Computer / Web Resources

Here are four web sites you should be familiar with:

I. NCBI - National Center for Biotechnology Information

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned:

> development of new algorithms and statistics with which to assess relationships among members of large data sets;
>
> analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures;
>
> development and implementation of tools that enable efficient access and management of different types of information

Entrez is a search and retrieval system that integrates information from databases at NCBI. These databases include nucleotide sequences, protein sequences, macromolecular structures, whole genomes, and **MEDLINE**, through **PubMed**.

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
```

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). Before

submitting a request, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes (e.g., N for unknown nucleic acid residue or X for unknown amino acid residue).
The nucleic acid codes supported are:

A --> adenosine	M --> A C (amino)
C --> cytidine	S --> G C (strong)
G --> guanine	W --> A T (weak)
T --> thymidine	B --> G T C
U --> uridine	D --> G A T
R --> G A (purine)	H --> A C T
Y --> T C (pyrimidine)	V --> G C A
K --> G T (keto)	N --> A G C T (any)
	- gap of indeterminate length

For those programs that use amino acid query sequences (BLASTP and TBLASTN), the accepted amino acid codes are:

A	alanine	P	proline
B	aspartate or asparagine	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	U	selenocysteine
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate or glutamine
L	leucine	X	any
M	methionine	*	translation stop
N	asparagine	-	gap of indeterminate length

BLAST (Basic Local Alignment Search Tool)

BLAST® (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. BLAST uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity (Altschul et al., 1990).

The **Expect value (E)** is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size. It decreases exponentially with the **Score (S) that is assigned to a match between two sequences.** Essentially, the **E value describes the random background noise that exists for matches between sequences.**

A. Nucleotide BLAST searches allow one to input nucleotide sequences and compare these against other nucleotides.

Standard nucleotide-nucleotide BLAST - Takes nucleotides sequences in FASTA format, GenBank Accession numbers or GI numbers and compares them against the NCBI nucleotide databases.

B. Protein BLAST allows one to input protein sequences and compare these against other protein sequences.

Standard protein-protein BLAST - Takes protein sequences in FASTA format, GenBank Accession numbers or GI numbers and compares them against the NCBI protein databases.

PSI-BLAST - Position Specific Iterated BLAST uses an iterative search in which sequences found in one round of searching are used to build a score model for the next round of searching. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform a second (etc.) BLAST search and the results of each "iteration" used to refine the profile. This iterative searching strategy results in increased sensitivity.

Pairwise BLAST performs a comparison between two sequences using the BLAST algorithm. Not that the program considers a "Sequence 1" to be the Query sequence and "Sequence 2" to be the Subject sequence. There are the following program options:

blastn - for nucleotide - nucleotide comparisons

blastp - for protein - protein comparisons

tblastn - compares the protein "Sequence 1" against the nucleotide "Sequence 2" which has been translated in all six reading frames

