity within other species, which may be useful in defining its function.

Part 6: Analyzing Taxonomy reports

An interesting feature of the Blastn output is the Taxonomy reports, which can provide you with valuable information of the taxonomic relationships among the records returned from a BLAST search. The taxonomy report link can be found just above the Blast Hits on the Results page.
Clicking on the Taxonomy Reports link on the BLAST results page will generate taxonomy reports in three formats: a Lineage Report, an Organism Report, and a Taxonomy Report.

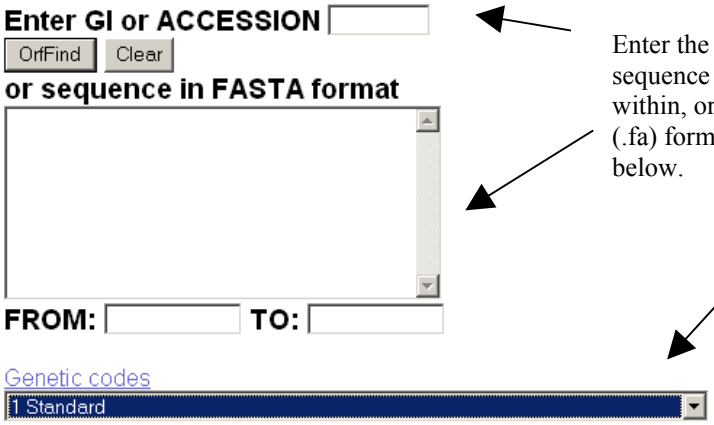


- The **Lineage Report** gives a simplified view of the relationships between the organisms generating database hits to the query sequence by showing how closely these organisms are related to a "focus organism", according to the taxonomy database. This focus organism is the organism giving the strongest BLAST hit and this will often be the source organism of the query sequence.
- In the **Organism Report**, the BLAST results are grouped into blocks by species. Within each species block, the records are sorted by BLAST score. The order of species blocks themselves is based on the BLAST score of the best hit within the block.
- The **Taxonomy Report** summarizes the relationships among all of the organisms found in the BLAST results. Using this report, it is easy to see how many records are found within broad taxonomic groups such as the mammalia, or the archaea.

Part 7: Using NCBI's Orf Finder

NCBI ORF finder website: http://ncbi.nih.gov/gorf/

"The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server." (NCBI website)

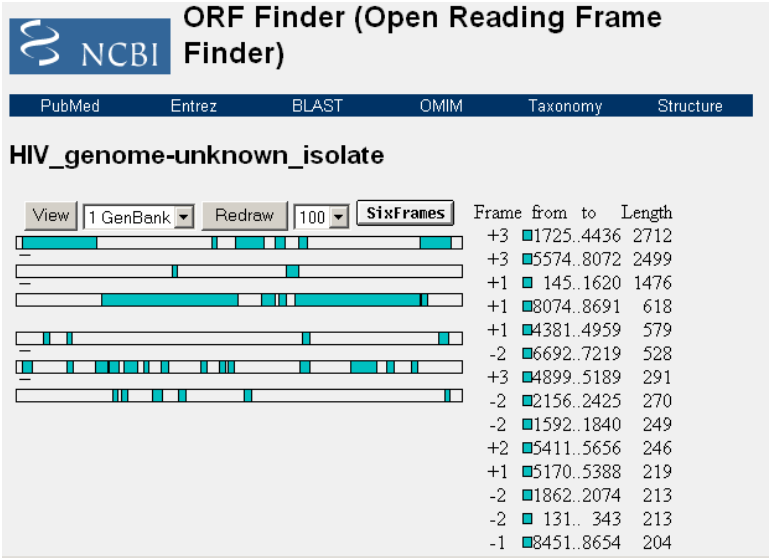


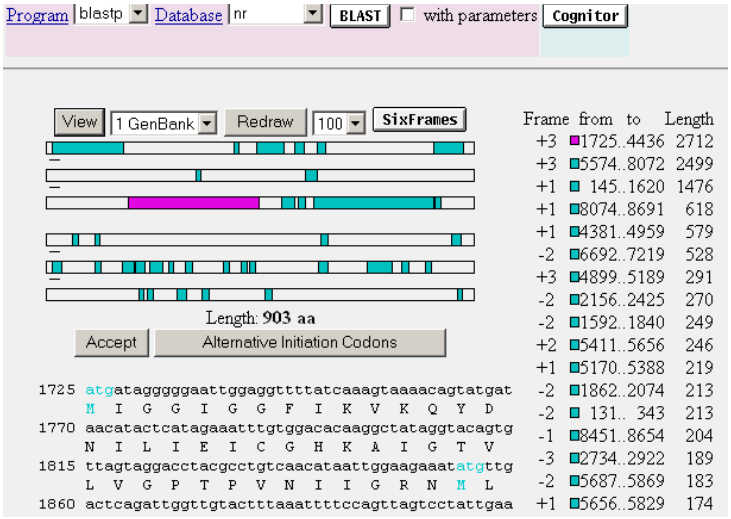
Enter the accession # of the sequence you would like to search within, or you can paste the fasta (.fa) format of nucleotides in the box below.

To limit your search to within a specific genetic code, you may select it from the below toolbar.

ORF Finder Results

The Orf Finder results page lists the largest open reading frames, along with a graphical view of their relative positions.





The graphical view shows all three reading frames both in the positive and the negative directions.

Clicking on an ORF will highlight it in the list and present you with its sequence.

You can also examine alternative initiation codons, as opposed to the deafult ATG codon that ORF Finder uses.

Tutorial: How to use ClustalW to perform multiple sequence alignments and build phylogenetic trees?

By Cedric Sam <cedric.sam@elf.mcgill.ca> Version 2.0, September 2005.

Part 1: Using the web-based version of ClustalW

For this tutorial, we will be showing you how to use the web interface for ClustalW hosted by the European Bioinformatics Institute. If you can't remember the website, Google will: just search for “clustalw” and you will get it as your first hit.

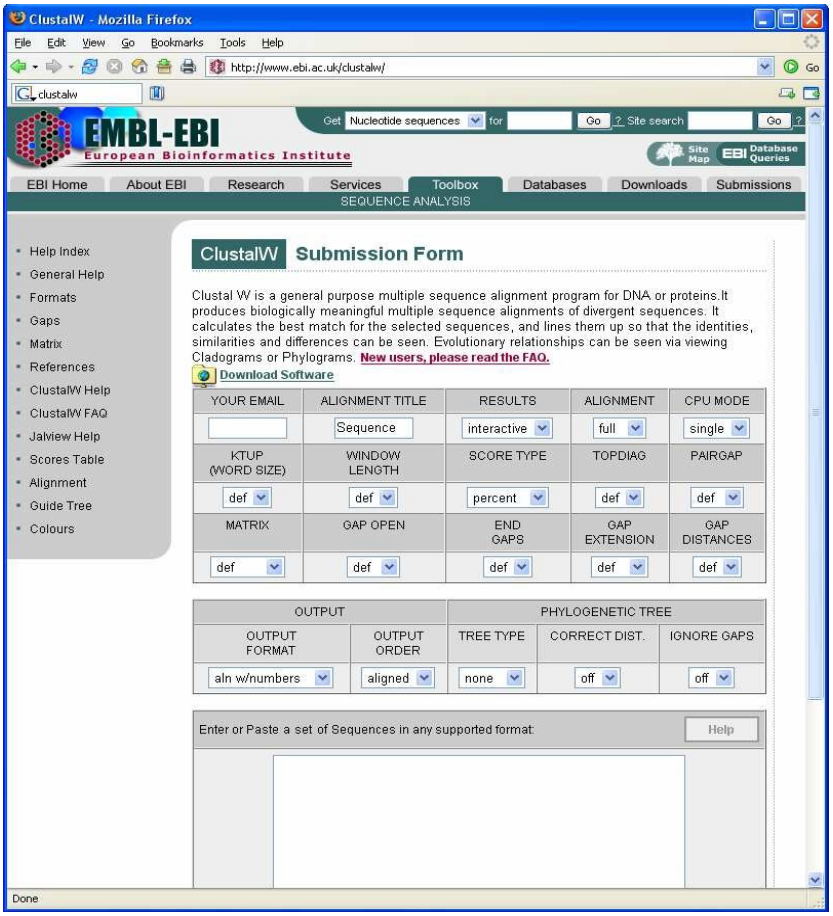
If you don't want to look, ClustalW can be found here: **http://www.ebi.ac.uk/clustalw/**

(ClustalW is a program. Many interfaces exist for it, and we show only show you the web version. You will also find ClustalW bundled with DS Gene, and another ClustalW called “ClustalX” which can be downloaded and run on your home computer!)

Part 2: The ClustalW form and importing data to it

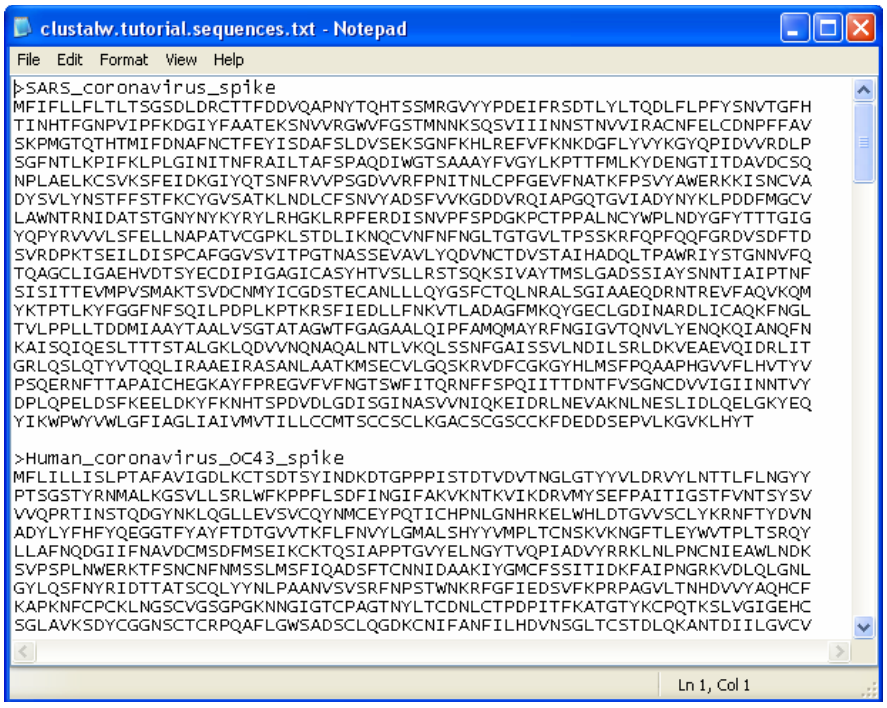
First, take your time and look around. The ClustalW form has many options to be toggled. For the purpose of our tutorial and upcoming exercise, we will keep all the default settings, minus trivial things such as the name we want to give to identify our query and an e-mail to send results to.

