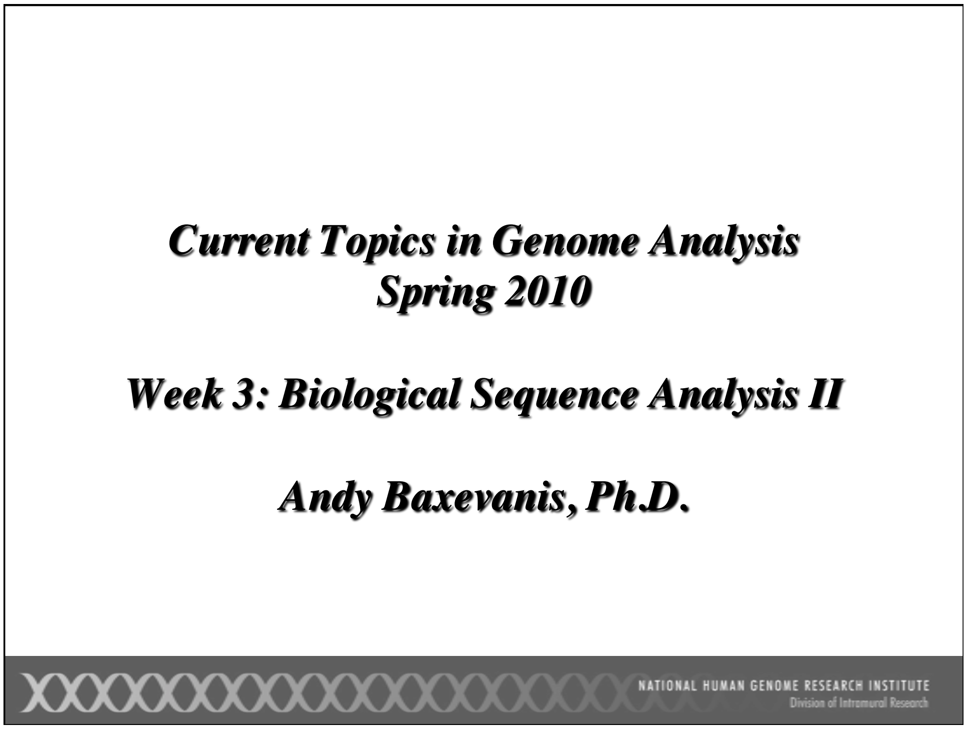
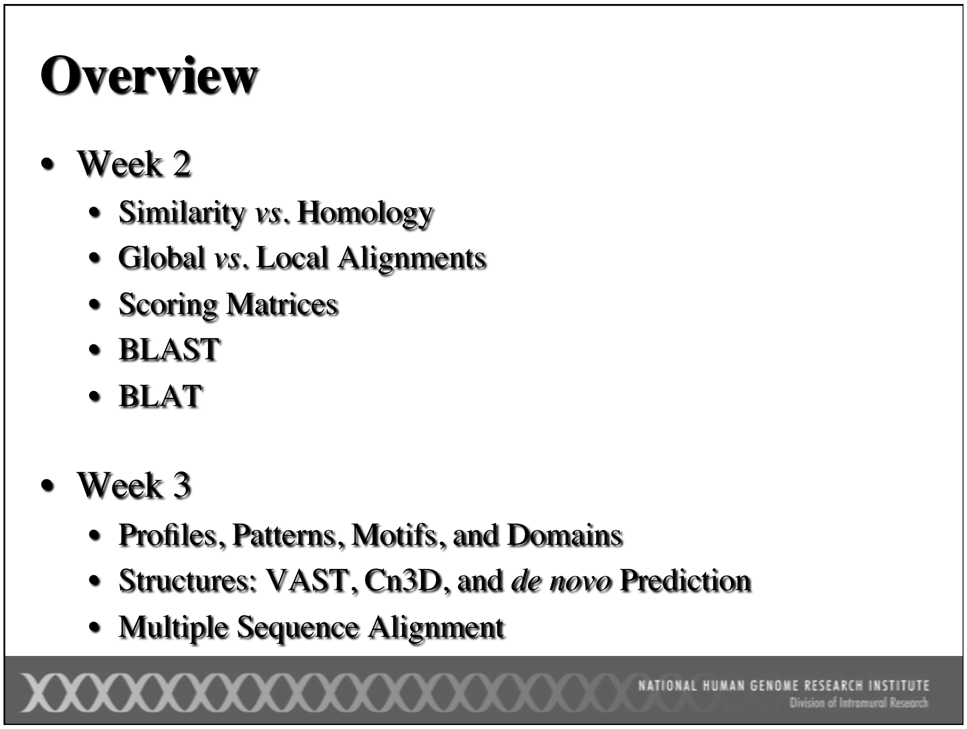


Current Topics in Genome Analysis Spring 2010

Week 3: Biological Sequence Analysis II

Andy Baxevanis, Ph.D.

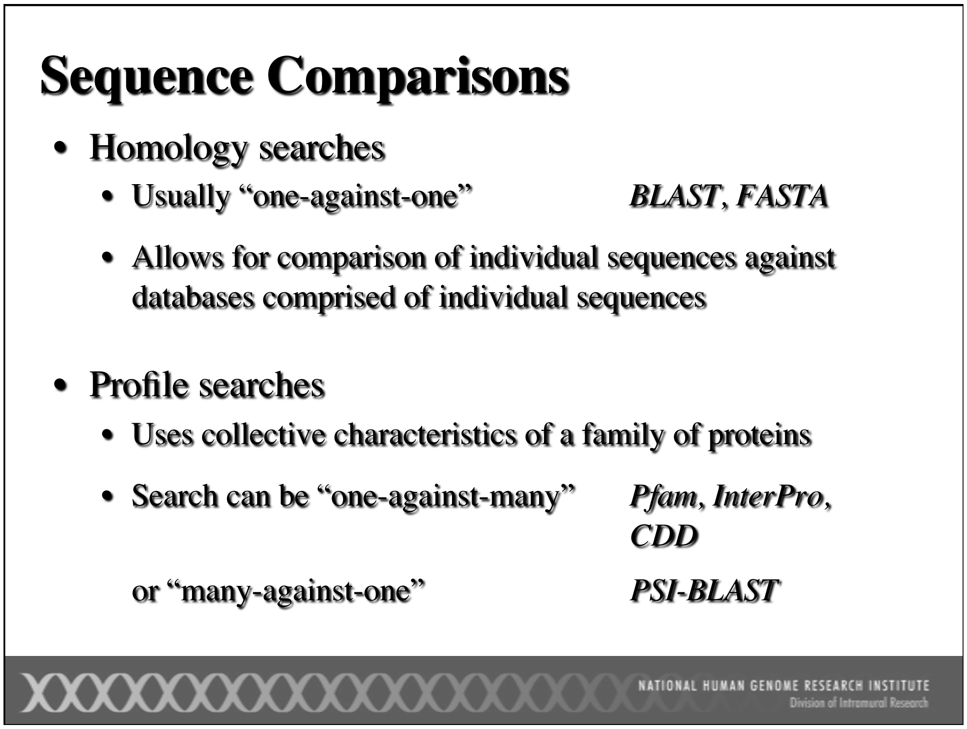
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

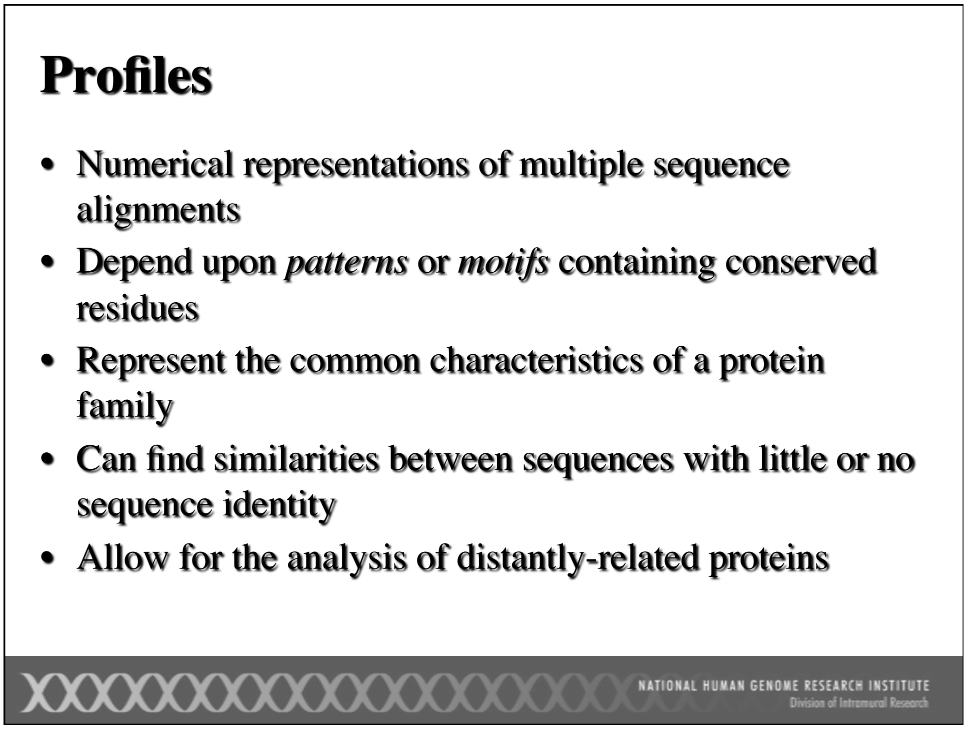


Sequence Comparisons

- 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*

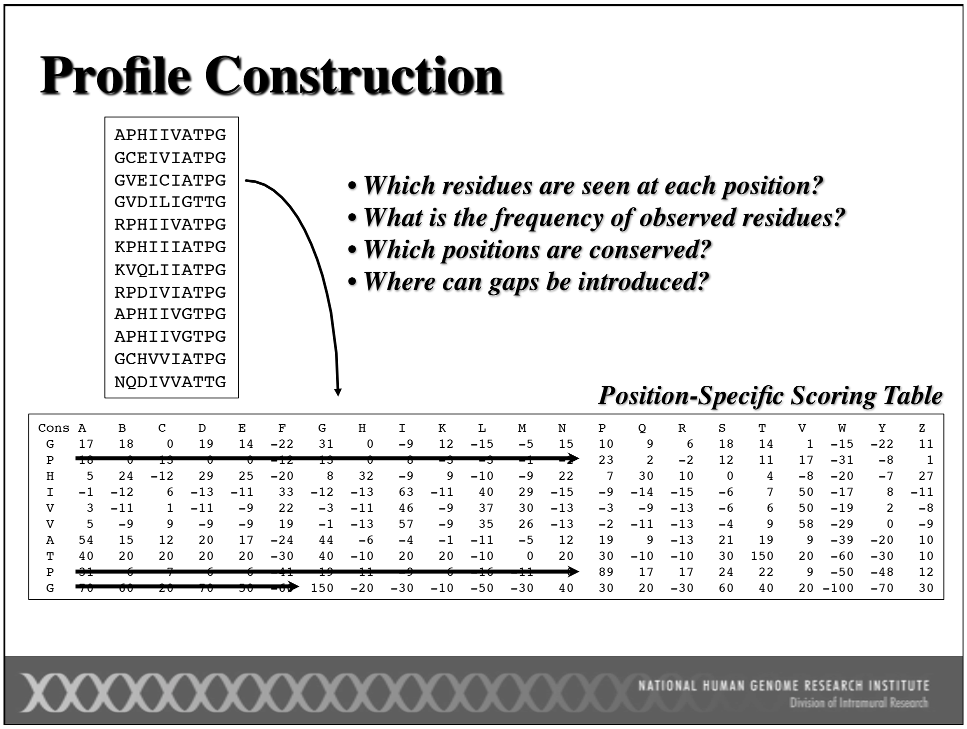
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Profile Construction

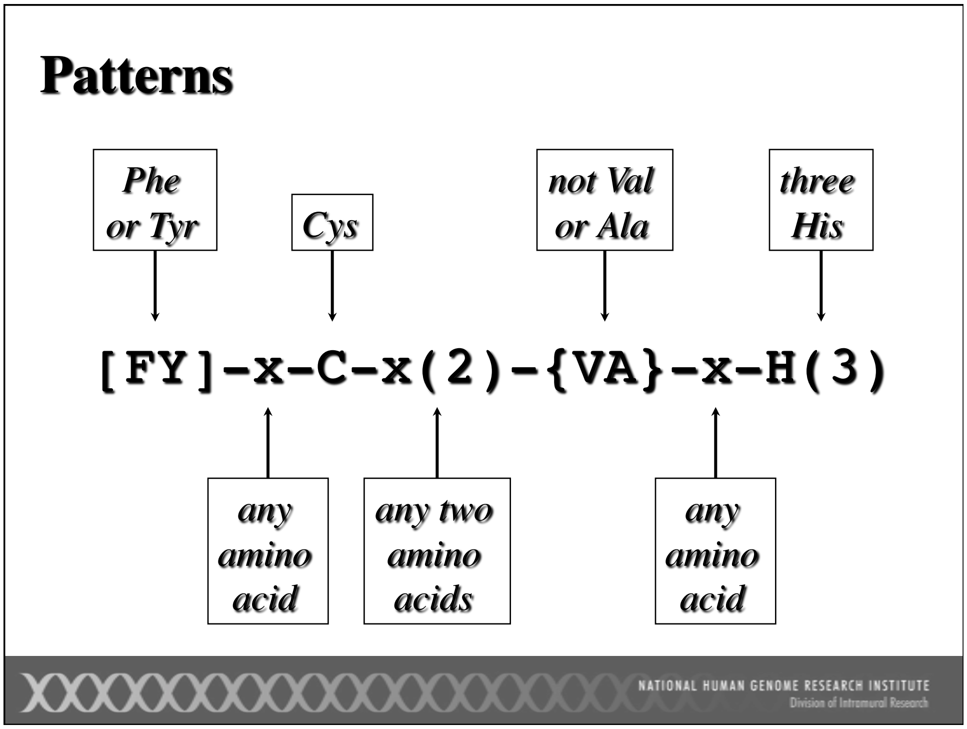
```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
```

- ***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~~	~~0~~	~~13~~	~~0~~	~~0~~	~~−12~~	~~13~~	~~0~~	~~8~~	~~−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	~~31~~	~~6~~	~~7~~	~~6~~	~~6~~	~~−41~~	~~19~~	~~11~~	~~−9~~	~~6~~	~~−16~~	~~−11~~	~~0~~	89	17	17	24	22	9	−50	−48	12
G	~~70~~	~~60~~	~~20~~	~~70~~	~~50~~	~~−60~~	150	−20	−30	−10	−50	−30	40	30	20	−30	60	40	20	−100	−70	30

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Patterns

Phe or Tyr

Cys

not Val or Ala

three His

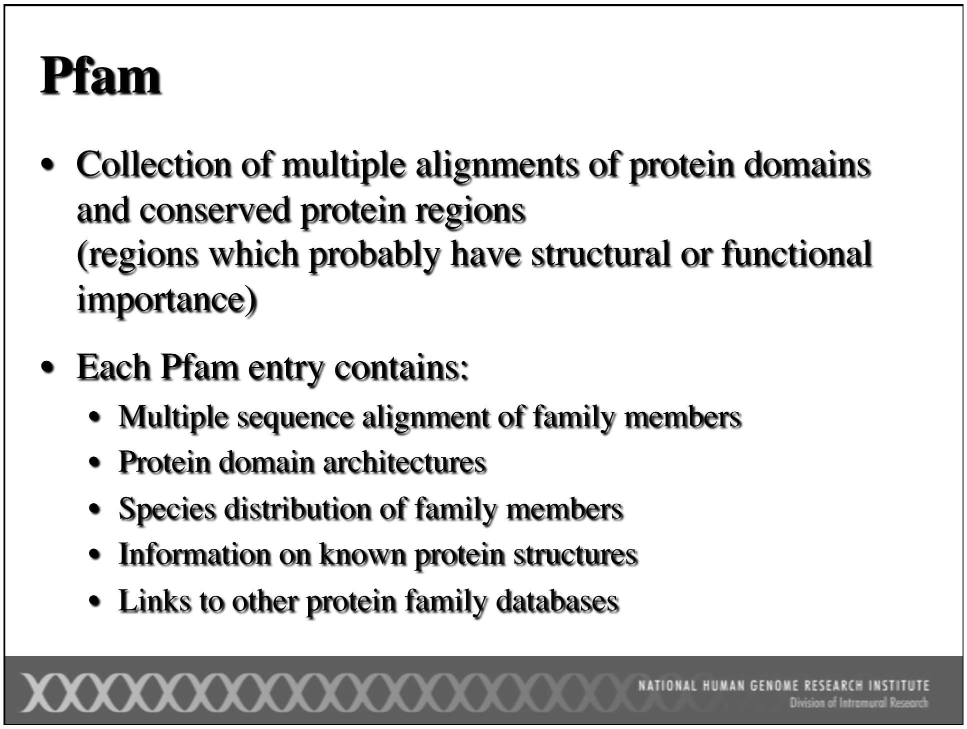
[FY]-x-C-x(2)-{VA}-x-H(3)

any amino acid

any two amino acids

any amino acid

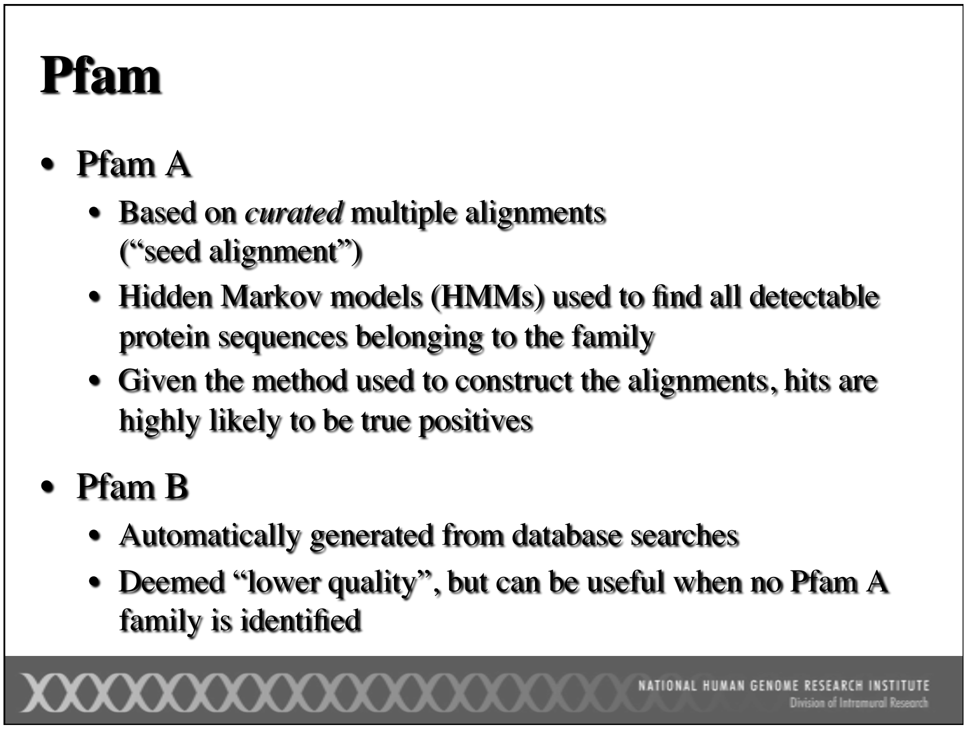
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



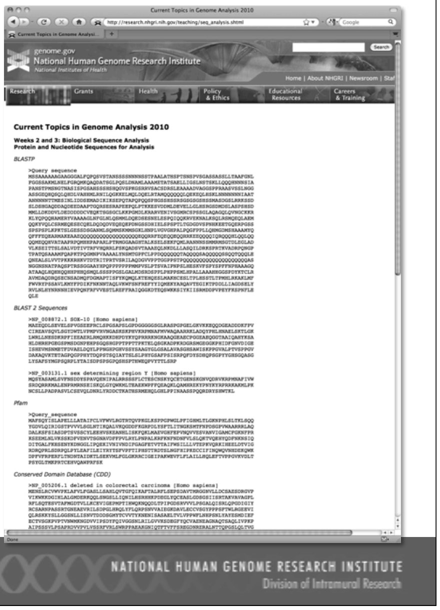
Pfam

- 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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Sequences Used in Examples

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



Current Topics in Genome Analysis 2010
http://research.nhgri.nih.gov/teaching/seq_analysis.shtml
Google
Current Topics in Genome Analysi...
genome.gov
National Human Genome Research Institute
National Institutes of Health
Search
Home | About NHGRI | Newsroom | Staf
Research
Grants
Health
Policy & Ethics
Educational Resources
Careers & Training
Done

Current Topics in Genome Analysis 2010

Weeks 2 and 3: Biological Sequence Analysis
Protein and Nucleotide Sequences for Analysis

BLASTP

```
>Query sequence
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNNSIA
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRAASVSSLNGG
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLNNNNNNIAAT
ANNNNNTTMESINLIDDSEMADIKIKSEPQTAPQPQQSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSD
SLDSHGAQDDAQDEEDAAPTGQRSESRAPEEPQLPTKRESVDDMLDEVELLGLHSRGSDMDSLASPSHSD
MMLLDKDDVLDEDDDDDCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKR
KLYQPQQHAMERYVAAAAGLNFGLNLQSMMLDQEDSESNELESPQIQQKRVEKNALKSQLRSMQEQLAEM
QQKYVQLCSRMEQESECQELDQDQDVEQEQEPDNGSSDHIELSPSPTLTGDGDVSPNHKEETGQERPGSS
SPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHPALPQGFPPLLQHMGDMSHAAAMYQ
QFFFEQEARMAKEAAEQQQQQQQQQQQQQQQQQEQQRRFEQEQQHQQRRKEEQQQQIQRQQQHLQQLQQ
QQMEQQEVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLEGLAD
VLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRKSPRTKVADRPQNGP
TPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQQTAQQQQSAQQQQSSQQTQQQLE
QNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDGVVPPTGGPPSTPQQQQQQQQQQQQQQQQQQQQAS
NGGNSNATPAQSPTRSSGGAAYHPQFPPPPPPMMPVSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQ
ATAAQLHQHHQQHHPHHQSMQLSSSPPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHGGSPDYKTCLR
AVMDAQDRQSECNSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKAKLMF
FWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAGDSELY
RVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRMDDPVPEYFKSPNFLE
QLE
```

BLAST 2 Sequences

```
>NP_008872.1 SOX-10 [Homo sapiens]
MAEEQDLSEVELSPVGSEEPRCLSPGSAPSLGPDGGGGGSGLRASPGPGELGKVKKEQQDGEADDDKFPV
CIREAVSQVLSGYDWTLVPMPVRVNGASKSKPHVKRPMNAFMVWAQAARRKLADQYPHLHNAELSKTLGK
LWRLLNESDKRPFIEEAERLRMQHKKDHPDYKYQPRRRKNGKAAQGEAECPGGEAEQGGTAAIQAHYKSA
HLDHRHPGEGSPMSDGNPEHPSGQSHGPPTPPTTPKTELQSGKADPKRDGRSMGEGGKPHIDFGNVDIGE
ISHEVMSNMETFDVAELDQYLPPNGHPGHVSSYSAAGYGLGSALAVASGHSAWISKPPGVALPTVSPPGV
DAKAQVKTETAGPQGPPHYTDQPSTSQIAYTSLSLPHYGSAFPSISRPQFDYSDHQPSGPYYGHSGQASG
LYSAFSYMGPSQRPLYTAISDPSPSGPQSHSPTHWEQPVYTTLSRP

>NP_003131.1 sex determining region Y [Homo sapiens]
MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKRPMNAFIVW
SRDQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKYRPRRKAKMLPK
NCSLLPADPASVLCSEVQLDNRLYRDDCTKATHSRMEHQLGHLPPINAASSPQQRDRYSHWTKL
```

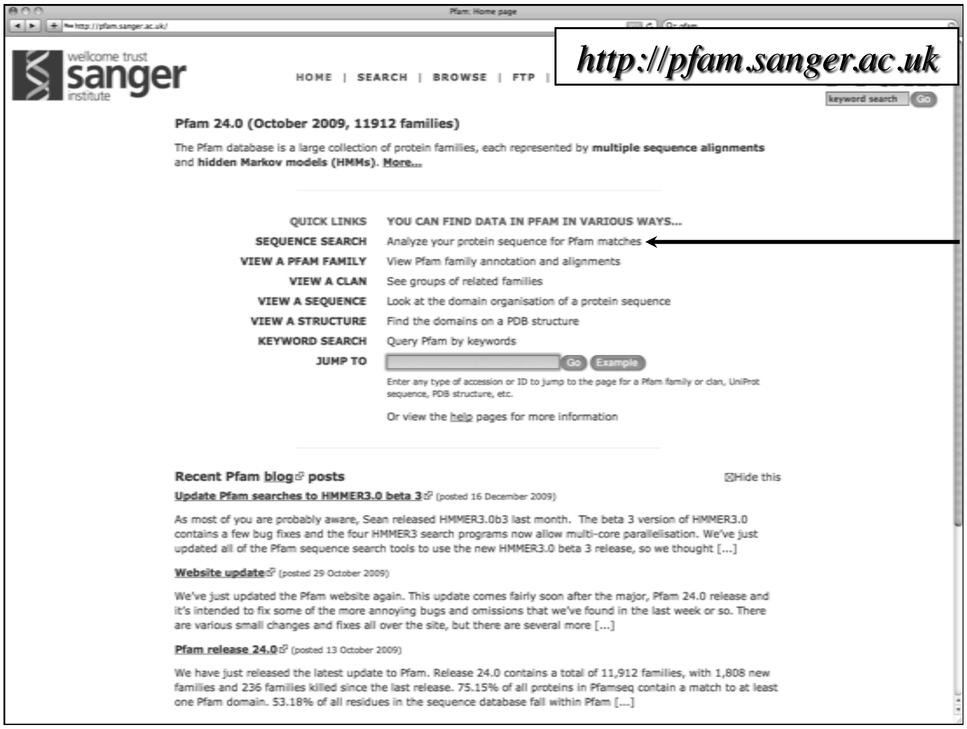
Pfam

```
>Query_sequence
MAFSQYISLAPELLLATAIFCLVFWVLRGTRTQVPKGLKSPPGPWGLPFIGHMLTLGKNPHLSLTKLSQQ
YGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFTLITNGKSMTFNPDSGPVWAARRRLAQ
DALKSFSIASDPTSVSSCYLEEHVSKEANHLISKFQKLMAEVGHFEPVNQVVESVANVIGAMCFGKNFPR
KSEEMLNLVKSSKDFVENVTSGNAVDFFPVLRYLPNPALKRFKNFNDNFVLSLQKTVQEHYQDFNKNSIQ
DITGALFKHSENYKDNGGLIPQEKIVNIVNDIFGAGFETVTTAIFWSILLLVTEPKVQRKIHEELDTVIG
RDRQPRLSDRPQLPYLEAFILEIYRYTSFVPFTIPHSTTRDTSLNGFHIPKECCIFINQWQVNHDEKQWK
DPFVFRPERFLTNDNTAIDKTLSEKVMLFGLGKRRCIGEIPAKWEVFLFLAILLHQLEFTVPPGVKVDLT
PSYGLTMKPRTCEHVQAWPRFSK
```

Conserved Domain Database (CDD)

```
>NP_005206.1 deleted in colorectal carcinoma [Homo sapiens]
MENSLRCVWVPKLAFVLFGASLLSAHLQVTGFQIKAFTALRFLSEPSDAVTMRGGNVLLDCSAESDRGVP
VIKWKKDGIHLALGMDERKQQLSNGSLLIQNILHSRHHKPDEGLYQCEASLGDSGSIISRTAKVAVAGPL
RFLSQTESVTAFMGDTVLLKCEVIGEPMPTIHWQKNQQDLTPIPGDSRVVVLPSGALQISRLQPGDIGIY
RCSARNPASSRTGNEAEVRILSDPGLHRQLYFLQRPSNVVAIEGKDAVLECCVSGYPPPSFTWLRGEEVI
QLRSKKYSLLGGSNLLISNVTDDDSGMYTCVVTYKNENISASAELTVLVPPWFLNHPSNLYAYESMDIEF
ECTVSGKPVPTVNWMKNGDVVIPSDYFQIVGGSNLRILGVVKSDEGFYQCVAENEAGNAQTSAQLIVPKP
AIPSSSVLPSAPRDVVPVLVSSRFVRLSWRPPAEAKGNIQTFTVFFSREGDNRERALNTTQPGSLQLTVG
```

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Pfam: Home page
http://pfam.sanger.ac.uk/
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP |
keyword search
Go

http://pfam.sanger.ac.uk

Pfam 24.0 (October 2009, 11912 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches ←
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	Go Example Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

Recent Pfam blog posts ☒Hide this

Update Pfam searches to HMMER3.0 beta 3 (posted 16 December 2009)

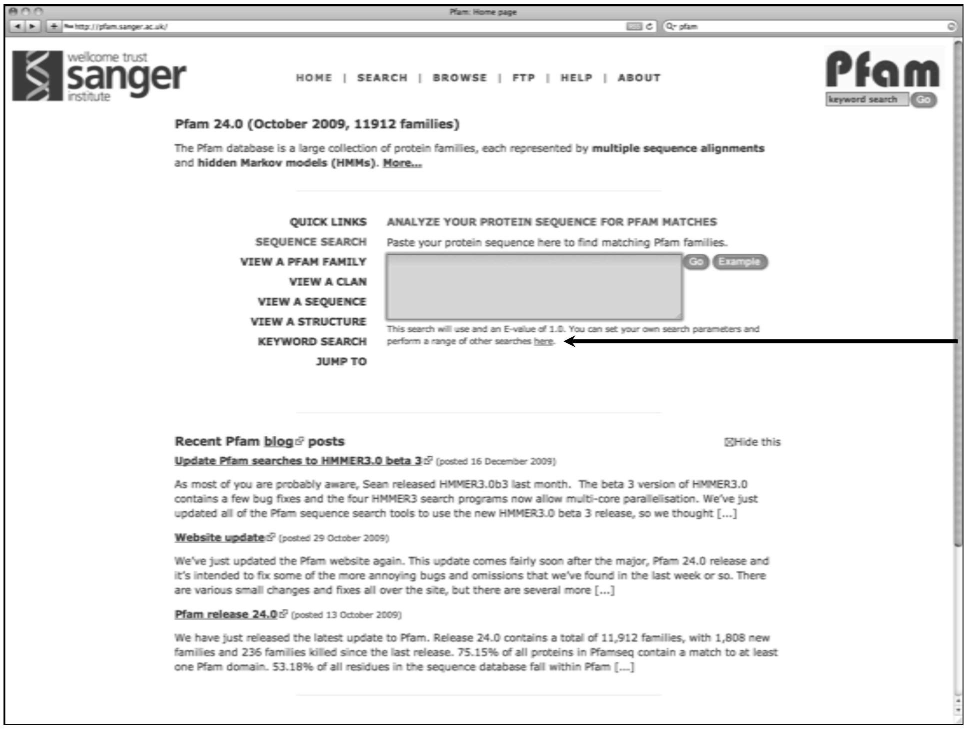
As most of you are probably aware, Sean released HMMER3.0b3 last month. The beta 3 version of HMMER3.0 contains a few bug fixes and the four HMMER3 search programs now allow multi-core parallelisation. We've just updated all of the Pfam sequence search tools to use the new HMMER3.0 beta 3 release, so we thought [...]

