


*Current Topics in Genome Analysis*  
*Spring 2010*

*Week 3: Biological Sequence Analysis II*

*Andy Baxevanis, Ph.D.*



NATIONAL HUMAN GENOME RESEARCH INSTITUTE  
Division of Intramural Research

## 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



NATIONAL HUMAN GENOME RESEARCH INSTITUTE  
Division of Intramural Research

## 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

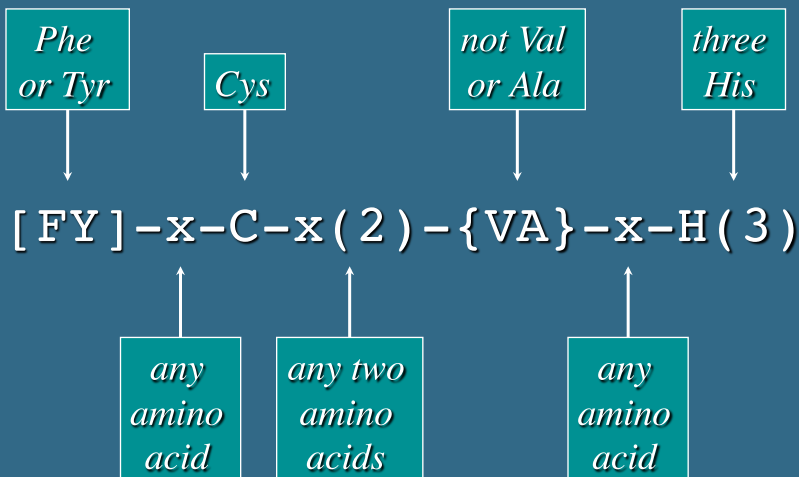
APHIIVATPG  
 GCEIVIAITPG  
 GVEICIAITPG  
 GVDILIGITPG  
 RPHEIIVATPG  
 KPHEIIVATPG  
 KVQLIIATPG  
 RPDIVIAITPG  
 APHEIIVGTPG  
 APHEIIVGTPG  
 GCHVVIATPG  
 NQDIVVATPG

- 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	16	0	13	0	0	-12	13	0	0	-3	-3	-1	-2	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	51	6	7	6	6	-11	13	11	-3	6	-16	-11	11	89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	30	60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

## Patterns



## 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://research.nhgri.nih.gov/teaching/seq_analysis.shtml)

The screenshot shows a web browser window with the URL [http://research.nhgri.nih.gov/teaching/seq\\_analysis.shtml](http://research.nhgri.nih.gov/teaching/seq_analysis.shtml). The page content includes a navigation bar, a title "Current Topics in Genome Analysis 2010", and a sub-heading "Weeks 2 and 3: Biological Sequence Analysis". Below this, there are sections for "Primer and Nucleotide Sequence for Analysis" and "BLAST". The BLAST section contains a large block of text, likely a sequence or search results. At the bottom of the page, there is a footer for the "NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research".



The screenshot shows the homepage of the Pfam database. The header includes the "welcome trust sanger institute" logo and navigation links for "HOME | SEARCH | BROWSE | FTP". A search bar is visible with the URL <http://pfam.sanger.ac.uk> and a "keyword search" button. The main content area is titled "Pfam 24.0 (October 2009, 11912 families)" and describes the database as a large collection of protein families. Below this, there are "QUICK LINKS" and "YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...". A red arrow points to the "ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES" link. The page also features a "Recent Pfam blog posts" section with several entries, including "Update Pfam searches to HMMER3.0 beta 3" and "Website update".



wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search

### Sequence search results

Show the detailed description of this results page.  
 We found 3 Pfam-A matches to your search sequence (1 significant and 2 insignificant) but we did not find any Pfam-B matches.

Show the search options and sequence that you submitted.  
 Return to the search form to look for Pfam domains on a new sequence.

**Significant Pfam-A Matches**  
 Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
p450	Cytochrome P450	Domain	n/a	41	505	41	500	1	452	345.6	2.8e-103	n/a	Show

**Insignificant Pfam-A Matches**  
 Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
COG7	Golgi complex component 7 (COG7)	Family	CL0294	189	308	247	296	317	366	11.0	0.056	n/a	Show
Sec8_exocyst	Sec8 exocyst complex component specific domain	Domain	CL0295	246	286	249	277	42	70	13.3	0.037	n/a	Show

Comments or questions on the site? Send a mail to [pfam-help@sanger.ac.uk](mailto:pfam-help@sanger.ac.uk)  
 The Wellcome Trust

wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search

### Sequence search results

Hide the detailed description of this results page.  
 The Pfam graphic below shows only the significant matches to your sequence. A significant match is one where the bits score is greater than or equal to the gathering threshold for the Pfam domain. Clicking on any of the domains in the image will take you to a page of information about that domain. Note that some Pfam-B domains may be obscured by overlapping Pfam-A domains, which are given higher priority when building the graphic.  
 Below are the details of the matches that were found. We separate Pfam-A matches into two tables, containing the significant and insignificant matches. Hits which do not start and end at the end points of the matching HMM are highlighted.  
 A small proportion of sequences within the enzymatic Pfam families have had their active sites experimentally determined. Using a strict set of rules, chosen to reduce the rate of false positives, we transfer experimentally determined active site residue data from a sequence within the same Pfam family to your query sequence. These are shown as "Predicted active sites". Full details of Pfam active site prediction process can be found in the accompanying paper.  
 For Pfam-A hits we show the alignments between your search sequence and the matching HMM. For Pfam-Bs the alignment is between your search sequence and the matching sequence from our library of Pfam-B sequences. 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 each table.  
 This alignment row for each hit shows the alignment between your sequence and the matching HMM. The alignment fragment includes the following rows:  
 #HMM: consensus of the HMM. Capital letters indicate the most conserved positions  
 #MATCH: the match between the query sequence and the HMM. A '+' indicates a positive score which can be interpreted as a conservative substitution  
 #PP: posterior probability. The degree of confidence in each individual aligned residue. 0 means 0-5%, 1 means 5-15% and so on; 9 means 85-95% and a '\*' means 95-100% posterior probability  
 #SEQ: query sequence. A '-' indicate deletions in the query sequence with respect to the HMM. Columns are coloured according to the posterior probability 0% 100%  
 You can bookmark this page and return to it later, but please use the URL that you can find in the "Search options" section below. Please note that old results may be removed after one week.  
 We found 3 Pfam-A matches to your search sequence (1 significant and 2 insignificant) but we did not find any Pfam-B matches.

Show the search options and sequence that you submitted.  
 Return to the search form to look for Pfam domains on a new sequence.

**Significant Pfam-A Matches**  
 Show or hide all alignments.

Family	Description	Entry type	Clan
p450	Cytochrome P450	Domain	n/a

```

#HMM  PpppTlpIvgnllqIq:keelheVrkiqkkygpifrkIqskpVvVIsapeavkveIikgkeefagrdcaIla:trkafkqkvI:fang..kxkklRfItptEtI:tf.....kI.sleelvoeeadIveIirkkag
#MATCH PpppIprI+g+1Iq++b1+kI++ygi+++++sgspvVvIag++k+1+kg++fng+rd++ ++gk++I+ +wRt+ +1sf +leev+ea+1+k+k
#PP 09999*****Dk*****9776555--589999999975555*****9779*****
#SEQ EPCWKLFFIGDMITLQ--KNRLSLTKISQQYQDVIQLRIGSTFVVVLSGLN1KQALVKGQDDPKRRLIYSPH--LTVKSNLTVLDEGFWAARRRLAQDLKSFIsaQdEavRSQYIIEHVSKEANHLISFQKILM
    
```

**Insignificant Pfam-A Matches**  
 Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope	Alignment	HMM	Bit score	E-value	Predicted active sites	Show/hide alignment	
				Start	End	Start	End	From	To		

**Family: p450 (PF00067)**

152 architectures 18883 sequences 2 interactions 1392 species 516 structures

### Summary

#### Cytochrome P450

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 regioselective and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].

#### Literature references

- 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](#)
- Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. [PUBMED:8405421](#)
- Nelson DR, Kamatani 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](#)
- Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. [PUBMED:2037557](#)
- Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. [PUBMED:3304150](#)
- Werck-Reichhart D, Feyereisen R; , Genome Biol 2000;1:REVIEWS3003.: Cytochromes P450: a success story. [PUBMED:11178272](#)

**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* (*Streptomyces erythraeus*).

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

**Family: p450 (PF00067)**

152 architectures 18883 sequences 2 interactions 1392 species 516 structures

### Domain organisation

Below is a listing of the unique domain organisations or architectures in which this domain is found. [More...](#)

**There are 16131 sequences with the following architecture: p450**  
 AVNA\_ASPPA [*Aspergillus parasiticus*] Averantin oxidoreductase EC=1.14.-.- (495 residues)

Show all sequences with this architecture.

