NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

Current Topics in Genome Analysis
Spring 2008

Week 3: Biological Sequence Analysis II

Andy Baxevanis, Ph.D.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES | NATIONAL INSTITUTES OF HEALTH | genome.gov/DIR

Overview

- Week 2
  - Similarity *vs*. Homology
  - Global *vs*. Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 3
  - Profiles, Patterns, Motifs, and Domains
  - Structures: VAST, Cn3D, and *de novo* Prediction
  - Multiple Sequence Alignment

Sequence Comparisons

- Homology searches
  - Usually "one-against-one" *BLAST, FASTA*
  - Allows for comparison of individual sequences against databases comprised of individual sequences
- Profile searches
  - Uses collective characteristics of a family of proteins
  - Search can be "one-against-many" *Pfam, InterPro, CDD*

    or "many-against-one" *PSI-BLAST*

Profiles

- Numerical representations of multiple sequence alignments
- Depend upon *patterns* or *motifs* containing conserved residues
- Represent the common characteristics of a protein family
- Can find similarities between sequences with little or no sequence identity
- Allow for the analysis of distantly-related proteins

Profile Construction

```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
```

- *Which residues are seen at each position?*
- *What is the frequency of observed residues?*
- *Which positions are conserved?*
- *Where can gaps be introduced?*

Position-Specific Scoring Table

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	18	8	15	8	8	-12	15	8	8	-5	-5	-1	7	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
P	31	6	7	6	6	11	19	11	9	6	16	11	7	89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	30	-8	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

Patterns

Phe *or* Tyr → [FY]

Cys → C

any amino acid → x

any two amino acids → x(2)

Val *or* Ala → {VA}

any amino acid → x

three His → H(3)

```
[FY]-x-C-x(2)-{VA}-x-H(3)
```

Pfam

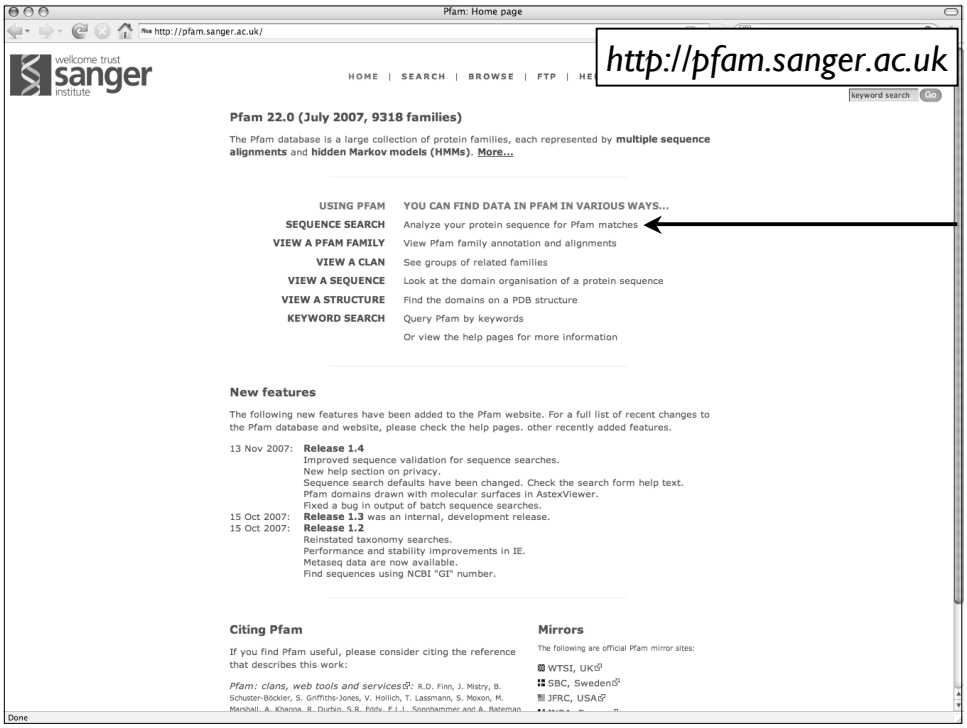
- Collection of multiple alignments of protein domains and conserved protein regions (regions which probably have structural or functional importance)
- Each Pfam entry contains:
  - Multiple sequence alignment of family members
  - Protein domain architectures
  - Species distribution of family members
  - Information on known protein structures
  - Links to other protein family databases

Pfam

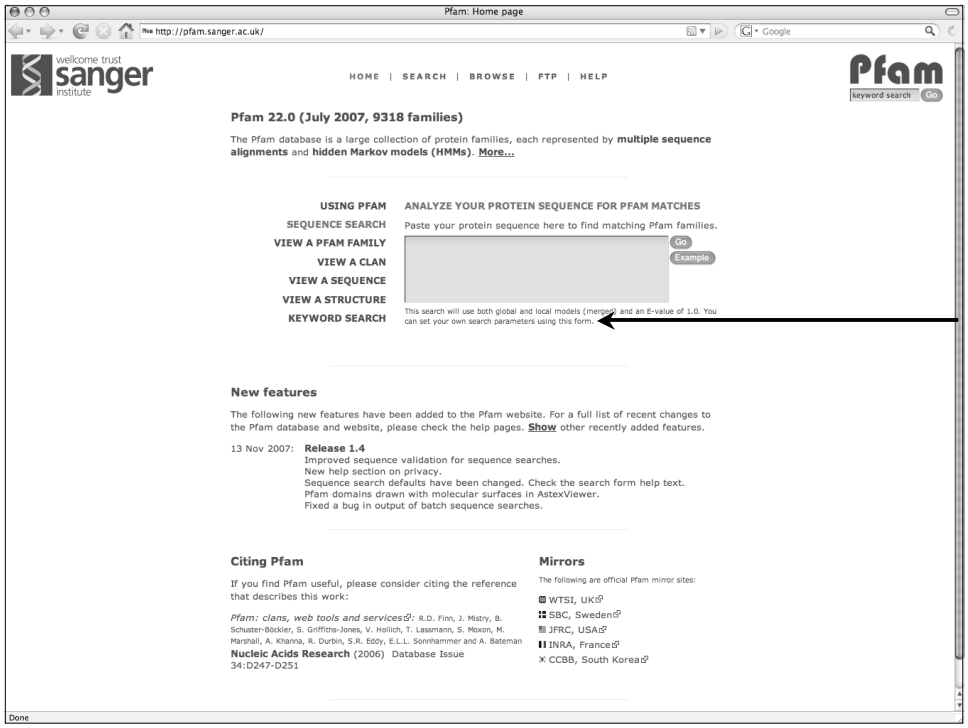
- Pfam A
  - Based on *curated* multiple alignments (“seed alignment”)
  - Hidden Markov models (HMMs) used to find all detectable protein sequences belonging to the family
  - Given the method used to construct the alignments, hits are highly likely to be true positives
- Pfam B
  - Automatically generated from database searches
  - Deemed “lower quality”, but can be useful when no Pfam A family is identified

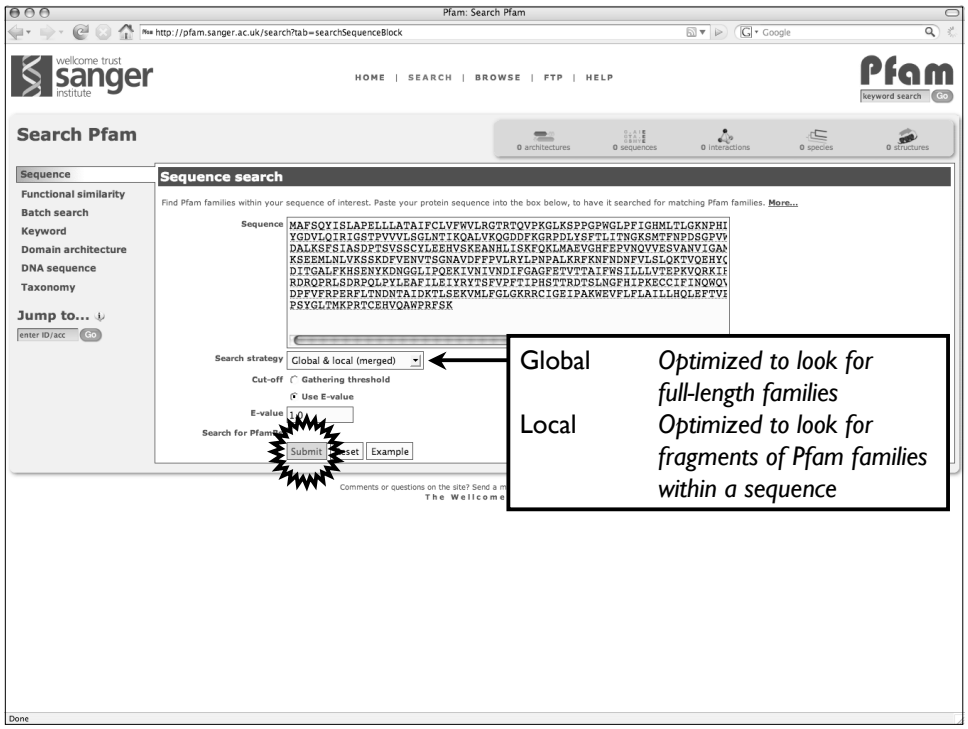
Sequences Used in Examples

http://research.nhgri.nih.gov/teaching/seq_analysis.shtml



http://pfam.sanger.ac.uk


Pfam: Home page
http://pfam.sanger.ac.uk/
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP
Pfam
keyword search Go
Pfam 22.0 (July 2007, 9318 families)
The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...
USING PFAM
SEQUENCE SEARCH
VIEW A PFAM FAMILY
VIEW A CLAN
VIEW A SEQUENCE
VIEW A STRUCTURE
KEYWORD SEARCH
ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES
Paste your protein sequence here to find matching Pfam families.
Go
Example
This search will use both global and local models (merged) and an E-value of 1.0. You can set your own search parameters using this form.
New features
The following new features have been added to the Pfam website. For a full list of recent changes to the Pfam database and website, please check the help pages. Show other recently added features.
13 Nov 2007: Release 1.4
Improved sequence validation for sequence searches.
New help section on privacy.
Sequence search defaults have been changed. Check the search form help text.
Pfam domains drawn with molecular surfaces in AstexViewer.
Fixed a bug in output of batch sequence searches.
Citing Pfam
If you find Pfam useful, please consider citing the reference that describes this work:
Pfam: clans, web tools and services: R.D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer and A. Bateman
Nucleic Acids Research (2006) Database Issue 34:D247-D251
Mirrors
The following are official Pfam mirror sites:
WTSI, UK
SBC, Sweden
JFRC, USA
INRA, France
CCBB, South Korea
Done


Pfam: Search Pfam
http://pfam.sanger.ac.uk/search?tab=searchSequenceBlock
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP
Pfam
keyword search Go
Search Pfam
0 architectures
0 sequences
0 interactions
0 species
0 structures
Sequence
Functional similarity
Batch search
Keyword
Domain architecture
DNA sequence
Taxonomy
Jump to...
enter ID/acc Go
Sequence search
Find Pfam families within your sequence of interest. Paste your protein sequence into the box below, to have it searched for matching Pfam families. More...
Sequence
Search strategy
Global & local (merged)
Cut-off
Gathering threshold
Use E-value
E-value
Search for Pfam
Submit
Example
Global
Optimized to look for full-length families
Local
Optimized to look for fragments of Pfam families within a sequence
Done

Pfam: Sequence search results

http://pfam.sanger.ac.uk/search/sequence/results?jobId=3881E3CA-C82D-11DC-8CCF-A4B270EDEBF6

HOME | SEARCH | BROWSE | FTP | HELP

Sequence search results

We found **1** Pfam-A match to your search sequence. You did not choose to search for Pfam-B matches.

The Pfam graphic below shows the arrangement of the domains on your search sequence. Clicking on any of the domains will take you to a page of information about that domain.

Below showing the details of the domains that were found. Rows containing significant hits are highlighted. Hits which do not start and end at the end points of the matching HMM are also highlighted.

For Pfam-A hits we show the alignments between your search sequence and the matching HMM. You can show individual alignments by clicking on the "Show" button in each row of the result table, or you can show all alignments using the links above the table.

You can bookmark this page and return to it later, but please note that old results will be removed after **one week**.

Return to the search form to look for Pfam domains on a new sequence.

Pfam-A Matches

Show or hide all alignments.

Pfam-A	Description	Entry type	Sequence		HMM		Bits score	E-value	Alignment mode	Show/hide alignment
			Start	End	From	To				
p450	Cytochrome P450	Domain	41	506	1	504	367.2	9.1e-113	fs	Hide

Comments or questions on the site? Send a mail to **pfam-help@sanger.ac.uk**

The Wellcome Trust

Done

Pfam: Family: p450 (PF00067)

http://pfam.sanger.ac.uk/family?acc=PF00067

Family: *p450* (PF00067)

70 architectures | 8703 sequences | 2 interactions | 1045 species | 148 structures

Summary
Domain organisation
Alignments
Trees
Curation & models
Species
Interactions
Structures

Jump to...

enter ID/acc Go

Summary

Cytochrome P450

Add annotation

Cytochrome P450s are haem-thiolate proteins [6] involved in the oxidative degradation of various compounds. They are particularly well known for their role in the degradation of environmental toxins and mutagens. They can be divided into 4 classes, according to the method by which electrons from NAD(P)H are delivered to the catalytic site. Sequence conservation is relatively low within the family - there are only 3 absolutely conserved residues - but their general topography and structural fold are highly conserved. The conserved core is composed of a coil termed the 'meander', a four-helix bundle, helices J and K, and two sets of beta-sheets. These constitute the haem-binding loop (with an absolutely conserved cysteine that serves as the 5th ligand for the haem iron), the proton-transfer groove and the absolutely conserved EXXR motif in helix K. While prokaryotic P450s are soluble proteins, most eukaryotic P450s are associated with microsomal membranes. their general enzymatic function is to catalyse regiospecific and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].

Example structure

PDB entry 1z10: CRYSTAL STRUCTURE OF HUMAN MICROSOMAL P450 2A6 WITH COUMARIN BOUND

1z10 View

Literature references

1. Werck-Reichhart D, Feyereisen R; , Genome Biol 2000;1:REVIEWS3003.: Cytochromes P450: a success story. PUBMED:11178272
2. Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. PUBMED:3304150
3. Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. PUBMED:2037557
4. Nelson DR, Kamataki T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, Gonzalez FJ, Coon MJ, Gunsalus IC, Gotoh O, et al; , DNA Cell Biol 1993;12:1-51.: The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. PUBMED:7678494
5. Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. PUBMED:8405421
6. Graham-Lorence S, Amarneh B, White RE, Peterson JA, Simpson ER; , Protein Sci 1995;4:1065-1080.: A three-dimensional model of aromatase cytochrome P450. PUBMED:7549871

Interpro entry IPR001128

Cytochrome P450 enzymes are a superfamily of haem-containing mono-oxygenases that are found in all kingdoms of life, and which show extraordinary diversity in their reaction chemistry. In mammals, these proteins are found primarily in microsomes of hepatocytes and other cell types, where they oxidise steroids, fatty acids and xenobiotics, and are important for the detoxification and clearance of various compounds, as well as for hormone synthesis and breakdown, cholesterol synthesis and vitamin D metabolism. In plants, these proteins are important for the biosynthesis of several compounds such as hormones, defensive compounds and fatty acids. In bacteria, they are important for several metabolic processes, such as the biosynthesis of antibiotic erythromycin in Saccharopolyspora erythraea.

Cytochrome P450 enzymes use haem to oxidise their substrates, using protons derived from NADH or NADPH to split the oxygen so a single atom can be added to a substrate. They also require electrons, which they receive from a variety of redox partners. In certain cases, cytochrome P450 can be fused to its redox partner to produce a bi-functional protein, such as with P450BM-3 from Bacillus megaterium PUBMED:17023115, which has haem and flavin domains.

Organisms produce many different cytochrome P450 enzymes (at least 58 in humans), which together with alternative splicing can provide a wide array of enzymes with different substrate and tissue specificities. Individual cytochrome P450 proteins follow the nomenclature: CYP, followed by a number (family), then a letter (subfamily), and another number (protein); e.g. CYP3A4 is the fourth protein in family 3, subfamily A. In general, family members should share >40% identity, while subfamily members should share >55% identity.

Cytochrome P450 proteins can also be grouped by two different schemes. One scheme was based on a taxonomic split: class I (prokaryotic/mitochondrial) and class II (eukaryotic microsomes). The other scheme was based on the number of components in the system: class B (3-components) and class E (2-components). These classes merge to a certain degree. Most prokaryotes and mitochondria (and fungal CYP55) have 3-component systems (class I/class B) - a FAD-containing flavoprotein (NAD(P)H-dependent reductase), an iron-sulphur protein and P450. Most eukaryotic microsomes have 2-component systems (class II/class E) - NADPH:P450 reductase (FAD and FMN-containing flavoprotein) and P450. There are exceptions to this scheme, such as 1-component systems that resemble class E enzymes PUBMED:16042601, PUBMED:15128046, PUBMED:8637843. The class E enzymes can be further subdivided into five sequence clusters, groups I-V, each of which may contain more than one cytochrome P450 family (eg, CYP1 and CYP2 are both found in group I). The divergence of the cytochrome P450 superfamily into B- and E-classes, and further divergence into stable clusters within the E-class, appears to be very ancient, occurring before the appearance of eukaryotes.

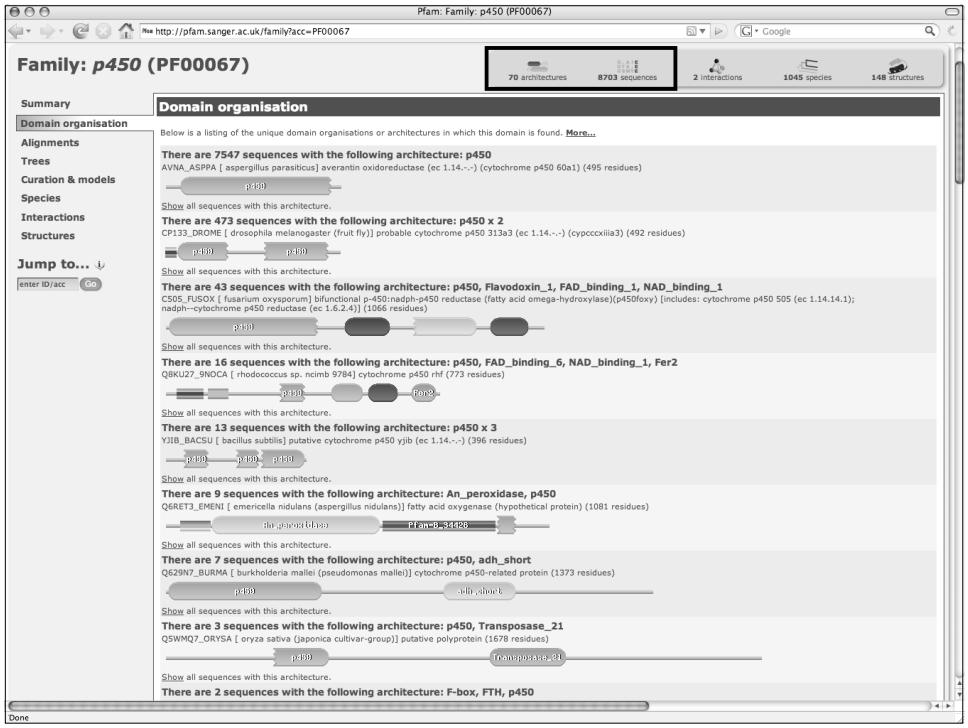
More information about these proteins can be found at Protein of the Month: Cytochrome P450 PUBMED:.