Website update (posted 29 October 2009)

We've just updated the Pfam website again. This update comes fairly soon after the major, Pfam 24.0 release and it's intended to fix some of the more annoying bugs and omissions that we've found in the last week or so. There are various small changes and fixes all over the site, but there are several more [...]

Pfam release 24.0 (posted 13 October 2009)

We have just released the latest update to Pfam. Release 24.0 contains a total of 11,912 families, with 1,808 new families and 236 families killed since the last release. 75.15% of all proteins in Pfamseq contain a match to at least one Pfam domain. 53.18% of all residues in the sequence database fall within Pfam [...]



Pfam 24.0 (October 2009, 11912 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. **More...**

ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES

Paste your protein sequence here to find matching Pfam families.

This search will use and an E-value of 1.0. You can set your own search parameters and perform a range of other searches here.

Recent Pfam blog posts

Update Pfam searches to HMMER3.0 beta 3 (posted 16 December 2009)

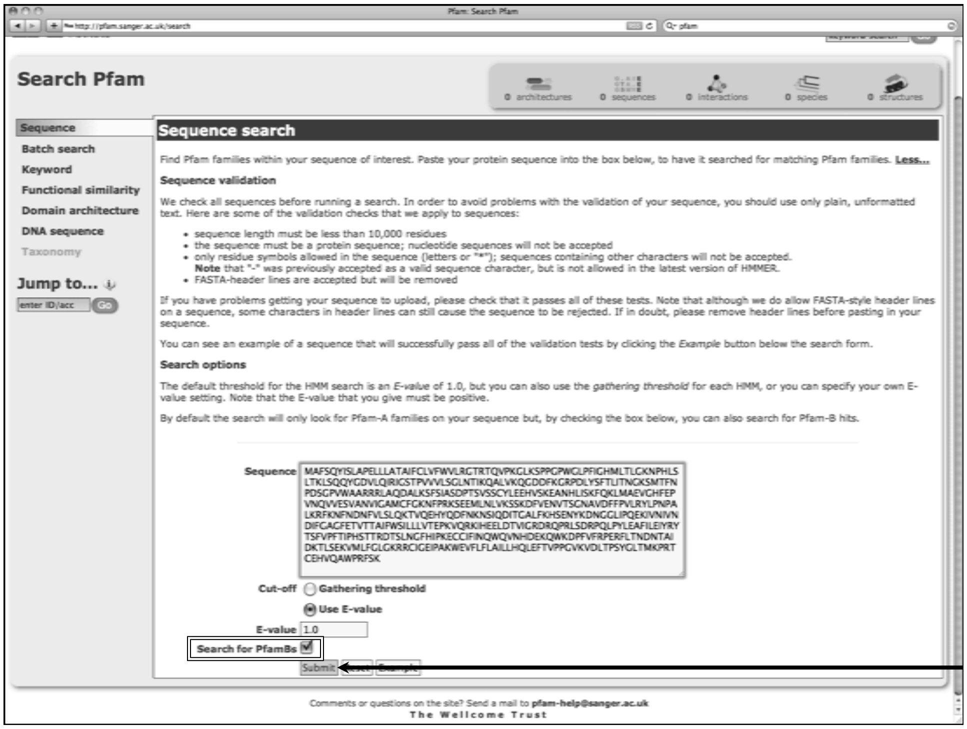
As most of you are probably aware, Sean released HMMER3.0b3 last month. The beta 3 version of HMMER3.0 contains a few bug fixes and the four HMMER3 search programs now allow multi-core parallelisation. We've just updated all of the Pfam sequence search tools to use the new HMMER3.0 beta 3 release, so we thought [...]

Website update (posted 29 October 2009)

We've just updated the Pfam website again. This update comes fairly soon after the major, Pfam 24.0 release and it's intended to fix some of the more annoying bugs and omissions that we've found in the last week or so. There are various small changes and fixes all over the site, but there are several more [...]

Pfam release 24.0 (posted 13 October 2009)

We have just released the latest update to Pfam. Release 24.0 contains a total of 11,912 families, with 1,808 new families and 236 families killed since the last release. 75.15% of all proteins in Pfamseq contain a match to at least one Pfam domain. 53.18% of all residues in the sequence database fall within Pfam [...]



Search Pfam

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. **Less...**

Sequence validation

We check all sequences before running a search. In order to avoid problems with the validation of your sequence, you should use only plain, unformatted text. Here are some of the validation checks that we apply to sequences:

- sequence length must be less than 10,000 residues
- the sequence must be a protein sequence; nucleotide sequences will not be accepted
- only residue symbols allowed in the sequence (letters or "*"); sequences containing other characters will not be accepted. **Note** that "-" was previously accepted as a valid sequence character, but is not allowed in the latest version of HMMER.
- FASTA-header lines are accepted but will be removed

If you have problems getting your sequence to upload, please check that it passes all of these tests. Note that although we do allow FASTA-style header lines on a sequence, some characters in header lines can still cause the sequence to be rejected. If in doubt, please remove header lines before pasting in your sequence.

You can see an example of a sequence that will successfully pass all of the validation tests by clicking the *Example* button below the search form.

Search options

The default threshold for the HMM search is an *E-value* of 1.0, but you can also use the *gathering threshold* for each HMM, or you can specify your own E-value setting. Note that the E-value that you give must be positive.

By default the search will only look for Pfam-A families on your sequence but, by checking the box below, you can also search for Pfam-B hits.

Sequence

```
MAFSQYISLAPELLLATAIFCLVFWVLRGTRTQVPKGLKSPPGPWGLPFIGHMLTLGKNPHLS
LTKLSQQYGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFTLITNGKSMTFN
PDSGPVWAARRRLAQDALKSFSIASDPTSVSSCYLEEHVSKEANHLISKFQKLMAEVGHFEP
VNQVVESVANVIGAMCFGKNFPRKSEEMLNLVKSSKDFVENVTSGNAVDFFPVLRYLPNPA
LKRFKNFNDNFVLSLQKTVQEHYQDFNKNSIQDITGALFKHSENYKDNGGLIPQEKIVNIVN
DIFGAGFETVTTAIFWSILLLVTEPKVQRKIHEELDTVIGRDRQPRLSDRPQLPYLEAFILEIYRY
TSFVPFTIPHSTTRDTSLNGFHIPKECCIFINQWQVNHDEKQWKDPFVFRPERFLTNDNTAI
DKTLSEKVMLFGLGKRRCIGEIPAKWEVFLFLAILLHQLEFTVPPGVKVDLTPSYGLTMKPRT
CEHVQAWPRFSK
```

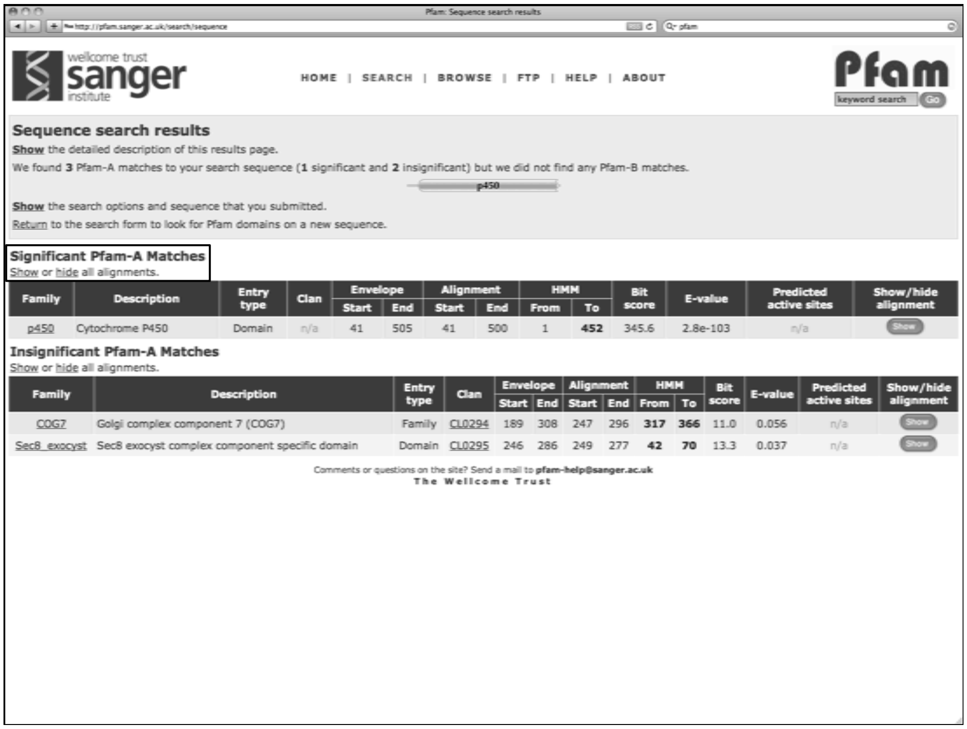
Cut-off ○ **Gathering threshold**

● **Use E-value**

E-value 1.0

Search for PfamBs ☑

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



Pfam: Sequence search results
http://pfam.sanger.ac.uk/search/sequence
pfam
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search
Go

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.

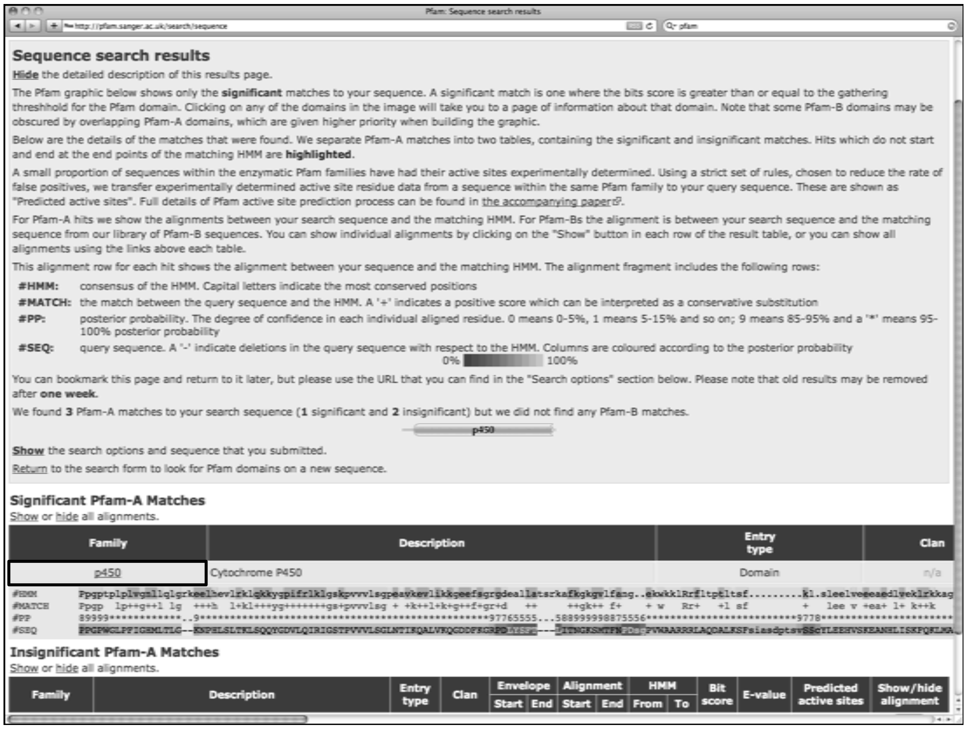
Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
p450	Cytochrome P450	Domain	n/a	41	505	41	500	1	**452**	345.6	2.8e-103	n/a	Show

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**
The Wellcome Trust



Pfam: Sequence search results
http://pfam.sanger.ac.uk/search/sequence
pfam

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 threshhold 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.

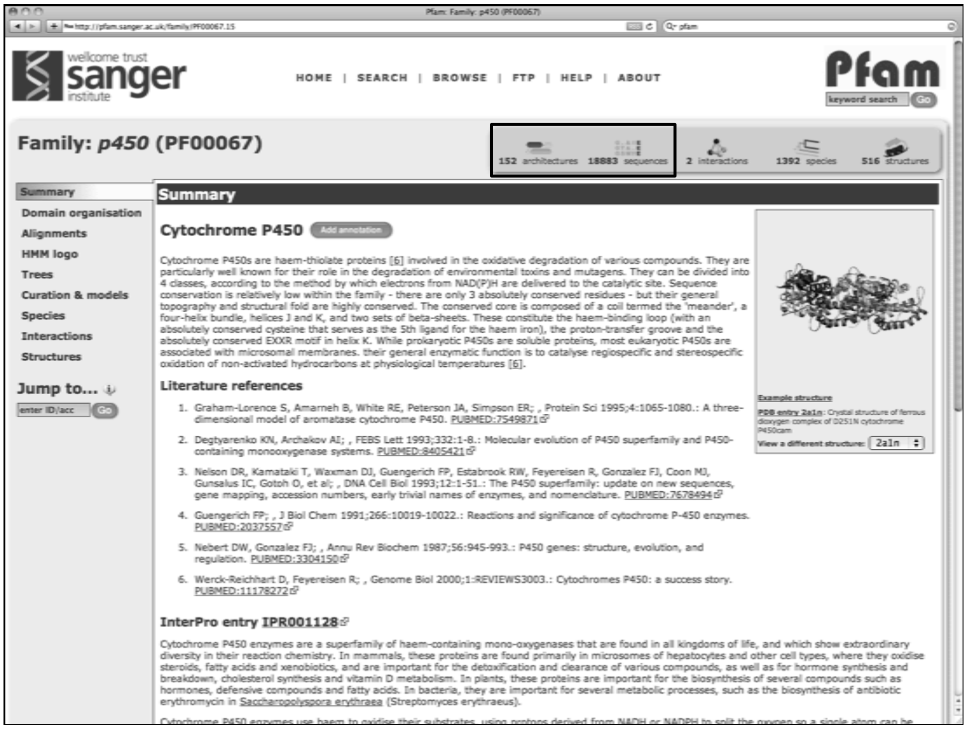
Family	Description	Entry type	Clan
p450	Cytochrome P450	Domain	n/a

```
#HMM     Ppgptplplvgnllqlgrkeelhevlrklqkkygpifrlklgskpvvvlsgpeavkevlikkgeefsgrgdeallatsrkafkgkgvlfang..ekwkklRrfltptltsf.........kl.sleelveeeaedlveklrkkag
#MATCH   Ppgp  lp++g++l lg  +++h  l+kl+++yg+++++++gs+pvvvlsg + +k++l+k+g++f+gr+d    ++       ++gk++ f+    + w    Rr+    +l sf          +    lee v +ea+ l+ k++k
#PP      89999*************..9**************************************************97765555...588999998875556***************************9778**********************
#SEQ     PPGPWGLPFIGHMLTLG--KNPHLSLTKLSQQYGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFT---LITNGKSMTFNPDsgPVWAARRRLAQDALKSFsiasdptsvSScYLEEHVSKEANHLISKFQKLMA
```

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)

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

Example structure

PDB entry 2a1n: Crystal structure of ferrous dioxygen complex of D251N cytochrome P450cam

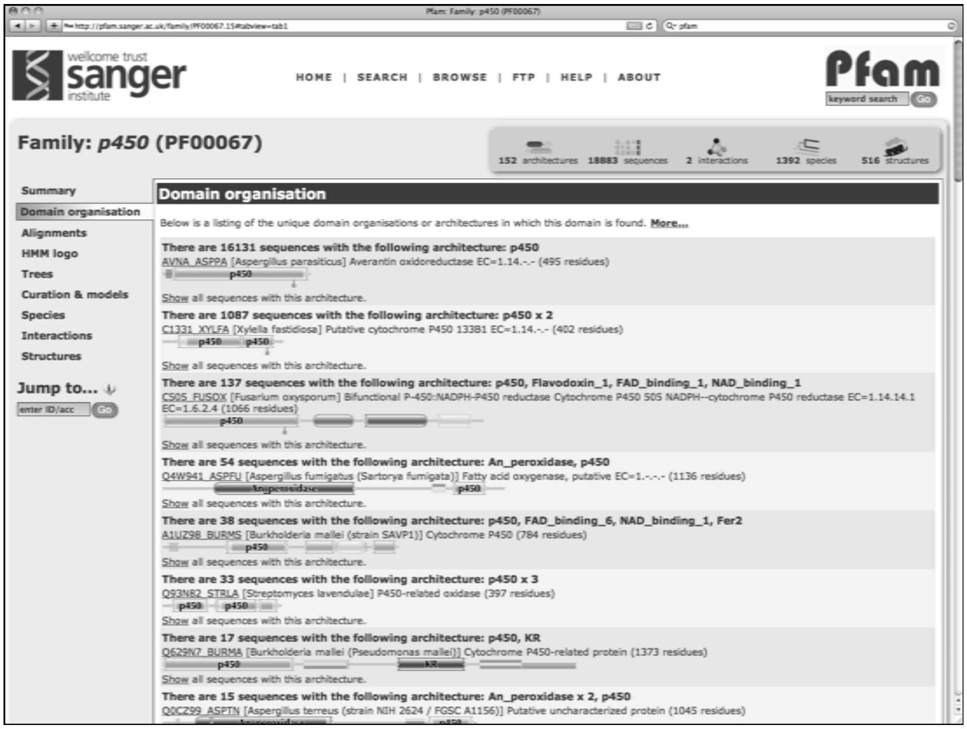
Literature references

1. 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
2. Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. PUBMED:8405421
3. 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
4. Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. PUBMED:2037557
5. Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. PUBMED:3304150
6. 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)

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
C1331_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
CS05_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.

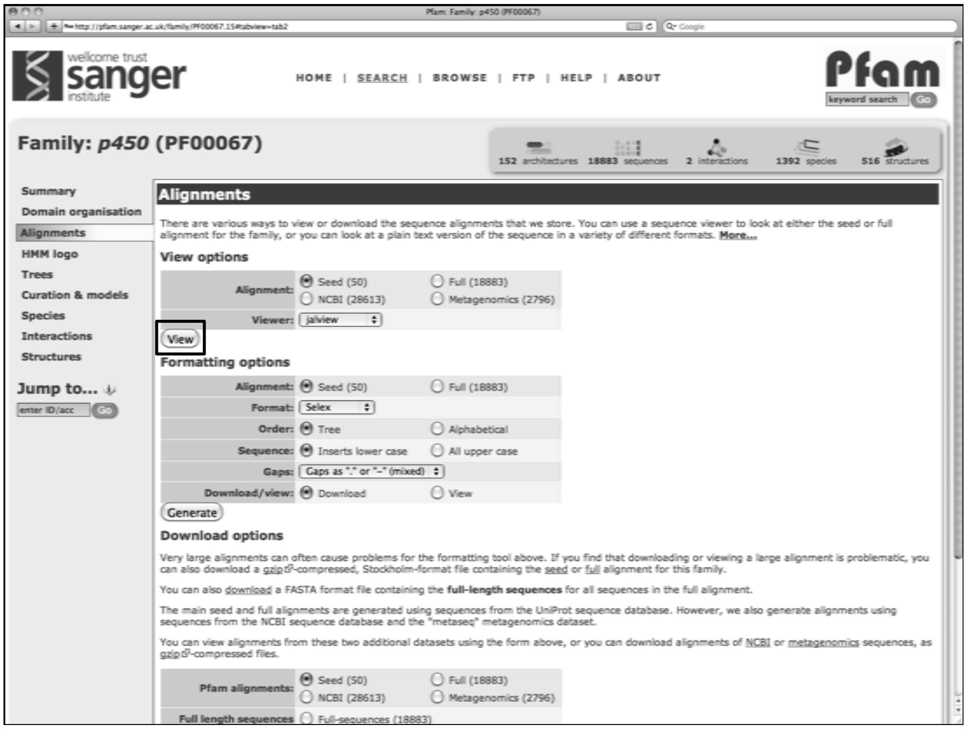
There are 54 sequences with the following architecture: An_peroxidase, p450
Q4W941_ASPFU [Aspergillus fumigatus (Sartorya fumigata)] Fatty acid oxygenase, putative EC=1.-.-.- (1136 residues)
Show all sequences with this architecture.

There are 38 sequences with the following architecture: p450, FAD_binding_6, NAD_binding_1, Fer2
A1UZ98_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
Q0CZ99_ASPTN [Aspergillus terreus (strain NIH 2624 / FGSC A1156)] Putative uncharacterized protein (1045 residues)



Family: *p450* (PF00067)

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) ○ NCBI (28613)	○ Full (18883) ○ Metagenomics (2796)
Viewer:	jalview	

View

Formatting options

Alignment:	● Seed (50)	○ Full (18883)
Format:	Selex	
Order:	● Tree	○ Alphabetical
Sequence:	● Inserts lower case	○ All upper case
Gaps:	Gaps as "." or "–" (mixed)	
Download/view:	● Download	○ View

Generate

Download options

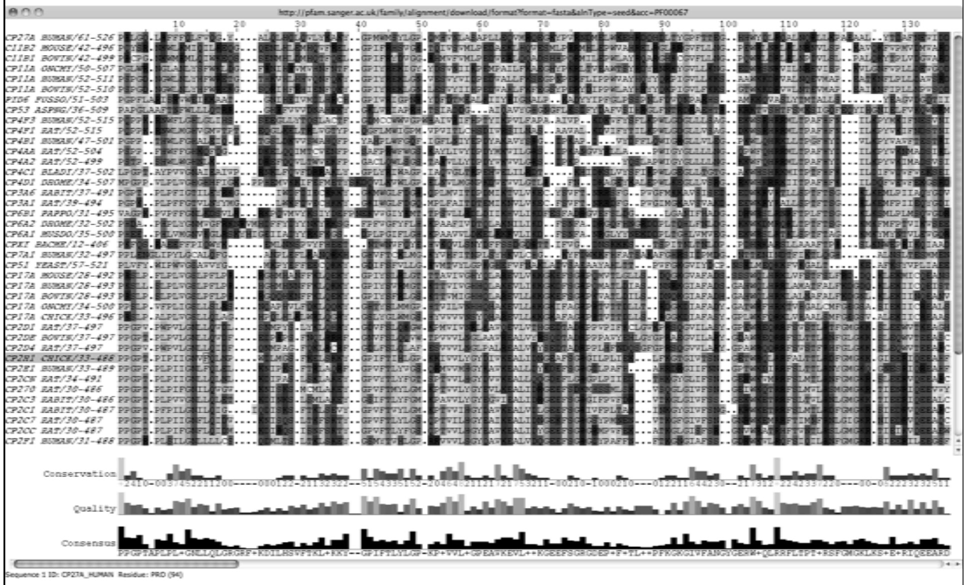
Very large alignments can often cause problems for the formatting tool above. If you find that downloading or viewing a large alignment is problematic, you can also download a gzip-compressed, Stockholm-format file containing the seed or full alignment for this family.

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

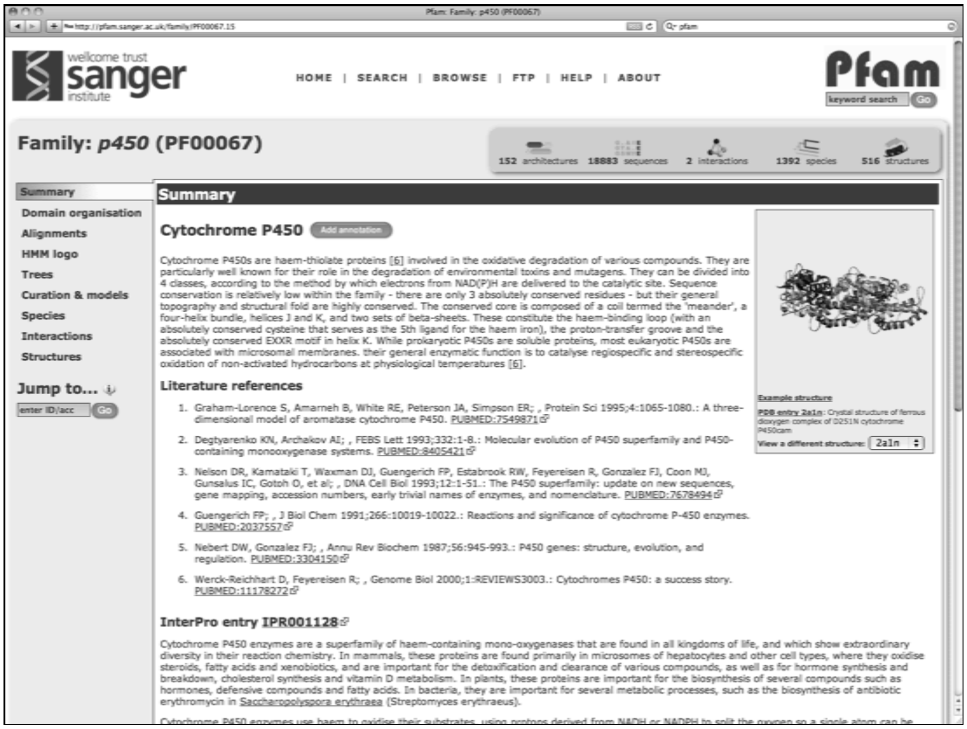
The main seed and full alignments are generated using sequences from the UniProt sequence database. However, we also generate alignments using sequences from the NCBI sequence database and the "metaseq" metagenomics dataset.

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.

Pfam alignments:	● Seed (50) ○ NCBI (28613)	○ Full (18883) ○ Metagenomics (2796)
Full length sequences	○ Full-sequences (18883)	



Sequence 1 ID: CP27A_HUMAN Residue: PRO (94)



Family: *p450* (PF00067)

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

Example structure

PDB entry 2a1n: Crystal structure of ferrous dioxygen complex of D251N cytochrome P450cam

View a different structure: 2a1n

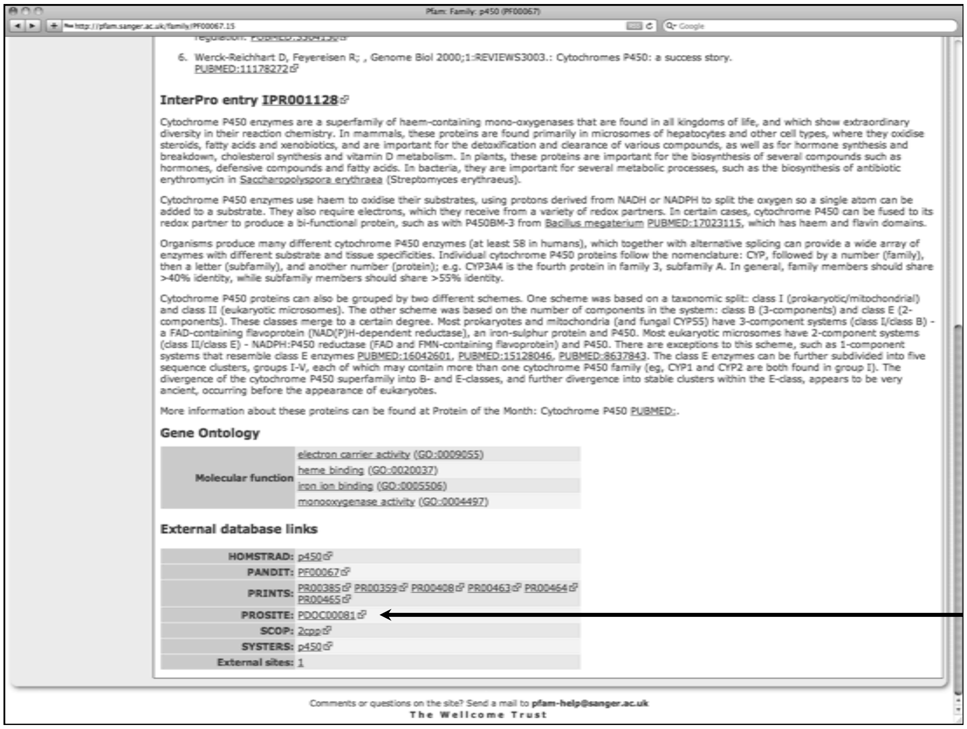
Literature references

1. 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
2. Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. PUBMED:8405421
3. 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
4. Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. PUBMED:2037557
5. Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. PUBMED:3304150
6. 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


Pfam: Family: p450 (PF00067)
http://pfam.sanger.ac.uk/family/PF00067.15
Google

6. 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) - 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:.

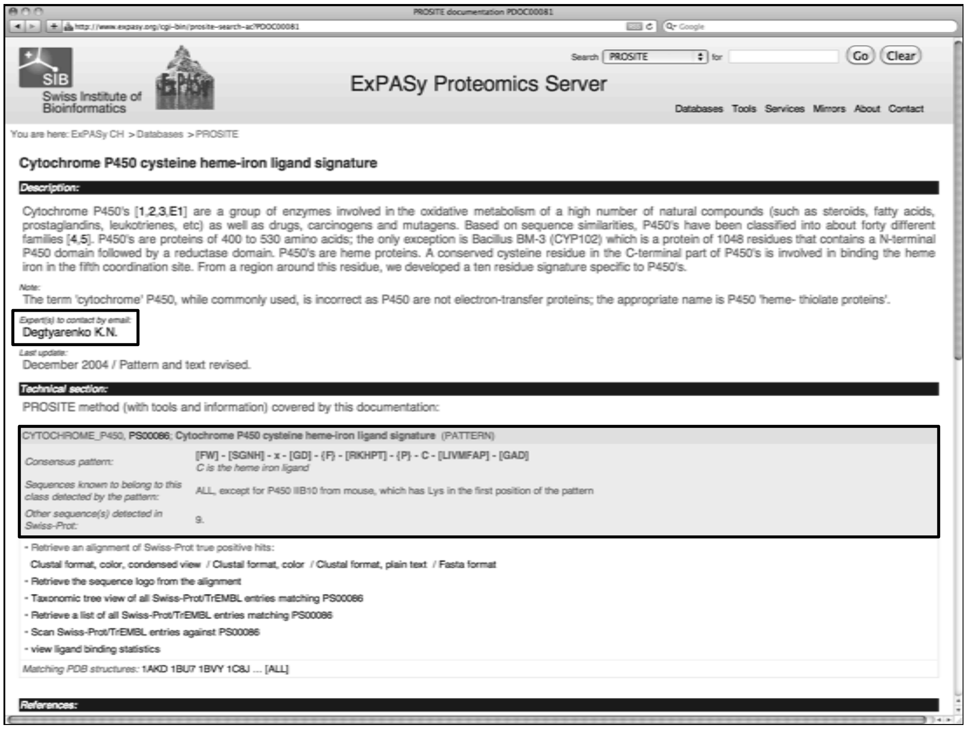
Gene Ontology

Molecular function	electron carrier activity (GO:0009055)
	heme binding (GO:0020037)
	iron ion binding (GO:0005506)
	monooxygenase activity (GO:0004497)

External database links

HOMSTRAD:	p450
PANDIT:	PF00067
PRINTS:	PR00385 PR00359 PR00408 PR00463 PR00464 PR00465
PROSITE:	PDOC00081
SCOP:	2cpp
SYSTERS:	p450
External sites:	1

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



PROSITE documentation PDOC00081

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

Google
Search PROSITE for
Go
Clear
SIB
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 530 amino acids; the only exception is Bacillus BM-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:
Degtyarenko K.N.