blastx - compares the nucleotide "Sequence 1" against the protein "Sequence 2"

tblastx - compares nucleotide "Sequence 1" translated in all six reading frames against the nucleotide "Sequence 2" translated in all six reading frames.

You can select several NCBI databases to compare your query sequences against. Note that some databases are specific to proteins or nucleotides and cannot be used in combination with certain BLAST programs (for example a blastn search against swissprot).

Proteins

Database	Description
nr	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 (no updates). These are uploaded to our system when they are received from EMBL.
patents	Protein sequences derived from the Patent division of GenBank.
yeast	Yeast (Saccharomyces cerevisiae) protein sequences. This database is not to be confused with a listing of all Yeast protein sequences. It is a database of the protein translations of the Yeast complete genome.
E. coli	E. coli (Escherichia coli) genomic CDS translations.
pdb	Sequences derived from the 3-dimensional structure Brookhaven Protein Data Bank.

Introduction to BLAST Output

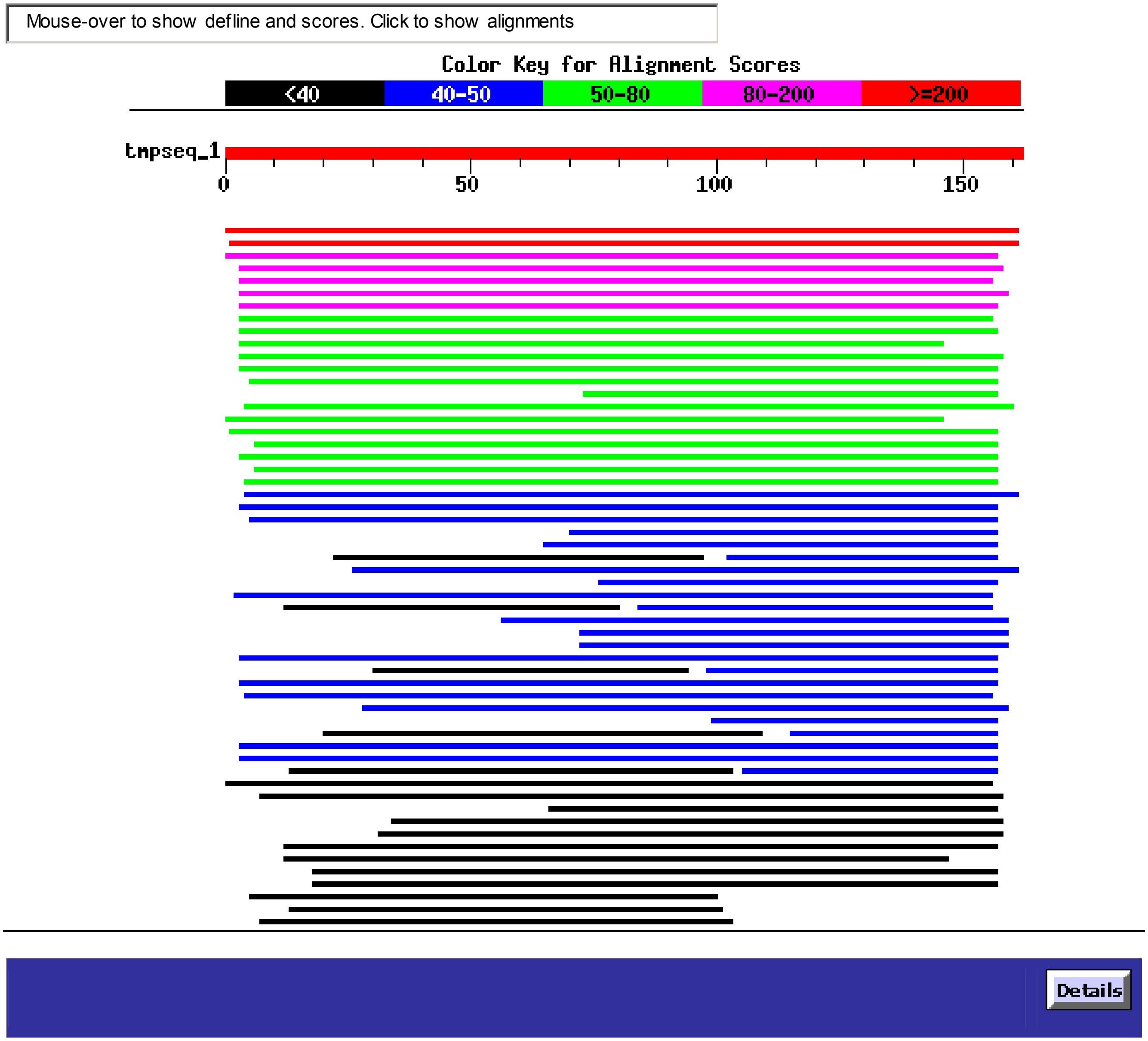
These are the results of a BLAST search of the non-redundant database using the uncharacterized protein, MJ0577, from *Methanococcus jannashii* as the query sequence. Listed below are each of the elements of BLAST output in their usual order. Explanatory notes have been added in light grey boxes. Scroll down the page and learn how to analyze BLAST output, step by step.

Query: = gi|2501594|sp|Q57997|Y577_METJA PROTEIN MJ0577 (162 letters)
Database: Non-redundant GenBank CDS translations+PDB+SwissProt+SPupdate+PIR 437,713 sequences; 134,605,311 total letters

Step 2. Graphical overview of database matches exceeding the E value threshold set in the query.

From the graphical overview, it is apparent that there are several high-scoring sequences that are highly related to the MJ0577 query. The top two bars (red) are matches to MJ0577 itself. The next five (pink) hits are probable homologs. The full-length green, blue and black bars may represent more distantly related homologs. Partial length bars reflect similarity between a portion of the query and a database entry. These regions of similarity may be a shared domain or motif. The nature of these lower scoring hits can be further explored by examining the corresponding alignment. To facilitate viewing any alignment of interest, each bar in the graphic and each E value in a description line (which are found below the graphic) is linked (click on it) to the corresponding alignment. Note, however, if the formatting options were set such that the number of alignments to be displayed is less than the number of descriptions (e.g. 100 descriptions; 50 alignments are the default values, and the values used in this tutorial), not all the links will be functional. If alignments for descriptions 51-100 are of interest, the number of alignments can be changed from 50 to 100 on the intermediate BLAST queue page. To see the new alignments, use the "format results" button once again.

Distribution of 95 Blast Hits on the Query Sequence



Mouse-over to show defline and scores. Click to show alignments

The description (also called definition) lines are listed below under the heading "Sequences producing significant alignments". The term "significant" simply refers to all those hits whose E value was less than the threshold. It does not imply *biological* significance.

Sequences producing significant alignments: Score(bits) E Value

