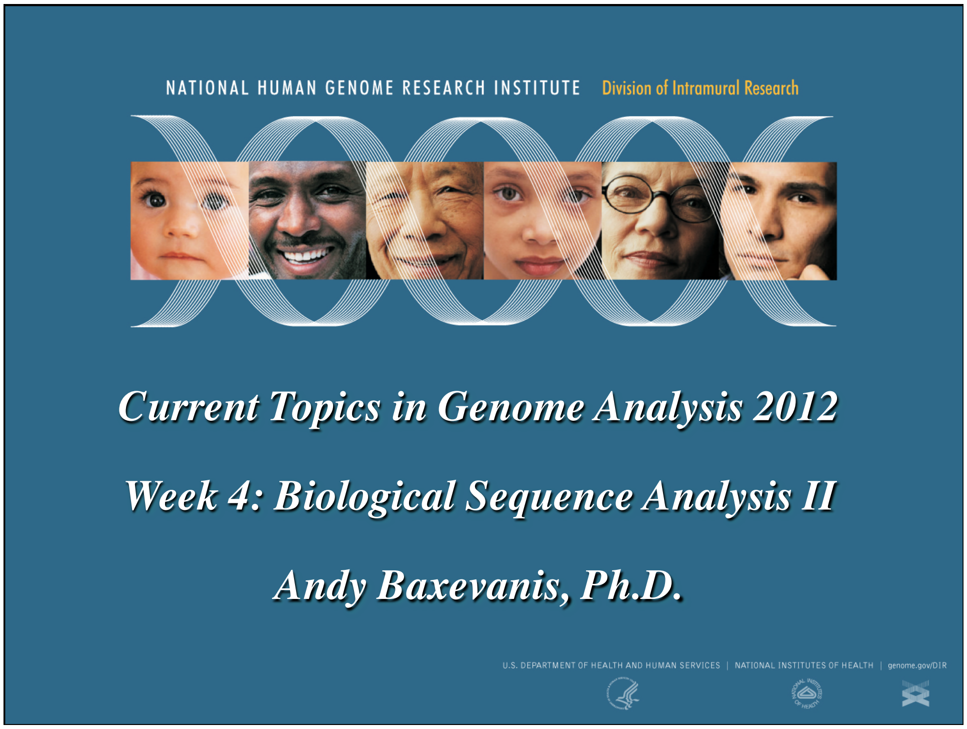


NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

Current Topics in Genome Analysis 2012

Week 4: Biological Sequence Analysis II

Andy Baxevanis, Ph.D.

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

JOHNS HOPKINS MEDICINE
CONTINUING MEDICAL EDUCATION

Current Topics in Genome Analysis 2012

Andy Baxevanis, Ph.D.

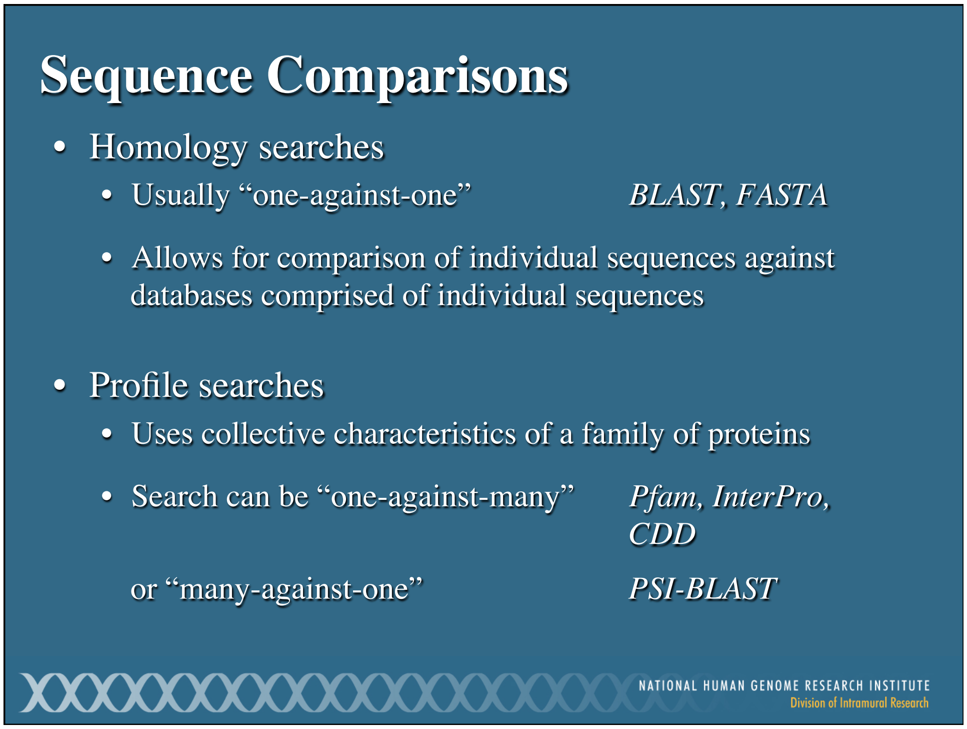
No Relevant Financial Relationships with Commercial Interests

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Overview

- Week 2
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 4
  - Profiles, Patterns, Motifs, and Domains
  - Structures: VAST, Cn3D, and *de novo* Prediction
  - Multiple Sequence Alignment

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

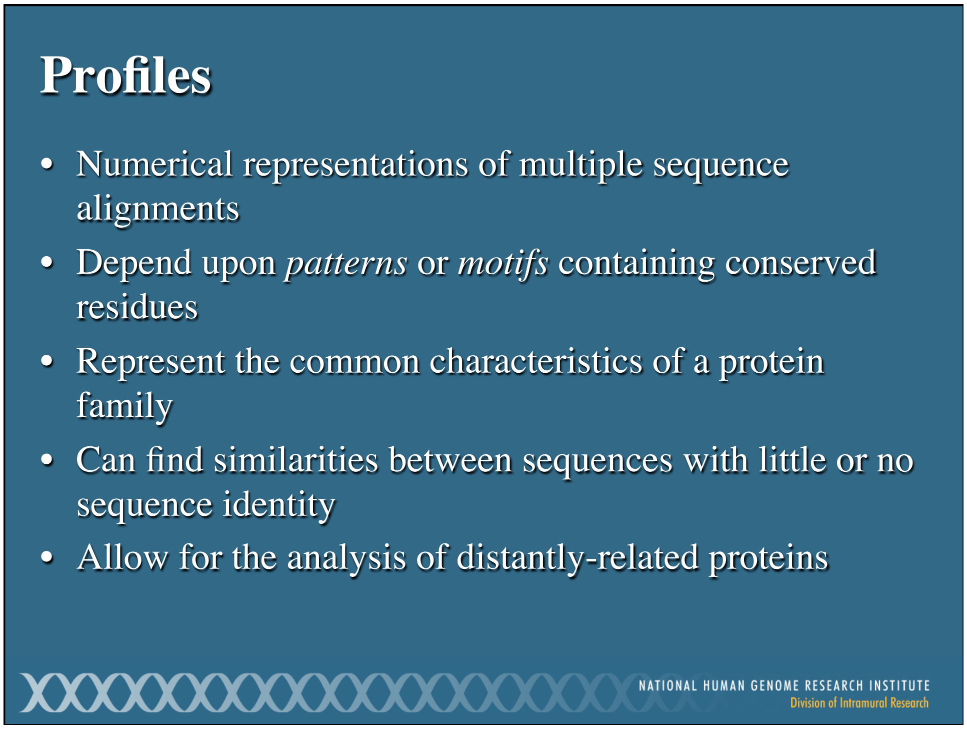


Sequence Comparisons

- 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*

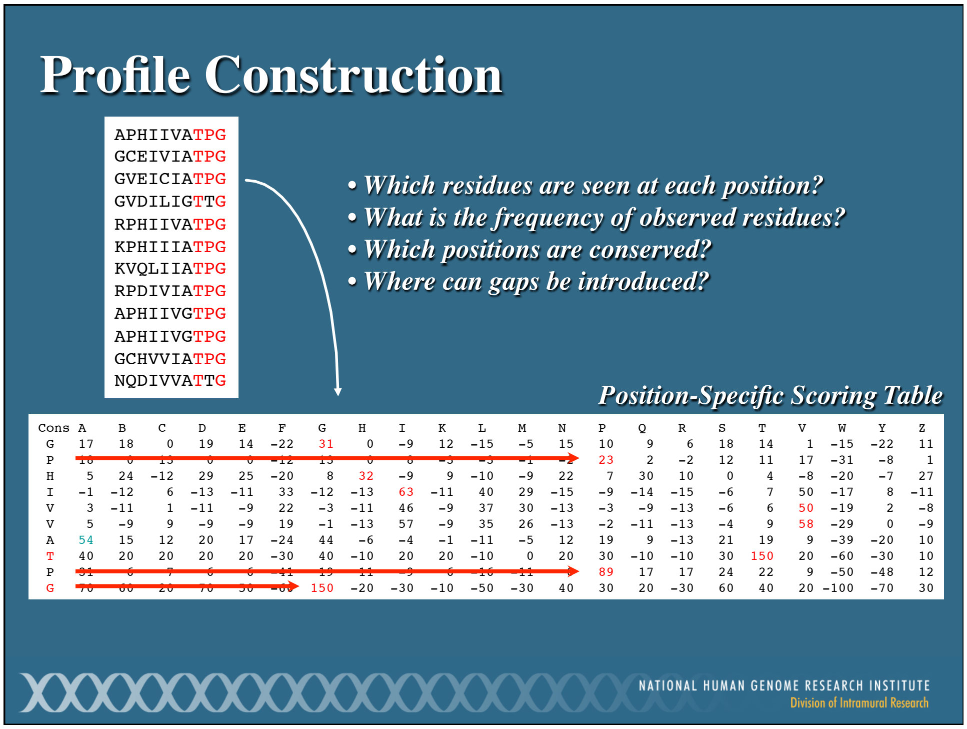
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

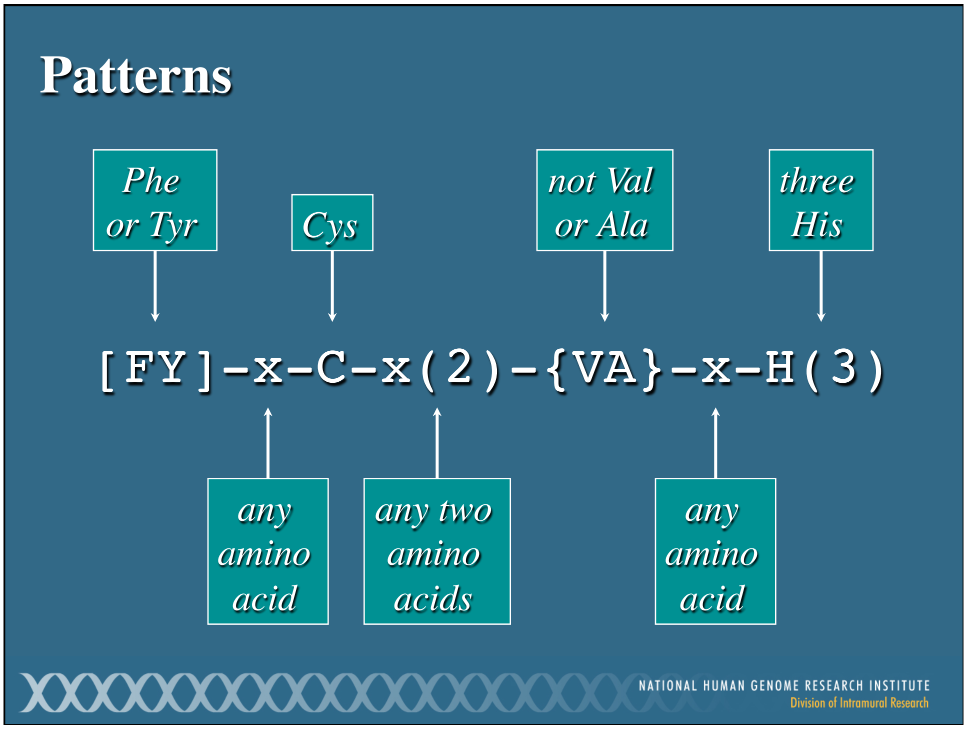
```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
```

- ***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~~	~~30~~	~~-60~~	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

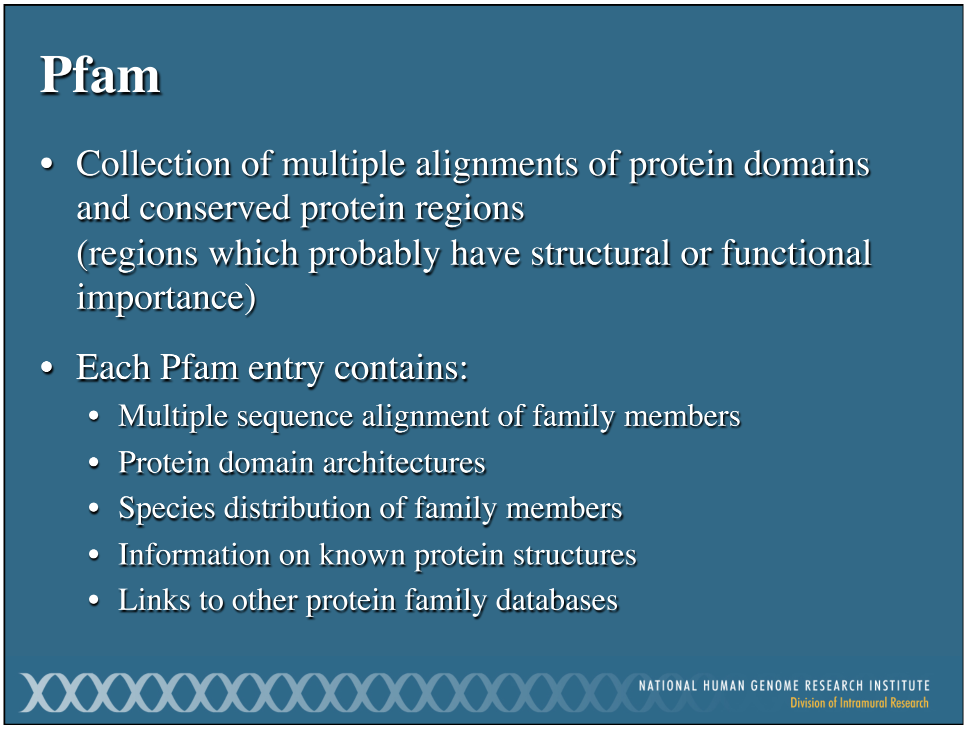
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Patterns

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

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

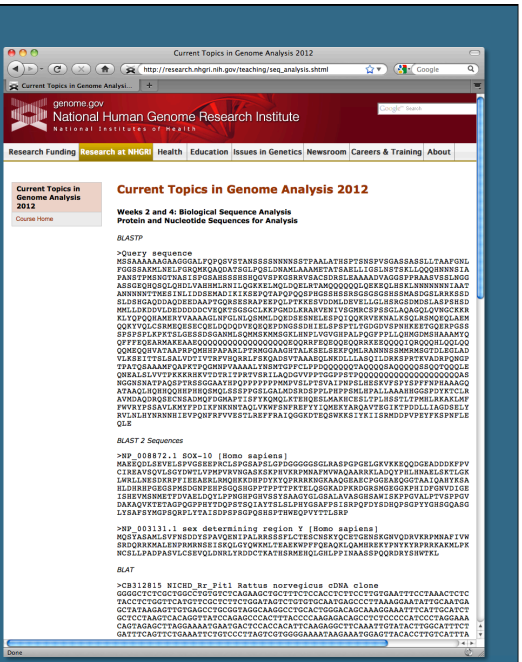
Pfam

- 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 2012
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
Google Search
Research Funding
Research at NHGRI
Health
Education
Issues in Genetics
Newsroom
Careers & Training
About
Current Topics in Genome Analysis 2012
Course Home
Done

Current Topics in Genome Analysis 2012

Weeks 2 and 4: Biological Sequence Analysis
Protein and Nucleotide Sequences for Analysis

BLASTP

```
>Query sequence
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAMETATSAELLIGSLNSTSKLLQQQHNNNSIA
PANSTPMSNGTNASISPGSAHSSSHSHQGVSPKGSRRVSACSDRSLEAAAADVAGGSPPRAASVSSLNGG
ASSGEQHQSQLQHDLVAHHMLRNILQGKKELMQLDQELRTAMQQQQQQLQEKEQLHSKLNNNNNNNIAAT
ANNNNNTTMESINLIDDSEMADIKIKSEPQTAPQPQQSPHGSSHSSRSGSGSGSHSSMASDGSLRRKSSD
SLDSHGAQDDAQDEEDAAPTGQRSESRAPEEPQLPTKKESVDDMLDEVELLGLHSRGSDMDSLASPSHSD
MMLLDKDDVLDEDDDDDCVEQKTSGSGCLKKPGMDLKRARVENIVSGMRCSPSSGLAQAGQLQVNGCKKR
KLYQPQQHAMERYVAAAAGLNFGLNLQSMMLDQEDSESNELESPQIQQKRVEKNALKSQLRSMQEQLAEM
QQKYVQLCSRMEQESECQELDQDQDVEQEQEPDNGSSDHIELSPSPTLTGDGDVSPNHKEETGQERPGSS
SPSPSPLKPKTSLGESSDSGANMLSQMMSKMMSGKLHNPLVGVGHPALPQGFPPLLQHMGDMSHAAAMYQ
QFFFEQEARMAKEAAEQQQQQQQQQQQQQQQQQQEQQRRFEQEQQEQQRRKEEQQQIQRQQQHLQQLQQ
QQMEQQHVATAAPRPQMHHPAPARLPTRMGGAAGHTALKSELSEKFQMLRANNNSSMMRMSGTDLEGLAD
VLKSEITTSLSALVDTIVTRFVHQRRLFSKQADSVTAAAEQLNKDLLLASQILDRKSPRTKVADRPQNGP
TPATQSAAAMFQAPKTPQGMNPVAAAALYNSMTGPFCLPPDQQQQQTAQQQSAQQQQSSQQTQQQLE
QNEALSLVVTPKKKRHKVTDTRITPRTVSRILAQDGVVPPTGGPPSTPQQQQQQQQQQQQQQQQQQQQAS
NGGNSNATPAQSPTRSSGGAAYHPQPPPPPPPMMPVSLPTSVAIPNPSLHESKVFSPYSPFFNPHAAAGQ
ATAAQLHQHHQQHHPHHQSMQLSSSPPGSLGALMDSRDSPPLPHPPSMLHPALLAAAHHGGSPDYKTCLR
AVMDAQDRQSECNSADMQFDGMAPTISFYKQMQLKTEHQESLMAKHCESLTPLHSSTLTPMHLRKAKLMF
FWVRYPSSAVLKMYFPDIKFNKNNTAQLVKWFSNFREFYYIQMEKYARQAVTEGIKTPDDLLIAGDSELY
RVLNLHYNRNNHIEVPQNFRFVVESTLREFFRAIQGGKDTEQSWKKSIYKIISRMDDPVPEYFKSPNFLE
QLE
```

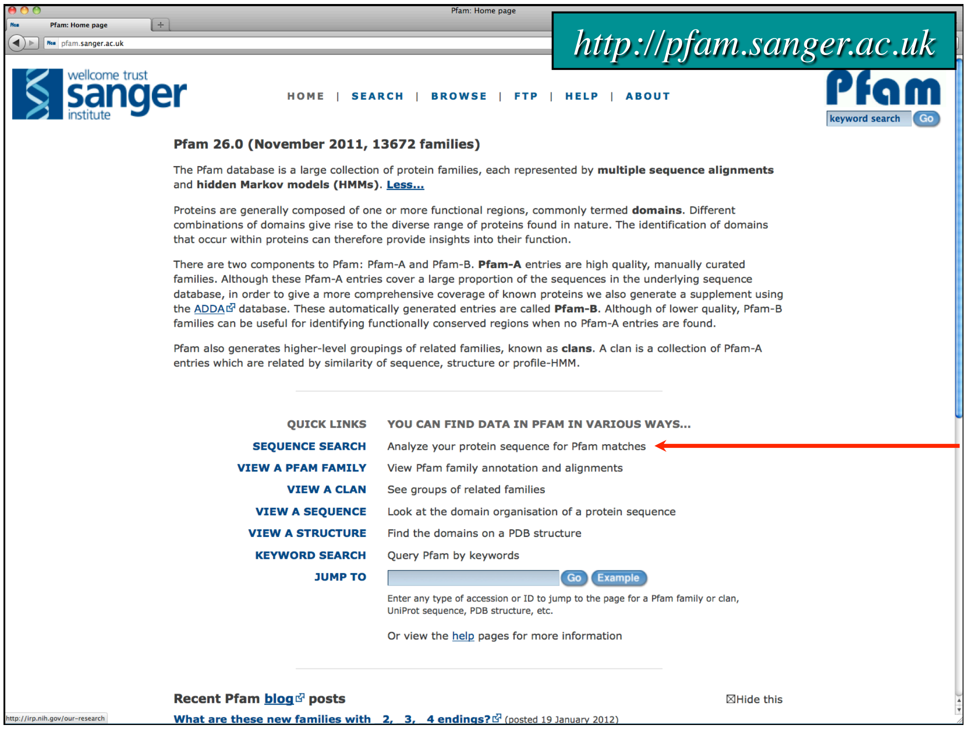
BLAST 2 Sequences

```
>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
```

BLAT

```
>CB312815 NICHD_Rr_Pit1 Rattus norvegicus cDNA clone
GGGGCTCTCGCTGGCCTGTGTCTCAGAAGCTGCTTTCTCCACCTCTTCCTTGTGAATTTCCTAAACTCTC
TACCTCTGGTTCATGTTCGCTCTTCTGGATAGTCTGTGTGCAATGAGCCCTTAAAGGAATATTGCAATGA
GCTATAAGAGTTGTGAGCCTGCGGTAGGCAAGGCCTGCACTGGGACAGCAAAGGAAATTTCATTGCATCT
GCTCCTAAGTCACAGGTTATCCAGAGCCCACTTTACCCCAAGAGACAGCCTCTCCCCCATCCCTAGGAAA
CAGTAGAGCTTAGGAAAATGAATGACTCCACCACATTCAAGAGGCTTCAAATTGTATACTTGGCATTTCT
GATTTCAGTTCTGAAATTCTGTCCCTTAGTCGTGGGGAAAATAAGAAATGGAGTTACACCTTGTCATTTA
```

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


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

http://pfam.sanger.ac.uk

Pfam 26.0 (November 2011, 13672 families)

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

Proteins are generally composed of one or more functional regions, commonly termed **domains**. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function.

There are two components to Pfam: Pfam-A and Pfam-B. **Pfam-A** entries are high quality, manually curated families. Although these Pfam-A entries cover a large proportion of the sequences in the underlying sequence database, in order to give a more comprehensive coverage of known proteins we also generate a supplement using the ADDA database. These automatically generated entries are called **Pfam-B**. Although of lower quality, Pfam-B families can be useful for identifying functionally conserved regions when no Pfam-A entries are found.

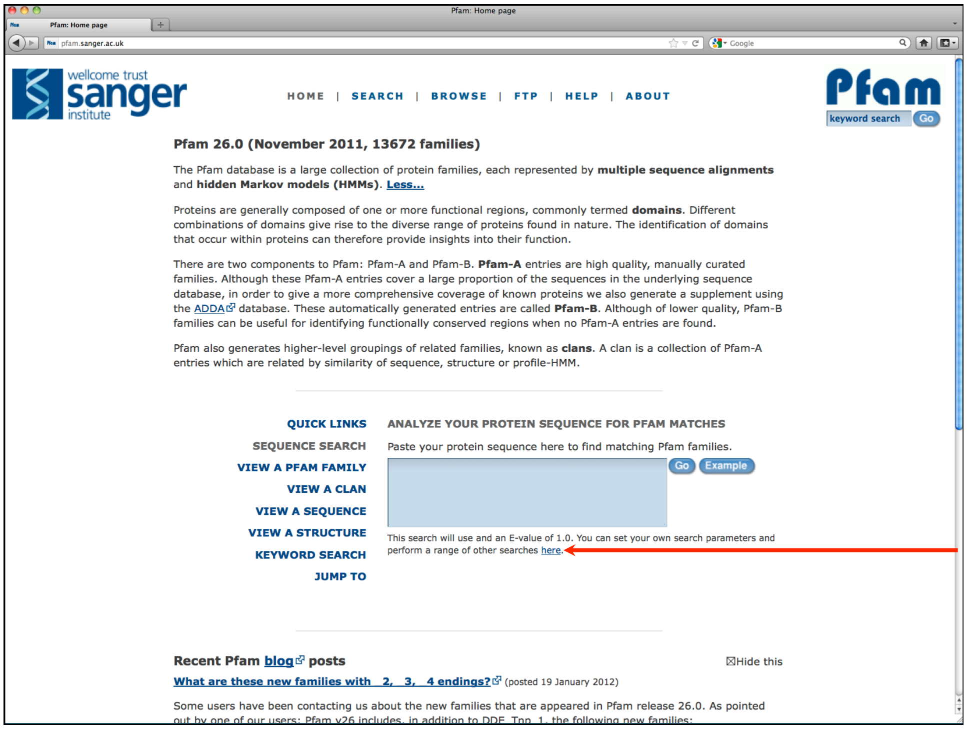
Pfam also generates higher-level groupings of related families, known as **clans**. A clan is a collection of Pfam-A entries which are related by similarity of sequence, structure or profile-HMM.

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

What are these new families with 2, 3, 4 endings? (posted 19 January 2012)

http://irp.nih.gov/our-research


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

Pfam 26.0 (November 2011, 13672 families)

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

Proteins are generally composed of one or more functional regions, commonly termed **domains**. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function.

There are two components to Pfam: Pfam-A and Pfam-B. **Pfam-A** entries are high quality, manually curated families. Although these Pfam-A entries cover a large proportion of the sequences in the underlying sequence database, in order to give a more comprehensive coverage of known proteins we also generate a supplement using the ADDA database. These automatically generated entries are called **Pfam-B**. Although of lower quality, Pfam-B families can be useful for identifying functionally conserved regions when no Pfam-A entries are found.

Pfam also generates higher-level groupings of related families, known as **clans**. A clan is a collection of Pfam-A entries which are related by similarity of sequence, structure or profile-HMM.

QUICK LINKS
SEQUENCE SEARCH
VIEW A PFAM FAMILY
VIEW A CLAN
VIEW A SEQUENCE
VIEW A STRUCTURE
KEYWORD SEARCH
JUMP TO

ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES

Paste your protein sequence here to find matching Pfam families.

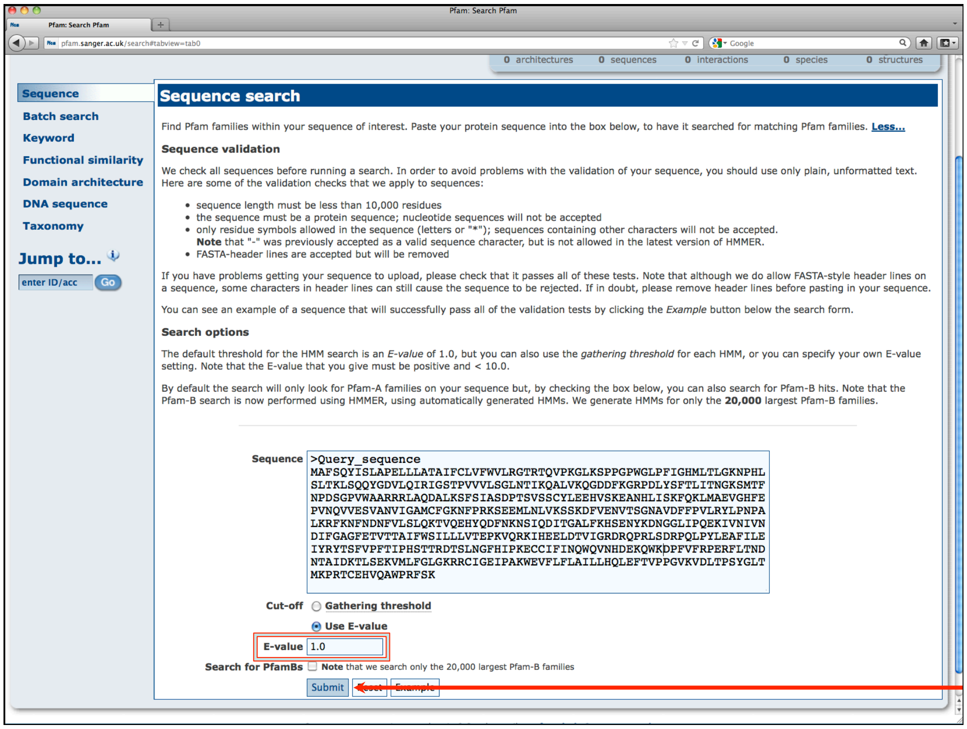
Go Example

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 ☒Hide this

What are these new families with 2, 3, 4 endings? (posted 19 January 2012)

Some users have been contacting us about the new families that are appeared in Pfam release 26.0. As pointed out by one of our users: Pfam v26 includes, in addition to DDE_Tnp_1, the following new families:



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 and < 10.0.

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. Note that the Pfam-B search is now performed using HMMER, using automatically generated HMMs. We generate HMMs for only the **20,000** largest Pfam-B families.