Last update:
December 2004 / Pattern and text revised.

Technical section:

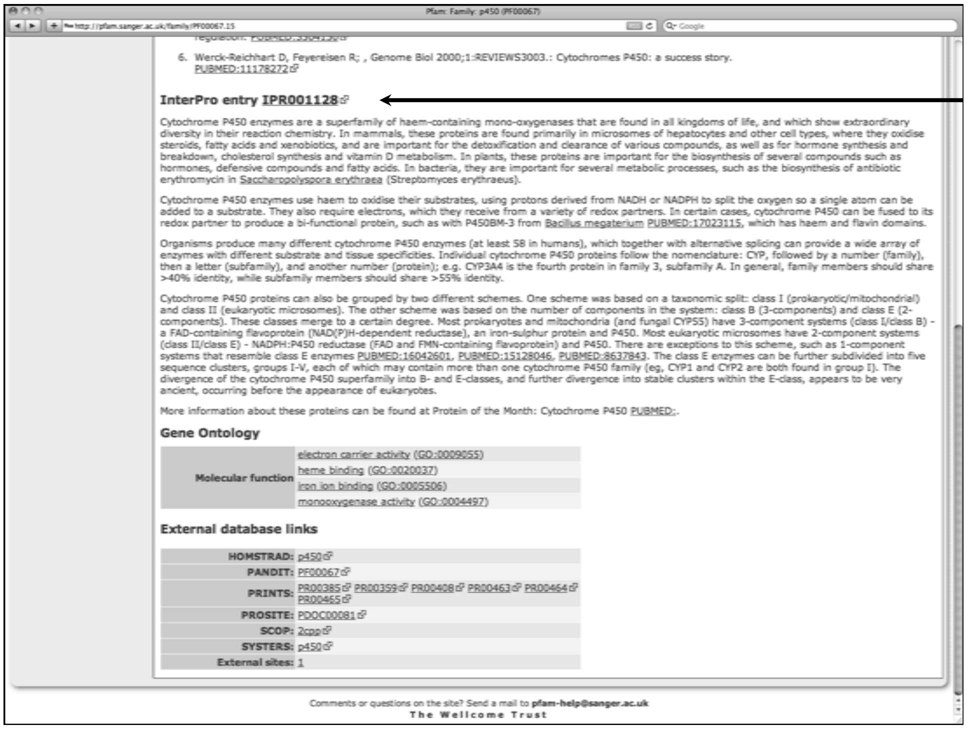
PROSITE method (with tools and information) covered by this documentation:

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 1C8J ... [ALL]

References:



Pfam: Family: p450 (PF00067)

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

6. 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) - 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:.

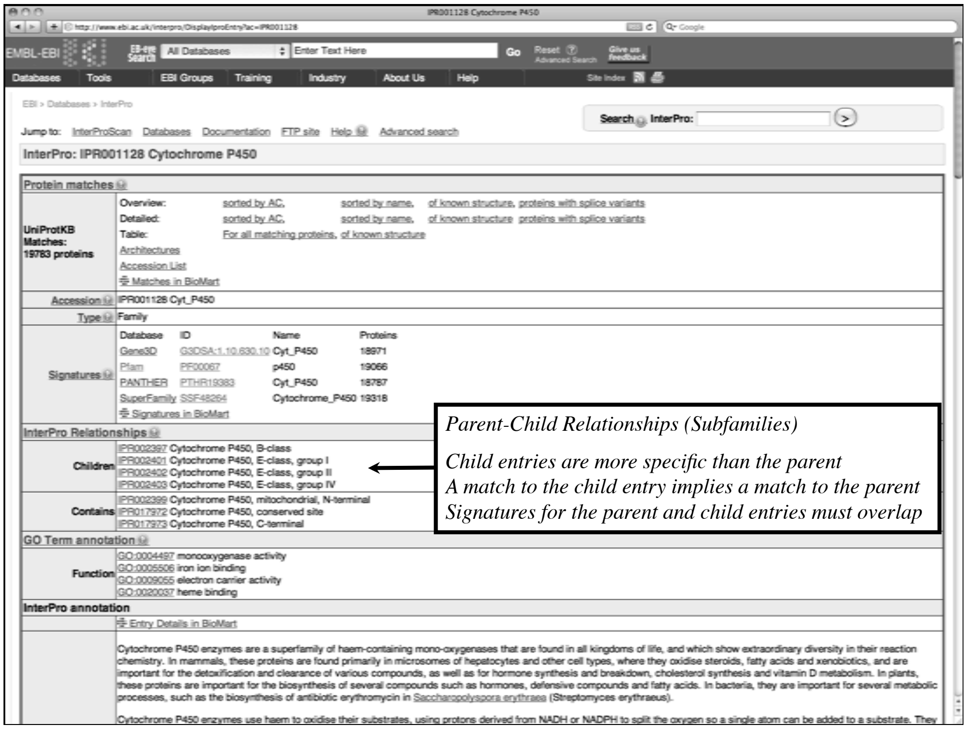
Gene Ontology

Molecular function	electron carrier activity (GO:0009055)
	heme binding (GO:0020037)
	iron ion binding (GO:0005506)
	monooxygenase activity (GO:0004497)

External database links

HOMSTRAD:	p450
PANDIT:	PF00067
PRINTS:	PR00385 PR00359 PR00408 PR00463 PR00464 PR00465
PROSITE:	PDOC00081
SCOP:	2cpp
SYSTERS:	p450
External sites:	1

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



EBI > Databases > InterPro

Jump to: InterProScan Databases Documentation FTP site Help Advanced search

Search InterPro: >

InterPro: IPR001128 Cytochrome P450

Protein matches

UniProtKB Matches: 19783 proteins

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		

Architectures
Accession List
Matches in BioMart

Accession: IPR001128 Cyt_P450

Type: Family

Signatures:

Database	ID	Name	Proteins
Gene3D	G3DSA:1.10.630.10	Cyt_P450	18971
Pfam	PF00067	p450	19066
PANTHER	PTHR19383	Cyt_P450	18787
SuperFamily	SSF48264	Cytochrome_P450	19318

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 monooxygenase activity
GO:0005506 iron ion binding
GO:0009055 electron carrier activity
GO:0020037 heme binding

InterPro annotation

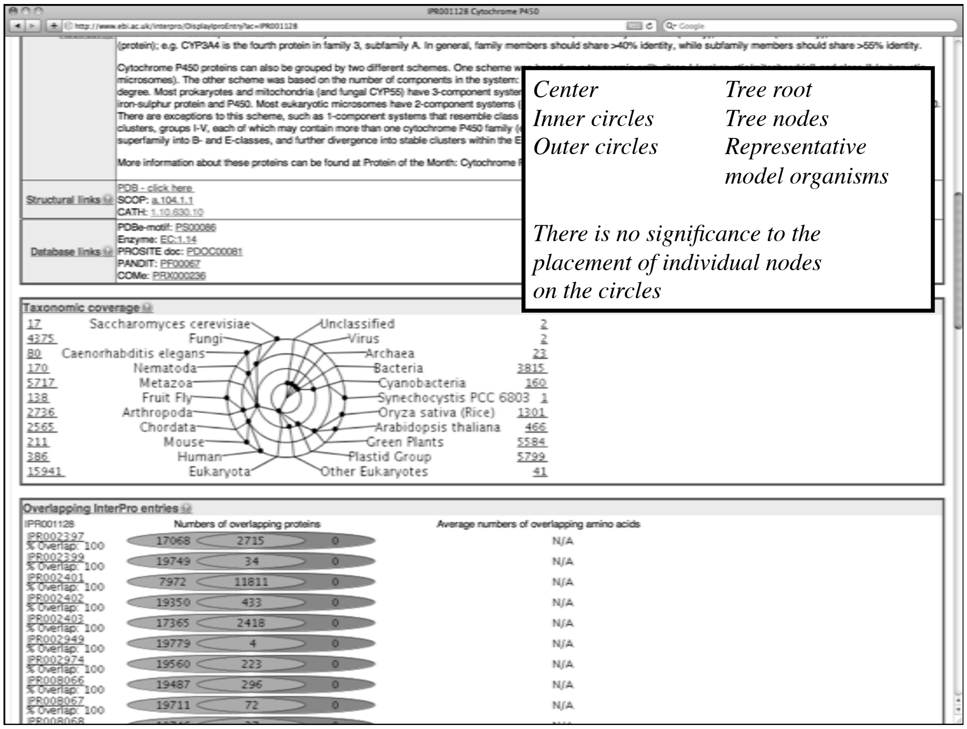
Entry Details in BioMart

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


IPR001128 Cytochrome P450
http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128
Google

(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 w
microsomes). The other scheme was based on the number of components in the system:
degree. Most prokaryotes and mitochondria (and fungal CYP55) have 3-component syster
iron-sulphur protein and P450. Most eukaryotic microsomes have 2-component systems (
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
superfamily into B- and E-classes, and further divergence into stable clusters within the E

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

Structural links
PDB - click here
SCOP: a.104.1.1
CATH: 1.10.630.10

Database links
PDBe-motif: PS00086
Enzyme: EC:1.14
PROSITE doc: PDOC00081
PANDIT: PF00067
COMe: PRX000236

Taxonomic coverage

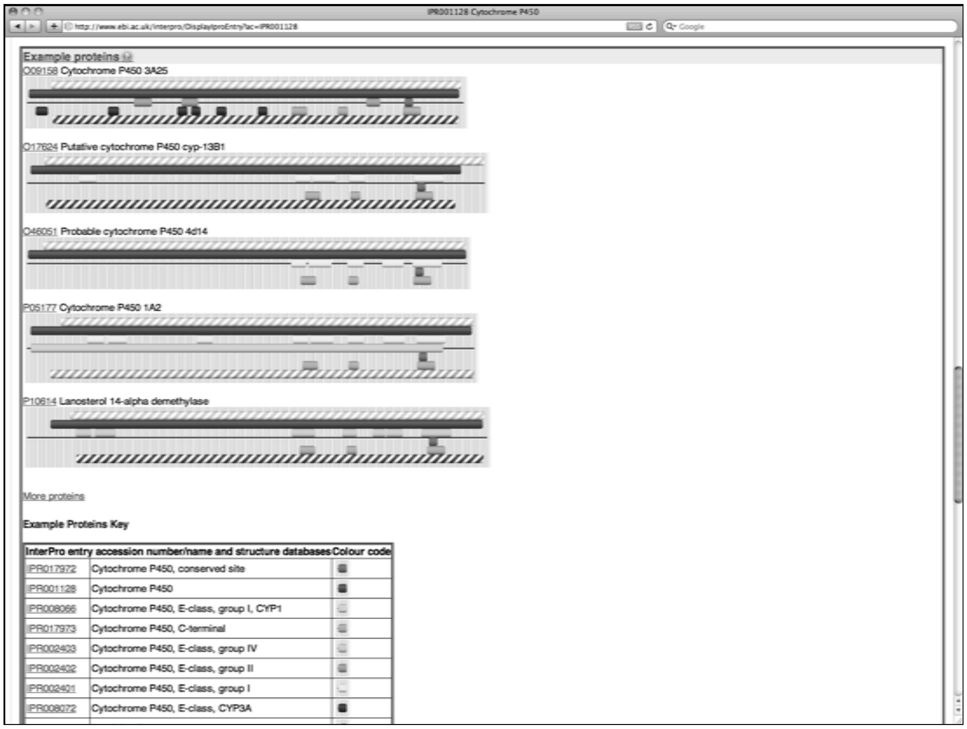
17	Saccharomyces cerevisiae	Unclassified	2
4375	Fungi	Virus	2
80	Caenorhabditis elegans	Archaea	23
170	Nematoda	Bacteria	3815
5717	Metazoa	Cyanobacteria	160
138	Fruit Fly	Synechocystis PCC 6803	1
2736	Arthropoda	Oryza sativa (Rice)	1301
2565	Chordata	Arabidopsis thaliana	466
211	Mouse	Green Plants	5584
386	Human	Plastid Group	5799
15941	Eukaryota	Other Eukaryotes	41

Overlapping InterPro entries

IPR001128	Numbers of overlapping proteins			Average numbers of overlapping amino acids
IPR002397 % Overlap: 100	17068	2715	0	N/A
IPR002399 % Overlap: 100	19749	34	0	N/A
IPR002401 % Overlap: 100	7972	11811	0	N/A
IPR002402 % Overlap: 100	19350	433	0	N/A
IPR002403 % Overlap: 100	17365	2418	0	N/A
IPR002949 % Overlap: 100	19779	4	0	N/A
IPR002974 % Overlap: 100	19560	223	0	N/A
IPR008066 % Overlap: 100	19487	296	0	N/A
IPR008067 % Overlap: 100	19711	72	0	N/A
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



IPR001128 Cytochrome P450

http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128

Example proteins

O09158 Cytochrome P450 3A25

O17624 Putative cytochrome P450 cyp-13B1

O46051 Probable cytochrome P450 4d14

P05177 Cytochrome P450 1A2

P10614 Lanosterol 14-alpha demethylase

More proteins

Example Proteins Key

InterPro entry accession number/name and structure databases		Colour code
IPR017972	Cytochrome P450, conserved site	■
IPR001128	Cytochrome P450	■
IPR008066	Cytochrome P450, E-class, group I, CYP1	□
IPR017973	Cytochrome P450, C-terminal	▨
IPR002403	Cytochrome P450, E-class, group IV	□
IPR002402	Cytochrome P450, E-class, group II	▨
IPR002401	Cytochrome P450, E-class, group I	□
IPR008072	Cytochrome P450, E-class, CYP3A	■

Further Reading

Identifying Protein Domains with the Pfam Database

UNIT 2.5

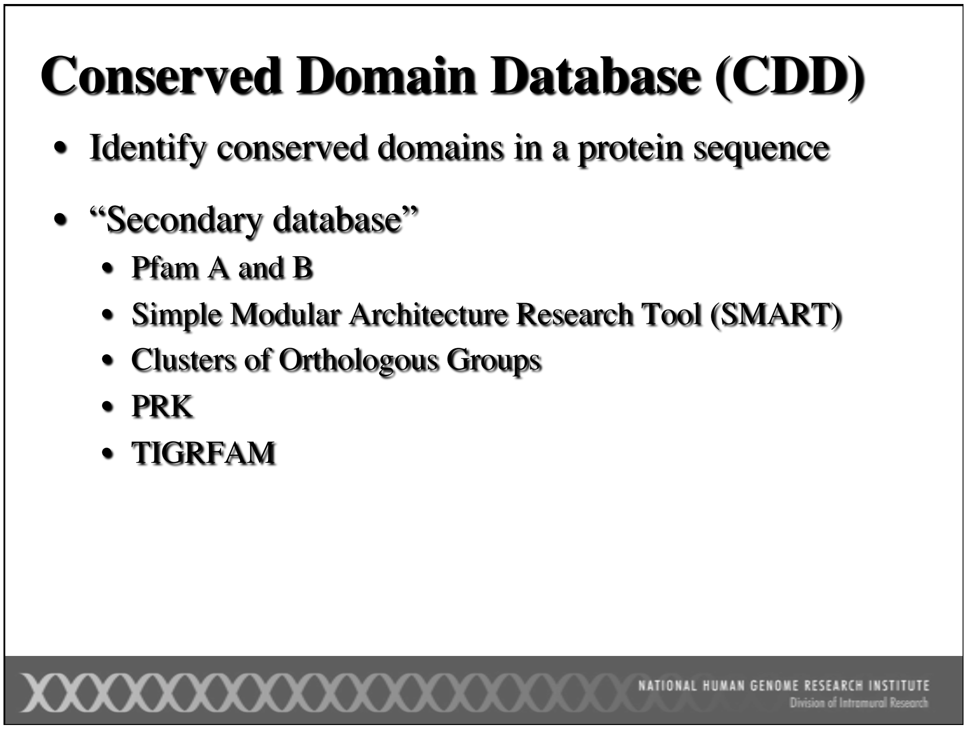
Current Protocols in Bioinformatics
Unit 2.5
Pfam

The InterPro Database and Tools for Protein Domain Analysis

UNIT 2.7

Current Protocols in Bioinformatics
Unit 2.7
InterPro

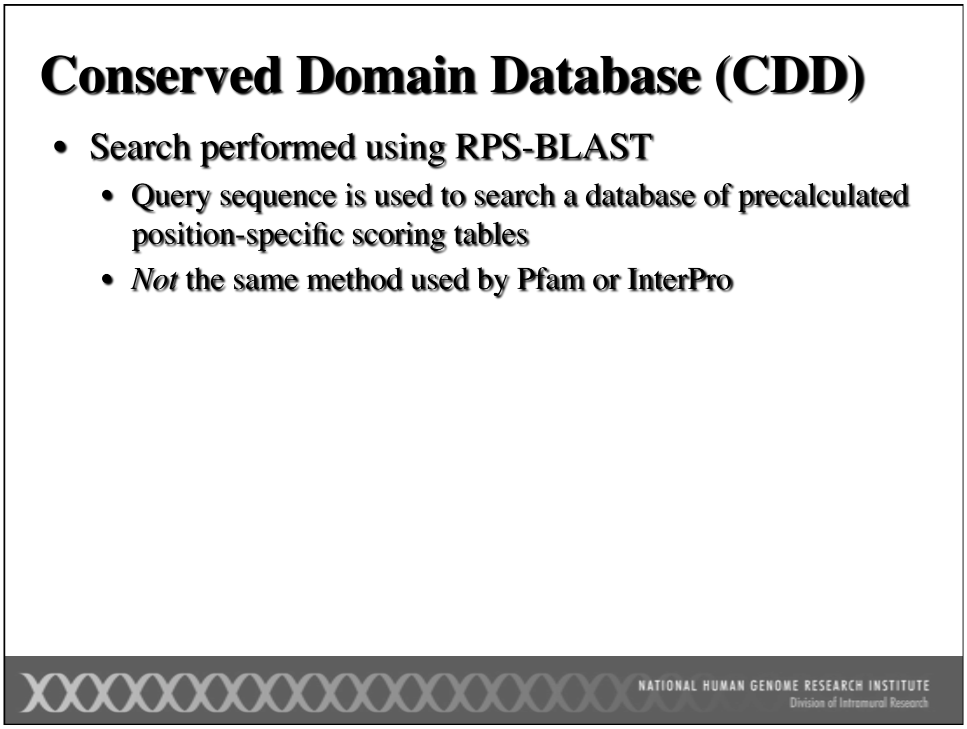
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

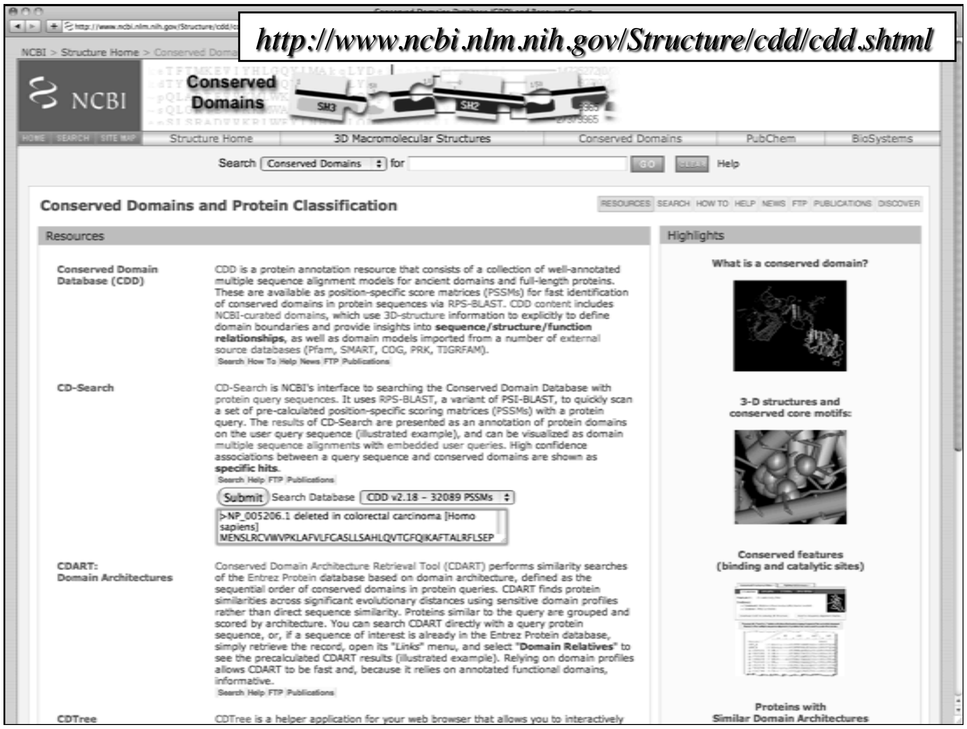
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

Conserved Domains and Protein Classification

Conserved Domain Database (CDD)

CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domains, which use 3D-structure information to explicitly to define domain boundaries and provide insights into **sequence/structure/function relationships**, as well as domain models imported from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAM).

CD-Search

CD-Search is NCBI's interface to searching the Conserved Domain Database with protein query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query. The results of CD-Search are presented as an annotation of protein domains on the user query sequence (illustrated example), and can be visualized as domain multiple sequence alignments with embedded user queries. High confidence associations between a query sequence and conserved domains are shown as **specific hits**.

```
>NP_005206.1 deleted in colorectal carcinoma [Homo
sapiens]
MENSLRCVWVPKLAFVLFCASLLSAHLQVTGFQIKAFTALRFLSEP
```

CDART: Domain Architectures

Conserved Domain Architecture Retrieval Tool (CDART) performs similarity searches of the Entrez Protein database based on domain architecture, defined as the sequential order of conserved domains in protein queries. CDART finds protein similarities across significant evolutionary distances using sensitive domain profiles rather than direct sequence similarity. Proteins similar to the query are grouped and scored by architecture. You can search CDART directly with a query protein sequence, or, if a sequence of interest is already in the Entrez Protein database, simply retrieve the record, open its "Links" menu, and select "**Domain Relatives**" to see the precalculated CDART results (illustrated example). Relying on domain profiles allows CDART to be fast and, because it relies on annotated functional domains, informative.

CDTree

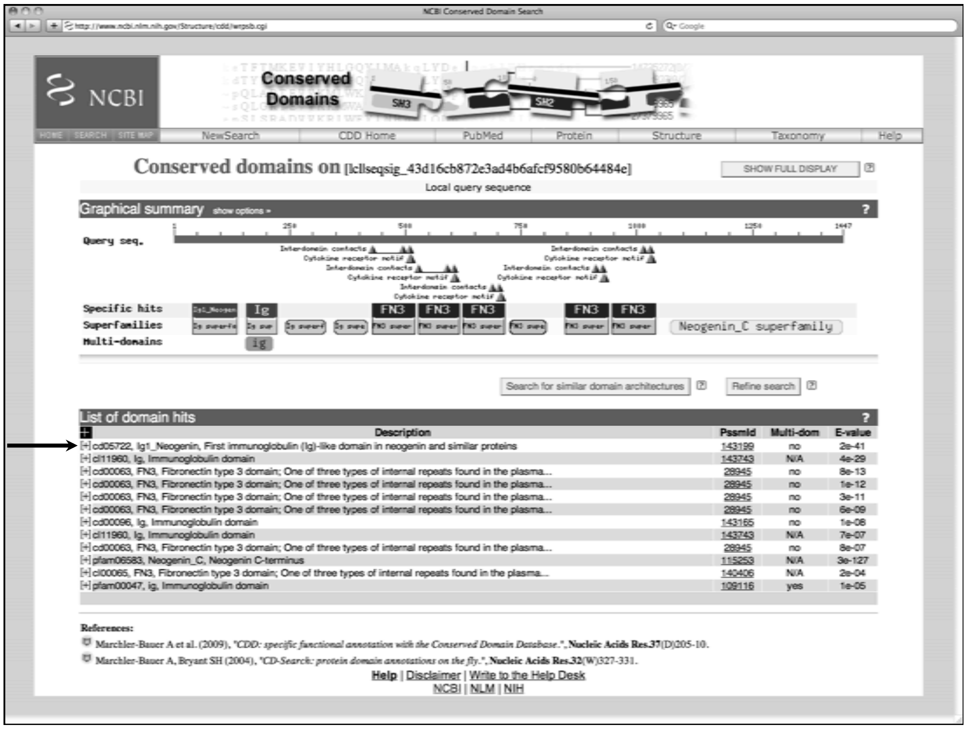
CDTree is a helper application for your web browser that allows you to interactively

What is a conserved domain?

3-D structures and conserved core motifs:

Conserved features (binding and catalytic sites)

Proteins with Similar Domain Architectures



Conserved domains on [lcl|seqsig_43d16cb872e3ad4b6afcf9580b64484e]

Local query sequence

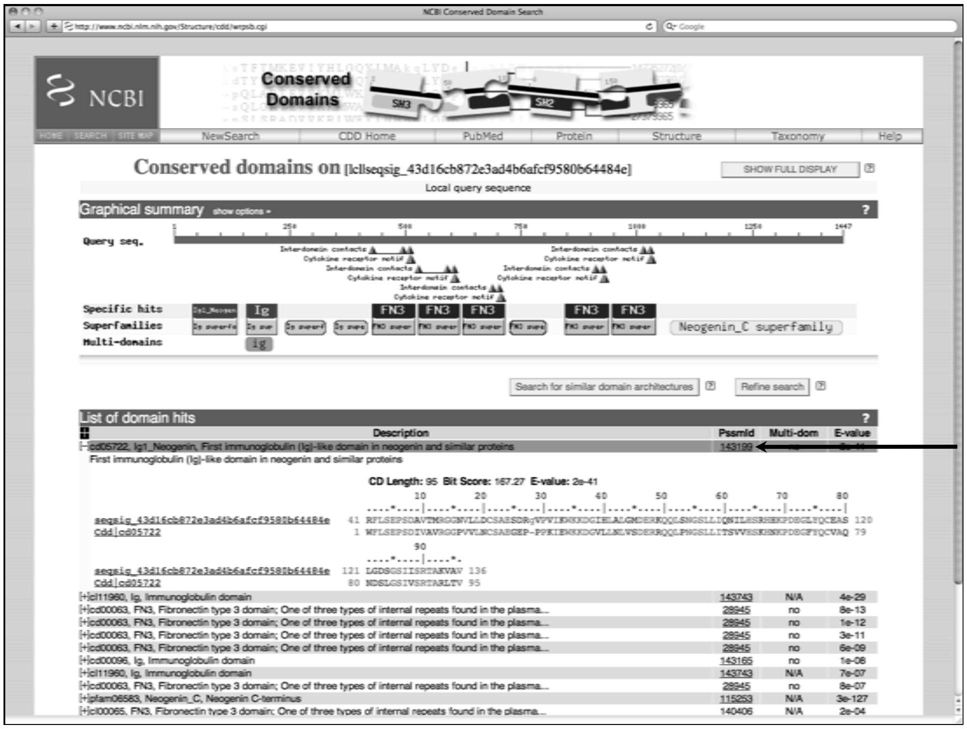
Description	PssmId	Multi-dom	E-value
[+] cd05722, Ig1_Neogenin, First immunoglobulin (Ig)-like domain in neogenin and similar proteins	143199	no	2e-41
[+] cl11960, Ig, Immunoglobulin domain	143743	N/A	4e-29
[+] cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the 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
[+] cd00096, Ig, Immunoglobulin domain	143165	no	1e-08
[+] cl11960, Ig, Immunoglobulin domain	143743	N/A	7e-07
[+] cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma...	28945	no	8e-07
[+] pfam06583, Neogenin_C, Neogenin C-terminus	115253	N/A	3e-127
[+] cl00065, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma...	140406	N/A	2e-04
[+] 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



Conserved domains on [lcl|seqsig_43d16cb872e3ad4b6afcf9580b64484e]

Local query sequence

Graphical summary

List of domain hits

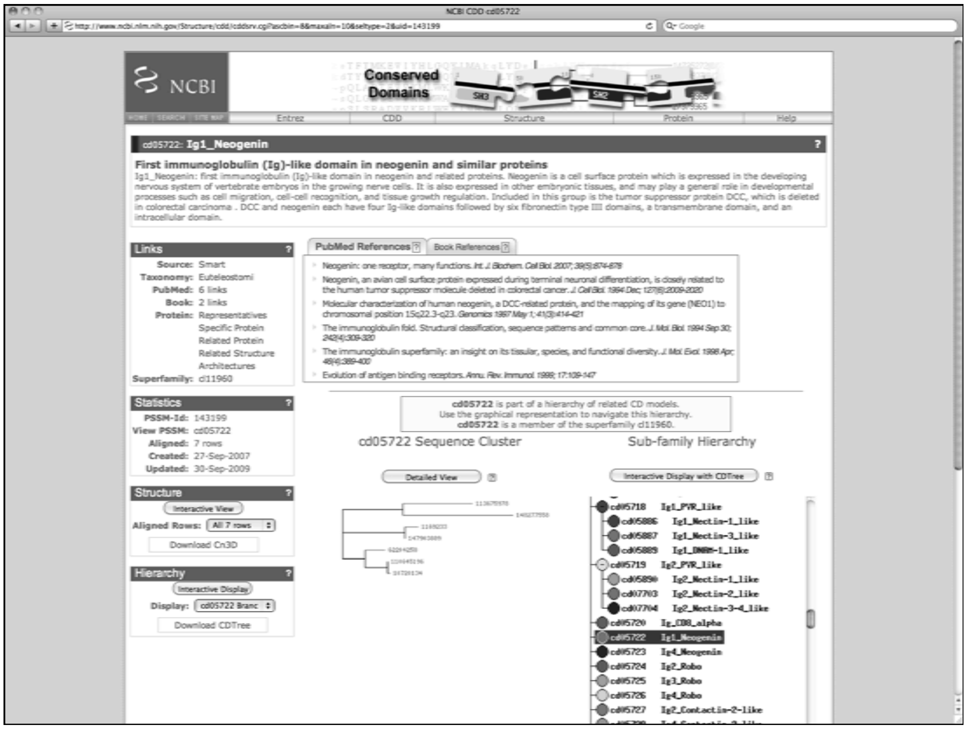
Description	PssmId	Multi-dom	E-value
[-]cd05722, Ig1_Neogenin, First immunoglobulin (Ig)-like domain in neogenin and similar proteins	143199	no	2e-41