For now, don't touch the output and phylogenetic tree sections. By default, ClustalW will output a “multiple sequence alignment” (or MSA), which takes several sequences, amino acid or nucleotide, and gives you the best alignment it can find between all the sequences. As part of the "alignment" process, and depending on the sequences given, gaps are inserted and "similar" letters aligned together. ClustalW is just one of many alignment programs. It has its strength and its weaknesses, but the details are beyond the scope of this course.



What sort of data does ClustalW take?

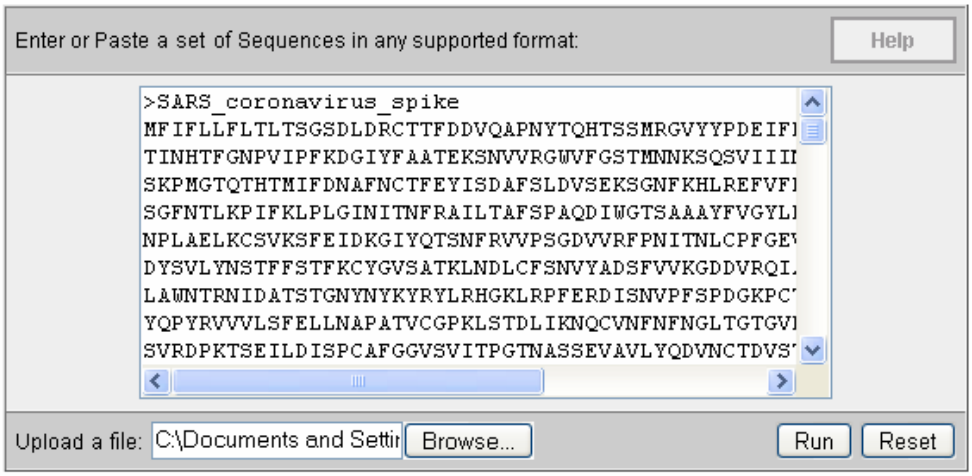
Many formats are supported by ClustalW, but we will use the format called "FASTA" (the name of another alignment program), a fairly standard and simple format to use. The FASTA format looks like this:



Each sequence is given as a block of text with a description header on a single line starting with a "Greater Than" symbol (">"). The first entry of the example given is called "SARS_coronavirus_spike" and the sequence goes like "MFIFLL...VKLHYT". The second ">" symbol indicates the start of the second sequence (you don't need to put a space between sequences and the next description line like we did here).

Enter sequences... ->

or Upload a file ->

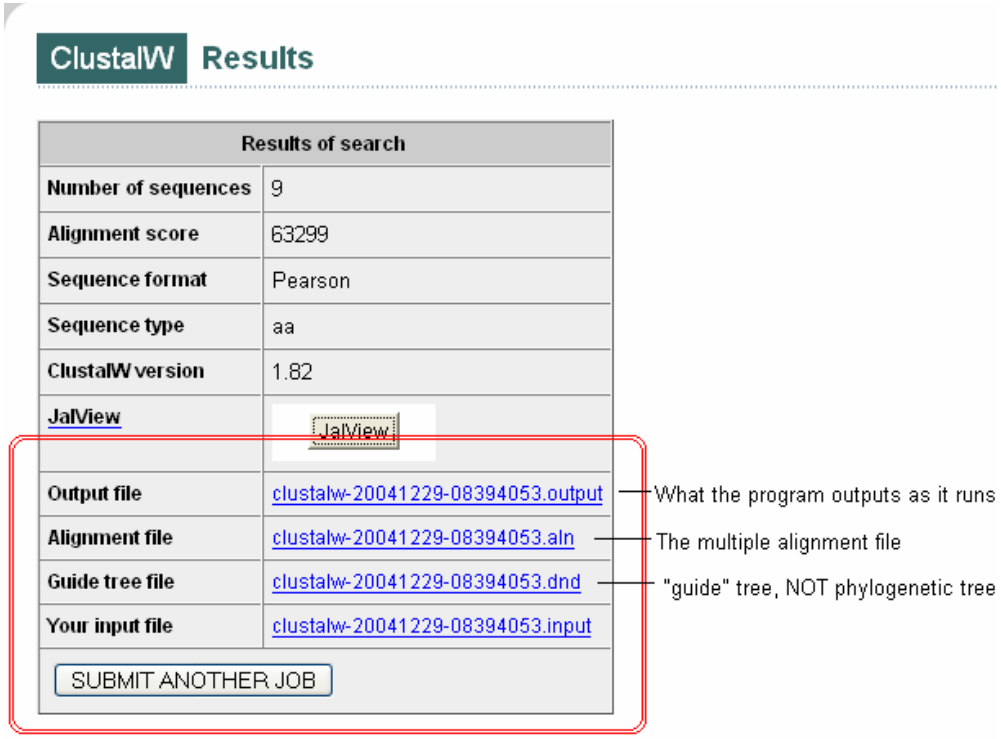


[Note: FASTA files are written as plain text. Plain text, as opposed to formatted text, consists only of characters without information pertaining to font, size, etc. Formatted text files, like Word documents, are encoded and can only be opened using specific programs, like MS Word, and cannot be interpreted by programs such as ClustalW. You must therefore use a plain text editor, like Windows' Notepad, to write your own FASTA files and save them with any Windows extension (.txt here). Remember this nuance, and always use plain text for anything that doesn't require formatting (such as sequences).]

For such short sequences, we might be waiting for a few minutes to get our results. For longer queries, it is possible that the query takes up to 20-30 minutes (especially if you send at peak periods - daytime for North American time zones). From the ClustalW form, it's possible to change the "Results" field (at the top) from the default "interactive" to "e-mail", where a link to your results is sent to you when your query has been processed.

Part 3: Interpreting the Multiple Sequence Alignment (MSA)

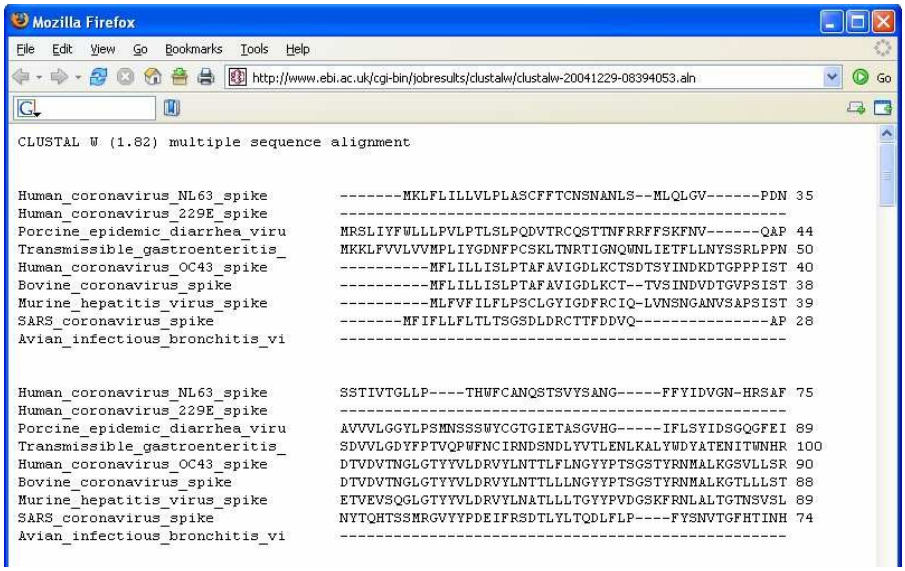
After processing, you get your first set of results: the MSA. Along with other data, it is displayed in your browser window as follows:



We're not showing the whole page here, but be aware of the output you get.

The "Output file" shows what the program outputs as it runs. ClustalW first does a pairwise alignment between each sequence inputted, and then puts them together for the multiple alignment. This file contains important data about the % identity between each sequence.

The "Alignment file" is the MSA itself. Symbols below each column of the MSA roughly indicates the level of identity for the aligned nucleotide/amino acid. No symbols are seen in the image because there is not enough identity in this section of the alignment.



The guide tree is constructed by ClustalW to infer a MSA. It is based on pairwise alignments, and is not a valid substitute of a "true phylogenetic tree" (itself built from a MSA).