Sequence

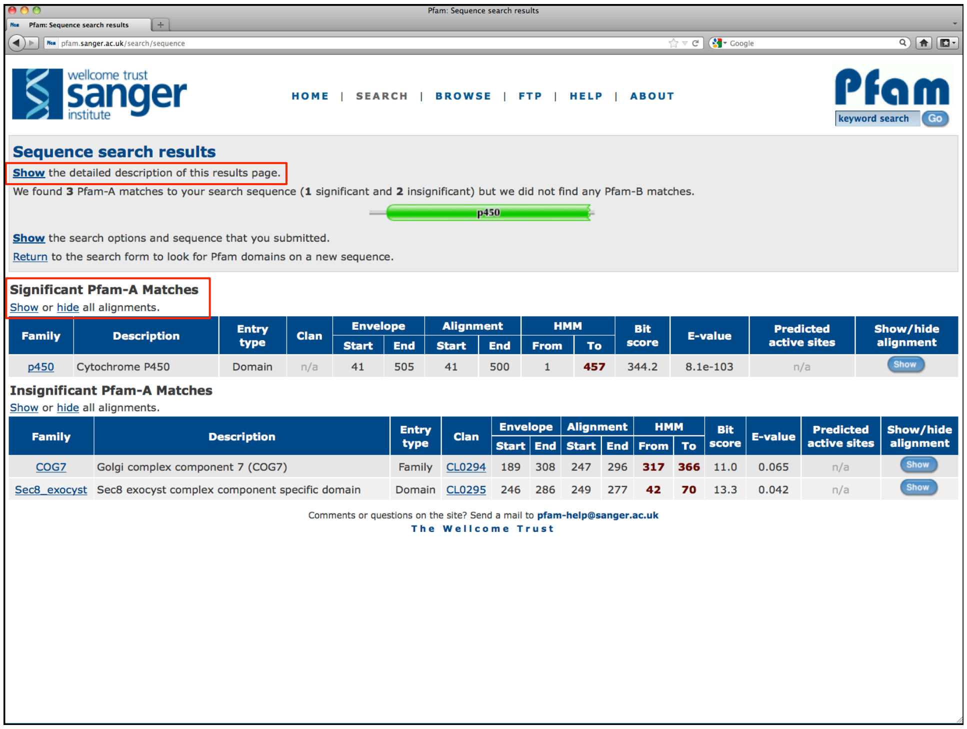
```
>Query_sequence
MAFSQYISLAPELLLATAIFCLVFWVLRGTRTQVPKGLKSPPGPWGLPFIGHMLTLGKNPHL
SLTKLSQQYGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFTLITNGKSMTF
NPDSGPVWAARRRLAQDALKSFSIASDPTSVSSCYLEEHVSKEANHLISKFQKLMAEVGHFE
PVNQVVESVANVIGAMCFGKNFPRKSEEMLNLVKSSKDFVENVTSGNAVDFFPVLRYLPNPA
LKRFKNFNDNFVLSLQKTVQEHYQDFNKNSIQDITGALFKHSENYKDNGGLIPQEKIVNIVN
DIFGAGFETVTTAIFWSILLLVTEPKVQRKIHEELDTVIGRDRQPRLSDRPQLPYLEAFILE
IYRYTSFVPFTIPHSTTRDTSLNGFHIPKECCIFINQWQVNHDEKQWKDPFVFRPERFLTND
NTAIDKTLSEKVMLFGLGKRRCIGEIPAKWEVFLFLAILLHQLEFTVPPGVKVDLTPSYGLT
MKPRTCEHVQAWPRFSK
```

Cut-off ○ **Gathering threshold**
● **Use E-value**

E-value 1.0

Search for PfamBs □ **Note** that we search only the 20,000 largest Pfam-B families

Submit Reset Example


Pfam: Sequence search results
pfam.sanger.ac.uk/search/sequence
Google
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search
Go
p450

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.

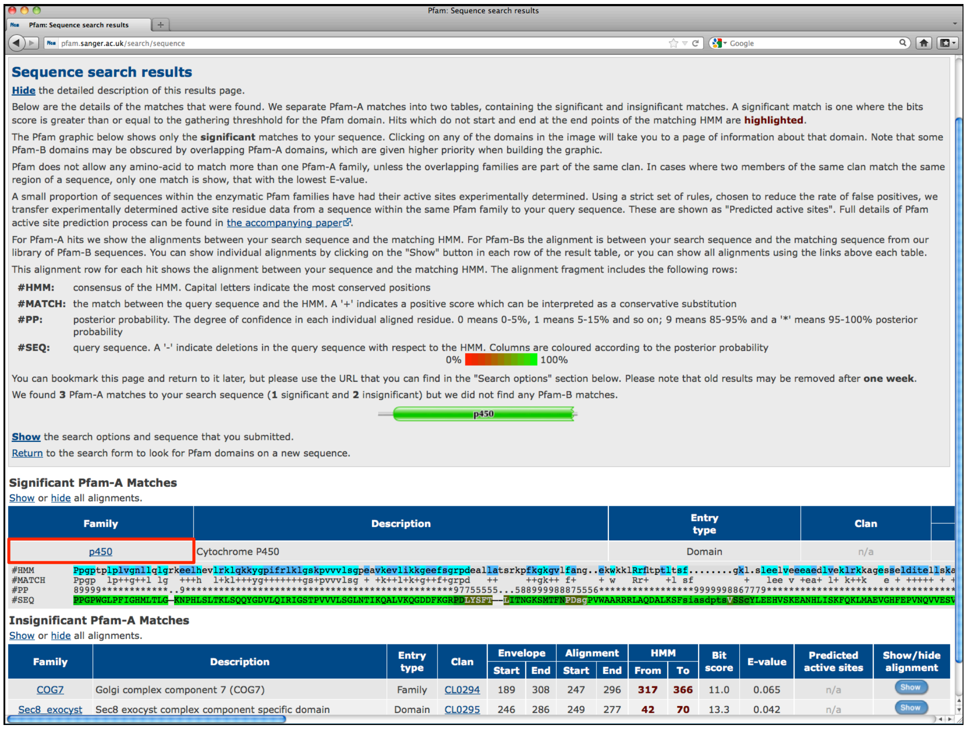
Family	Description	Entry type	Clan	Envelope		Alignment		HMM		Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To				
p450	Cytochrome P450	Domain	n/a	41	505	41	500	1	457	344.2	8.1e-103	n/a	Show

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	0.065	n/a	Show
Sec8_exocyst	Sec8 exocyst complex component specific domain	Domain	CL0295	246	286	249	277	42	70	13.3	0.042	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
Pfam: Sequence search results
pfam.sanger.ac.uk/search/sequence
Google
p450
Show
Show

Sequence search results

Hide the detailed description of this results page.

Below are the details of the matches that were found. We separate Pfam-A matches into two tables, containing the significant and insignificant matches. A significant match is one where the bits score is greater than or equal to the gathering threshhold for the Pfam domain. Hits which do not start and end at the end points of the matching HMM are **highlighted**.

The Pfam graphic below shows only the **significant** matches to your sequence. 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.

Pfam does not allow any amino-acid to match more than one Pfam-A family, unless the overlapping families are part of the same clan. In cases where two members of the same clan match the same region of a sequence, only one match is show, that with the lowest E-value.

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.

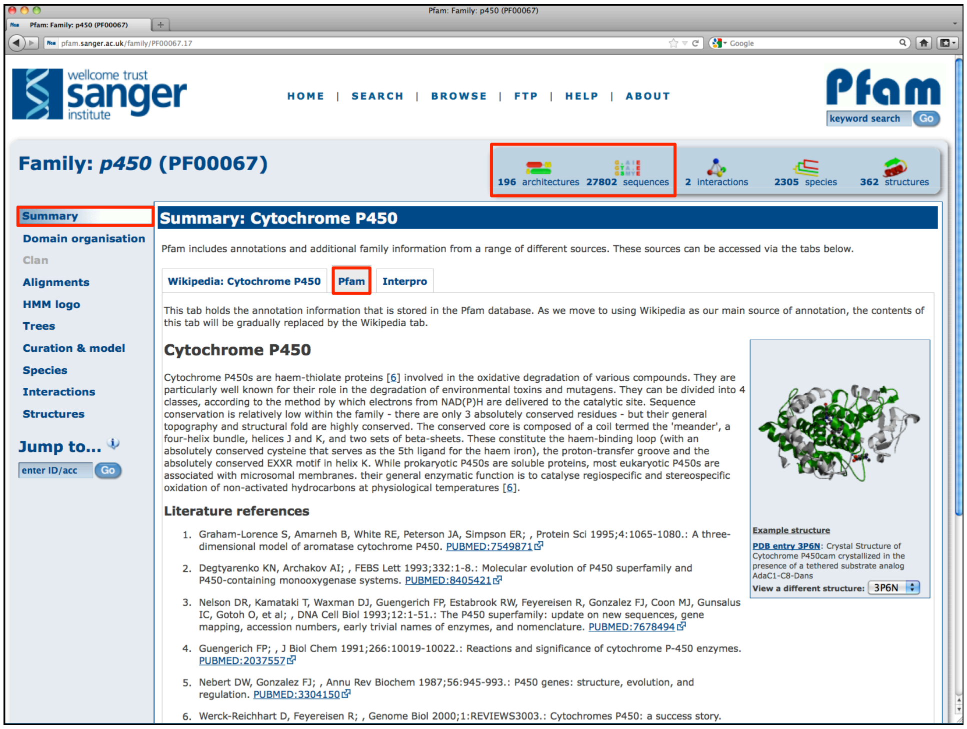
Family	Description	Entry type	Clan
p450	Cytochrome P450	Domain	n/a

```
#HMM      Ppgptplplvgnllqlgrkeelhevlrklqkkygpifrlklgskpvvvlsgpeavkevlikkgeefsgrpdeallatsrkpfkgkgvlfang..ekwkklRrfltptltsf........gkl.sleelveeeaedlveklrkkagesselditellska
#MATCH    Ppgp  lp++g++l lg  +++h  l+kl+++yg+++++++gs+pvvvlsg + +k++l+k+g++f+grpd   ++      ++gk++ f+    + w   Rr+   +l sf           +   lee v +ea+ l+ k++k   e + +++++ + +
#PP       89999************..9*************************************************97755555...588999988875556***************999999886777 9*************************************
#SEQ      PPGPWGLPFIGHMLTLG--KNPHLSLTKLSQQYGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFT---LITNGKSMTFNPDsgPVWAARRRLAQDALKSFsiasdptsVSScYLEEHVSKEANHLISKFQKLMAEVGHFEPVNQVVESV
```

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	0.065	n/a	Show
Sec8_exocyst	Sec8 exocyst complex component specific domain	Domain	CL0295	246	286	249	277	42	70	13.3	0.042	n/a	Show



Family: *p450* (PF00067)

Summary: Cytochrome P450

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

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].

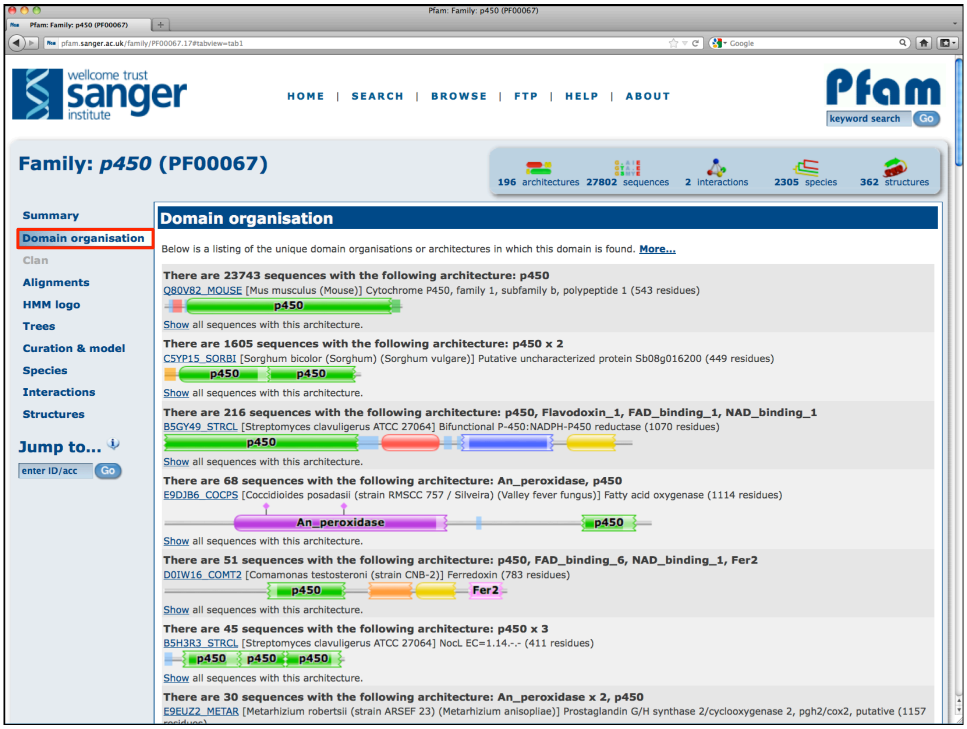
Example structure

PDB entry 3P6N: Crystal Structure of Cytochrome P450cam crystallized in the presence of a tethered substrate analog AdaC1-C8-Dans

View a different structure: 3P6N

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.



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 23743 sequences with the following architecture: p450
Q80V82_MOUSE [Mus musculus (Mouse)] Cytochrome P450, family 1, subfamily b, polypeptide 1 (543 residues)
Show all sequences with this architecture.

There are 1605 sequences with the following architecture: p450 x 2
C5YP15_SORBI [Sorghum bicolor (Sorghum) (Sorghum vulgare)] Putative uncharacterized protein Sb08g016200 (449 residues)
Show all sequences with this architecture.

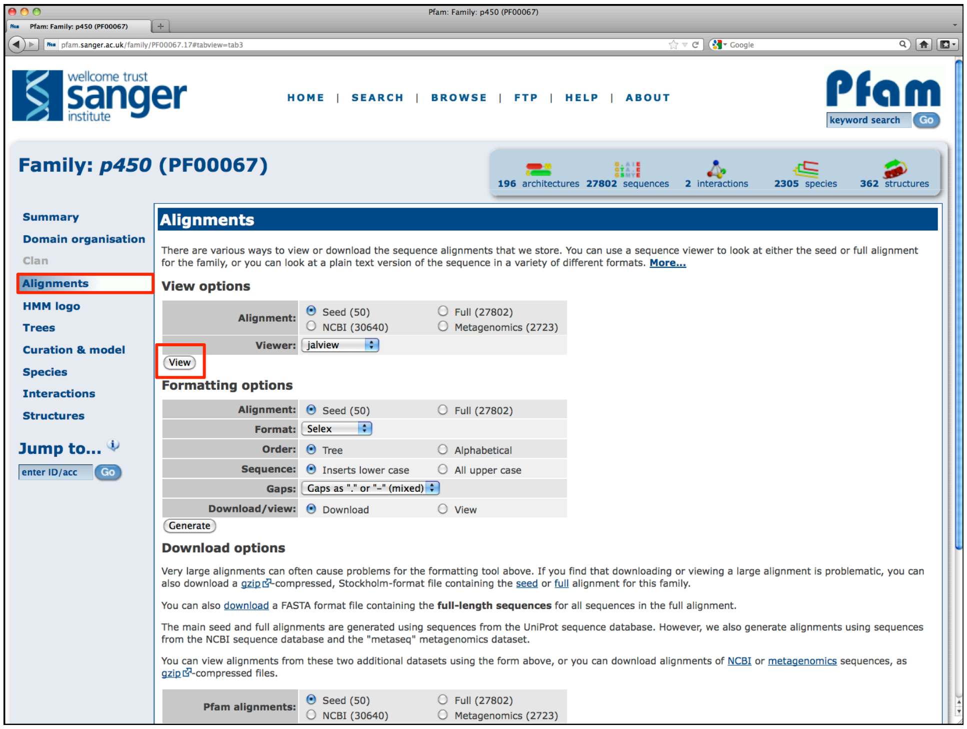
There are 216 sequences with the following architecture: p450, Flavodoxin_1, FAD_binding_1, NAD_binding_1
B5GY49_STRCL [Streptomyces clavuligerus ATCC 27064] Bifunctional P-450:NADPH-P450 reductase (1070 residues)
Show all sequences with this architecture.

There are 68 sequences with the following architecture: An_peroxidase, p450
E9DJB6_COCPS [Coccidioides posadasii (strain RMSCC 757 / Silveira) (Valley fever fungus)] Fatty acid oxygenase (1114 residues)
Show all sequences with this architecture.

There are 51 sequences with the following architecture: p450, FAD_binding_6, NAD_binding_1, Fer2
D0IW16_COMT2 [Comamonas testosteroni (strain CNB-2)] Ferredoxin (783 residues)
Show all sequences with this architecture.

There are 45 sequences with the following architecture: p450 x 3
B5H3R3_STRCL [Streptomyces clavuligerus ATCC 27064] NocL EC=1.14.-.- (411 residues)
Show all sequences with this architecture.

There are 30 sequences with the following architecture: An_peroxidase x 2, p450
E9EUZ2_METAR [Metarhizium robertsii (strain ARSEF 23) (Metarhizium anisopliae)] Prostaglandin G/H synthase 2/cyclooxygenase 2, pgh2/cox2, putative (1157 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 (30640)	○ Full (27802) ○ Metagenomics (2723)
Viewer:	jalview	

View

Formatting options

Alignment:	● Seed (50)	○ Full (27802)
Format:	Selex	
Order:	● Tree	○ Alphabetical
Sequence:	● Inserts lower case	○ All upper case
Gaps:	Gaps as "." or "-" (mixed)	
Download/view:	● Download	○ View

Generate

Download options

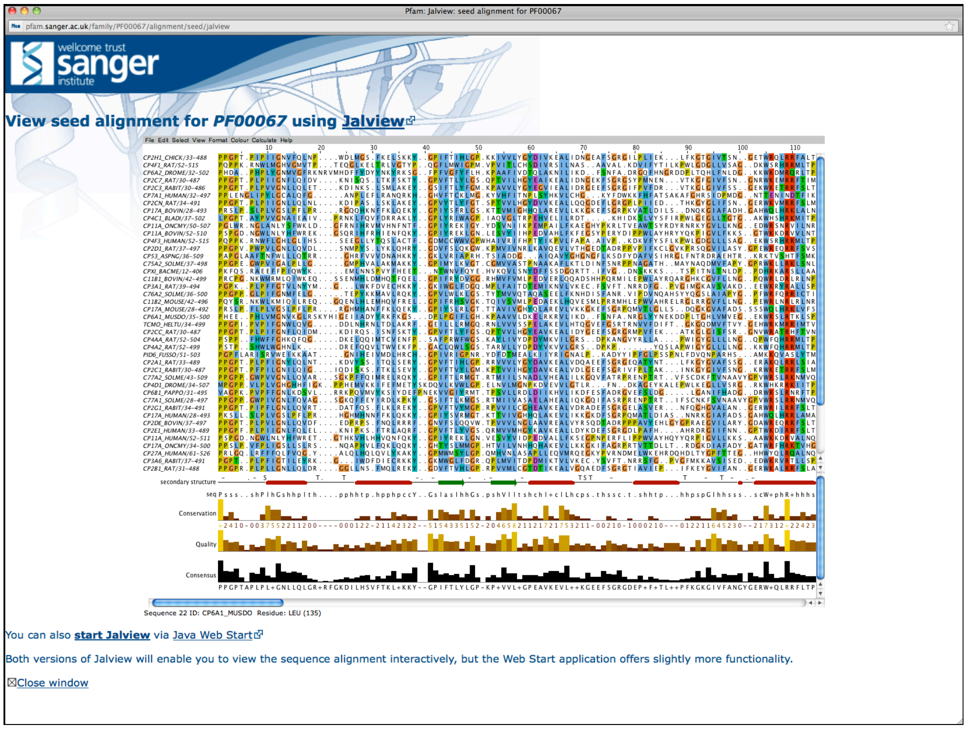
Very large alignments can often cause problems for the formatting tool above. If you find that downloading or viewing a large alignment is problematic, you can also download a gzip-compressed, Stockholm-format file containing the seed or full alignment for this family.

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 (30640)	○ Full (27802) ○ Metagenomics (2723)


Pfam: Jalview: seed alignment for PF00067
pfam.sanger.ac.uk/family/PF00067/alignment/seed/jalview
wellcome trust sanger institute
File Edit Select View Format Colour Calculate Help
secondary structure
Conservation
Quality
Consensus

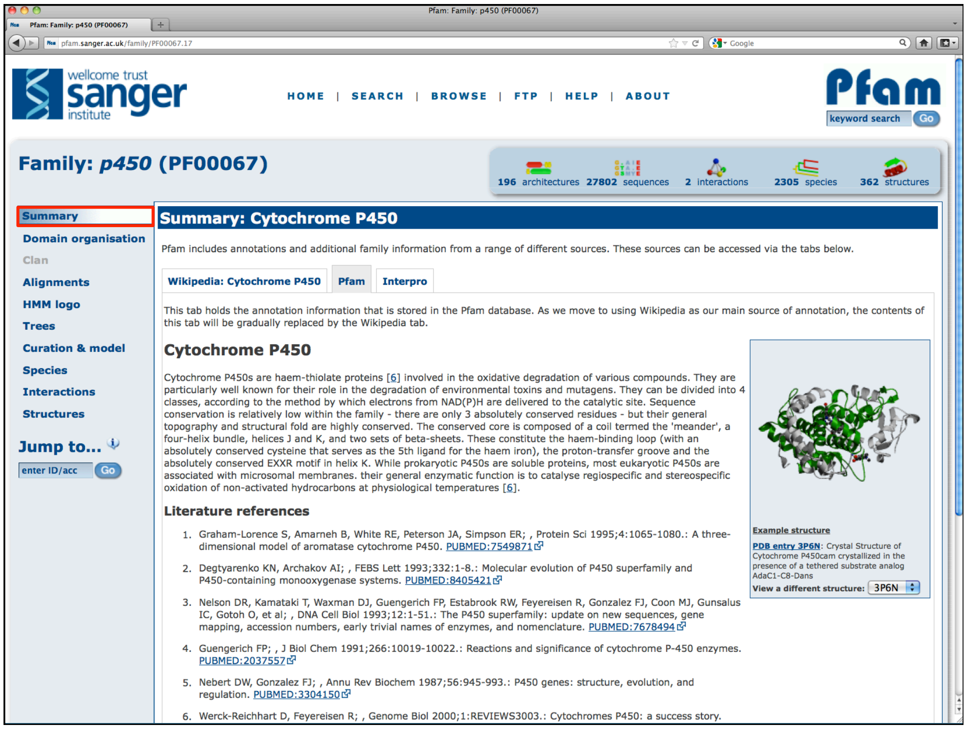
View seed alignment for *PF00067* using Jalview

Sequence 22 ID: CP6A1_MUSDO Residue: LEU (135)

You can also start Jalview via Java Web Start

Both versions of Jalview will enable you to view the sequence alignment interactively, but the Web Start application offers slightly more functionality.

☒Close window



Family: *p450* (PF00067)

Summary: Cytochrome P450

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

Cytochrome P450

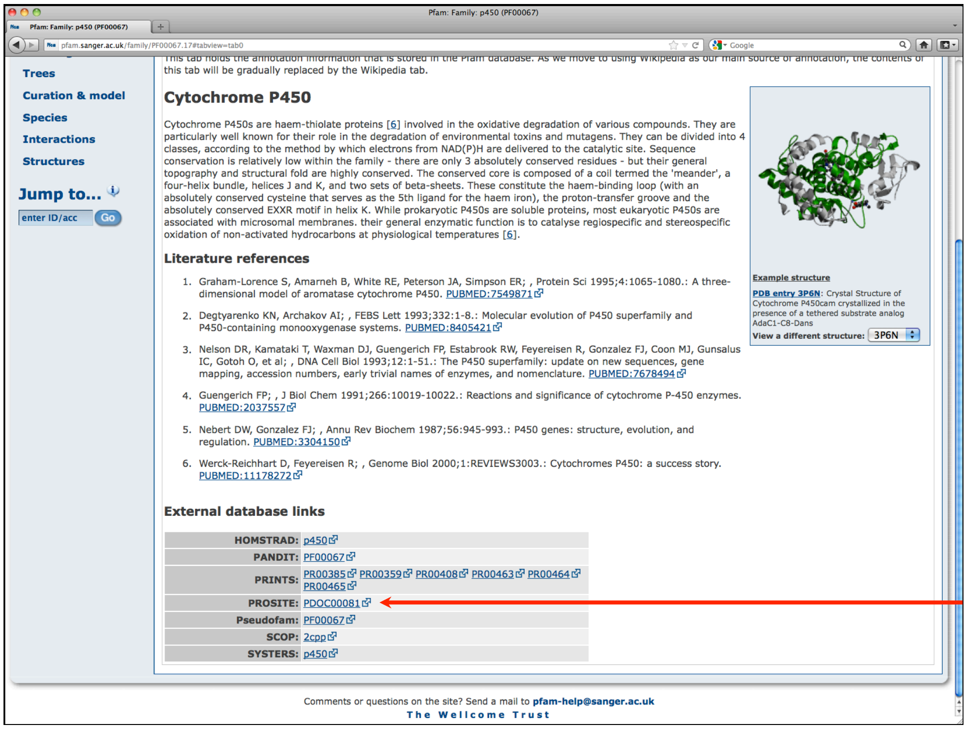
Cytochrome P450s are haem-thiolate proteins [6] involved in the oxidative degradation of various compounds. They are particularly well known for their role in the degradation of environmental toxins and mutagens. They can be divided into 4 classes, according to the method by which electrons from NAD(P)H 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 3P6N: Crystal Structure of Cytochrome P450cam crystallized in the presence of a tethered substrate analog AdaC1-C8-Dans

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.


Pfam: Family: p450 (PF00067)
Pfam: Family: p450 (PF00067)
pfam.sanger.ac.uk/family/PF00067.17#tabview=tab0
Google
Trees
Curation & model
Species
Interactions
Structures
Jump to...
enter ID/acc
Go

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

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 3P6N: Crystal Structure of Cytochrome P450cam crystallized in the presence of a tethered substrate analog AdaC1-C8-Dans

View a different structure: 3P6N

Literature references

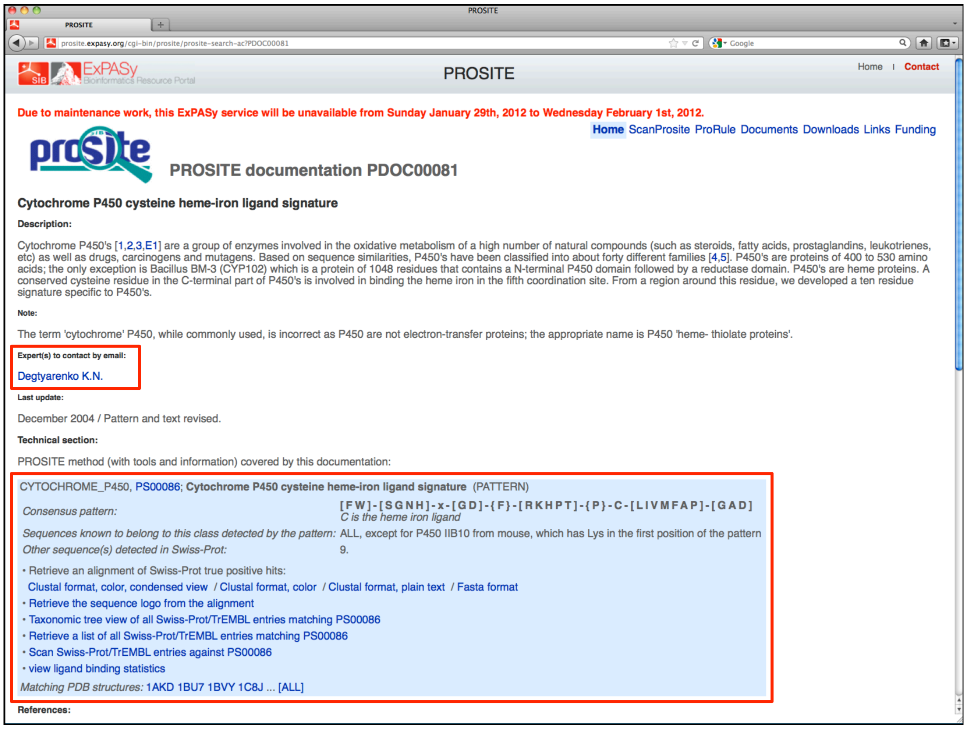
1. Graham-Lorence S, Amarneh B, White RE, Peterson JA, Simpson ER; , Protein Sci 1995;4:1065-1080.: A three-dimensional model of aromatase cytochrome P450. PUBMED:7549871
2. Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase 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

External database links

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

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

The Wellcome Trust


PROSITE
PROSITE
prosite.expasy.org/cgi-bin/prosite/prosite-search-ac?PDOC00081
Google
SIB ExPASy Bioinformatics Resource Portal
Home
Contact
Home
ScanProsite
ProRule
Documents
Downloads
Links
Funding
prosite

PROSITE

Due to maintenance work, this ExPASy service will be unavailable from Sunday January 29th, 2012 to Wednesday February 1st, 2012.

PROSITE documentation PDOC00081

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:

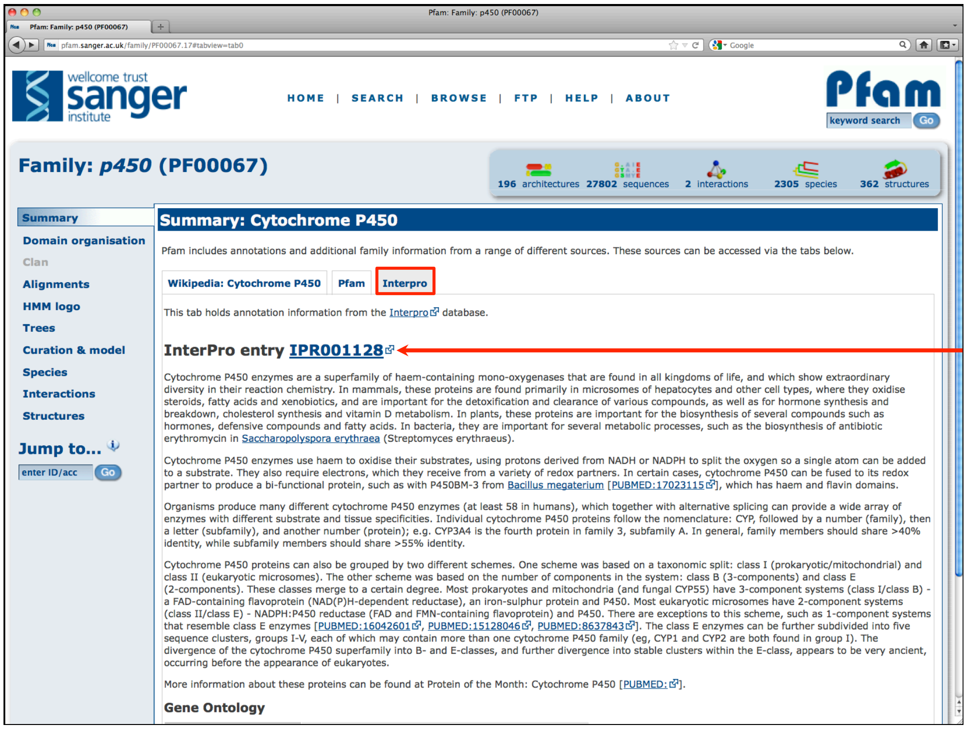
CYTOCHROME_P450, PS00086; **Cytochrome P450 cysteine heme-iron ligand signature** (PATTERN)

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:



Family: *p450* (PF00067)

Summary: Cytochrome P450

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

This tab holds annotation information from the Interpro database.