```
1.  sp|Q57997|Y577_METJA  PROTEIN MJ0577 >gi|2128018|pir||A64372...   314  2e-85
2.  pdb|1MJH|    Structure-Based Assignment Of The Biochemical F...   272  1e-72
```

The description lines reveal that the sequence in the database with greatest similarity to MJ0577 is MJ0577 itself. The second hit is to the database entry associated with the determination of the MJ0577 structure. The score for the structure entry is somewhat lower and E value somewhat higher because (see pairwise alignment - the missing residues will appear as dashes) a certain number of residues were omitted from this database entry because they were disordered in the structure. To further evaluate any hits of potential interest, examine the corresponding alignments. Click on the score to the right of the description line to jump down to the alignment for any hits of interest.

```
3.  dbj|BAA29916|   (AP000003) 170aa long hypothetical protein [P...   107  6e-23
4.  sp|Q57951|Y531_METJA  HYPOTHETICAL PROTEIN MJ0531 >gi|212801...    91  4e-18
7.  gi|2621194   (AE000803) conserved protein [Methanobacterium t...    80  7e-15
```

The set of entries corresponding to the pink bars in the overview are to orthologous sequences in two other Archaeal species, *Pyrococcus horikoshii* (#3) and *Methanobacterium thermoautotrophicum* (#5,6,7)). Interestingly, the top hits also include sets of paralogs in *Methanococcus jannaschii* (#1,4) and *Methanobacterium thermoautotrophicum* (#5,6,7). To examine the relationship among these sequences, they have been subject to multiple alignment using the multiple alignment algorithm, ClustalW. The alignment can be seen here. A multiple alignment can also be generated using the "flat query-anchored with identities" formatting option in BLAST. A multiple alignment generated by BLAST is simply a compilation of all pairwise alignments, whereas Clustal W uses a progressive approach, first aligning the most similar sequences, and then adding more distant sequences to the alignment. Therefore, the two types of alignments may not be identical. Consult the Details button for more on how to generate a multiple alignment from within BLAST.