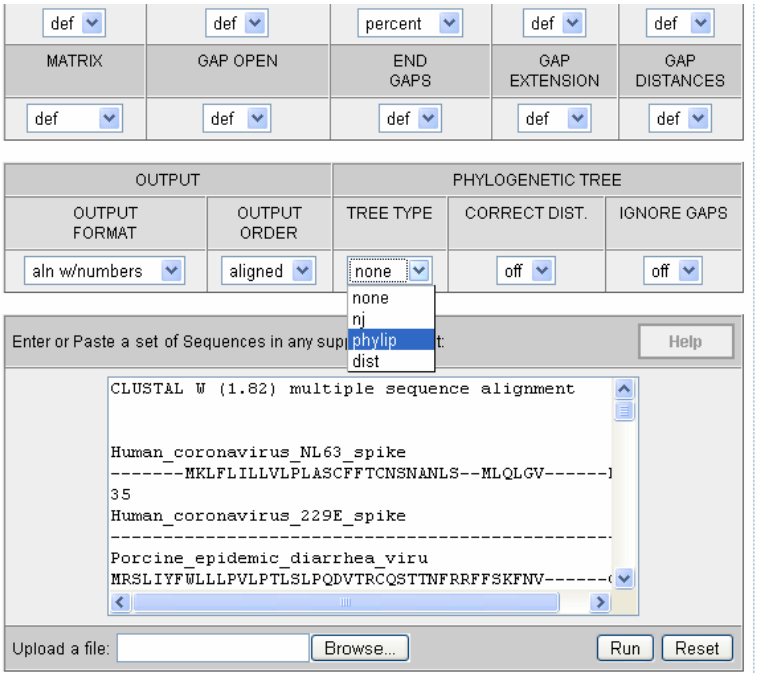
Part 4: Building the phylogenetic tree

Now save your multiple alignment, because we need it for the next step.

Depending on the system you work on, you may open you .aln directly from your browser window. Copy its contents to the Clipboard (or Notepad, if you're afraid to lose it). Otherwise, you should save you .aln file. Right-click on the link at the top of the page.

Alternatively, you may also copy the alignment as seen on the results webpage. ClustalW is smart and will interpret it (but only if you didn't copy any junk before and after the alignment!).

Now return to your original ClustalW form (http://www.ebi.ac.uk/clustalw/) and paste your multiple sequence alignment, as ugly as it might be. (Or choose to upload the .aln file you saved - it's always a good idea to save every file you use in a safe place. Like a good experiment in a real lab requires you to keep track of anything you do in a lab book.)



The only modification you have to make is at the level of "tree type" in the Phylogenetic Tree section. This will tell ClustalW that we don't want the default MSA, but rather a phylogenetic tree, as the output. "Phylip" is one of the existing tree formats, which we'll show you briefly on the next page.

Press run, and after the usual wait screen, you will get the following results page:

ClustalW Results

Results of search	
Number of sequences	9
Sequence format	Clustal
Sequence type	aa
ClustalW version	1.82
Output file	clustalw-20041229-12475333.output
Phylip tree file	clustalw-20041229-12475333.ph
Your input file	clustalw-20041229-12475333.input

SUBMIT ANOTHER JOB

Phylip tree file

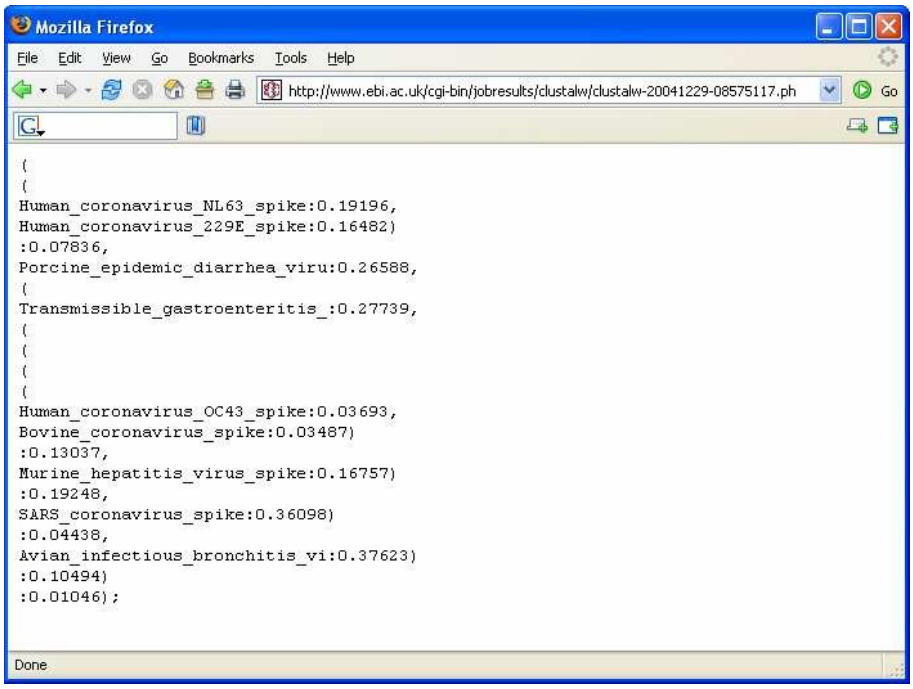
Again, the "output file" is a semi-misnomer: it is what the program ClustalW outputs while it runs. Here, nothing really "useful" comes out of it, but the length of the sequences and the name of the input format. The ".ph" file is what really interests us.

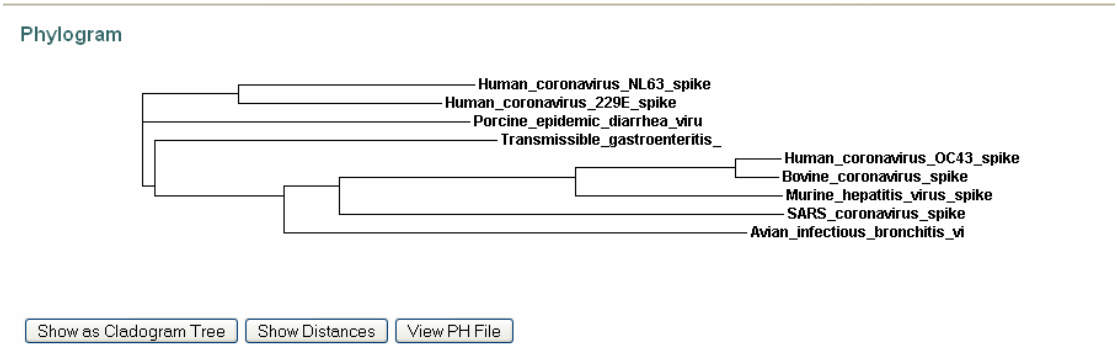
Every ClustalW results page comes with a java applet displaying a simple representation of the tree (the ph file). So, what is actually the tree? How do you represent a tree, if not as something visual?

Here's the tree data -->

Not that you need to understand the format of the .ph file, but it insightful to know that so little is used to define the appearance of the tree.

At the bottom is a representation of the tree by the java applet built into the ClustalW web version webpage.

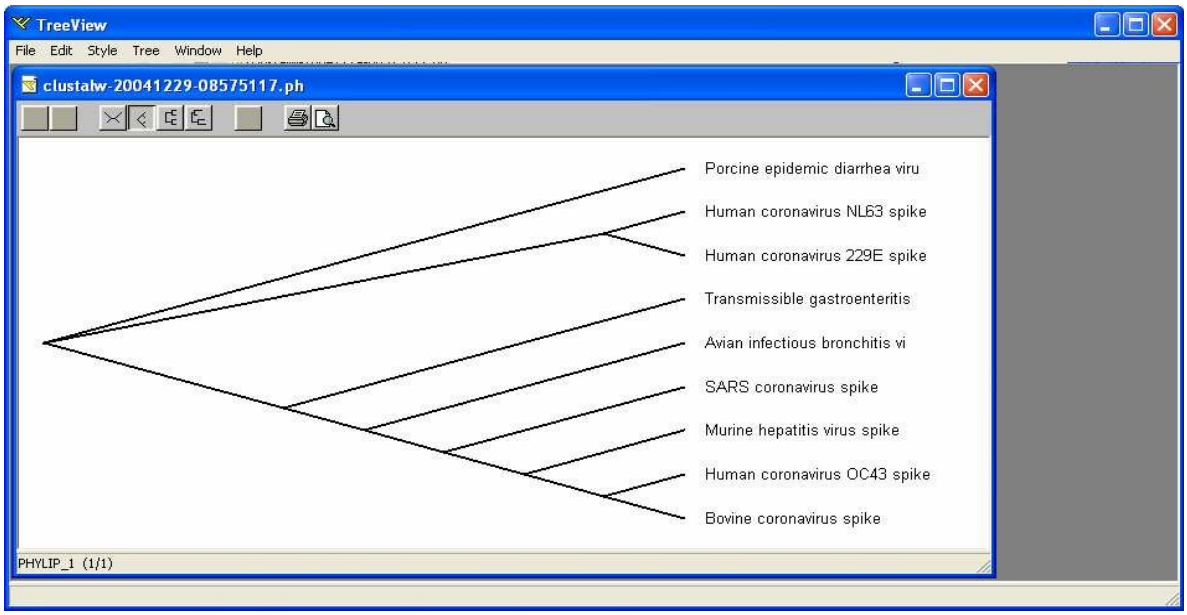




Part 5: Using TreeView to view tree files (.ph)

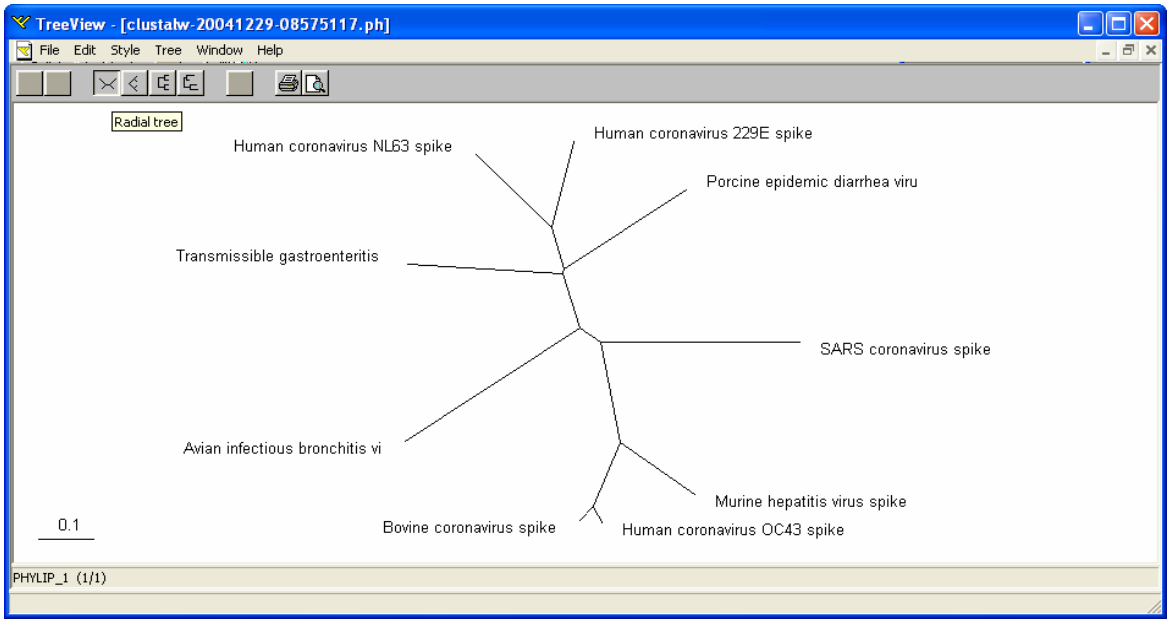
The next step is to view .ph files in a program somewhat more flexible than the ClustalW webpage's java applet. The program we use is called TreeView (while the Phylip suite contains a tree viewing utility with more viewing options, TreeView is much easier to manipulate).