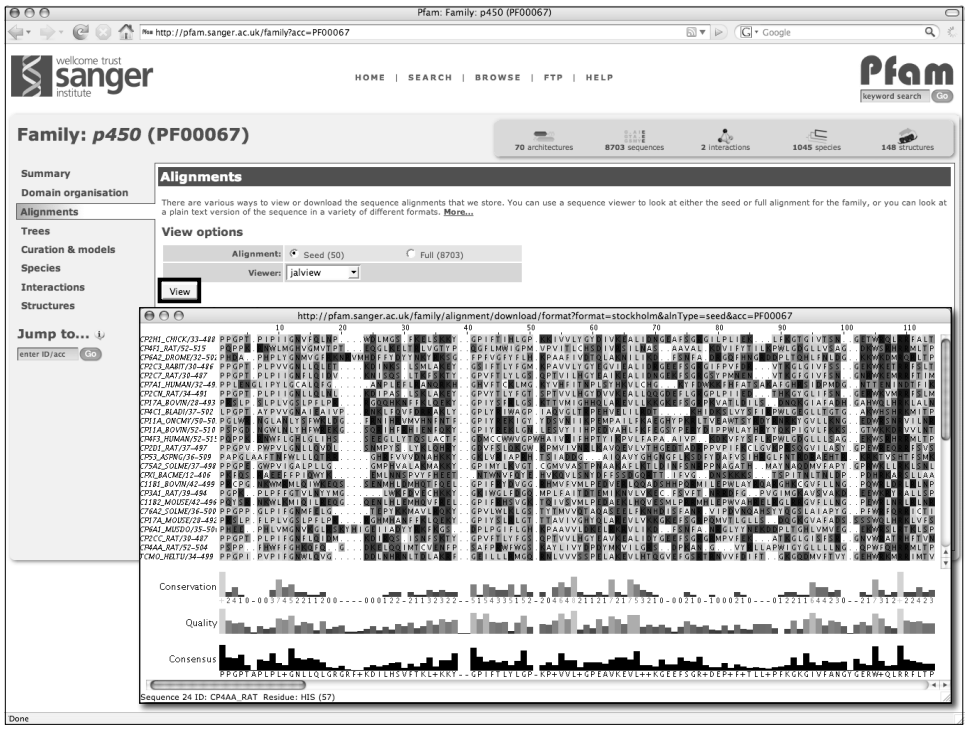
Gene Ontology

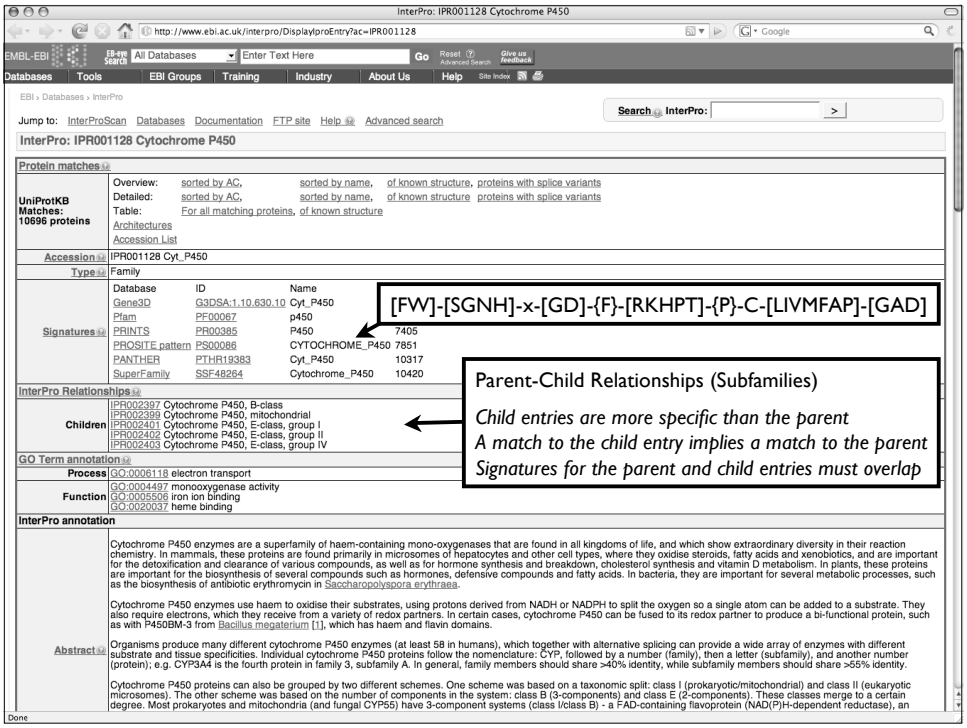
Molecular function	heme binding (GO:0020037)
Molecular function	iron ion binding (GO:0005506)
Biological process	electron transport (GO:0006118)
Molecular function	monooxygenase activity (GO:0004497)

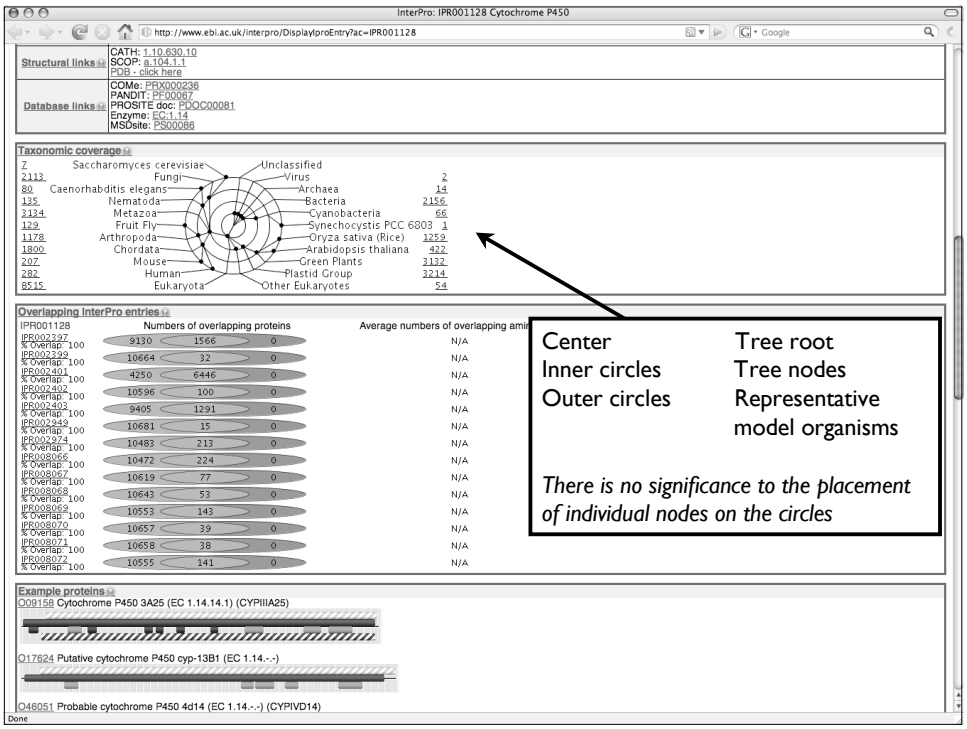
External database links

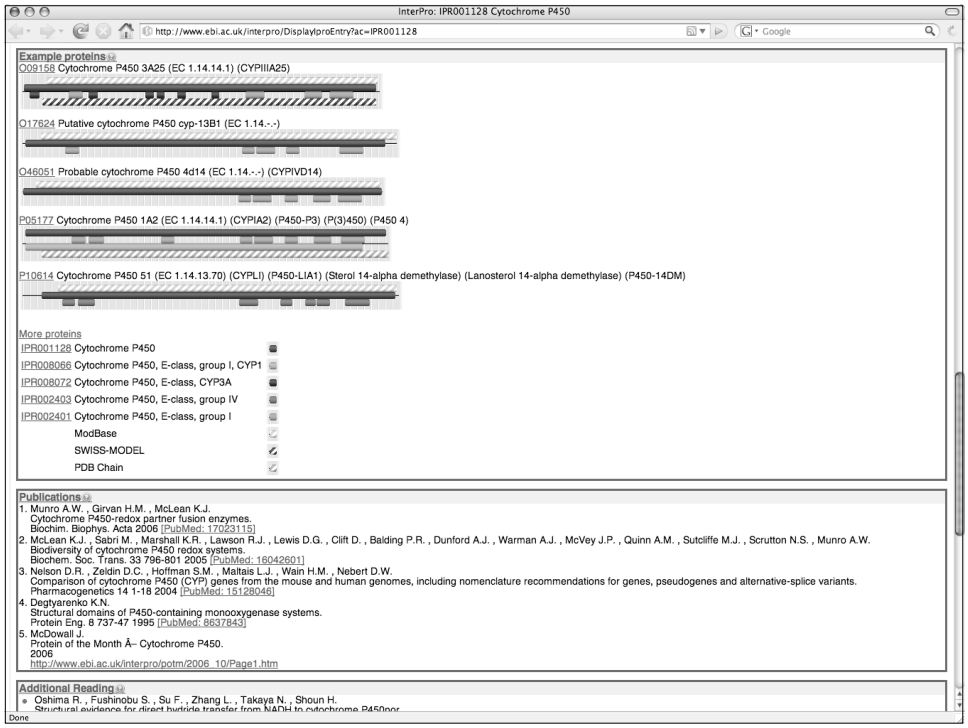
Done


Pfam: Family: p450 (PF00067)
http://pfam.sanger.ac.uk/family?acc=PF00067
Family: p450 (PF00067)
70 architectures
8703 sequences
2 interactions
1045 species
148 structures
Summary
Domain organisation
Alignments
Trees
Curation & models
Species
Interactions
Structures
Jump to...
Domain organisation
Below is a listing of the unique domain organisations or architectures in which this domain is found. More...
There are 7547 sequences with the following architecture: p450
There are 473 sequences with the following architecture: p450 x 2
There are 43 sequences with the following architecture: p450, Flavodoxin_1, FAD_binding_1, NAD_binding_1
There are 16 sequences with the following architecture: p450, FAD_binding_6, NAD_binding_1, Fer2
There are 13 sequences with the following architecture: p450 x 3
There are 9 sequences with the following architecture: An_peroxidase, p450
There are 7 sequences with the following architecture: p450, adh_short
There are 3 sequences with the following architecture: p450, Transposase_21
There are 2 sequences with the following architecture: F-box, FTH, p450
Done


Pfam: Family: p450 (PF00067)
http://pfam.sanger.ac.uk/family?acc=PF00067
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP
Pfam
keyword search
Family: p450 (PF00067)
70 architectures
8703 sequences
2 interactions
1045 species
148 structures
Summary
Domain organisation
Alignments
Trees
Curation & models
Species
Interactions
Structures
Jump to...
Alignments
There are various ways to view or download the sequence alignments that we store. You can use a sequence viewer to look at either the seed or full alignment for the family, or you can look at a plain text version of the sequence in a variety of different formats. More...
View options
Alignment: Seed (50) Full (8703)
Viewer: jalview
View
http://pfam.sanger.ac.uk/family/alignment/download/format?format=stockholm&alnType=seed&acc=PF00067
Conservation
Quality
Consensus
Sequence 24 ID: CP4AA_RAT Residue: HIS (57)
Done


InterPro: IPR001128 Cytochrome P450
http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128
Jump to: InterProScan Databases Documentation FTP site Help Advanced search
Protein matches
UniProtKB Matches: 10696 proteins
Accession IPR001128 Cyt_P450
Type Family
Signatures
[FW]-[SGNH]-x-[GD]-{F}-[RKHPT]-{P}-C-[LIVMFAP]-[GAD]
Parent-Child Relationships (Subfamilies)
Child entries are more specific than the parent
A match to the child entry implies a match to the parent
Signatures for the parent and child entries must overlap
InterPro Relationships
Children
GO Term annotation
Process
Function
InterPro annotation
Abstract


InterPro: IPR001128 Cytochrome P450
Structural links
Database links
Taxonomic coverage
Overlapping InterPro entries
Example proteins
Center
Inner circles
Outer circles
Tree root
Tree nodes
Representative model organisms
There is no significance to the placement of individual nodes on the circles



Current Protocols in Bioinformatics

CPBI Unit 2.5
Pfam

Identifying Protein Domains with the Pfam Database

UNIT 2.5

Hundreds of thousands of protein sequences are now known and the deluge of data shows no signs of slowing. The sequence analysis of proteins may seem like a perpetual task. However, the majority of protein sequences appear to fall into a few thousand protein families (Chothia, 1992). Very often these families are representative of proteins at the domain level, where domains are discrete structural units that are frequently found in different protein contexts. Pfam is a database of such protein domain families (Sonnhammer et al., 1997; Bateman et al., 2002), with each family represented by multiple sequence alignments and profile hidden Markov models (HMMs). In addition, each family has associated annotations, literature references, and links to other databases. The entries in Pfam are available via the Web and in flatfile format.

The use of Pfam by molecular biologists as a protein information resource and analysis tool is widespread. Genome sequencing projects, including human and fly, have used Pfam extensively for large-scale functional annotation of genomic data, while smaller groups devoted to a single protein or biochemical pathway frequently use Pfam for their analyses. The multiple sequence alignments around which Pfam families are built are important for understanding both protein structure and function, and form the basis for techniques such as secondary structure prediction, fold recognition, and phylogenetic analysis, and can aid mutation design. This unit contains detailed information on how to access and utilize the information present in the Pfam database, namely the families, multiple alignments, and annotation. Details on running Pfam, both remotely (see Basic Protocol) and locally (see Alternate Protocol 1) are presented. The unit also includes a brief discussion of analyzing genomic DNA with Pfam (see Alternate Protocol 2). Information on underlying flatfiles is also presented (see Appendix at the end of this unit).

ANALYZING A PROTEIN SEQUENCE WITH Pfam VIA THE WEB

BASIC PROTOCOL

A primary use of the Pfam database is to determine what domains are in a protein of interest. This section describes several approaches for carrying out this analysis. In a typical session any one of them or several in various combinations may be used, depending on the kinds of questions asked and the types of information sought. A later section describes more advanced analyses that would typically follow these basic ones (see Suggestions for Further Analysis).

Necessary Resources

Hardware

Workstation with network connection

Software

Javascript-capable browser (e.g., Netscape 4.0, Internet Explorer 4.0)

Files

Protein sequence of interest in appropriate format (e.g., FASTA; APPENDIX 1B)

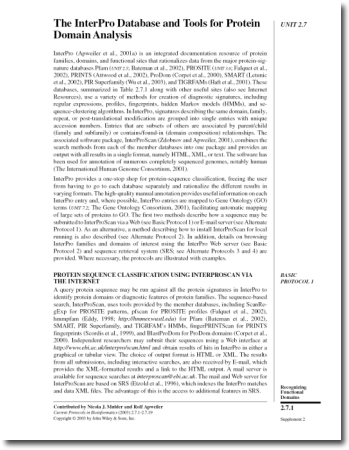
The example sequences used below are available at the Current Protocols Web site (http://www.3.interscience.wiley.com/c_p/cpbi_samples/analysfiles.htm).

Contributed by Robert Finn, Sam Griffiths-Jones, and Alex Bateman

Current Protocols in Bioinformatics (2003) 2.5.1-2.5.18
Copyright © 2003 by John Wiley & Sons, Inc.

Recognizing Functional Domains
2.5.1
Supplement 1

CPBI Unit 2.7
InterPro



The InterPro Database and Tools for Protein Domain Analysis

UNIT 2.7

InterPro (Apweiler et al., 2001a) is an integrated documentation resource of protein families, domains, and functional sites that rationalizes data from the major protein signature databases Pfam (UNIT 2.5; Bateman et al., 2002), PROSITE (UNIT 2.6; Falquet et al., 2002), PRINTS (Attwood et al., 2002), ProDom (Corpet et al., 2000), SMART (Letunic et al., 2002), PIR Superfamily (Wu et al., 2003), and TIGRFAMs (Haft et al., 2001). These databases, summarized in Table 2.7.1 along with other useful sites (also see Internet Resources), use a variety of methods for creation of diagnostic signatures, including regular expressions, profiles, fingerprints, hidden Markov models (HMMs), and sequence-clustering algorithms. In InterPro, signatures describing the same domain, family, repeat, or post-translational modification are grouped into single entries with unique accession numbers. Entries that are subsets of others are associated by parent/child (family and subfamily) or contains/found in (domain composition) relationships. The associated software package, InterProScan (Zdobnov and Apweiler, 2001), combines the search methods from each of the member databases into one package and provides an output with all results in a single format, namely HTML, XML, or text. The software has been used for annotation of numerous completely sequenced genomes, notably human (The International Human Genome Consortium, 2001).

InterPro provides a one-stop shop for protein-sequence classification, freeing the user from having to go to each database separately and rationalize the different results in varying formats. The high-quality manual annotation provides useful information on each InterPro entry and, where possible, InterPro entries are mapped to Gene Ontology (GO) terms (UNIT 7.2; The Gene Ontology Consortium, 2001), facilitating automatic mapping of large sets of proteins to GO. The first two methods describe how a sequence may be submitted to InterProScan via a Web (see Basic Protocol 1) or E-mail server (see Alternate Protocol 1). As an alternative, a method describing how to install InterProScan for local running is also described (see Alternate Protocol 2). In addition, details on browsing InterPro families and domains of interest using the InterPro Web server (see Basic Protocol 2) and sequence retrieval system (SRS; see Alternate Protocols 3 and 4) are provided. Where necessary, the protocols are illustrated with examples.