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).

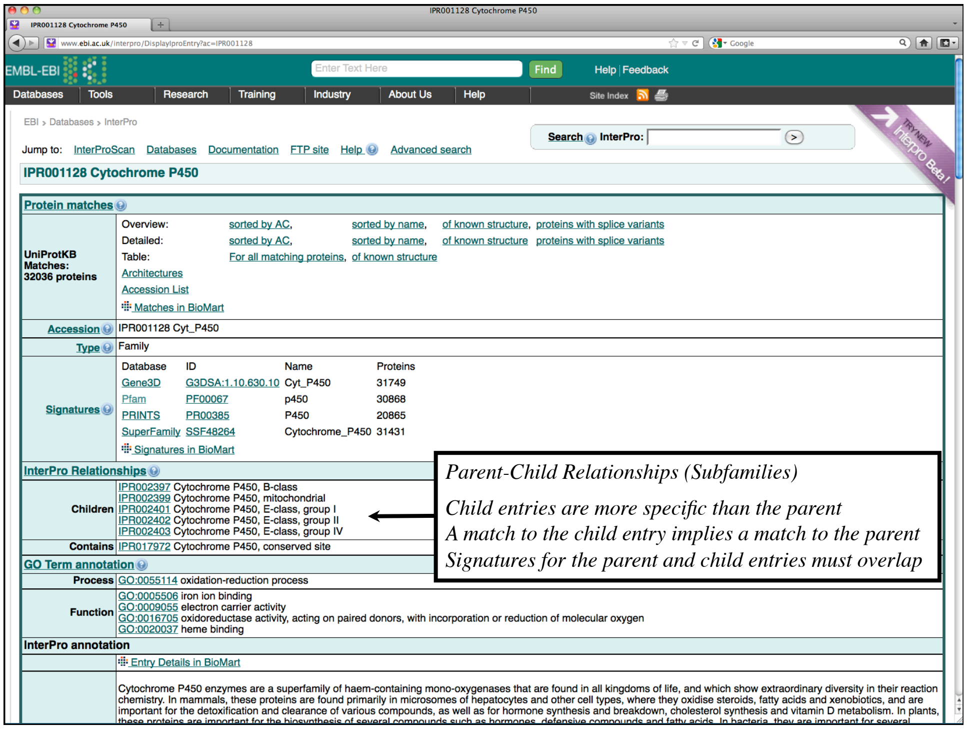
Cytochrome P450 enzymes use haem to oxidise their substrates, using protons derived from NADH or NADPH to split the oxygen so a single atom can be added to a substrate. They also require electrons, which they receive from a variety of redox partners. In certain cases, cytochrome P450 can be fused to 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

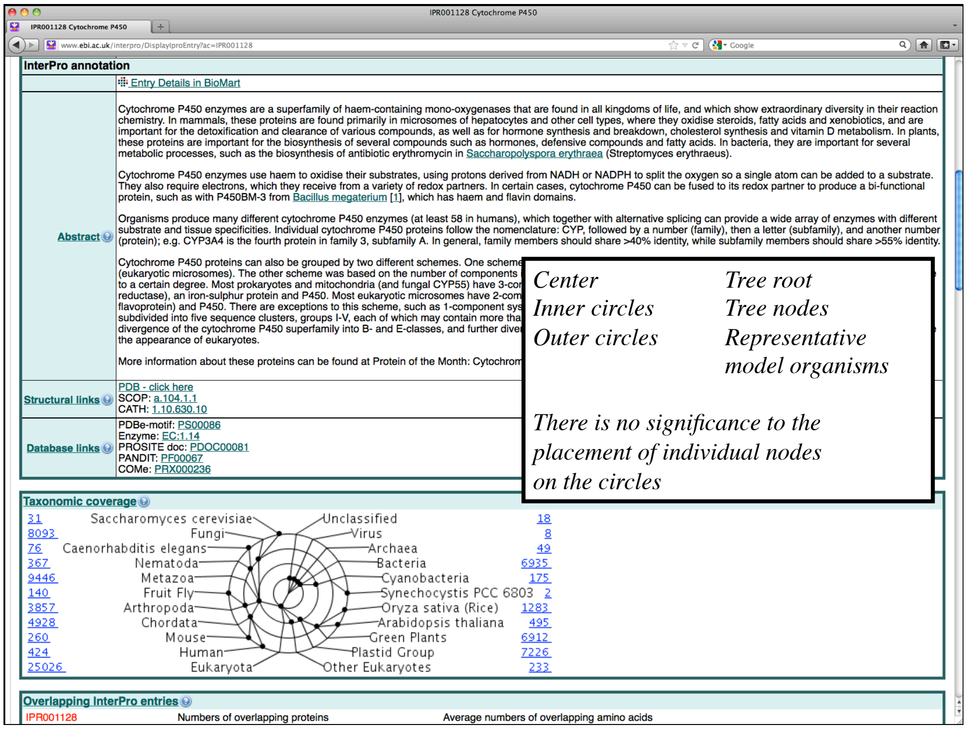

IPR001128 Cytochrome P450
www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128
EMBL-EBI
Enter Text Here
Find
Help | Feedback
Databases
Tools
Research
Training
Industry
About Us
Help
Site Index
EBI › Databases › InterPro
Search InterPro:
TRY NEW Interpro Beta!
Jump to: InterProScan Databases Documentation FTP site Help Advanced search

IPR001128 Cytochrome P450

Protein matches	
UniProtKB Matches: 32036 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 / 31749 Pfam / PF00067 / p450 / 30868 PRINTS / PR00385 / P450 / 20865 SuperFamily / SSF48264 / Cytochrome_P450 / 31431 Signatures in BioMart
InterPro Relationships	
Children	IPR002397 Cytochrome P450, B-class IPR002399 Cytochrome P450, mitochondrial IPR002401 Cytochrome P450, E-class, group I IPR002402 Cytochrome P450, E-class, group II IPR002403 Cytochrome P450, E-class, group IV
Contains	IPR017972 Cytochrome P450, conserved site
GO Term annotation	
Process	GO:0055114 oxidation-reduction process
Function	GO:0005506 iron ion binding GO:0009055 electron carrier activity GO:0016705 oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen 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

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

InterPro annotation

Entry Details in BioMart

Abstract

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 [1], 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
(eukaryotic microsomes). The other scheme was based on the number of components
to a certain degree. Most prokaryotes and mitochondria (and fungal CYP55) have 3-cor
reductase), an iron-sulphur protein and P450. Most eukaryotic microsomes have 2-com
flavoprotein) and P450. There are exceptions to this scheme, such as 1-component sys
subdivided into five sequence clusters, groups I-V, each of which may contain more tha
divergence of the cytochrome P450 superfamily into B- and E-classes, and further diver
the appearance of eukaryotes.

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

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

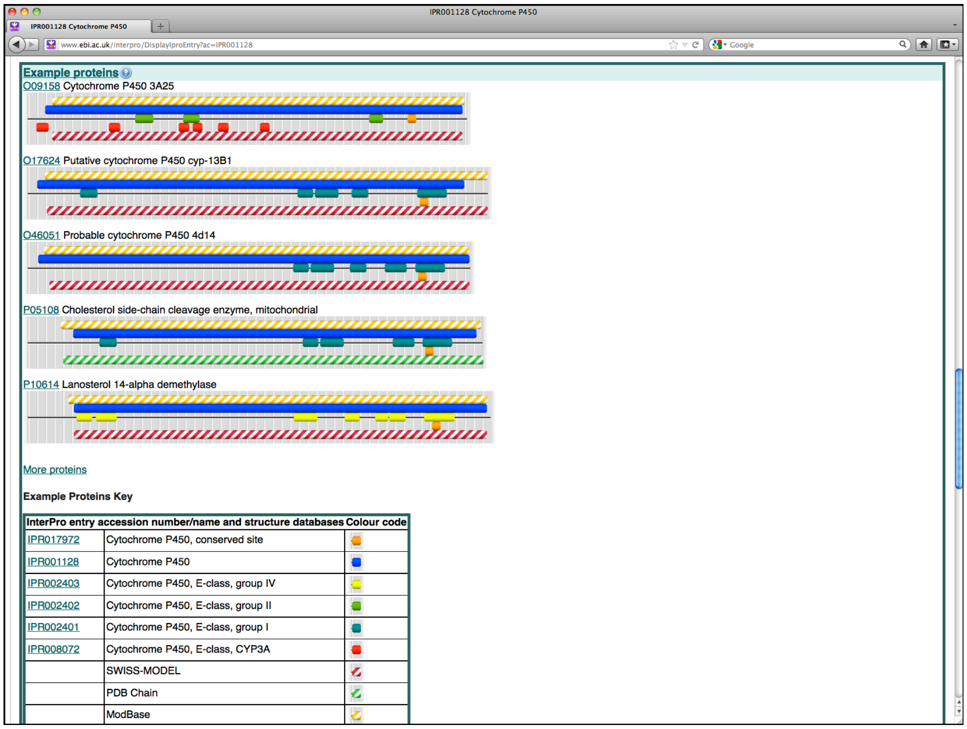
Taxonomic coverage

Overlapping InterPro entries

IPR001128 Numbers of overlapping proteins Average numbers of overlapping amino acids

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

Example proteins

O09158 Cytochrome P450 3A25

O17624 Putative cytochrome P450 cyp-13B1

O46051 Probable cytochrome P450 4d14

P05108 Cholesterol side-chain cleavage enzyme, mitochondrial

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	🟦
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	🟥
	SWISS-MODEL	▨
	PDB Chain	▨
	ModBase	▨

Further Reading

Identifying Protein Domains with the Pfam Database — *UNIT 2.5*

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

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

ANALYZING A PROTEIN SEQUENCE WITH Pfam VIA THE WEB — *BASIC PROTOCOL*

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

Necessary Resources

Hardware

Workstation with network connection

Software

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

Files

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

The example sequences used below are available at the Current Protocols Web site (http://www3.interscience.wiley.com/c_p/cpbi_sampledatafiles.htm).

Contributed by Robert Finn, Sam Griffiths-Jones, and Alex Bateman
Current Protocols in Bioinformatics (2003) 2.5.1-2.5.19
Copyright © 2003 by John Wiley & Sons, Inc.

Recognizing Functional Domains
2.5.1
Supplement 1

Current Protocols in Bioinformatics
Unit 2.5
Identifying Protein Domains with the Pfam Database

The InterPro Database and Tools for Protein Domain Analysis — *UNIT 2.7*

InterPro (Apweiler et al., 2001a) is an integrated documentation resource of protein families, domains, and functional sites that rationalizes data from the major protein-signature databases Pfam (UNIT 2.5; Bateman et al., 2002), PROSITE (UNIT 3.6; Falquet et al., 2002), PRINTS (Attwood et al., 2002), ProDom (Corpet et al., 2000), SMART (Letunic et al., 2002), PIR Superfamily (Wu et al., 2003), and TIGRFAMs (Haft et al., 2001). These databases, summarized in Table 2.7.1 along with other useful sites (also see Internet Resources), use a variety of methods for creation of diagnostic signatures, including regular expressions, profiles, fingerprints, hidden Markov models (HMMs), and sequence-clustering algorithms. In InterPro, signatures describing the same domain, family, repeat, or post-translational modification are grouped into single entries with unique accession numbers. Entries that are subsets of others are associated by parent/child (family and subfamily) or contains/found-in (domain composition) relationships. The associated software package, InterProScan (Zdobnov and Apweiler, 2001), combines the search methods from each of the member databases into one package and provides an output with all results in a single format, namely HTML, XML, or text. The software has been used for annotation of numerous completely sequenced genomes, notably human (The International Human Genome Consortium, 2001).

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

PROTEIN SEQUENCE CLASSIFICATION USING INTERPROSCAN VIA THE INTERNET — *BASIC PROTOCOL 1*

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

Contributed by Nicola J. Mulder and Rolf Apweiler
Current Protocols in Bioinformatics (2003) 2.7.1-2.7.19
Copyright © 2003 by John Wiley & Sons, Inc.

Recognizing Functional Domains
2.7.1
Supplement 2

Current Protocols in Bioinformatics
Unit 2.7
The InterPro Database and Tools for Protein Domain Analysis

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Conserved Domain Database (CDD)

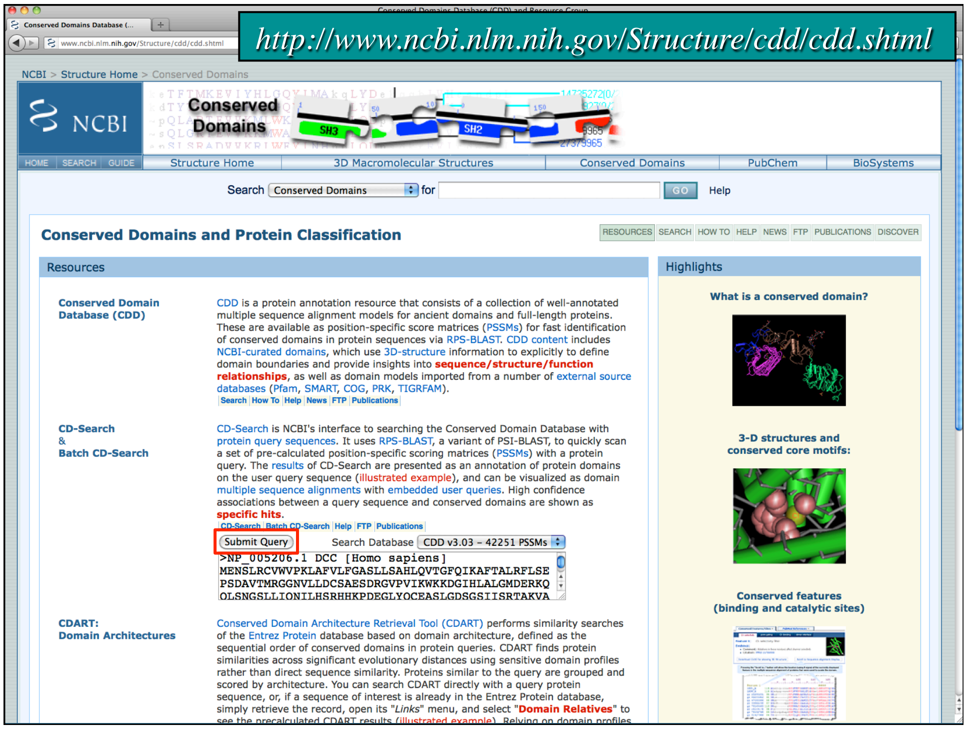
- Identify conserved domains in a protein sequence
- "Secondary database"
  - Pfam A (not Pfam B)
  - Simple Modular Architecture Research Tool (SMART)
  - COG (orthologous prokaryotic protein families)
  - KOG (eukaryotic equivalent of COG)
  - PRK ("protein clusters" of related protein RefSeq entries)
  - 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 Database (...
www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
NCBI > Structure Home > Conserved Domains
NCBI
Conserved Domains
SH3
SH2
HOME SEARCH GUIDE
Structure Home
3D Macromolecular Structures
Conserved Domains
PubChem
BioSystems
Search Conserved Domains for
GO
Help
RESOURCES SEARCH HOW TO HELP NEWS FTP PUBLICATIONS DISCOVER

Conserved Domains and Protein Classification

Resources

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).

Search | How To | Help | News | FTP | Publications

CD-Search & Batch 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**.

CD-Search | Batch CD-Search | Help | FTP | Publications

Submit Query   Search Database: CDD v3.03 – 42251 PSSMs

```
>NP_005206.1 DCC [Homo sapiens]
MENSLRCVWVPKLAFVLFGASLLSAHLQVTGFQIKAFTALRFLSE
PSDAVTMRGGNVLLDCSAESDRGVPVIKWKKDGIHLALGMDERKQ
QLSNGSLLIQNILHSRHHKPDEGLYQCEASLGDSGSIISRTAKVA
```

CDART: Domain Architectures

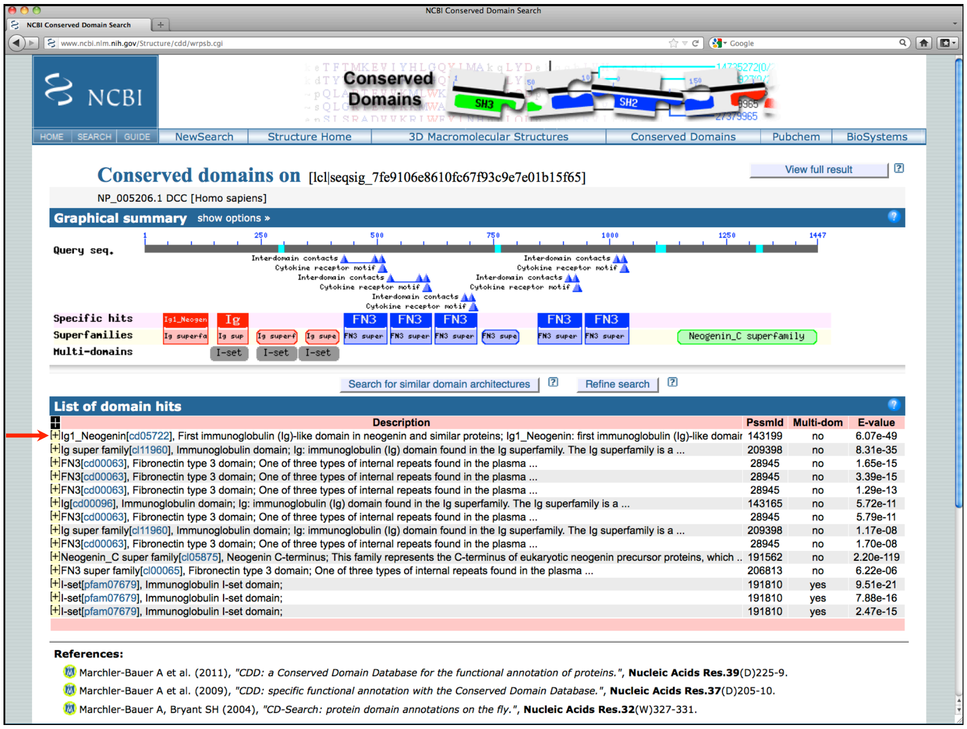
Conserved Domain Architecture Retrieval Tool (CDART) performs similarity searches of the Entrez Protein database based on domain architecture, defined as the sequential order of conserved domains in protein queries. CDART finds protein similarities across significant evolutionary distances using sensitive 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

Highlights

What is a conserved domain?

3-D structures and conserved core motifs:

Conserved features (binding and catalytic sites)



Conserved domains on [lcl|seqsig_7fe9106e8610fc67f93c9e7e01b15f65]

NP_005206.1 DCC [Homo sapiens]

List of domain hits

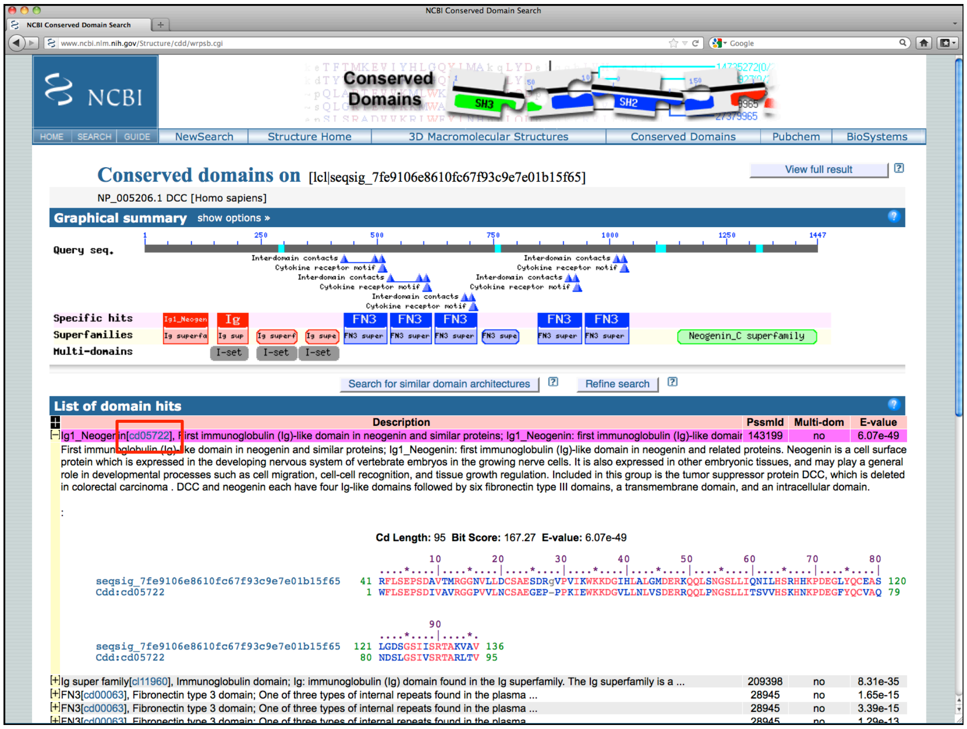
Description	PssmId	Multi-dom	E-value
[+]Ig1_Neogenin[cd05722], First immunoglobulin (Ig)-like domain in neogenin and similar proteins; Ig1_Neogenin: first immunoglobulin (Ig)-like domair	143199	no	6.07e-49
[+]Ig super family[cl11960], Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig superfamily is a ...	209398	no	8.31e-35
[+]FN3[cd00063], Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	28945	no	1.65e-15
[+]FN3[cd00063], Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	28945	no	3.39e-15
[+]FN3[cd00063], Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	28945	no	1.29e-13
[+]Ig[cd00096], Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig superfamily is a ...	143165	no	5.72e-11
[+]FN3[cd00063], Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	28945	no	5.79e-11
[+]Ig super family[cl11960], Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig superfamily is a ...	209398	no	1.17e-08
[+]FN3[cd00063], Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	28945	no	1.70e-08
[+]Neogenin_C super family[cl05875], Neogenin C-terminus; This family represents the C-terminus of eukaryotic neogenin precursor proteins, which ..	191562	no	2.20e-119
[+]FN3 super family[cl00065], Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	206813	no	6.22e-06
[+]I-set[pfam07679], Immunoglobulin I-set domain;	191810	yes	9.51e-21
[+]I-set[pfam07679], Immunoglobulin I-set domain;	191810	yes	7.88e-16
[+]I-set[pfam07679], Immunoglobulin I-set domain;	191810	yes	2.47e-15

References:

Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.

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.



Conserved domains on [lcl|seqsig_7fe9106e8610fc67f93c9e7e01b15f65]

NP_005206.1 DCC [Homo sapiens]

List of domain hits

Description	PssmId	Multi-dom	E-value
[–]Ig1_Neogenin[cd05722], First immunoglobulin (Ig)-like domain in neogenin and similar proteins; Ig1_Neogenin: first immunoglobulin (Ig)-like domair	143199	no	6.07e-49

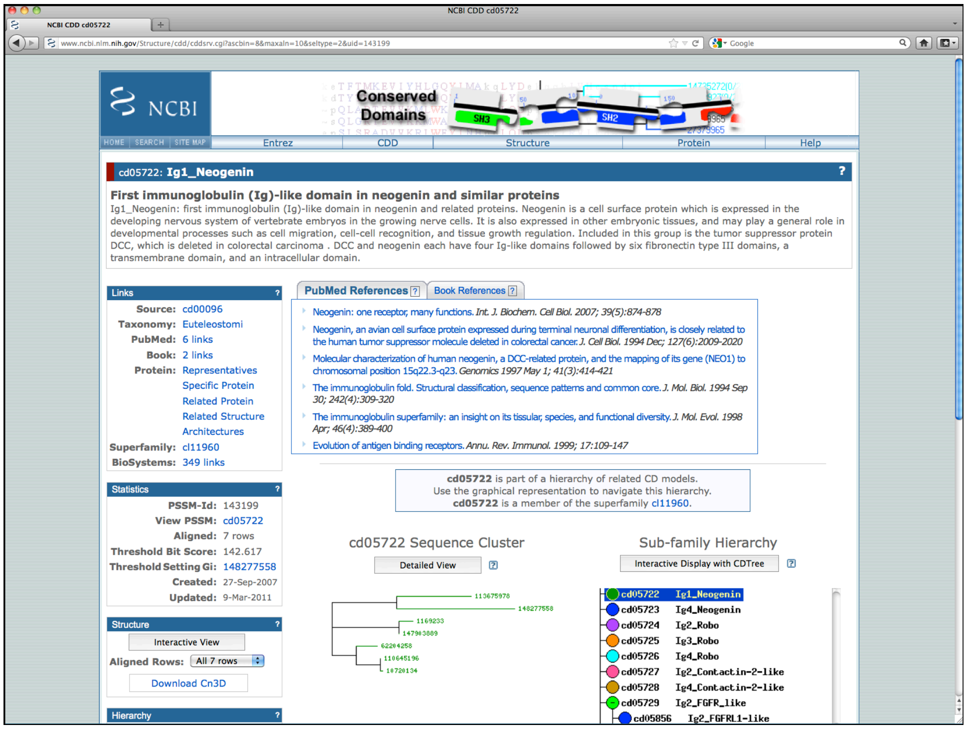
First immunoglobulin (Ig)-like domain in neogenin and similar proteins; Ig1_Neogenin: first immunoglobulin (Ig)-like domain in neogenin and related proteins. Neogenin is a cell surface protein which is expressed in the developing nervous system of vertebrate embryos in the growing nerve cells. It is 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.

:

Cd Length: 95 **Bit Score:** 167.27 **E-value:** 6.07e-49

```
                                              10        20        30        40        50        60        70        80
                                      ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
seqsig_7fe9106e8610fc67f93c9e7e01b15f65   41 RFLSEPSDAVTMRGGNVLLDCSAESDRgVPVIKWKKDGIHLALGMDERKQQLSNGSLLIQNILHSRHHKPDEGLYQCEAS 120
Cdd:cd05722                                1 WFLSEPSDIVAVRGGPVVLNCSAEGEP-PPKIEWKKDGVLLNLVSDERRQQLPNGSLLITSVVHSKHNKPDEGFYQCVAQ 79

                                              90
                                      ....*....|....*.
seqsig_7fe9106e8610fc67f93c9e7e01b15f65  121 LGDSGSIISRTAKVAV 136
Cdd:cd05722                               80 NDSLGSIVSRTARLTV 95
```

Description	PssmId	Multi-dom	E-value
[+]Ig super family[cl11960], Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig superfamily is a ...	209398	no	8.31e-35
[+]FN3[cd00063], Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	28945	no	1.65e-15
[+]FN3[cd00063], Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	28945	no	3.39e-15
[+]FN3[cd00063], Fibronectin type 3 domain; One of three types of internal repeats found in the plasma	28945	no	1.29e-13



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:	cd00096
Taxonomy:	Euteleostomi
PubMed:	6 links
Book:	2 links
Protein:	Representatives Specific Protein Related Protein Related Structure Architectures
Superfamily:	cl11960
BioSystems:	349 links

PubMed References | Book 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*

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.