**There are 1087 sequences with the following architecture: p450 x 2**  
 CI331\_XYLFA [*Xylella fastidiosa*] Putative cytochrome P450 133B1 EC=1.14.-.- (402 residues)

Show all sequences with this architecture.

**There are 137 sequences with the following architecture: p450, Flavodoxin\_1, FAD\_binding\_1, NAD\_binding\_1**  
 C805\_FUSOX [*Fusarium oxysporum*] Bifunctional P-450:NADPH-P450 reductase Cytochrome P450 505 NADPH-cytochrome P450 reductase EC=1.14.14.1 EC=1.6.2.4 (1066 residues)

Show all sequences with this architecture.

**There are 54 sequences with the following architecture: An\_peroxidase, p450**  
 Q4W941\_ASPFU [*Aspergillus fumigatus* (Sartorya fumigata)] Fatty acid oxygenase, putative EC=1.-.-.- (1136 residues)

Show all sequences with this architecture.

**There are 38 sequences with the following architecture: p450, FAD\_binding\_6, NAD\_binding\_1, Fer2**  
 A1U298\_BURMS [*Burkholderia mallei* (strain SAVP1)] Cytochrome P450 (784 residues)

Show all sequences with this architecture.

**There are 33 sequences with the following architecture: p450 x 3**  
 Q93N82\_STRLA [*Streptomyces lavendulae*] P450-related oxidase (397 residues)

Show all sequences with this architecture.

**There are 17 sequences with the following architecture: p450, KR**  
 Q629N7\_BURMA [*Burkholderia mallei* (*Pseudomonas mallei*)] Cytochrome P450-related protein (1373 residues)

Show all sequences with this architecture.

**There are 15 sequences with the following architecture: An\_peroxidase x 2, p450**  
 QQCZ99\_ASPTN [*Aspergillus terreus* (strain NIH 2624 / FGSC A1156)] Putative uncharacterized protein (1045 residues)



wellcome trust sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

### Family: p450 (PF00067)

152 architectures 18883 sequences 2 interactions 1392 species 516 structures

**Summary**

**Domain organisation**

**Alignments**

HMM logo

Trees

Curation & models

Species

Interactions

Structures

Jump to...  
enter ID/acc Go

#### 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 (18883)  
 NCBI (28613)  Metagenomics (2796)

Viewer:

**View**

#### Formatting options

Alignment:  Seed (50)  Full (18883)

Format:

Order:  Tree  Alphabetical

Sequence:  Inserts lower case  All upper case

Gaps:

Download/view:  Download  View

**Generate**

#### Download options

Very large alignments can often cause problems for the formatting tool above. If you find that downloading or viewing a large alignment is problematic, you can also download a gzip-compressed, Stockholm-format file containing the **seed** or **full** alignment for this family.

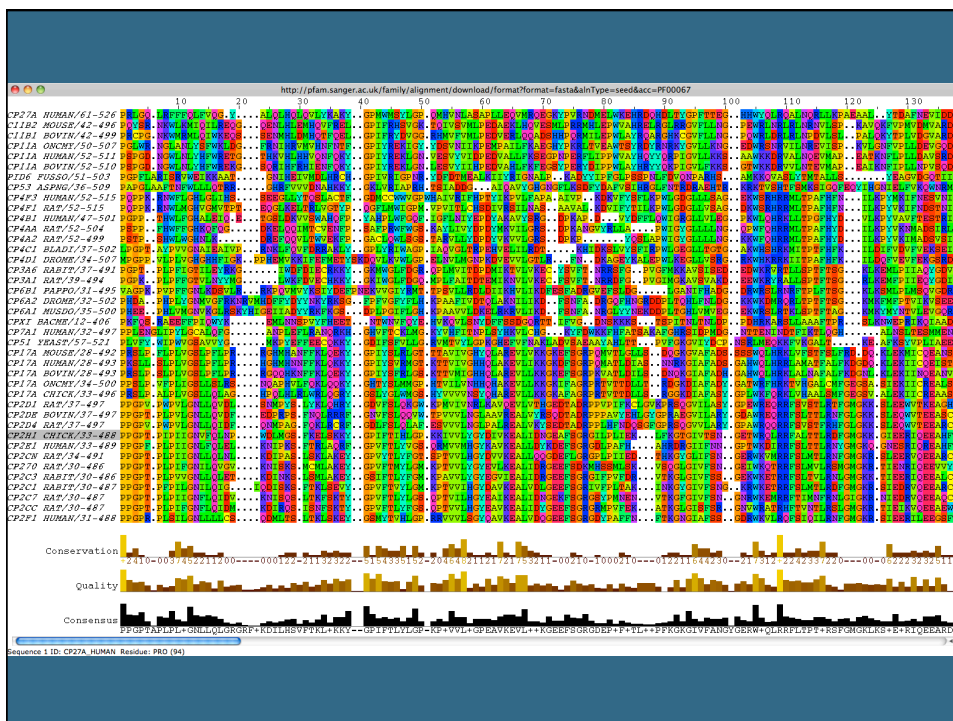
You can also **download** a FASTA format file containing the **full-length sequences** for all sequences in the full alignment.

The main seed and full alignments are generated using sequences from the UniProt sequence database. However, we also generate alignments using sequences from the NCBI sequence database and the "metaseq" metagenomics dataset.

You can view alignments from these two additional datasets using the form above, or you can download alignments of **NCBI** or **metagenomics** sequences, as gzip-compressed files.

Ffam alignments:  Seed (50)  Full (18883)  
 NCBI (28613)  Metagenomics (2796)

Full length sequences  Full-sequences (18883)



Family: p450 (PF00067)

wellcome trust sanger institute | HOME | SEARCH | BROWSE | FTP | HELP | ABOUT | Pfam keyword search

152 architectures 18863 sequences 2 interactions 1392 species 516 structures

**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 regioselective and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].

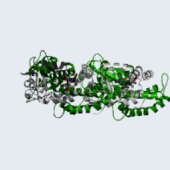
**Literature references**

- Graham-Lorence S, Amarnah B, White RE, Peterson JA, Simpson ER; , Protein Sci 1995;4:1065-1080.: A three-dimensional model of aromatase cytochrome P450. [PUBMED:7549871](#)
- Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. [PUBMED:8405421](#)
- 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](#)
- Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. [PUBMED:2037557](#)
- Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. [PUBMED:3304150](#)
- Werck-Reichhart D, Feyereisen R; , Genome Biol 2000;1:REVIEWS3003.: Cytochromes P450: a success story. [PUBMED:11178272](#)

**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* (*Streptomyces erythraeus*).

*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*



**Example structure**  
 PDB entry 2a1a: Crystal structure of ferrous dioxygen complex of D251N cytochrome P450cam  
 View a different structure:

Family: p450 (PF00067)

wellcome trust sanger institute | HOME | SEARCH | BROWSE | FTP | HELP | ABOUT | Pfam keyword search

152 architectures 18863 sequences 2 interactions 1392 species 516 structures

**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 regioselective and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].

**Literature references**

- Graham-Lorence S, Amarnah B, White RE, Peterson JA, Simpson ER; , Protein Sci 1995;4:1065-1080.: A three-dimensional model of aromatase cytochrome P450. [PUBMED:7549871](#)
- Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. [PUBMED:8405421](#)
- 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](#)
- Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. [PUBMED:2037557](#)
- Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. [PUBMED:3304150](#)
- Werck-Reichhart D, Feyereisen R; , Genome Biol 2000;1:REVIEWS3003.: Cytochromes P450: a success story. [PUBMED:11178272](#)

**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* (*Streptomyces erythraeus*).

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) - NAD(P)H:450 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	<a href="#">electron carrier activity (GO:0009055)</a> <a href="#">heme binding (GO:0020037)</a> <a href="#">iron ion binding (GO:0005506)</a> <a href="#">monooxygenase activity (GO:0004497)</a>
--------------------	---

**External database links**

HOMSTRAD:	<a href="#">p450</a>
PANDIT:	<a href="#">PF00067</a>
PRINITS:	<a href="#">PR00385</a> <a href="#">PR00359</a> <a href="#">PR00408</a> <a href="#">PR00463</a> <a href="#">PR00464</a>
PROSITE:	<a href="#">PDOC00081</a>
SCOP:	<a href="#">2cnp</a>
SYSTEMS:	<a href="#">p450</a>
External sites:	1

Comments or questions on the site? Send a mail to [pfam-help@sanger.ac.uk](mailto:pfam-help@sanger.ac.uk)  
 The Wellcome Trust

PROSITE documentation PDOC00081

http://www.expasy.org/cgi-bin/prosite-search-ac?PDOC00081

Swiss Institute of Bioinformatics

ExpASY Proteomics Server

Databases Tools Services Mirrors About Contact

You are here: ExpASY CH > Databases > PROSITE

### Cytochrome P450 cysteine heme-iron ligand signature

**Description:**

Cytochrome P450's [1,2,3,E1] are a group of enzymes involved in the oxidative metabolism of a high number of natural compounds (such as steroids, fatty acids, prostaglandins, leukotrienes, etc) as well as drugs, carcinogens and mutagens. Based on sequence similarities, P450's have been classified into about forty different families [4,5]. P450's are proteins of 400 to 500 amino acids; the only exception is *Bacillus EM-3* (CYP102) which is a protein of 1048 residues that contains a N-terminal P450 domain followed by a reductase domain. P450's are heme proteins. A conserved cysteine residue in the C-terminal part of P450's is involved in binding the heme iron in the fifth coordination site. From a region around this residue, we developed a ten residue signature specific to P450's.