TreeView has a Windows version that can be downloaded from this website:
http://taxonomy.zoology.gla.ac.uk/rod/treeview.html



Clicking one of the buttons on the top will allow you to change the "view" of the tree. Here, the same tree as before is now view as an unrooted tree, more appropriate with this example

(different species of coronaviruses with no specified evolutionary ancestry).



Tutorial: How to use PDB and Jmol to find and manipulate three-dimensional Structures

By Cedric Sam <cedric.sam@elf.mcgill.ca>, Version 2.5 (March 2006)

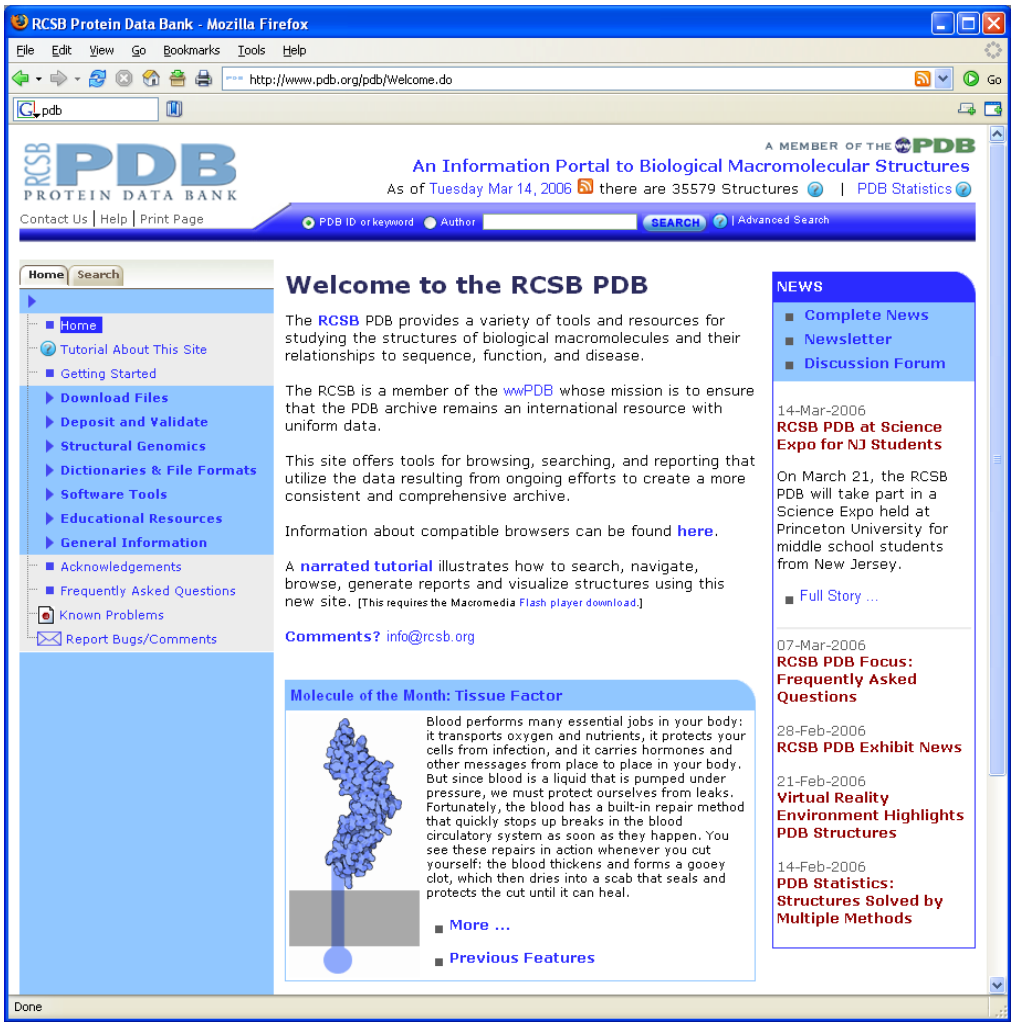
Part 1: Protein Data Bank (PDB) to find structures

PDB homepage: http://www.rcsb.org/pdb (or search "pdb beta" on Google)

[As of September 2005, this tutorial shows the use of the beta site of PDB, found at http://pdbbeta.rcsb.org/pdb/]

PDB is a repository for 3-D structures of biological relevance. Although PDB means Protein Data Bank, it is also a database where you can find structures of nucleic acids and other macromolecules (although proteins are by far the most well-represented category).

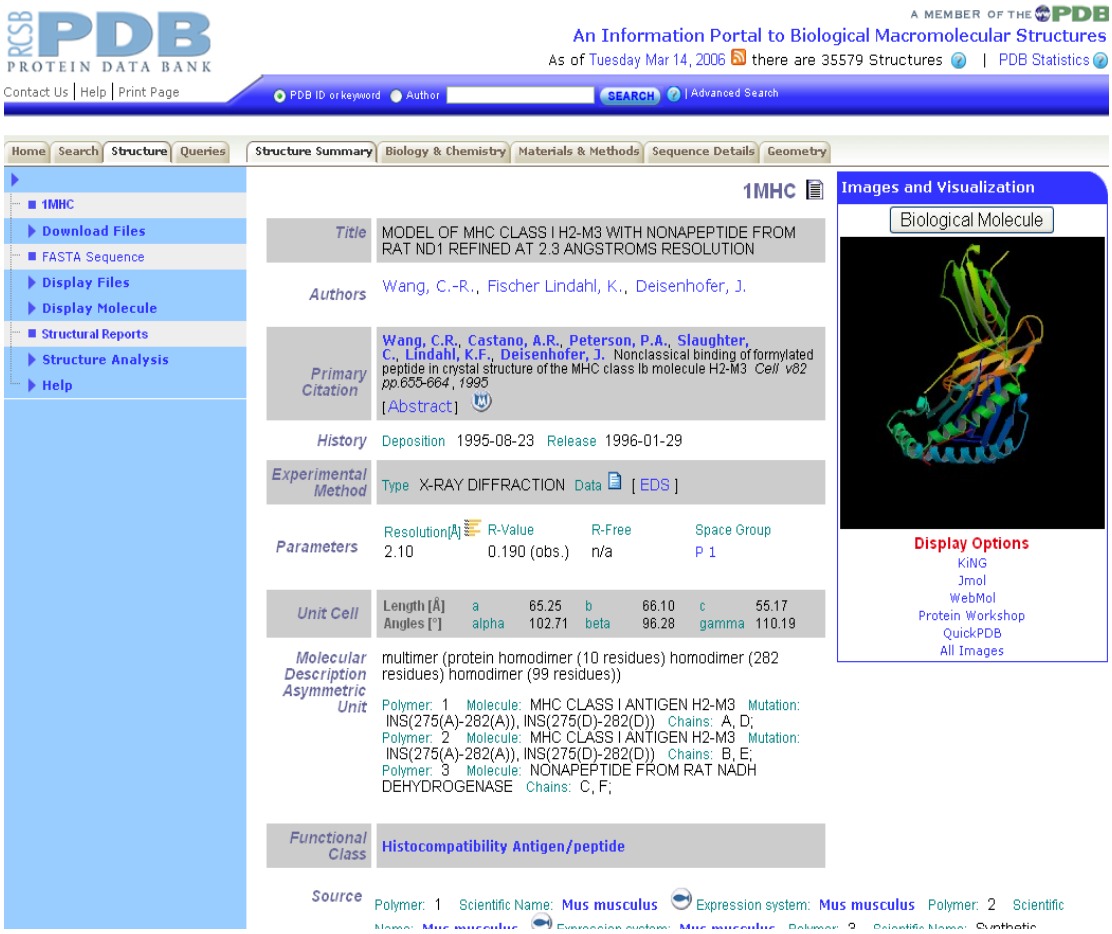
Just to illustrate the use of PDB and molecular visualization tools, we will use a major histocompatibility complex class 1 (MHC-I) molecule from mice throughout this tutorial.



Step 1 [Getting started]: Open a browser window and google “PDB”. Then, click on the first link (the site's address is http://www.rcsb.org/pdb, but it's easier to find through Google).

Step 2 [Searching by PDB ID]: Now that you are on the PDB website, use the main search form at the top of the page to find the struture you want to study. If you know the PDB ID (a unique four-character ID for all structures found in PDB), you may input it in the main search form. Otherwise, you may search the PDB database by keywords and browse the results for a suitable structure.