Statistics

PSSM-Id:	143199
View PSSM:	cd05722
Aligned:	7 rows
Threshold Bit Score:	142.617
Threshold Setting Gi:	148277558
Created:	27-Sep-2007
Updated:	9-Mar-2011

Structure

Interactive View

Aligned Rows: All 7 rows

Download Cn3D

Hierarchy

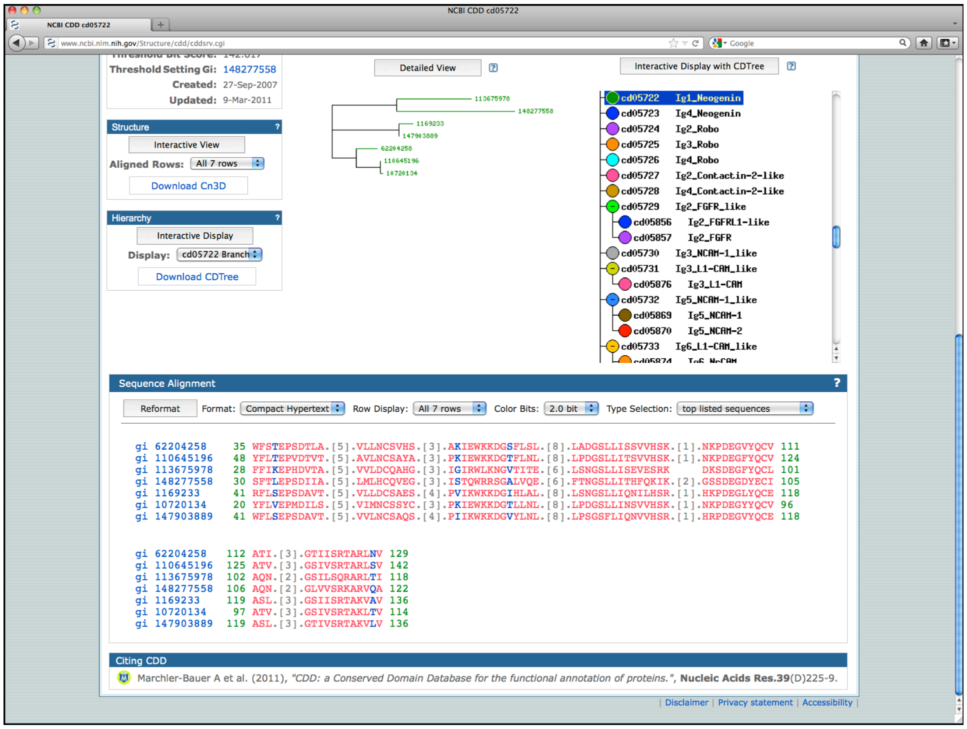
cd05722 Sequence Cluster

Detailed View

Sub-family Hierarchy

Interactive Display with CDTree

```
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
```



Threshold Setting Gi:	148277558
Created:	27-Sep-2007
Updated:	9-Mar-2011

Structure

Interactive View

Aligned Rows: All 7 rows

Download Cn3D

Detailed View

Interactive Display with CDTree

Hierarchy

Interactive Display

Display: cd05722 Branch

Download CDTree

```
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
  cd05857  Ig2_FGFR
cd05730  Ig3_NCAM-1_like
cd05731  Ig3_L1-CAM_like
  cd05876  Ig3_L1-CAM
cd05732  Ig5_NCAM-1_like
  cd05869  Ig5_NCAM-1
  cd05870  Ig5_NCAM-2
cd05733  Ig6_L1-CAM_like
  cd05874  Ig6_NrCAM
```

Sequence Alignment

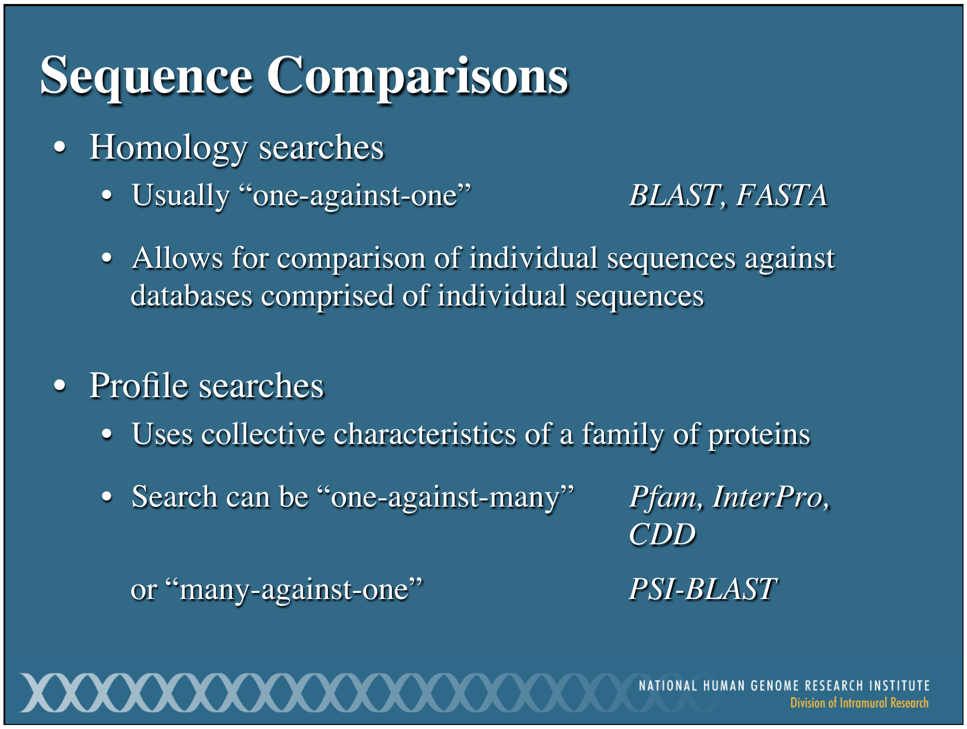
Reformat | Format: Compact Hypertext | Row Display: All 7 rows | Color Bits: 2.0 bit | Type Selection: top listed sequences

```
gi 62204258     35 WFSTEPSDTLA.[5].VLLNCSVHS.[3].AKIEWKKDGSFLSL.[8].LADGSLLISSVVHSK.[1].NKPDEGVYQCV 111
gi 110645196    48 YFLTEPVDTVT.[5].AVLNCSAYA.[3].PKIEWKKDGTFLNL.[8].LPDGSLLITSVVHSK.[1].NKPDEGFYQCV 124
gi 113675978    28 FFIKEPHDVTA.[5].VVLDCQAHG.[3].IGIRWLKNGVTITE.[6].LSNGSLLISEVESRK      DKSDEGFYQCL 101
gi 148277558    30 SFTLEPSDIIA.[5].LMLHCQVEG.[3].ISTQWRRSGALVQE.[6].FTNGSLLITHFQKIK.[2].GSSDEGDYECI 105
gi 1169233      41 RFLSEPSDAVT.[5].VLLDCSAES.[4].PVIKWKKDGIHLAL.[8].LSNGSLLIQNILHSR.[1].HKPDEGLYQCE 118
gi 10720134     20 YFLVEPMDILS.[5].VIMNCSSYC.[3].PKIEWKKDGTLLNL.[8].LPDGSLLINSVVHSK.[1].NKPDEGYYQCV 96
gi 147903889    41 WFLSEPSDAVT.[5].VVLNCSAQS.[4].PIIKWKKDGVYLNL.[8].LPSGSFLIQNVVHSR.[1].HRPDEGVYQCE 118


gi 62204258    112 ATI.[3].GTIISRTARLNV 129
gi 110645196   125 ATV.[3].GSIVSRTARLSV 142
gi 113675978   102 AQN.[2].GSILSQRARLTI 118
gi 148277558   106 AQN.[2].GLVVSRKARVQA 122
gi 1169233     119 ASL.[3].GSIISRTAKVAV 136
gi 10720134     97 ATV.[3].GSIVSRTAKLTV 114
gi 147903889   119 ASL.[3].GTIVSRTAKVLV 136
```

Citing CDD

Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.

| Disclaimer | Privacy statement | Accessibility |

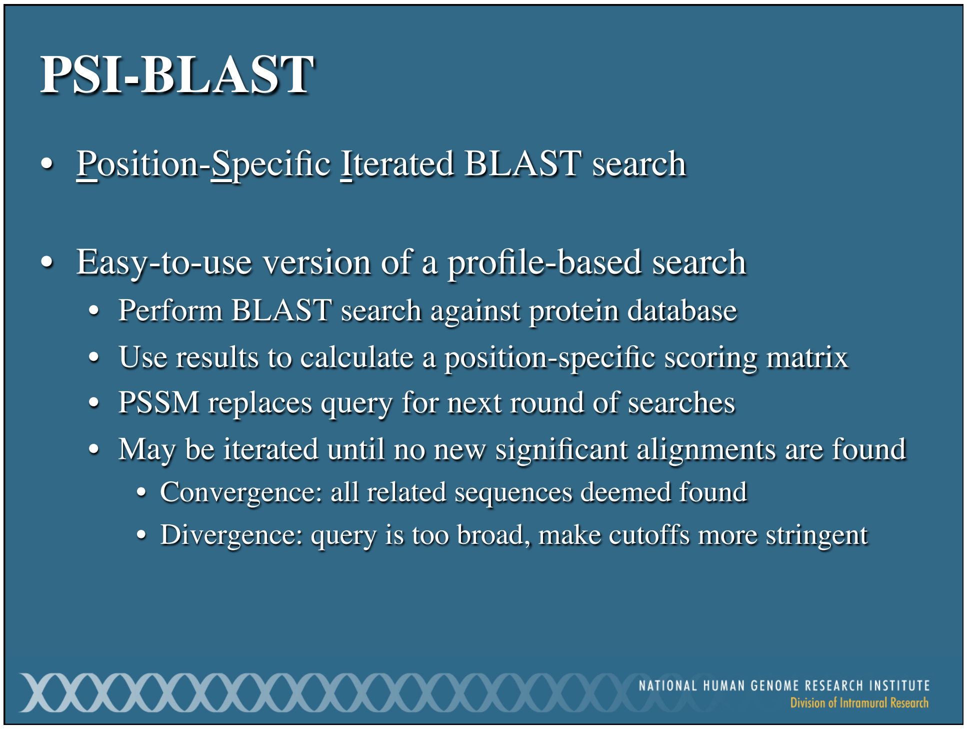


Sequence Comparisons

- 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*

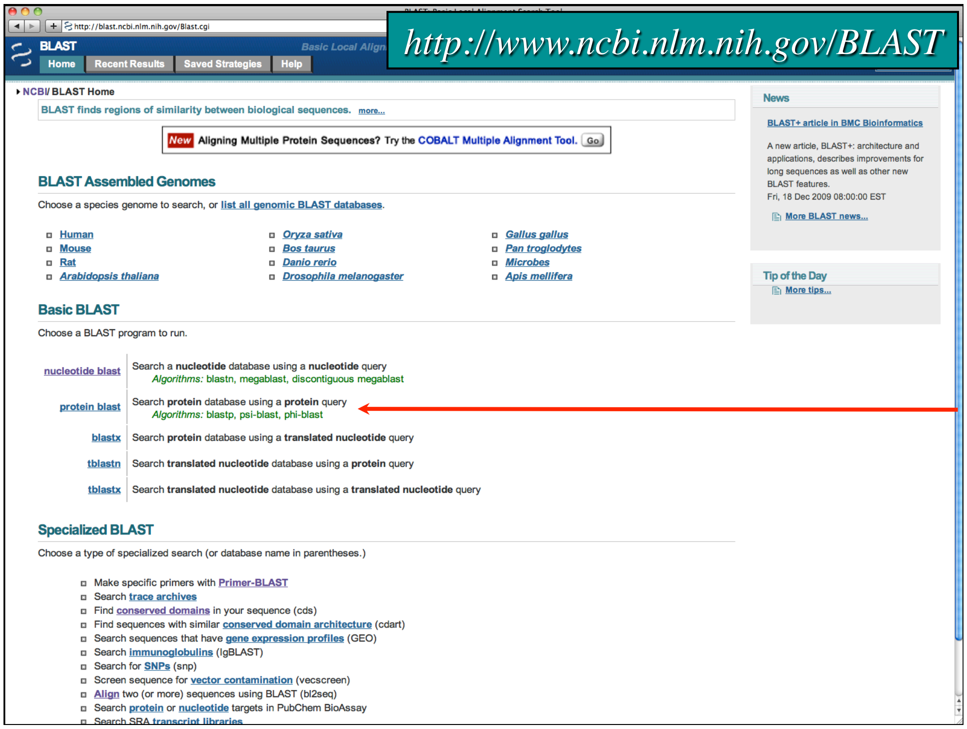
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

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

BLAST — Basic Local Align…

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST Home

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

New Aligning Multiple Protein Sequences? Try the COBALT Multiple Alignment Tool. Go

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

News

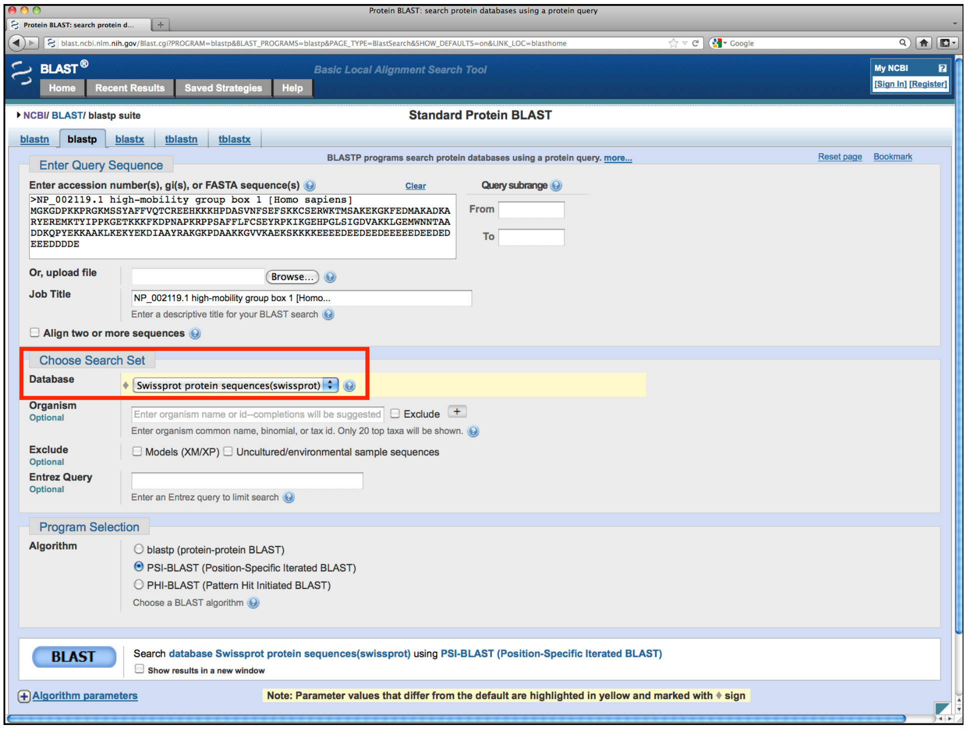
BLAST+ article in BMC Bioinformatics

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...

Tip of the Day

More tips...



Protein BLAST: search protein databases using a protein query

blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome

BLAST® — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help — My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more... — Reset page — Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) — Clear

```
>NP_002119.1 high-mobility group box 1 [Homo sapiens]
MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKFEDMAKADKA
RYEREMKTYIPPKGETKKKFKDPNAPKRPPSAFFLFCSEYRPKIKGEHPGLSIGDVAKKLGEMWNNTAA
DDKQPYEKKAAKLKEKYEKDIAAYRAKGKPDAAKKGVVKAEKSKKKKEEEEDEEDEEDEEEEEDEEDED
EEEDDDDE
```

Query subrange — From: — To:

Or, upload file — Browse...

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) □ Uncultured/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

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

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

Swiss-Prot

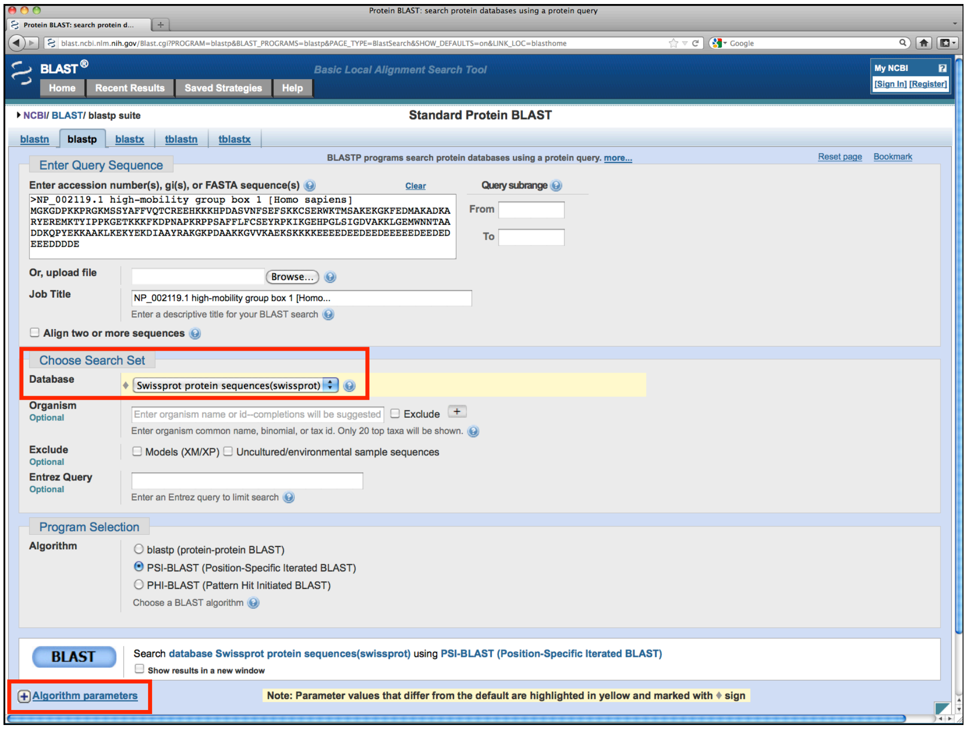
- *Goal:* Provide a single reference sequence for each protein sequence
- Distinguishing Features
  - Non-redundancy
  - Ongoing curation by EBI staff and *external experts*
  - Expert annotation includes editing/updates of

KW	Keyword lines
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
Protein BLAST: search protein d...
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]
NCBI/ BLAST/ blastp suite
blastn
blastp
blastx
tblastn
tblastx
Reset page
Bookmark
Clear
Browse...
BLAST
Algorithm parameters

Standard Protein BLAST

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

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>NP_002119.1 high-mobility group box 1 [Homo sapiens]
MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKFEDMAKADKA
RYEREMKTYIPPKGETKKKFKDPNAPKRPPSAFFLFCSEYRPKIKGEHPGLSIGDVAKKLGEMWNNTAA
DDKQPYEKKAAKLKEKYEKDIAAYRAKGKPDAAKKGVVKAEKSKKKKEEEEDEEDEEDEEEEEDEEDED
EEEDDDDE
```

Query subrange
From
To

Or, upload file

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) □ Uncultured/environmental sample sequences

Entrez Query (Optional)
Enter an Entrez query to limit search

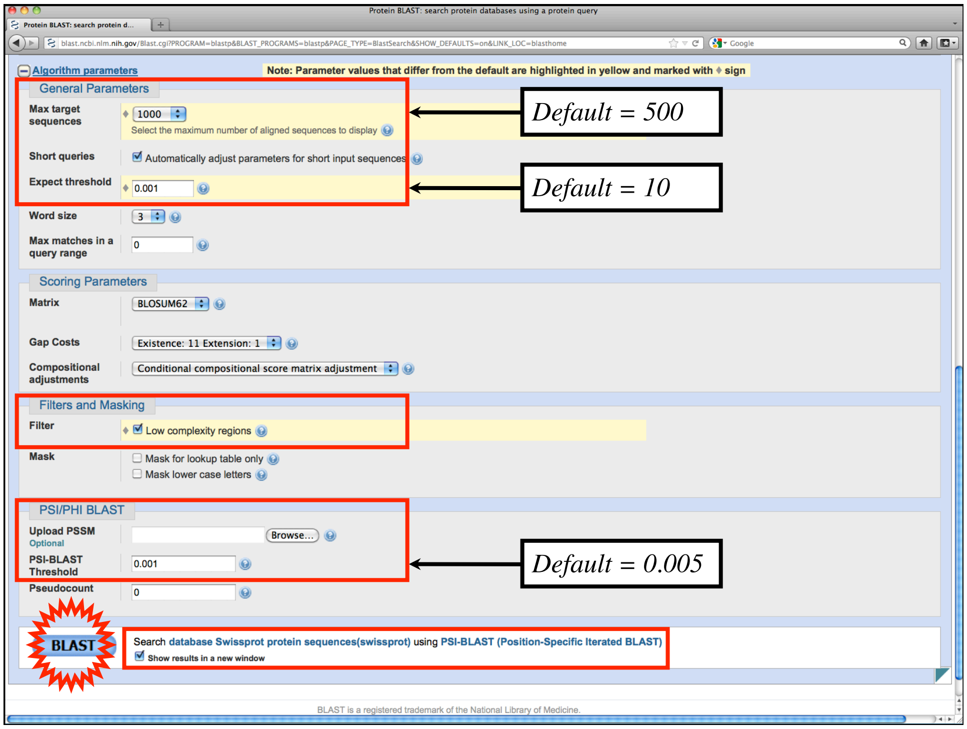
Program Selection

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

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



Protein BLAST: search protein databases using a protein query
Protein BLAST: search protein d...
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome
Google
BLAST

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
- Max matches in a query range: 0

Default = 500

Default = 10

Scoring Parameters

- Matrix: BLOSUM62
- Gap Costs: Existence: 11 Extension: 1
- Compositional adjustments: Conditional compositional score matrix adjustment

Filters and Masking

- Filter: ♦ ☑ Low complexity regions
- Mask: ☐ Mask for lookup table only; ☐ Mask lower case letters

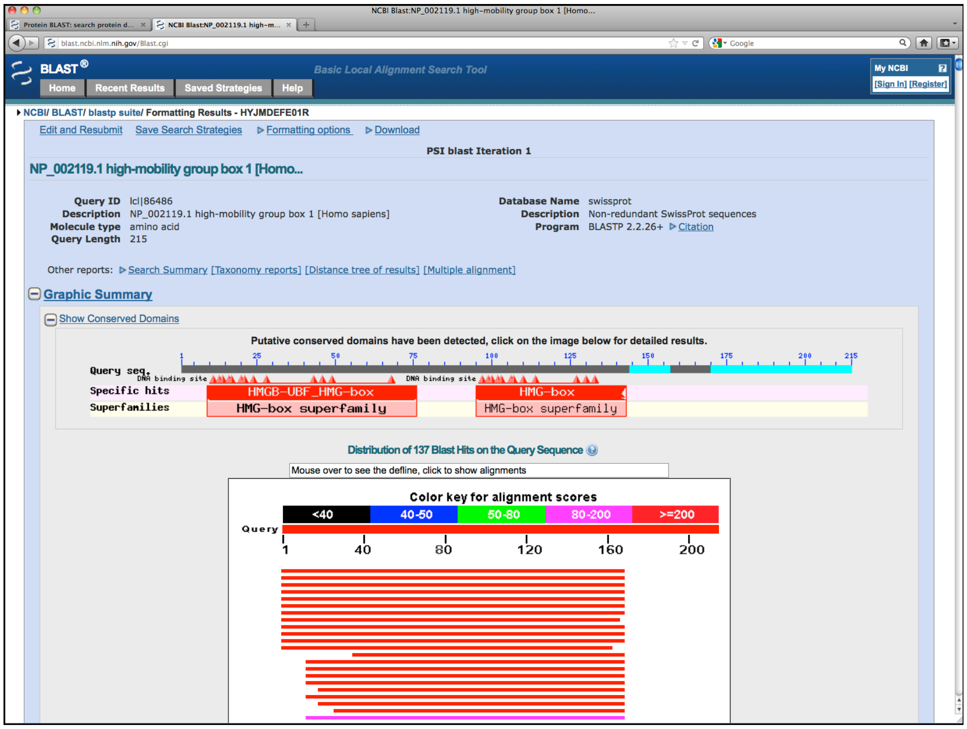
PSI/PHI BLAST

- Upload PSSM (Optional): Browse...
- PSI-BLAST Threshold: 0.001
- Pseudocount: 0

Default = 0.005

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

BLAST is a registered trademark of the National Library of Medicine.



NCBI/ BLAST/ blastp suite/ Formatting Results - HYJMDEFE01R

PSI blast Iteration 1

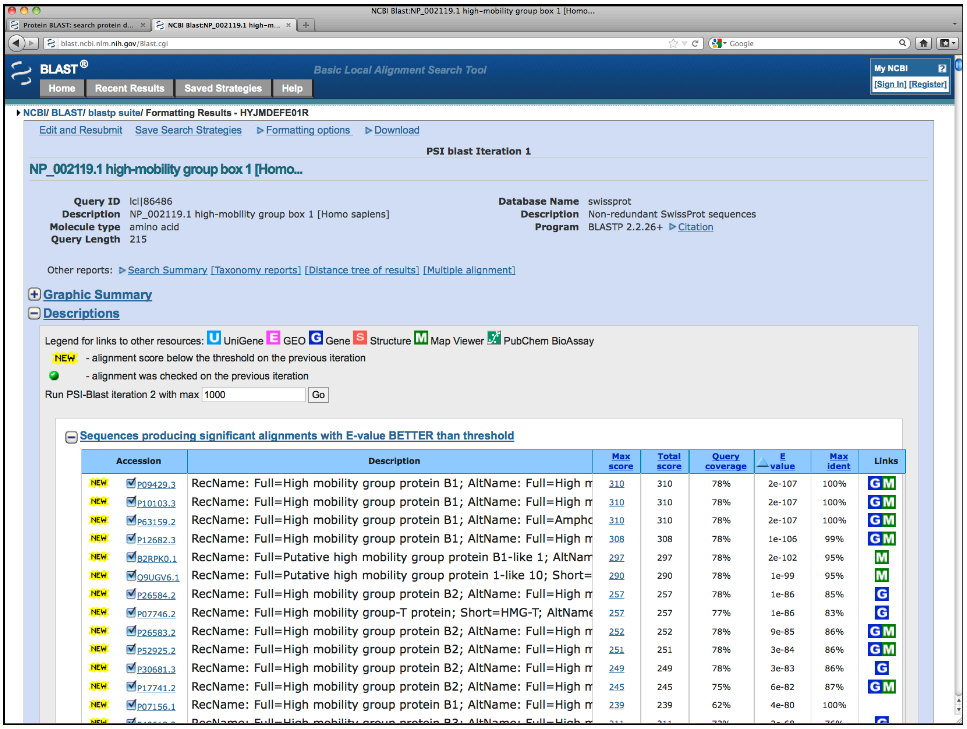
NP_002119.1 high-mobility group box 1 [Homo...

Query ID	lcl|86486	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 2.2.26+ Citation
Query Length	215		

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

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

Distribution of 137 Blast Hits on the Query Sequence


NCBI Blast:NP_002119.1 high-mobility group box 1 [Homo...
Protein BLAST: search protein d...
NCBI Blast:NP_002119.1 high-m...
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 - HYJMDEFE01R

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|86486
Description NP_002119.1 high-mobility group box 1 [Homo sapiens]
Molecule type amino acid
Query Length 215

Database Name swissprot
Description Non-redundant SwissProt sequences
Program BLASTP 2.2.26+ ▷ Citation

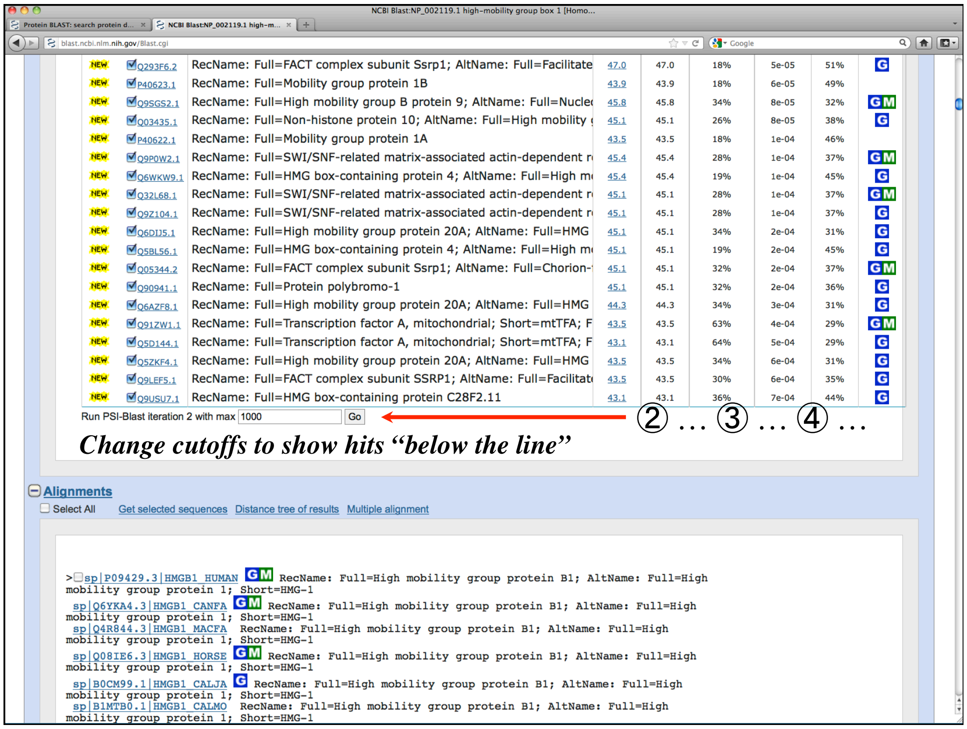
Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

⊞ Graphic Summary
⊟ Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay
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 producing significant alignments with E-value BETTER than threshold

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NEW ☑ P09429.3	RecName: Full=High mobility group protein B1; AltName: Full=High m	310	310	78%	2e-107	100%	G M
NEW ☑ P10103.3	RecName: Full=High mobility group protein B1; AltName: Full=High m	310	310	78%	2e-107	100%	G M
NEW ☑ P63159.2	RecName: Full=High mobility group protein B1; AltName: Full=Amphc	310	310	78%	2e-107	100%	G M
NEW ☑ P12682.3	RecName: Full=High mobility group protein B1; AltName: Full=High m	308	308	78%	1e-106	99%	G M
NEW ☑ B2RPK0.1	RecName: Full=Putative high mobility group protein B1-like 1; AltNam	297	297	78%	2e-102	95%	M
NEW ☑ Q9UGV6.1	RecName: Full=Putative high mobility group protein 1-like 10; Short=	290	290	78%	1e-99	95%	M
NEW ☑ P26584.2	RecName: Full=High mobility group protein B2; AltName: Full=High m	257	257	78%	1e-86	85%	G
NEW ☑ P07746.2	RecName: Full=High mobility group-T protein; Short=HMG-T; AltName	257	257	77%	1e-86	83%	G
NEW ☑ P26583.2	RecName: Full=High mobility group protein B2; AltName: Full=High m	252	252	78%	9e-85	86%	G M
NEW ☑ P52925.2	RecName: Full=High mobility group protein B2; AltName: Full=High m	251	251	78%	3e-84	86%	G M
NEW ☑ P30681.3	RecName: Full=High mobility group protein B2; AltName: Full=High m	249	249	78%	3e-83	86%	G
NEW ☑ P17741.2	RecName: Full=High mobility group protein B2; AltName: Full=High m	245	245	75%	6e-82	87%	G M
NEW ☑ P07156.1	RecName: Full=High mobility group protein B1; AltName: Full=High m	239	239	62%	4e-80	100%	


NCBI Blast:NP_002119.1 high-mobility group box 1 [Homo...
Protein BLAST: search protein d...
NCBI Blast:NP_002119.1 high-m...
blast.ncbi.nlm.nih.gov/Blast.cgi
Google

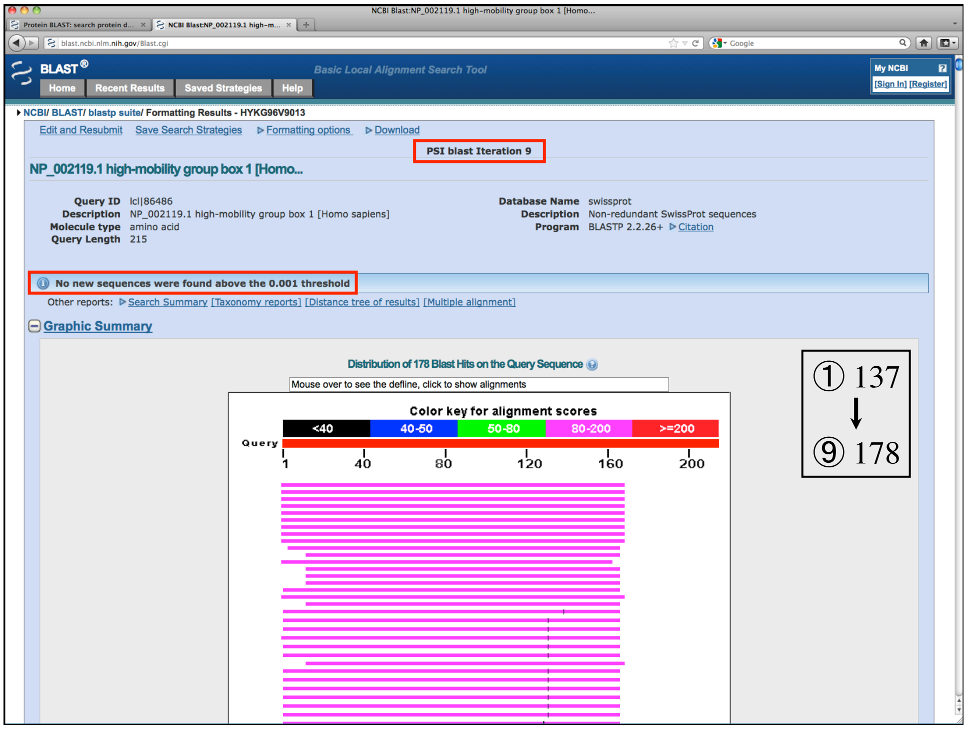
Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
NEW ☑ Q293F6.2	RecName: Full=FACT complex subunit Ssrp1; AltName: Full=Facilitate	47.0	47.0	18%	5e-05	51%	G
NEW ☑ P40623.1	RecName: Full=Mobility group protein 1B	43.9	43.9	18%	6e-05	49%	
NEW ☑ Q9SGS2.1	RecName: Full=High mobility group B protein 9; AltName: Full=Nucle	45.8	45.8	34%	8e-05	32%	G M
NEW ☑ Q03435.1	RecName: Full=Non-histone protein 10; AltName: Full=High mobility	45.1	45.1	26%	8e-05	38%	G
NEW ☑ P40622.1	RecName: Full=Mobility group protein 1A	43.5	43.5	18%	1e-04	46%	
NEW ☑ Q9P0W2.1	RecName: Full=SWI/SNF-related matrix-associated actin-dependent r	45.4	45.4	28%	1e-04	37%	G M
NEW ☑ Q6WKW9.1	RecName: Full=HMG box-containing protein 4; AltName: Full=High m	45.4	45.4	19%	1e-04	45%	G
NEW ☑ Q32L68.1	RecName: Full=SWI/SNF-related matrix-associated actin-dependent r	45.1	45.1	28%	1e-04	37%	G M
NEW ☑ Q9Z104.1	RecName: Full=SWI/SNF-related matrix-associated actin-dependent r	45.1	45.1	28%	1e-04	37%	G
NEW ☑ Q6DIJ5.1	RecName: Full=High mobility group protein 20A; AltName: Full=HMG	45.1	45.1	34%	2e-04	31%	G
NEW ☑ Q5BL56.1	RecName: Full=HMG box-containing protein 4; AltName: Full=High m	45.1	45.1	19%	2e-04	45%	G
NEW ☑ Q05344.2	RecName: Full=FACT complex subunit Ssrp1; AltName: Full=Chorion-	45.1	45.1	32%	2e-04	37%	G M
NEW ☑ Q90941.1	RecName: Full=Protein polybromo-1	45.1	45.1	32%	2e-04	36%	G
NEW ☑ Q6AZF8.1	RecName: Full=High mobility group protein 20A; AltName: Full=HMG	44.3	44.3	34%	3e-04	31%	G
NEW ☑ Q91ZW1.1	RecName: Full=Transcription factor A, mitochondrial; Short=mtTFA; F	43.5	43.5	63%	4e-04	29%	G M
NEW ☑ Q5D144.1	RecName: Full=Transcription factor A, mitochondrial; Short=mtTFA; F	43.1	43.1	64%	5e-04	29%	G
NEW ☑ Q5ZKF4.1	RecName: Full=High mobility group protein 20A; AltName: Full=HMG	43.5	43.5	34%	6e-04	31%	G
NEW ☑ Q9LEF5.1	RecName: Full=FACT complex subunit SSRP1; AltName: Full=Facilitat	43.5	43.5	30%	6e-04	35%	G
NEW ☑ Q9USU7.1	RecName: Full=HMG box-containing protein C28F2.11	43.1	43.1	36%	7e-04	44%	G

Run PSI-Blast iteration 2 with max 1000 Go ← ② … ③ … ④ …

Change cutoffs to show hits "below the line"

⊟ Alignments
☐ Select All   Get selected sequences   Distance tree of results   Multiple alignment