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

CD Length: 95 **Bit Score:** 167.27 **E-value:** 2e-41

```
                                                    10        20        30        40        50        60        70        80
                                            ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
seqsig_43d16cb872e3ad4b6afcf9580b64484e   41 RFLSEPSDAVTMRGGNVLLDCSAESDRgVPVIEWKKDGIHLALGMDERKQQLSNGSLLIQNILHSRHHKPDEGLYQCEAS 120
Cdd|cd05722                                1 WFLSEPSDIVAVRGGPVVLNCSAEGEP-PPKIEWKKDGVLLNLVSDERRQQLPNGSLLITSVVHSKHNKPDEGFYQCVAQ 79
                                                    90
                                            ....*....|....*.
seqsig_43d16cb872e3ad4b6afcf9580b64484e  121 LGDSGSIISRTAKVAV 136
Cdd|cd05722                               80 NDSLGSIVSRTARLTV 95
```

Description	PssmId	Multi-dom	E-value
[+]cl11960, Ig, Immunoglobulin domain	143743	N/A	4e-29
[+]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
[+]cd00096, Ig, Immunoglobulin domain	143165	no	1e-08
[+]cl11960, Ig, Immunoglobulin domain	143743	N/A	7e-07
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma...	28945	no	8e-07
[+]pfam06583, Neogenin_C, Neogenin C-terminus	115253	N/A	3e-127
[+]cl00065, FN3, Fibronectin type 3 domain; One of three types of internal repeats found in the plasma...	140406	N/A	2e-04



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: cl11960

PubMed References

- Neogenin: one receptor, many functions. *Int. J. Biochem. Cell Biol. 2007; 39(5):874-878*
- 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):2009-2020*
- 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 1; 41(3):414-421*
- The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol. 1994 Sep 30; 242(4):309-320*
- The immunoglobulin superfamily: an insight on its tissular, species, and functional diversity. *J. Mol. Evol. 1998 Apr; 46(4):389-400*
- Evolution of antigen binding receptors. *Annu. Rev. Immunol. 1999; 17:109-147*

Statistics

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

Structure

Aligned Rows: All 7 rows

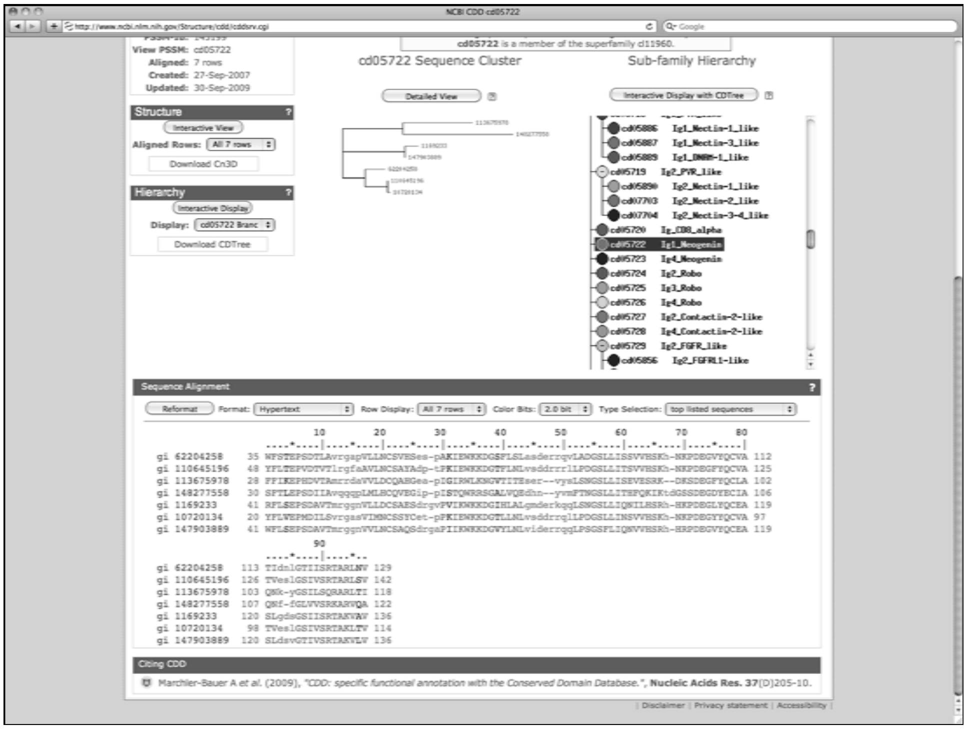
Hierarchy

Display: cd05722 Branc

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 cl11960.

cd05722 Sequence Cluster

Sub-family Hierarchy



View PSSM: cd05722
Aligned: 7 rows
Created: 27-Sep-2007
Updated: 30-Sep-2009

cd05722 is a member of the superfamily cl11960.

cd05722 Sequence Cluster

Sub-family Hierarchy

Structure

Aligned Rows: All 7 rows

Hierarchy

Display: cd05722 Branc

```
cd05886   Ig1_Nectin-1_like
cd05887   Ig1_Nectin-3_like
cd05889   Ig1_DNAM-1_like
cd05719   Ig2_PVR_like
cd05890   Ig2_Nectin-1_like
cd07703   Ig2_Nectin-2_like
cd07704   Ig2_Nectin-3-4_like
cd05720   Ig_CD8_alpha
cd05722   Ig1_Neogenin
cd05723   Ig4_Neogenin
cd05724   Ig2_Robo
cd05725   Ig3_Robo
cd05726   Ig4_Robo
cd05727   Ig2_Contactin-2-like
cd05728   Ig4_Contactin-2-like
cd05729   Ig2_FGFR_like
cd05856   Ig2_FGFRL1-like
```

Sequence Alignment

Format: Hypertext   Row Display: All 7 rows   Color Bits: 2.0 bit   Type Selection: top listed sequences

```
                      10        20        30        40        50        60        70        80
              ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 62204258    35 WFSTEPSDTLAvrgapVLLNCSVEGes-pAKIEWKKDGSFLSLasderrqvLADGSLLISSVVHSKh-NKPDEGVYQCVA 112
gi 110645196   48 YFLTEPVDTVTlrgfaAVLNCSAYAdp-tPKIEWKKDGTFLNLvsddrrrlLPDGSLLITSVVHSKh-NKPDEGFYQCVA 125
gi 113675978   28 FFIKEPHDVTAmrrdaVVLDCQAHGea-pIGIRWLKNGWTITEser--vysLSNGSLLISEVESRK--DKSDEGFYQCLA 102
gi 148277558   30 SFTLEPSDIIAvqqqpLMLHCQVEGip-pISTQWRRSGALVQEdhn--yvmFTNGSLLITHFQKIKtdGSSDEGDYECIA 106
gi 1169233     41 RFLSEPSDAVTmrggnVLLDCSAESdrgvPVIKWKKDGIHLALgmderkqqLSNGSLLIQNILHSRh-HKPDEGLYQCEA 119
gi 10720134    20 YFLWEPMDILSvrgasVIMNCSSYCet-pPKIEWKKDGTLLNLvsddrrqlLPDGSLLINSVVHSKh-NKPDEGYYQCVA 97
gi 147903889   41 WFLSEPSDAVTmrggnVVLNCSAQSdrgaPIIKWKKDGWYLNLviderrqqLPSGSFLIQNVVHSRh-HRPDEGVYQCEA 119

                      90
              ....*....|....*..
gi 62204258   113 TIdnlGTIISRTARLNV 129
gi 110645196  126 TVeslGSIVSRTARLSV 142
gi 113675978  103 QNk-yGSILSQRARLTI 118
gi 148277558  107 QNf-fGLVVSRKARVQA 122
gi 1169233    120 SLgdsGSIISRTAKVAV 136
gi 10720134    98 TVeslGSIVSRTAKLTV 114
gi 147903889  120 SLdsvGTIVSRTAKVLV 136
```

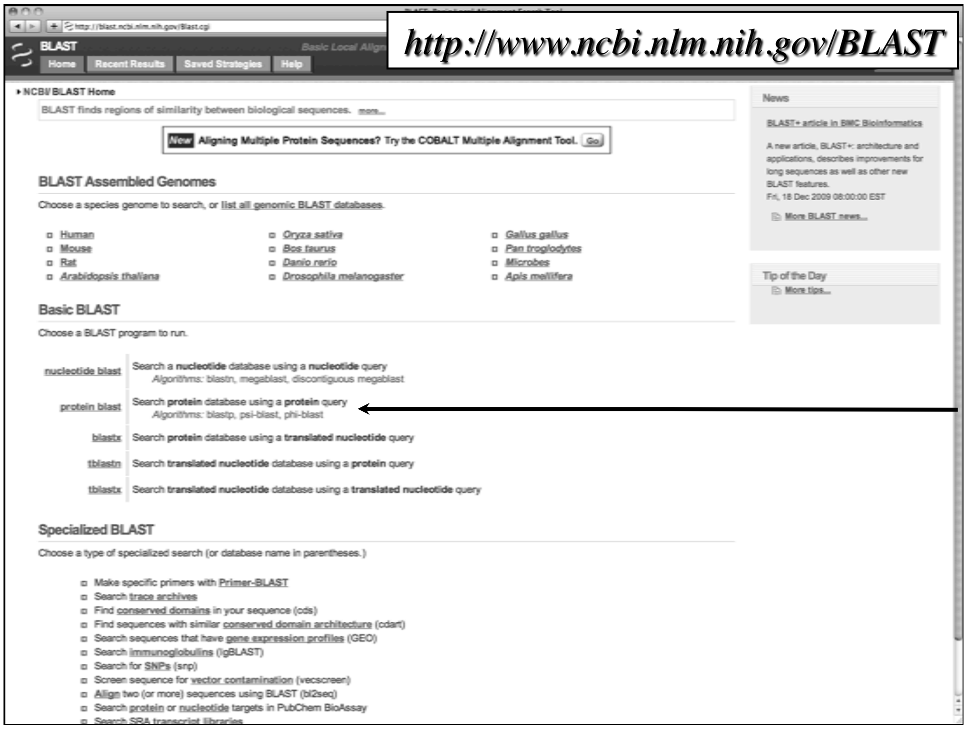
Citing CDD

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

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



http://www.ncbi.nlm.nih.gov/BLAST

▸ NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

Aligning Multiple Protein Sequences? Try the **COBALT Multiple Alignment Tool.**

BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

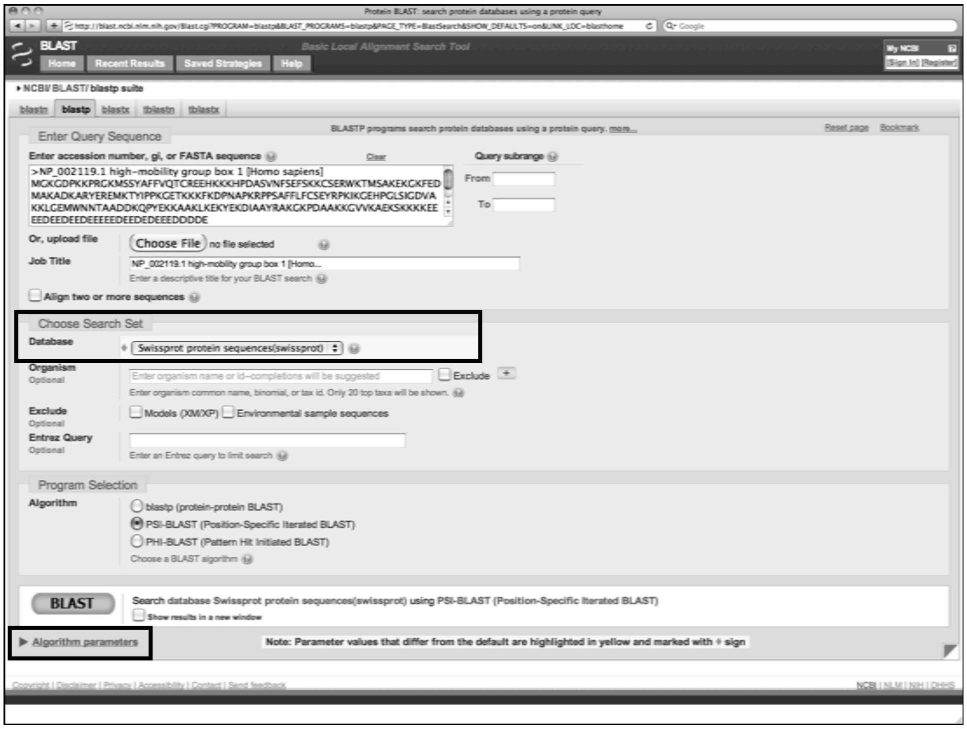
Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search for SNPs (snp)
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA transcript libraries

BLAST+ article in BMC Bioinformatics

A new article, BLAST+: architecture and applications, describes improvements for long sequences as well as other new BLAST features.
Fri, 18 Dec 2009 08:00:00 EST

More BLAST news...


Protein BLAST: search protein databases using a protein query
http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome
Google
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
blastn
blastp
blastx
tblastn
tblastx
Reset page
Bookmark
Clear
Choose File
BLAST
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

▸ NCBI/ BLAST/ blastp suite

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
>NP_002119.1 high-mobility group box 1 [Homo sapiens]
MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKFED
MAKADKARYEREMKTYIPPKGETKKKFKDPNAPKRPPSAFFLFCSEYRPKIKGEHPGLSIGDVA
KKLGEMWNNTAADDKQPYEKKAAKLKEKYEKDIAAYRAKGKPDAAKKGVVKAEKSKKKKEE
EEDEEDEEDEEEEEDEEDEDEEEDDDDE
```

Query subrange From □ To □

Or, upload file Choose File no file selected

Job Title NP_002119.1 high-mobility group box 1 [Homo...
Enter a descriptive title for your BLAST search

□ **Align two or more sequences**

Choose Search Set

Database Swissprot protein sequences(swissprot)

Organism *Optional* Enter organism name or id--completions will be suggested □ Exclude +
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude *Optional* □ Models (XM/XP) □ Environmental sample sequences

Entrez Query *Optional*
Enter an Entrez query to limit search

Program Selection

Algorithm
○ blastp (protein-protein BLAST)
● PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
Choose a BLAST algorithm

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

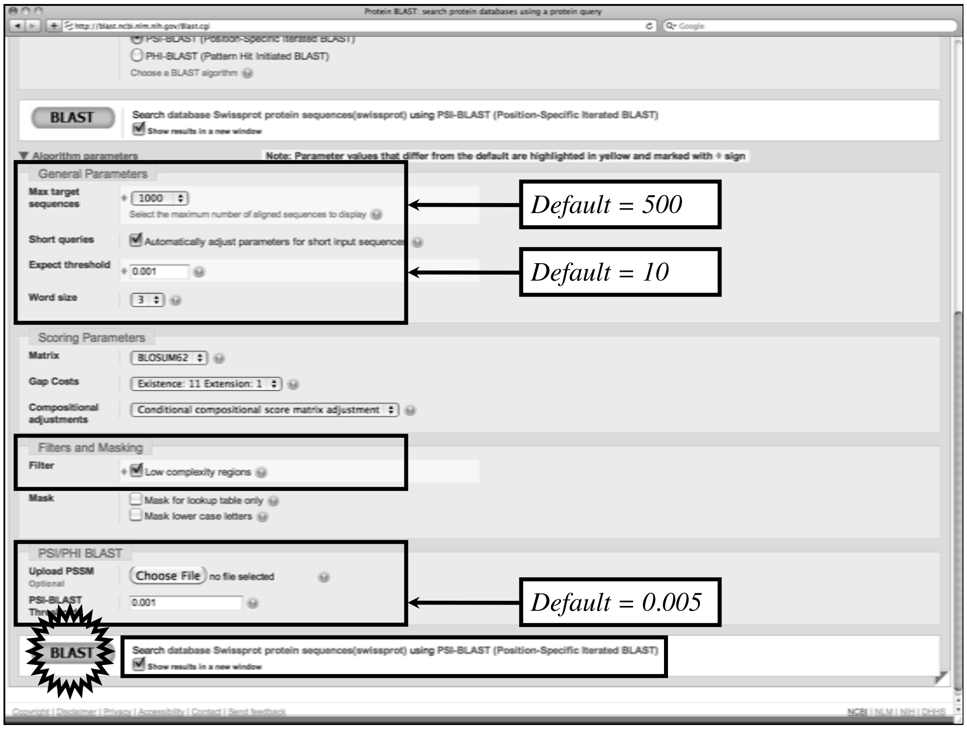
▸ Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

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`

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


Protein BLAST: search protein databases using a protein query
http://blast.ncbi.nlm.nih.gov/Blast.cgi
Google
BLAST
Choose File
Copyright | Disclaimer | Privacy | Accessibility | Contact | Send feedback
NCBI | NLM | NIH | DHHS

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

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 ♦ sign

General Parameters
Max target sequences ♦ 1000 — Select the maximum number of aligned sequences to display
Short queries ☑ Automatically adjust parameters for short input sequences
Expect threshold ♦ 0.001
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

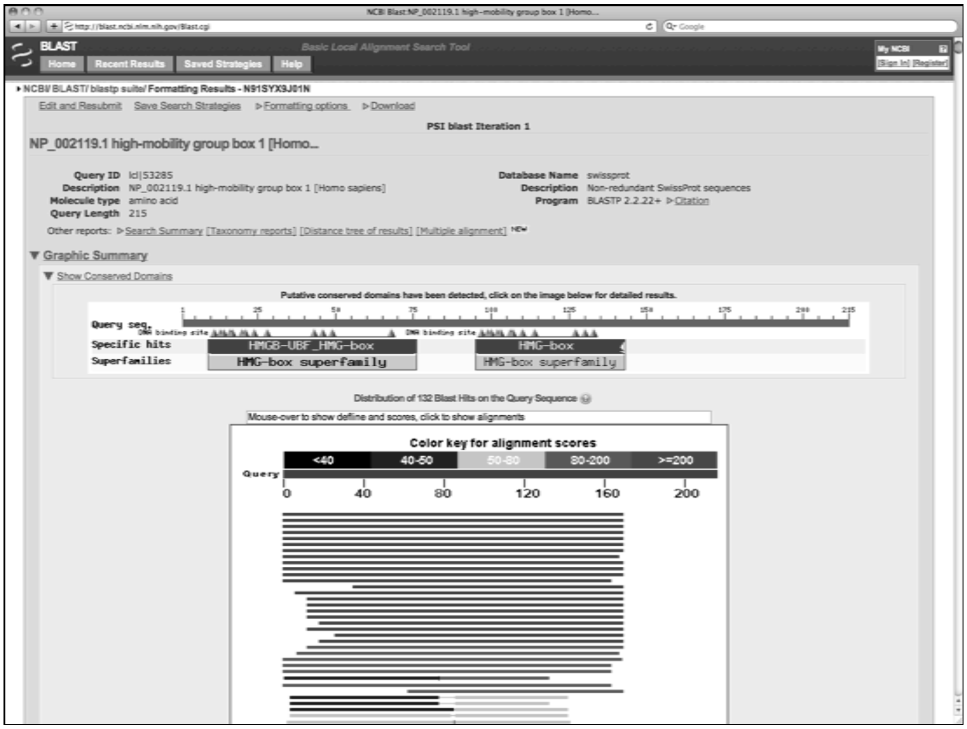
PSI/PHI BLAST
Upload PSSM Optional — no file selected
PSI-BLAST Threshold 0.001

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

Default = 500

Default = 10

Default = 0.005


NCBI Blast:NP_002119.1 high-mobility group box 1 [Homo...
http://blast.ncbi.nlm.nih.gov/Blast.cgi
Google
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - N91SYX9J01N

Edit and Resubmit Save Search Strategies ▷Formatting options ▷Download

PSI blast Iteration 1

NP_002119.1 high-mobility group box 1 [Homo...

Query ID	lcl|53285	**Database Name**	swissprot
Description	NP_002119.1 high-mobility group box 1 [Homo sapiens]	**Description**	Non-redundant SwissProt sequences
Molecule type	amino acid	**Program**	BLASTP 2.2.22+ ▷Citation
Query Length	215		

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

▼ Graphic Summary

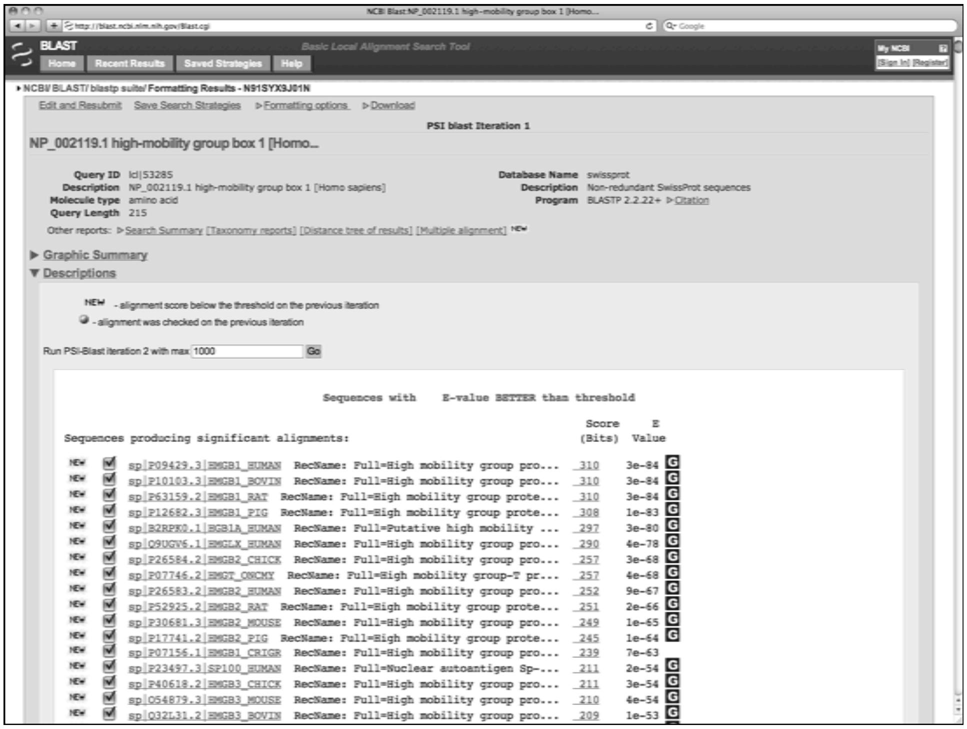
▼ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 132 Blast Hits on the Query Sequence

Mouse-over to show define 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


NCBI Blast:NP_002119.1 high-mobility group box 1 [Homo...
http://blast.ncbi.nlm.nih.gov/Blast.cgi
Google
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - N91SYX9J01N

Edit and Resubmit Save Search Strategies ▷Formatting options ▷Download

PSI blast Iteration 1

NP_002119.1 high-mobility group box 1 [Homo...

Query ID	lcl|53285	**Database Name**	swissprot
Description	NP_002119.1 high-mobility group box 1 [Homo sapiens]	**Description**	Non-redundant SwissProt sequences
Molecule type	amino acid	**Program**	BLASTP 2.2.22+ ▷Citation
Query Length	215		

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

▶ Graphic Summary

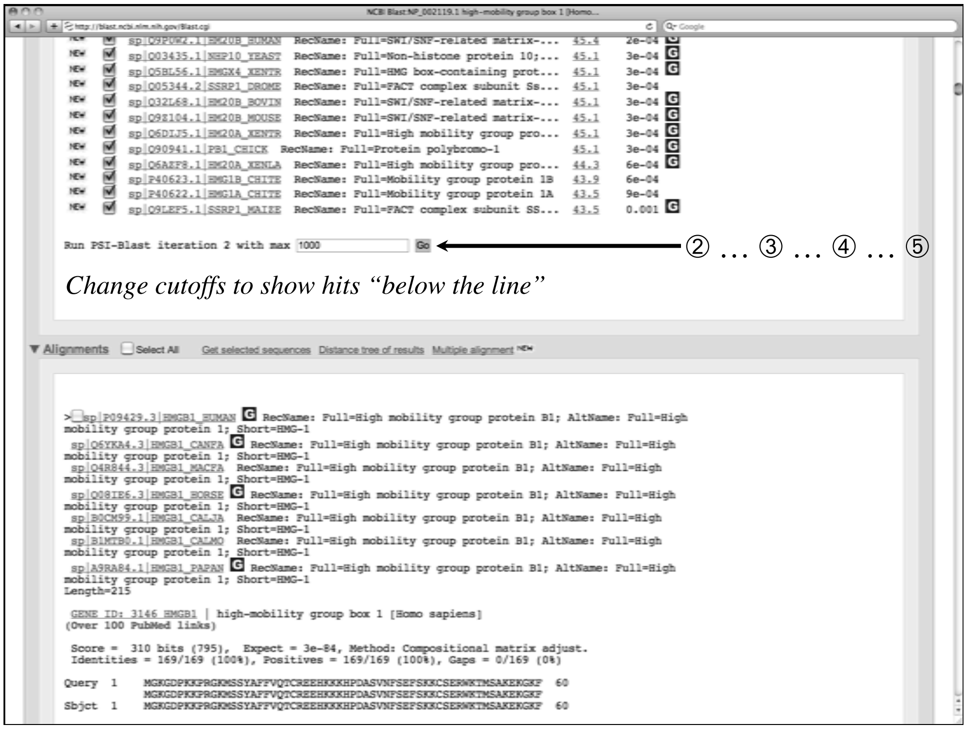
▼ Descriptions

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 Go

```
                          Sequences with    E-value BETTER than threshold

                                                                          Score    E
Sequences producing significant alignments:                              (Bits)  Value

NEW ☑ sp|P09429.3|HMGB1_HUMAN  RecName: Full=High mobility group pro...   310    3e-84 G
NEW ☑ sp|P10103.3|HMGB1_BOVIN  RecName: Full=High mobility group pro...   310    3e-84 G
NEW ☑ sp|P63159.2|HMGB1_RAT  RecName: Full=High mobility group prote...   310    3e-84 G
NEW ☑ sp|P12682.3|HMGB1_PIG  RecName: Full=High mobility group prote...   308    1e-83 G
NEW ☑ sp|B2RPK0.1|HGB1A_HUMAN  RecName: Full=Putative high mobility ...   297    3e-80 G
NEW ☑ sp|Q9UGV6.1|HMGLX_HUMAN  RecName: Full=High mobility group pro...   290    4e-78 G
NEW ☑ sp|P26584.2|HMGB2_CHICK  RecName: Full=High mobility group pro...   257    3e-68 G
NEW ☑ sp|P07746.2|HMGT_ONCMY  RecName: Full=High mobility group-T pr...   257    4e-68 G
NEW ☑ sp|P26583.2|HMGB2_HUMAN  RecName: Full=High mobility group pro...   252    9e-67 G
NEW ☑ sp|P52925.2|HMGB2_RAT  RecName: Full=High mobility group prote...   251    2e-66 G
NEW ☑ sp|P30681.3|HMGB2_MOUSE  RecName: Full=High mobility group pro...   249    1e-65 G
NEW ☑ sp|P17741.2|HMGB2_PIG  RecName: Full=High mobility group prote...   245    1e-64 G
NEW ☑ sp|P07156.1|HMGB1_CRIGR  RecName: Full=High mobility group pro...   239    7e-63
NEW ☑ sp|P23497.3|SP100_HUMAN  RecName: Full=Nuclear autoantigen Sp-...   211    2e-54 G
NEW ☑ sp|P40618.2|HMGB3_CHICK  RecName: Full=High mobility group pro...   211    3e-54 G
NEW ☑ sp|O54879.3|HMGB3_MOUSE  RecName: Full=High mobility group pro...   210    4e-54 G
NEW ☑ sp|Q32L31.2|HMGB3_BOVIN  RecName: Full=High mobility group pro...   209    1e-53 G
```

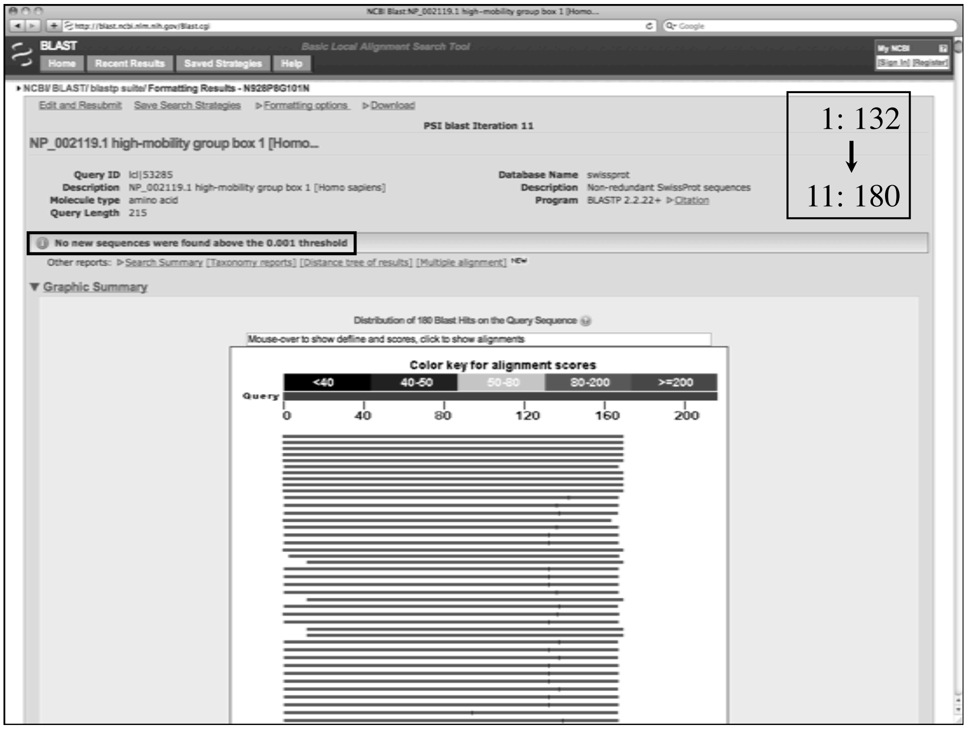