```
8.  gi|2622163   (AE000877) conserved protein [Methanobacterium t...    79  2e-14
9.  sp|P42297|YXIE_BACSU  HYPOTHETICAL 15.9 KD PROTEIN IN BGLH-W...    76  1e-13
10. sp|Q50777|YB54_METTM  HYPOTHETICAL 16.1 KD PROTEIN IN MTR RE...    66  2e-10
11. gi|2648791   (AE000981) conserved hypothetical protein [Archa...    65  3e-10
20. sp|P39177|UP12_ECOLI  UNKNOWN PROTEIN FROM 2D-PAGE (SPOTS PR...    55  2e-07
21. sp|P74148|YD88_SYNY3  HYPOTHETICAL 17.3 KD PROTEIN SLL1388 >...    52  2e-06
```

The set of entries corresponding to the green bars in the overview are to more distantly related Archaeal sequences. The scores and E values are respectable and most of the alignments extend the length of the query.

Since all of the closest homologs of MJ0577 are unannotated, it is of interest to examine a few of the highest scoring annotated sequence in the BLAST hit list. The first sequences in the hit list (scroll down to find these) with positive identification are gi|2648945, whose annotation indicates that this sequence encodes a cationic amino acid transporter and gi|1787640, a putative filament protein. The scores and E values of these hits put the significance of the alignments in the "twilight zone" with respect to significance. Since the transporter entry is 780 aa long, it is apparent that MJ0577 is not itself a homolog of a transporter. It may however, share a domain with this family of proteins. The filament protein is harder to dismiss, however manual filtering of the coiled-coil region in the query (MJ0577) causes the filament protein hit to drop off the hit list (E > 10). This could be an indication that the sequence similarity between MJ0577 and the putative filament protein is non-specific.

The results of this BLAST search have revealed that MJ0577 is one member of a moderate size family of proteins that are found in Archaea and in Eubacteria. To gain insight into the function of the MJ0577 family of proteins, a more sensitive search with the profile based tool, PSI-BLAST, would be one way to proceed.

```
sp|P45680|YFMU_COXBU  HYPOTHETICAL 15.8 KD PROTEIN IN FMU-RP...    49  2e-05
gb|AAF11689.1|AE002048_9  (AE002048) hypothetical protein [D...    48  3e-05
gi|2160182  (AC000132) ESTs gb|ATTS1236,gb|T43334,gb|N97019,...    45  3e-04
gb|AAF11911.1|AE002067_3  (AE002067) conserved hypothetical ...    43  0.001
sp|P44195|YDAA_HAEIN  HYPOTHETICAL PROTEIN HI1426                  41  0.005


gi|1787640  (AE000234) putative filament protein [Escherichi...    41  0.005
gi|2649611  (AE001036) conserved hypothetical protein [Archa...    37  0.073
gi|3493653  (AF083219) unknown [Azospirillum brasilense]           37  0.096
emb|CAB53147.1|  (AL109962) hypothetical protein SCJ1.29c [S...    36  0.17
gb|AAD07556.1|  (AE000563) H. pylori predicted coding region...    34  0.83
```

The description list is truncated at E = 1.0, as requested on the query page. It is interesting to note that even had the E value been left at the default of 10, the list would have truncated at E = 1.9. In this case the list would have been limited by our selection of 100 as the number of description lines to report. When hits with higher E values are of interest, they can be viewed by returning to the query page and resetting both the E value and the description options.

Step 4. Pairwise alignments

Below are the alignments of each "significant" hit to the query. These are shown below in pairwise format, because that option was selected while setting up the query. The query-anchored format is another useful way to inspect the relationship of hits to a query. There will be as many alignments here as were specified on the query page. One efficient way to inspect alignments of interest is to use the links from specific entries from within the graphic or from within the list of descriptions. It may also be useful to scroll through the alignments to inspect the overall match quality. For example, if the aligned residues are all hydrophobic, it may indicate that a transmembrane or coiled-coil domain in the query is causing non-specific hits. Scroll through the alignments now, or skip over the alignments and go to Step 5.