```
>☐sp|P09429.3|HMGB1_HUMAN GM RecName: Full=High mobility group protein B1; AltName: Full=High
mobility group protein 1; Short=HMG-1
 sp|Q6YKA4.3|HMGB1_CANFA GM 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 GM RecName: Full=High mobility group protein B1; AltName: Full=High
mobility group protein 1; Short=HMG-1
 sp|B0CM99.1|HMGB1_CALJA G 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
```


NCBI Blast:NP_002119.1 high-mobility group box 1 [Homo...
Protein BLAST: search protein d...
NCBI Blast:NP_002119.1 high-m...
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 - HYKG96V9013

Edit and Resubmit Save Search Strategies ▷ Formatting options ▷ Download

PSI blast Iteration 9

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

Query ID	lcl|86486	**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 2.2.26+ ▷ Citation
Query Length	215		

ⓘ No new sequences were found above the 0.001 threshold

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

⊟ Graphic Summary

Distribution of 178 Blast Hits on the Query Sequence

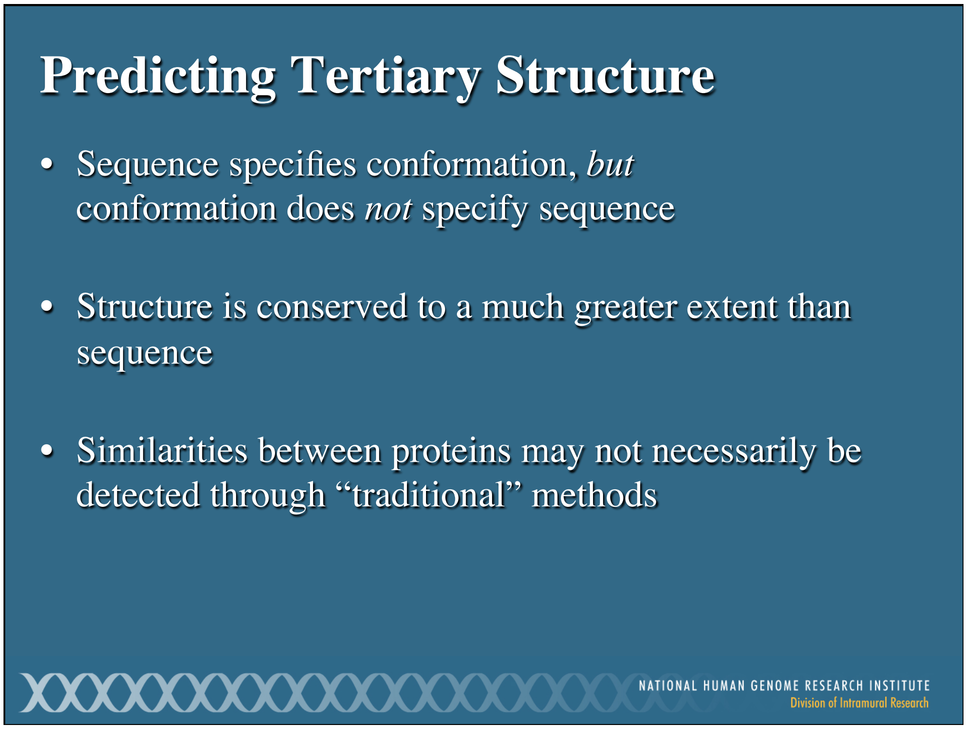
Mouse over to see the defline, click to show alignments

① 137
↓
⑨ 178

Overview

- Week 2
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 4
  - Profiles, Patterns, Motifs, and Domains
  - Structures: VAST, Cn3D, and *de novo* Prediction
  - Multiple Sequence Alignment

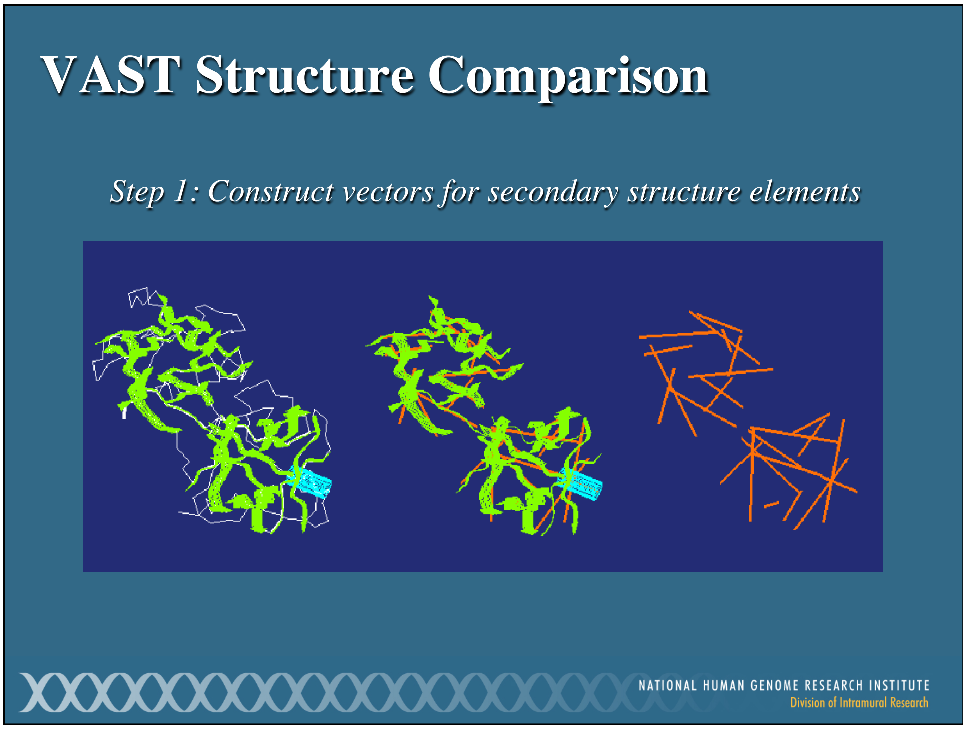
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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



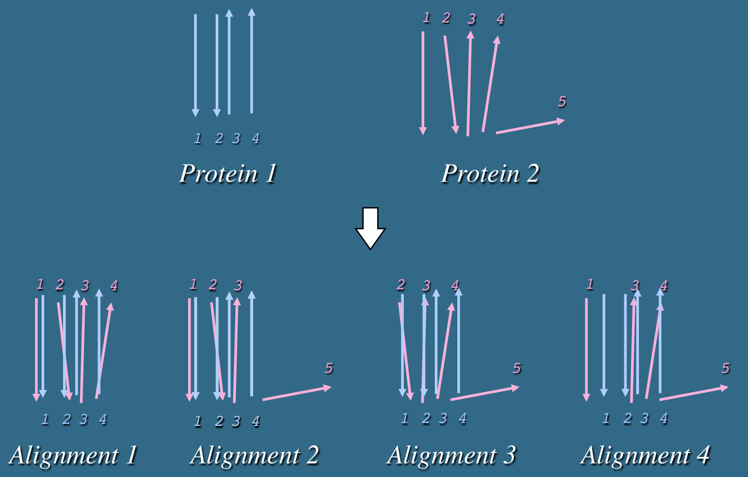
VAST Structure Comparison

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



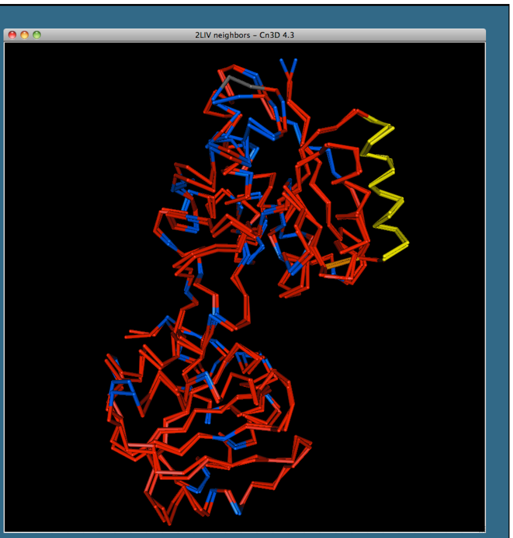
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

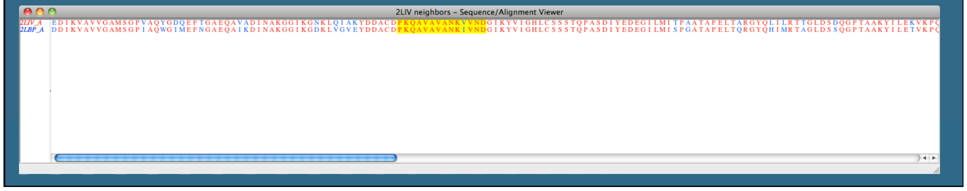
Cn3D Viewer
2LIV vs. 2LBP

Rendering: Tubes

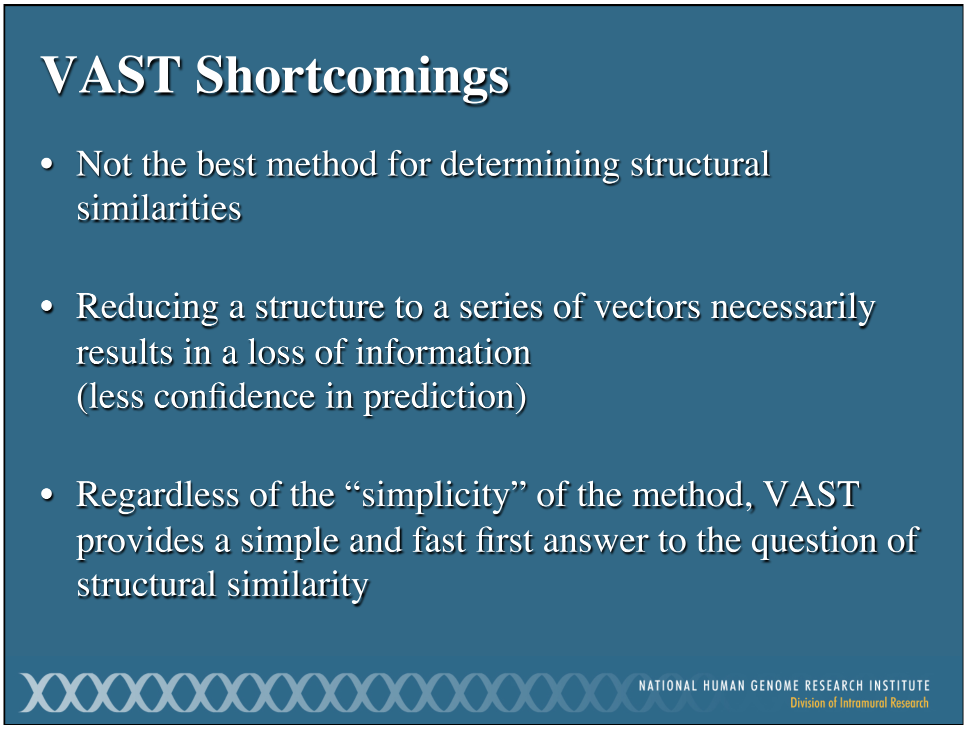
Coloring: Identity

Red	*matches*
Blue	*mismatches*
Yellow	*highlighted*





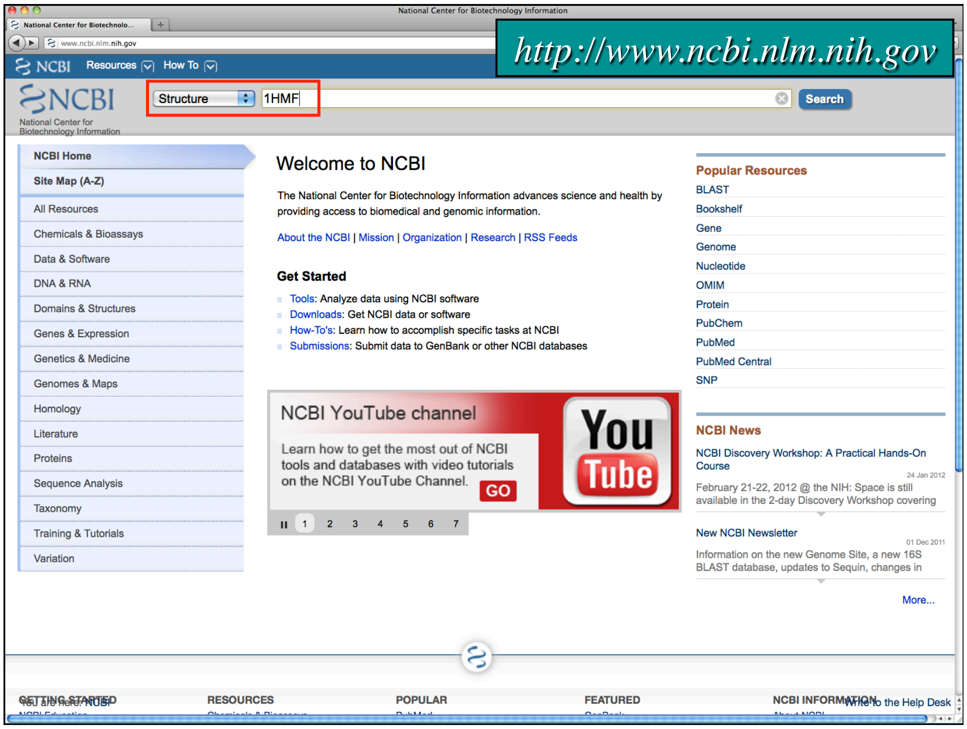
```
2LIV_A  E D I K V A V V G A M S G P V A Q Y G D Q E F T G A E Q A V A D I N A K G G I K G N K L Q I A K Y D D A C D P K Q A V A V A N K V V N D G I K Y V I G H L C S S S T Q P A S D I Y E D E G I L M I T P A A T A P E L T A R G Y Q L I L R T T G L D S D Q G P T A A K Y I L E K V K P Q
2LBP_A  D D I K V A V V G A M S G P I A Q W G I M E F N G A E Q A I K D I N A K G G I K G D K L V G V E Y D D A C D P K Q A V A V A N K I V N D G I K Y V I G H L C S S S T Q P A S D I Y E D E G I L M I S P G A T A P E L T Q R G Y Q H I M R T A G L D S S Q G P T A A K Y I L E T V K P Q
```



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, VAST provides a simple and fast first answer to the question of structural similarity

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

Welcome to NCBI

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

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

NCBI YouTube channel

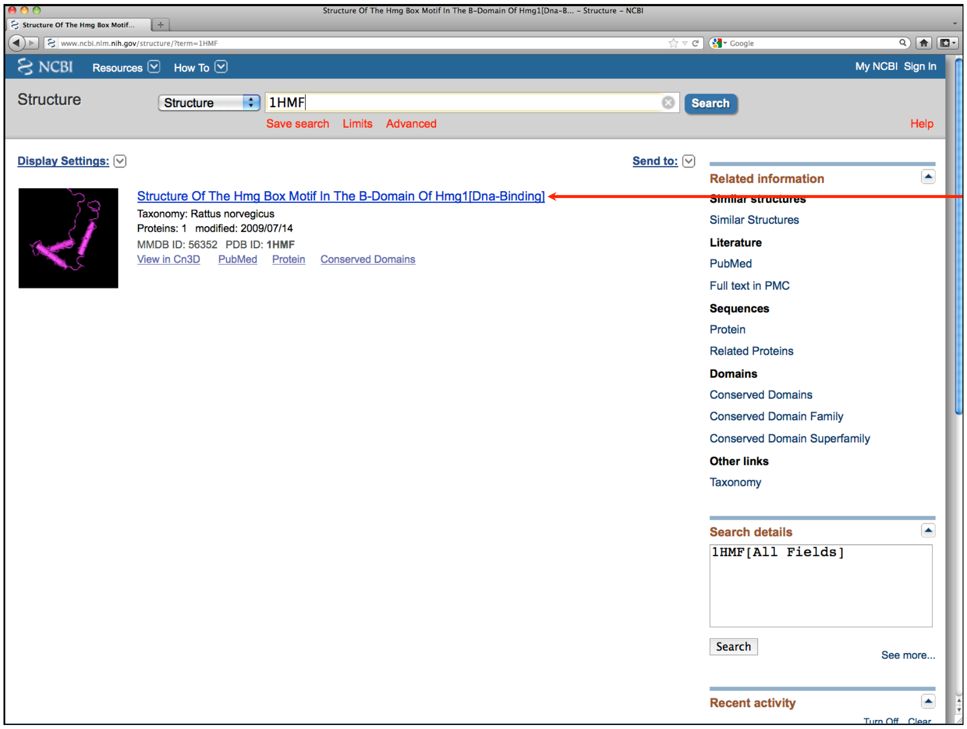
Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel.

NCBI News

NCBI Discovery Workshop: A Practical Hands-On Course
24 Jan 2012
February 21-22, 2012 @ the NIH: Space is still available in the 2-day Discovery Workshop covering

New NCBI Newsletter
01 Dec 2011
Information on the new Genome Site, a new 16S BLAST database, updates to Sequin, changes in

More...



Structure: 1HMF

Structure Of The Hmg Box Motif In The B-Domain Of Hmg1[Dna-Binding]
Taxonomy: Rattus norvegicus
Proteins: 1 modified: 2009/07/14
MMDB ID: 56352 PDB ID: **1HMF**

Related information

Similar structures
Similar Structures

Literature
PubMed
Full text in PMC

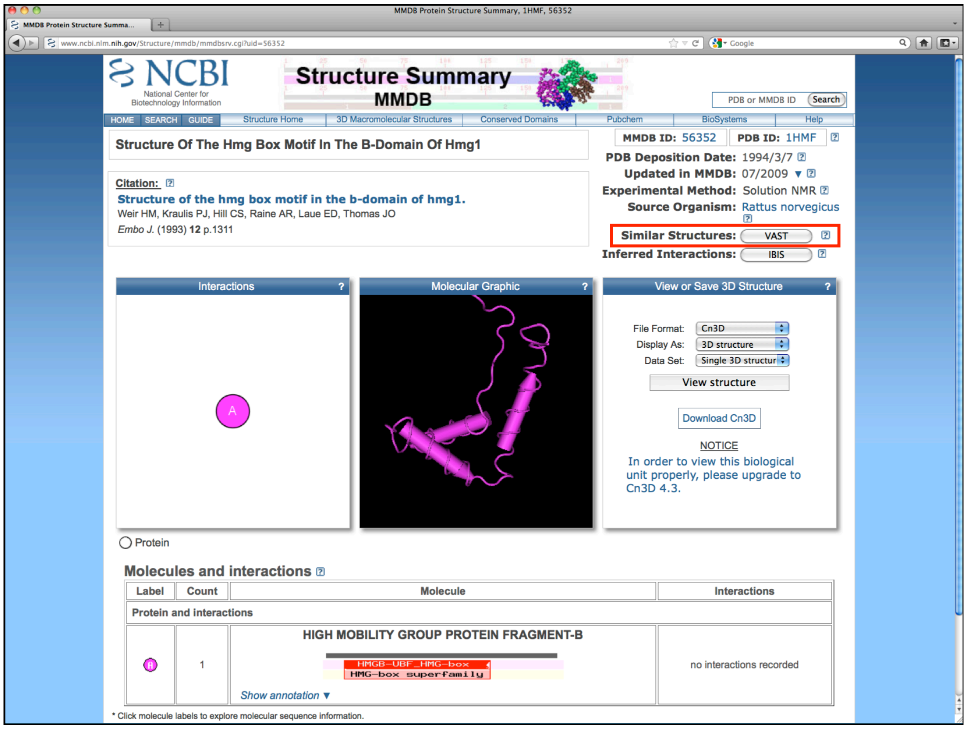
Sequences
Protein
Related Proteins

Domains
Conserved Domains
Conserved Domain Family
Conserved Domain Superfamily

Other links
Taxonomy

Search details