PROTEIN SEQUENCE CLASSIFICATION USING INTERPROSCAN VIA THE INTERNET

BASIC PROTOCOL 1

A query protein sequence may be run against all the protein signatures in InterPro to identify protein domains or diagnostic features of protein families. The sequence-based search, InterProScan, uses tools provided by the member databases, including ScanProGEnz for PROSITE patterns, pfscan for PROSITE profiles (Falquet et al., 2002), hmmpfam (Eddy, 1998; http://hmmer.wustl.edu/) for Pfam (Bateman et al., 2002), SMART, PIR Superfamily, and TIGRFAM + HMMs, fingerPRINTScan for PRINTS fingerprints (Scordis et al., 1999), and BlastProDom for ProDom domains (Corpet et al., 2000). Independent researchers may submit their sequences using a Web interface at http://www.ebi.ac.uk/interpro/scan.html and obtain results of hits in InterPro in either a graphical or tabular view. The choice of output format is HTML or XML. The results from all submissions, including interactive searches, are also received by E-mail, which provides the XML-formatted results and a link to the HTML output. A mail server is available for sequence searches at interproscan@ebi.ac.uk. The mail and Web server for InterProScan are based on SRS (Etzold et al., 1996), which indexes the InterPro matches and data XML files. The advantage of this is the access to additional features in SRS.

Contributed by Nicola J. Mulder and Rolf Apweiler

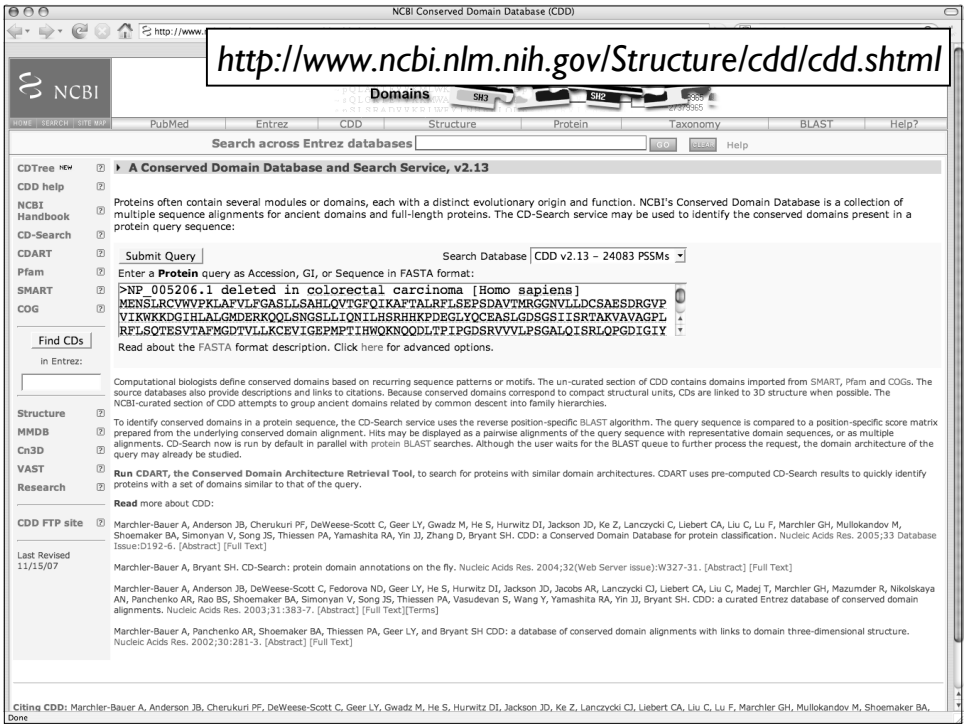
Current Protocols in Bioinformatics (2003) 2.7.1-2.7.39
Copyright © 2003 by John Wiley & Sons, Inc.

Recognizing Functional Domains
2.7.1
Supplement 2

http://nihlibrary.nih.gov
Search "Online Journals" for "Current Protocols in Bioinformatics"

Conserved Domain Database (CDD)

- Identify conserved domains in a protein sequence
- "Secondary database"
  - Pfam A and B
  - Simple Modular Architecture Research Tool (SMART)
  - Clusters of Orthologous Groups
- Search performed using RPS-BLAST
  - Query sequence is used to search a database of precalculated position-specific scoring tables
  - *Not* the same method used by Pfam or InterPro



NCBI Conserved Domain Search

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

NCBI **Conserved Domains**

NewSearch | PubMed | Nucleotide | Protein | Structure | Taxonomy | Help

Query sequence: [(local sequence)lcl|1]

Concise Result | Full Result | Show Search Information

Click on the **colored bar** for a conserved domain to **view your query sequence** within the multiple sequence alignment for that domain. To see only the sequences used to generate the domain, click on its **PSSMID** in the tabular summary.

Descriptions

Title	PssmId	Multi-Dom	E-value
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ...	28983	No	3e-15
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ...	28945	No	8e-13
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ...	28945	No	1e-12
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ...	28983	No	9e-12
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ...	28945	No	3e-11
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ...	28983	No	3e-10
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ...	28945	No	6e-09
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ...	28945	No	8e-07
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ...	28983	No	1e-06
pfam06583, Neogenin_C, Neogenin C-terminus. This family represents the C-terminus of e...	87114	No	3e-96
pfam07686, V-set, Immunoglobulin V-set domain. This domain is found in antibodies as w...	87333	Yes	2e-04

Search for similar domain architectures

CD Search Reference:
Marchler-Bauer A, Bryant SH (2004), "*CD-Search: protein domain annotations on the fly.*", **Nucleic Acids Res.32**(W)327-331.

Help | Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Done

NCBI Conserved Domain Search

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

NewSearch | PubMed | Nucleotide | Protein | Structure | Taxonomy | Help

Query sequence: [(local sequence)lcl|1]

Concise Result | Full Result | Show Search Information

Click on the **colored bar** for a conserved domain to **view your query sequence** within the multiple sequence alignment for that domain. To see only the sequences used to generate the domain, click on its **PSSMID** in the tabular summary.

Descriptions

Title	PssmId	Multi-Dom	E-value
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ...	28983	No	3e-15

cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members are components of neural cell adhesion molecules (N-CAM L1), Fasciclin II and the insect immune protein Hemolin. The subfamily also includes receptor domains such as as the extracelluar ligand binding domain of Fibroblast Growth Factor Receptor 2. Members are phylogenetically diverse, occuring throughout metazoa, and are not components of the adaptive immune system molecules found in jawed vertebrates. A predominant feature of most Ig domains is a disulfide bridge connecting 2 beta-sheets with a Trp packing against the disulfide bond..

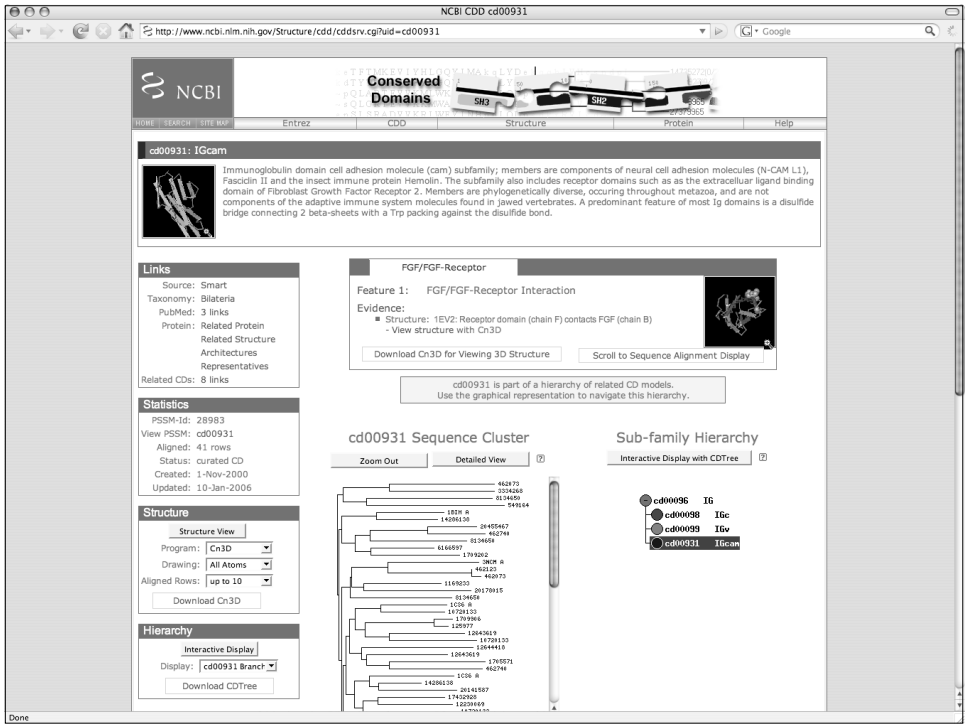
CD Length: 89, **Pct. Aligned:** 100, **Bit Score:** 79.775909, **E-value:** 3e-15

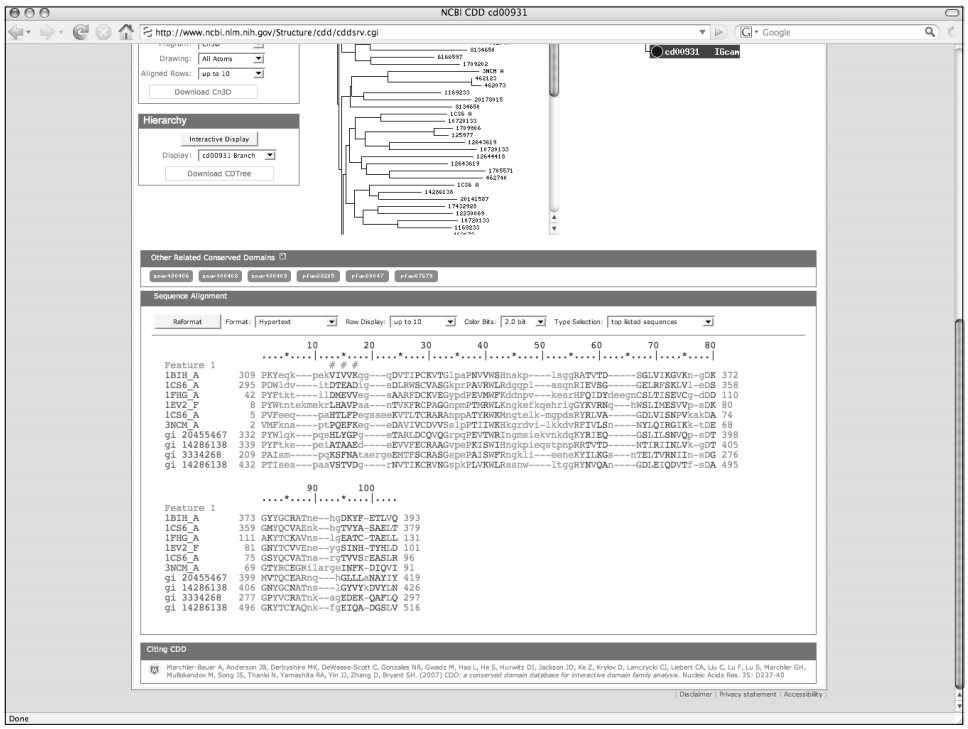
```
                 10        20        30        40        50        60        70        80
            ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1       330 PPWFLNHPSNLYAYESMDIEFECTVSGKPVPTVNWMKNGDVVIPSD--YFQIVGGSNLRILGVVKSDEGFYQCVAENEAG 407
cd00931   1 PTFTQKPPPDTVVAGGEDVTLECRASGNPPPTITWLKNGKPLSLLDgrYTVLDNNGTLTISNVTKEDAGTYTCVATNSAG 80

            ....*....
1       408 NAQTSAQLI 416
cd00931  81 GASASARLT 89
```

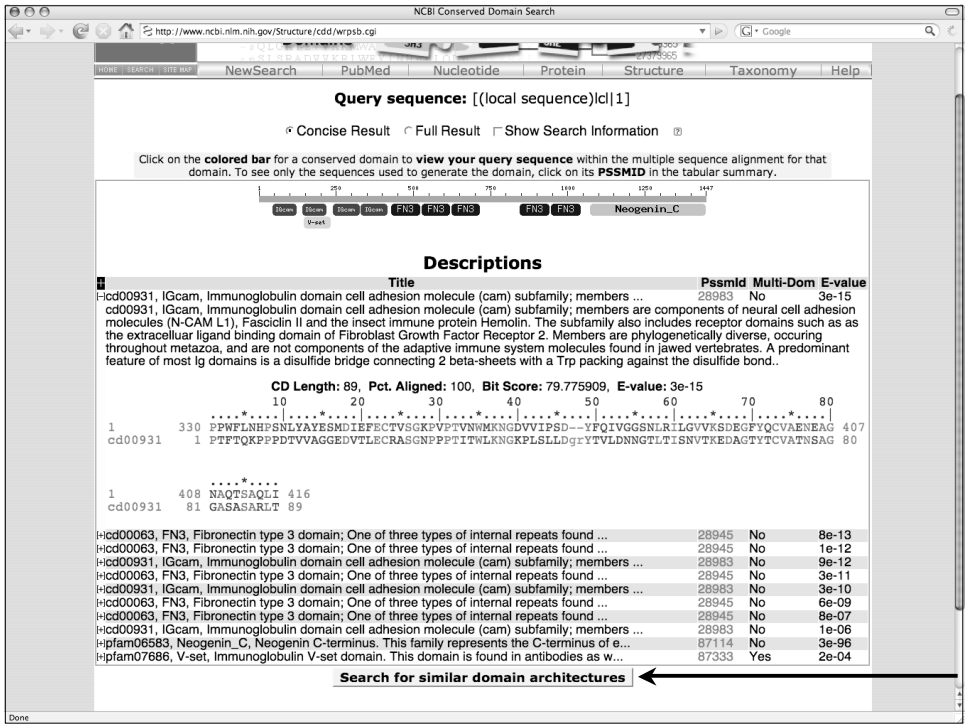
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ...	28945	No	8e-13
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ...	28945	No	1e-12
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ...	28983	No	9e-12
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ...	28945	No	3e-11
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ...	28983	No	3e-10
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ...	28945	No	6e-09
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ...	28945	No	8e-07
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ...	28983	No	1e-06
pfam06583, Neogenin_C, Neogenin C-terminus. This family represents the C-terminus of e...	87114	No	3e-96
pfam07686, V-set, Immunoglobulin V-set domain. This domain is found in antibodies as w...	87333	Yes	2e-04

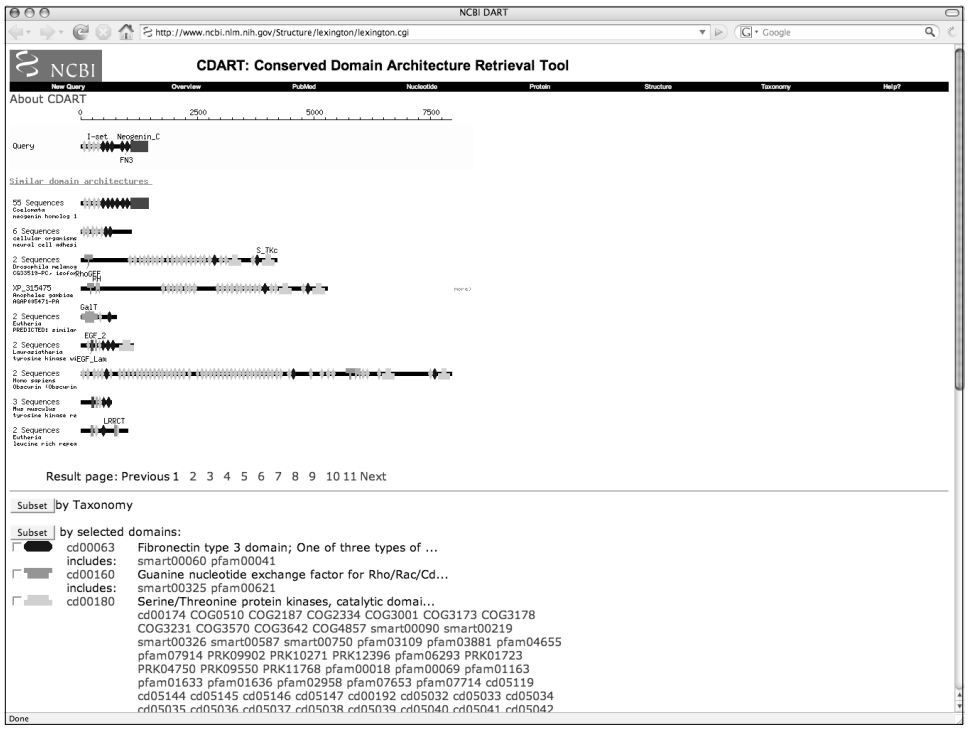
Search for similar domain architectures

Done


NCBI CDD cd00931
http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=cd00931
NCBI
Conserved Domains
SH3
SH2
Entrez
CDD
Structure
Protein
Help
cd00931: IGcam
Immunoglobulin domain cell adhesion molecule (cam) subfamily; members are components of neural cell adhesion molecules (N-CAM L1), Fasciclin II and the insect immune protein Hemolin. The subfamily also includes receptor domains such as as the extracellular ligand binding domain of Fibroblast Growth Factor Receptor 2. Members are phylogenetically diverse, occuring throughout metazoa, and are not components of the adaptive immune system molecules found in jawed vertebrates. A predominant feature of most Ig domains is a disulfide bridge connecting 2 beta-sheets with a Trp packing against the disulfide bond.
Links
Source: Smart
Taxonomy: Bilateria
PubMed: 3 links
Protein: Related Protein
Related Structure
Architectures
Representatives
Related CDs: 8 links
FGF/FGF-Receptor
Feature 1: FGF/FGF-Receptor Interaction
Evidence:
Structure: 1EV2: Receptor domain (chain F) contacts FGF (chain B)
- View structure with Cn3D
Download Cn3D for Viewing 3D Structure
Scroll to Sequence Alignment Display
cd00931 is part of a hierarchy of related CD models.
Use the graphical representation to navigate this hierarchy.
Statistics
PSSM-Id: 28983
View PSSM: cd00931
Aligned: 41 rows
Status: curated CD
Created: 1-Nov-2000
Updated: 10-Jan-2006
cd00931 Sequence Cluster
Zoom Out
Detailed View
Sub-family Hierarchy
Interactive Display with CDTree
cd00096 IG
cd00098 IGc
cd00099 IGv
cd00931 IGcam
Structure
Structure View
Program: Cn3D
Drawing: All Atoms
Aligned Rows: up to 10
Download Cn3D
Hierarchy
Interactive Display
Display: cd00931 Branch
Download CDTree
Done