```
  sp|Q57997|Y577_METJA  MJ0577 - Methanococcus jannaschii
>gi|5107801|pdb|1MJH|A
            Chain A, Structure-Based Assignment Of The Biochemical
            Function Of Hypothetical Protein Mj0577: A Test Case Of
            Structural Genomics >gi|5107802|pdb|1MJH|B Chain B,
            Structure-Based Assignment Of The Biochemical Function
            Of Hypothetical Protein Mj0577: A Test Case Of
```

```
           Structural Genomics >gi|1591284 (U67506) conserved
           hypothetical protein [Methanococcus jannaschii]
           Length = 162

 Score =  314 bits (796), Expect = 2e-85

 Identities = 162/162 (100%), Positives = 162/162 (100%)


Query: 1   MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA 60
           MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA
Sbjct: 1   MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA 60

Query: 61  GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEG 120
           GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEG
Sbjct: 61  GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEG 120

Query: 121 VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS 162
           VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS
Sbjct: 121 VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS 162
 pdb|1MJH|   Structure-Based Assignment Of The Biochemical Function Of
           Hypothetical Protein Mj0577: A Test Case Of Structural
           Genomics
           Length = 287

 Score =  272 bits (687), Expect = 1e-72
 Identities = 145/161 (90%), Positives = 145/161 (90%), Gaps = 16/161 (9%)