Step 3 [The Structure Summary Page]: Searching for "1MHC" will lead you to the structure's webpage. Various information is given on the "Structure Summary Page" of each structure in PDB:



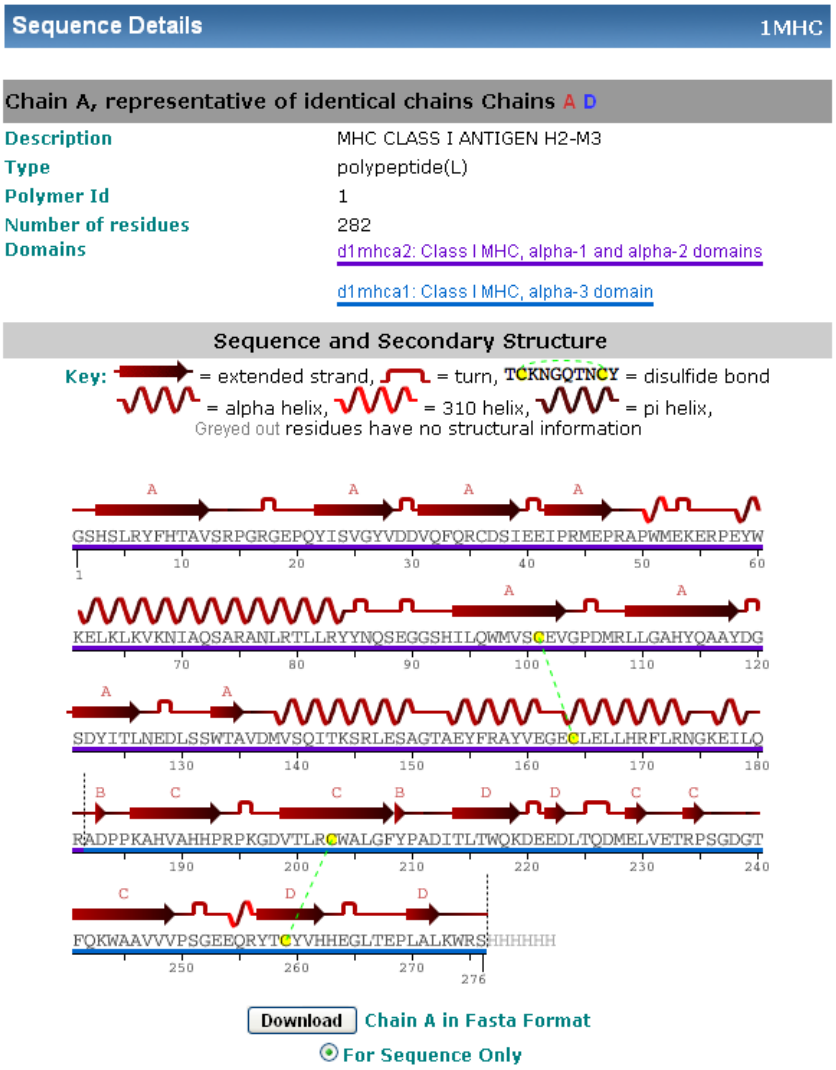
- *Title*: A description of the structure.
- *Primary citation*: Reference published when this structure was submitted.
- *Molecular Description*: A summary of the structure's chains (a single structure can be made of several polypeptide chains).
- *Source*: The organism from which the protein originally comes from, how it was amplified for crystallization, etc.
- SCOP Classification: A manual classification of similar structures into hierarchized categories.

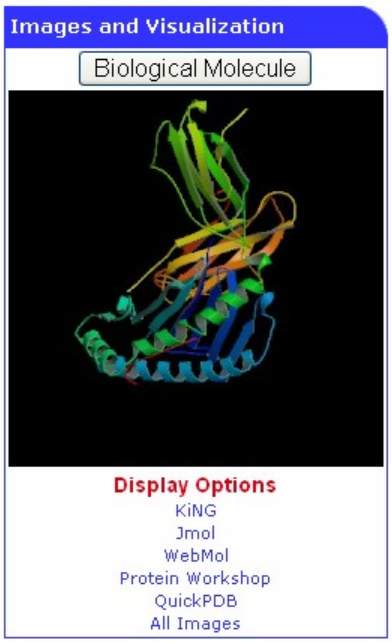
Step 4 [The Structure Explorer bar]: At the top of each PDB entry page, you will also find the "Structure Explorer" bar, which you will use to find more information about a structure, as well to download the structure for viewing in RasMol, an external program which allows you to manipulate a structure, and make cosmetical changes to it.

Step 5 [viewing the structure's sequence]: Before we go download the structure, we will look over some of the features of the Structure Explorer. The first one is called "**Sequence Details**", which shows the amino acid sequence of each chain of the structure file. From the top menu, select *Sequence Details*.

The sequence of the structure can be downloaded in "Fasta" format from the button below the secondary structure overview.

All sorts of data concerning the protein sequence can be found on this page. Occasionally, the structure will have a link going to the corresponding Swiss-Prot page, which would contain curated data on the protein.





Step 6 [viewing the structure]:

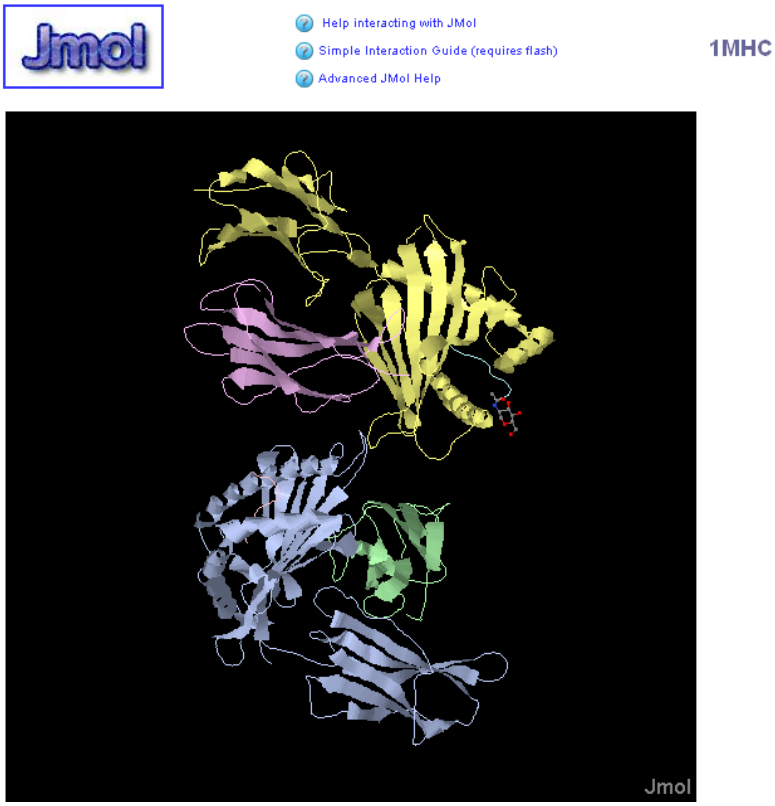
If you are using an external molecular visualization tool, you may choose to download the PDB file from the "Download Files" sub-menu, or under one of the.

However, our preferred mode of viewing would be with one of the web-based viewers, using Jmol for instance, under "Images and Visualization" on the right-side menu.

Part 2: Using Jmol to visualize PDB files

Step 1 [getting the structure]:

From the Structure Explorer page, you have clicked on one of the links leading to Jmol. The applet will load and display the structure in its most basic view, in ribbons without annotation. Each single chain is coloured differently, and non-protein compounds are shown in sticks and balls.



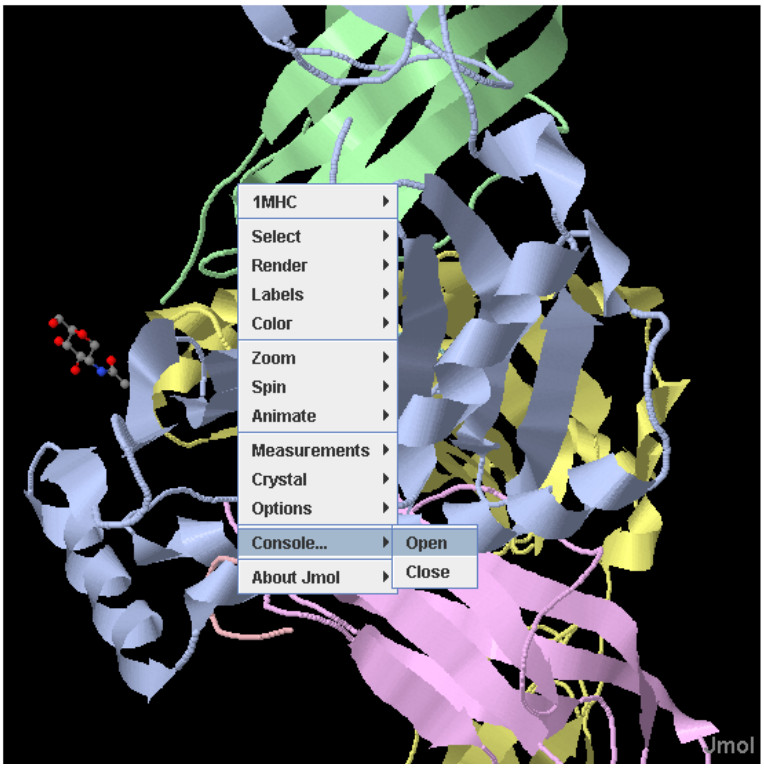
Pressing the middle wheel button allows you to zoom in and out the molecule.

Pressing CTRL and dragging with the right button allows to move the molecule around (translation) within the window.

Double-click on an atom (easier to perform when the molecules are in spacefill mode) will display a meter that will show the distance between it and any other atom you are pointing to (and to subsequent atoms as well).

Single-click on an atom and the status bar of your browser (view>status bar) or the console (if opened) will display details on what you have just clicked.