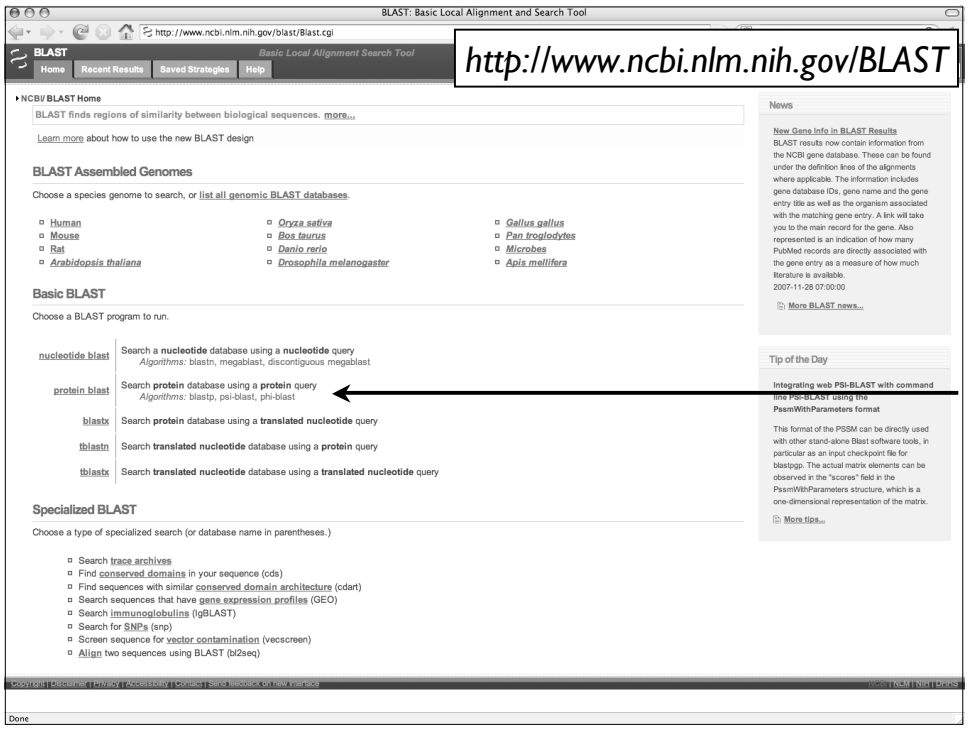

NCBI CDD cd00931
http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi
Drawing: All Atoms
Aligned Rows: up to 10
Download Cn3D
Hierarchy
Interactive Display
Display: cd00931 Branch
Download CDTree
cd00931 IGcam
Other Related Conserved Domains
Sequence Alignment
Reformat
Format: Hypertext
Row Display: up to 10
Color Bits: 2.0 bit
Type Selection: top listed sequences
Feature 1
1BIH_A 309 PKYeqk---pekVIVVKg---qDVTIPCKVTGlpaPNVVWSHnakp----lsggRATVTD-----SGLVIKGVKn-gDK 372
1CS6_A 295 PDWldv----itDTEADig---sDLRWSCVASGkprPAVRWLRdqqpl---asqnRIEVSG-----GELRFSKLVl-eDS 358
1FHG_A 42 PYFtkt----ilDMEVVeg---sAARFDCKVEGypdPEVMWFKddnpv---kesrHFQIDYdeegnCSLTISEVCg-dDD 110
1EV2_F 8 PYWtntekmekrLHAVPaa---nTVKFRCPAGGnpmPTMRWLKngkefkqehrigGYKVRNq---hWSLIMESVVp-sDK 80
1CS6_A 5 PVFeeq----paHTLFPegSaaeeKVTLTCRARAnppATYRWKNngtalk-mgpdsRYRLVA-----GDLVISNPVkakDA 74
3NCM_A 2 VMFkna----ptPQEFKeg---eDAVIVCDVVSslpPTIIWKHkgrdvi-lkkdvRFIVLSn----NYLQIRGIKk-tDE 68
gi 20455467 332 PYWlqk---pqsHLYGPg----eTARLDCQVQGrpqPEVTWRIngmsiekvnkdqKYRIEQ-----GSLILSNVQp-sDT 398
gi 14286138 339 PYFtke---peiATAAEd----eEVVFECRAAGvpePKISWIHngkpieqstpnpRKTVTD-----NTIRIINLVk-gDT 405
gi 3334268 209 PAIsm-----pqKSFNAtaergeEMTFSCRASGspePAISWFRngkli----eeneKYILKGs---nTELTVRNIIn-sDG 276
gi 14286138 432 PTIsea---paaVSTVDg----rNVTIKCRVNGspkPLVKWLRasnw----ltggRYNVQAn----GDLEIQDVTf-sDA 495
Citing CDD
Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH. (2007) CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res. 35: D237-40
Disclaimer | Privacy statement | Accessibility
Done

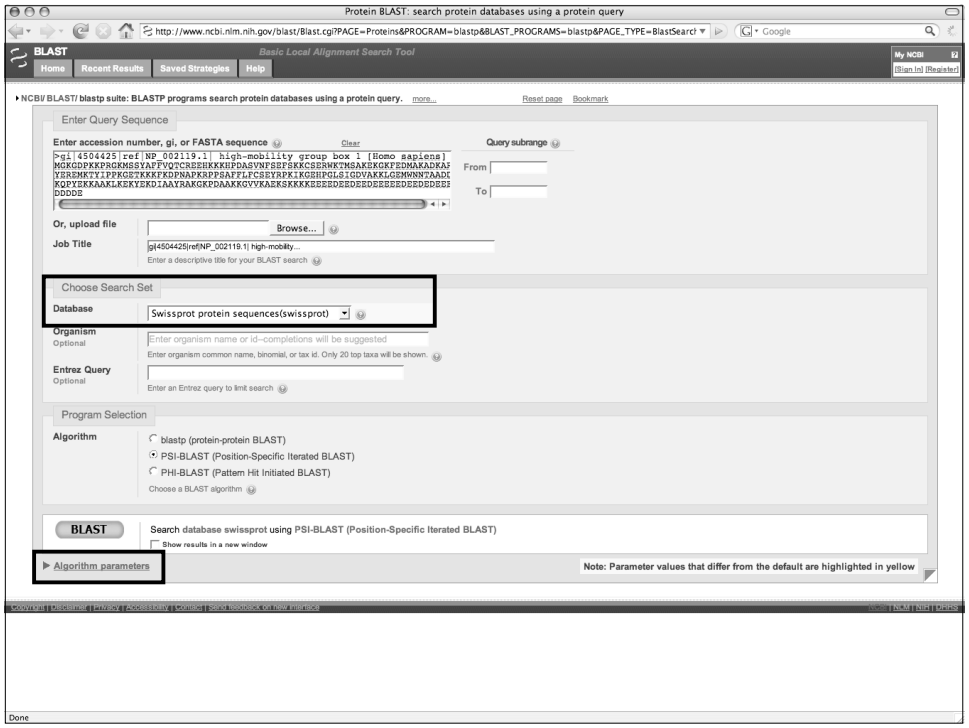

NCBI Conserved Domain Search
http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
NewSearch
PubMed
Nucleotide
Protein
Structure
Taxonomy
Help
Query sequence: [(local sequence)lcl|1]
Concise Result
Full Result
Show Search Information
Click on the colored bar for a conserved domain to view your query sequence within the multiple sequence alignment for that domain. To see only the sequences used to generate the domain, click on its PSSMID in the tabular summary.
Neogenin_C
Descriptions
Title
PssmId
Multi-Dom
E-value
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 3e-15
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members are components of neural cell adhesion molecules (N-CAM L1), Fasciclin II and the insect immune protein Hemolin. The subfamily also includes receptor domains such as as the extracelluar ligand binding domain of Fibroblast Growth Factor Receptor 2. Members are phylogenetically diverse, occuring throughout metazoa, and are not components of the adaptive immune system molecules found in jawed vertebrates. A predominant feature of most Ig domains is a disulfide bridge connecting 2 beta-sheets with a Trp packing against the disulfide bond..
CD Length: 89, Pct. Aligned: 100, Bit Score: 79.775909, E-value: 3e-15
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 8e-13
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 1e-12
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 9e-12
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 3e-11
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 3e-10
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 6e-09
cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 8e-07
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 1e-06
pfam06583, Neogenin_C, Neogenin C-terminus. This family represents the C-terminus of e... 87114 No 3e-96
pfam07686, V-set, Immunoglobulin V-set domain. This domain is found in antibodies as w... 87333 Yes 2e-04
Search for similar domain architectures
Done

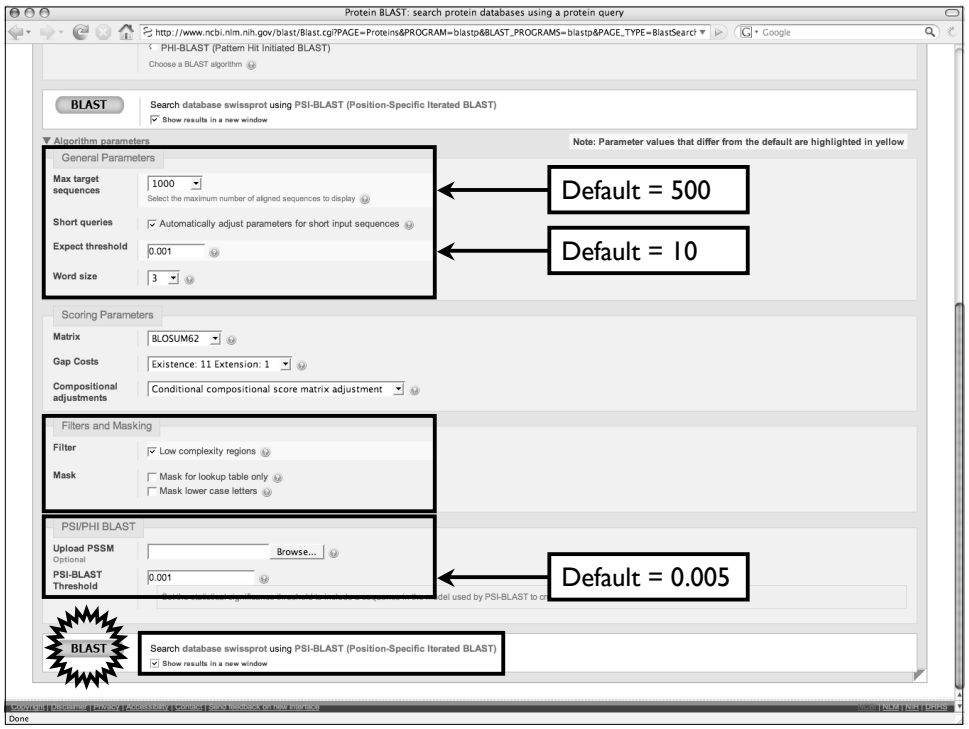

NCBI DART
http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi
NCBI
CDART: Conserved Domain Architecture Retrieval Tool
New Query
Overview
PubMed
Nucleotide
Protein
Structure
Taxonomy
Help?
About CDART
Query
I-set
Neogenin_C
FN3
Similar domain architectures
Result page: Previous 1 2 3 4 5 6 7 8 9 10 11 Next
Subset by Taxonomy
Subset by selected domains:
cd00063 Fibronectin type 3 domain; One of three types of ...
includes: smart00060 pfam00041
cd00160 Guanine nucleotide exchange factor for Rho/Rac/Cd...
includes: smart00325 pfam00621
cd00180 Serine/Threonine protein kinases, catalytic domai...
cd00174 COG0510 COG2187 COG2334 COG3001 COG3173 COG3178
COG3231 COG3570 COG3642 COG4857 smart00090 smart00219
smart00326 smart00587 smart00750 pfam03109 pfam03881 pfam04655
pfam07914 PRK09902 PRK10271 PRK12396 pfam06293 PRK01723
PRK04750 PRK09550 PRK11768 pfam00018 pfam00069 pfam01163
pfam01633 pfam01636 pfam02958 pfam07653 pfam07714 cd05119
cd05144 cd05145 cd05146 cd05147 cd00192 cd05032 cd05033 cd05034
Done

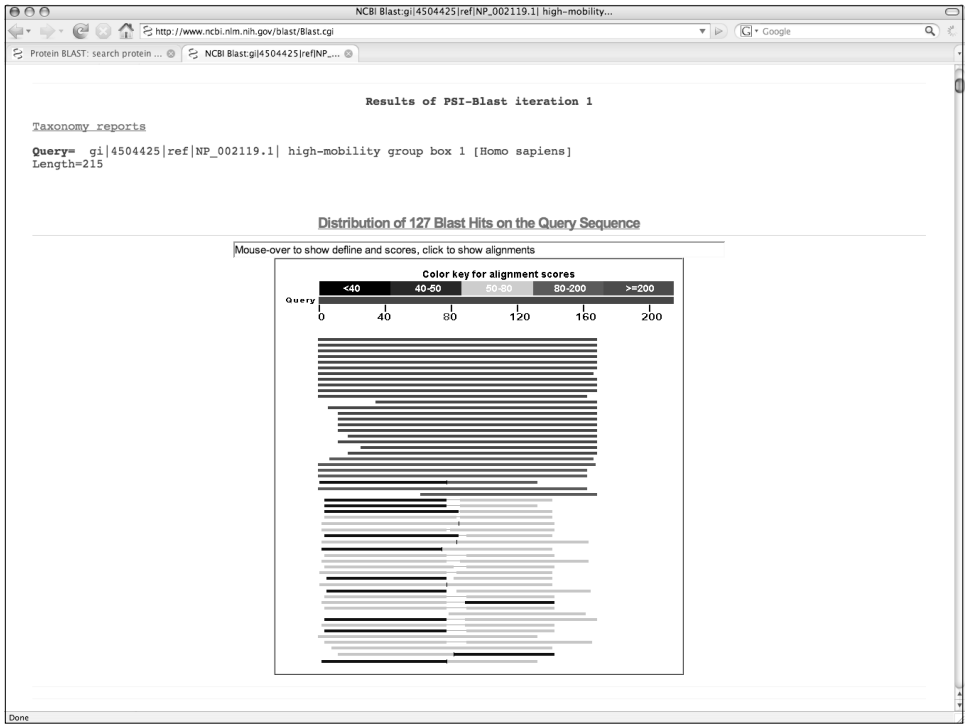
PSI-BLAST

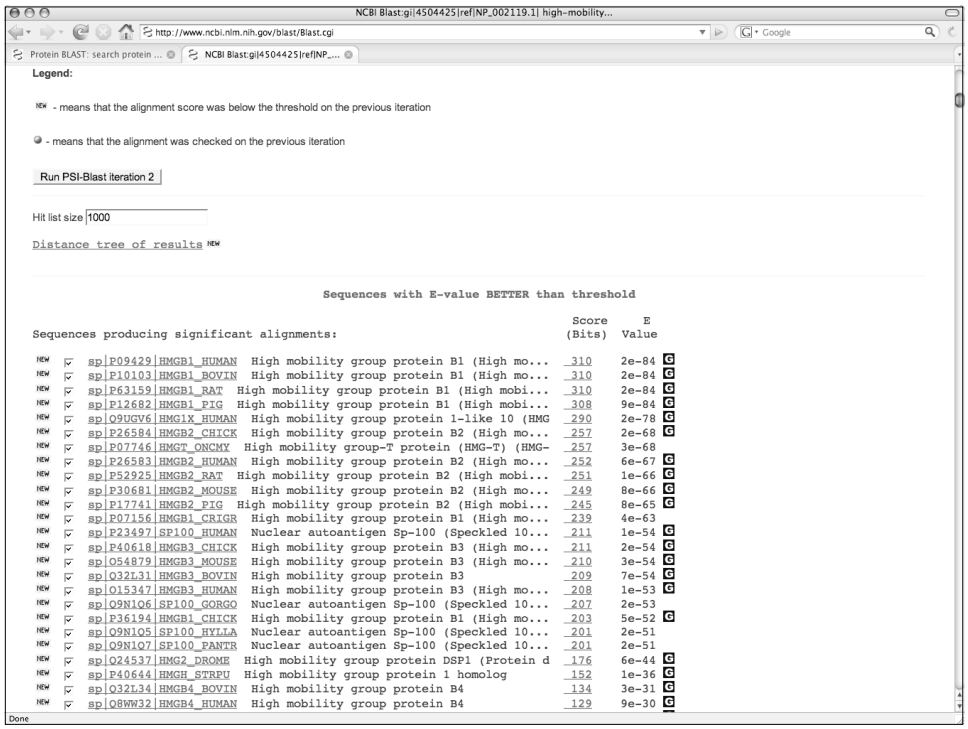
- Position-Specific Iterated BLAST search
- Easy-to-use version of a profile-based search
  - Perform BLAST search against protein database
  - Use results to calculate a position-specific scoring matrix
  - PSSM replaces query for next round of searches
  - May be iterated until no new significant alignments are found
    - Convergence – all related sequences deemed found
    - Divergence – query is too broad, make cutoffs more stringent




Protein BLAST: search protein databases using a protein query
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
Enter Query Sequence
Enter accession number, gi, or FASTA sequence
Query subrange
Or, upload file
Job Title
Choose Search Set
Database
Swissprot protein sequences(swissprot)
Organism
Optional
Entrez Query
Optional
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
BLAST
Search database swissprot using PSI-BLAST (Position-Specific Iterated BLAST)
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow


Protein BLAST: search protein databases using a protein query
BLAST
Search database swissprot using PSI-BLAST (Position-Specific Iterated BLAST)
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow
General Parameters
Max target sequences
1000
Short queries
Expect threshold
0.001
Word size
3
Default = 500
Default = 10
Scoring Parameters
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
Compositional adjustments
Conditional compositional score matrix adjustment
Filters and Masking
Filter
Low complexity regions
Mask
Mask for lookup table only
Mask lower case letters
PSI/PHI BLAST
Upload PSSM
Optional
PSI-BLAST Threshold
0.001
Default = 0.005
BLAST
Search database swissprot using PSI-BLAST (Position-Specific Iterated BLAST)


NCBI Blast:gi|4504425|ref|NP_002119.1| high-mobility...
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Results of PSI-Blast iteration 1
Taxonomy reports
Query= gi|4504425|ref|NP_002119.1| high-mobility group box 1 [Homo sapiens]
Length=215
Distribution of 127 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
40
80
120
160
200
Done


NCBI Blast:gi|4504425|ref|NP_002119.1| high-mobility...
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Legend:
NEW - means that the alignment score was below the threshold on the previous iteration
- means that the alignment was checked on the previous iteration
Run PSI-Blast iteration 2
Hit list size 1000
Distance tree of results NEW
Sequences with E-value BETTER than threshold
Sequences producing significant alignments: Score (Bits) E Value
NEW sp|P09429|HMGB1_HUMAN High mobility group protein B1 (High mo... 310 2e-84 G
NEW sp|P10103|HMGB1_BOVIN High mobility group protein B1 (High mo... 310 2e-84 G
NEW sp|P63159|HMGB1_RAT High mobility group protein B1 (High mobi... 310 2e-84 G
NEW sp|P12682|HMGB1_PIG High mobility group protein B1 (High mobi... 308 9e-84 G
NEW sp|Q9UGV6|HMG1X_HUMAN High mobility group protein 1-like 10 (HMG 290 2e-78 G
NEW sp|P26584|HMGB2_CHICK High mobility group protein B2 (High mo... 257 2e-68 G
NEW sp|P07746|HMGT_ONCMY High mobility group-T protein (HMG-T) (HMG- 257 3e-68
NEW sp|P26583|HMGB2_HUMAN High mobility group protein B2 (High mo... 252 6e-67 G
NEW sp|P52925|HMGB2_RAT High mobility group protein B2 (High mobi... 251 1e-66 G
NEW sp|P30681|HMGB2_MOUSE High mobility group protein B2 (High mo... 249 8e-66 G
NEW sp|P17741|HMGB2_PIG High mobility group protein B2 (High mobi... 245 8e-65 G
NEW sp|P07156|HMGB1_CRIGR High mobility group protein B1 (High mo... 239 4e-63
NEW sp|P23497|SP100_HUMAN Nuclear autoantigen Sp-100 (Speckled 10... 211 1e-54 G
NEW sp|P40618|HMGB3_CHICK High mobility group protein B3 (High mo... 211 2e-54 G
NEW sp|O54879|HMGB3_MOUSE High mobility group protein B3 (High mo... 210 3e-54 G
NEW sp|Q32L31|HMGB3_BOVIN High mobility group protein B3 209 7e-54 G
NEW sp|O15347|HMGB3_HUMAN High mobility group protein B3 (High mo... 208 1e-53 G
NEW sp|Q9N1Q6|SP100_GORGO Nuclear autoantigen Sp-100 (Speckled 10... 207 2e-53
NEW sp|P36194|HMGB1_CHICK High mobility group protein B1 (High mo... 203 5e-52 G
NEW sp|Q9N1Q5|SP100_HYLLA Nuclear autoantigen Sp-100 (Speckled 10... 201 2e-51
NEW sp|Q9N1Q7|SP100_PANTR Nuclear autoantigen Sp-100 (Speckled 10... 201 2e-51
NEW sp|Q24537|HMG2_DROME High mobility group protein DSP1 (Protein d 176 6e-44 G
NEW sp|P40644|HMGH_STRPU High mobility group protein 1 homolog 152 1e-36 G
NEW sp|Q32L34|HMGB4_BOVIN High mobility group protein B4 134 3e-31 G
NEW sp|Q8WW32|HMGB4_HUMAN High mobility group protein B4 129 9e-30 G
Done

NCBI Blast:gi|4504425|ref|NP_002119.1| high-mobility...