**Note:**  
 The term 'cytochrome' P450, while commonly used, is incorrect as P450 are not electron-transfer proteins; the appropriate name is P450 'heme- thiolate proteins'.

**Expert(s) to contact by email:**  
 Deglyarenko K.N.

**Last update:**  
 December 2004 / Pattern and text revised.

**Technical section:**

PROSITE method (with tools and information) covered by this documentation:

CYTOCHROME\_P450, PS00086; Cytochrome P450 cysteine heme-iron ligand signature (PATTERN)

**Consensus pattern:** [FW] - [SGNH] - x - [GD] - [F] - [RKHPT] - [P] - C - [LIVMFAP] - [GAD]  
*C is the heme iron ligand*

**Sequences known to belong to this class detected by the pattern:**  
 ALL, except for P450 IIB10 from mouse, which has Lys in the first position of the pattern

**Other sequence(s) detected in Swiss-Prot:**  
 9.

- Retrieve an alignment of Swiss-Prot true positive hits:  
[Clustal format, color, condensed view](#) / [Clustal format, color](#) / [Clustal format, plain text](#) / [Fasta format](#)
- Retrieve the sequence logo from the alignment
- Taxonomic tree view of all Swiss-Prot/TrEMBL entries matching PS00086
- Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS00086
- Scan Swiss-Prot/TrEMBL entries against PS00086
- view ligand binding statistics

**Matching PDB structures:** 1AKD 1BU7 1BVY 1C6J ... [ALL]

**References:**

pfam: Family: p450 (PF00067)

http://pfam.sanger.ac.uk/family/PF00067.15

6. Werck-Reichhart D, Feyerherren R; , *Genome Biol* 2000;1:REVIEWS3003.: Cytochromes P450: a success story. [PUBMED:11178272](#)

**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* (*Streptomyces erythraeus*).

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 >5% 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**

<b>Molecular function</b>	<a href="#">electron carrier activity (GO:0009055)</a>
	<a href="#">heme binding (GO:0020037)</a>
	<a href="#">iron ion binding (GO:0005506)</a>
	<a href="#">monooxygenase activity (GO:0004497)</a>

**External database links**

<b>HOMSTRAD:</b>	<a href="#">p450</a>
<b>PANDIT:</b>	<a href="#">PF00067</a>
<b>PRINTS:</b>	<a href="#">PR00385</a> <a href="#">PR00359</a> <a href="#">PR00408</a> <a href="#">PR00463</a> <a href="#">PR00464</a>
<b>PROSITE:</b>	<a href="#">PDOC00081</a>
<b>SCOP:</b>	<a href="#">2cpp</a>
<b>SYSTERS:</b>	<a href="#">p450</a>
<b>External sites:</b>	1

Comments or questions on the site? Send a mail to [pfam-help@sanger.ac.uk](mailto:pfam-help@sanger.ac.uk)

The Wellcome Trust

**InterPro: IPR001128 Cytochrome P450**

**Protein matches**  
 Overview: sorted by AC, sorted by name, of known structure, proteins with splice variants  
 Detailed: sorted by AC, sorted by name, of known structure, proteins with splice variants  
 Table: For all matching proteins, of known structure

**UniProtKB Matches:**  
 19783 proteins  
 Architectures  
 Accession List  
 # Matches in BioMart

**Accession**  
 IPR001128 Cyt\_P450

**Type**  
 Family

Database	ID	Name	Proteins
Gene3D	G3DSA_1.10.630.10	Cyt_P450	18971
Pfam	PF00067	p450	19068
PANTHER	PTHR19393	Cyt_P450	18787
SuperFamily	SF48264	Cytochrome_P450	19318

**Signatures**  
 # Signatures in BioMart

**InterPro Relationships**

- Children**
  - IPR002397 Cytochrome P450, B-class
  - IPR002401 Cytochrome P450, E-class, group I
  - IPR002402 Cytochrome P450, E-class, group II
  - IPR002403 Cytochrome P450, E-class, group IV
- Contains**
  - IPR002399 Cytochrome P450, mitochondrial, N-terminal
  - IPR017972 Cytochrome P450, conserved site
  - IPR017973 Cytochrome P450, C-terminal

**GO Term annotation**

**Function**  
 GO:0004497 monoxygenase activity  
 GO:0005506 iron ion binding  
 GO:0009055 electron carrier activity  
 GO:0020037 heme binding

**InterPro annotation**  
 # Entry Details in BioMart

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* (Streptomyces erythraeus).

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

*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

(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 >65% identity.

Cytochrome P450 proteins can also be grouped by two different schemes. One scheme is based on the number of components in the system (microsomes). The other scheme was based on the number of components in the system: degree. Most prokaryotes and mitochondria (and fungal CYP50) have 3-component systems (iron-sulphur protein and P450). Most eukaryotic microsomes have 2-component systems (cytochrome P450 and cytochrome b5). There are exceptions to this scheme, such as 1-component systems that resemble class clusters, groups I-V, each of which may contain more than one cytochrome P450 family (e.g. CYP3A4 and CYP3A5) and further divergence into stable clusters within the B- and E-classes.

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

**Structural links**  
 PDB: [click here](#)  
 SCOP: [a104.1.1](#)  
 CATH: [1.10.630.10](#)

**Database links**  
 PDB-motif: [P500088](#)  
 Enzyme: [EC:1.14](#)  
 PROSITE doc: [PDOC00081](#)  
 PANDIT: [PF00067](#)  
 COME: [PFX000236](#)

**Taxonomic coverage**

Organism	Count
17 Saccharomyces cerevisiae	
4375 Fungi	
80 Caenorhabditis elegans	
170 Nematoda	
5717 Metazoa	
138 Fruit Fly	
2736 Arthropoda	
2585 Chordata	
211 Mouse	
386 Human	
15941 Eukaryota	
Unclassified	
2 Virus	
2 Archaea	
23 Bacteria	
3815 Cyanobacteria	
160 Synechocystis PCC 6803	
1 Oryza sativa (Rice)	
1301 Arabidopsis thaliana	
456 Green Plants	
5584 Plastid Group	
5799 Other Eukaryotes	
41	

**Overlapping InterPro entries**

IPR001128	Numbers of overlapping proteins	Average numbers of overlapping amino acids
IPR002397	17068	2715
IPR002399	19749	34
IPR002401	7972	11811
IPR002402	19350	433
IPR002403	17365	2418
IPR002404	19779	4
IPR002474	19560	223
IPR002906	19487	296
IPR008067	19711	72
IPR008068		

*Center* Tree root  
*Inner circles* Tree nodes  
*Outer circles* Representative model organisms

There is no significance to the placement of individual nodes on the circles

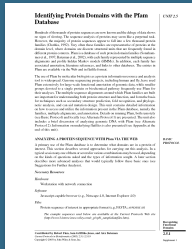
**Example proteins**

- O09158 Cytochrome P450 3A25
- Q17824 Putative cytochrome P450 cyp-13B1
- O46051 Probable cytochrome P450 4d14
- P05177 Cytochrome P450 1A2
- P10614 Lanosterol 14-alpha demethylase

**Example Proteins Key**

InterPro entry accession number/name and structure databases	Colour code
IPR017972 Cytochrome P450, conserved site	Green
IPR001128 Cytochrome P450	Red
IPR008066 Cytochrome P450, E-class, group I, CYP1	Blue
IPR017973 Cytochrome P450, C-terminal	Orange
IPR002403 Cytochrome P450, E-class, group IV	Pink
IPR002402 Cytochrome P450, E-class, group II	Yellow
IPR002401 Cytochrome P450, E-class, group I	Light Green
IPR008072 Cytochrome P450, E-class, CYP3A	Light Blue

## Further Reading



*Current Protocols in Bioinformatics*  
 Unit 2.5  
 Pfam



*Current Protocols in Bioinformatics*  
 Unit 2.7  
 InterPro

## 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
  - PRK
  - TIGRFAM



## Conserved Domain Database (CDD)

- 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



<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>

NCBI Conserved Domain Search

Conserved domains on [lclseqsig\_43d16cb872e3ad4b6afc9580b64484c]

Local query sequence

Graphical summary

Query seq.