Query: 2   SVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVAG 61
           SVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIK
Sbjct: 143 SVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIK------------- 189
```

Step 5.   Beyond BLAST

The absence of annotated hits among the list of sequences bearing similarity to MJ0577 is unsatisfying. To continue the search for clues to the function of MJ0577 weak, yet significant alignments may be sought using profile-based searching available using PSI-BLAST. Find out more about MJ0577 function in the PSI-BLAST tutorial.

****

PSI – BLAST

Position specific iterative BLAST (PSI-BLAST) refers to a feature of BLAST 2.0 in which a profile (or position specific scoring matrix, PSSM) is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search. The PSSM is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform a second (etc.) BLAST search and the results of each "iteration" used to refine the profile. This iterative searching strategy results in increased sensitivity.

This PSI-BLAST tutorial uses as an example, the uncharacterized archaebacterial protein, MJ0577, from Methanococcus jannaschii. The tutorial illustrates the potential for PSI-BLAST searches to identify even weak (subtle) homologies to annotated entries in the database. It demonstrates that PSI-BLAST is an important tool for predicting both biochemical activities and function from sequence relationships.

A BLAST search of this sequence revealed a number of probable homologs but no hits that provided useful information about the function or biochemical activities of this protein. Since this PSI-BLAST tutorial builds on the example in the BLAST tutorial, it is most useful to tackle the two tutorials in order. Go back to the BLAST tutorial now, if you missed it before.

The PSI-BLAST search in this tutorial has two purposes:
(1) to identify distant relatives of the MJ0577 family and
(2) to gain insight into the function of this family of proteins.

II. Computational Services at EMBL (European Molecular Biology Laboratory)

Sequence Annotation

ensembl Automatic annotation on eukaryotic genomes. Human data are available.

Sequence Search & Retrieval

Bioccelerator	Fast Smith-Waterman/Profile/Frameshifting Searches (e.g. for ESTs)
BLAST2	Advanced **BLAST2** Search Server