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

Protein BLAST: search protein ... | NCBI Blast:gi|4504425|ref|NP_....

NEW	sp|Q32L68.1|HM20B_BOVIN	SWI/SNF-related matrix-associated act...	45.1	2e-04
NEW	sp|Q9Z104.1|HM20B_MOUSE	SWI/SNF-related matrix-associated act...	45.1	2e-04
NEW	sp|Q6DIJ5|HM20A_XENTR	High mobility group protein 20A (HMG bo...	45.1	2e-04
NEW	sp|Q90941|PB1_CHICK	Protein polybromo-1	45.1	2e-04
NEW	sp|Q6AZF8|HM20A_XENLA	High mobility group protein 20A (HMG bo...	44.3	4e-04
NEW	sp|P40623|HMG1B_CHITE	Mobility group protein 1B	43.9	4e-04
NEW	sp|P40622|HMG1A_CHITE	Mobility group protein 1A	43.5	6e-04
NEW	sp|Q9LEF5|SSRP1_MAIZE	FACT complex subunit SSRP1 (Facilitates...	43.5	6e-04
NEW	sp|Q91ZW1|TFAM_RAT	Transcription factor A, mitochondrial precurs	43.5	7e-04
NEW	sp|Q5ZKF4|HM20A_CHICK	High mobility group protein 20A (HMG bo...	43.5	7e-04
NEW	sp|Q5D144|TFAM_PIG	Transcription factor A, mitochondrial precurs	43.1	7e-04
NEW	sp|Q9USU7.1|YHHB_SCHPO	HMG box-containing protein C28F2.11	43.1	8e-04

Run PSI-Blast iteration 2 ← ② … ③ … ④ … ⑤

Alignments

Get selected sequences | Select all | Deselect all | Distance tree of results

```
>sp|P09429|HMGB1_HUMAN High mobility group protein B1 (High mobility group protein 1)
(HMG-1)
 sp|Q6YKA4|HMGB1_CANFA High mobility group protein B1 (High mobility group protein 1)
(HMG-1)
 sp|Q4R844|HMGB1_MACFA  High mobility group protein B1 (High mobility group protein 1)
(HMG-1)
 sp|Q08IE6|HMGB1_HORSE High mobility group protein B1 (High mobility group protein 1)
(HMG-1)
Length=215

 GENE ID: 3146 HMGB1 | high-mobility group box 1 [Homo sapiens]
(Over 100 PubMed links)

 Score =  310 bits (795),  Expect = 2e-84, Method: Compositional matrix adjust.
 Identities = 169/169 (100%), Positives = 169/169 (100%), Gaps = 0/169 (0%)

Query  1    MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKF  60
            MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKF
Sbjct  1    MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKF  60
```

Done

NCBI Blast:gi|4504425|ref|NP_002119.1| high-mobility...

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

Protein BLAST: search protein ... | NCBI Blast:gi|4504425|ref|NP_....

Results of PSI-Blast iteration 5

No new sequences were found above the 0.001 threshold!

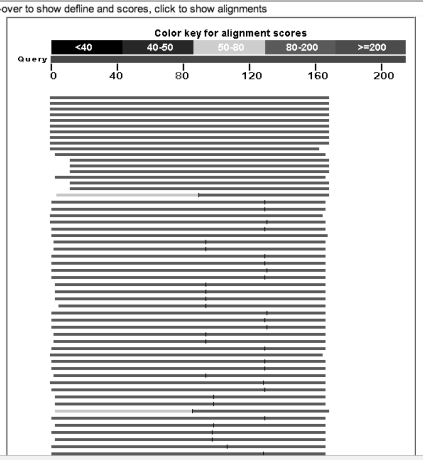
① 127 ↓ ⑤ 183

Taxonomy reports

Query= gi|4504425|ref|NP_002119.1| high-mobility group box 1 [Homo sapiens]
Length=215

Distribution of 183 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments



Done

Overview

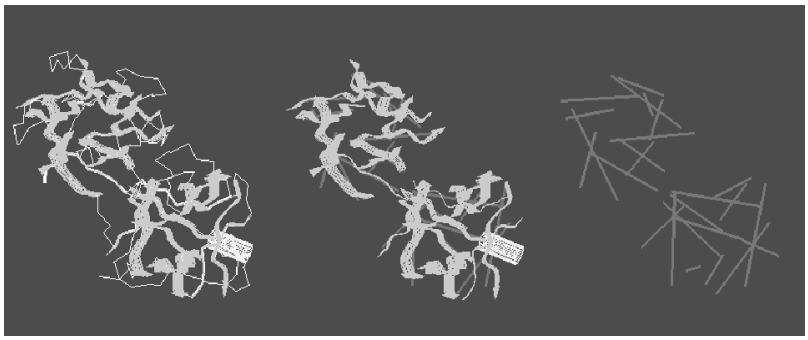
- Week 2
  - Similarity *vs*. Homology
  - Global *vs*. Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 3
  - Profiles, Patterns, Motifs, and Domains
  - Structures: VAST, Cn3D, and *de novo* Prediction
  - Multiple Sequence Alignment

Predicting Tertiary Structure

- Sequence specifies conformation, *but* conformation does *not* specify sequence
- Structure is conserved to a much greater extent than sequence
- Similarities between proteins may not necessarily be detected through "traditional" methods

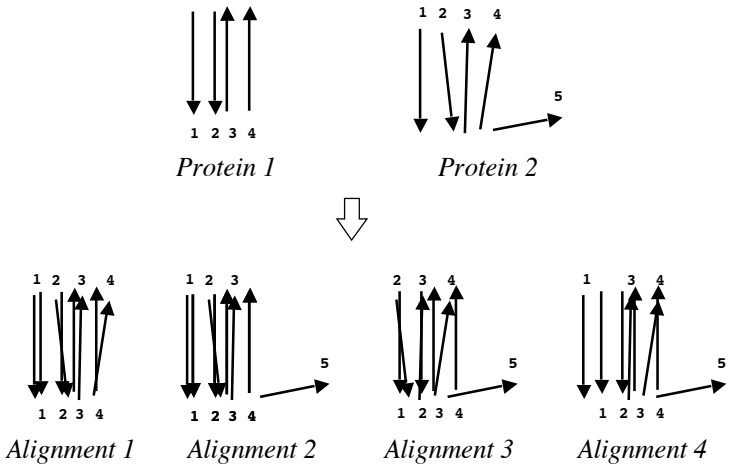
VAST Structure Comparison

Step 1: Construct vectors for secondary structure elements



VAST Structure Comparison

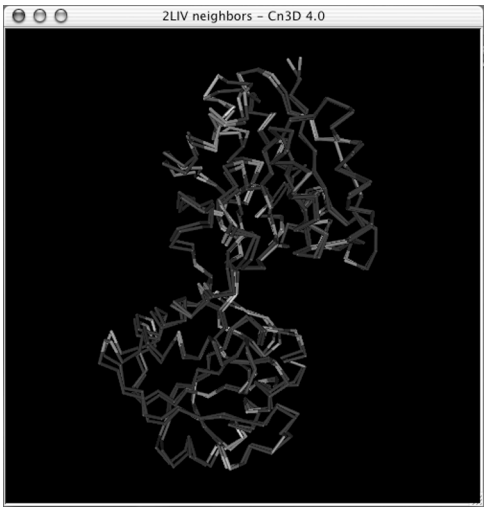
Step 2: Optimally align structure element vectors

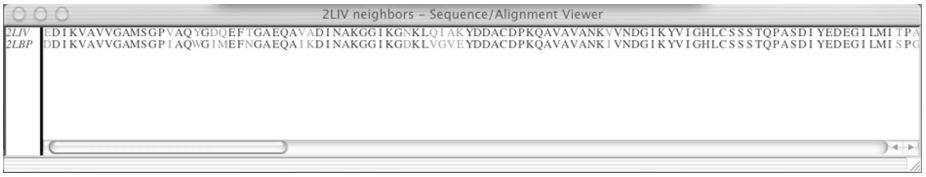


Cn3D Viewer

Rendering: Tubes

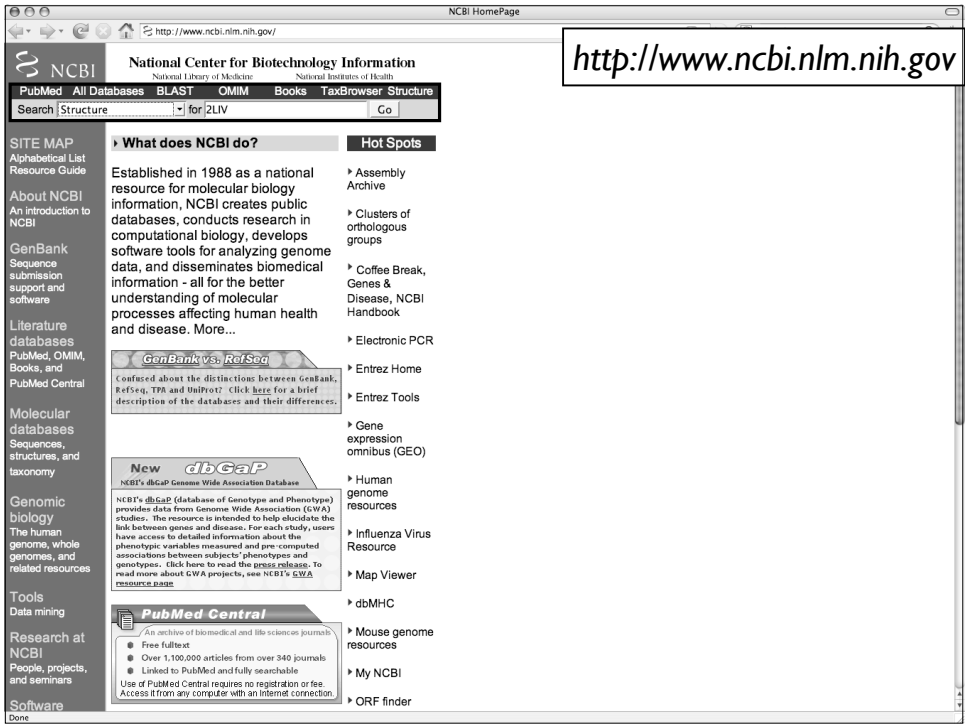
Coloring: Identity
Red – matches
Blue – mismatches



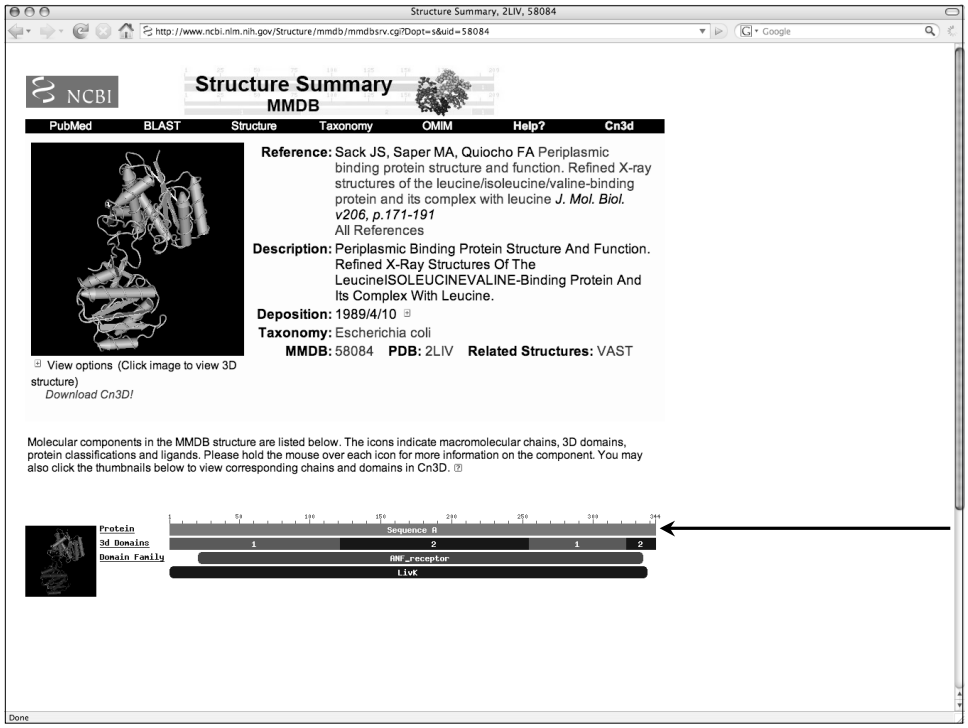


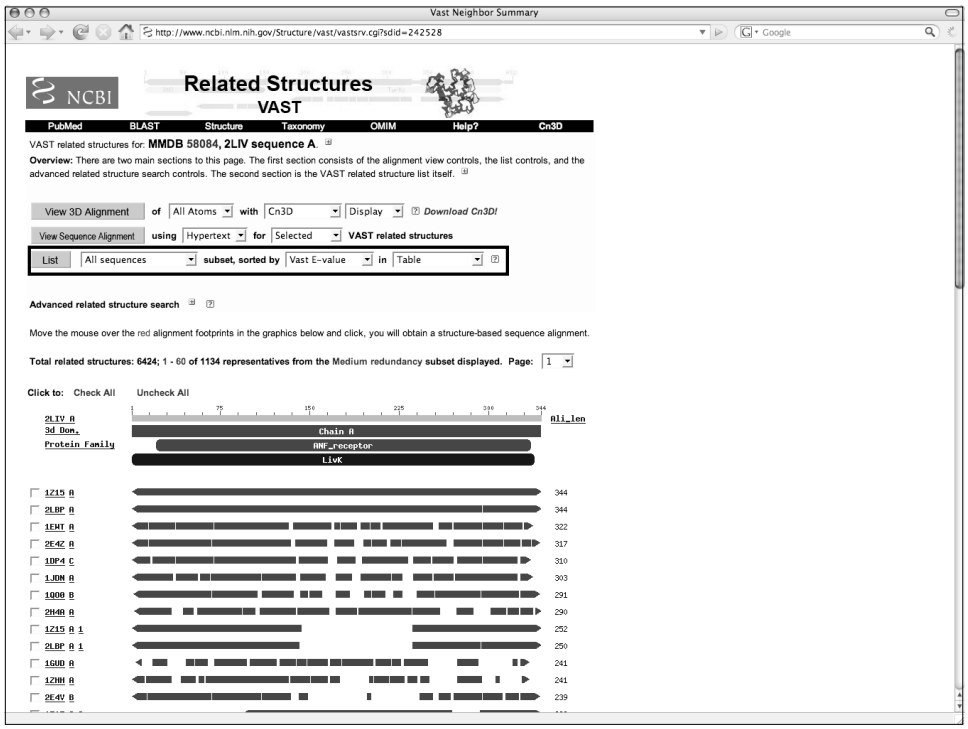
VAST Shortcomings

- Not the best method for determining structural similarities
- Reducing a structure to a series of vectors necessarily results in a loss of information (less confidence in prediction)
- Regardless of the "simplicity" of the method, provides a simple and fast first answer to the question of structural similarity


NCBI HomePage
http://www.ncbi.nlm.nih.gov/
http://www.ncbi.nlm.nih.gov
NCBI
National Center for Biotechnology Information
PubMed All Databases BLAST OMIM Books TaxBrowser Structure
Search Structure for 2LIV Go
SITE MAP
Alphabetical List
Resource Guide
About NCBI
An introduction to NCBI
GenBank
Sequence submission support and software
Literature databases
PubMed, OMIM, Books, and PubMed Central
Molecular databases
Sequences, structures, and taxonomy
Genomic biology
The human genome, whole genomes, and related resources
Tools
Data mining
Research at NCBI
People, projects, and seminars
Software
What does NCBI do?
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. More...
GenBank vs. RefSeq
Confused about the distinctions between GenBank, RefSeq, TPA and UniProt? Click here for a brief description of the databases and their differences.
New dbGaP
NCBI's dbGaP Genome Wide Association Database
PubMed Central
An archive of biomedical and life sciences journals
Free fulltext
Over 1,100,000 articles from over 340 journals
Linked to PubMed and fully searchable
Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.
Hot Spots
Assembly Archive
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
Electronic PCR
Entrez Home
Entrez Tools
Gene expression omnibus (GEO)
Human genome resources
Influenza Virus Resource
Map Viewer
dbMHC
Mouse genome resources
My NCBI
ORF finder
Done