Specific hits:

Superfamilies:

Multi-domains:

List of domain hits

Hit	Description	PssmId	Multi-dom	E-value
[h] cd05722, Ig1_Neogenin, First immunoglobulin (Ig)-like domain in neogenin and similar proteins		143139	no	2e-41
[h] d11980, Ig, Immunoglobulin domain		143743	N/A	4e-29
[h] cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma...		28945	no	8e-13
[h] cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma...		28945	no	1e-12
[h] cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma...		28945	no	3e-11
[h] cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma...		28945	no	6e-09
[h] cd00066, Ig, Immunoglobulin domain		143105	no	1e-08
[h] d11980, Ig, Immunoglobulin domain		143743	N/A	7e-07
[h] cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma...		28945	no	8e-07
[h] pfam06883, Neogenin_C, Neogenin C-terminus		115253	N/A	3e-127
[h] cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma...		140406	N/A	2e-04
[h] pfam00047, Ig, Immunoglobulin domain		109116	yes	1e-05

References:

- Marchler-Bauer A et al. (2009), "CDD: specific functional annotation with the Conserved Domain Database.", *Nucleic Acids Res.*37(D)205-10.
- 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

NCBI Conserved Domain Search

Conserved domains on [lc1seqsig\_43d16cb872e3ad4b6afcf9580b64484e]

Graphical summary

Query seq. 1 250 500 750 1000 1250 1497

Specific hits: Ig1\_Neogenin, Ig superfamily, FN3, FN3 superfamily, Neogenin\_C superfamily

ID	Description	Pssmid	Multi-dom	E-value
cd05722	Ig1_Neogenin, First immunoglobulin (Ig)-like domain in neogenin and similar proteins	143199		
cd00063	FN3, Fibronectin type 3 domain: One of three types of internal repeats found in the plasma...	28945	no	8e-13
cd00063	FN3, Fibronectin type 3 domain: One of three types of internal repeats found in the plasma...	28945	no	1e-12
cd00063	FN3, Fibronectin type 3 domain: One of three types of internal repeats found in the plasma...	28945	no	3e-11
cd00063	FN3, Fibronectin type 3 domain: One of three types of internal repeats found in the plasma...	28945	no	6e-09
cd00066	Ig, Immunoglobulin domain	143185	no	1e-08
cd00063	FN3, Fibronectin type 3 domain: One of three types of internal repeats found in the plasma...	143743	N/A	7e-07
cd00063	FN3, Fibronectin type 3 domain: One of three types of internal repeats found in the plasma...	28945	no	9e-07
cd05722	Ig1_Neogenin, First immunoglobulin (Ig)-like domain in neogenin and similar proteins	115283	N/A	3e-127
cd00063	FN3, Fibronectin type 3 domain: One of three types of internal repeats found in the plasma...	140406	N/A	2e-04

NCBI CDD cd05722

cd05722: Ig1\_Neogenin

First immunoglobulin (Ig)-like domain in neogenin and similar proteins

Ig1\_Neogenin: first immunoglobulin (Ig)-like domain in neogenin and related proteins. Neogenin is a cell surface protein which is expressed in the developing nervous system of vertebrate embryos in the growing nerve cells. It is also expressed in other embryonic tissues, and may play a general role in developmental processes such as cell migration, cell-cell recognition, and tissue growth regulation. Included in this group is the tumor suppressor protein DCC, which is deleted in colorectal carcinoma. DCC and neogenin each have four Ig-like domains followed by six fibronectin type III domains, a transmembrane domain, and an intracellular domain.

Links: Source: Smart; Taxonomy: Euteleostomi; PubMed: 6 links; Book: 2 links; Protein: Representatives, Specific Protein, Related Protein, Related Structure, Architectures; Superfamily: d11960

Statistics: PSSM-Id: 143199; View PSSM: cd05722; Aligned: 7 rows; Created: 27-Sep-2007; Updated: 30-Sep-2009

Structure: Interactive View; Aligned Rows: All 7 rows; Download Cn3D

Hierarchy: Interactive Display; Display: cd05722 Branch; Download CDTree

PubMed References:

- Neogenin: one receptor, many functions. *Int J Biochem Cell Biol.* 2007; 39(5):674-678
- Neogenin, an avian cell surface protein expressed during terminal neuronal differentiation, is closely related to the human tumor suppressor molecule deleted in colorectal cancer. *J Cell Biol.* 1994 Dec; 127(6):2029-2030
- Molecular characterization of human neogenin, a DCC-related protein, and the mapping of its gene (NEO1) to chromosomal position 15q22.3-q23. *Genomics* 1997 May 7; 41(3):414-421
- The immunoglobulin fold. Structural classification, sequence patterns and common core. *J Mol Biol.* 1994 Sep 30; 242(4):309-337
- The immunoglobulin superfamily: an insight on its tissue, species, and functional diversity. *J Mol Biol.* 1998 Apr; 46(6):989-1000
- Evolution of antigen binding receptors. *Annu Rev Immunol.* 1999; 17:109-147

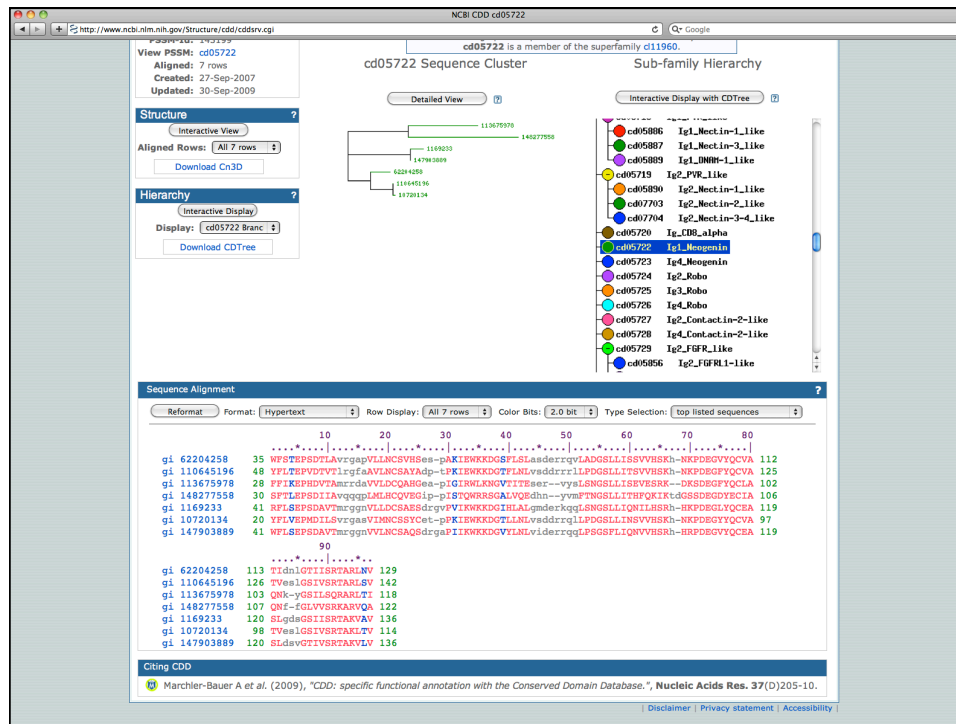
cd05722 is part of a hierarchy of related CD models. Use the graphical representation to navigate this hierarchy. cd05722 is a member of the superfamily d11960.

cd05722 Sequence Cluster

Sub-family Hierarchy

- cd05718 Ig1\_PVR\_Like
- cd05886 Ig1\_Nectin-1\_Like
- cd05887 Ig1\_Nectin-3\_Like
- cd05889 Ig1\_DNHR-1\_Like
- cd05719 Ig2\_PVR\_Like
- cd05890 Ig2\_Nectin-1\_Like
- cd07703 Ig2\_Nectin-2\_Like
- cd07704 Ig2\_Nectin-3-4\_Like
- cd05720 Ig2\_DB\_alpha
- cd05722 Ig1\_Neogenin**
- cd05723 Ig1\_Neogenin
- cd05724 Ig2\_Robo
- cd05725 Ig3\_Robo
- cd05726 Ig4\_Robo
- cd05727 Ig2\_Contactin-2\_Like
- cd05728 Ig2\_Contactin-3\_Like





## 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



## Swiss-Prot

- *Goal:* Provide a single reference sequence for each protein sequence
- Distinguishing Features
  - Non-redundancy
  - Integration with other databases (db\_xref)
  - Ongoing curation by EBI staff and *external experts*
  - Expert annotation includes editing/updates of
    - CC** Comment lines
    - FT** Feature table
  - Distinct accession series  
**[OPQ] 12345**



Protein BLAST: search protein databases using a protein query

PSI-BLAST (Position-Specific Iterated BLAST)  
PHI-BLAST (Pattern Hit Initiated BLAST)  
Choose a BLAST algorithm

**BLAST** Search database Swissprot protein sequences(swissprot) using PSI-BLAST (Position-Specific Iterated BLAST)  
 Show results in a new window