```
NEW ☑ sp|Q9P0W2.1|HM20B_HUMAN  RecName: Full=SWI/SNF-related matrix-...  45.4   2e-04 G
NEW ☑ sp|Q03435.1|NHP10_YEAST  RecName: Full=Non-histone protein 10;...  45.1   3e-04 G
NEW ☑ sp|Q5BL56.1|HMGX4_XENTR  RecName: Full=HMG box-containing prot...  45.1   3e-04 G
NEW ☑ sp|Q05344.2|SSRP1_DROME  RecName: Full=FACT complex subunit Ss...  45.1   3e-04
NEW ☑ sp|Q32L68.1|HM20B_BOVIN  RecName: Full=SWI/SNF-related matrix-...  45.1   3e-04 G
NEW ☑ sp|Q9Z104.1|HM20B_MOUSE  RecName: Full=SWI/SNF-related matrix-...  45.1   3e-04 G
NEW ☑ sp|Q6DIJ5.1|HM20A_XENTR  RecName: Full=High mobility group pro...  45.1   3e-04 G
NEW ☑ sp|Q90941.1|PB1_CHICK  RecName: Full=Protein polybromo-1           45.1   3e-04 G
NEW ☑ sp|Q6AZF8.1|HM20A_XENLA  RecName: Full=High mobility group pro...  44.3   6e-04 G
NEW ☑ sp|P40623.1|HMG1B_CHITE  RecName: Full=Mobility group protein 1B   43.9   6e-04
NEW ☑ sp|P40622.1|HMG1A_CHITE  RecName: Full=Mobility group protein 1A   43.5   9e-04
NEW ☑ sp|Q9LEF5.1|SSRP1_MAIZE  RecName: Full=FACT complex subunit SS...  43.5   0.001 G

Run PSI-Blast iteration 2 with max 1000 Go
```

← ② … ③ … ④ … ⑤

Change cutoffs to show hits “below the line”

```
>□sp|P09429.3|HMGB1_HUMAN G RecName: Full=High mobility group protein B1; AltName: Full=High
mobility group protein 1; Short=HMG-1
 sp|Q6YKA4.3|HMGB1_CANFA G RecName: Full=High mobility group protein B1; AltName: Full=High
mobility group protein 1; Short=HMG-1
 sp|Q4R844.3|HMGB1_MACFA  RecName: Full=High mobility group protein B1; AltName: Full=High
mobility group protein 1; Short=HMG-1
 sp|Q08IE6.3|HMGB1_HORSE G RecName: Full=High mobility group protein B1; AltName: Full=High
mobility group protein 1; Short=HMG-1
 sp|B0CM99.1|HMGB1_CALJA  RecName: Full=High mobility group protein B1; AltName: Full=High
mobility group protein 1; Short=HMG-1
 sp|B1MTB0.1|HMGB1_CALMO  RecName: Full=High mobility group protein B1; AltName: Full=High
mobility group protein 1; Short=HMG-1
 sp|A9RA84.1|HMGB1_PAPAN G 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    MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKF  60
            MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKF
Sbjct  1    MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKF  60
```



NCBI Blast:NP_002119.1 high-mobility group box 1 [Homo...
http://blast.ncbi.nlm.nih.gov/Blast.cgi
Google
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
[Sign In] [Register]
▸ NCBI/ BLAST/ blastp suite/ Formatting Results - N928P8G101N
Edit and Resubmit
Save Search Strategies
▷Formatting options
▷Download

PSI blast Iteration 11

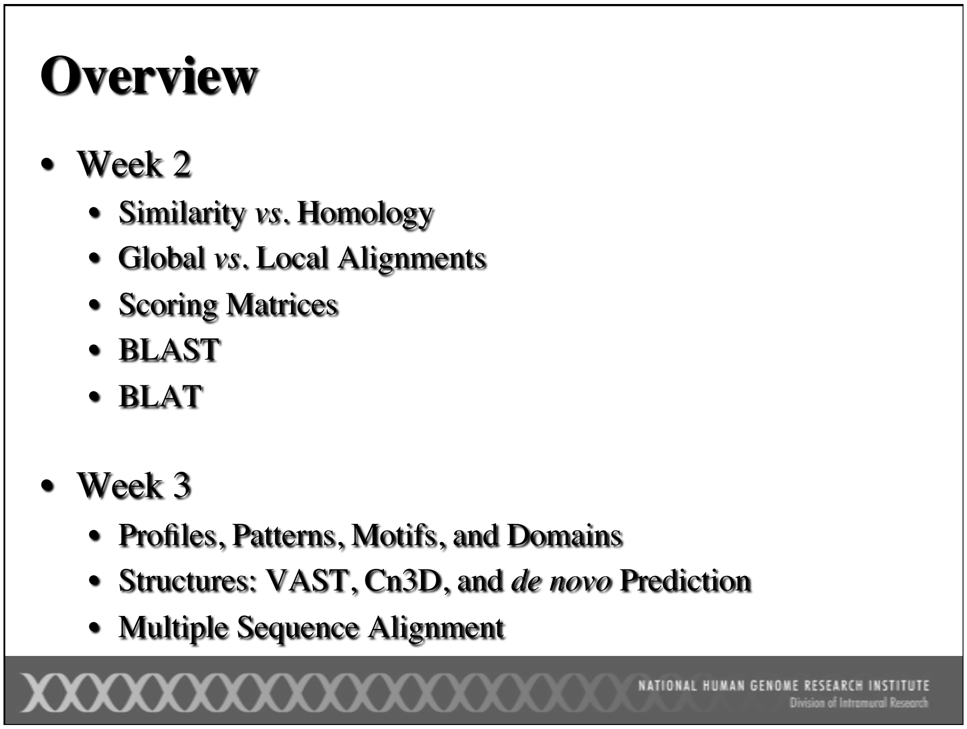
NP_002119.1 high-mobility group box 1 [Homo...

Query ID	lcl|53285	Database Name	swissprot
Description	NP_002119.1 high-mobility group box 1 [Homo sapiens]	Description	Non-redundant SwissProt sequences
Molecule type	amino acid	Program	BLASTP 2.2.22+ ▷Citation
Query Length	215		

ⓘ No new sequences were found above the 0.001 threshold

1: 132
↓
11: 180

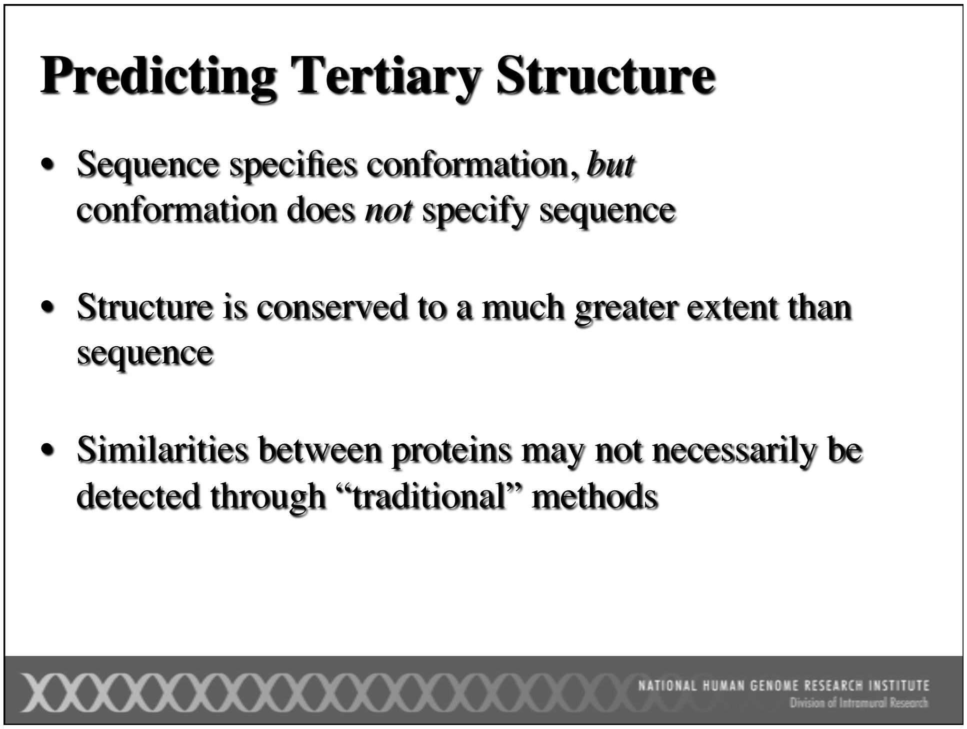
Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] NEW
▼ Graphic Summary
Distribution of 180 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
40
80
120
160
200



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

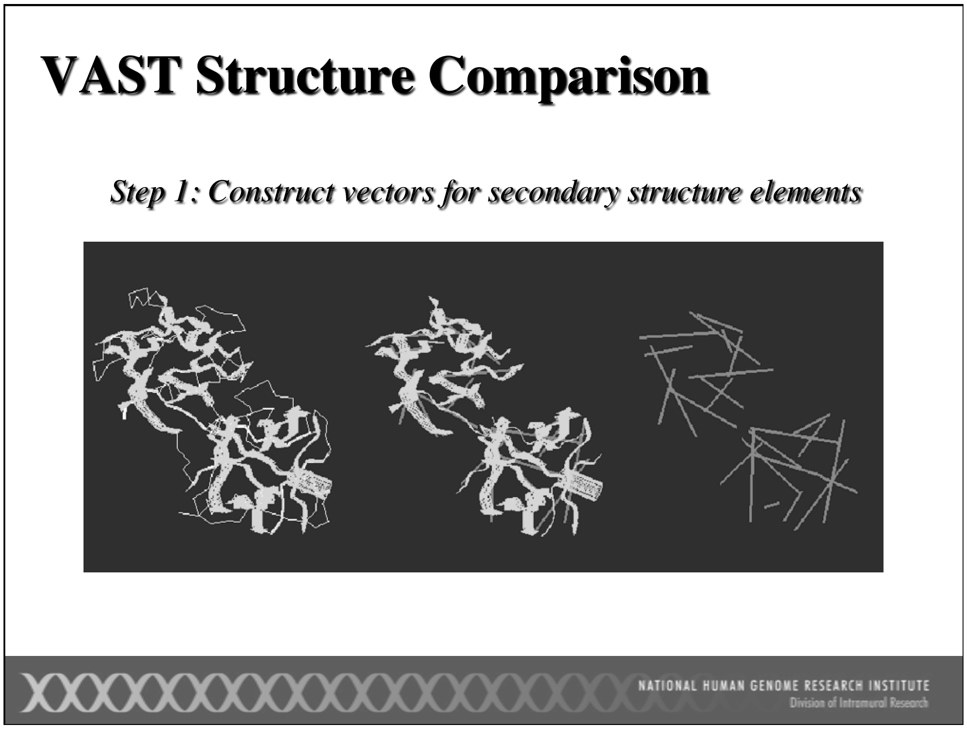
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

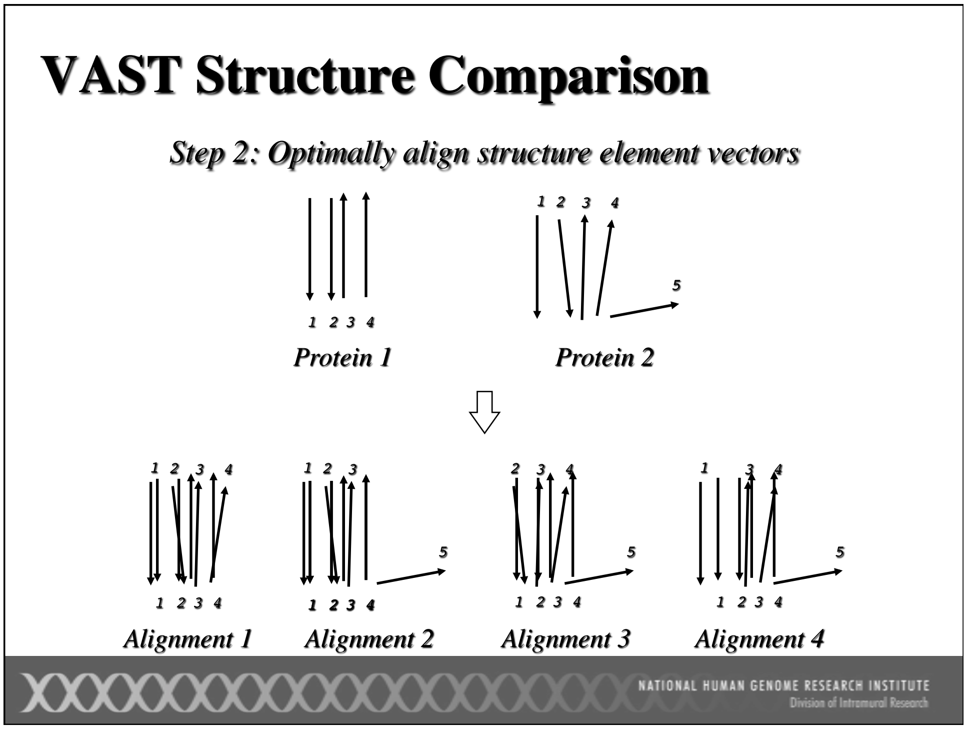
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



VAST Structure Comparison

Step 1: Construct vectors for secondary structure elements

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



VAST Structure Comparison

Step 2: Optimally align structure element vectors

Protein 1

Protein 2

Alignment 1 *Alignment 2* *Alignment 3* *Alignment 4*

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

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

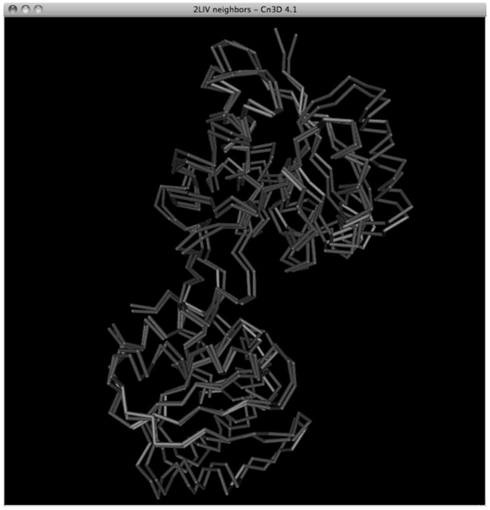
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

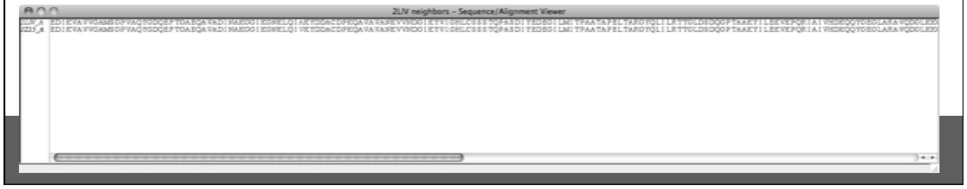
Cn3D Viewer

Rendering: Tubes

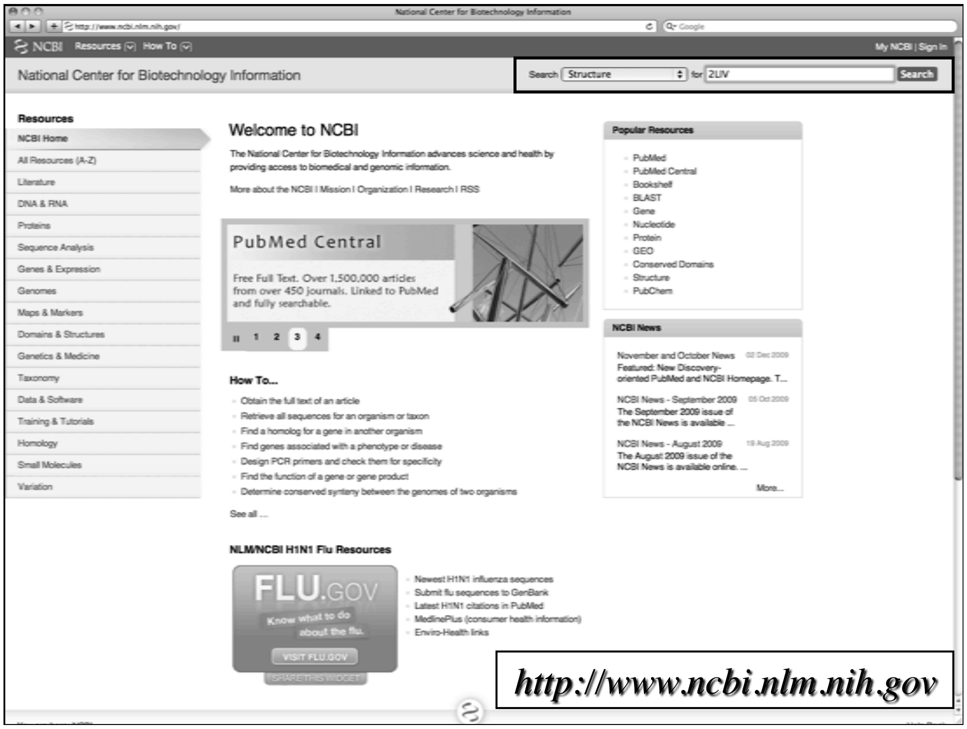
Coloring: Identity

Red – matches
Blue – mismatches