Structure Result
http://www.ncbi.nlm.nih.gov/sites/entrez?db=structure&cmd=search&term=2LIV
NCBI
Structure
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Search Structure for 2LIV Go Clear Save Search
Limits Preview/Index History Clipboard Details
Display Summary Show 20 Sort by Send to Download Cn3D
All: 1 Bacterial: 1 Eukaryotic: 0 Ligand: 0 NMR: 0 X-ray: 1
1: 2LIV
Related Structures, Literature, Domains, Other Links
Periplasmic Binding Protein Structure And Function. Refined X-Ray Structures Of The LeucineISOLEUCINEVALINE-Binding Protein And Its Complex With Leucine [mmdbId:58084]
About Entrez
Entrez Structure
Help | FAQ
Structure Research
MMDB
CDD
PDBeast
Cn3D
VAST
VAST Search
Research
Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Privacy Statement | Freedom of Information Act | Disclaimer
Done


Structure Summary, 2LIV, 58084
http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?Dopt=s&uid=58084
NCBI
Structure Summary
MMDB
PubMed BLAST Structure Taxonomy OMIM Help? Cn3d
Reference: Sack JS, Saper MA, Quiocho FA Periplasmic binding protein structure and function. Refined X-ray structures of the leucine/isoleucine/valine-binding protein and its complex with leucine J. Mol. Biol. v206, p.171-191
All References
Description: Periplasmic Binding Protein Structure And Function. Refined X-Ray Structures Of The LeucineISOLEUCINEVALINE-Binding Protein And Its Complex With Leucine.
Deposition: 1989/4/10
Taxonomy: Escherichia coli
MMDB: 58084 PDB: 2LIV Related Structures: VAST
View options (Click image to view 3D structure)
Download Cn3D!
Molecular components in the MMDB structure are listed below. The icons indicate macromolecular chains, 3D domains, protein classifications and ligands. Please hold the mouse over each icon for more information on the component. You may also click the thumbnails below to view corresponding chains and domains in Cn3D.
Protein
3d Domains
Domain Family
Sequence A
ANF_receptor
LivK
Done


Vast Neighbor Summary
http://www.ncbi.nlm.nih.gov/Structure/vast/vastsrv.cgi?sdid=242528
NCBI
Related Structures
VAST
PubMed BLAST Structure Taxonomy OMIM Help? Cn3D
VAST related structures for: MMDB 58084, 2LIV sequence A.
Overview: There are two main sections to this page. The first section consists of the alignment view controls, the list controls, and the advanced related structure search controls. The second section is the VAST related structure list itself.
View 3D Alignment of All Atoms with Cn3D Display Download Cn3D!
View Sequence Alignment using Hypertext for Selected VAST related structures
List All sequences subset, sorted by Vast E-value in Table
Advanced related structure search
Move the mouse over the red alignment footprints in the graphics below and click, you will obtain a structure-based sequence alignment.
Total related structures: 6424; 1 - 60 of 1134 representatives from the Medium redundancy subset displayed. Page: 1
Click to: Check All Uncheck All
2LIV A
3d Dom.
Protein Family
Chain A
ANF_receptor
LivK
Ali_Len
1Z15 A 344
2LBP A 344
1EWT A 322
2E4Z A 317
1DP4 C 310
1JDN A 303
1QOO B 291
2H4B A 290
1Z15 A 1 252
2LBP A 1 250
1GUD A 241
1ZHH A 241
2E4V B 239

Vast Neighbor Summary

http://www.ncbi.nlm.nih.gov/Structure/vast/vastsrv.cgi?reqid=&sdid=242528&allbfid=14573001%2C24251101%2C4883801%2C

NCBI

Related Structures
VAST

PubMed BLAST Structure Taxonomy OMIM Help? Cn3D

VAST related structures for: **MMDB 58084, 2LIV sequence A**.

Overview: There are two main sections to this page. The first section consists of the alignment view controls, the list controls, and the advanced related structure search controls. The second section is the VAST related structure list itself.

View 3D Alignment of All Atoms with Cn3D Display Download Cn3D!

View Sequence Alignment using Hypertext for Selected VAST related structures

List All sequences subset, sorted by Vast E-value in Table

Advanced related structure search

1 - 60 of 6424 related structures displayed. Page: 1

Click to: Check All Uncheck All

	PDB C D	Ali. Len	Score	E_Val	Rmsd	%Id	MMDB Date	LHM	GSP	Description
☐	1Z15 A	344	42.1	10e-48.8	1.3	99.7	10/2005	0.0	0.4	Crystal Structure Analysis Of Periplasmic LeuILEVAL-Binding Protein In Superopen Form
☑	2LBP A	344	39.8	10e-44.6	0.9	79.1	10/2007	0.2	0.3	Structure Of The L-Leucine-Binding Protein Refined At 2.4 Angstroms Resolution And Comparison With The Leu(Slash) Ile(Slash)val-Binding Protein Structure
☐	1USG A	343	40.1	10e-42.4	2.0	79.0	01/2004	0.2	0.6	L-Leucine-Binding Protein, Apo Form
☐	1JDP B	302	29.8	10e-22.5	4.3	13.6	10/2001	6.2	1.5	Crystal Structure Of HormoneRECEPTOR COMPLEX
☐	1YK0 A	314	29.7	10e-22.3	4.5	15.0	05/2006	6.0	1.5	Structure Of Natriuretic Peptide Receptor-C Complexed With Atrial Natriuretic Peptide
										Structure Of Natriuretic Peptide Receptor-C

P-value ≤ 0.001
and
% Identity > 25
over at least 20 residues

Read the descriptions!

Cn3D Viewer

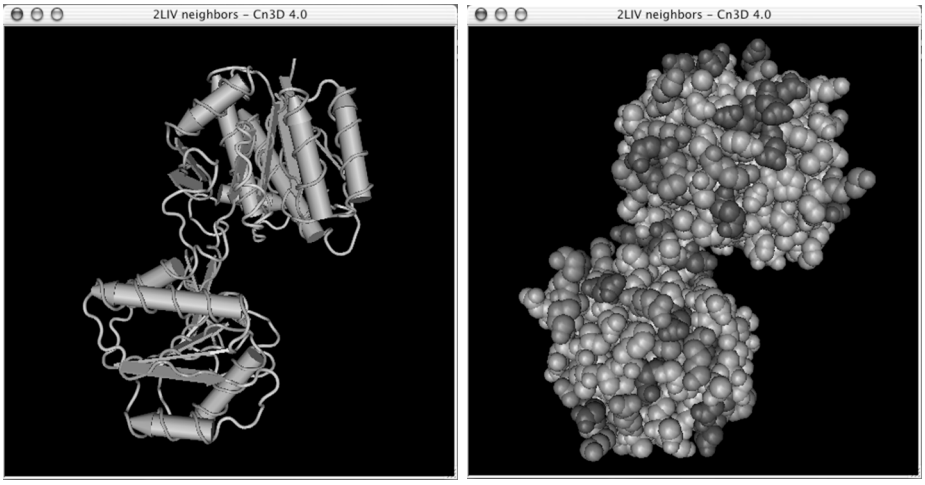
Rendering: Tubes

Coloring: Identity
Red – matches
Blue – mismatches

2LIV neighbors – Cn3D 4.0

2LIV neighbors – Sequence/Alignment Viewer

2LIV EDIKVAVVGAMSGPVAQYGDQEFTGAEQAVADINAKGGIKGNKLQIAKYDDACDPKQAVAVANKVVNDGIKYVIGHLCSSSTQPASDIYEDEGILMITPA
2LBP DDIKVAVVGAMSGPIAQWGIMEFNGAEQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVANKIVNDGIKYVIGHLCSSSTQPASDIYEDEGILMISPG



Worms	**Rendering**	Spacefill
Secondary Structure	**Coloring**	Charge

Current Protocols in Bioinformatics

CPBI Unit 1.3
Entrez and Cn3D

Searching the NCBI Databases Using Entrez

UNIT 1.3

One of the most widely used interfaces for the retrieval of information from biological databases is the NCBI Entrez system. Entrez capitalizes on the fact that three are pre-existing, logical relationships between the individual entries found in numerous public databases. For example, a paper in MEDLINE (or, more properly, PubMed) may describe the sequencing of a gene whose sequence appears in GenBank. The nucleotide sequence, in turn, may code for a protein product whose sequence is stored in the protein databases. The three-dimensional structure of that protein may be known, and the coordinates for that structure may appear in the structure database. Finally, the gene may have been mapped to a specific region of a given chromosome, with that information being stored in a mapping database. The existence of such natural connections, mostly biological in nature, argued for the development of a method through which all the information about a particular biological entity could be found without having to sequentially visit and query disparate databases.

Basic Protocols 1 and 2 describe simple, text-based searches, illustrating the types of information that can be retrieved through the Entrez system; Basic Protocol 2 also illustrates the use of Cn3D, a viewer that is used to visualize three-dimensional structures. The Alternate Protocol builds upon Basic Protocol 1, using additional, built-in features of the Entrez system, as well as alternative ways of issuing the initial query. The Support Protocol reviews how to save frequently issued queries.

QUERYING ENTREZ

BASIC PROTOCOL 1

The Entrez Web interface is located at *http://www.ncbi.nlm.nih.gov/Entrez*. Most of the Web pages at the NCBI Web site provide a direct link to Entrez, either as a blue bar running across the top of the page or in the left-hand sidebar. Entrez queries can also be issued from the NCBI home page (*http://www.ncbi.nlm.nih.gov*). The best way to illustrate the integrated nature of the Entrez system and to drive home the power of neighboring (see Commentary) is by considering three biological examples, described in Basic Protocols 1 and 2 and the Alternate Protocol.

Necessary Resources

Software

An up-to-date Web browser, such as Netscape Communicator, MS Internet Explorer, Apple Safari, or Mozilla Firefox

Select and search an Entrez database

1. Begin at the Entrez home page (*http://www.ncbi.nlm.nih.gov/Entrez*).
2. In the "Search across databases" text box, enter the following:

   `DOT AND "Vogelstein H"`

   Using Boolean operators such as AND, OR, and NOT is the simplest way to query the Entrez system. Note that all Boolean operators must be capitalized for the query to return the expected results.

Contributed by Andreas D. Baxevanis
Current Protocols in Bioinformatics (2006) 1.3.1-1.3.24
Copyright © 2006 by John Wiley & Sons, Inc.

Using Biological Databases
1.3.1
Supplement 13

http://nihlibrary.nih.gov
Search "Online Journals" for "Current Protocols in Bioinformatics"

SWISS-MODEL

- Automated comparative protein modelling server
- Web front-end at *http://www.expasy.org/swissmod*
- Results returned by E-mail

BLAST search to find similarities in PDB *by sequence*

↓

Select templates with sequence identity > 25% and projected model size > 20 amino acids

↓

Generate models

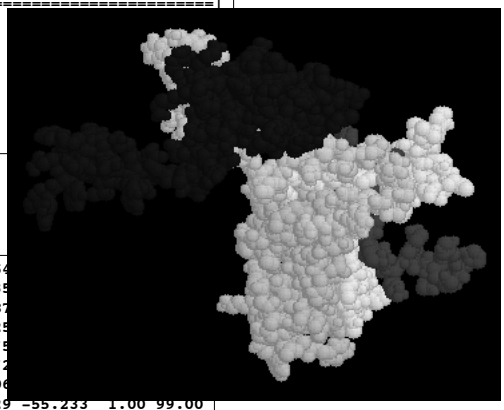
↓

Do energy minimization

↓

Generate PDB file for new protein model

```
21DJH.pdb: 42.77 % identity
21DJG.pdb: 42.77 % identity
11DJG.pdb: 42.22 % identity
11QAS.pdb: 44.17 % identity
11QAT.pdb: 43.52 % identity
21QAT.pdb: 43.52 % identity
21QAS.pdb: 43.52 % identity

Target:          |==================================================|
21DJH.pdb        |        ———-
21DJG.pdb        |        ———-
11DJG.pdb        |        ———-
11QAS.pdb        |        ———-
11QAT.pdb        |        ———-
21QAT.pdb        |        ———-
21QAS.pdb        |        ———-
```



```
ATOM      1  H1  SER     1      24.219  22.954
ATOM      2  H2  SER     1      24.770  21.435
ATOM      3  N   SER     1      24.355  22.187
ATOM      4  H3  SER     1      23.466  21.925
ATOM      5  CA  SER     1      25.266  22.675
ATOM      6  CB  SER     1      24.826  24.072
ATOM      7  OG  SER     1      24.857  25.006
ATOM      8  HG  SER     1      24.717  25.929 -55.233  1.00 99.00
ATOM      9  C   SER     1      25.471  21.750 -53.751  1.00 25.00
ATOM     10  O   SER     1      25.923  22.169 -52.684  1.00 25.00
ATOM     11  N   LYS     2      25.227  20.460 -53.972  1.00 25.00
ATOM     12  H   LYS     2      24.961  20.142 -54.878  1.00 99.00
ATOM     13  CA  LYS     2      25.366  19.408 -52.943  1.00 25.00
ATOM     14  CB  LYS     2      24.003  18.772 -52.622  1.00 25.00
```

Structural Modeling Software

- Modeller *http://www.salilab.org/modeller/*
- DeepView *http://us.expasy.org/spdbv/*
- WHAT IF *http://swift.cmbi.kun.nl*

Current Topics in Genome Analysis

Week 14
Tuesday, April 15, 2008

Protein Structure Analysis and Protein-Protein Interactions

David Wishart, Ph.D.
Departments of Computing Science and Biological Sciences
University of Alberta

Overview

- Week 2
  - Similarity *vs*. Homology
  - Global *vs*. Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 3
  - Profiles, Patterns, Motifs, and Domains
  - Structures: VAST, Cn3D, and *de novo* Prediction
  - Multiple Sequence Alignment

Why do multiple sequence alignments?

- Identify conserved regions, patterns, and domains
  - Experimental design
  - Predicting structure and function
  - Identifying new members of protein families
- Perform phylogenetic analysis
- Generate position-specific scoring matrices for subsequent searches ("many-against-one" or "one against many")
- Bolster confidence in secondary structure predictions

Considerations

- Absolute sequence similarity
  Create the alignment by lining up as many common characters as possible
- Conservation
  Take into account residues that can substitute for one another and not adversely affect the function of the protein
- Structural similarity
  Knowledge of the secondary or tertiary structure of the proteins being aligned can be used to fine-tune the alignment

General Guidelines

- As with most analyses, concentrate on the protein level rather than on the nucleotide level
  - More informative
  - Less prone to inaccurate alignment ("20 *vs.* 4")
  - Can "translate back" to nucleotide sequences *after* doing the alignment

General Guidelines

- Use a reasonable number of sequences to avoid technical difficulties
  - ***Global*** alignment method: compute time increases exponentially as sequences are added to the set
  - Most alignment algorithms are ineffective on huge data sets (and may yield inaccurate alignments)
  - Phylogenetic studies resulting from inordinately large data sets are almost impossible
  - Good starting point: 10-15 sequences
  - Ballpark upper limit: 50 sequences

General Guidelines

- Selecting sequences for alignment
  - Sequences should be of about the same length
  - Use closely-related sequences to determine "required" amino acids
  - Use more divergent sequences to study evolutionary relationships
  - Good starting point: use sequences that are 30-70% similar to most of the other sequences in the data set
  - The most informative alignments result when the sequences in the data set are not "too similar", but also not "too different"

General Guidelines

- Iterative process
  - Perform alignment on small set of sequences
  - Examine the quality of the alignment
  - If alignment good, can add new sequences to data set, then realign
  - If alignment not good, remove any sequences that result in the inclusion of long gaps, then realign

Interpretation

- Absolutely-conserved positions are ***required*** for proper structure and function
- Relatively well-conserved positions are able to tolerate limited amounts of change and not adversely affect the structure or function of the protein
- Non-conserved positions may "mutate freely," and these mutations can possibly give rise to proteins with new functions

Interpretation

- Gap-free blocks probably correspond to regions of secondary structure
- Gap-rich blocks probably correspond to unstructured or loop regions

ClustalW2

- Automatic multiple alignment of nucleotide or amino acid sequences
- Implementations
  - Client versions
    command-line text menu system, all platforms
  - Web-based version
    http://www.ebi.ac.uk/clustalw2

Progressive Alignment

- Align two sequences at a time
- Gradually build up the multiple sequence alignment by merging larger and larger sub-alignments, clustering on the basis of similarity
- Uses protein scoring matrices and gap penalties to calculate alignments having the best score
- Major advantages of method
  - Very fast
  - Alignments generally of high quality

Progressive Alignment

```
>sequence A
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST
>sequence B
VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDSLN
>sequence C
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH
>sequence D
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSH
```

Progressive Alignment

1. Calculate a similarity score (percent identity) between every pair of sequences to drive the alignment

   For N sequences, this requires the calculation of $[N \times (N - 1)] / 2$ pairwise alignments

Sequences	Alignments
4	6
10	45
25	300
50	1,225
100	4,950

Progressive Alignment

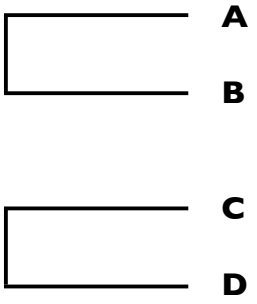
```
>sequence A
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST
>sequence B
VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDSLN
>sequence C
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH
>sequence D
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSH
```

%ID	A	B	C	D
A	100			
B	80	100		
C	44	40	100	
D	40	40	92	100

Progressive Alignment

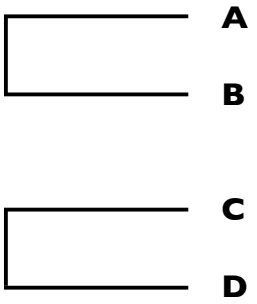
2. Derive a dendrogram (guide tree) based on the pairwise comparisons (`.dnd` file)

   Can infer from tree that A and B share greater similarity with each other than with C or D



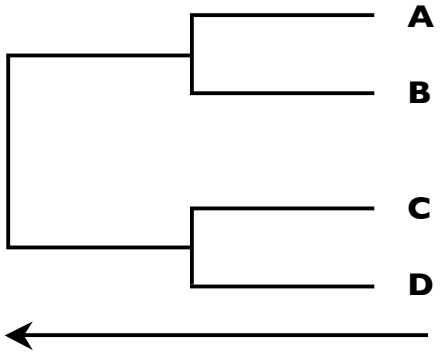
Progressive Alignment

3. Align A with B → alignment AB (fixed)
4. Align C with D → alignment CD (fixed)
5. Represent alignments AB and CD as *single sequences*



Progressive Alignment

6. Align "sequence" AB with "sequence" CD
7. Continue following the branching order of the tree, from the tips to the root, merging each new pair of "sequences"