**Algorithm parameters** Note: Parameter values that differ from the default are highlighted in yellow and marked with a sign

**General Parameters**

- Max target sequences: 1000 (Default = 500)
- Short queries:  Automatically adjust parameters for short input sequences
- Expect threshold: 0.001 (Default = 10)
- Word size: 3

**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

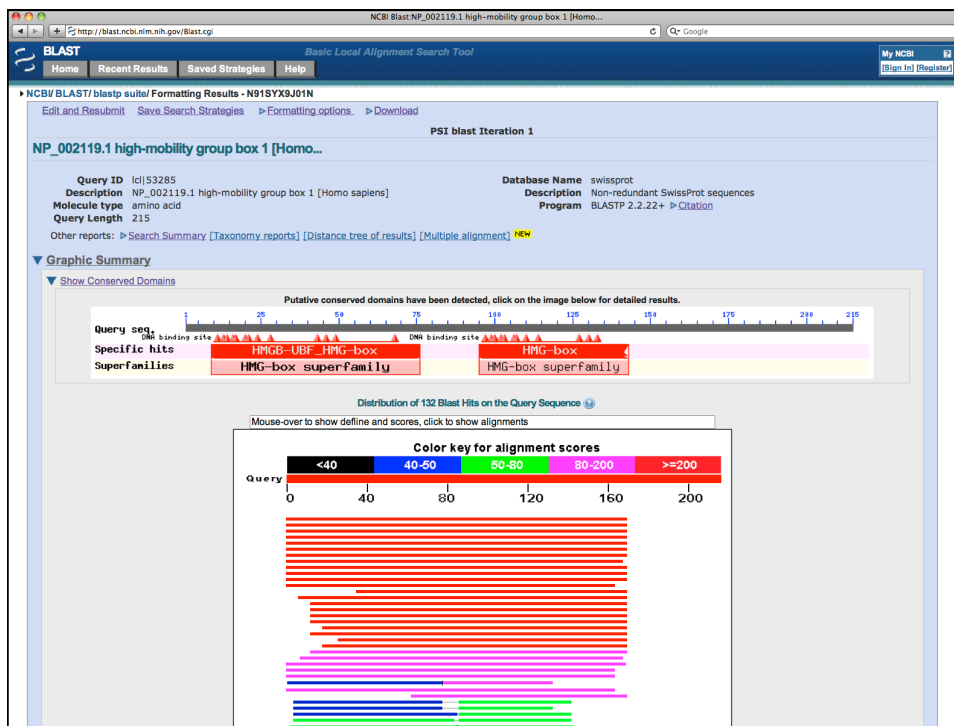
**PSI/PHI BLAST**

- Upload PSM: Choose File (no file selected)
- PSI-BLAST Threshold: 0.001 (Default = 0.005)

**BLAST** Search database Swissprot protein sequences(swissprot) using PSI-BLAST (Position-Specific Iterated BLAST)  
 Show results in a new window

Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback

NCBI | NLM | NIH | DPHS



NCBI BLAST Iteration 1

NP\_002119.1 high-mobility group box 1 [Homo...]

Query ID: |c|53285  
 Description: NP\_002119.1 high-mobility group box 1 [Homo sapiens]  
 Molecule type: amino acid  
 Query Length: 215

Database Name: swissprot  
 Description: Non-redundant SwissProt sequences  
 Program: BLASTP 2.2.22+ > Citation

Other reports: > Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] **NEW**

**Graphic Summary**

**Descriptions**

**NEW** - alignment score below the threshold on the previous iteration  
 ✓ - alignment was checked on the previous iteration

Run PSI-Blast Iteration 2 with max: 1000

Sequences with E-value BETTER than threshold

Sequences producing significant alignments:	Score (Bits)	E Value
<b>NEW</b> ✓ <a href="#">sp P09429.3 HMGGB1_HUMAN</a> RecName: Full=High mobility group pro...	310	3e-84
<b>NEW</b> ✓ <a href="#">sp P10103.3 HMGGB1_BOVIN</a> RecName: Full=High mobility group pro...	310	3e-84
<b>NEW</b> ✓ <a href="#">sp P63159.2 HMGGB1_RAT</a> RecName: Full=High mobility group prote...	310	3e-84
<b>NEW</b> ✓ <a href="#">sp P12682.3 HMGGB1_PIG</a> RecName: Full=High mobility group prote...	308	1e-83
<b>NEW</b> ✓ <a href="#">sp B2RPK0.1 HMG1A_HUMAN</a> RecName: Full=Putative high mobility ...	297	3e-80
<b>NEW</b> ✓ <a href="#">sp O9UGV6.1 HMG1X_HUMAN</a> RecName: Full=High mobility group pro...	290	4e-78
<b>NEW</b> ✓ <a href="#">sp P26584.2 HMG2B_CHICK</a> RecName: Full=High mobility group pro...	257	3e-68
<b>NEW</b> ✓ <a href="#">sp P07746.2 HMG7_ONCMY</a> RecName: Full=High mobility group-T pr...	257	4e-68
<b>NEW</b> ✓ <a href="#">sp P26583.2 HMG2B_HUMAN</a> RecName: Full=High mobility group pro...	252	9e-67
<b>NEW</b> ✓ <a href="#">sp P52925.2 HMG2B_RAT</a> RecName: Full=High mobility group prote...	251	2e-66
<b>NEW</b> ✓ <a href="#">sp P30681.3 HMG2B_MOUSE</a> RecName: Full=High mobility group pro...	249	1e-65
<b>NEW</b> ✓ <a href="#">sp P17741.2 HMG2B_PIG</a> RecName: Full=High mobility group prote...	245	1e-64
<b>NEW</b> ✓ <a href="#">sp P07156.1 HMGGB1_CRIGR</a> RecName: Full=High mobility group pro...	239	7e-63
<b>NEW</b> ✓ <a href="#">sp P23497.3 SP100_HUMAN</a> RecName: Full=Nuclear autoantigen Sp...	211	2e-54
<b>NEW</b> ✓ <a href="#">sp P49618.2 HMG2B_CHICK</a> RecName: Full=High mobility group pro...	211	3e-54
<b>NEW</b> ✓ <a href="#">sp O54879.3 HMG2B_MOUSE</a> RecName: Full=High mobility group pro...	210	4e-54
<b>NEW</b> ✓ <a href="#">sp O32L31.2 HMG2B_BOVIN</a> RecName: Full=High mobility group pro...	209	1e-53

Run PSI-Blast iteration 2 with max 1000   ← ② ... ③ ... ④ ... ⑤

*Change cutoffs to show hits "below the line"*

```

>sp|P09429.3|HMGB1_HUMAN RecName: Full=High mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1
sp|Q0YK4.3|HMGB1_CANFA RecName: Full=High mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1
sp|Q4R84.3|HMGB1_MACEA RecName: Full=High mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1
sp|Q081E6.3|HMGB1_HORSE RecName: Full=High mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1
sp|B0C999.1|HMGB1_CALJA RecName: Full=High mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1
sp|B1W8B0.1|HMGB1_COLEO RecName: Full=High mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1
sp|A9R84.1|HMGB1_PAPAN RecName: Full=High mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1
Length=215

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

Score = 310 bits (795), Expect = 3e-84, Method: Compositional matrix adjust.
Identities = 169/169 (100%), Positives = 169/169 (100%), Gaps = 0/169 (0%)

Query 1  MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKCCSERWKTMSAKEKGF 60
          MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKCCSERWKTMSAKEKGF
Sbjct 1  MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKCCSERWKTMSAKEKGF 60
    
```

PSI blast Iteration 11

NP\_002119.1 high-mobility group box 1 [Homo...]