Gene2EST	BLAST2 server for searching EST databases
Peptide Search	Protein identification by peptide mapping or peptide sequencing
PROPSEARCH	Compositional search using a sequence
PROPSEARCH	Compositional search using experimental Amino Acid Analysis data
SMART	Simple Modular Architecture Research Tool: domain annotation in proteins.
SRS5	Network Browser for Databanks in Molecular Biology. Fulltext search in EMBL, GENBANK, ...
STRING	Search Tool for Recurring Instances of Neighbouring Genes

Additional sequence search and data retrieval tools at the EBI (FASTA, BLAST, PROSITE,...)
Mutations databases/links at the EBI

EMBL database files via FTP

Sequence Alignment

ClustalW	Multiple Sequence Alignment (at the EBI)

Some multiple alignments are available via ftp at the EBI (submission).

Sequence Databases

(To search the EMBL DNA databases please see the Sequence search tools)

EMBL	sequence retrieval by accession number (at EBI)

EMBL sequence database

The EMBL DNA sequence database is maintained at the EMBL outstation EBI.

- simple entry retrieval by accession number
- ftp access to the EMBL database files
- sequence submission

Structure Comparison

DALI	Comparison of protein structures in 3D (at EBI)

Structure Prediction

AGADIR	An algorithm to predict the helical content of peptides
FOLD-X	A tool to calculate the folding pathways of proteins and the effect of a point mutation on the stability of a protein
PredictProtein	The PredictProtein server **EvalSec** evaluation of prediction accuracy **PHDsec** prediction of secondary structure **PHDacc** prediction of residue solvent accessibility **PHDhtm** prediction of transmembrane helices **PHDhtm** prediction of transmembrane helices **PHDthreader** prediction-based threading
SSCP	Secondary structural content prediction from amino acid composition

StructureVerification

BIOTECH	Protein structure verification

Information Services

- Genequiz: automated analysis of biological sequences. Analysis of the complete genomes of yeast, Methanococcus jannaschii, Haemophilus influenzae, Mycoplasma genitalium (at EBI)
- EMBL Anonymous FTP server (ftp.embl-heidelberg.de)
- Cell Biophysics Programme Anonymous FTP server (ftp.pi.embl-heidelberg.de)
- Molecular Biology Databases at ftp.embl-heidelberg.de/pub/databases
- EMBL Virus Structure Resource