Progressive Alignment: Advantages

- Do "easier" alignments between highly-related sequences first
- Use information regarding conservation at each position to help with more difficult alignments between more distantly-related sequences later on in process

Progressive Alignment: Disadvantages

- If initial alignments are made on distantly related sequences, there may be errors in the initial alignments
- Once an alignment is "fixed", it is not reconsidered, so any errors in the early alignments may propagate through subsequent alignments
- New version of ClustalW2 does provide a "remove first" iteration scheme to attempt to improve alignments

ClustalW2 Output

- Pairwise scores
- Multiple sequence alignment (`.aln`)
  - Alternative formats available:
    GCG, Phylip, PIR, GDE

ClustalW2 Output

- Cladogram
  - Tree assumed to be an estimate of a phylogeny
  - Branches are of equal length
  - Cladograms show common ancestry, but do not provide an indication of the amount of "evolutionary time" separating taxa
- Phylogram
  - Tree that is assumed to be an estimate of phylogeny
  - Branch lengths proportional to the amount of inferred evolutionary change

ClustalW2 Conservation Patterns

- Conservation patterns in multiple sequence alignments usually follow the following rules:

`[WYF]`	Aromatics
`[KRH]`	Basic side chains (+)
`[DE]`	Acidic side chains (–)
`[GP]`	Ends of helices
`[HS]`	Catalytic sites
`[C]`	Cysteine cross-bridges

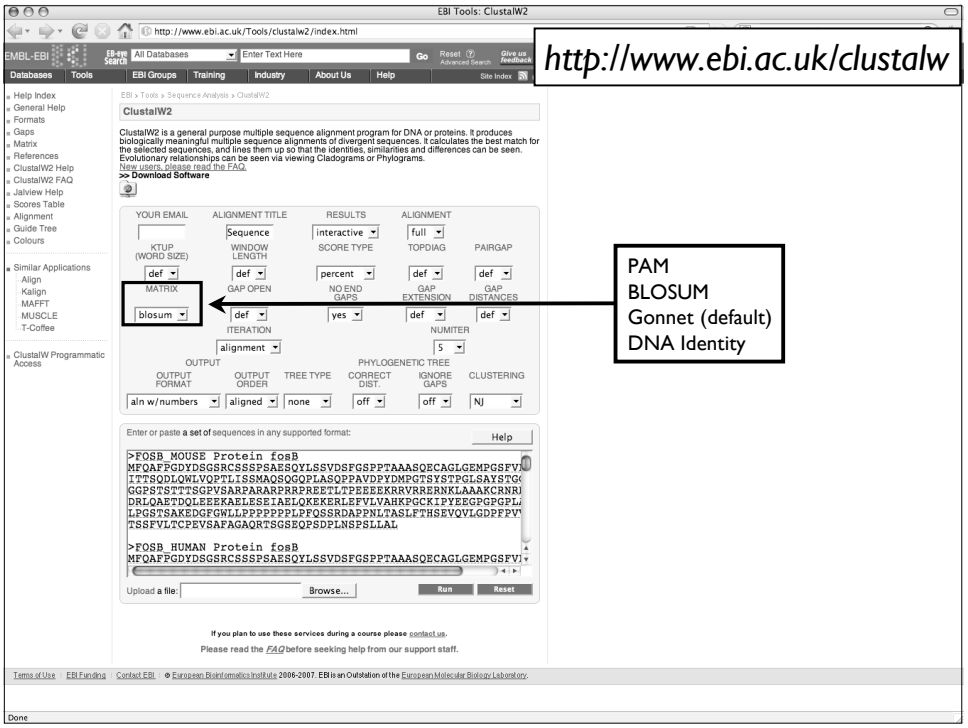
ClustalW2 Conservation Patterns

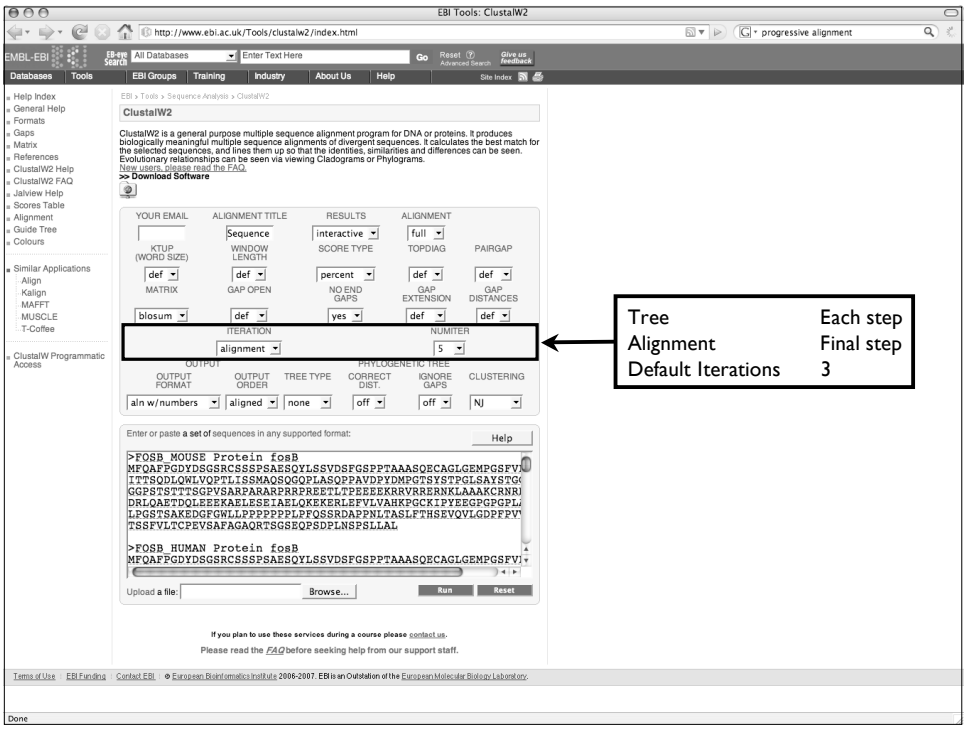
- Interpretation is *empirical* — there is no parallel to the *E*-values seen in BLAST searches to assess "significance"

*	entirely conserved column (want in at least 10% of positions)
:	"conserved" (according to color table)
.	"semi-conserved"

ClustalW Colors

`AVFPMILW`	**Red**	Small
`DE`	**Blue**	Acidic
`RK`	**Magenta**	Basic
`STYHCNGQ`	**Green**	


EBI Tools: ClustalW2
http://www.ebi.ac.uk/clustalw
PAM
BLOSUM
Gonnet (default)
DNA Identity


EBI Tools: ClustalW2
Tree Each step
Alignment Final step
Default Iterations 3

ClustalW2

http://www.ebi.ac.uk/cgi-bin/clustalw2/result?tool=clustalw2&jobid=clustalw2-20080123-02310884&treendisp=hide&treetyp progressive alignment

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- Help
- General Help
- Formats
- Gaps
- Matrix
- References
- ClustalW2 Help
- ClustalW2 FAQ
- Jalview Help
- Scores Table
- Alignment
- Guide Tree
- Colours

ClustalW2 Results

Results of search	
Number of sequences	5
Alignment score	1076
Sequence format	Pearson
Sequence type	aa
JalView	Start Jalview
Output file	clustalw2-20080123-02310884.output
Alignment file	clustalw2-20080123-02310884.aln
Guide tree file	clustalw2-20080123-02310884.dnd
Your input file	clustalw2-20080123-02310884.input

SUBMIT ANOTHER JOB

To save a result file right-click the file link in the above table and choose "Save Target As".
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

Scores Table

Sort by | Sequence Number | View Output File

SeqA	Name	Len(aa)	SeqB	Name	Len(aa)	Score
1	POSB_MOUSE	338	2	POSB_HUMAN	338	95
1	POSB_MOUSE	338	3	POS_CHICK	367	43
1	POSB_MOUSE	338	4	POS_RAT	380	43
1	POSB_MOUSE	338	5	POS_MOUSE	380	44
2	POSB_HUMAN	338	3	POS_CHICK	367	43
2	POSB_HUMAN	338	4	POS_RAT	380	43
2	POSB_HUMAN	338	5	POS_MOUSE	380	45
3	POS_CHICK	367	4	POS_RAT	380	74
3	POS_CHICK	367	5	POS_MOUSE	380	75
4	POS_RAT	380	5	POS_MOUSE	380	96

PLEASE NOTE: Some scores may be missing from the above table if the alignment was done using multiple CPU mode. Please check the output.

Sort by | Sequence Number | View Output File

Alignment

Applet ClustalTree started

ClustalW2

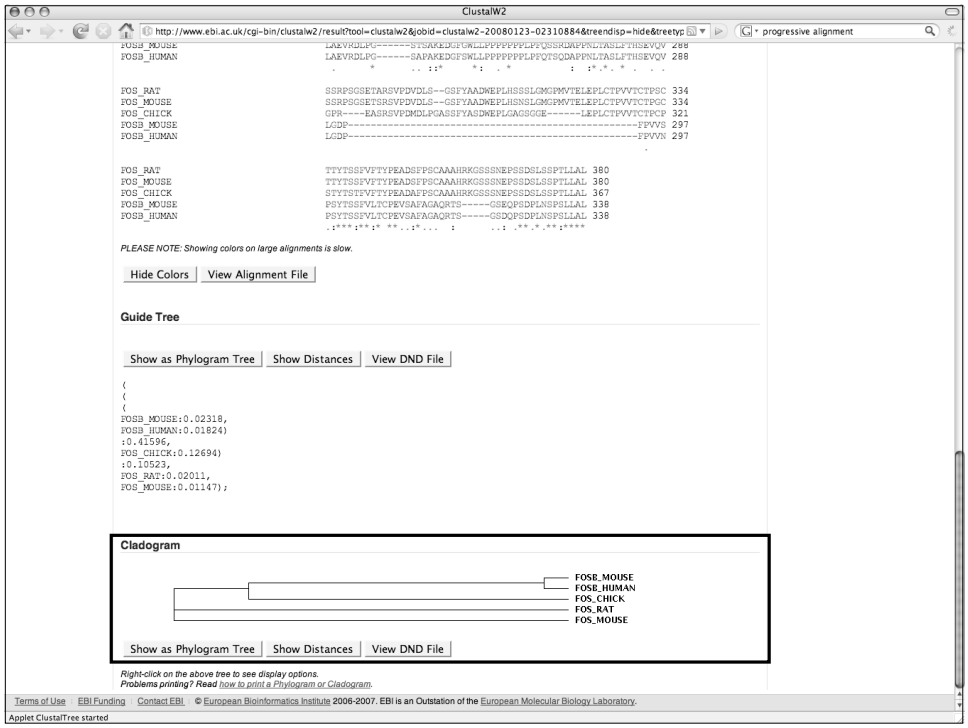
http://www.ebi.ac.uk/cgi-bin/clustalw2/result?tool=clustalw2&jobid=clustalw2-20080123-02310884&treendisp=hide&treetyp progressive alignment

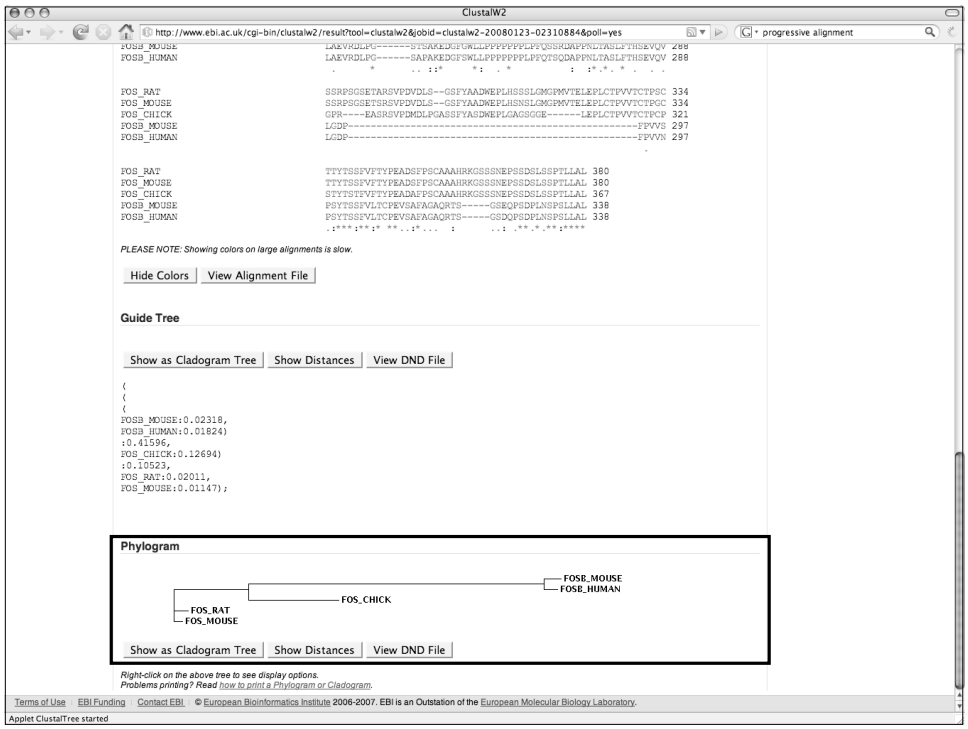
Alignment

Hide Colors | View Alignment File

```
CLUSTAL W 2.0 multiple sequence alignment


POS_RAT                         MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
POS_MOUSE                       MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
POS_CHICK                       MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANF 60
POSB_MOUSE                      -MFQAFPGDYDSGSRCSSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSF 54
POSB_HUMAN                      -MFQAFPGDYDSGSRCSSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSF 54
                                 *:..* .:*:: *  .**.*    : : * *..***.*  ... :.: *:.*.  ...*

POS_RAT                         IPTVTAISTSPDLQWLVQPTLVSSVAPSQ---TRAPHPYGLPTPSTGAYARAGVVKTMSG 117
POS_MOUSE                       IPTVTAISTSPDLQWLVQPTLVSSVAPSQ---TRAPHPYGLPTQSAGAYARAGMVKTVSG 117
POS_CHICK                       VPTVTAISTSPDLQWLVQPTLISSVAPSQ---NRG-HPYGVPAPAPPAAYSRPAVLKAPG 116
POSB_MOUSE                      VPTVTAITTSQDLQWLVQPTLISSMAQSQQQPLASQPPAVDPYDMPGTSYSTPGLSAYST 114
POSB_HUMAN                      VPTVTAITTSQDLQWLVQPTLISSMAQSQQQPLASQPPVVDPYDMPGTSYSTPGMSGYSS 114
                                :******:** *********:**:* **     .   *   *   . :      :  .

POS_RAT                         GRAQSIG--------------------RRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
POS_MOUSE                       GRAQSIG--------------------RRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
POS_CHICK                       GRGQSIG--------------------RRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 155
POSB_MOUSE                      GGASGSGGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRN 174
POSB_HUMAN                      GGASGSGGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRN 174
                                * ... *                      **  : * *:********:******:*******

POS_RAT                         RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPNDLGPPEEM 216
POS_MOUSE                       RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGPPEEM 216
POS_CHICK                       RRRELTDTLQAETDQLEEEKSALQAEIANLLKEKEKLEFILAAHRPACKMPEELRFSEEL 215
POSB_MOUSE                      RRRELTDRLQAETDQLEEEKAELSEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGP 234
POSB_HUMAN                      RRRELTDRLQAETDQLEEEKAELSEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGP 234
                                ****** ********:**: *::***:* ****:***:*.**:*.**:* :

POS_RAT                         SVTSLDLTGGLPEATTPESEEAFTPLLNDPEPKPSLEPVKNISNMELKAEPFDDFLFPA 276
POS_MOUSE                       SVASLDLTGGLPEASTPESEEAFTPLLNDPEPKPSLEPVKSISNVELKAEPFDDFLFPA 276
POS_CHICK                       AAATALDLG---APSPAAAEEAFALPLMTEAPPAVPPK-EPSGSGLELKAEPFDELLFSA 271
POSB_MOUSE                      LAEVRDLPG------STSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQV 288
POSB_HUMAN                      LAEVRDLPG------SAPAKEDGFSWLLPPPPPPLPFQYSQDAPPNLTASLFTHSEVQV 288
                                 .    *      .. ::*    *:  . *            :  :*.*. * .  . .

POS_RAT                         SSRPSGSETARSVPDVDLS--GSFYAADWEPLHSSSLGMGPMVTELEPLCTPVVTCTPSC 334
POS_MOUSE                       SSRPSGSETSRSVPDVDLS--GSFYAADWEPLHSNSLGMGPMVTELEPLCTPVVTCTPGC 334
POS_CHICK                       GPR----EASRSVPDMDLPGASSFYASDWEPLGAGSGGE------LEPLCTPVVTCTPCP 321
POSB_MOUSE                      LGDP-----------------------------------------------------PPVVS 297
POSB_HUMAN                      LGDP-----------------------------------------------------PPVVN 297
                                                                                             .

POS_RAT                         TTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 380
POS_MOUSE                       TTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 380
POS_CHICK                       STYTSTFVFTYPEADAFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 367
POSB_MOUSE                      PSYTSSFVLTCPEVSAFAGAQRTS-----GSEQPSDPLNSPSLLAL 338
POSB_HUMAN                      PSYTSSFVLTCPEVSAFAGAQRTS-----GSDQPSDPLNSPSLLAL 338
                                .:***:**:* **..:*...  :      ..: .**.*.**:****
```