```
2LIV_A  EDIKVAVVGAMSGPVAQYGDQEFTGAEQAVADINAKGGIKGNKLQIAKYDDACDPKQAVAVANKVVNDGIKYVIGHLCSSSTQPASDIYEDEGILMITPAATAPELTARGYQLILRTTGLDSDQGPTAAKYILEKVKPQRIAIVHDKQQYGEGLARAVQDGLKK
1Z15_A  EDIKVAVVGAMSGPVAQYGDQEFTGAEQAVADINAKGGIKGNKLQIVKYDDACDPKQAVAVANKVVNDGIKYVIGHLCSSSTQPASDIYEDEGILMITPAATAPELTARGYQLILRTTGLDSDQGPTAAKYILEKVKPQRIAIVHDKQQYGEGLARAVQDGLKK
```



National Center for Biotechnology Information

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

NCBI Resources How To My NCBI | Sign In

National Center for Biotechnology Information

Search Structure for 2LIV Search

Resources

- NCBI Home
- All Resources (A-Z)
- Literature
- DNA & RNA
- Proteins
- Sequence Analysis
- Genes & Expression
- Genomes
- Maps & Markers
- Domains & Structures
- Genetics & Medicine
- Taxonomy
- Data & Software
- Training & Tutorials
- Homology
- Small Molecules
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

More about the NCBI | Mission | Organization | Research | RSS

PubMed Central

Free Full Text. Over 1,500,000 articles from over 450 journals. Linked to PubMed and fully searchable.

II 1 2 3 4

How To...

- Obtain the full text of an article
- Retrieve all sequences for an organism or taxon
- Find a homolog for a gene in another organism
- Find genes associated with a phenotype or disease
- Design PCR primers and check them for specificity
- Find the function of a gene or gene product
- Determine conserved synteny between the genomes of two organisms

See all ...

NLM/NCBI H1N1 Flu Resources

FLU.GOV Know what to do about the flu. VISIT FLU.GOV SHARE THIS WIDGET

- Newest H1N1 influenza sequences
- Submit flu sequences to GenBank
- Latest H1N1 citations in PubMed
- MedlinePlus (consumer health information)
- Enviro-Health links

Popular Resources

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

NCBI News

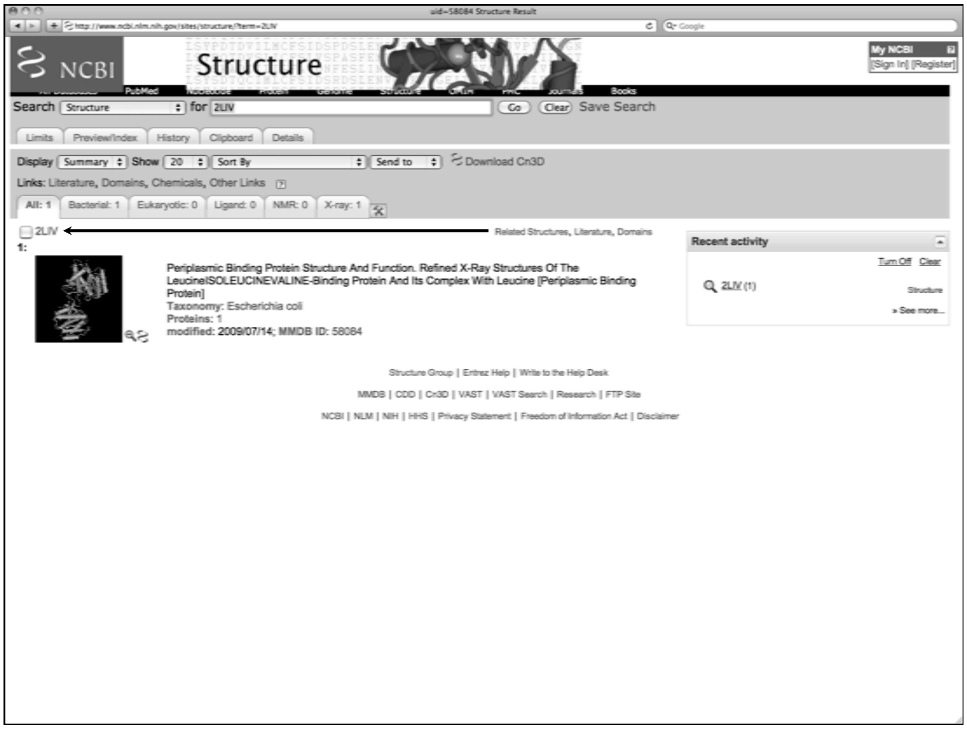
November and October News 02 Dec 2009
Featured: New Discovery-oriented PubMed and NCBI Homepage. T...

NCBI News - September 2009 05 Oct 2009
The September 2009 issue of the NCBI News is available ...

NCBI News - August 2009 19 Aug 2009
The August 2009 issue of the NCBI News is available online. ...

More...

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



uid=58084 Structure Result

http://www.ncbi.nlm.nih.gov/sites/structure/?term=2LIV

NCBI Structure

My NCBI [Sign In] [Register]

PubMed Nucleotide Protein Genome Structure OMIM PMC Journals 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

Links: Literature, Domains, Chemicals, Other Links

All: 1 Bacterial: 1 Eukaryotic: 0 Ligand: 0 NMR: 0 X-ray: 1

□ 2LIV ← Related Structures, Literature, Domains

1:

Periplasmic Binding Protein Structure And Function. Refined X-Ray Structures Of The LeucineISOLEUCINEVALINE-Binding Protein And Its Complex With Leucine [Periplasmic Binding Protein]
Taxonomy: Escherichia coli
Proteins: 1
modified: 2009/07/14; MMDB ID: 58084

Recent activity

Turn Off Clear

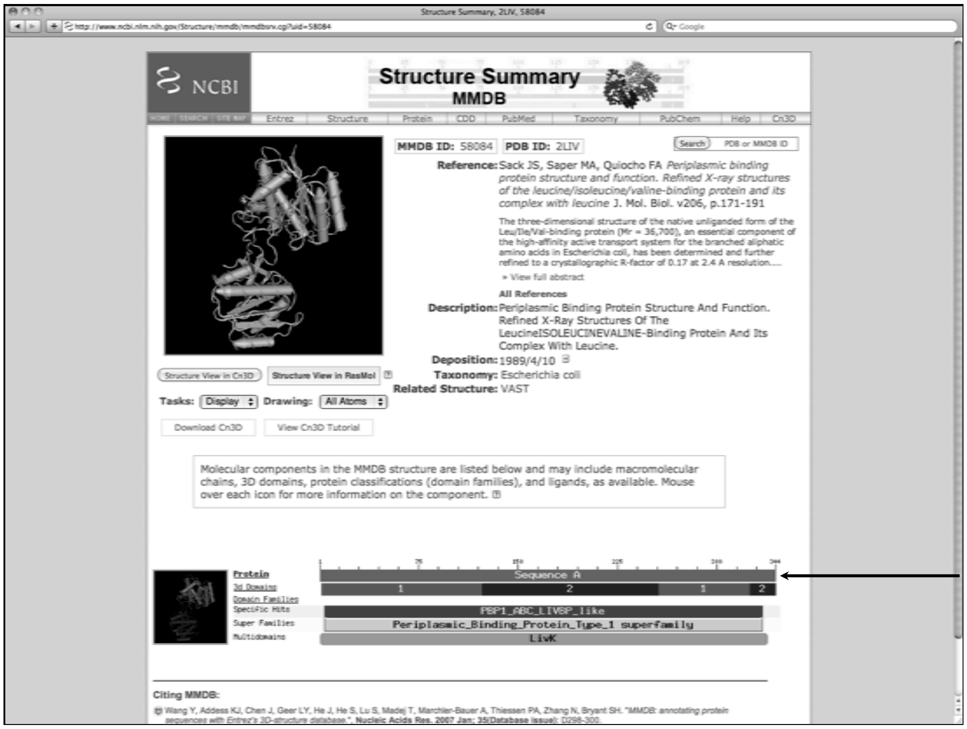
2LIV (1) Structure

» See more...

Structure Group | Entrez Help | Write to the Help Desk

MMDB | CDD | Cn3D | VAST | VAST Search | Research | FTP Site

NCBI | NLM | NIH | HHS | Privacy Statement | Freedom of Information Act | Disclaimer



MMDB ID: 58084 **PDB ID:** 2LIV

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

The three-dimensional structure of the native unliganded form of the Leu/Ile/Val-binding protein (Mr = 36,700), an essential component of the high-affinity active transport system for the branched aliphatic amino acids in Escherichia coli, has been determined and further refined to a crystallographic R-factor of 0.17 at 2.4 A resolution....

» View full abstract

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

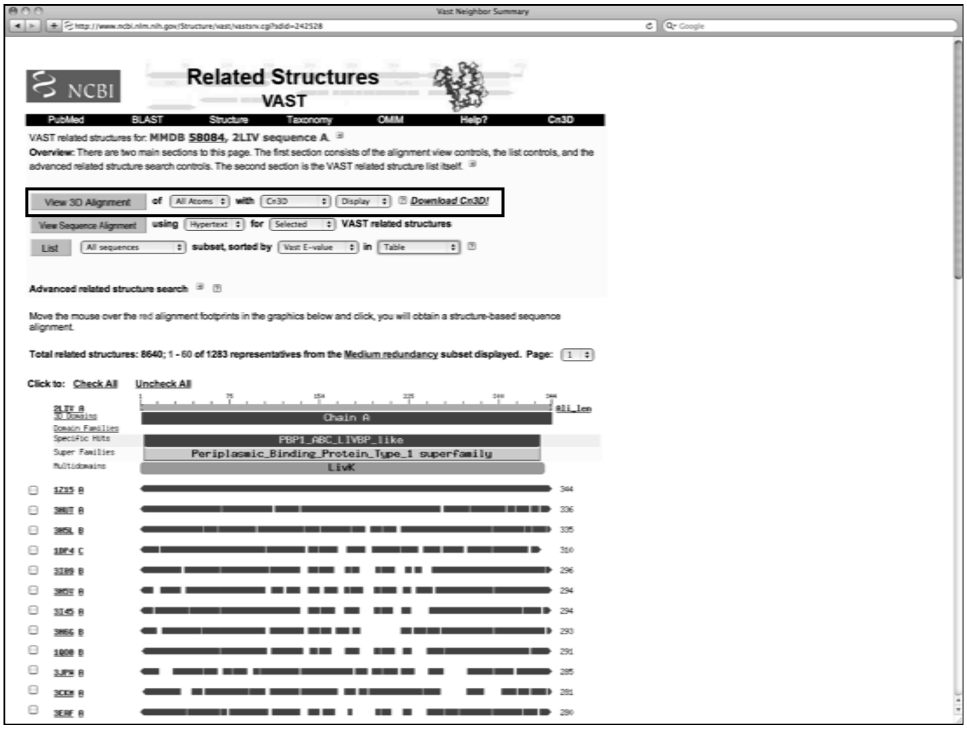
Related Structure: VAST

Tasks: Display **Drawing:** All Atoms

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.

Citing MMDB:

Wang Y, Addess KJ, Chen J, Geer LY, He J, He S, Lu S, Madej T, Marchler-Bauer A, Thiessen PA, Zhang N, Bryant SH. "MMDB: *annotating protein sequences with Entrez's 3D-structure database*.", **Nucleic Acids Res. 2007 Jan; 35(Database issue):** D298-300.



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

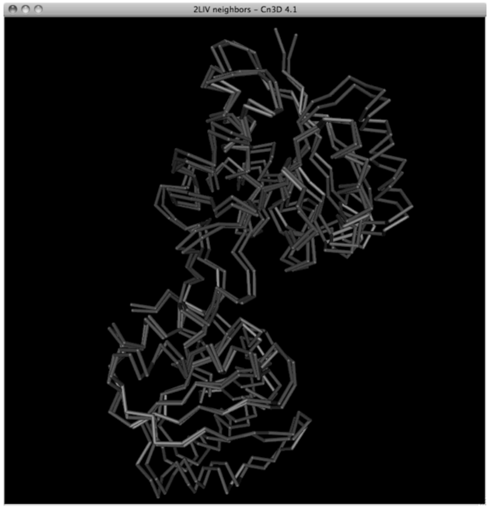
Click to: Check All Uncheck All

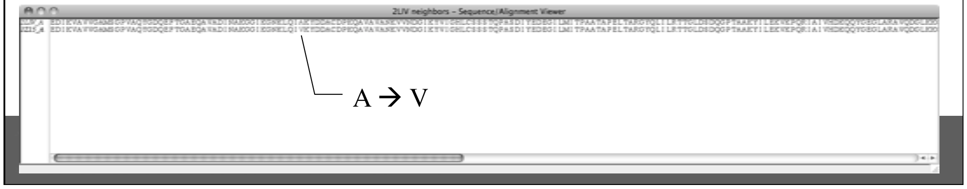
Cn3D Viewer

Rendering: Tubes

Coloring: Identity

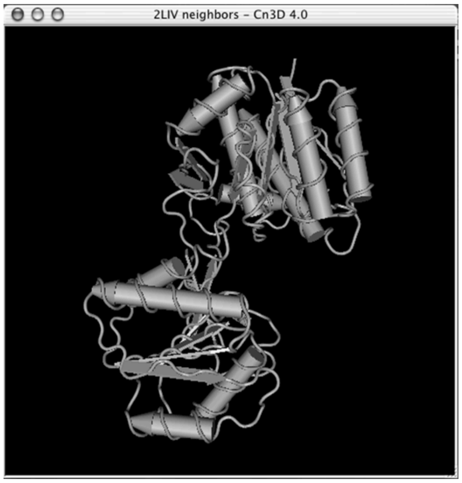
Red – matches
Blue – mismatches

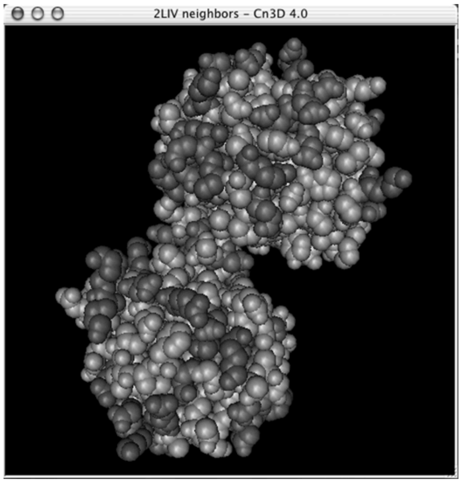




[illegible]

A → V





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

Further Reading

Searching the NCBI Databases Using Entrez

UNIT 1.3

Current Protocols in Bioinformatics
Unit 1.3
Entrez and Cn3D

An Introduction to Modeling Structure from Sequence

UNIT 5.1

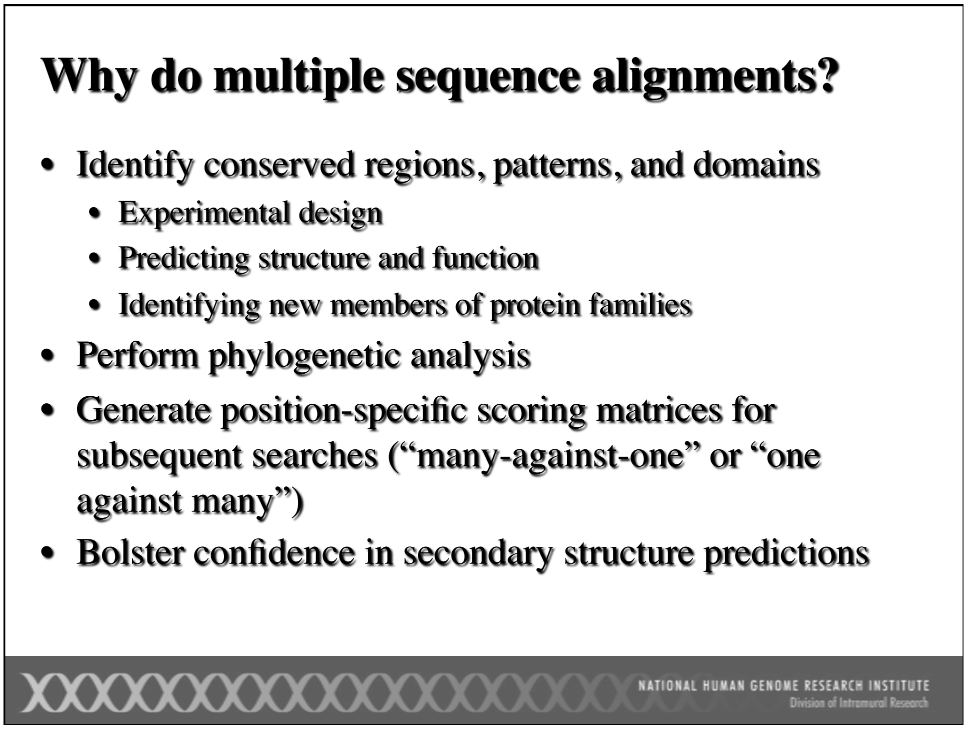
Current Protocols in Bioinformatics
Unit 5.1
An Introduction to Modeling Protein Structure from Sequence

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

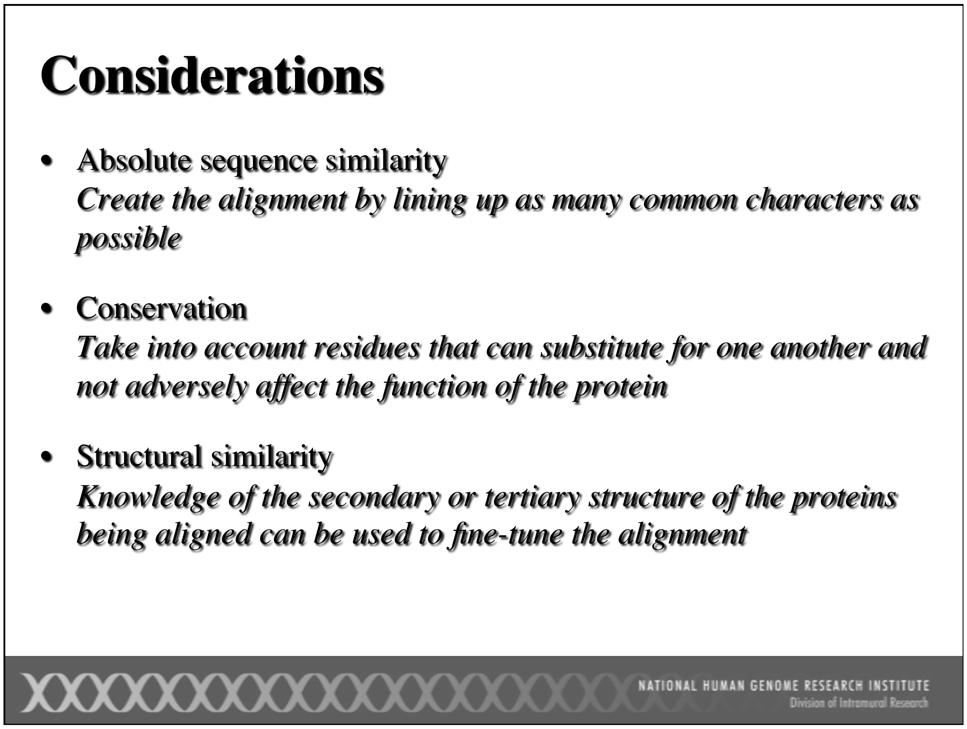
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



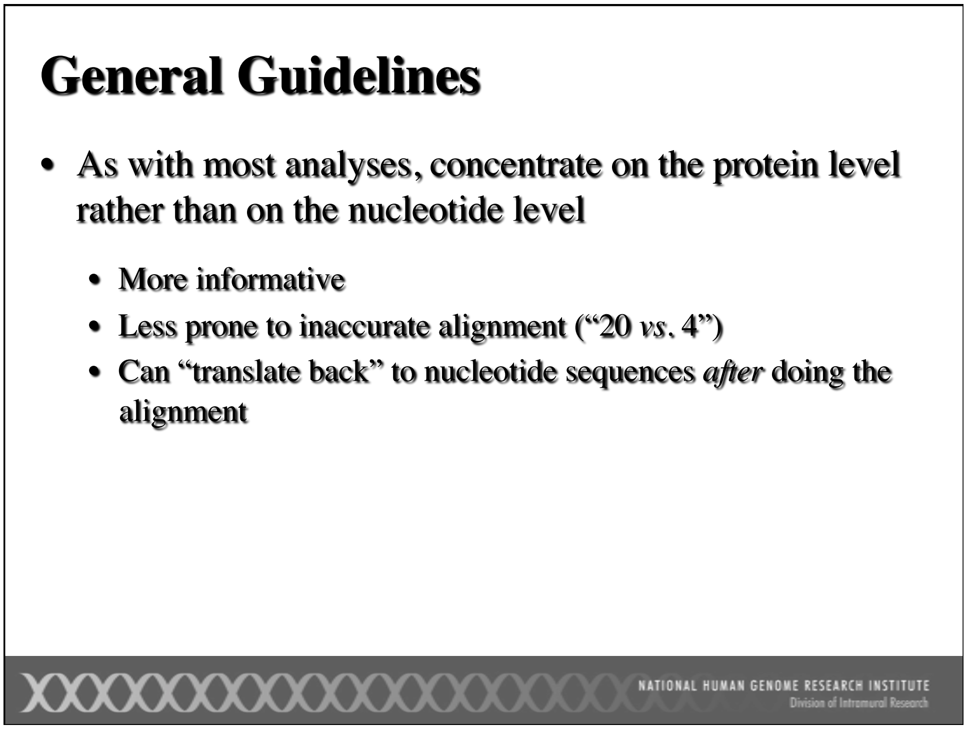
Considerations

- 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

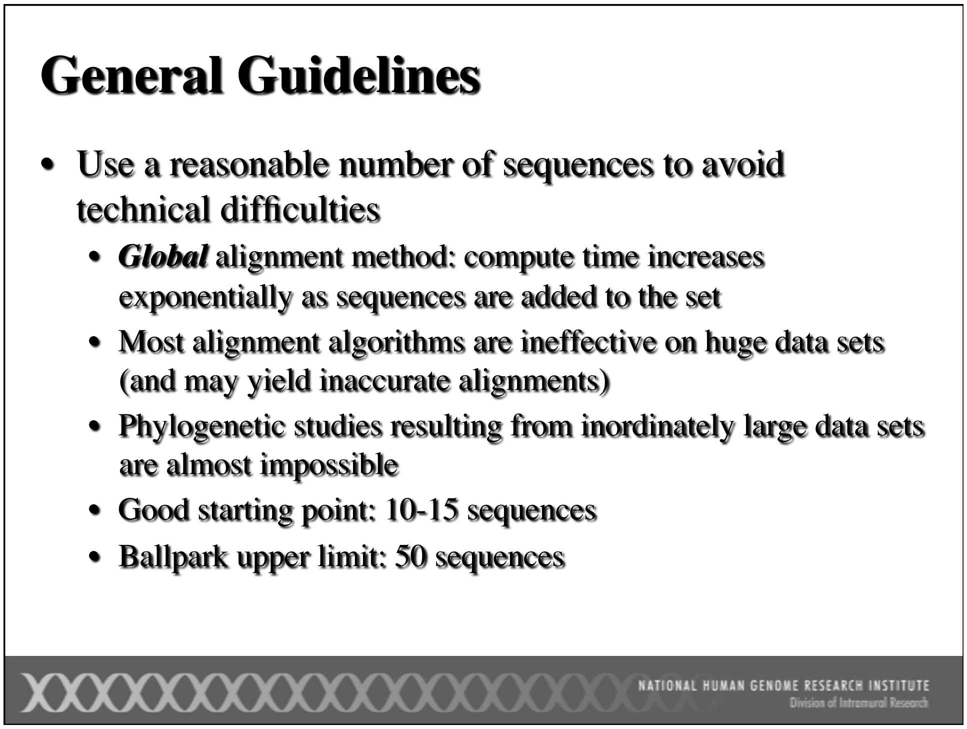
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

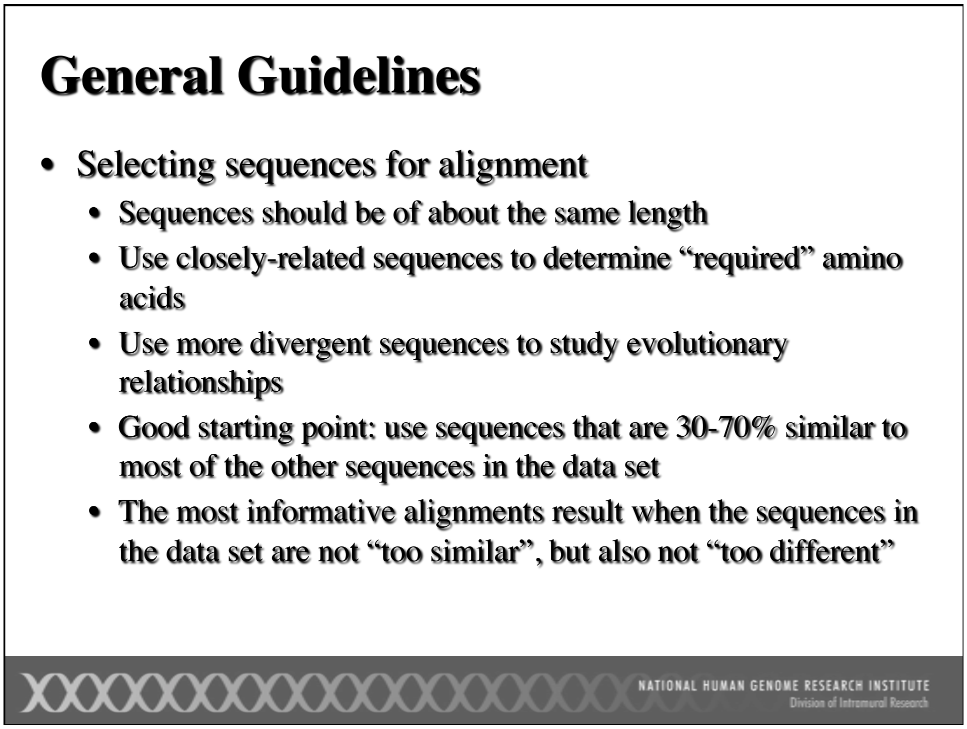
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

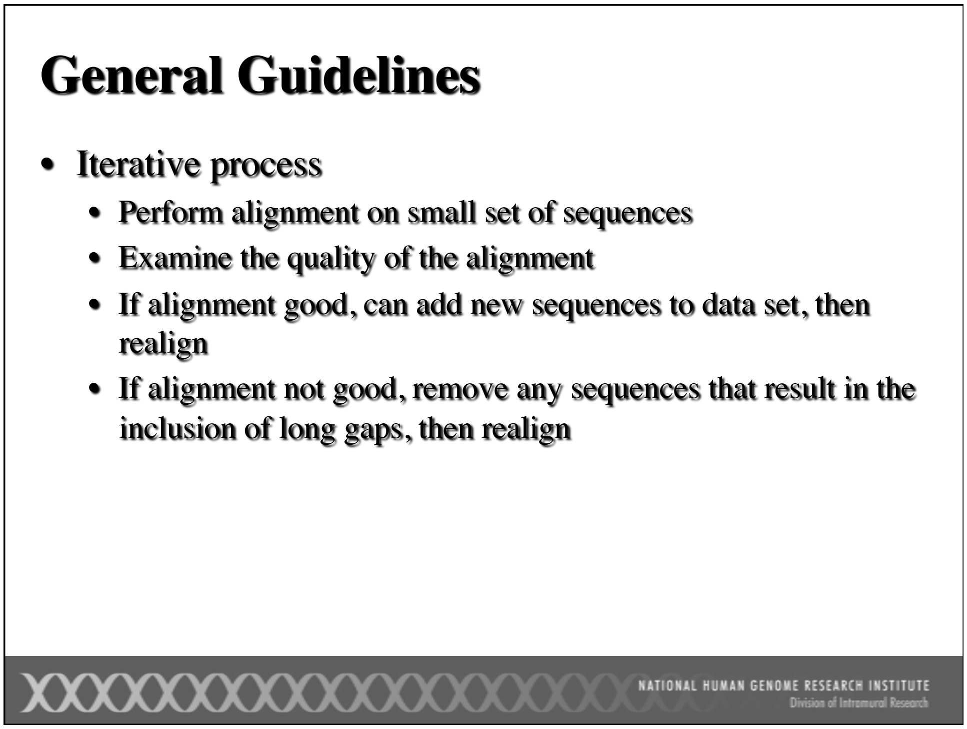
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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"

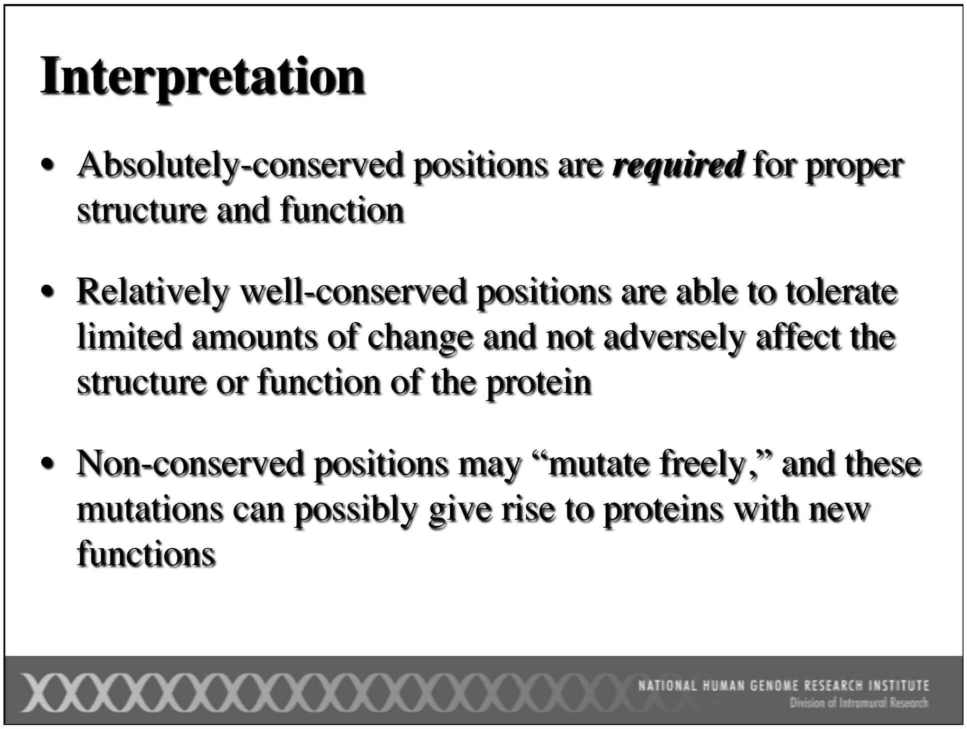
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

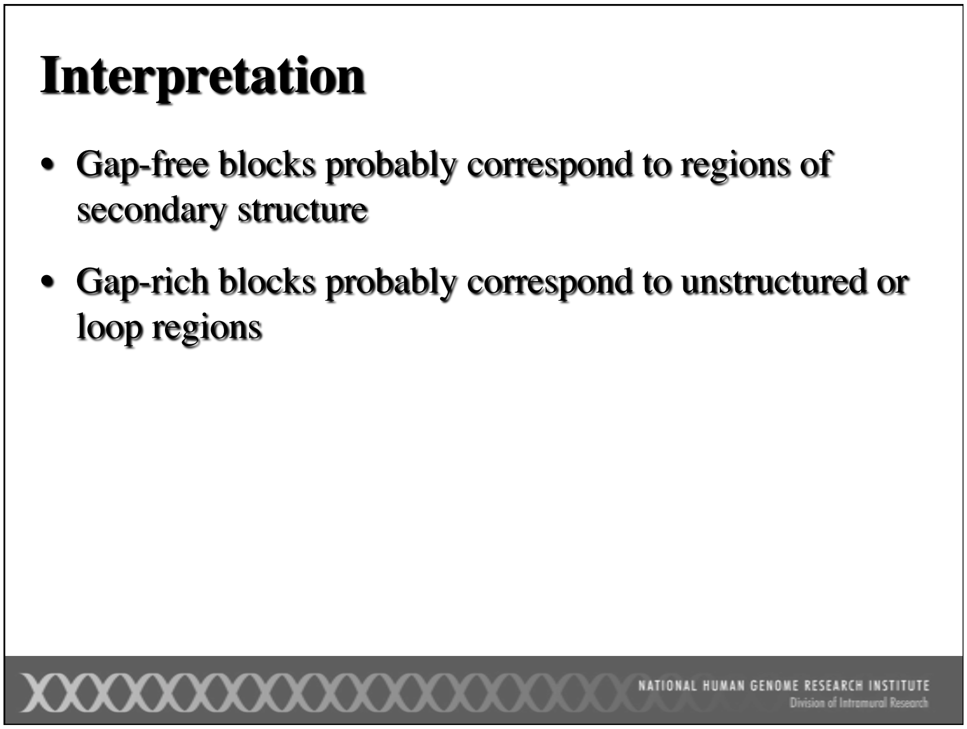
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

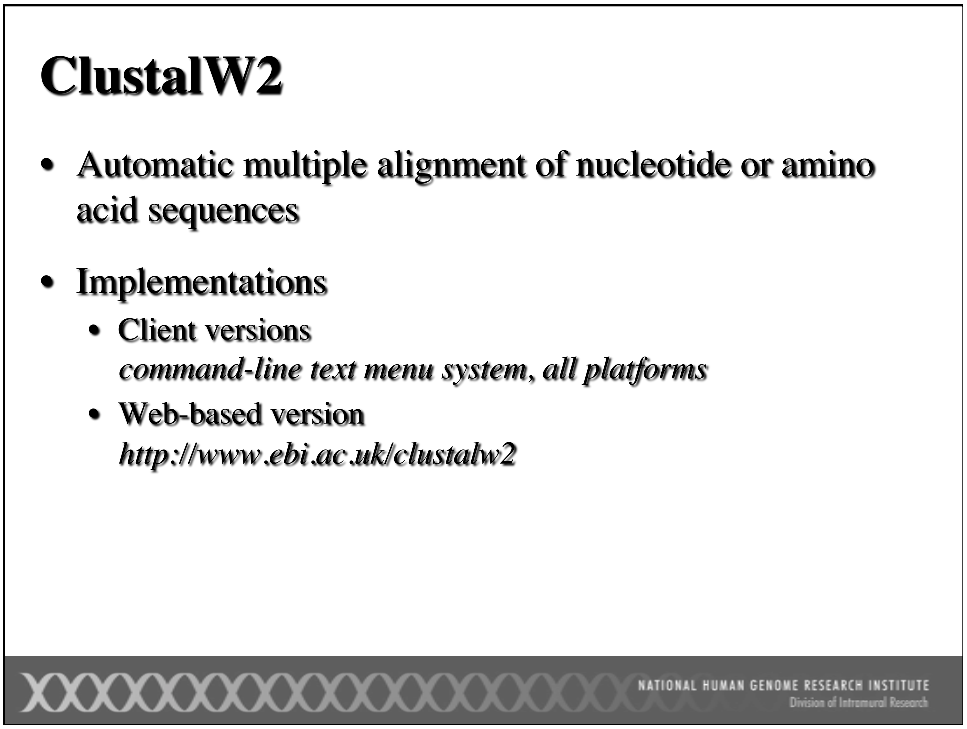
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Interpretation

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

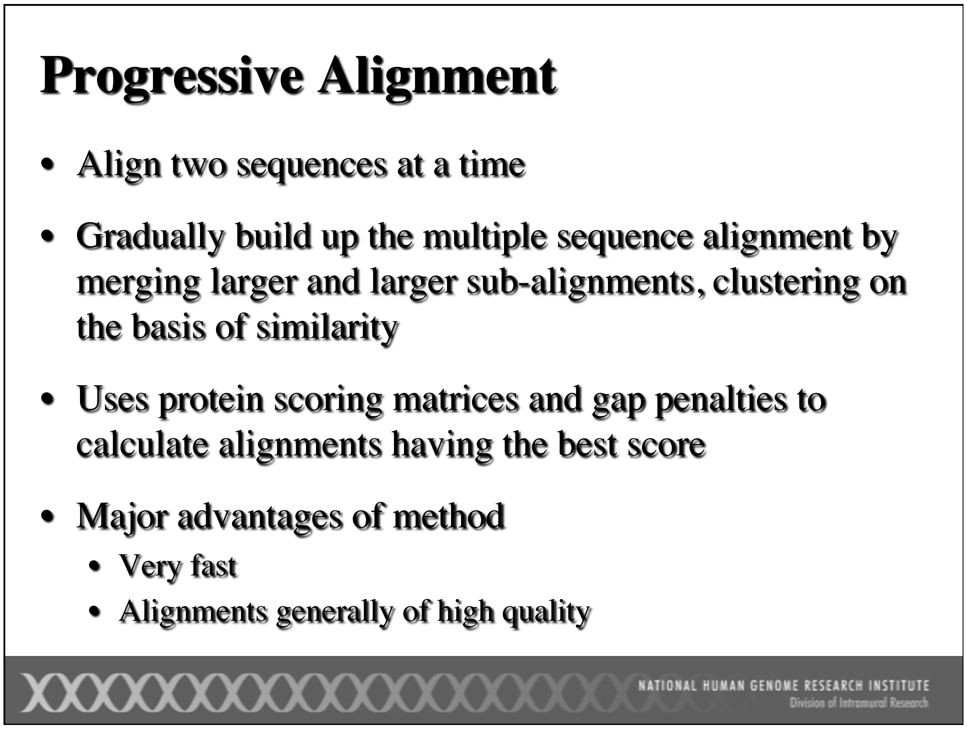
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

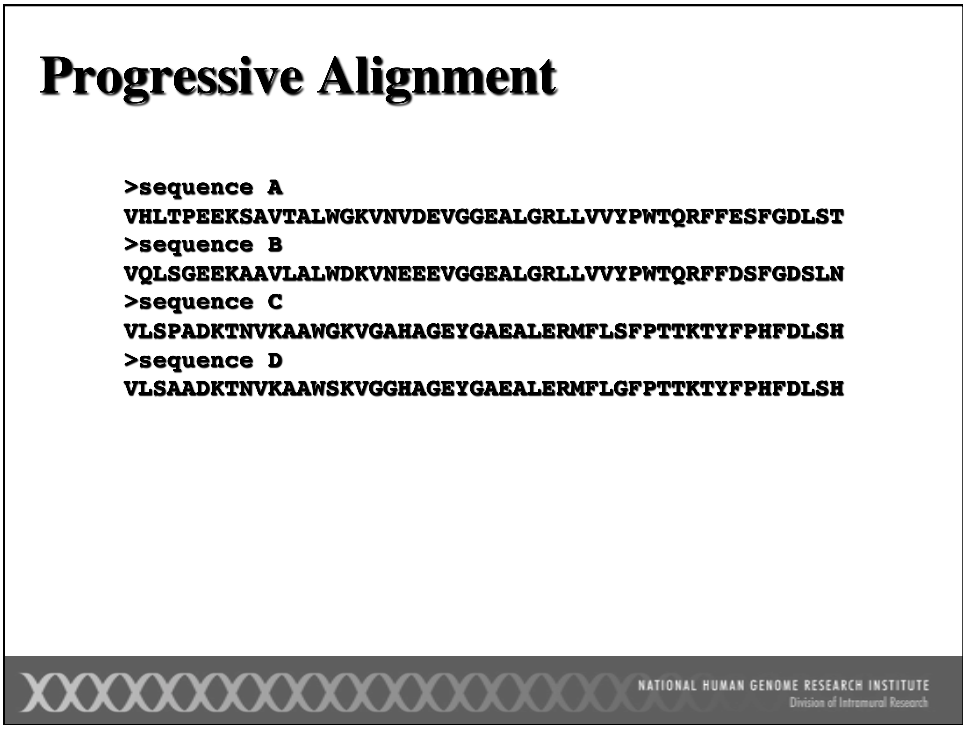
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

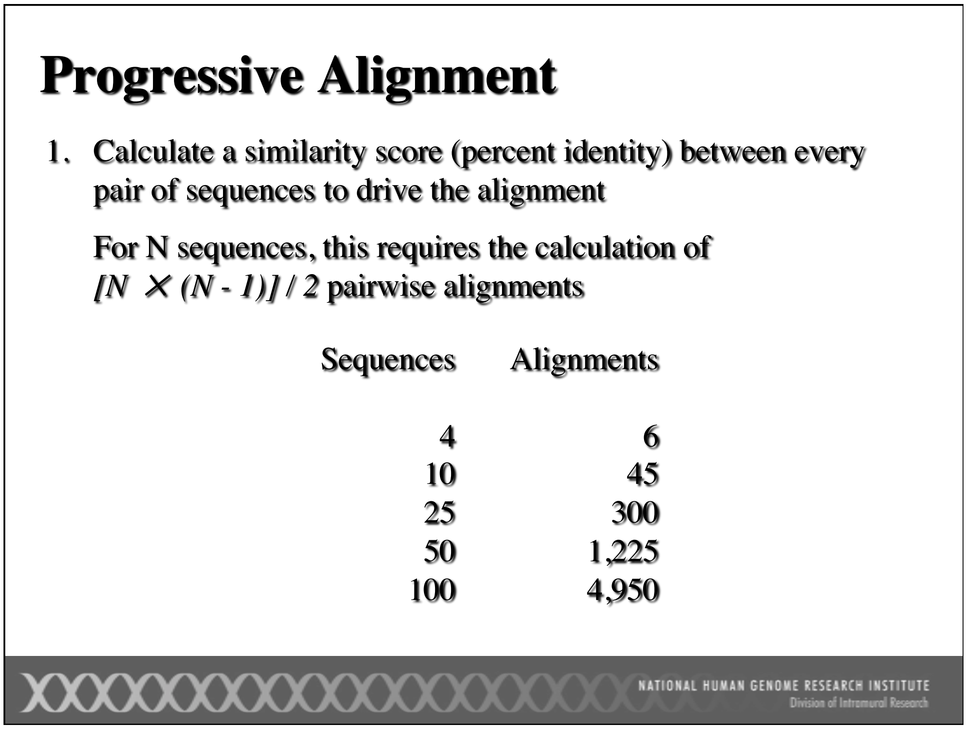
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Progressive Alignment

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



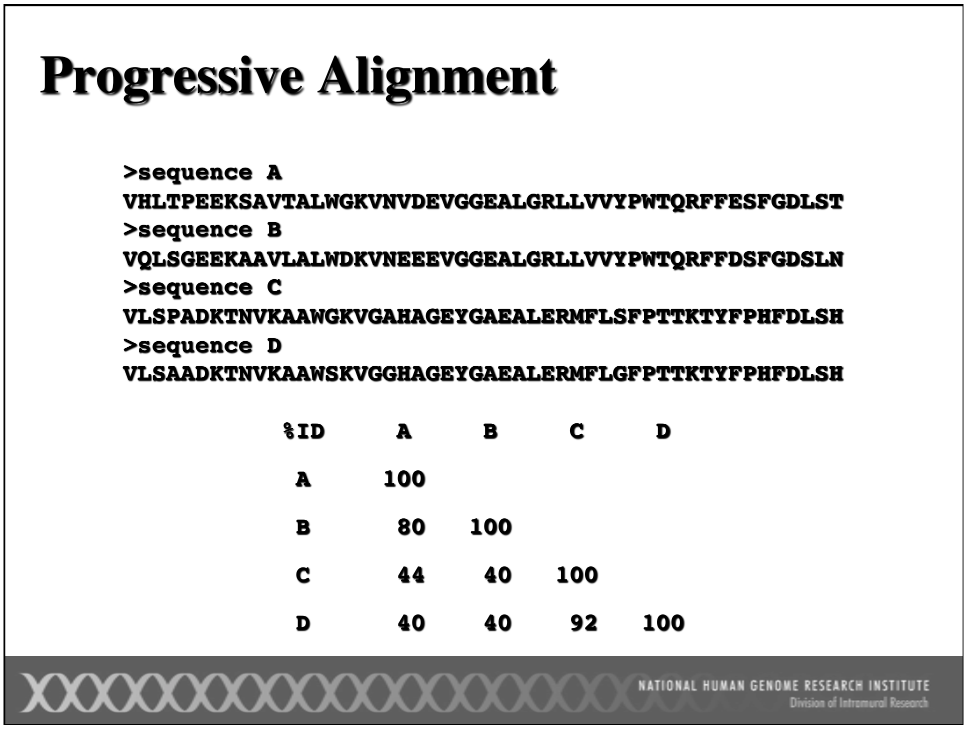
Progressive Alignment

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 × (N - 1)] / 2* pairwise alignments

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

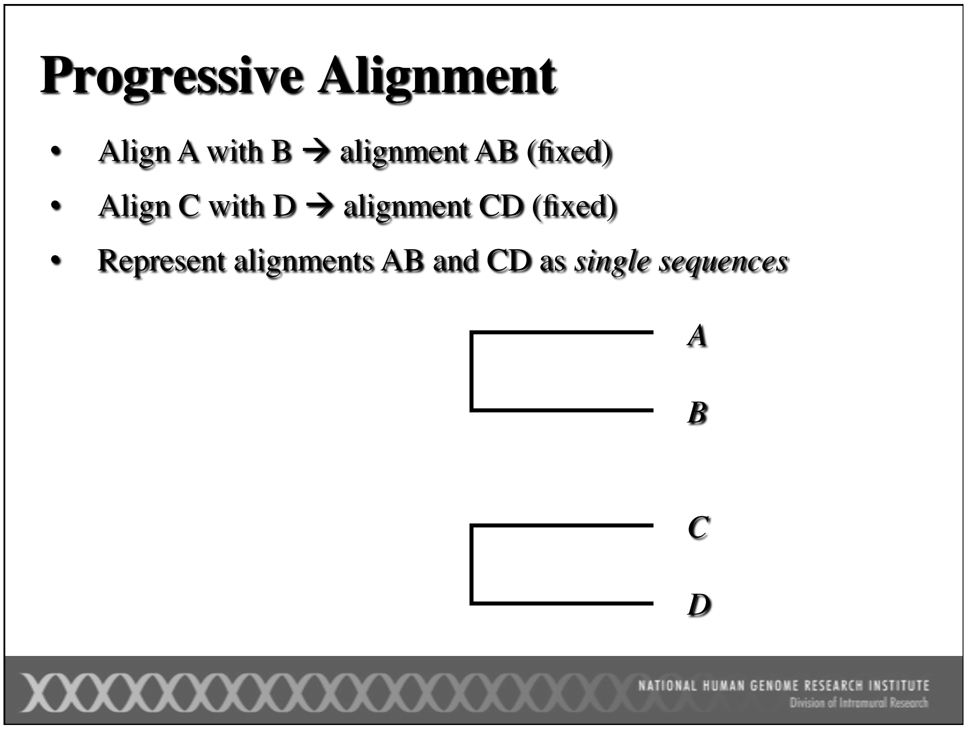


Progressive Alignment