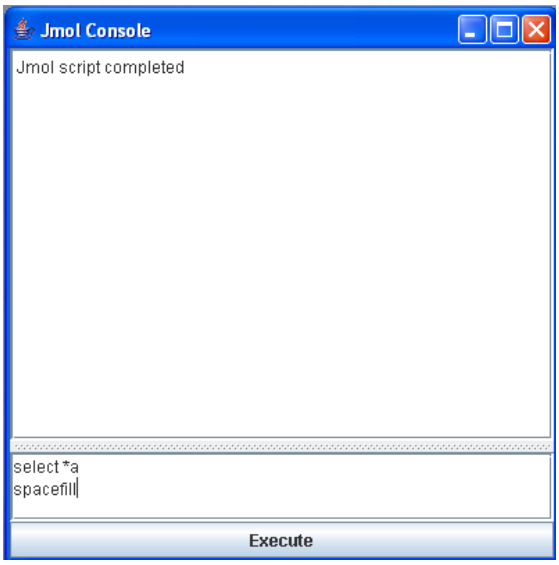
The right-click menu allows you to perform many manipulations on the appearance of the molecule. The context menu can also be access by single-clicking the Jmol logo.



Step 2 [manipulating the image]:

By default, all of the molecule (or sometimes, just the protein chains) are selected. The application keeps in memory what molecules you have selected, and performs the rendering commands on these only.

By using the context menu, you may select a whole group of molecules under the "Select" sub-menu. However, for greater flexibility (for selecting a range of amino acids for instance), you must use the Jmol Console instead.



Step 3 [using the console]:

In our example '1MHC', we know from the Sequence Details of the structure that it made up of four polypeptide chains. By using the Console, we can *select* a particular chain, and then perform various aesthetic changes on the selection. The syntax of the scripting commands is fully described on the Jmol documentation website: http://jmol.sourceforge.net/docs/. (The scripting documentation also shows interactive example of some commands)

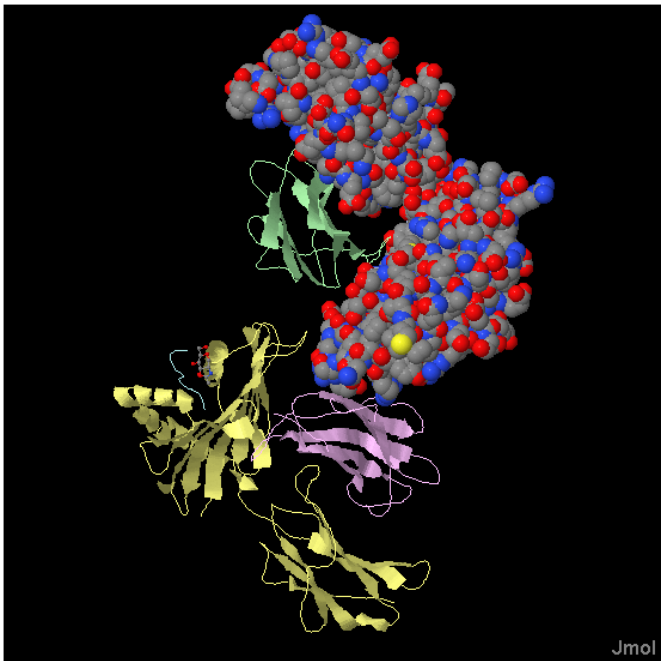
For instance, if we wanted to select all of chain A, and show the electromagnetic contours of the molecules, we would have to execute "select *a" and "spacefill".

We can also decide to **select by amino acid**, by using the 'select' command by appending a number range to the chain letter. For instance, to select amino acids 300 to 400 in chain B, we would call the command "select 300-400b". If successful (if the range delimiters exist - verify with the sequence details page), the console will display that it has just selected a certain number of molecules.

[The chain name is optional, and if omitted, it will select molecules on all chains. This might not be important because many structures have one single polypeptide chain, or many subunits of the same polypeptidic chain.]

To put emphasis on these selected molecules, the user can use various view customization commands, such as 'color <colorname>' and 'spacefill'.

To select co-crystallized compounds (peptides, single nucleic acids, chemicals, solvent, etc), it is more convenient to use the context menu, because their selection name may not be standard. If these chemicals are listed, it will be under "Chemical Component" on the main page.



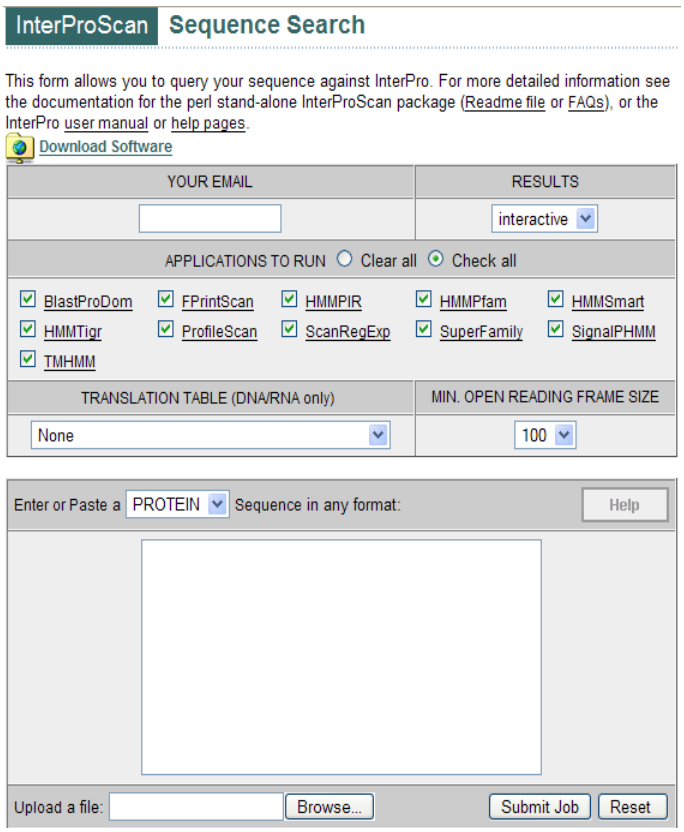
Tutorial: How to use InterPro to find conserved protein domains?

By Cedric Sam <cedric.sam@elf.mcgill.ca> and Abrar Khan <abrar.khan@mail.mcgill.ca>
Version 2, August 2005.

InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.

Part 1: Using InterProScan to search InterPro

Go to **http://www.ebi.ac.uk/InterProScan/** (case-sensitive)

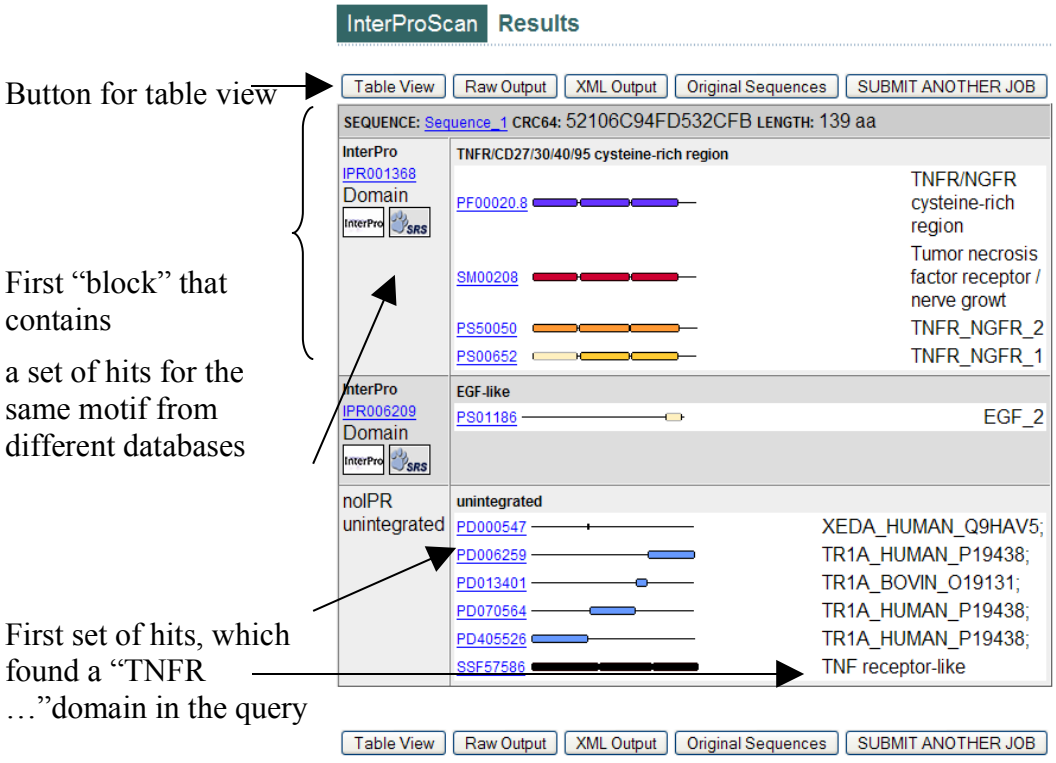


- Enter an e-mail address if you want the results sent to your inbox.
- InterPro integrates data from various Protein Family database, the most notorious of which are ProSite (a product of the Swiss Bioinformatics Institute) and Pfam (originally developed by the Sanger Institute in the UK). It's OK to choose the default options(*).
- This is where you paste your sequence. You would typically use a protein sequence, but the system will take a nucleotide sequence, or even multiple protein sequences. You may also use a file containing all sequences already.

(*) HMMPfam looks in Pfam; ScanRegExp looks in Prosite; TMHMM predicts transmembrane domain; and SignalPHMM predicts the presence of signal peptides.