```
1HMF[All Fields]
```


MMDB Protein Structure Summary, 1HMF, 56352
MMDB Protein Structure Summa...
www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?uid=56352
Google
NCBI
National Center for Biotechnology Information
Structure Summary
MMDB
PDB or MMDB ID
Search
HOME
SEARCH
GUIDE
Structure Home
3D Macromolecular Structures
Conserved Domains
Pubchem
BioSystems
Help
Interactions
Molecular Graphic
View or Save 3D Structure
A
View structure
Download Cn3D
Show annotation ▼

Structure Of The Hmg Box Motif In The B-Domain Of Hmg1

Citation:
Structure of the hmg box motif in the b-domain of hmg1.
Weir HM, Kraulis PJ, Hill CS, Raine AR, Laue ED, Thomas JO
Embo J. (1993) **12** p.1311

MMDB ID: 56352 **PDB ID:** 1HMF
PDB Deposition Date: 1994/3/7
Updated in MMDB: 07/2009
Experimental Method: Solution NMR
Source Organism: Rattus norvegicus
Similar Structures: VAST
Inferred Interactions: IBIS

File Format: Cn3D
Display As: 3D structure
Data Set: Single 3D structur

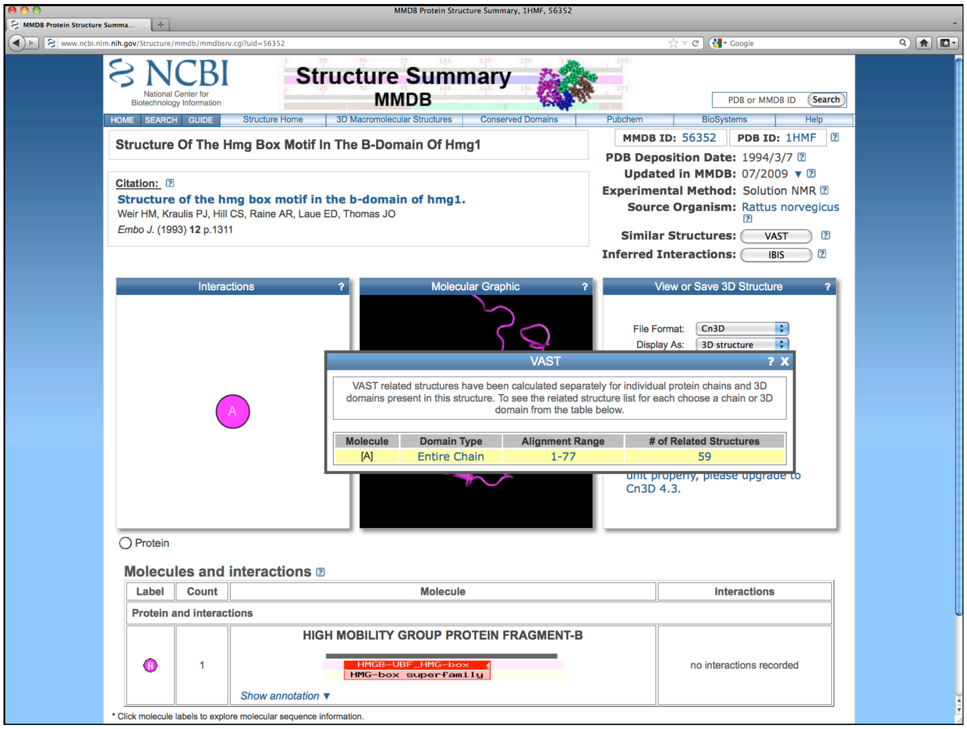
NOTICE
In order to view this biological unit properly, please upgrade to Cn3D 4.3.

○ Protein

Molecules and interactions

Label	Count	Molecule	Interactions
Protein and interactions			
A	1	HIGH MOBILITY GROUP PROTEIN FRAGMENT-B HMGB-UBF_HMG-box HMG-box superfamily	no interactions recorded

* Click molecule labels to explore molecular sequence information.



MMDB Protein Structure Summary, 1HMF, 56352

www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?uid=56352

Structure Of The Hmg Box Motif In The B-Domain Of Hmg1

MMDB ID: 56352 **PDB ID:** 1HMF

PDB Deposition Date: 1994/3/7
Updated in MMDB: 07/2009
Experimental Method: Solution NMR
Source Organism: Rattus norvegicus
Similar Structures: VAST
Inferred Interactions: IBIS

Citation:
Structure of the hmg box motif in the b-domain of hmg1.
Weir HM, Kraulis PJ, Hill CS, Raine AR, Laue ED, Thomas JO
Embo J. (1993) **12** p.1311

File Format: Cn3D
Display As: 3D structure

unit properly, please upgrade to Cn3D 4.3.

VAST

VAST related structures have been calculated separately for individual protein chains and 3D domains present in this structure. To see the related structure list for each choose a chain or 3D domain from the table below.

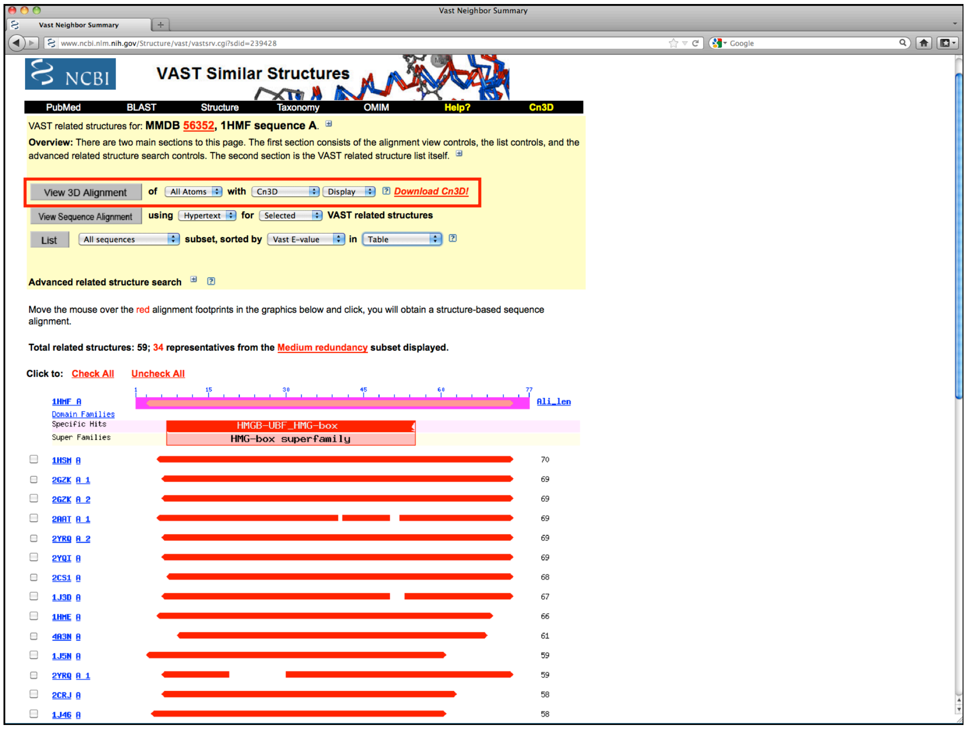
Molecule	Domain Type	Alignment Range	# of Related Structures
[A]	Entire Chain	1-77	59

○ Protein

Molecules and interactions

Label	Count	Molecule	Interactions
Protein and interactions			
A	1	HIGH MOBILITY GROUP PROTEIN FRAGMENT-B HMGB-UBF_HMG-box HMG-box superfamily Show annotation ▼	no interactions recorded

* Click molecule labels to explore molecular sequence information.



Vast Neighbor Summary

www.ncbi.nlm.nih.gov/Structure/vast/vastsrv.cgi?sdid=239428

VAST related structures for: **MMDB 56352, 1HMF 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

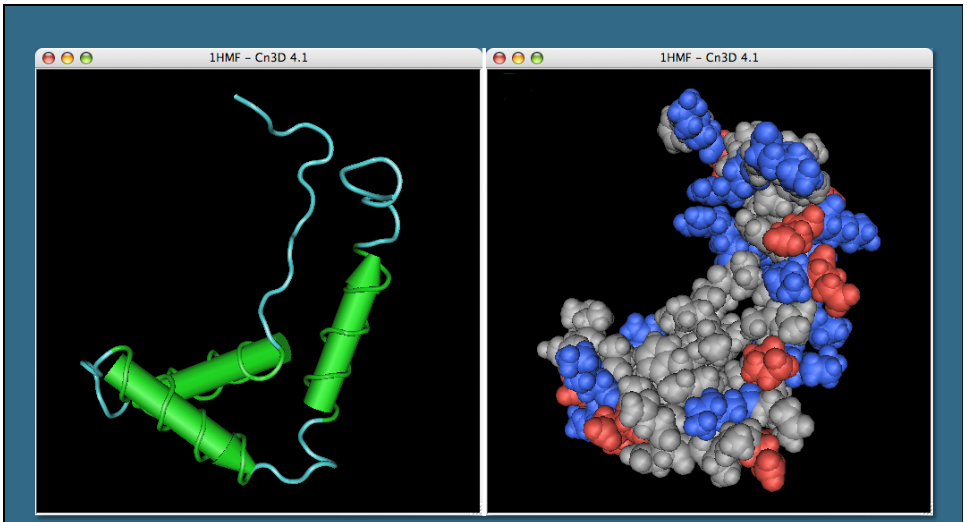
Advanced related structure search

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: 59; 34 representatives from the Medium redundancy subset displayed.

Click to: Check All Uncheck All

	Ali_len
□ 1HSM A	70
□ 2GZK A 1	69
□ 2GZK A 2	69
□ 2AAT A 1	69
□ 2YRQ A 2	69
□ 2YQI A	69
□ 2CS1 A	68
□ 1J3D A	67
□ 1HME A	66
□ 4A3N A	61
□ 1J5N A	59
□ 2YRQ A 1	59
□ 2CRJ A	58
□ 1J46 A	58

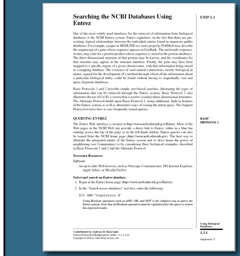


Worms — ***RENDERING*** — *Space Fill*

Secondary Structure — ***COLORING*** — *Charge*

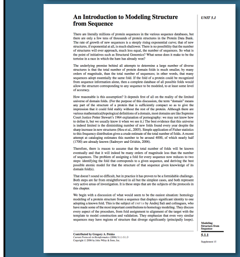
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Further Reading



Searching the NCBI Databases Using Entrez UNIT 1.3

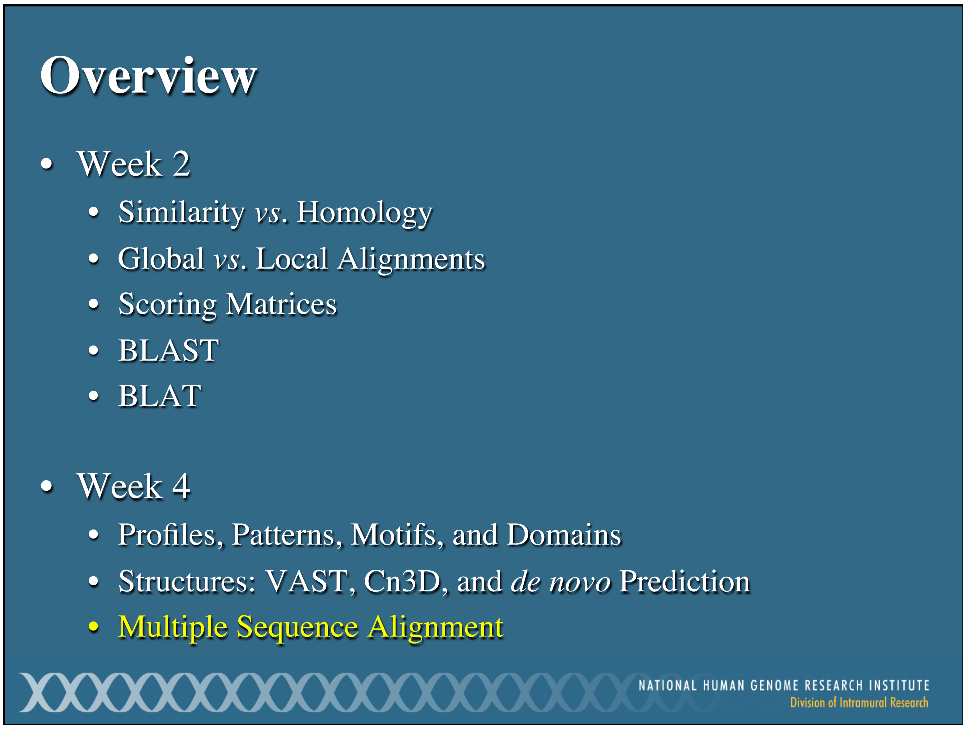
Current Protocols in Bioinformatics
Unit 1.3
Entrez and Cn3D



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

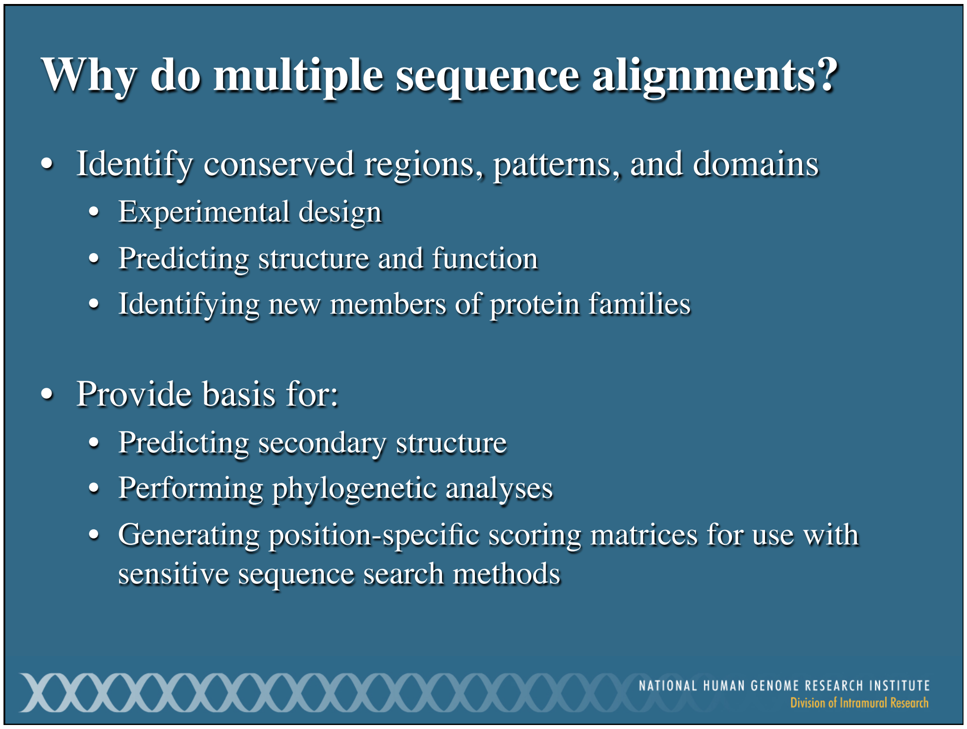


Overview

- Week 2
  - Similarity *vs*. Homology
  - Global *vs*. Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT

- Week 4
  - Profiles, Patterns, Motifs, and Domains
  - Structures: VAST, Cn3D, and *de novo* Prediction
  - Multiple Sequence Alignment

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

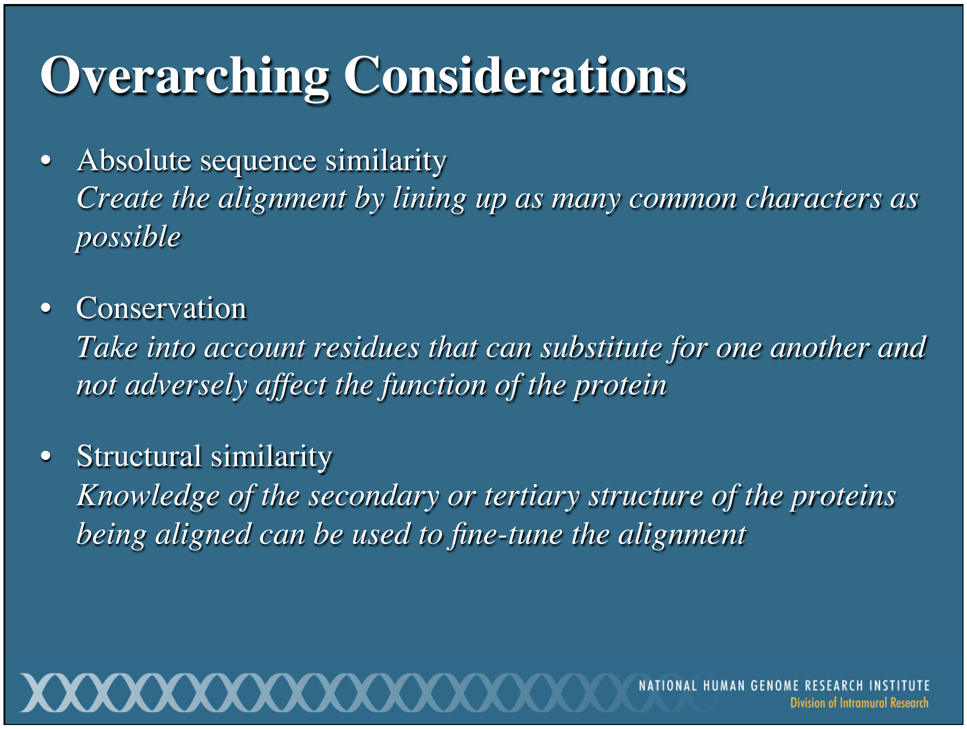


Why do multiple sequence alignments?

- Identify conserved regions, patterns, and domains
  - Experimental design
  - Predicting structure and function
  - Identifying new members of protein families

- Provide basis for:
  - Predicting secondary structure
  - Performing phylogenetic analyses
  - Generating position-specific scoring matrices for use with sensitive sequence search methods

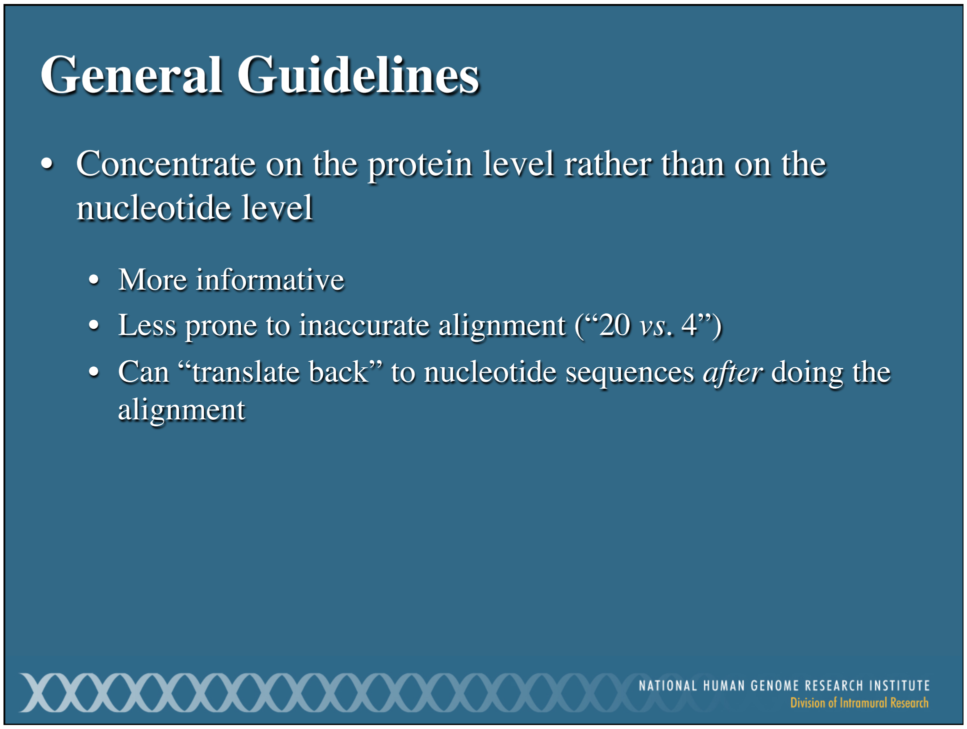
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Overarching 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

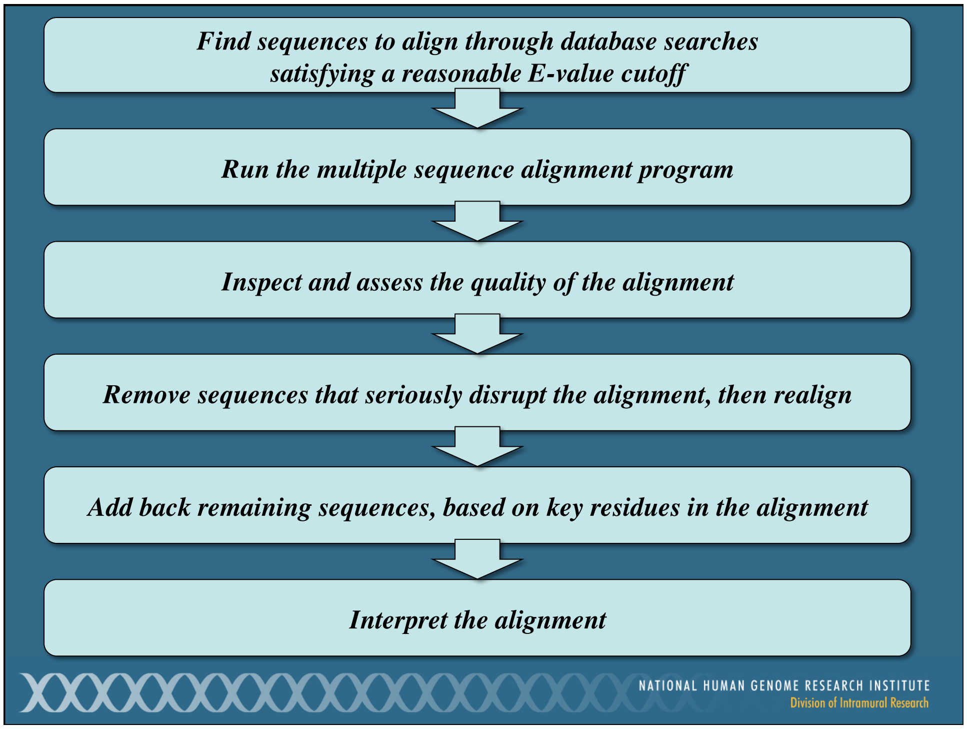
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



General Guidelines

- 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



Find sequences to align through database searches satisfying a reasonable E-value cutoff

Run the multiple sequence alignment program

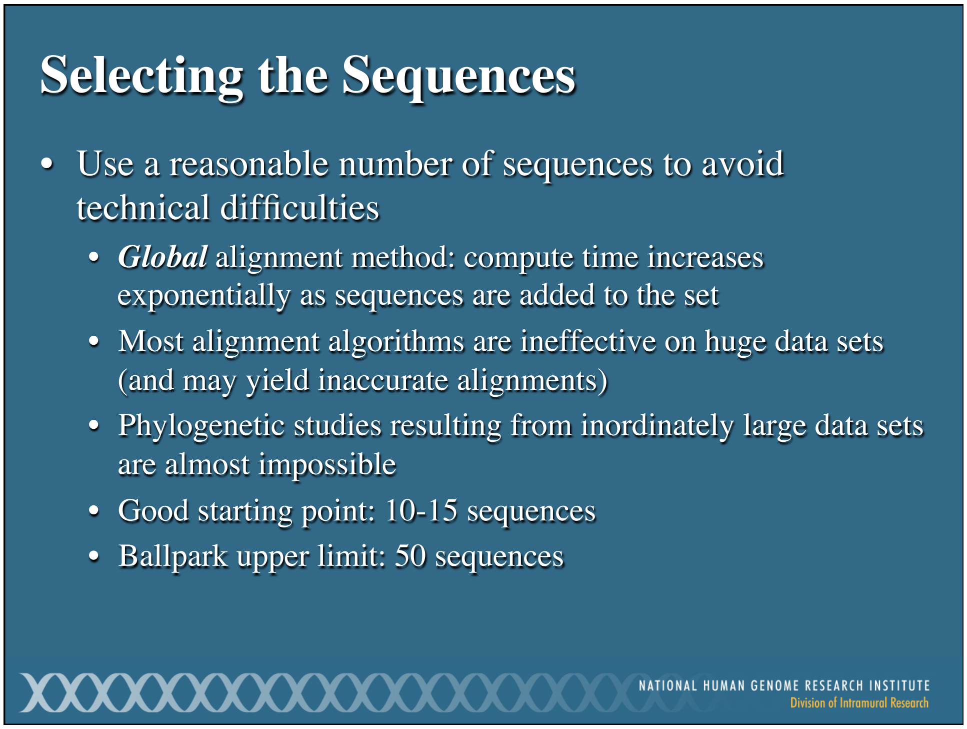
Inspect and assess the quality of the alignment

Remove sequences that seriously disrupt the alignment, then realign

Add back remaining sequences, based on key residues in the alignment

Interpret the alignment

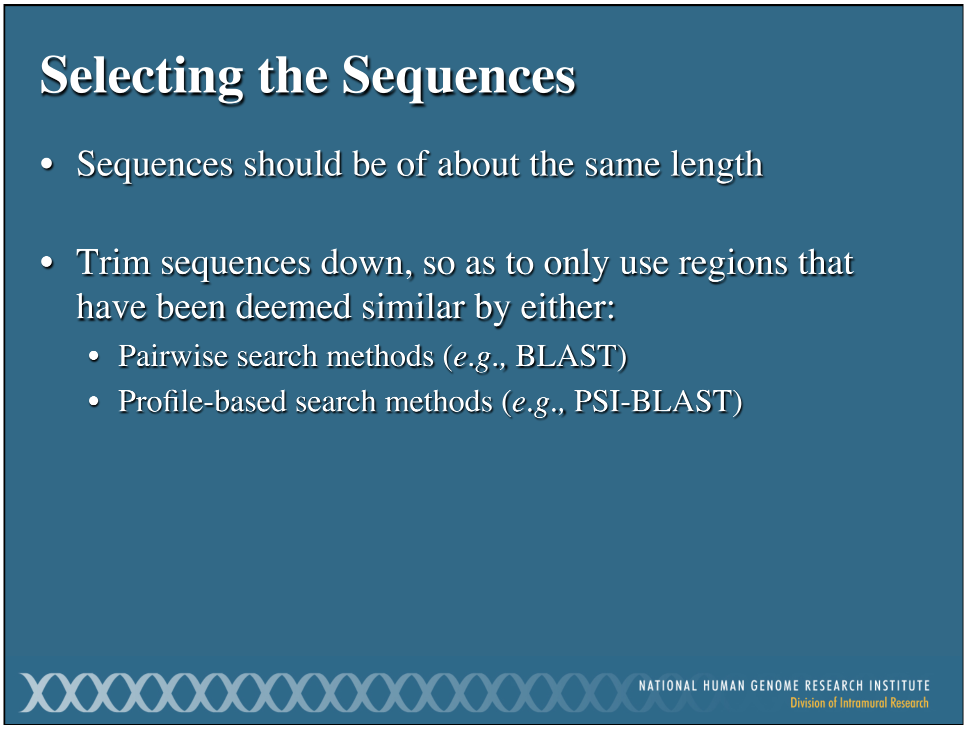
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Selecting the Sequences

- 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

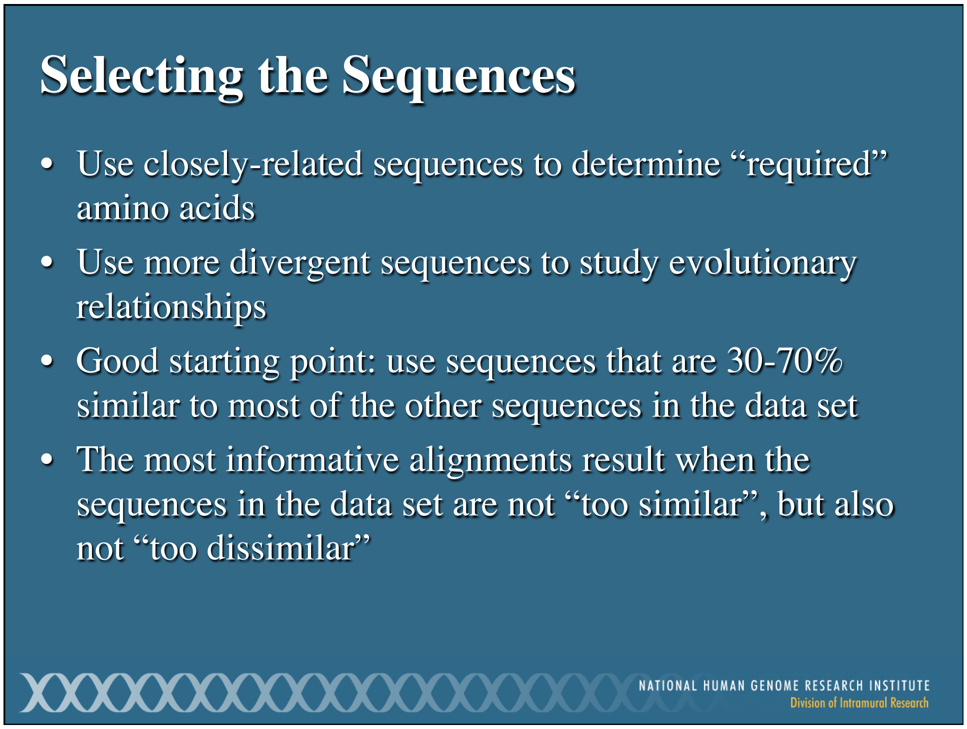
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Selecting the Sequences

- Sequences should be of about the same length
- Trim sequences down, so as to only use regions that have been deemed similar by either:
  - Pairwise search methods (*e.g.*, BLAST)
  - Profile-based search methods (*e.g.*, PSI-BLAST)

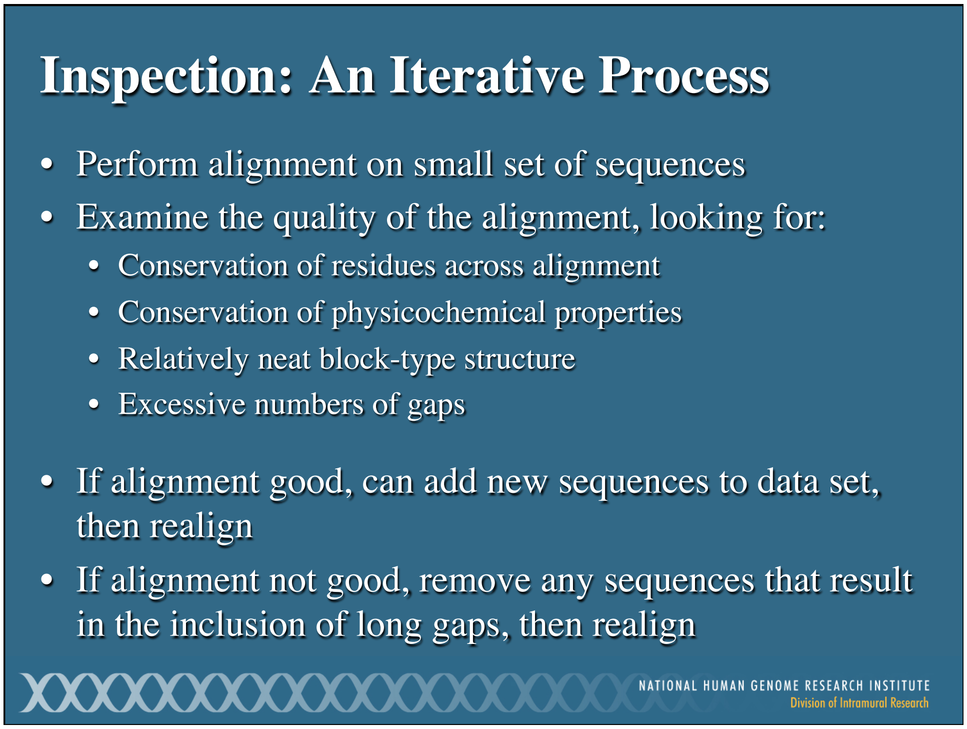
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Selecting the Sequences

- 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 dissimilar”

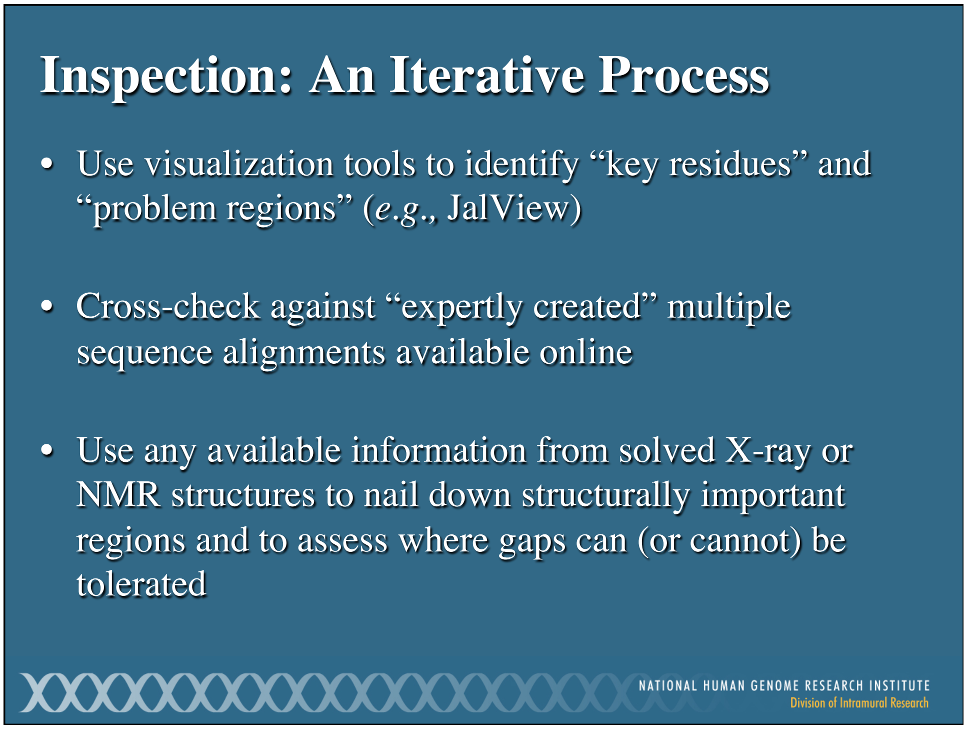
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Inspection: An Iterative Process

- Perform alignment on small set of sequences
- Examine the quality of the alignment, looking for:
  - Conservation of residues across alignment
  - Conservation of physicochemical properties
  - Relatively neat block-type structure
  - Excessive numbers of gaps
- 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

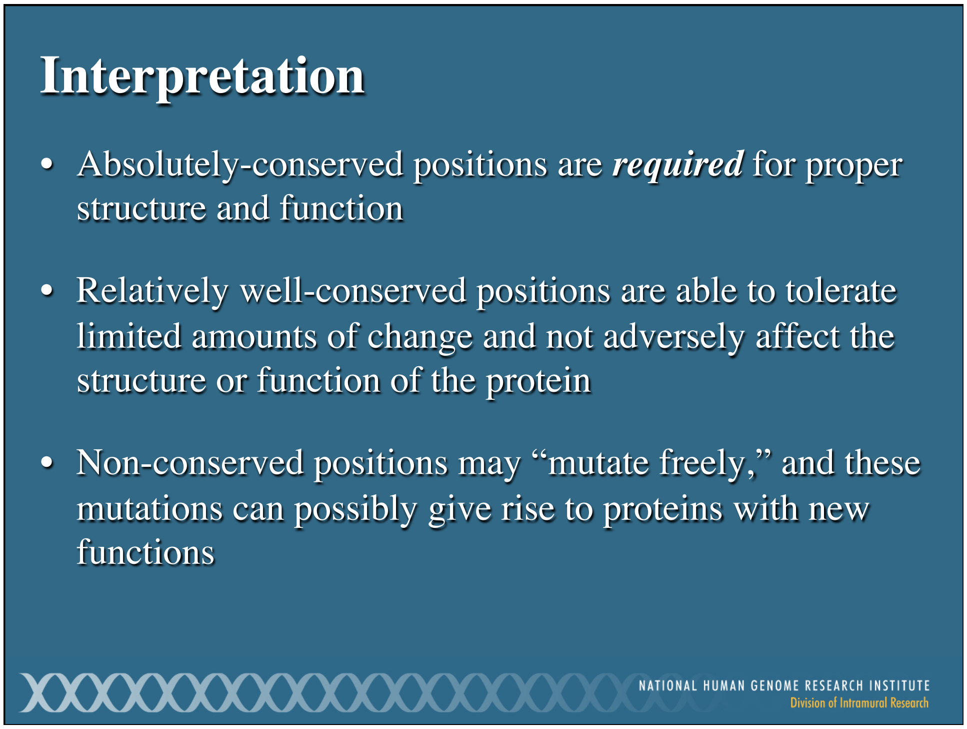
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Inspection: An Iterative Process

- Use visualization tools to identify "key residues" and "problem regions" (*e.g.*, JalView)
- Cross-check against "expertly created" multiple sequence alignments available online
- Use any available information from solved X-ray or NMR structures to nail down structurally important regions and to assess where gaps can (or cannot) be tolerated

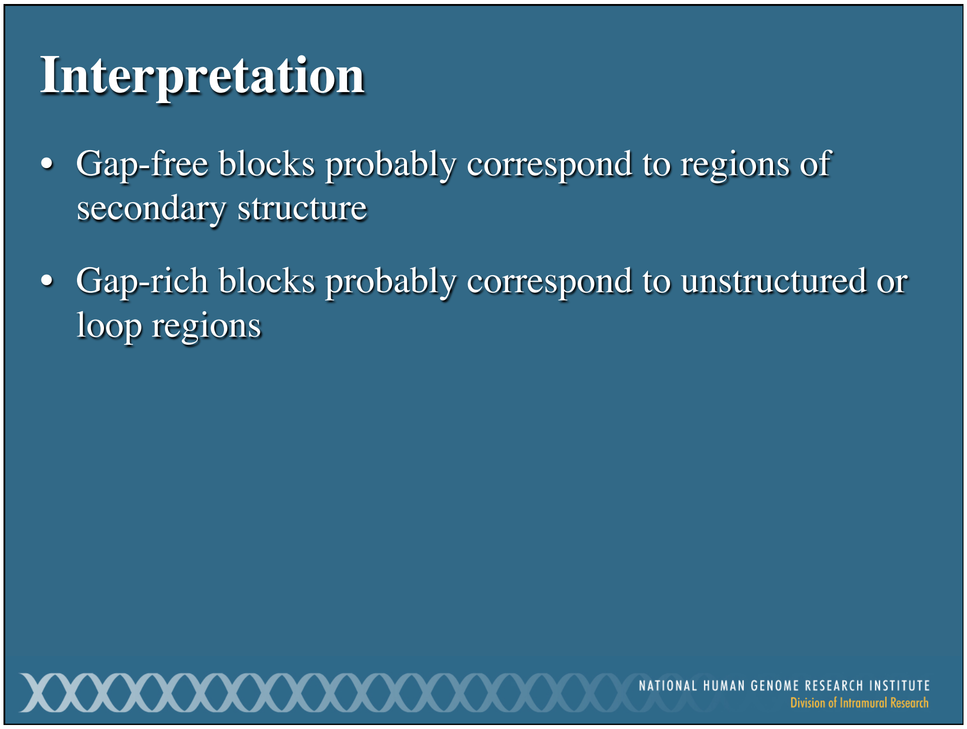
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

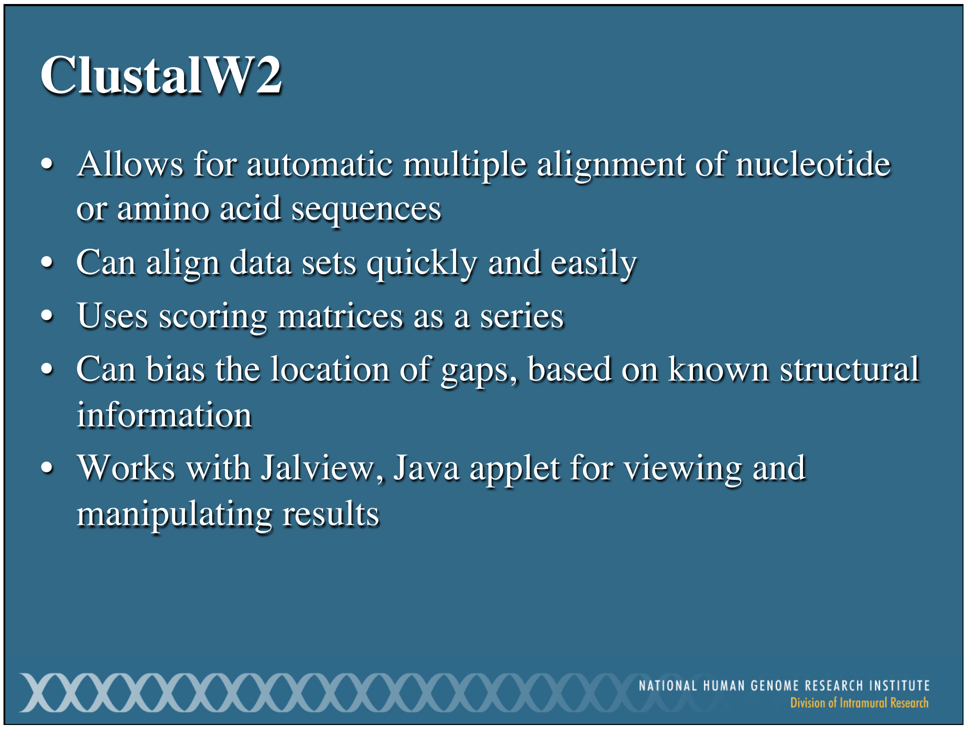
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Interpretation

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

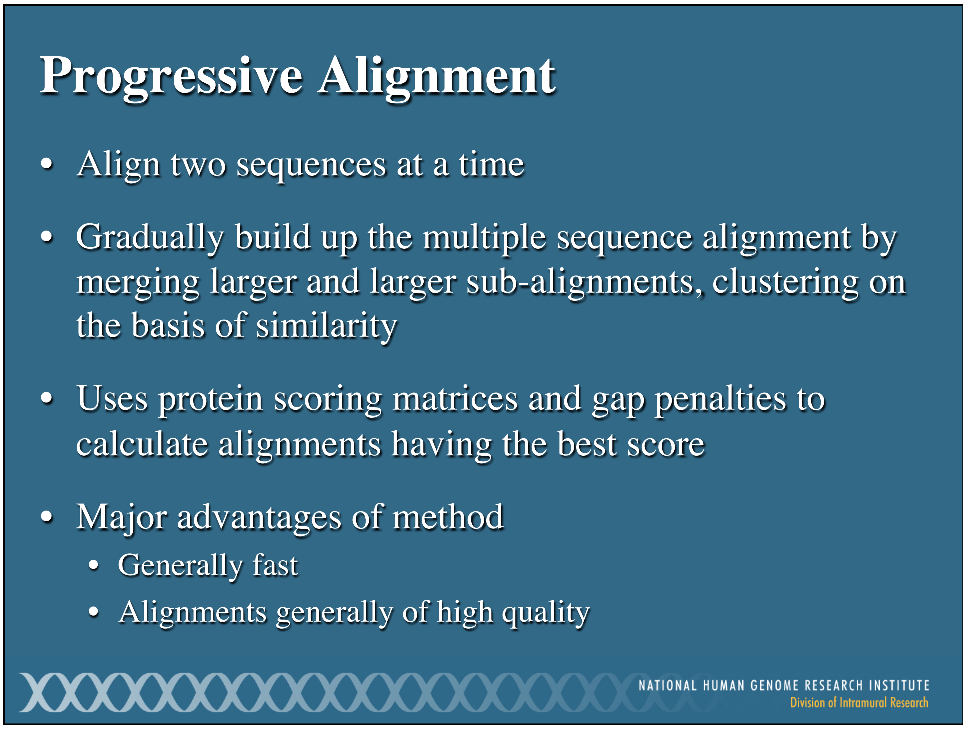
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



ClustalW2

- Allows for automatic multiple alignment of nucleotide or amino acid sequences
- Can align data sets quickly and easily
- Uses scoring matrices as a series
- Can bias the location of gaps, based on known structural information
- Works with Jalview, Java applet for viewing and manipulating results

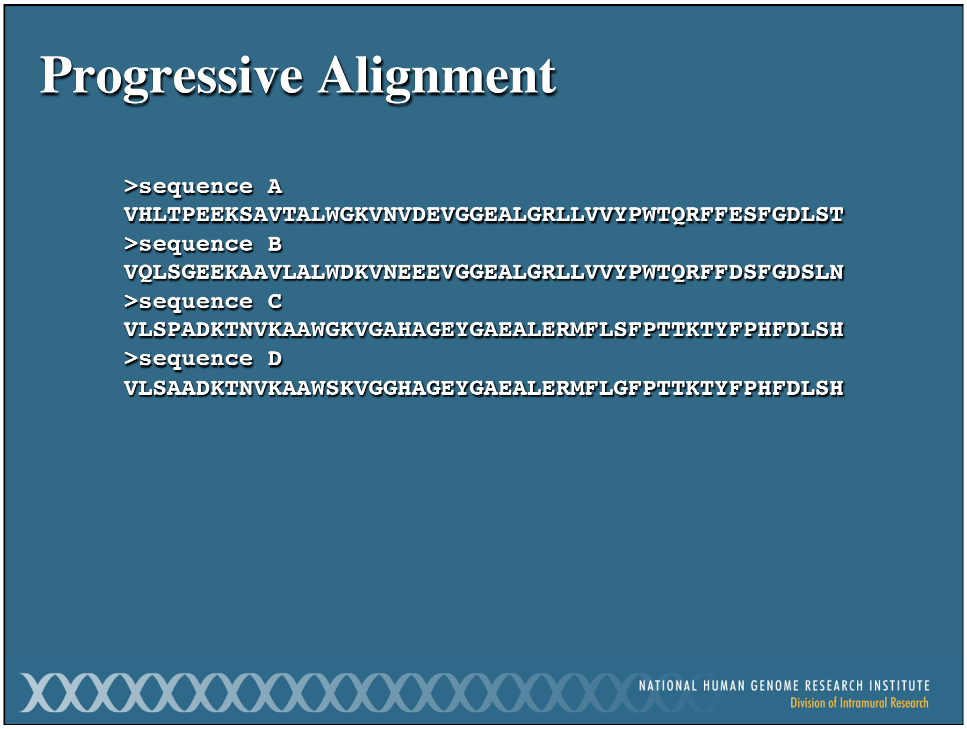
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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
  - Generally fast
  - Alignments generally of high quality

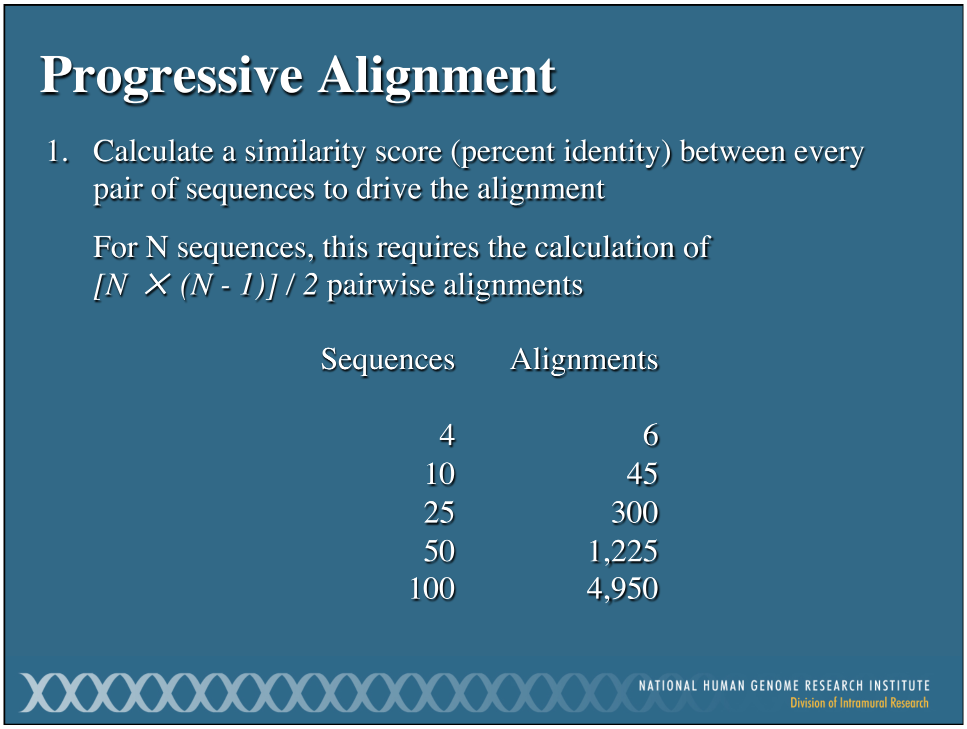
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Progressive Alignment

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



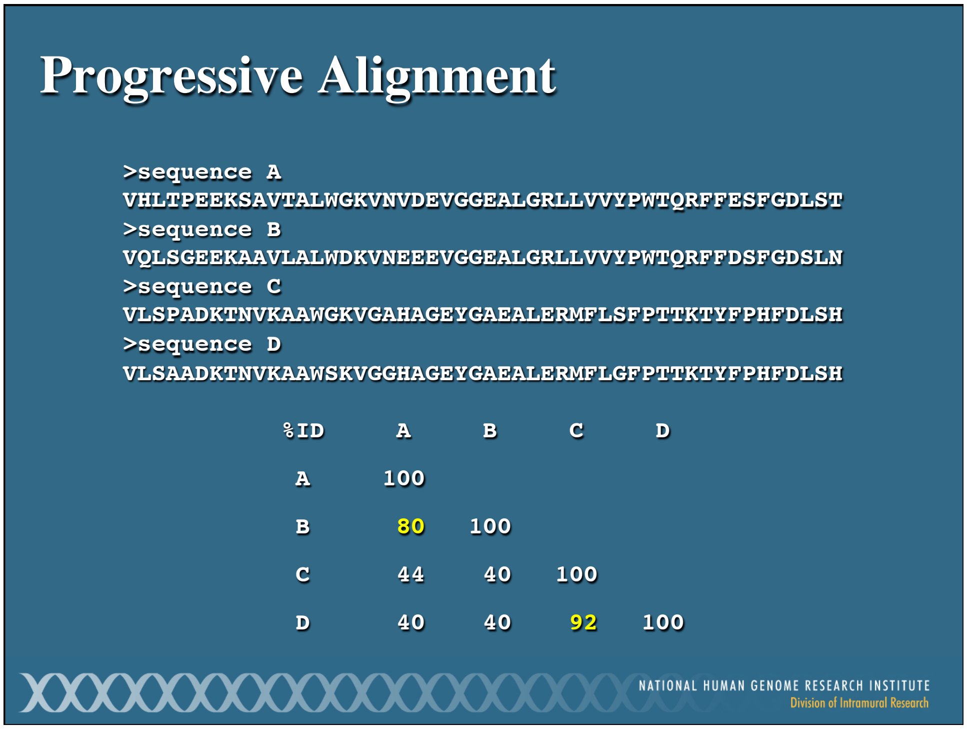
Progressive Alignment

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

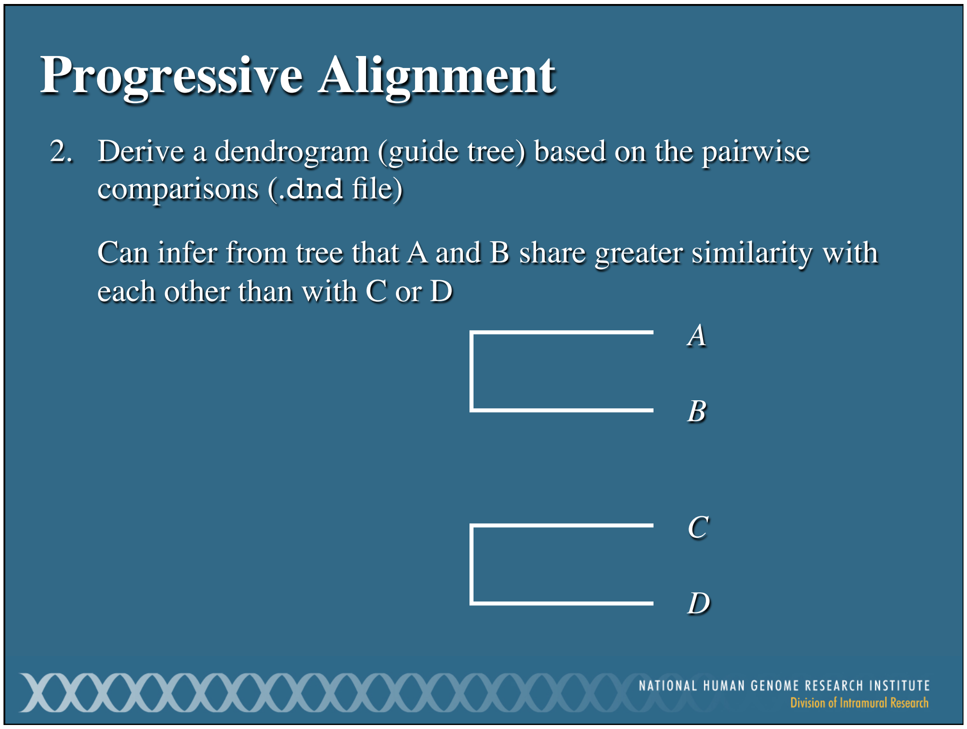


Progressive Alignment