```
>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

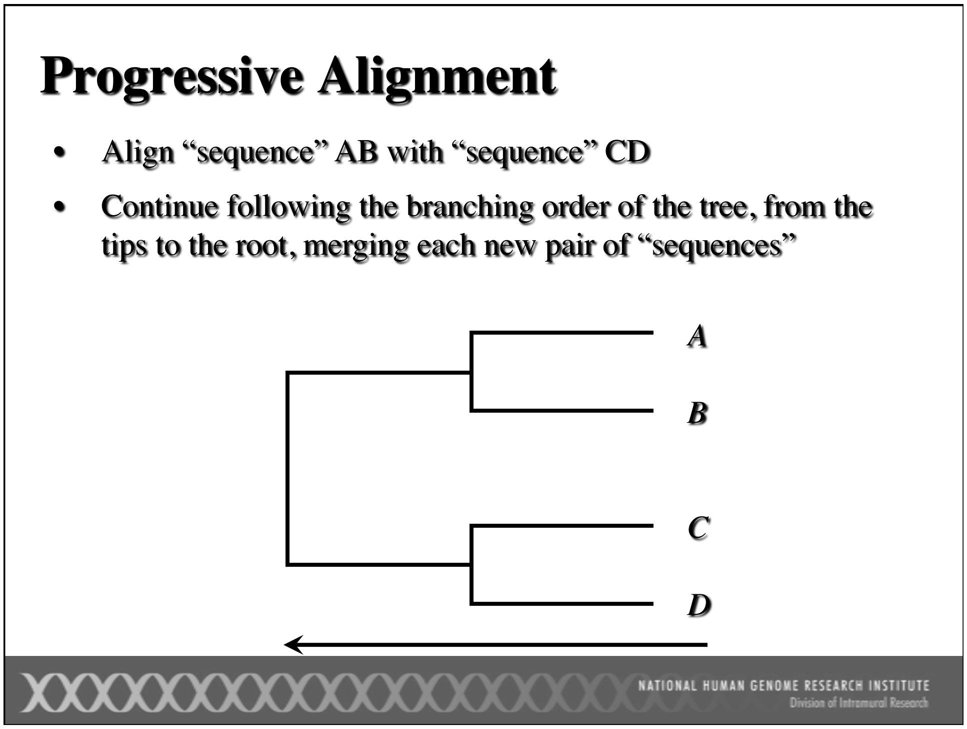
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Progressive Alignment

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

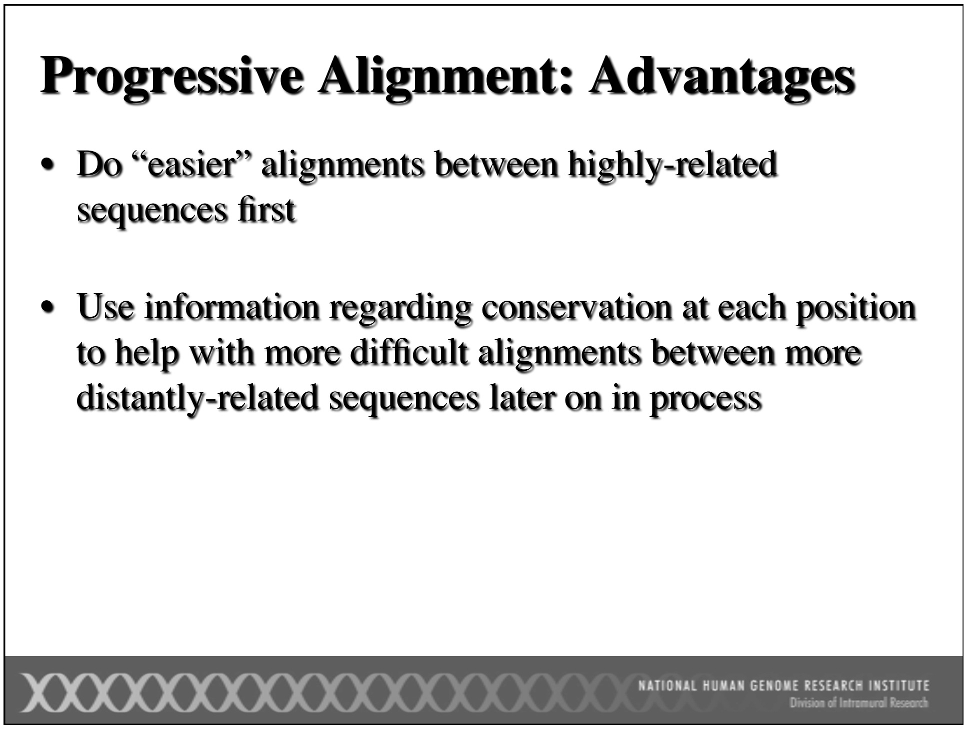
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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"

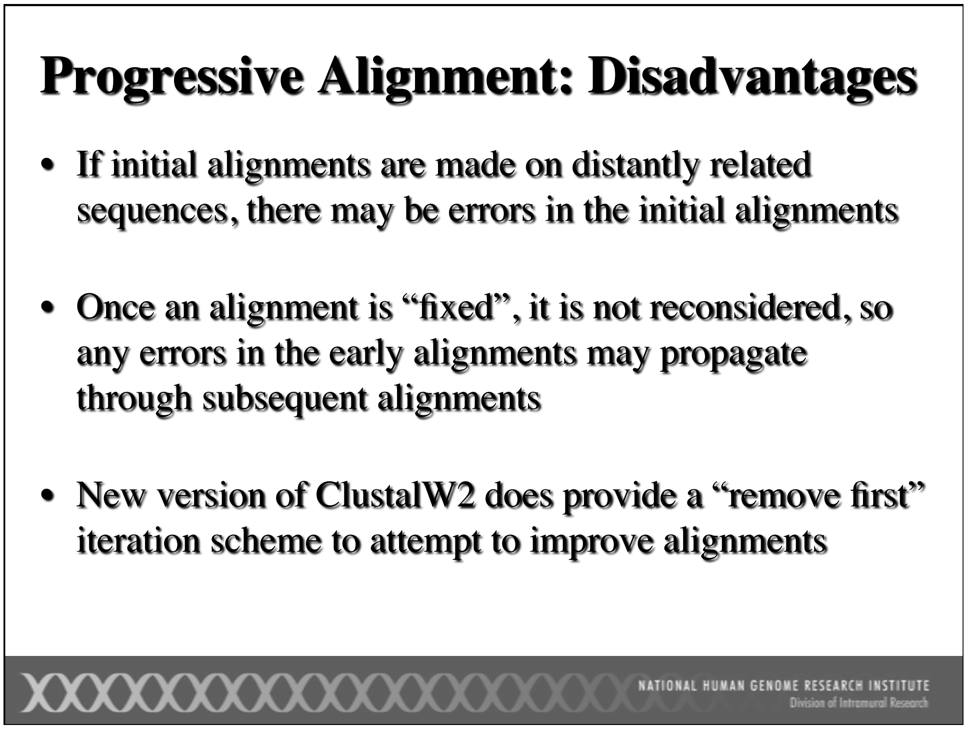
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

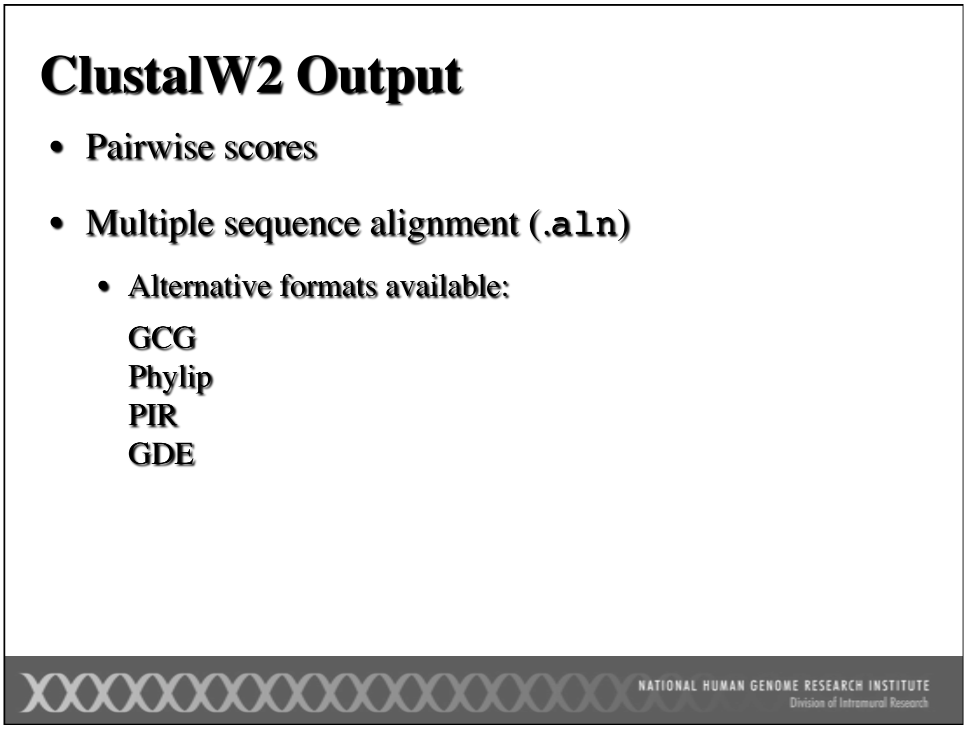
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

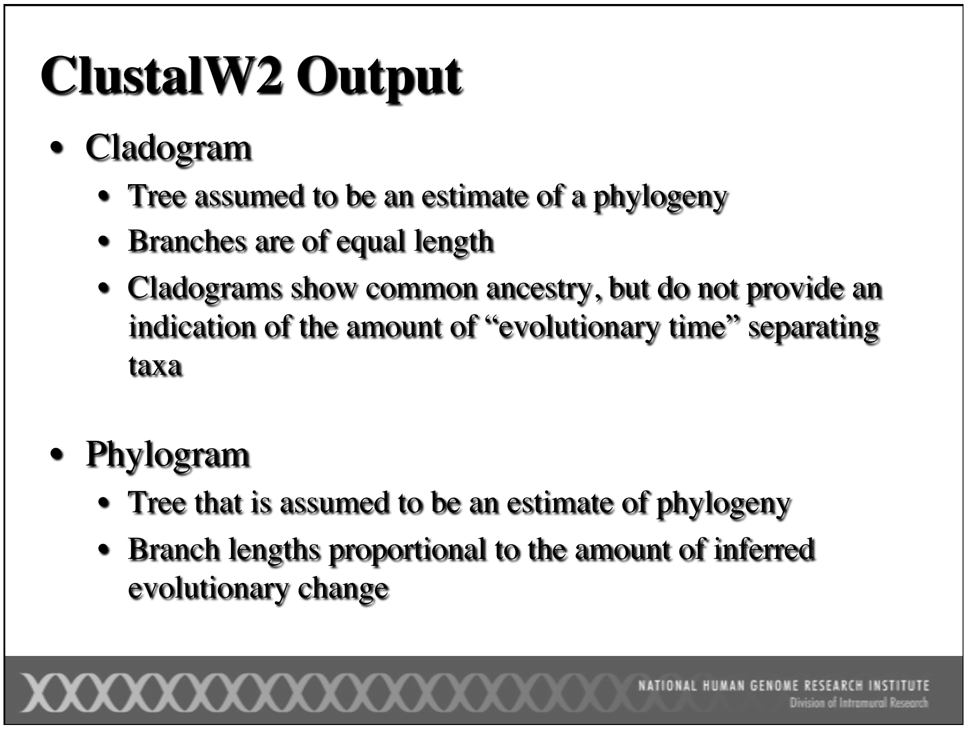


ClustalW2 Output

- Pairwise scores
- Multiple sequence alignment (`.aln`)
  - Alternative formats available:

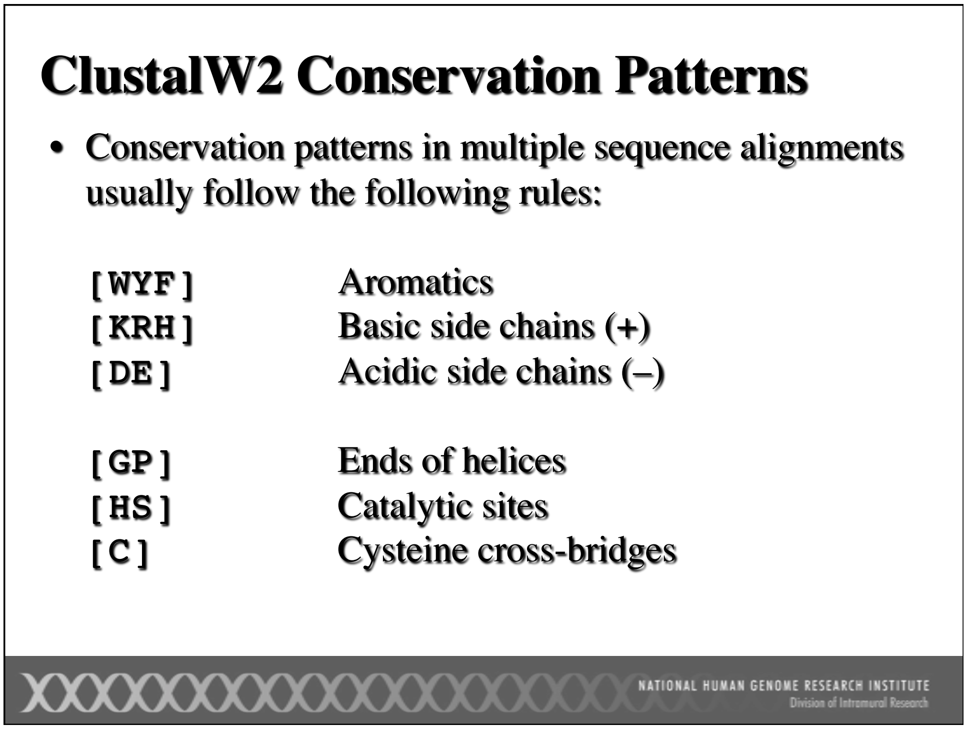
    GCG
    Phylip
    PIR
    GDE

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



ClustalW2 Output

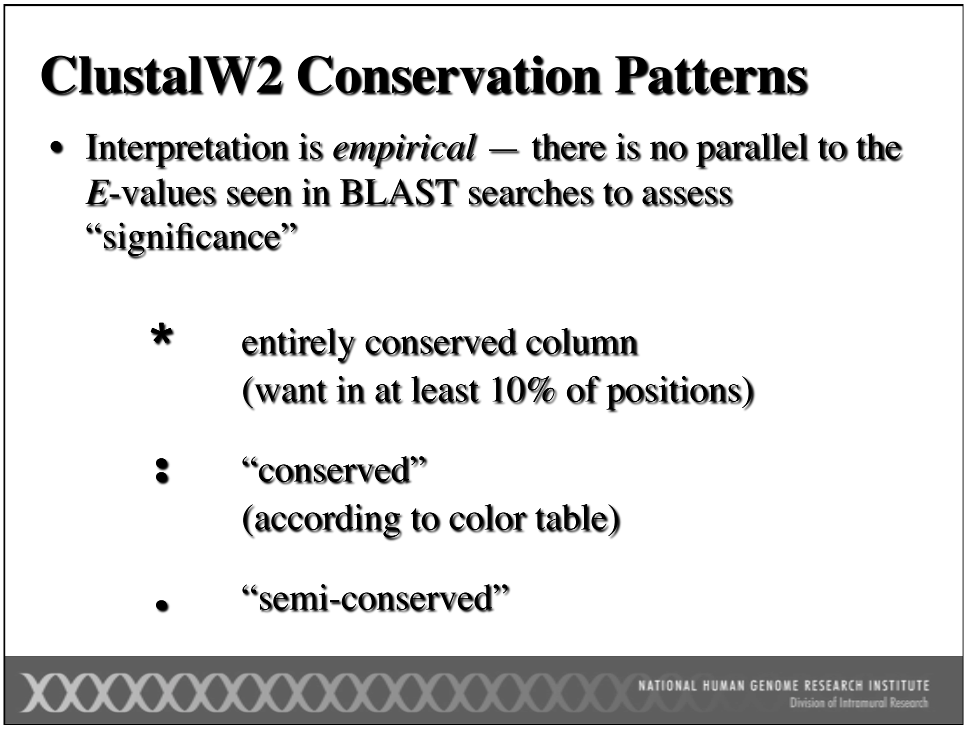
- Cladogram
  - Tree assumed to be an estimate of a phylogeny
  - Branches are of equal length
  - Cladograms show common ancestry, but do not provide an indication of the amount of “evolutionary time” separating taxa
- Phylogram
  - Tree that is assumed to be an estimate of phylogeny
  - Branch lengths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

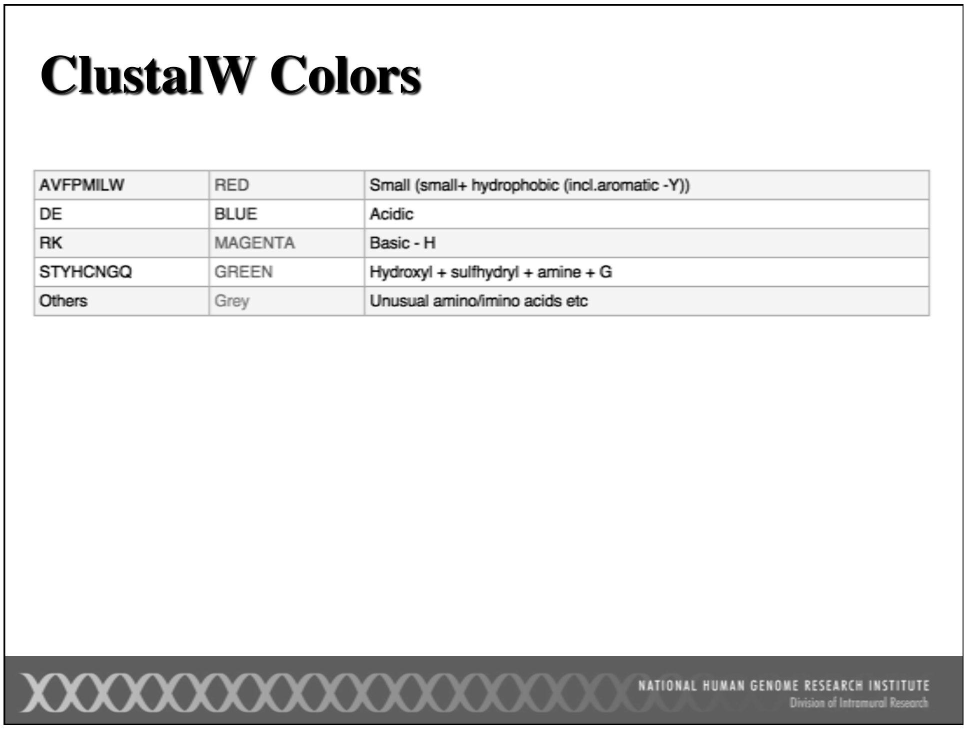


ClustalW2 Conservation Patterns

- 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"

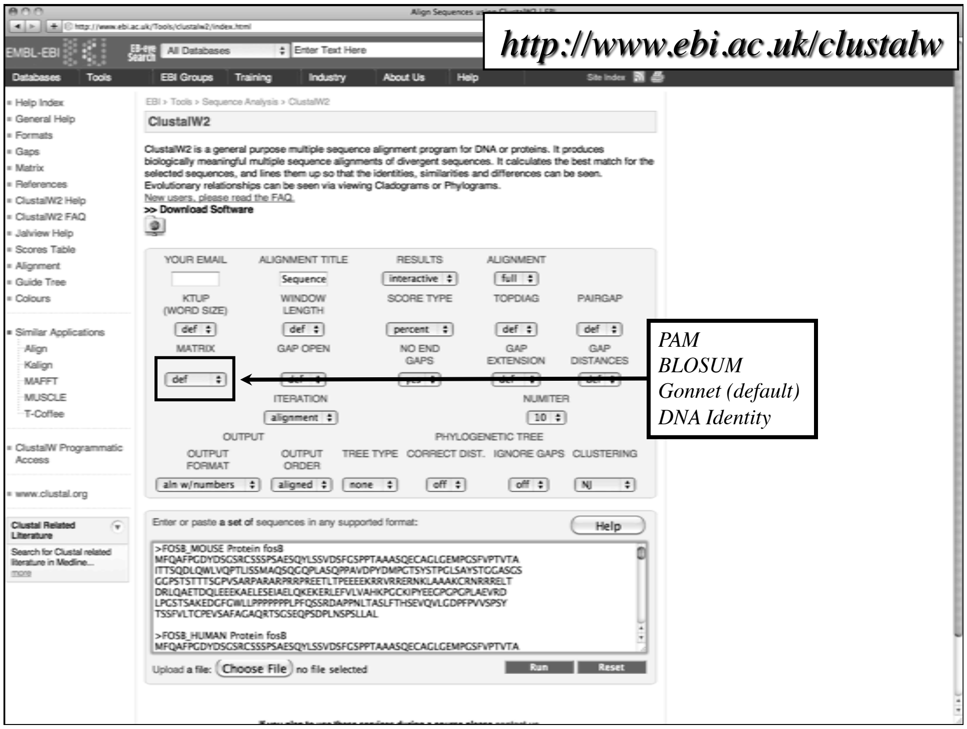
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



ClustalW Colors

AVFPMILW	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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

Align Sequences using ClustalW2 | EBI

http://www.ebi.ac.uk/Tools/clustalw2/index.html

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	RESULTS	ALIGNMENT	
	Sequence	interactive	full	
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	NO END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	yes	def	def
	ITERATION		NUMITER	
	alignment		10	

OUTPUT / PHYLOGENETIC TREE

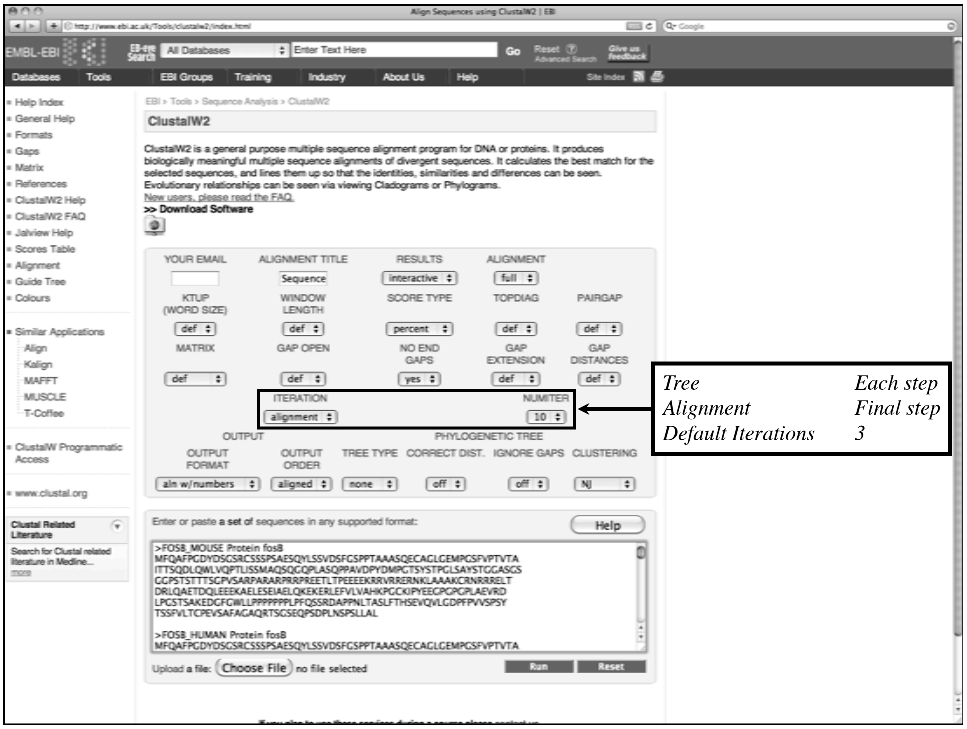
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS	CLUSTERING
aln w/numbers	aligned	none	off	off	NJ

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

```
>FOSB_MOUSE Protein fosB
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL

>FOSB_HUMAN Protein fosB
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
```

Upload a file: no file selected

PAM
BLOSUM
Gonnet (default)
DNA Identity



Align Sequences using ClustalW2 | EBI

http://www.ebi.ac.uk/Tools/clustalw2/index.html

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	RESULTS	ALIGNMENT	
	Sequence	interactive	full	
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	NO END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	yes	def	def
	ITERATION		NUMITER	
	alignment		10	

OUTPUT / PHYLOGENETIC TREE

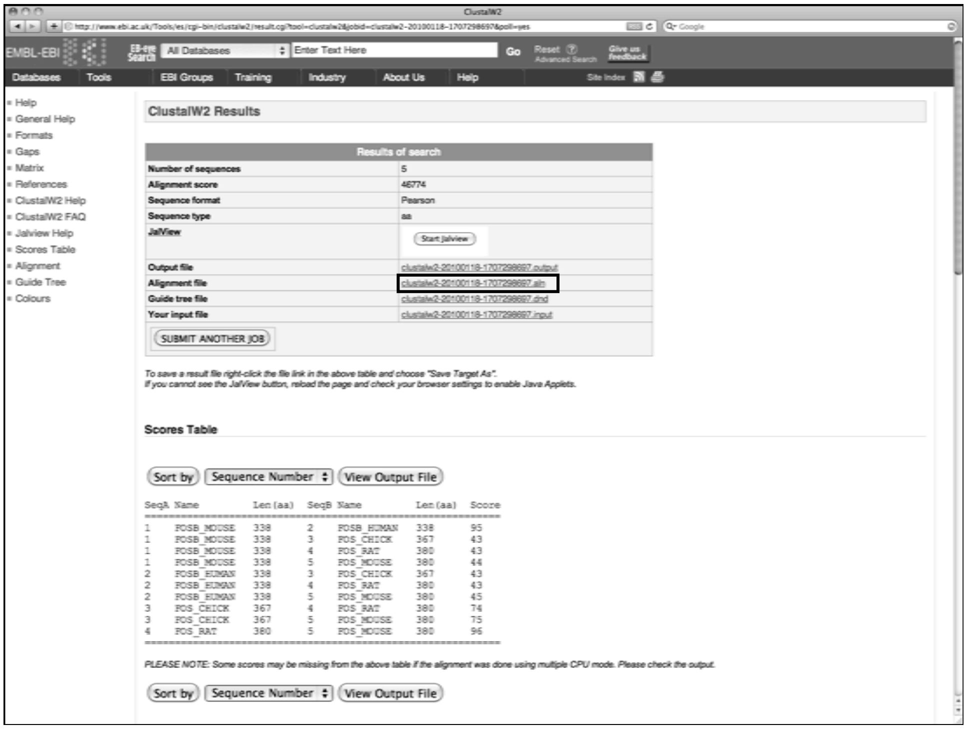
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS	CLUSTERING
aln w/numbers	aligned	none	off	off	NJ

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

```
>FOSB_MOUSE Protein fosB
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL

>FOSB_HUMAN Protein fosB
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
```