Query ID |cl|53285  
 Description NP\_002119.1 High-mobility group box 1 [Homo sapiens]  
 Molecule type amino acid  
 Query Length 215

Database Name swissprot  
 Description Non-redundant SwissProt sequences  
 Program BLASTP 2.2.22+ [Citation](#)

**No new sequences were found above the 0.001 threshold**

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Multiple alignment](#) [NEW](#)

**Graphic Summary**

Distribution of 180 Blast Hits on the Query Sequence

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

Color key for alignment scores

Score Range	Color
<40	Black
40-50	Blue
50-80	Green
80-200	Pink
>=200	Red

Query 0 40 80 120 160 200

1: 132  
 ↓  
 11: 180

## 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



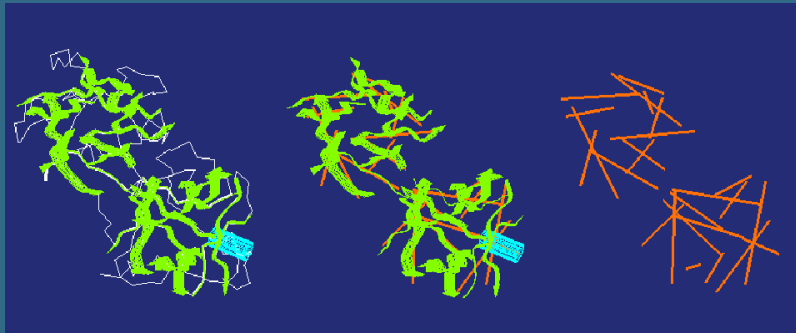
## Predicting Tertiary Structure

- 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



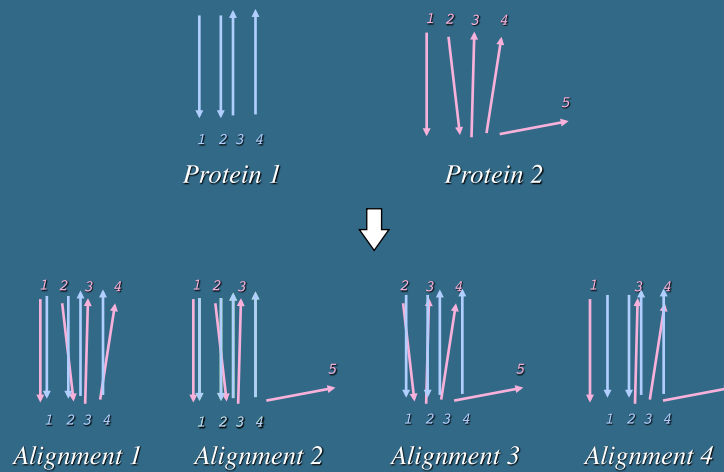
## VAST Structure Comparison

*Step 1: Construct vectors for secondary structure elements*



## VAST Structure Comparison

*Step 2: Optimally align structure element vectors*



## 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

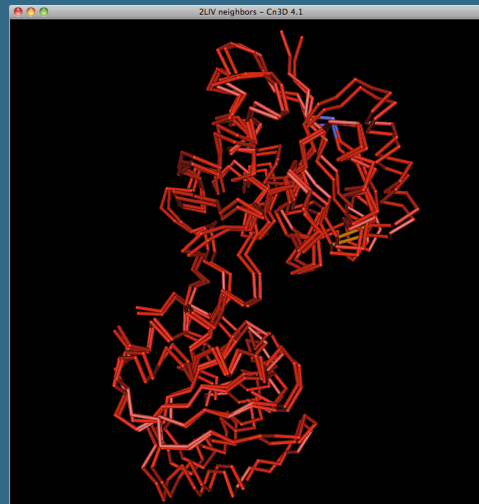
*Cn3D Viewer*

*Rendering: Tubes*

*Coloring: Identity*

*Red – matches*

*Blue – mismatches*



```
2LIV EDIENAVVQAKSEDFVAYVYDQDFPSAEGAVADINAGDITKQNSLFDI...
Neighbor EDIENAVVQAKSEDFVAYVYDQDFPSAEGAVADINAGDITKQNSLFDI...
```



The screenshot shows the NCBI homepage. At the top, there is a search bar with the text "Search Structure for 2LV" and a "Search" button. The page is divided into several sections: "Resources" on the left with a vertical menu; "Welcome to NCBI" in the center with a "PubMed Central" banner; "Popular Resources" on the right with a list of links; and "NCBI News" below it. At the bottom, there is a "FLU.GOV" widget and a URL "http://www.ncbi.nlm.nih.gov".

The screenshot shows the NCBI Structure search results page for the query "2LV". The page features a search bar at the top with "Structure" and "2LV" entered. Below the search bar, there are navigation tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The main content area displays the search results for "2LV", including a 3D protein structure visualization, a description of the protein, and its taxonomy. The description reads: "Periplasmic Binding Protein Structure And Function. Refined X-Ray Structures Of The Leucine/SOLEUCINEVALINE-Binding Protein And Its Complex With Leucine [Periplasmic Binding Protein]". The taxonomy is listed as "Escherichia coli". There are also links for "Structure Group", "Entrez Help", and "Write to the Help Desk".

Structure Summary, 2LIV, 58084

http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?uid=58084

NCBI Structure Summary MMDB

MMDB ID: 58084 PDB ID: 2LIV

Reference: Sack JS, Saper MA, Qulocho 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

Description: Periplasmic Binding Protein Structure And Function. Refined X-Ray Structures Of The Leucine/ISOLEUCINE/VALINE-Binding Protein And Its Complex With Leucine.

Deposition: 1989/4/10

Taxonomy: Escherichia coli

Related Structure: VAST

Molecular components in the MMDB structure are listed below and may include macromolecular chains, 3D domains, protein classifications (domain families), and ligands, as available. Mouse over each icon for more information on the component.

Protein 3D Domains Domain Families Specific Hits Super Families Multidomains

Sequence: 1 2

FBP1\_FBP\_L\_IVBP\_Like

Periplasmic\_Binding\_Protein\_Type\_1 superfamily

LIVK

Citing MMDB: Wang Y, Address KJ, Chen J, Geer LY, He J, He S, Lu S, Madej T, Marchler-Bauer A, Thissen PA, Zhang N, Bryant SH. "MMDB: annotating protein sequences with Entrez's 3D-structure database." Nucleic Acids Res. 2007 Jan; 35(Database Issue): 6298-300.

Vast Neighbor Summary

http://www.ncbi.nlm.nih.gov/Structure/vast/vastsrv.cgi?uid=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: 8640; 1 - 60 of 1283 representatives from the Medium redundancy subset displayed. Page: 1

Click to: Check All Uncheck All

2LIV A Chain A all len

3D Domains

Domain Families

Specific Hits

Super Families

Multidomains

1Z19 B 344

3MKT B 336

3HSL B 335

1D74 C 310

3T89 B 296

3H2V B 294

3L45 B 294

3H66 B 293

1A98 B 291

3JPH B 285

3CKH B 281

3E8E B 280

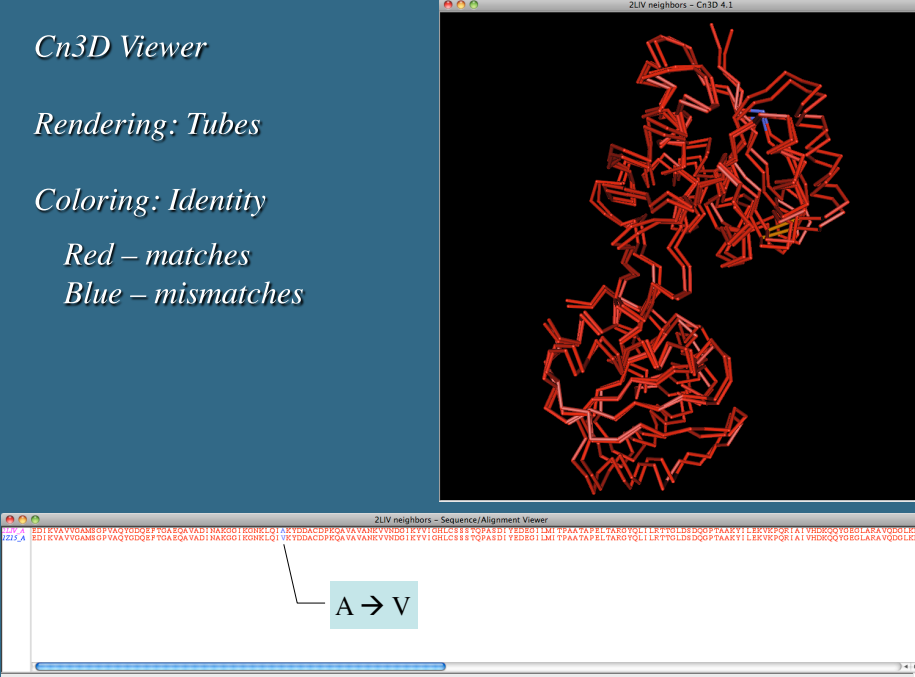
*Cn3D Viewer*

*Rendering: Tubes*

*Coloring: Identity*

*Red – matches*

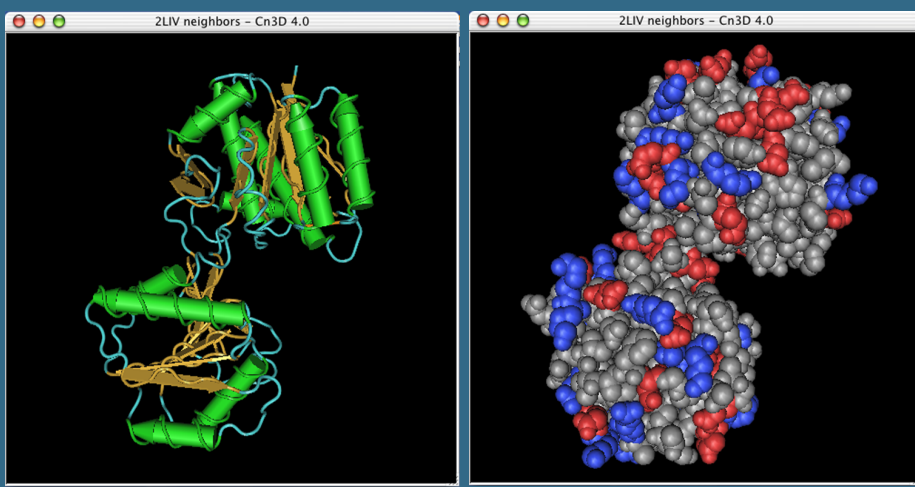
*Blue – mismatches*



2LIV neighbors - Cn3D 4.1

2LIV neighbors - Sequence/Alignment Viewer

A → V



2LIV neighbors - Cn3D 4.0

2LIV neighbors - Cn3D 4.0

*Worms*

*Secondary Structure*

*Rendering*

*Coloring*

*Spacefill*

*Charge*

NATIONAL HUMAN GENOME RESEARCH INSTITUTE  
Division of Intramural Research

## Further Reading



*Current Protocols in Bioinformatics*  
Unit 1.3  
*Entrez and Cn3D*



*Current Protocols in Bioinformatics*  
Unit 5.1  
*An Introduction to Modeling Protein Structure from Sequence*