```
>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

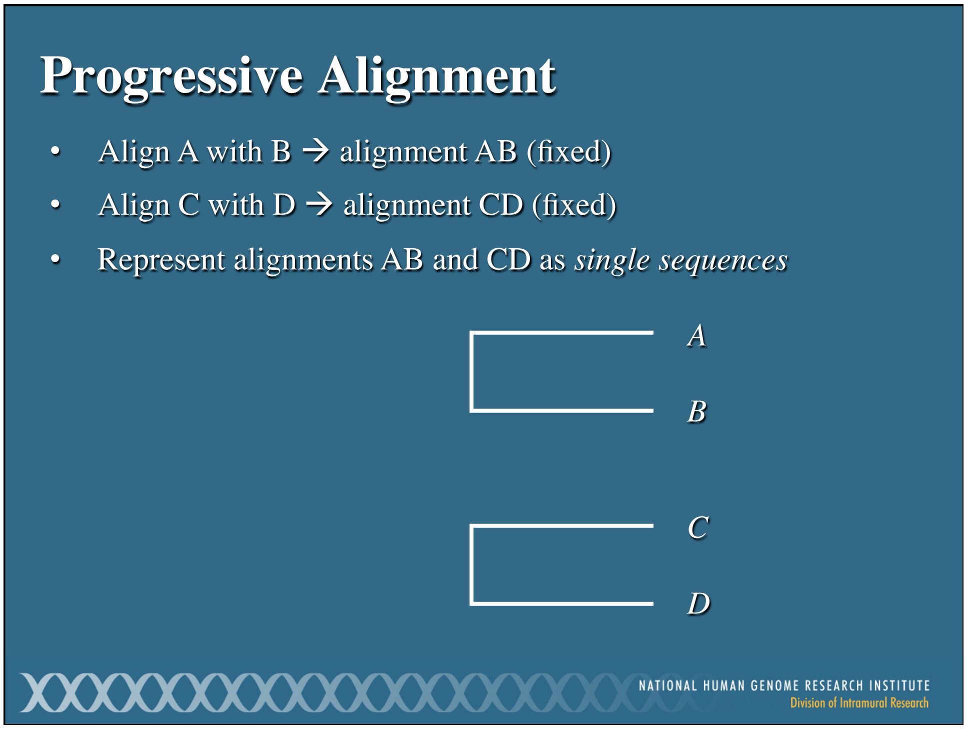


Progressive Alignment

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

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

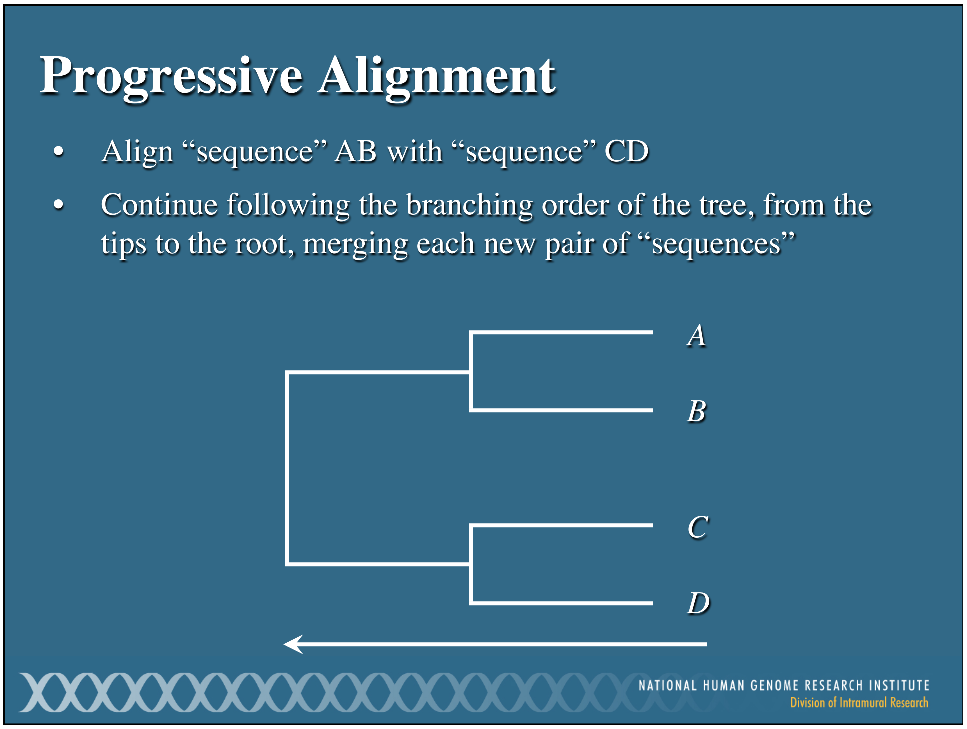
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Progressive Alignment

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

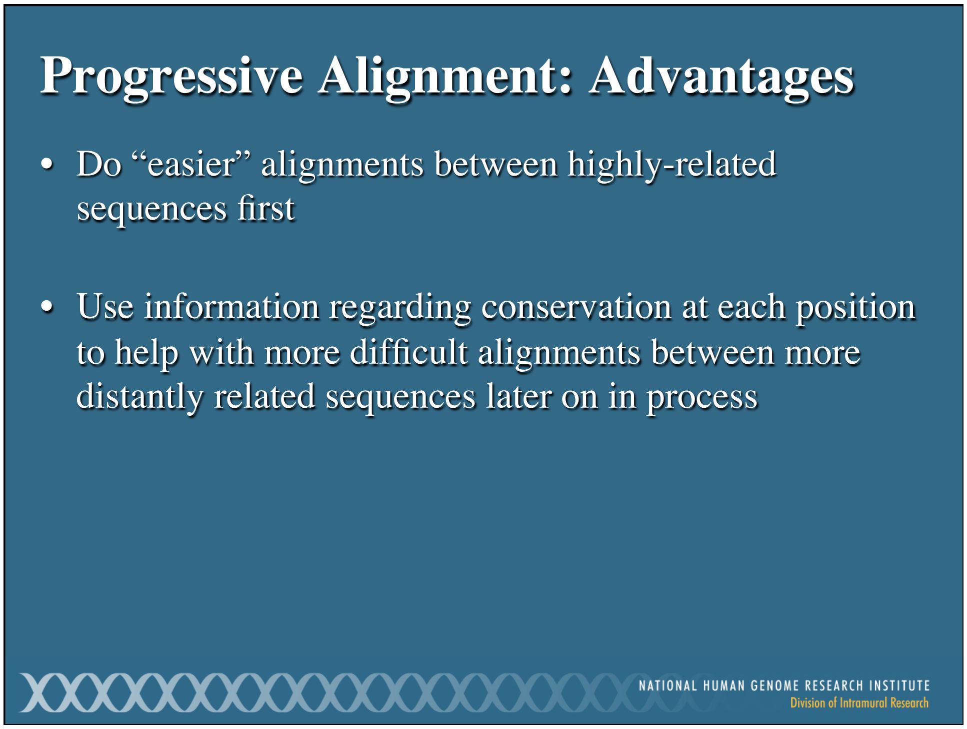
A
B
C
D
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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”

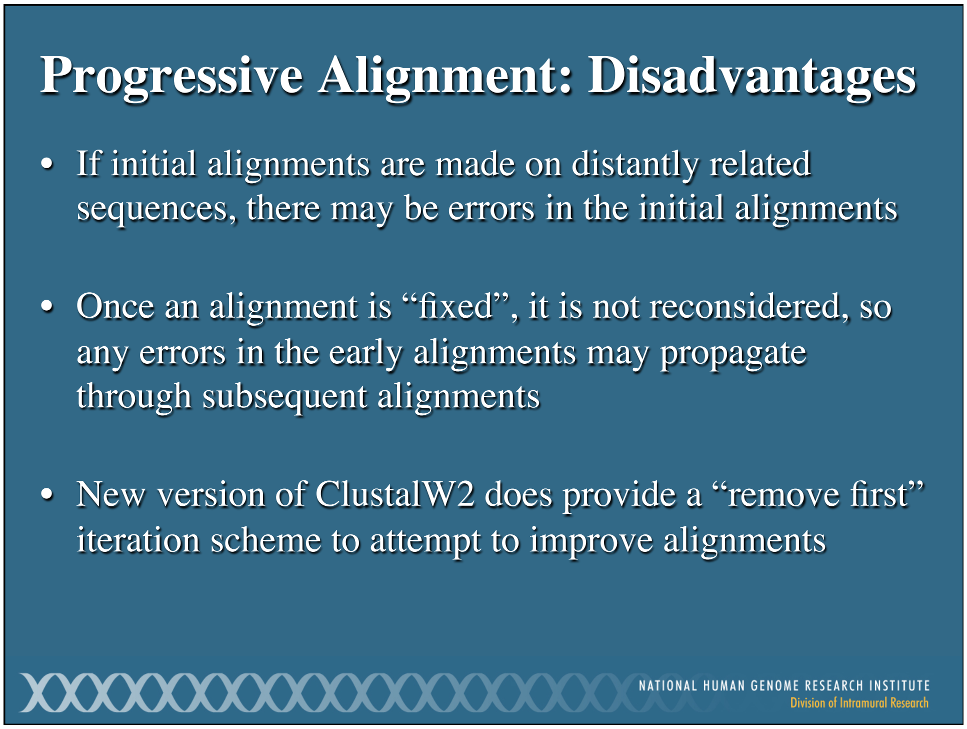
A
B
C
D
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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

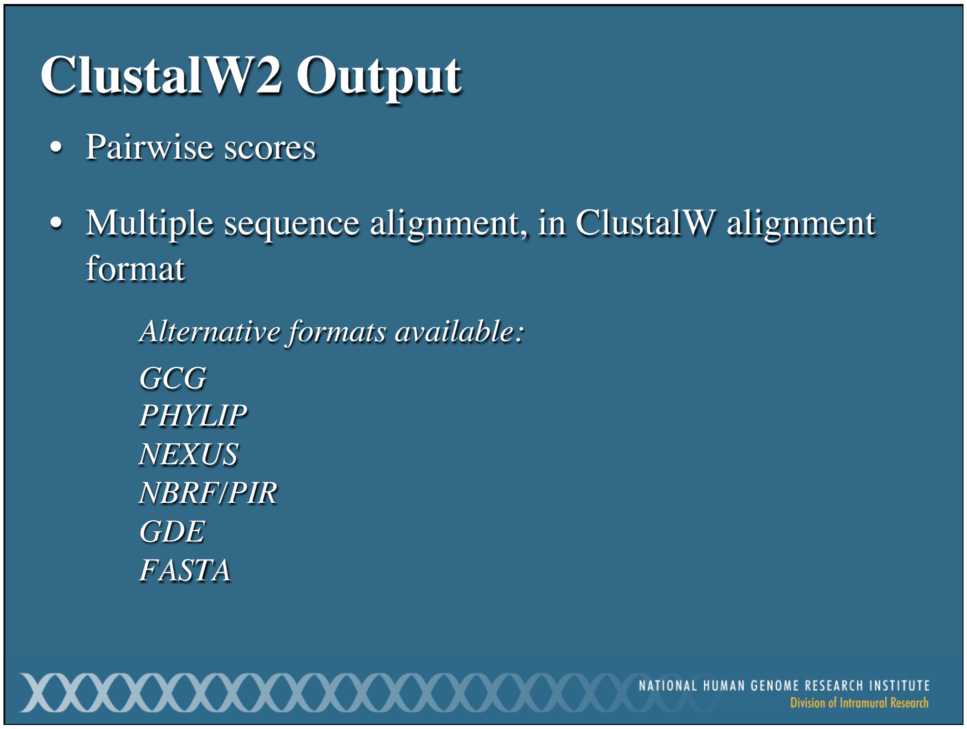
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