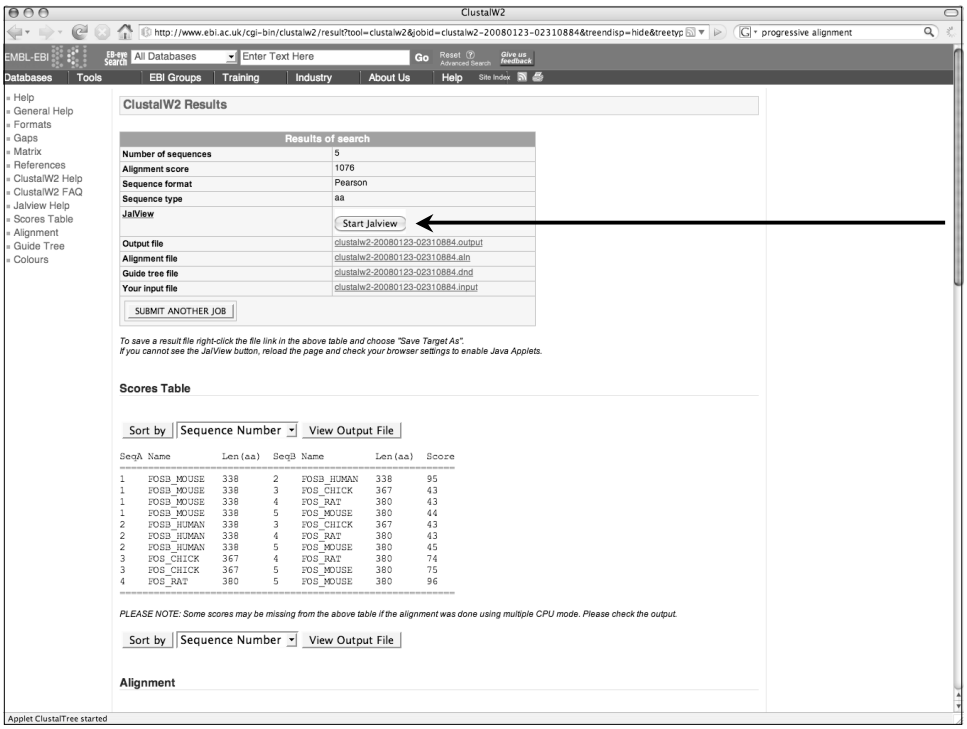
Applet ClustalTree started


ClustalW2
http://www.ebi.ac.uk/cgi-bin/clustalw2/result?tool=clustalw2&jobid=clustalw2-20080123-02310884&treendisp=hide&treetyp
progressive alignment
FOSB_MOUSE
FOSB_HUMAN
FOS_RAT
FOS_MOUSE
FOS_CHICK
PLEASE NOTE: Showing colors on large alignments is slow.
Hide Colors
View Alignment File
Guide Tree
Show as Phylogram Tree
Show Distances
View DND File
(
(
(
FOSB_MOUSE:0.02318,
FOSB_HUMAN:0.01824)
:0.41596,
FOS_CHICK:0.12694)
:0.10523,
FOS_RAT:0.02011,
FOS_MOUSE:0.01147);
Cladogram
FOSB_MOUSE
FOSB_HUMAN
FOS_CHICK
FOS_RAT
FOS_MOUSE
Show as Phylogram Tree
Show Distances
View DND File
Right-click on the above tree to see display options.
Problems printing? Read how to print a Phylogram or Cladogram.
Terms of Use
EBI Funding
Contact EBI
© European Bioinformatics Institute 2006-2007. EBI is an Outstation of the European Molecular Biology Laboratory.
Applet ClustalTree started

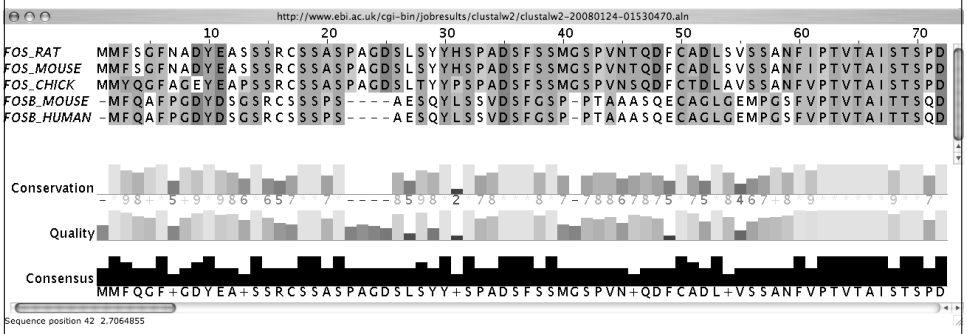

ClustalW2
http://www.ebi.ac.uk/cgi-bin/clustalw2/result?tool=clustalw2&jobid=clustalw2-20080123-02310884&poll=yes
progressive alignment
FOSB_MOUSE
FOSB_HUMAN
FOS_RAT
FOS_MOUSE
FOS_CHICK
PLEASE NOTE: Showing colors on large alignments is slow.
Hide Colors
View Alignment File
Guide Tree
Show as Cladogram Tree
Show Distances
View DND File
(
(
(
FOSB_MOUSE:0.02318,
FOSB_HUMAN:0.01824)
:0.41596,
FOS_CHICK:0.12694)
:0.10523,
FOS_RAT:0.02011,
FOS_MOUSE:0.01147);
Phylogram
FOSB_MOUSE
FOSB_HUMAN
FOS_CHICK
FOS_RAT
FOS_MOUSE
Show as Cladogram Tree
Show Distances
View DND File
Right-click on the above tree to see display options.
Problems printing? Read how to print a Phylogram or Cladogram.
Terms of Use
EBI Funding
Contact EBI
© European Bioinformatics Institute 2006-2007. EBI is an Outstation of the European Molecular Biology Laboratory.
Applet ClustalTree started

Jalview

- Java applet available within ClustalW2 results
- Used to manually edit ClustalW2 alignments
- Color residues based on various properties
- Pairwise alignment of selected sequences
- Consensus sequence calculations
- Removal of redundant sequences
- Calculation of phylogenetic trees
- Color PostScript output

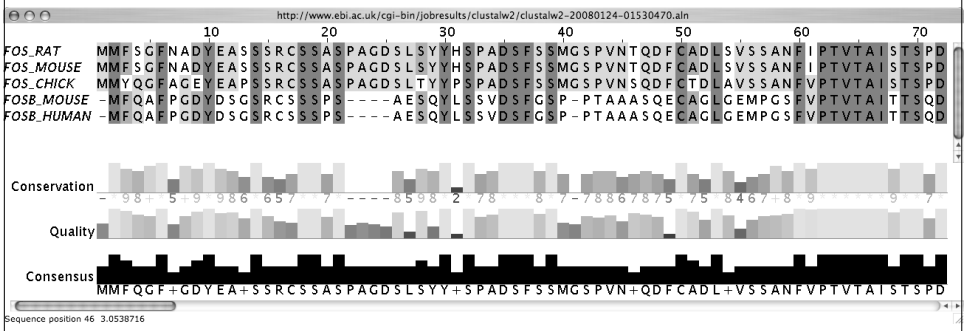


Default view



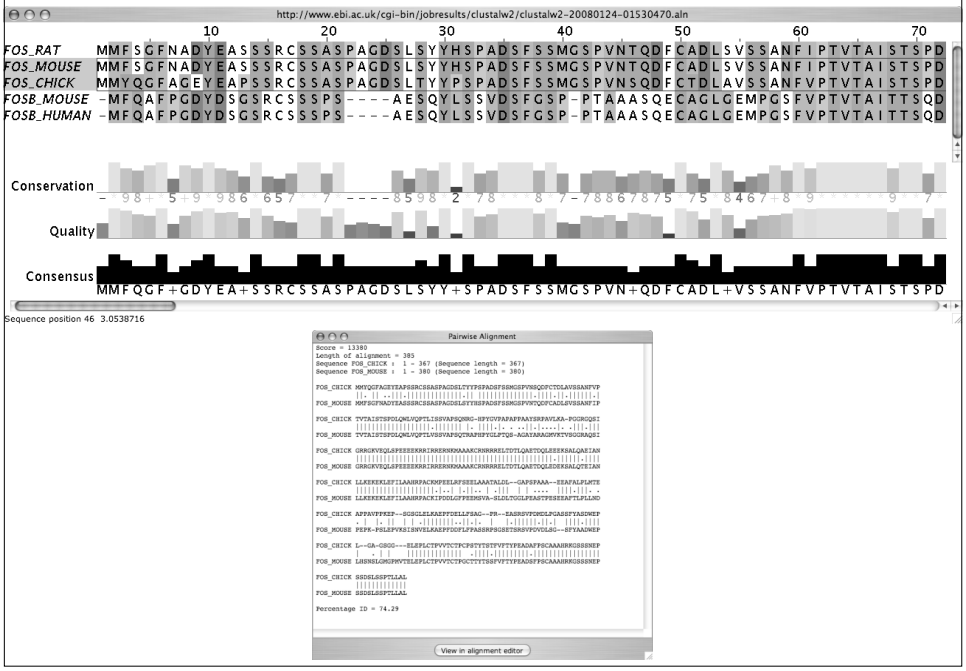
Conservation	*Conservation of total alignment (indication of percent identity)*
Quality	*Alignment quality, based on BLOSUM62 scores*
Consensus	*Based on percent identity*

Colour → Percentage Identity

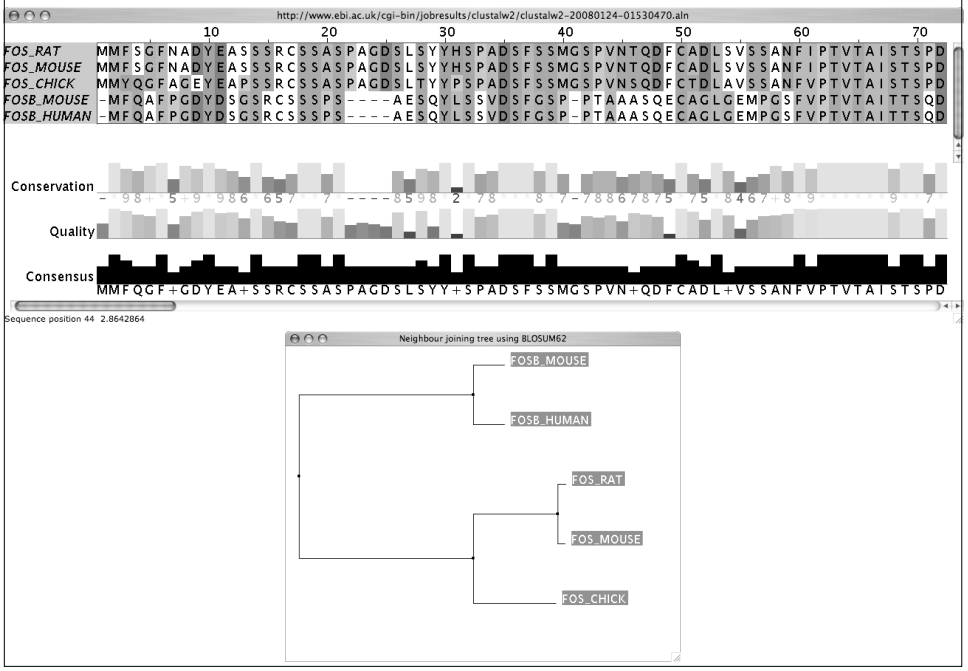


Agreement	***Background Color***
81 - 100%	Dark blue
61 - 80%	Medium blue
41 - 60%	Light blue
≤ 40%	White

Calculate → Pairwise Alignments



Calculate → Calculate Tree → Neighbour Joining Using BLOSUM62



Current Protocols in Bioinformatics

CPBI Unit 2.3
ClustalW

Multiple Sequence Alignment Using ClustalW and ClustalX

UNIT 2.3

The Clustal programs are widely used for carrying out automatic multiple alignment of sets of nucleotide or amino acid sequences. The most familiar version is ClustalW (Thompson et al., 1994), which uses a simple text menu system that is portable to more or less all computer systems. ClustalX (Thompson et al., 1997) features a graphical user interface and some powerful graphical utilities for aiding the interpretation of alignments, and is the preferred version for interactive usage. ClustalW and ClustalX are developed in parallel, and the same version-numbering system is used for both in order to synchronize changes (e.g., bug fixes, improvements, and additions). In January 2002, the latest version for both programs was 1.81. The programs can both be run interactively, but the protocols below give instructions on how to do this using ClustalX. Alternatively, ClustalW supports a full command-line interface which allows it to be used automatically as part of larger analyses (e.g., it can be run from scripts). In the simplest usage (see Basic Protocol), the programs are employed to take a set of homologous sequences (all DNA/RNA or all protein) and to produce a single multiple alignment. This covers the vast majority of Clustal usage and will be sufficient for most cases. Nonetheless, Clustal also has extensive facilities for adding sequences to existing alignments, merging existing alignments (so-called profile alignment as described in the Alternate Protocol), realignment of sections of alignment, detecting and fixing alignment errors, and basic phylogenetic analysis. Users may run Clustal remotely from several sites using the Web, or the programs may be downloaded to be run locally on PCs, Macintosh, or Unix computers (Support Protocol).

USING CLUSTALW AND CLUSTALX TO DO MULTIPLE ALIGNMENTS

BASIC PROTOCOL

The programs ClustalW and ClustalX provide alternative user interfaces to the Clustal multiple alignment software. The alignments produced by the two programs are exactly the same; the only difference between ClustalW and ClustalX is the way in which the user interacts with the program. ClustalW is now mainly used as a command-line program by Web servers and automatic batch systems, although the program does provide text menus which can be used to input sequences and perform multiple alignments. Most users who run Clustal interactively now use the graphical interface provided by ClustalX. This protocol therefore uses ClustalX (here on a Silicon Graphics Unix workstation) to illustrate the basic multiple alignment procedure. Although the example given here uses protein sequences, the same protocol can be performed with nucleic acid sequences.

Necessary Resources

Hardware

Unix (including Linux) workstation (e.g., Sun, Alpha, Silicon Graphics, PC), PC with MS Windows, or Power Macintosh

Software

ClustalW or ClustalX program (see Support Protocol)

Files

Sequences can be input to both ClustalW and ClustalX in one of seven file formats. All sequences must be in the same file. The formats that are automatically recognized are: NBRF/PIR, EMBL/Swiss-Prot, Pearson (FASTA; *APPENDIX 1B*), Clustal, GCG/MSF, GCG9/RSF, and GDE flat file. The sequences

Contributed by Julie D. Thompson, Toby J. Gibson, and Des G. Higgins
Current Protocols in Bioinformatics (2003) 2.3.1-2.3.22
Copyright © 2003 by John Wiley & Sons, Inc.

Navigating Multiple Sequence Domains
2.3.1

CPBI Unit 3.8
T-Coffee

Computing Multiple Sequence/Structure Alignments with the T-Coffee Package

UNIT 3.8

This unit describes how to assemble a multiple sequence alignment using the T-Coffee multiple sequence alignment package (Notredame et al., 2000). Although T-Coffee is slower than its close relative ClustalW (*UNIT 2.3*; Thompson et al., 1994), it is better suited for aligning distantly related sequences and mixing sequences and structures. T-Coffee is also much more flexible than most methods because it makes it possible to combine many alternative alignments into a single one, based on an estimate of consistency between the alignments. The protocols below show how such a combination can be done and how alternative alignment methods can be added as separate modules to the T-Coffee process.

This unit assumes that the user wants to align a set of sequences that are more or less homologous over their entire length (see Basic Protocol 1). These sequences may have been gathered using any appropriate database search strategy. Given such a data set, the user can assemble a multiple alignment in order to carry out family modeling (profile; see Basic Protocol 2), structural modeling (see Basic Protocol 3), or phylogenetic modeling (tree computation). This multiple sequence alignment may also be used to analyze the potential effect of nsSNPs (Non-Synonymous Single Nucleotide Polymorphisms) or simply to check whether a specific sequence is a true member of the family.

Most of this unit assumes that the user is familiar with the Unix environment (without being a specialist or a programmer; see *APPENDIX 1C* & *APPENDIX 1D*). The last section (see the Appendix at the end of the unit) is a bit more demanding in terms of computer skills, but it should be relatively straightforward to anyone with a basic knowledge of the scripting language Perl.

T-Coffee is very appropriate for generating high-quality alignments, but it is more demanding in terms of resources than other, similar programs. Given a standard 2-GHz PC with 500 Mb of memory, one should not hope to align more than 100 sequences that are up to 2000 residues long when using the default mode. This figure is simply an indication, since the memory requirement depends on the relatedness of the sequences being considered (close sequences require less time and less memory). Given these limitations, it is often a good strategy to start with a rapid multiple sequence alignment method (such as ClustalW; *UNIT 2.3*) in order to quickly identify potential problems within the data set, before refining the results with T-Coffee. Nevertheless, the authors provide two alternative strategies that make it possible to bypass some of the limitations of T-Coffee regarding memory usage.

In this unit, protocols are presented on how to use T-Coffee in a Unix/Linux environment, taking advantage of the rich command-line options of this program. Yet, for those who prefer clickable Web interfaces, much of what is presented here can also be executed by using the Web tool T-coffee@igs (Poirot et al., 2003), available at *http://igs-server.cnrs-mrs.fr/Tcoffee/*. This service is provided to the community by the CNRS and HP servers. Other online versions of this software exist, and an exhaustive list is maintained on the T-Coffee home page (accessible from the Tcoffee@igs server at the above-mentioned URL).

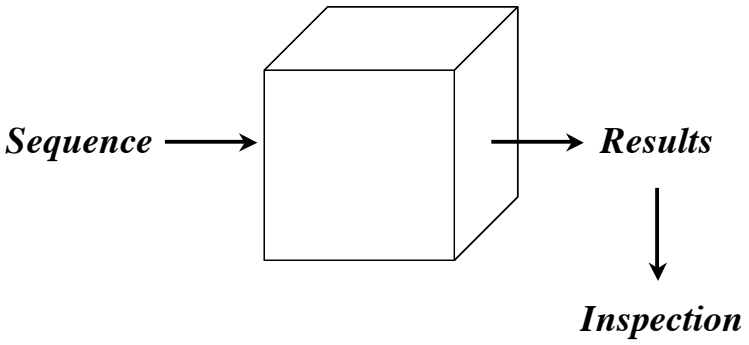
NOTE: Investigators who are unfamiliar with the Unix environment are encouraged to read *APPENDIX 1C* & *APPENDIX 1D*.

Contributed by Cedric Notredame and Karsten Suhre
Current Protocols in Bioinformatics (2003) 3.8.1-3.8.28
Copyright © 2003 by John Wiley & Sons, Inc.

Finding Similarities and Identifying Homologies
3.8.1
Supplement 4

http://nihlibrary.nih.gov
Search "Online Journals" for "Current Protocols in Bioinformatics"

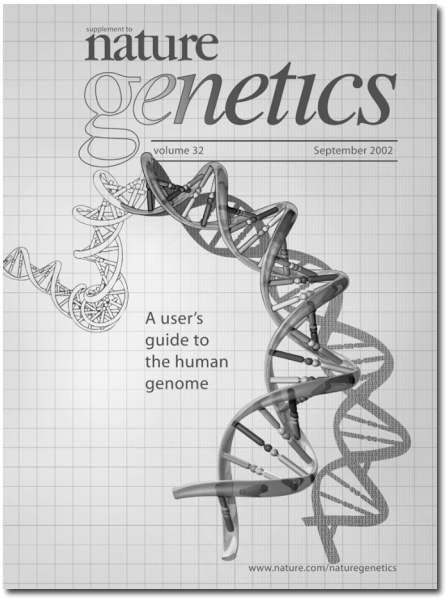
Understanding Analyses



A User's Guide to the Human Genome II

http://www.nature.com/ng/supplements/

Commentary: Keeping Biology in Mind



Current Topics in Genome Analysis

Next Lecture:

Mining Data from Genome Browsers

Tyra Wolfsberg, Ph.D.
National Human Genome Research Institute
National Institutes of Health