## 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
VQLSGEKA AVLALWDKVN EEEVGG EALGRLLVVYPWTQRFFDSFGDSL N
>sequence C
VLSPADKTNVKAAWGKVGAHAGEYGA EALERMF LSFPTTKTYFPHFDLSH
>sequence D
VLSAADKTNVKA AWSKVGGHAGEYGA EALERMFLGFP TTKTYFPHFDLSH
```



## 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
VHLTPEEKSAVTALWGKVNVEVGGEEALGRLLVVYPWTQRFFESFGDLST
>sequence B
VQLSGEEKAAVLALWDKVNEEEVGGEEALGRLLVVYPWTQRFFDSFGDSL
>sequence C
VLSPADKTNVKAAWGKVGAAHAGEYGAEALERMFSLFPTTKTYFPHFDLSH
>sequence D
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPPTTKTYFPHFDLSH
```

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



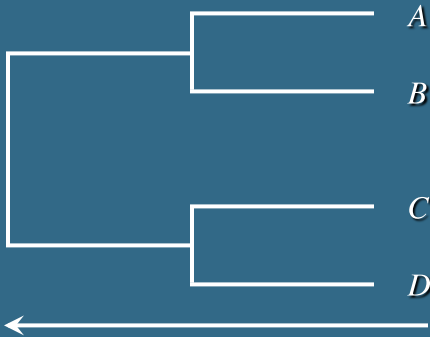
## Progressive Alignment

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



## Progressive Alignment

- Align “sequence” AB with “sequence” CD
- 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

AVFPMLW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RK	MAGENTA	Basic - H
STYHCNGQ	GREEN	Hydroxyl + sulfhydryl + amine + G
Others	Grey	Unusual amino/imino acids etc



<http://www.ebi.ac.uk/clustalw>

EMBL-EBI EBI Groups Training Industry About Us Help

EBI > Tools > Sequence Analysis > ClustalW2

### ClustalW2

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.  
 New users, please read the FAQ.  
[Download Software](#)

YOUR EMAIL:

ALIGNMENT TITLE: Sequence

RESULTS: interactive

ALIGNMENT: full

KTUP (WORD SIZE): def

WINDOW LENGTH: def

SCORE TYPE: percent

TOPDIAG: def

PAIRGAP: def

MATRIX: **def** ←

GAP OPEN: def

NO END GAPS: yes

GAP EXTENSION: def

GAP DISTANCES: def

ITERATION: alignment

NUMITER: 10

OUTPUT FORMAT: aln w/numbers

OUTPUT ORDER: aligned

TREE TYPE: none

CORRECT DIST: off

PHYLOGENETIC TREE: off

IGNORE GAPS: off

CLUSTERING: NJ

Enter or paste a set of sequences in any supported format:

```
>F05B_MOUSE Protein fosB
MFQAFPGDYDSCRCSSPSAESQYLLSVDVDFGSPPTAAASQECAGLCEMPCSFVPTVTA
ITTSQDLQWLQVFTLSSMAQSQCPILASQPPAVDYPMPKTSYTPGLSAYSTCGASG
GGPSTSTTSGPVSARPARARRRPRELTPPEEKRRVRRRNKLAACKRNRRREL
DRLQAEITDQLEEKAELEIAELQKERLRFVLVAVHKPCCKIPYEEGCPGLAEVRD
LPCSTSAKEDGFCWLLPPIPPPIPLFGQSSRDAPPNLASLFTHSEVQLGDFPVPVSPY
TSSVLTCEVSAFAGARTSGSEQSPDPLNSPLLAL

>F05B_HUMAN Protein fosB
MFQAFPGDYDSCRCSSPSAESQYLLSVDVDFGSPPTAAASQECAGLCEMPCSFVPTVTA
```

Upload a file:  no file selected

PAM  
 BLOSUM  
 Gonnet (default)  
 DNA Identity

<http://www.ebi.ac.uk/clustalw>

EMBL-EBI EBI Groups Training Industry About Us Help

EBI > Tools > Sequence Analysis > ClustalW2

### ClustalW2

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.  
 New users, please read the FAQ.  
[Download Software](#)

YOUR EMAIL:

ALIGNMENT TITLE: Sequence

RESULTS: interactive

ALIGNMENT: full

KTUP (WORD SIZE): def

WINDOW LENGTH: def

SCORE TYPE: percent

TOPDIAG: def

PAIRGAP: def

MATRIX: def

GAP OPEN: def

NO END GAPS: yes

GAP EXTENSION: def

GAP DISTANCES: def

ITERATION: **alignment** ←

NUMITER: 10

OUTPUT FORMAT: aln w/numbers

OUTPUT ORDER: aligned

TREE TYPE: none

CORRECT DIST: off

PHYLOGENETIC TREE: off

IGNORE GAPS: off

CLUSTERING: NJ

Enter or paste a set of sequences in any supported format:

```
>F05B_MOUSE Protein fosB
MFQAFPGDYDSCRCSSPSAESQYLLSVDVDFGSPPTAAASQECAGLCEMPCSFVPTVTA
ITTSQDLQWLQVFTLSSMAQSQCPILASQPPAVDYPMPKTSYTPGLSAYSTCGASG
GGPSTSTTSGPVSARPARARRRPRELTPPEEKRRVRRRNKLAACKRNRRREL
DRLQAEITDQLEEKAELEIAELQKERLRFVLVAVHKPCCKIPYEEGCPGLAEVRD
LPCSTSAKEDGFCWLLPPIPPPIPLFGQSSRDAPPNLASLFTHSEVQLGDFPVPVSPY
TSSVLTCEVSAFAGARTSGSEQSPDPLNSPLLAL

>F05B_HUMAN Protein fosB
MFQAFPGDYDSCRCSSPSAESQYLLSVDVDFGSPPTAAASQECAGLCEMPCSFVPTVTA
```

Upload a file:  no file selected

Tree Alignment Default Iterations

Each step Final step 3



ClustalW2 Results

Results of search

Number of sequences	5
Alignment score	48774
Sequence format	Pearson
Sequence type	aa

Output file: [clustalw2-20100118-170729697.out](#)  
 Alignment file: [clustalw2-20100118-170729697.ali](#)  
 Guide tree file: [clustalw2-20100118-170729697.dnd](#)  
 Your input file: [clustalw2-20100118-170729697.inp](#)

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 FOSB_MOUSE	338	2 FOSB_HUMAN	338	95
1 FOSB_MOUSE	338	3 FOS_CHICK	367	43
1 FOSB_MOUSE	338	4 FOS_RAT	380	43
1 FOSB_MOUSE	338	5 FOS_MOUSE	380	44
2 FOSB_HUMAN	338	3 FOS_CHICK	367	43
2 FOSB_HUMAN	338	4 FOS_RAT	380	43
2 FOSB_HUMAN	338	5 FOS_MOUSE	380	45
3 FOS_CHICK	367	4 FOS_RAT	380	74
3 FOS_CHICK	367	5 FOS_MOUSE	380	75
4 FOS_RAT	380	5 FOS_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