Upload a file: no file selected

Tree	*Each step*
Alignment	*Final step*
Default Iterations	*3*



ClustalW2 Results

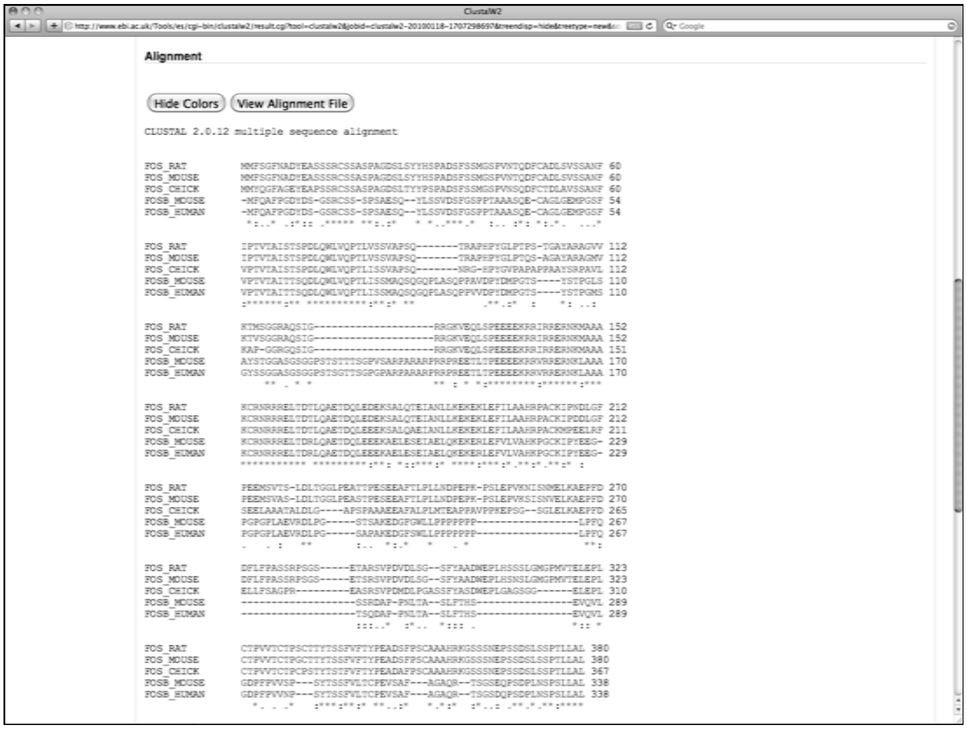
Results of search	
Number of sequences	5
Alignment score	46774
Sequence format	Pearson
Sequence type	aa
JalView	Start Jalview
Output file	clustalw2-20100118-1707298697.output
Alignment file	clustalw2-20100118-1707298697.aln
Guide tree file	clustalw2-20100118-1707298697.dnd
Your input file	clustalw2-20100118-1707298697.input

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

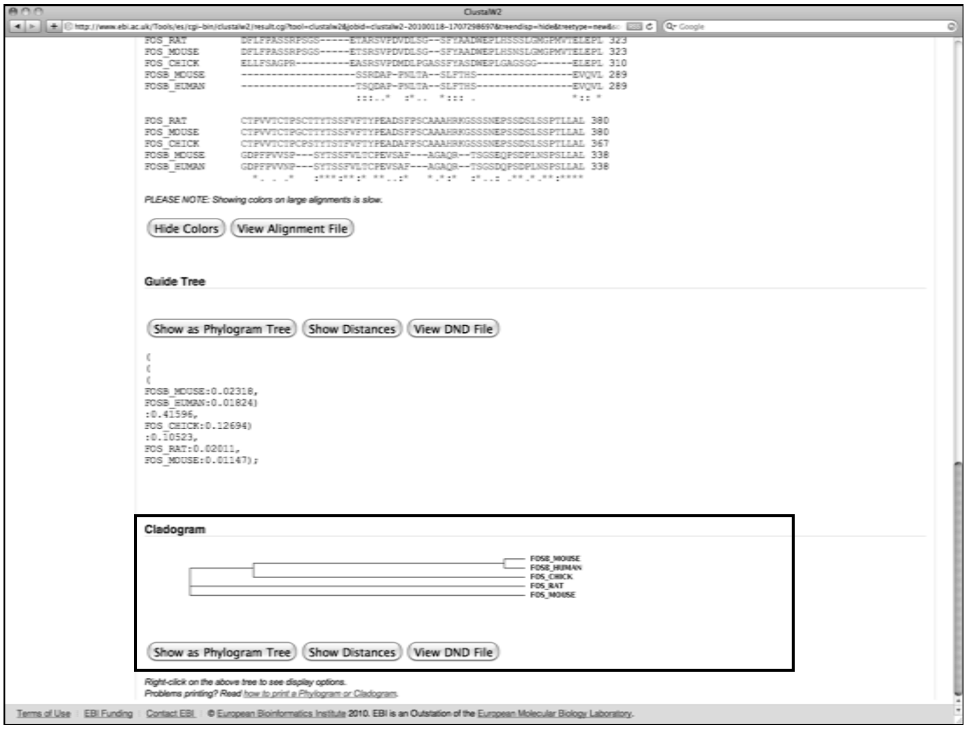
```
SeqA Name         Len(aa)  SeqB Name         Len(aa)  Score
===========================================================
1    FOSB_MOUSE   338      2    FOSB_HUMAN   338      95
1    FOSB_MOUSE   338      3    FOS_CHICK    367      43
1    FOSB_MOUSE   338      4    FOS_RAT      380      43
1    FOSB_MOUSE   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.


ClustalW2
http://www.ebi.ac.uk/Tools/es/cgi-bin/clustalw2/result.cgi?tool=clustalw2&jobid=clustalw2-20100118-1707298697&treendisp=hide&treetype=new&s
Hide Colors
View Alignment File

Alignment

```
CLUSTAL 2.0.12 multiple sequence alignment


FOS_RAT         MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_MOUSE       MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_CHICK       MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANF 60
FOSB_MOUSE      -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
FOSB_HUMAN      -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSF 54
                 *:..* .:*:: .***** **:.:*   * *..***.*   :.. :*: *:.*.  ...*

FOS_RAT         IPTVTAISTSPDLQWLVQPTLVSSVAPSQ-------TRAPHPYGLPTPS-TGAYARAGVV 112
FOS_MOUSE       IPTVTAISTSPDLQWLVQPTLVSSVAPSQ-------TRAPHPYGLPTQS-AGAYARAGMV 112
FOS_CHICK       VPTVTAISTSPDLQWLVQPTLISSVAPSQ-------NRG-HPYGVPAPAPPAAYSRPAVL 112
FOSB_MOUSE      VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTS----YSTPGLS 110
FOSB_HUMAN      VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTS----YSTPGMS 110
                :******:** **********:**:* **              .**.:*  :    *: ..:

FOS_RAT         KTMSGGRAQSIG--------------------RRGKVEQLSPEEEEKRRIRRERNKMAAA 152
FOS_MOUSE       KTVSGGRAQSIG--------------------RRGKVEQLSPEEEEKRRIRRERNKMAAA 152
FOS_CHICK       KAP-GGRGQSIG--------------------RRGKVEQLSPEEEEKRRIRRERNKMAAA 151
FOSB_MOUSE      AYSTGGASGSGGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAA 170
FOSB_HUMAN      GYSSGGASGSGGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAA 170
                    ** . * *                    ** : * *:********:******:***

FOS_RAT         KCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPNDLGF 212
FOS_MOUSE       KCRNRRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGF 212
FOS_CHICK       KCRNRRRELTDTLQAETDQLEEEKSALQAEIANLLKEKEKLEFILAAHRPACKMPEELRF 211
FOSB_MOUSE      KCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEG- 229
FOSB_HUMAN      KCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEG- 229
                *********** *********:**: *::***:* ****:***:*.**:*.**:* :

FOS_RAT         PEEMSVTS-LDLTGGLPEATTPESEEAFTLPLLNDPEPK-PSLEPVKNISNMELKAEPFD 270
FOS_MOUSE       PEEMSVAS-LDLTGGLPEASTPESEEAFTLPLLNDPEPK-PSLEPVKSISNVELKAEPFD 270
FOS_CHICK       SEELAAATALDLG----APSPAAAEEAFALPLMTEAPPAVPPKEPSG--SGLELKAEPFD 265
FOSB_MOUSE      PGPGPLAEVRDLPG-----STSAKEDGFGWLLPPPPPPP-----------------LPFQ 267
FOSB_HUMAN      PGPGPLAEVRDLPG-----SAPAKEDGFSWLLPPPPPPP-----------------LPFQ 267
                .   . :  **       :..  *:.*   *    .*                 **:

FOS_RAT         DFLFPASSRPSGS-----ETARSVPDVDLSG--SFYAADWEPLHSSSLGMGPMVTELEPL 323
FOS_MOUSE       DFLFPASSRPSGS-----ETSRSVPDVDLSG--SFYAADWEPLHSNSLGMGPMVTELEPL 323
FOS_CHICK       ELLFSAGPR---------EASRSVPDMDLPGASSFYASDWEPLGAGSGG------ELEPL 310
FOSB_MOUSE      ------------------SSRDAP-PNLTA--SLFTHS-----------------EVQVL 289
FOSB_HUMAN      ------------------TSQDAP-PNLTA--SLFTHS-----------------EVQVL 289
                                  :::..*  :*..  *::: .               *:: *

FOS_RAT         CTPVVTCTPSCTTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 380
FOS_MOUSE       CTPVVTCTPGCTTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 380
FOS_CHICK       CTPVVTCTPCPSTYTSTFVFTYPEADAFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 367
FOSB_MOUSE      GDPFFPVVSP---SYTSSFVLTCPEVSAF---AGAQR--TSGSEQPSDPLNSPSLLAL 338
FOSB_HUMAN      GDPFFPVVNP---SYTSSFVLTCPEVSAF---AGAQR--TSGSDQPSDPLNSPSLLAL 338
                 *. . .*   :***:**:* **..:*   *.*:*  :*..: .**.*.**:****
```


ClustalW2
http://www.ebi.ac.uk/Tools/es/cgi-bin/clustalw2/result.cgi?tool=clustalw2&jobid=clustalw2-20100118-1707298697&treendisp=hide&treetype=new&...

```
FOS_RAT        DFLFPASSRPSGS-----ETARSVPDVDLSG--SFYAADWEPLHSSSLGMGPMVTELEPL 323
FOS_MOUSE      DFLFPASSRPSGS-----ETSRSVPDVDLSG--SFYAADWEPLHSNSLGMGPMVTELEPL 323
FOS_CHICK      ELLFSAGPR---------EASRSVPDMDLPGASSFYASDWEPLGAGSGG------ELEPL 310
FOSB_MOUSE     -------------------SSRDAP-PNLTA--SLFTHS----------------EVQVL 289
FOSB_HUMAN     -------------------TSQDAP-PNLTA--SLFTHS----------------EVQVL 289
                                  :::..*  :*..  *::: .                *:: *

FOS_RAT        CTPVVTCTPSCTTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 380
FOS_MOUSE      CTPVVTCTPGCTTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 380
FOS_CHICK      CTPVVTCTPCPSTYTSTFVFTYPEADAFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 367
FOSB_MOUSE     GDPFPVVSP---SYTSSFVLTCPEVSAF---AGAQR--TSGSEQPSDPLNSPSLLAL 338
FOSB_HUMAN     GDPFPVVNP---SYTSSFVLTCPEVSAF---AGAQR--TSGSDQPSDPLNSPSLLAL 338
                 *. . .*   :***:**:* **..:*   *.*:*  :*..: .**.*.**:****
```

PLEASE NOTE: Showing colors on large alignments is slow.

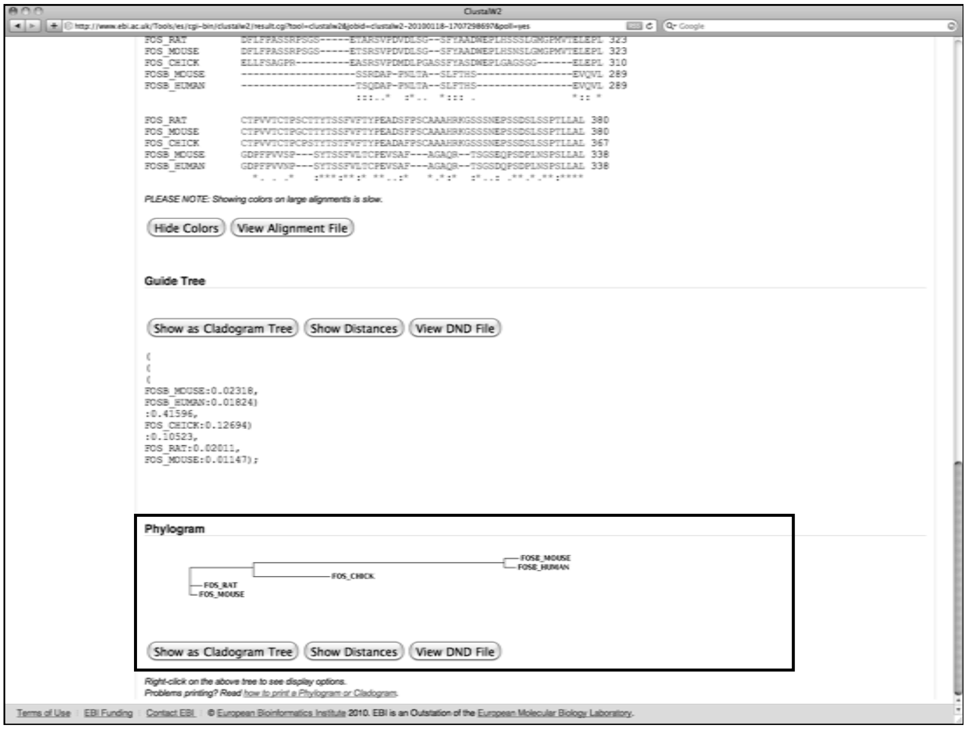
Guide Tree

```
(
(
(
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

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 2010. EBI is an Outstation of the European Molecular Biology Laboratory.


ClustalW2
http://www.ebi.ac.uk/Tools/es/cgi-bin/clustalw2/result.cgi?tool=clustalw2&jobid=clustalw2-20100118-1707298697&poll=yes

```
FOS_RAT        DFLFPASSRPSGS-----ETARSVPDVDLSG--SFYAADWEPLHSSSLGMGPMVTELEPL 323
FOS_MOUSE      DFLFPASSRPSGS-----ETSRSVPDVDLSG--SFYAADWEPLHSNSLGMGPMVTELEPL 323
FOS_CHICK      ELLFSAGPR---------EASRSVPDMDLPGASSFYASDWEPLGAGSGG------ELEPL 310
FOSB_MOUSE     -------------------SSRDAP-PNLTA--SLFTHS----------------EVQVL 289
FOSB_HUMAN     -------------------TSQDAP-PNLTA--SLFTHS----------------EVQVL 289
                                  :::..*  :*..  *::: .                *:: *

FOS_RAT        CTPVVTCTPSCTTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 380
FOS_MOUSE      CTPVVTCTPGCTTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 380
FOS_CHICK      CTPVVTCTPCPSTYTSTFVFTYPEADAFPSCAAAHRKGSSSNEPSSDSLSSPTLLAL 367
FOSB_MOUSE     GDPFPVVSP---SYTSSFVLTCPEVSAF---AGAQR--TSGSEQPSDPLNSPSLLAL 338
FOSB_HUMAN     GDPFPVVNP---SYTSSFVLTCPEVSAF---AGAQR--TSGSDQPSDPLNSPSLLAL 338
                 *. . .*   :***:**:* **..:*   *.*:*  :*..: .**.*.**:****
```

PLEASE NOTE: Showing colors on large alignments is slow.

Guide Tree

```
(
(
(
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

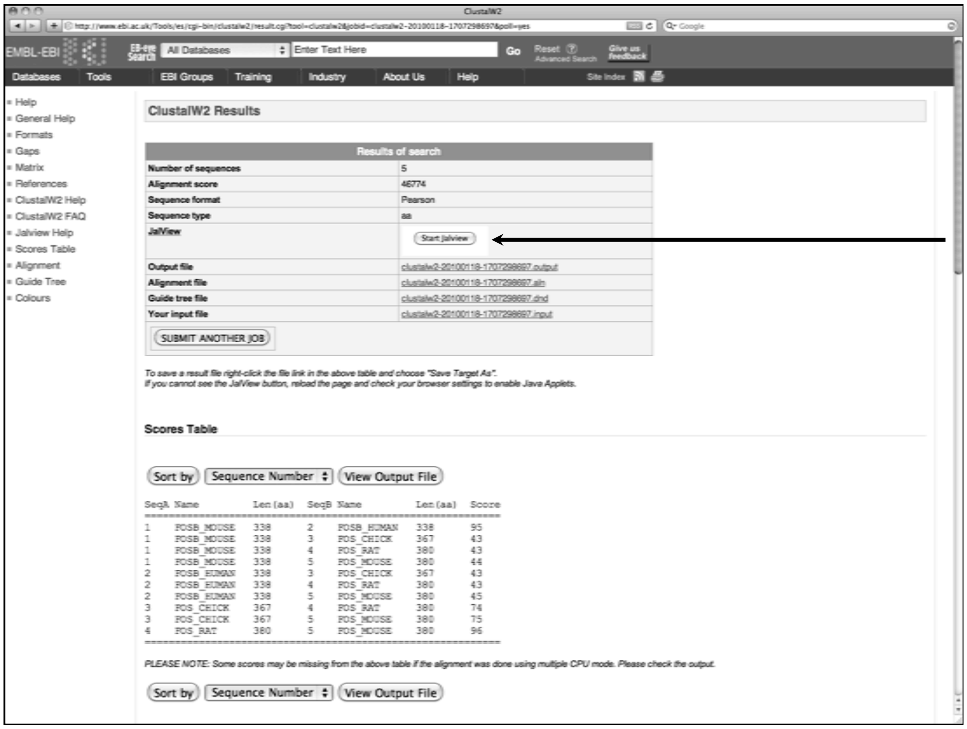
Right-click on the above tree to see display options.
Problems printing? Read how to print a Phylogram or Cladogram.

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2010. EBI is an Outstation of the European Molecular Biology Laboratory.

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



ClustalW2 Results

Results of search	
Number of sequences	5
Alignment score	46774
Sequence format	Pearson
Sequence type	aa
JalView	Start Jalview
Output file	clustalw2-20100118-1707298697.output
Alignment file	clustalw2-20100118-1707298697.aln
Guide tree file	clustalw2-20100118-1707298697.dnd
Your input file	clustalw2-20100118-1707298697.input

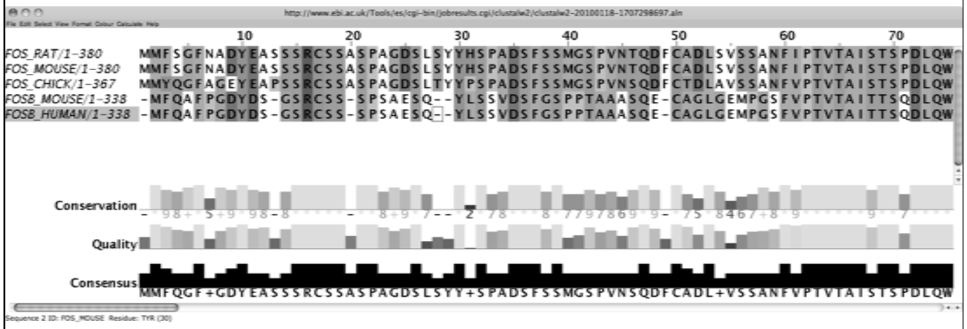
To save a result file right-click the file link in the above table and choose "Save Target As".
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

Scores Table

```
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.

Default view



http://www.ebi.ac.uk/Tools/es/cgi-bin/jobresults.cgi/clustalw2/clustalw2-20100118-1707298697.aln
File Edit Select View Format Colour Calculate Help

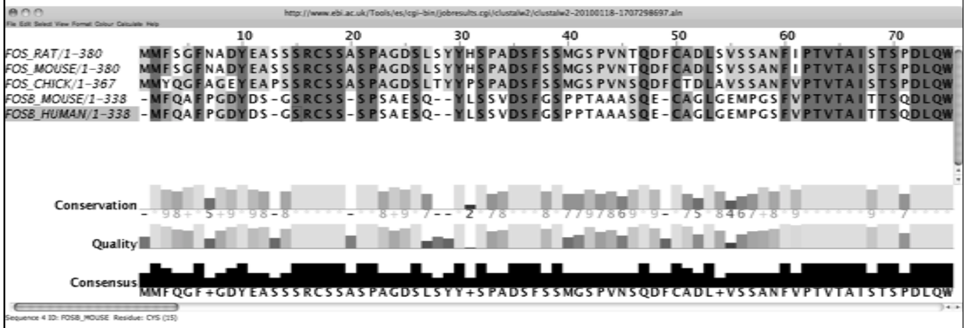
```
                     10        20        30        40        50        60        70
FOS_RAT/1-380     MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQW
FOS_MOUSE/1-380   MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQW
FOS_CHICK/1-367   MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANFVPTVTAISTSPDLQW
FOSB_MOUSE/1-338  -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSFVPTVTAITTSQDLQW
FOSB_HUMAN/1-338  -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSFVPTVTAITTSQDLQW

Conservation      -*98+**5+9*98-8*****-**8+9*7--*2*78***8*7797869*9-*75*8467+8*9******9*7****
Quality
Consensus         MMFQGF+GDYEASSSRCSSASPAGDSLSYY+SPADSFSSMGSPVNSQDFCADL+VSSANFVPTVTAISTSPDLQW

Sequence 2 ID: FOS_MOUSE  Residue: TYR (30)
```

Conservation Conservation of total alignment (indication of percent identity)

Quality Alignment quality, based on BLOSUM62 scores

Consensus Based on percent identity

Colour → Percentage Identity

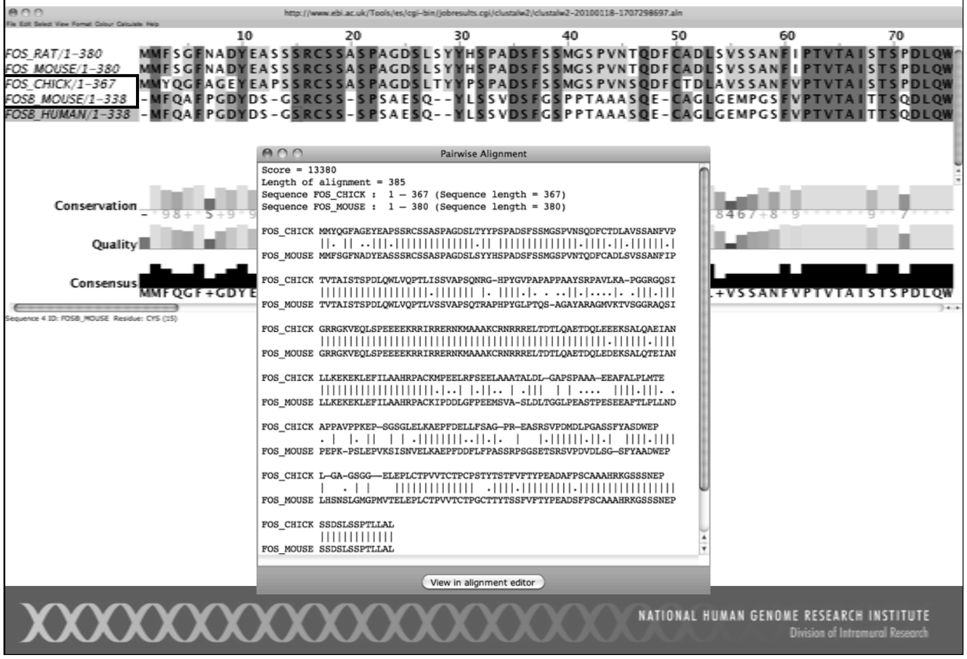


http://www.ebi.ac.uk/Tools/es/cgi-bin/jobresults.cgi/clustalw2/clustalw2-20100118-1707298697.aln
File Edit Select View Format Colour Calculate Help

```
                     10        20        30        40        50        60        70
FOS_RAT/1-380     MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQW
FOS_MOUSE/1-380   MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQW
FOS_CHICK/1-367   MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANFVPTVTAISTSPDLQW
FOSB_MOUSE/1-338  -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSFVPTVTAITTSQDLQW
FOSB_HUMAN/1-338  -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSFVPTVTAITTSQDLQW

Conservation      -*98+**5+9*98-8*****-**8+9*7--*2*78***8*7797869*9-*75*8467+8*9******9*7****
Quality
Consensus         MMFQGF+GDYEASSSRCSSASPAGDSLSYY+SPADSFSSMGSPVNSQDFCADL+VSSANFVPTVTAISTSPDLQW

Sequence 4 ID: FOSB_MOUSE  Residue: CYS (15)
```

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

Calculate → Pairwise Alignments

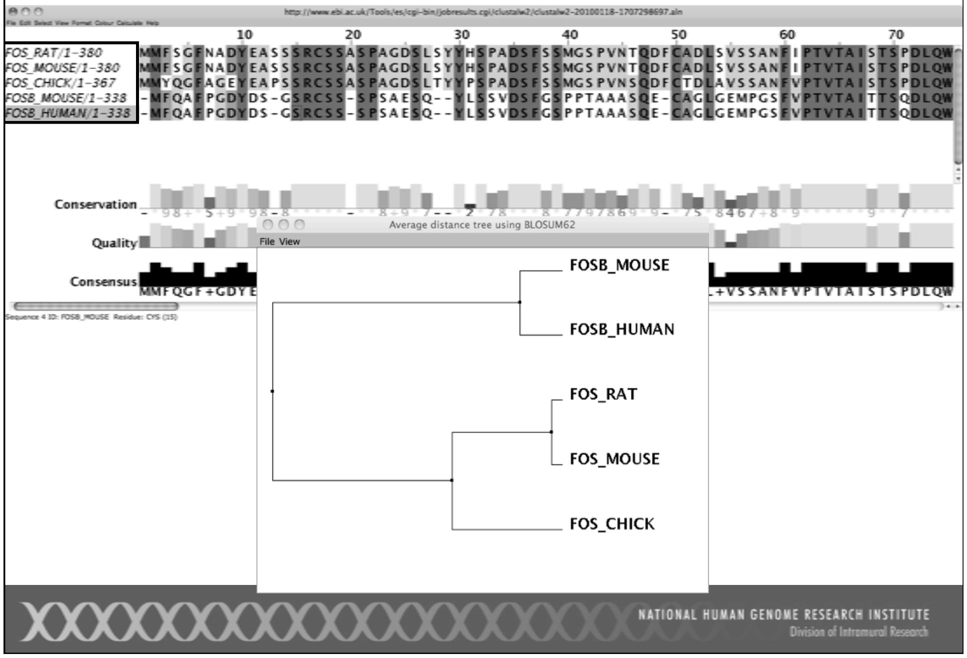

http://www.ebi.ac.uk/Tools/es/cgi-bin/jobresults.cgi/clustalw2/clustalw2-20100118-1707298697.aln
File Edit Select View Format Colour Calculate Help
10
20
30
40
50
60
70
Conservation
Quality
Consensus
Sequence 4 ID: FOSB_MOUSE Residue: CYS (15)
Pairwise Alignment
View in alignment editor

```
FOS_RAT/1-380     MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQW
FOS_MOUSE/1-380   MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQW
FOS_CHICK/1-367   MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANFVPTVTAISTSPDLQW
FOSB_MOUSE/1-338  -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSFVPTVTAITTSQDLQW
FOSB_HUMAN/1-338  -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSFVPTVTAITTSQDLQW
```

```
Score = 13380
Length of alignment = 385
Sequence FOS_CHICK :  1 — 367 (Sequence length = 367)
Sequence FOS_MOUSE :  1 — 380 (Sequence length = 380)

FOS_CHICK MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANFVP
          ||. || ..|||.||||||||||||||.|| ||||||||||||||.||||.||.||||||.|
FOS_MOUSE MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIP

FOS_CHICK TVTAISTSPDLQWLVQPTLISSVAPSQNRG-HPYGVPAPAPPAAYSRPAVLKA-PGGRGQSI
          |||||||||||||||||||.||||||| |. ||||.|. . ..||.|....|. .|||.|||
FOS_MOUSE TVTAISTSPDLQWLVQPTLVSSVAPSQTRAPHPYGLPTQS-AGAYARAGMVKTVSGGRAQSI

FOS_CHICK GRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEEEKSALQAEIAN
          |||||||||||||||||||||||||||||||||||||||||||||||||||.|||||.||||
FOS_MOUSE GRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDEKSALQTEIAN

FOS_CHICK LLKEKEKLEFILAAHRPACKMPEELRFSEELAAATALDL—GAPSPAAA—EEAFALPLMTE
          ||||||||||||||||||||.|..| |.||.. | .|||  | | ....  ||||.|||. .
FOS_MOUSE LLKEKEKLEFILAAHRPACKIPDDLGFPEEMSVA-SLDLTGGLPEASTPESEEAFTLPLLND

FOS_CHICK APPAVPPKEP—SGSGLELKAEPFDELLFSAG—PR—EASRSVPDMDLPGASSFYASDWEP
          . |  |. ||  | | .||||||||..||.|.  |   |.||||||.||.|  ||||.||||
FOS_MOUSE PEPK-PSLEPVKSISNVELKAEPFDDFLFPASSRPSGSETSRSVPDVDLSG—SFYAADWEP

FOS_CHICK L—GA-GSGG——ELEPLCTPVVTCTPCPSTYTSTFVFTYPEADAFPSCAAAHRKGSSSNEP
          |   . | |    ||||||||||||||  .||||.||||||||||.|||||||||||||||||
FOS_MOUSE LHSNSLGMGPMVTELEPLCTPVVTCTPGCTTYTSSFVFTYPEADSFPSCAAAHRKGSSSNEP

FOS_CHICK SSDSLSSPTLLAL
          |||||||||||||
FOS_MOUSE SSDSLSSPTLLAL
```

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Calculate → Calculate Tree → Neighbour Joining Using BLOSUM62


http://www.ebi.ac.uk/Tools/es/cgi-bin/jobresults.cgi/clustalw2/clustalw2-20100118-1707298697.aln
File Edit Select View Format Colour Calculate Help
10
20
30
40
50
60
70
Conservation
Quality
Consensus
Sequence 4 ID: FOSB_MOUSE Residue: CYS (15)
Average distance tree using BLOSUM62
File View
FOSB_MOUSE
FOSB_HUMAN
FOS_RAT
FOS_MOUSE
FOS_CHICK

```
FOS_RAT/1-380     MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQW
FOS_MOUSE/1-380   MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQW
FOS_CHICK/1-367   MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANFVPTVTAISTSPDLQW
FOSB_MOUSE/1-338  -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSFVPTVTAITTSQDLQW
FOSB_HUMAN/1-338  -MFQAFPGDYDS-GSRCSS-SPSAESQ--YLSSVDSFGSPPTAAASQE-CAGLGEMPGSFVPTVTAITTSQDLQW
```

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

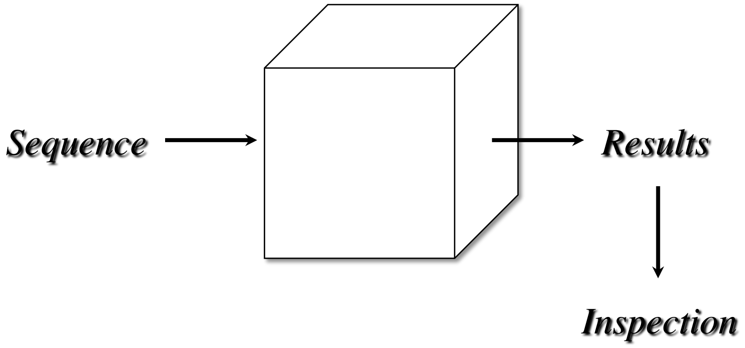
Further Reading

Current Protocols in Bioinformatics
Unit 2.3
ClustalW

Current Protocols in Bioinformatics
Unit 3.8
T-Coffee

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Understanding Analyses



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research