- Sequence Alerting sequence alerting system searches in several databases for news on (homologues of) a sequence and informs by email, if it has detected a new relative

Local Servers

NMR Group	

Software

ADS	Analytically Defined molecular Surfaces
ECM	Effective (potential derived) Charges for Macromolecules in solvent
LIGIN	Molecular docking using surface complementarity
SDA	Simulation of Diffusional Association
TRAJAN	A Tool to Analyze Trajectories from Molecular Simulations
MolSurfer	2D Maps to Navigate 3D Structures of Proteins and their Complexes
WHAT IF	WHAT IF is a versatile protein structure analysis program that can be used for mutant prediction, structure verification, molecular graphics

- Database of Molecular Biology Software (at EBI)

Misc

pI	Isoelectric Point computation
pKa	scripts for protein pKa calculations with the UHBD program
WebMol	Java based PDB viewer

Sequences related to 4-OT:

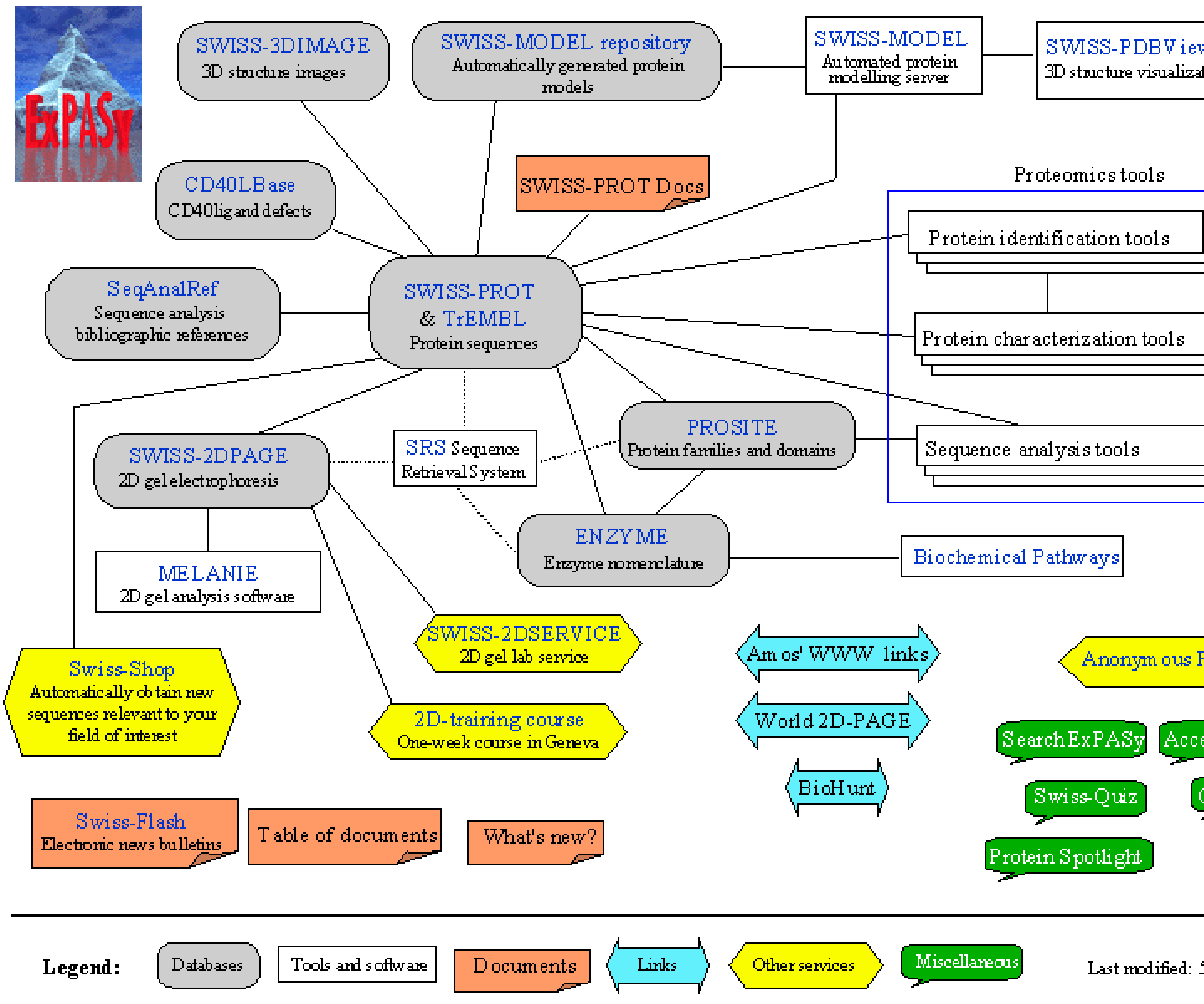
```
>Pseudomonas putida
PFAQIYMIEGRTEAQKKAVIEKVSQALVEATGAPMANVRVWIQEVPKENWGIAGVSAKEL
>dehalogenase
PFIECHIATGLSVARKQQLIRDVIDVTNKSIGSDPKIINVLLVEHAEANMSISGRI
```

III. ExPASy - (Expert Protein Analysis System) proteomics server

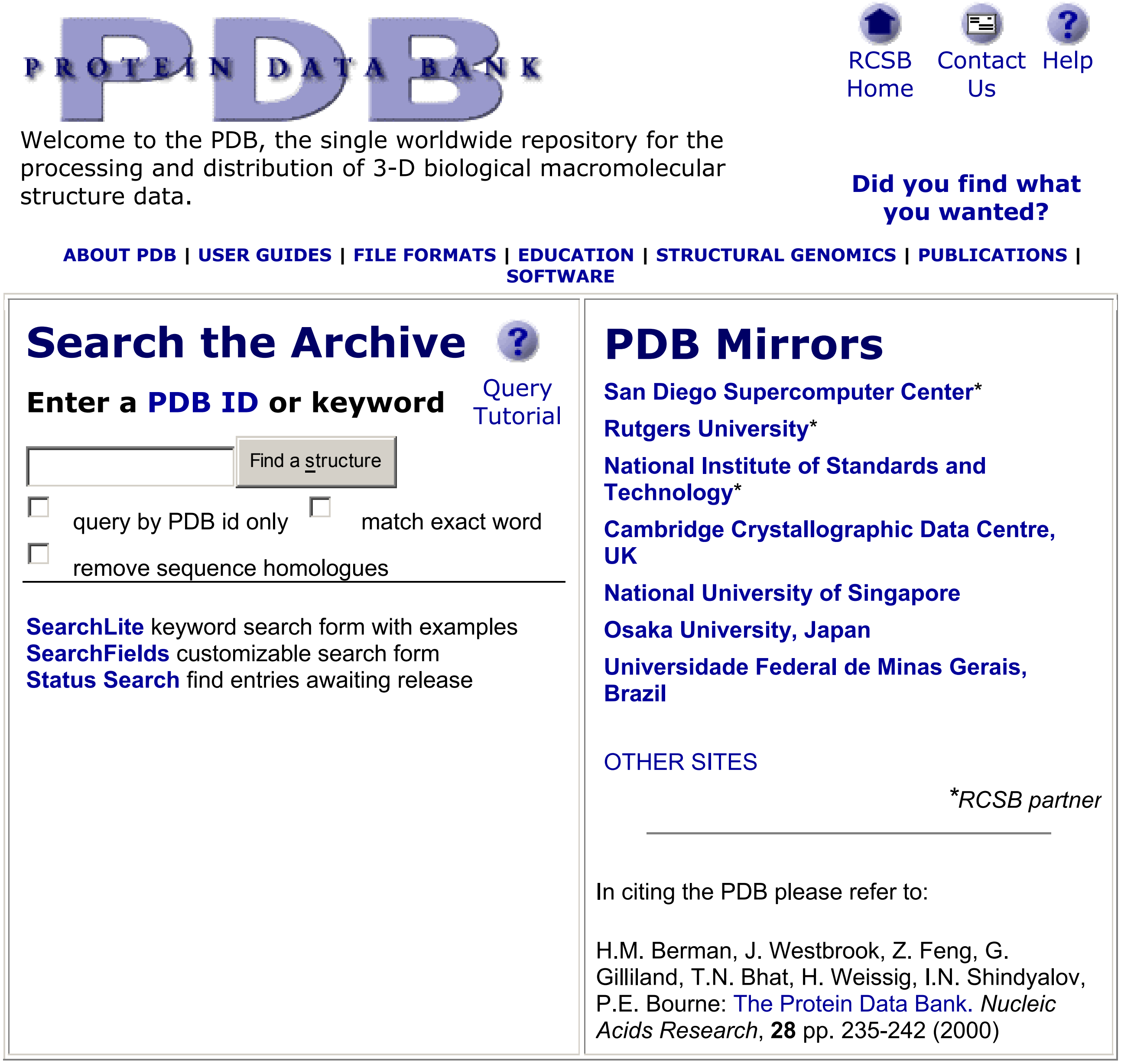
-

SWISS-MODEL – automatic generation of protein models from sequence, if there is a close sequence neighbor whose structure is already known.

ExPASy site map



IV. PDB (Protein Data Bank)
- home of protein & nucleic acid structures



PROTEIN DATA BANK
PDB
RCSB Home
Contact Us
Help
ABOUT PDB | USER GUIDES | FILE FORMATS | EDUCATION | STRUCTURAL GENOMICS | PUBLICATIONS | SOFTWARE
Query Tutorial
Find a structure
OTHER SITES

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

Did you find what you wanted?

Search the Archive

Enter a PDB ID or keyword

□ query by PDB id only □ match exact word

□ remove sequence homologues

SearchLite keyword search form with examples
SearchFields customizable search form
Status Search find entries awaiting release

PDB Mirrors

San Diego Supercomputer Center*
Rutgers University*
National Institute of Standards and Technology*
Cambridge Crystallographic Data Centre, UK
National University of Singapore
Osaka University, Japan
Universidade Federal de Minas Gerais, Brazil

**RCSB partner*

In citing the PDB please refer to:

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)