Hide Colors View Alignment File

CLUSTAL 2.0.12 multiple sequence alignment

```

FOS_RAT      MMFSGFNADYEAASSHCSSAPAGDSLSYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_MOUSE   MMFSGFNADYEAASSHCSSAPAGDSLSYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_CHICK   MMYGFAGEYAPSSHCSSAPAGDSLSYHSPADSFSSMGSPVNSQDFCTGLAVSSANF 60
FOSB_MOUSE  -MFQAFPGDYB-GS-CSS-SFSAESQ--YLSVDFSGSPPTAAASQD-CAGLGMPPGSF 54
FOSB_HUMAN  -MFQAFPGDYB-GS-CSS-SFSAESQ--YLSVDFSGSPPTAAASQD-CAGLGMPPGSF 54
*..* .:.; ***** *.: * *..*..* *.: *.: *.: *.: *.: *.: *.: *.:

FOS_RAT      IPTVTAISTSDLQMLVQPTLVSSVAFSQ-----THAPHYGLPTFS-TGAYAAAGVV 112
FOS_MOUSE   IPTVTAISTSDLQMLVQPTLVSSVAFSQ-----THAPHYGLPTFS-AGAYAAAGVV 112
FOS_CHICK   VPTVTAISTSDLQMLVQPTLVSSVAFSQ-----NIG-HPYGFAPAPFAAYSPAVL 112
FOSB_MOUSE  VPTVTAISTSDLQMLVQPTLVSSMAQQQPLASQPPVDPVYMPGTS---YSTPGLS 110
FOSB_HUMAN  VPTVTAISTSDLQMLVQPTLVSSMAQQQPLASQPPVDPVYMPGTS---YSTPGLS 110
*****.*****.*****.*****.*****.*****.*****.*****.

FOS_RAT      NTVSGGHAQSIG-----RGRVQQLSPEEEKRRIRHNNMAAA 152
FOS_MOUSE   NTVSGGHAQSIG-----RGRVQQLSPEEEKRRIRHNNMAAA 152
FOS_CHICK   PAP-GGQSIG-----RGRVQQLSPEEEKRRIRHNNMAAA 151
FOSB_MOUSE  AYSTGGASGGSPSTSTTSGPVAIPAAPRPRPEETLTPPEEKRRIRHNNMAAA 170
FOSB_HUMAN  GYSSGASGGSPSTSTTSGPVAIPAAPRPRPEETLTPPEEKRRIRHNNMAAA 170
*..* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

FOS_RAT      KGNRRHLDTLQAEQDLEEKALQTEIANLLEKLEFLIAHRPACIPNDLGF 212
FOS_MOUSE   KGNRRHLDTLQAEQDLEEKALQTEIANLLEKLEFLIAHRPACIPDDLGF 212
FOS_CHICK   KGNRRHLDTLQAEQDLEEKALQTEIANLLEKLEFLIAHRPACIPPELGF 211
FOSB_MOUSE  KGNRRHLDTLQAEQDLEEKALQTEIANLLEKLEFLIVAHFGCIPYEEG- 229
FOSB_HUMAN  KGNRRHLDTLQAEQDLEEKALQTEIANLLEKLEFLIVAHFGCIPYEEG- 229
*****.*****.*****.*****.*****.*****.*****.*****.

FOS_RAT      PEEMSVTS-LDLTGGLEPAFTPESEEAFTLLPNDPEPK-PSLEPVNISMNLEAEPFD 270
FOS_MOUSE   PEEMSVAS-LDLTGGLEPAFTPESEEAFTLLPNDPEPK-PSLEPVNISMNLEAEPFD 270
FOS_CHICK   SEELAAATLDLQ----APSPAAAEAFALPMTAPAPVPEPFSG--SGLLEAEPFD 265
FOSB_MOUSE  PEGFLAVYLDLQ----STGAEQDGLLPPPPPP-----LFFQ 267
FOSB_HUMAN  PGGFLAVYLDLQ----SAPAEQDGLLPPPPPP-----LFFQ 267
. . . * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

FOS_RAT      DFLFPASSRFGS-----ETASVVPVDLSG--SFYAADWELHSSNLSGMPVTELEPL 323
FOS_MOUSE   DFLFPASSRFGS-----ETASVVPVDLSG--SFYAADWELHSSNLSGMPVTELEPL 323
FOS_CHICK   ELFLSAGPR-----EASVVPVDLPGASSYADWELPLGASGG-----ELEPL 310
FOSB_MOUSE  -----SSDAP-PLNLA-SLFTHS-----EIVQL 289
FOSB_HUMAN  -----TSQAP-PLNLA-SLFTHS-----EIVQL 289
*.: * * * * * * * * * * * * * * * * * * * * * * * * * * * *

FOS_RAT      CTPVVTCTPCCTVYTSFVFTYFPAESFSCAAHRKGGSSNEPSSDLSLSPILLAL 380
FOS_MOUSE   CTPVVTCTPCCTVYTSFVFTYFPAESFSCAAHRKGGSSNEPSSDLSLSPILLAL 380
FOS_CHICK   CTPVVTCTPCCTVYTSFVFTYFPAESFSCAAHRKGGSSNEPSSDLSLSPILLAL 367
FOSB_MOUSE  GDFPFVVD---SYTSFVLTCEVSAF---AGAQF---TSGDQDFDLSNPSILLAL 338
FOSB_HUMAN  GDFPFVVD---SYTSFVLTCEVSAF---AGAQF---TSGDQDFDLSNPSILLAL 338
*..* * * * * * * * * * * * * * * * * * * * * * * * * * * *
    
```

The screenshot displays the ClustalW2 web interface. At the top, there are sequence alignments for four species: FOS\_RAT, FOS\_MOUSE, FOS\_CHICK, and FOS\_HUMAN. The alignments are color-coded to show conserved regions. Below the alignments, there are buttons for 'Hide Colors' and 'View Alignment File'. A 'Guide Tree' section shows a phylogenetic tree with distance values for each species. A red box highlights the 'Cladogram' section, which shows a simplified tree structure with labels for FOS\_MOUSE, FOS\_HUMAN, FOS\_CHICK, FOS\_RAT, and FOS\_MOUSE. Below the cladogram are buttons for 'Show as Phylogram Tree', 'Show Distances', and 'View DND File'. At the bottom, there is a note about right-clicking for display options and a footer with terms of use and funding information.

This screenshot is identical to the one above, showing the ClustalW2 interface with sequence alignments and a Guide Tree. However, a red box highlights the 'Phylogram' section instead of the Cladogram. The phylogram shows a tree structure where the branches are proportional to the genetic distance between species. The labels for the species are FOS\_MOUSE, FOS\_HUMAN, FOS\_CHICK, FOS\_RAT, and FOS\_MOUSE. The same buttons for 'Show as Cladogram Tree', 'Show Distances', and 'View DND File' are present below the phylogram.

## 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



**ClustalW2 Results**

Results of search

Number of sequences	5
Alignment score	48774
Sequence format	Pearson
Sequence type	aa

**JalView**

[Start Jalview](#)

**Output file** [clustalw2-20100118-1707298697\\_output](#)

**Alignment file** [clustalw2-20100118-1707298697aln](#)

**Guide tree file** [clustalw2-20100118-1707298697.dnd](#)

**Your input file** [clustalw2-20100118-1707298697.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	FOSB_MOUSE 338	2	FOSB_HUMAN 338	95
1	FOSB_MOUSE 338	3	FOSB_CHICK 367	43
1	FOSB_MOUSE 338	4	FOSB_RAT 380	43
1	FOSB_MOUSE 338	5	FOSB_MOUSE 380	44
2	FOSB_HUMAN 338	3	FOSB_CHICK 367	43
2	FOSB_HUMAN 338	4	FOSB_RAT 380	43
2	FOSB_HUMAN 338	5	FOSB_MOUSE 380	45
3	FOSB_CHICK 367	4	FOSB_RAT 380	74
3	FOSB_CHICK 367	5	FOSB_MOUSE 380	75
4	FOSB_RAT 380	5	FOSB_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](#)

**Default view**

Conservation      Conservation of total alignment (indication of percent identity)

Quality            Alignment quality, based on BLOSUM62 scores

Consensus        Based on percent identity

NATIONAL HUMAN GENOME RESEARCH INSTITUTE  
 Division of Intramural Research

**Colour → Percentage Identity**

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE  
 Division of Intramural Research

**Calculate → Pairwise Alignments**

Score = 13380  
 Length of alignment = 385  
 Sequence FOS\_CHICK : 1 - 367 (Sequence length = 367)  
 Sequence FOS\_MOUSE : 1 - 380 (Sequence length = 380)

Conservation: 9 8 + 5 + 9 - 9  
 Quality: 8 4 6 7 + 8 - 9 - - - 9 - 7 - - -  
 Consensus: MMFQGF+GDYE

Sequence 4 ID: FOSB\_MOUSE Residue: CYS (15)

NATIONAL HUMAN GENOME RESEARCH INSTITUTE  
 Division of Intramural Research

**Calculate → Calculate Tree → Neighbour Joining Using BLOSUM62**

Average distance tree using BLOSUM62

File View

- FOSB\_MOUSE
- FOSB\_HUMAN
- FOS\_RAT
- FOS\_MOUSE
- FOS\_CHICK

Sequence 4 ID: FOSB\_MOUSE Residue: CYS (15)

NATIONAL HUMAN GENOME RESEARCH INSTITUTE  
 Division of Intramural Research

## Further Reading



*Current Protocols in Bioinformatics*  
Unit 2.3  
*ClustalW*



*Current Protocols in Bioinformatics*  
Unit 3.8  
*T-Coffee*



## Understanding Analyses