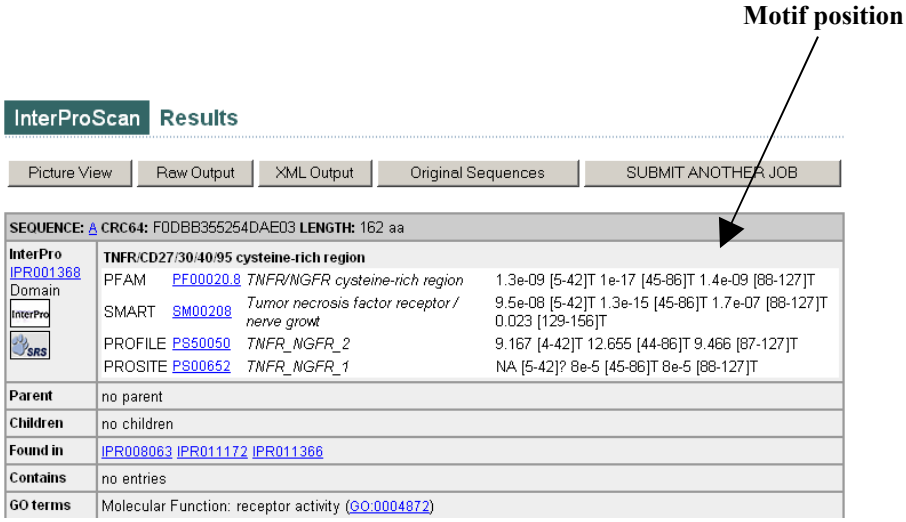
Part 2: Gathering the results

After InterProScan has looked through the database using the programs you selected, you will get a set of results, as shown below.



This picture shows the default "graphical" view of the results. Each block represents a set of hits from several programs/databases for one documented protein domain/family.

In this example, the first hit is for TNFR cysteine-rich domains, which are said in the literature to be repeated four times in members of the TNFR superfamily of receptors, which we used here in our example. Boxes show the relative location of each conserved domain (so, we only see three repeated domains, but this is probably because this is a truncated version of the protein don't appear). If you are using Internet Explorer, you may hover on each rectangle to obtain numerical values for the start-end amino acids of the hit, as well as an "E-value" determining the goodness of the hit (lower is better). If you use a different browser, you may need to click the **Table View** (see figure below) button to see these details. The hits in Table View are sorted by the InterPro accession number, which has the form of IPRXXXXXX with X being a digit.

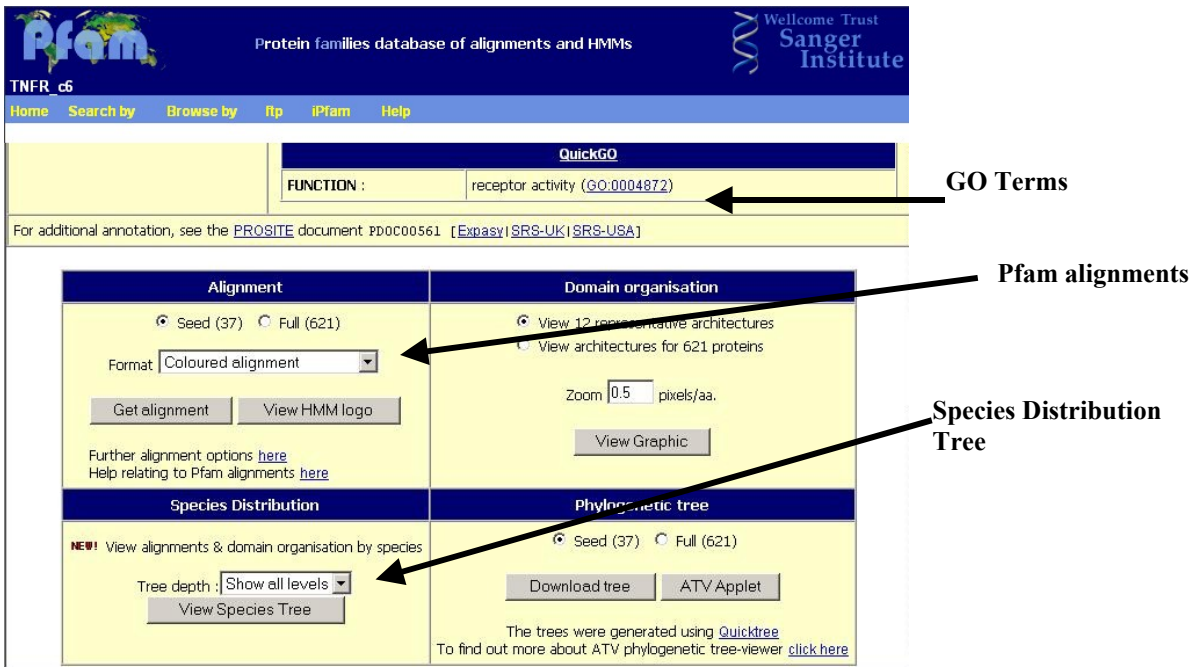


The ID next to each row represents the families found using each different program. Each link leads you to a description of the domain found. To simplify things, we can limit ourselves to the InterPro description (ID starting with "IPR...") since all domains listed in one block are equivalent. Numbers to the right represent the location of the domain within the sequence and the letters beside the numbers signify the status of the hit T for True, F for False positive or (?) for unknown. For hits with Negative (N) and partial (P) status the positions are undefined and cannot be shown in graphical view.

Part 3: Gathering information from the PFAM database

Pfam is a database of protein domain families. Pfam contains curated multiple sequence alignments for each family, as well as profile hidden Markov models (profile HMMs) for finding these domains in new sequences. Pfam contains functional annotation, literature references and database links for each family.

Pfam is a member of the InterPro consortium and has, like the other member databases, contributed annotation and families to the InterPro project. InterPro aims to provide an integrated view of the diverse protein family databases and one of its strengths is that a comprehensive set of annotations has been created through the merging of information from each member.



GO Terms: This link is a browser of the Gene Ontology at the EBI. It is a site that describes gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

Pfam Alignment: This link leads to the output alignment file that Pfam uses to determine the domains within a query sequence

Species Distribution Tree: This link leads to a level-based visualization of the phylogenetic tree, and allows the user to view alignments and& domain organisation by species.

Glossary

(Selected from the 2can glossary on the EBI website: http://www.ebi.ac.uk/2can/glossary/)

A

Accession number: An identifier supplied by the curators of the major biological databases upon submission of a novel entry that uniquely identifies that sequence (or other) entry.

Algorithm: A series of steps defining a procedure or formula for solving a problem, which can be coded into a programming language and executed. Bioinformatics algorithms typically are used to process, store, analyze, visualize and make predictions from biological data.

Analogy: Reasoning by which the function of a novel gene or protein sequence may be deduced from comparisons with other gene or protein sequences of known function. Identifying analogous or homologous genes via similarity searching and alignment is one of the chief uses of Bioinformatics.

Annotation: A combination of comments, notations, references, and citations, either in free format or utilizing a controlled vocabulary, that together describe all the experimental and inferred information about a gene or protein. Annotations can also be applied to the description of other biological systems. Batch, automated annotation of bulk biological sequence is one of the key uses of Bioinformatics tools.

B

Bioinformatics:
1.The field of endeavor that relates to the collection, organization and analysis of large amounts of biological data using networks of computers and databases (usually with reference to the genome project and DNA sequence information).
2. Computational biology, sometimes, is used interchangeably with the term

C

Cluster: The grouping of similar objects in a multidimensional space. Clustering is used for constructing new features which are abstractions of the existing features of those objects. The quality of the clustering depends crucially on the distance metric in the space. In bioinformatics, clustering is performed on sequences, high-throughput expression and other experimental data. Clusters of partial or complete gene sequences can be used to identify the complete (contiguous) sequence and to better identify its function. Clustering expression data enables the researcher to discern patterns of co-regulation in groups of genes.

Complexity (of gene sequence): The term "low complexity sequence" may be thought of as synonymous with regions of locally biased amino acid composition. In these regions, the sequence composition deviates from the random model that underlies the calculation of the statistical significance (P-value) of an alignment. Such alignments among low complexity sequences are statistically but not biologically significant, i.e., one cannot infer homology (common ancestry) or functional similarity.

Conformation: The precise three-dimensional arrangement of atoms and bonds in a molecule describing its geometry and hence its molecular function.

Consensus sequence: A single sequence delineated from an alignment of multiple constituent sequences that represents a "best fit" for all those sequences. A "voting" or other selection procedure is used to determine which residue (nucleotide or amino acid) is placed at a given position in the event that not all of the constituent sequences have the identical residue at that position.

D

Database: Any file system by which data gets stored following a logical process.

Deletion: A chromosomal alteration in which a portion of the chromosome or the underlying DNA is lost.

Domain (protein): A region of special biological interest within a single protein sequence. However, a domain may also be defined as a region within the three-dimensional structure of a protein that may encompass regions of several distinct protein sequences that accomplishes a specific function. A domain class is a group of domains that share a common set of well-defined properties or characteristics.

E

F

FASTA format: A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length. An example sequence in FASTA format is:

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRT
QIWQKHRTSNDSALILLNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWC
HFPSNWKGAWKEVKEEIVNLPKERYRGTNDPKRIFFQRQWGDPETANLWFNCHGEFFYCK
MDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVACHIRSVIIWLETISKK
TYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGF
APTEVRRYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNL
LAAVEAQQQMLKLTIWGVK
```

A FASTA file can also contain multiple sequences:

```
>VECTOR32    Synthetic vector sequence #32
ATGAGCGGCGGCCCCATGGGCGGCAGGCCCGGCGGCAGGGGCGCCCCCGCCGTGCAGCAG
AACATCCCCAGCACCCTGCTGCAGGACCACGAGAACCAGAGGCTGTTCGAGATGCTGGGC
>VECTOR33    Synthetic vector sequence #33
ACGAGCGGCGGTCCCATGGGCGCCAGGCCCGGCGGCAGGGGCGCTGCCGCCGTGCAGCAC
ATCATCCCCAGCACCCTGCAGCAGGACCACGAGTACCAGAGGCTGTTCGAGATGCTGGGC
>VECTOR34    Synthetic vector sequence #34
GTGAGCGGCGGCTACTTGGGCGGCAGGCCCGGCGGCAGGGGCGCCCACGCCGTGCAGCAG
```

Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes with these exceptions: lower-case letters are accepted and are mapped into upper-case; a single hyphen or dash can be used to represent a gap of indeterminate length; and in amino acid sequences, U and * are acceptable letters (see below). Invalid characters (digits, blanks) are automatically removed.

Frameshift: A deletion, substitution, or duplication of one or more bases that causes the reading-frame of a structural gene to shift from the normal series of triplets.

G

Gaps (affine gaps): A gap is defined as any maximal, consecutive run of spaces in a single string of a given alignment. Gaps help create alignments that better conform to underlying biological models and more closely fit patterns that one expects to find in meaningful alignment. The idea is to take in account the number of continuous gaps and not only the number of spaces when calculating an alignment. Affine gaps contain a component for gap insertion and a component for gap extension, where the extension penalty is usually much lower than the insertion penalty. This mimics biological reality as multiple gaps would imply multiple mutations, but a single mutation can lead to a long gap quite easily.

Gap penalties: The penalty applied to a similarity score for the introduction of an insertion or deletion gap, the extension of a gap, or both. Gap penalties are usually subtracted from a cumulative score being determined for the comparison of two or more sequences via an optimization algorithm that attempts to maximize that score.

Gene: Classically, a unit of inheritance. In practice, a gene is a segment of DNA on a chromosome that encodes a protein and all the regulatory sequences (promoter) required to control expression of that protein.

Gene families: Subsets of genes containing homologous sequences which usually correlate with a common function.

H

Heterodimer: Protein composed of 2 different chains or subunits.

Homeobox: A highly conserved region in a homeotic gene composed of 180 bases (60 amino acids) that specifies a protein domain (the homeodomain) that serves as a master genetic regulatory element in cell differentiation during development in species as diverse as worms, fruitflies, and humans.

Homeodomain: A 60 amino-acid protein domain coded for by the homeobox region of a homeotic gene.

Homology: (strict) Two or more biological species, systems or molecules that share a common evolutionary ancestor. (general) Two or more gene or protein sequences that share a significant degree of similarity, typically measured by the amount of identity (in the case of DNA), or conservative replacements (in the case of protein), that they register along their lengths. Sequence "homology" searches are typically performed with a query DNA or protein sequence to identify known genes or gene products that share significant similarity and hence might inform on the ancestry, heritage and possible function of the query gene.

I

in silico (biology): (Lit. computer mediated). The use of computers to simulate, process, or analyse a biological experiment.

Iteration: A series of steps in an algorithm whereby the processing of data is performed repetitively until the result exceeds a particular threshold. Iteration is often used in multiple sequence alignments whereby each set of pairwise alignments are compared with every other, starting with the most similar pairs and progressing to the least similar, until there are no longer any sequence-pairs remaining to be aligned.

J

Junk DNA: Term used to describe the excess DNA that is present in the genome beyond that required to encode proteins. A misleading term since these regions are likely to be involved in gene regulation, and other as yet unidentified functions.

K

L

Library: A large collection of compounds, peptides, cDNAs or genes which may be screened in order to isolate cognate molecules.

M

Map unit: A measure of genetic distance between two linked genes that corresponds to a recombination frequency of 1%.

Motif: A conserved element of a protein sequence alignment that usually correlates with a particular function. Motifs are generated from a local multiple protein sequence alignment corresponding to a region whose function or structure is known. It is sufficient that it is conserved, and is hence likely to be predictive of any subsequent occurrence of such a structural/functional region in any other novel protein sequence.

Multigene family: A set of genes derived by duplication of an ancestral gene, followed by independent mutational events resulting in a series of independent genes either clustered together on a chromosome or dispersed throughout the genome.

Multiple (sequence) alignment: A Multiple Alignment of k sequences is a rectangular array, consisting of characters taken from the alphabet A, that satisfies the following conditions: There are exactly k rows; ignoring the gap character, row number i is exactly the sequence sI; and each column contains at least one character different from "-". In practice multiple sequence alignments include a cost/weight function, that defines the penalty for the insertion of gaps (the "-" character) and weights identities and conservative substitutions accordingly. Multiple alignment algorithms attempt to create the optimal alignment defined as the one with the lowest cost/weight score.

Mutation: An inheritable alteration to the genome that includes genetic (point or single base) changes, or larger scale alterations such as chromosomal deletions or rearrangements.

N

Naked DNA: Pure, isolated DNA devoid of any proteins that may bind to it.

O

Open reading frame (ORF): Any stretch of DNA that potentially encodes a protein. Open reading frames start with a start codon, and end with a termination codon. No termination codons may be present internally. The identification of an ORF is the first indication that a segment of DNA may be part of a functional gene.

Operator: A segment of DNA that interacts with the products of regulatory genes and facilitates the transcription of one or more structural genes.

Operon: A unit of transcription consisting of one or more structural genes, an operator, and a promoter.

Ortholog: Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes. (See also Paralogs.)

P

PAM matrix: PAM (percent accepted mutation) and BLOSUM (blocks substitution matrix) are matricies that define scores for each of the 210 possible amino acid substitutions. The scores are based on empirical substitution frequencies observed in alignments of database sequences and in general reflect similar physiochemical properties (e.g. a substitution of leucine for isoleucine, two amino acids of similar hydrophobicity and size, will score higher than a substitution of leucine for glutamine)

Paralog: Paralogs are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

Parameters: Parameters are user-selectable values, typically experimentally determined, that govern the boundaries of an algorithm or program. For instance, selection of the appropriate input parameters governs the success of a search algorithm. Some of the most common search parameters in bioinformatics tools include the stringency of an alignment search tool, and the weights (penalties) provided for mismatches and gaps.

Protein families: Sets of proteins that share a common evolutionary origin reflected by their relatedness in function which is usually reflected by similarities in sequence, or in primary, secondary or tertiary structure. Subsets of proteins with related structure and function.

Q

Query (sequence): A DNA, RNA of protein sequence used to search a sequence database in order to identify close or remote family members (homologs) of known function, or sequences with similar active sites or regions (analogs), from whom the function of the query may be deduced.

R

Reading frame: A sequence of codons beginning with an intiation codon and ending with a termination codon,

typically of at least 150 bases (50 amino acids) coding for a polypeptide or protein chain (see ORF and URF).

Repeats (repeat sequences): Repeat sequences and approximate repeats occur throughout the DNA of higher organisms (mammals). For example, the Alu sequences of length about 300 characters, appear hundreds of thousands of times in Human DNA with about 87% homology to a consensus Alu string. Some short substrings such as TATA-boxes, poly-A and (TG)* also appear more often than by chance. Repeat sequences may also occur within genes, as mutations or alterations to those genes. Repetitive sequences, especially mobile elements, have many applications in genetic research. DNA transposons and retroposons are routinely used for insertional mutagenesis, gene mapping, gene tagging, and gene transfer in several model systems.

Repetitive elements: Repetitive elements provide important clues about chromosome dynamics, evolutionary forces, and mechanisms for exchange of genetic information between organisms The most ubiquitous class of repetitive elements in the DNA sequence in primate genomes is the Alu family of interspersed repeats which have arisen in the last 65 million years of evolution Alu repeats belong to a class of sequences defined as short interspersed elements (SINEs). Approximately 500,000 Alu SINEs exist within the human genome, representing about 5% of the genome by mass.

S

Selectivity: Selectivity of bioinformatics similarity search algorithms is defined as the significance threshold for reporting database sequence matches. As an example, for BLAST searches, the parameter E is interpreted as the upper bound on the expected frequency of chance occurrence of a match within the context of the entire database search. E may be thought of as the number of matches one expects to observe by chance alone during the database search.

Sensitivity: Sensitivity of bioinformatics similarity search algorithms centers around two areas: First, how well can the method detect biologically meaningful relationships between two related sequences in the presence of mutations and sequencing errors; Secondly how does the heuristic nature of the algorithm affect the probability that a matching sequence will not be detected. At the user's discretion, the speed of most similarity search programs can be sacrificed in exchange for greater sensitivity - with an emphasis on detecting lower scoring matches.

Similarity (homology) search: Given a newly sequenced gene, there are two main approaches to the prediction of structure and function from the amino acid sequence. Homology methods are the most powerful and are based on the detection of significant extended sequence similarity to a protein of known structure, or of a sequence pattern characteristic of a protein family. Statistical methods are less successful but more general and are based on the derivation of structural preference values for single residues, pairs of residues, short oligopeptides or short sequence patterns. The transfer of structure/function information to a potentially homologous protein is straightforward when the sequence similarity is high and extended in length, but the assessment of the structural significance of sequence similarity can be difficult when sequence similarity is weak or restricted to a short region.

Structure prediction: Algorithms that predict the secondary, tertiary and sometimes even quarternary structure of proteins from their sequences. Determining protein structure from sequence has been dubbed "the second half of the Genetic Code" since it is the folded tertiary structure of a protein that governs how it functions as a gene product. As yet most structure prediction methods are only partially successful, and typically work best for certain well-defined classes of proteins.

Substitution matrix: A model of protein evolution at the sequence level resulting in the development of a set of widely used substitution matrices. These are frequently called Dayhoff, MDM (Mutation Data Matrix), BLOSUM or PAM (Percent Accepted Mutation) matrices. They are derived from global alignments of closely related sequences. Matrices for greater evolutionary distances are extrapolated from those for lesser ones.

T

U

Unidentified reading frame (URF): An open reading frame encoding a protein of undefined function.

V

Variation (genetic): Variation in genetic sequences and the detection of DNA sequence variants genome-wide allow studies relating the distribution of sequence variation to a population history. This in turn allows one to determine the density of SNPS or other markers needed for gene mapping studies. Quantitation of these variations together with analytical tools for studying sequence variation also relate genetic variations to phenotype.

W

Weight matrix: The density of binding sites in a gene or sequence can be used to derive a ratio of density for each element in a pattern of interest. The combined individual density ratios of all elements are then collectively used to build a scoring profile known as a weight matrix. This profile can be used to test the prediction of the identification of the selected pattern and the ability of the algorithm to discriminate them from non-pattern sequences.

X

Y

Z