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



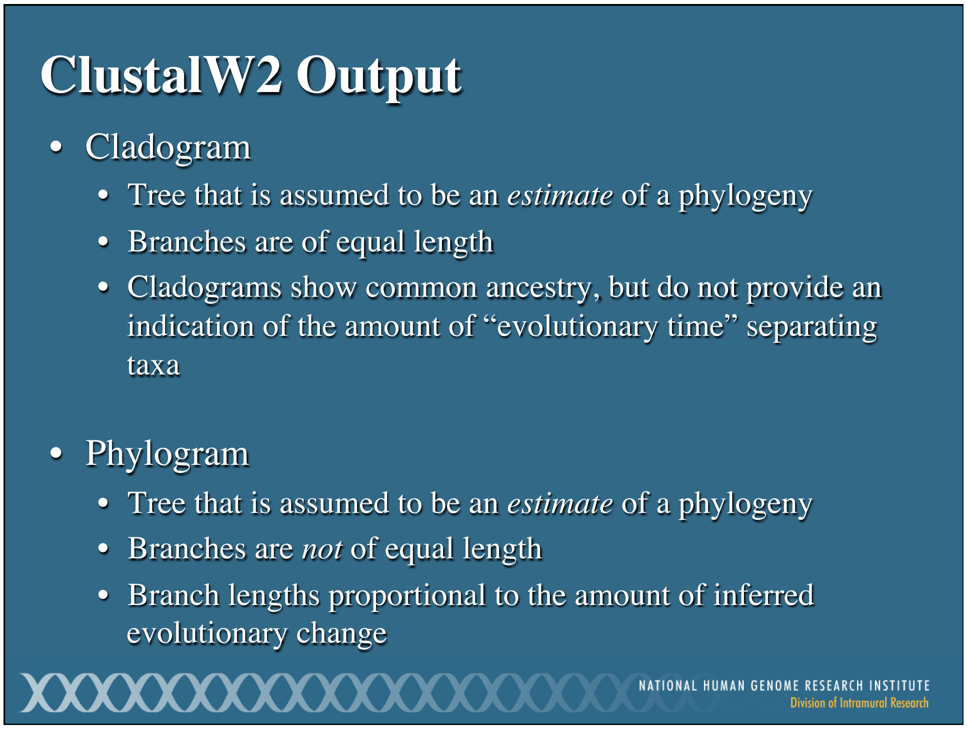
ClustalW2 Output

- Pairwise scores
- Multiple sequence alignment, in ClustalW alignment format

Alternative formats available:

GCG
PHYLIP
NEXUS
NBRF/PIR
GDE
FASTA

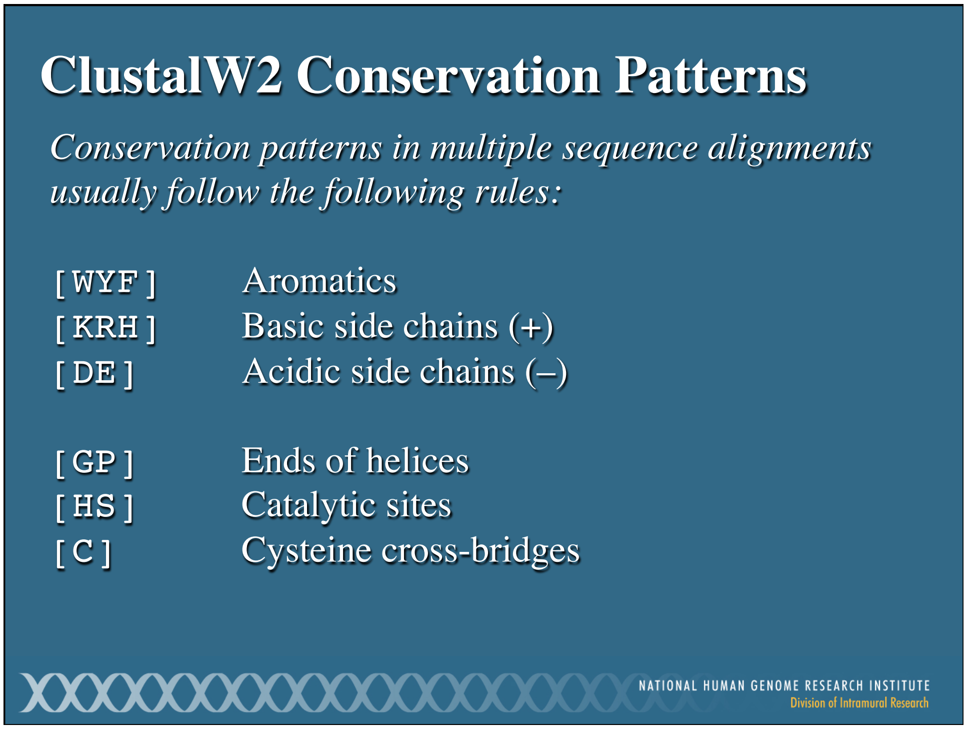
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



ClustalW2 Output

- Cladogram
  - Tree that is 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 a phylogeny
  - Branches are *not* of equal length
  - Branch lengths proportional to the amount of inferred evolutionary change

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

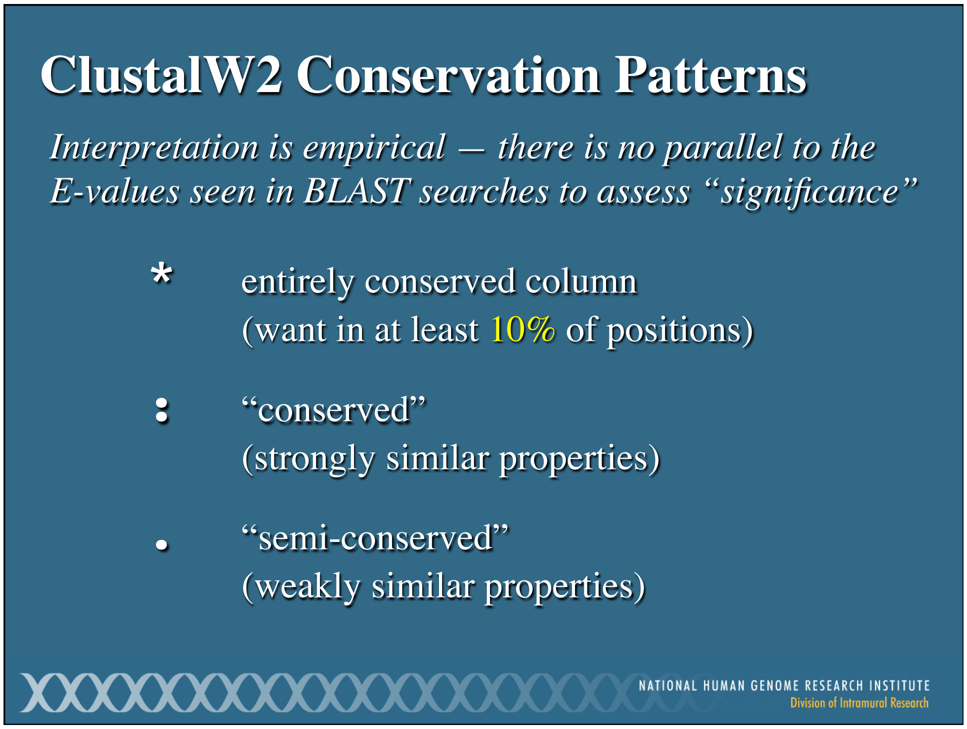


ClustalW2 Conservation Patterns

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

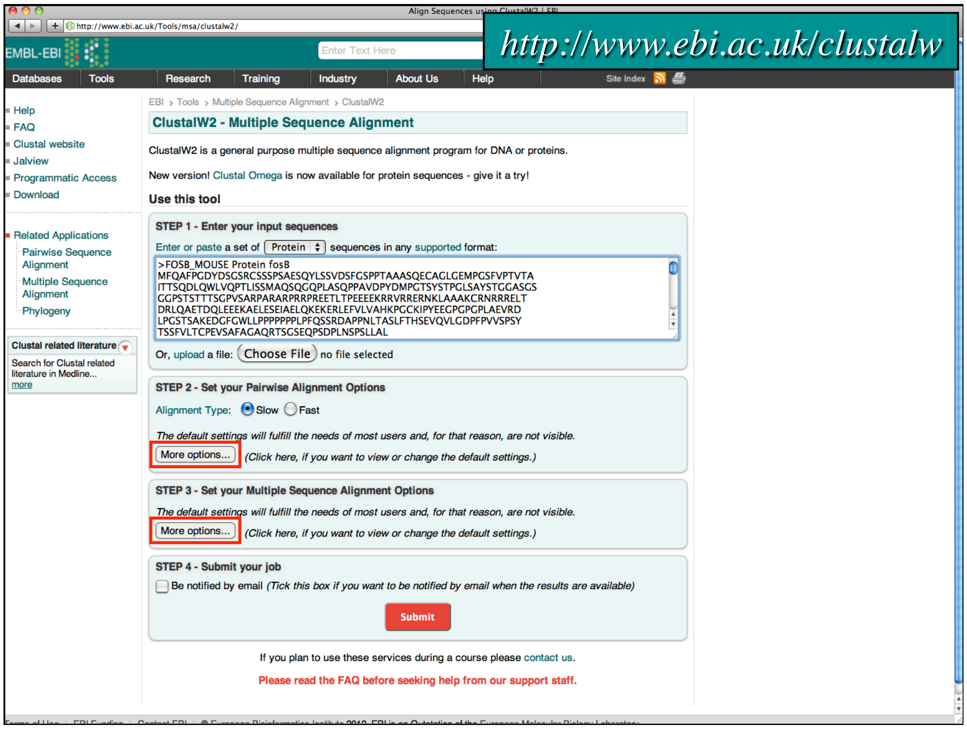


ClustalW2 Conservation Patterns

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" (strongly similar properties)
.	"semi-conserved" (weakly similar properties)

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


Align Sequences using ClustalW2 | EBI
http://www.ebi.ac.uk/Tools/msa/clustalw2/
EMBL-EBI
Enter Text Here
Databases
Tools
Research
Training
Industry
About Us
Help
Site Index
Help
FAQ
Clustal website
Jalview
Programmatic Access
Download
Related Applications
Pairwise Sequence Alignment
Multiple Sequence Alignment
Phylogeny
Clustal related literature
Search for Clustal related literature in Medline...
more
EBI › Tools › Multiple Sequence Alignment › ClustalW2

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

ClustalW2 - Multiple Sequence Alignment

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.

New version! Clustal Omega is now available for protein sequences - give it a try!

Use this tool

STEP 1 - Enter your input sequences

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

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

Or, upload a file: Choose File no file selected

STEP 2 - Set your Pairwise Alignment Options

Alignment Type: ● Slow ○ Fast

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Set your Multiple Sequence Alignment Options

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 4 - Submit your job

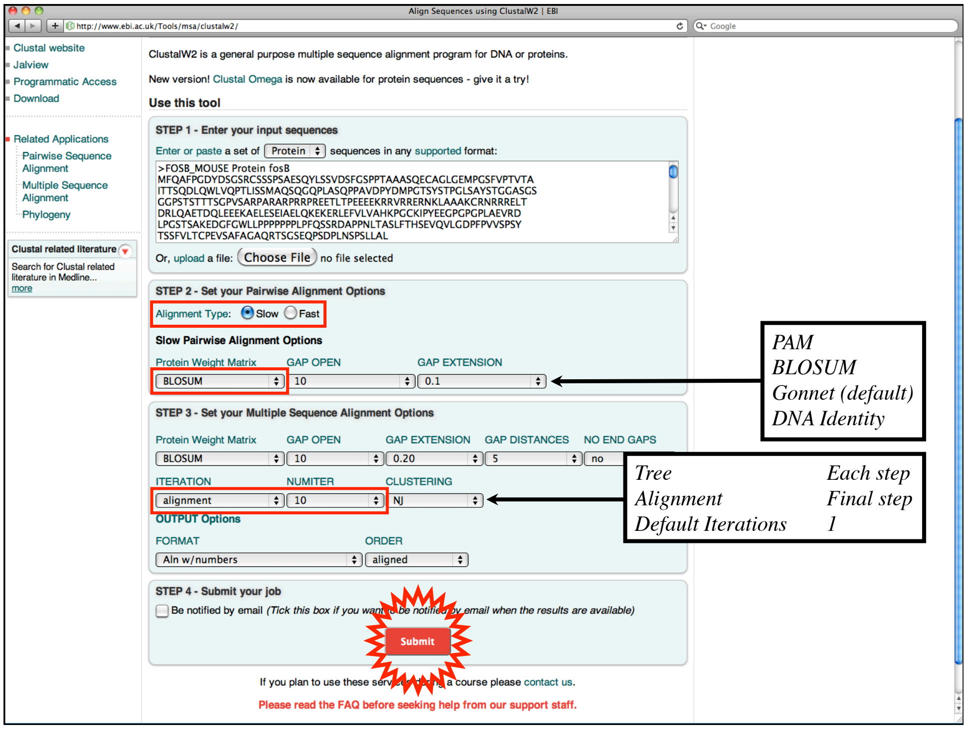
□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

If you plan to use these services during a course please contact us.

Please read the FAQ before seeking help from our support staff.

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2012. EBI is an Outstation of the European Molecular Biology Laboratory.


Align Sequences using ClustalW2 | EBI
http://www.ebi.ac.uk/Tools/msa/clustalw2/
Google
Clustal website
Jalview
Programmatic Access
Download
Related Applications
Pairwise Sequence Alignment
Multiple Sequence Alignment
Phylogeny
Clustal related literature
Search for Clustal related literature in Medline...
more

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.

New version! Clustal Omega is now available for protein sequences - give it a try!

Use this tool

STEP 1 - Enter your input sequences

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

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

Or, upload a file: Choose File no file selected

STEP 2 - Set your Pairwise Alignment Options

Alignment Type: ● Slow ○ Fast

Slow Pairwise Alignment Options

Protein Weight Matrix: BLOSUM | GAP OPEN: 10 | GAP EXTENSION: 0.1

PAM
BLOSUM
Gonnet (default)
DNA Identity

STEP 3 - Set your Multiple Sequence Alignment Options

Protein Weight Matrix: BLOSUM | GAP OPEN: 10 | GAP EXTENSION: 0.20 | GAP DISTANCES: 5 | NO END GAPS: no

ITERATION: alignment | NUMITER: 10 | CLUSTERING: NJ

Tree	Each step
Alignment	Final step
Default Iterations	1

OUTPUT Options

FORMAT: Aln w/numbers | ORDER: aligned

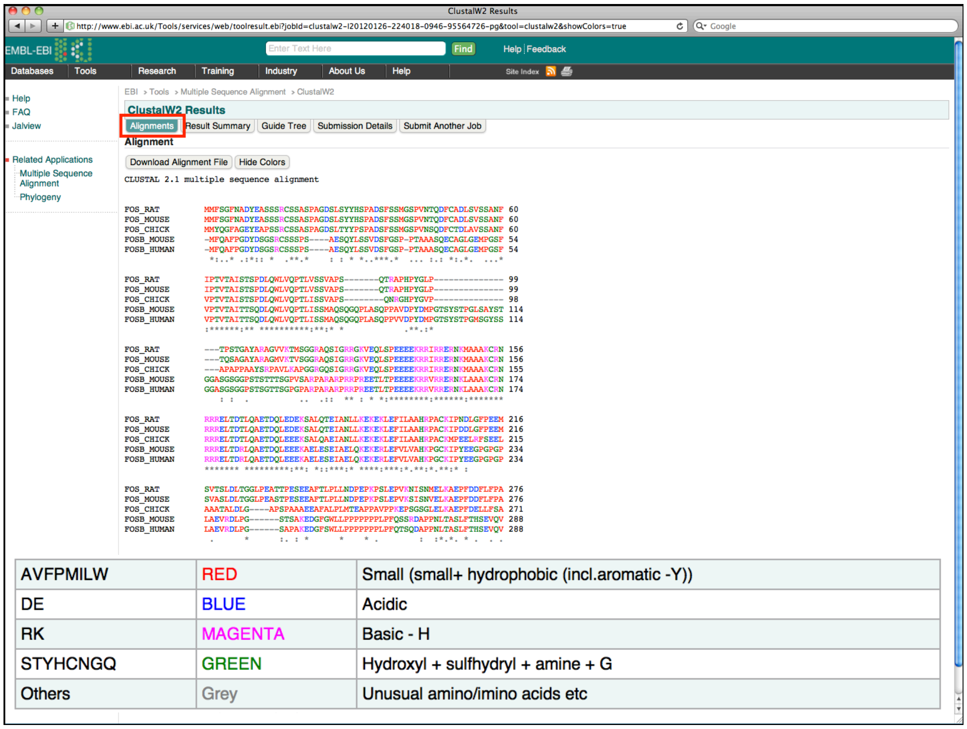
STEP 4 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

If you plan to use these services during a course please contact us.

Please read the FAQ before seeking help from our support staff.

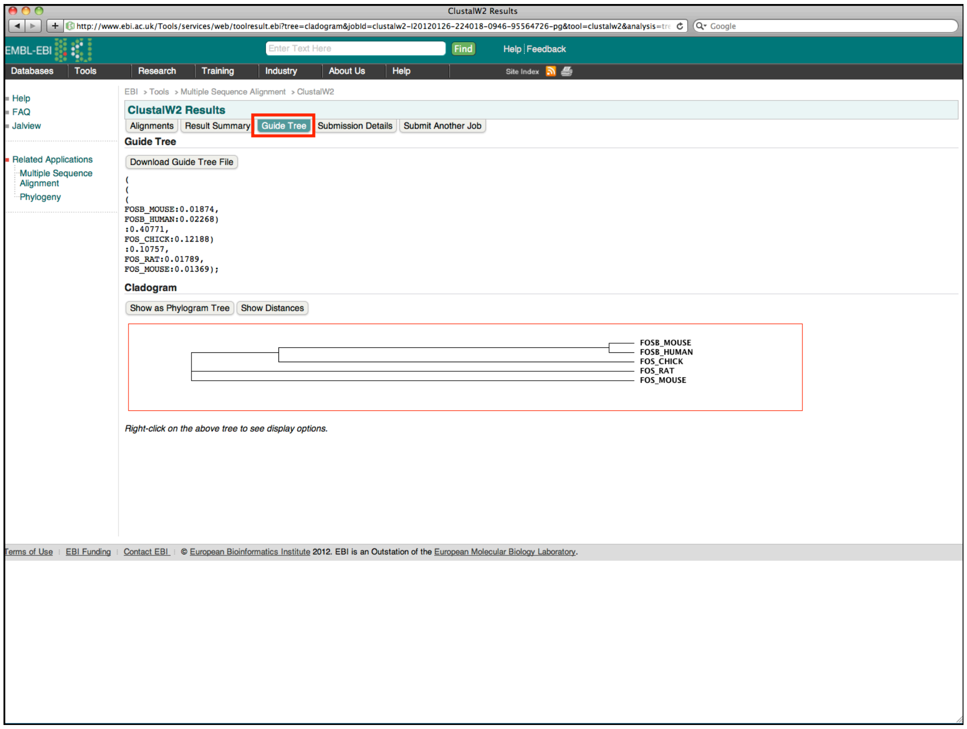


ClustalW2 Results

Alignment

```
CLUSTAL 2.1 multiple sequence alignment


FOS_RAT         MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_MOUSE       MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANF 60
FOS_CHICK       MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANF 60
FOSB_MOUSE      -MFQAFPGDYDSGSRCSSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSF 54
FOSB_HUMAN      -MFQAFPGDYDSGSRCSSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSF 54
                 *:..* .:*:: *  .**.*    : : * *..***.*  ... :.: *:.*.  ...*

FOS_RAT         IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGLP-------------- 99
FOS_MOUSE       IPTVTAISTSPDLQWLVQPTLVSSVAPS-------QTRAPHPYGLP-------------- 99
FOS_CHICK       VPTVTAISTSPDLQWLVQPTLISSVAPS--------QNRGHPYGVP-------------- 98
FOSB_MOUSE      VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYST 114
FOSB_HUMAN      VPTVTAITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSS 114
                :******:** **********:**:* *            .**.:*

FOS_RAT         ---TPSTGAYARAGVVKTMSGGRAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
FOS_MOUSE       ---TQSAGAYARAGMVKTVSGGRAQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 156
FOS_CHICK       ---APAPPAAYSRPAVLKAPGGRGQSIGRRGKVEQLSPEEEEKRRIRRERNKMAAAKCRN 155
FOSB_MOUSE      GGASGSGGPSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRN 174
FOSB_HUMAN      GGASGSGGPSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRN 174
                   : :  .          ..  .::  ** : * *:********:******:*******

FOS_RAT         RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPNDLGFPEEM 216
FOS_MOUSE       RRRELTDTLQAETDQLEDEKSALQTEIANLLKEKEKLEFILAAHRPACKIPDDLGFPEEM 216
FOS_CHICK       RRRELTDTLQAETDQLEEEKSALQAEIANLLKEKEKLEFILAAHRPACKMPEELRFSEEL 215
FOSB_MOUSE      RRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGP 234
FOSB_HUMAN      RRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGP 234
                ******* *********:**: *::***:* ****:***:*.**:*.**:* :

FOS_RAT         SVTSLDLTGGLPEATTPESEEAFTLPLLNDPEPKPSLEPVKNISNMELKAEPFDDFLFPA 276
FOS_MOUSE       SVASLDLTGGLPEASTPESEEAFTLPLLNDPEPKPSLEPVKSISNVELKAEPFDDFLFPA 276
FOS_CHICK       AAATALDLG----APSPAAAEEAFALPLMTEAPPAVPPKEPSGSGLELKAEPFDELLFSA 271
FOSB_MOUSE      LAEVRDLPG------STSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQV 288
FOSB_HUMAN      LAEVRDLPG------SAPAKEDGFSWLLPPPPPPPLPFQTSQDAPPNLTASLFTHSEVQV 288
                 .     *     :. : *      *    * .        :  :*.*. * .  . .
```

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



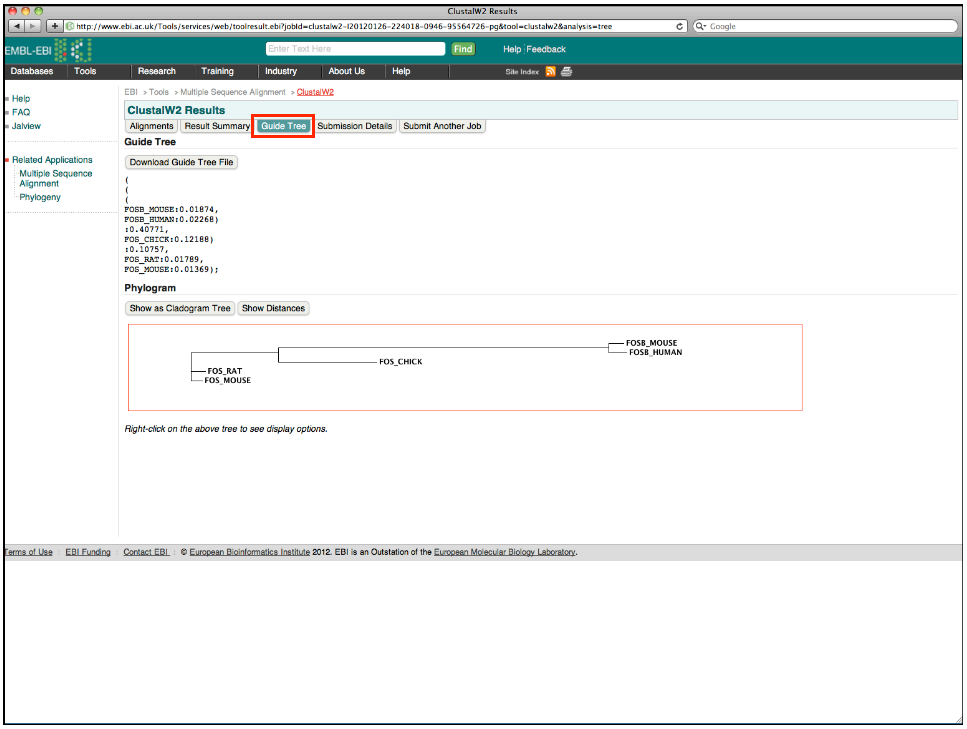
ClustalW2 Results

Guide Tree

```
(
(
(
FOSB_MOUSE:0.01874,
FOSB_HUMAN:0.02268)
:0.40771,
FOS_CHICK:0.12188)
:0.10757,
FOS_RAT:0.01789,
FOS_MOUSE:0.01369);
```

Cladogram

Right-click on the above tree to see display options.


ClustalW2 Results
http://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalw2-I20120126-224018-0946-95564726-pg&tool=clustalw2&analysis=tree
Google
EMBL-EBI
Enter Text Here
Find
Help | Feedback
Databases
Tools
Research
Training
Industry
About Us
Help
Site Index
Help
FAQ
Jalview
Related Applications
Multiple Sequence Alignment
Phylogeny
EBI > Tools > Multiple Sequence Alignment > ClustalW2
Alignments
Result Summary
Guide Tree
Submission Details
Submit Another Job
Download Guide Tree File
Show as Cladogram Tree
Show Distances
FOSB_MOUSE
FOSB_HUMAN
FOS_CHICK
FOS_RAT
FOS_MOUSE
Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2012. EBI is an Outstation of the European Molecular Biology Laboratory.

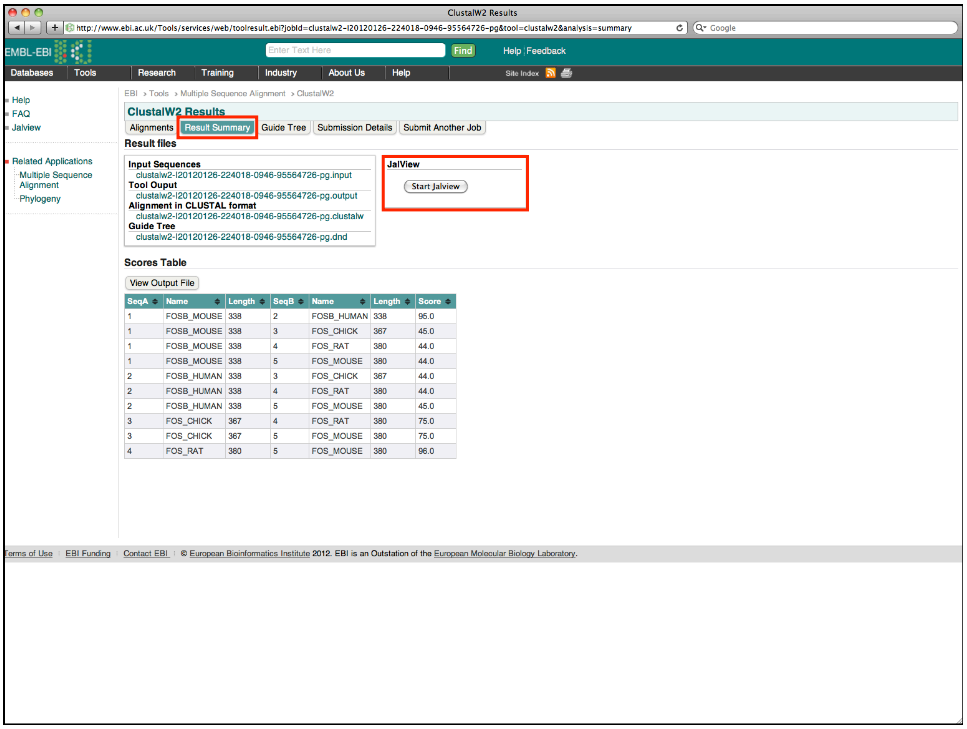
ClustalW2 Results

Guide Tree

```
(
(
(
FOSB_MOUSE:0.01874,
FOSB_HUMAN:0.02268)
:0.40771,
FOS_CHICK:0.12188)
:0.10757,
FOS_RAT:0.01789,
FOS_MOUSE:0.01369);
```

Phylogram

Right-click on the above tree to see display options.


ClustalW2 Results
http://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=clustalw2-I20120126-224018-0946-95564726-pg&tool=clustalw2&analysis=summary
Google
EMBL-EBI
Enter Text Here
Find
Help | Feedback
Databases
Tools
Research
Training
Industry
About Us
Help
Site Index
Help
FAQ
Jalview
Related Applications
Multiple Sequence Alignment
Phylogeny
EBI > Tools > Multiple Sequence Alignment > ClustalW2
Alignments
Result Summary
Guide Tree
Submission Details
Submit Another Job
Start Jalview
View Output File
Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2012. EBI is an Outstation of the European Molecular Biology Laboratory.

ClustalW2 Results

Result files

Input Sequences
clustalw2-I20120126-224018-0946-95564726-pg.input
Tool Ouput
clustalw2-I20120126-224018-0946-95564726-pg.output
Alignment in CLUSTAL format
clustalw2-I20120126-224018-0946-95564726-pg.clustalw
Guide Tree
clustalw2-I20120126-224018-0946-95564726-pg.dnd

JalView

Scores Table

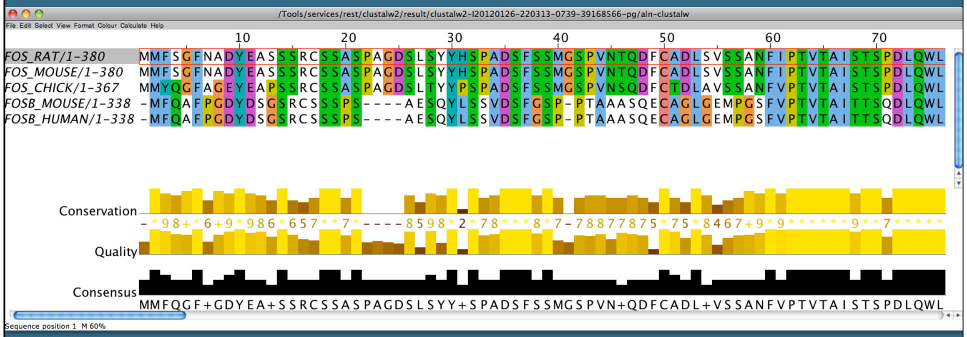
SeqA	Name	Length	SeqB	Name	Length	Score
1	FOSB_MOUSE	338	2	FOSB_HUMAN	338	95.0
1	FOSB_MOUSE	338	3	FOS_CHICK	367	45.0
1	FOSB_MOUSE	338	4	FOS_RAT	380	44.0
1	FOSB_MOUSE	338	5	FOS_MOUSE	380	44.0
2	FOSB_HUMAN	338	3	FOS_CHICK	367	44.0
2	FOSB_HUMAN	338	4	FOS_RAT	380	44.0
2	FOSB_HUMAN	338	5	FOS_MOUSE	380	45.0
3	FOS_CHICK	367	4	FOS_RAT	380	75.0
3	FOS_CHICK	367	5	FOS_MOUSE	380	75.0
4	FOS_RAT	380	5	FOS_MOUSE	380	96.0

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Default view



/Tools/services/rest/clustalw2/result/clustalw2-I20120126-220313-0739-39168566-pg/aln-clustalw
File Edit Select View Format Colour Calculate Help

```
                         10        20        30        40        50        60        70
FOS_RAT/1-380    MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQWL
FOS_MOUSE/1-380  MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQWL
FOS_CHICK/1-367  MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANFVPTVTAISTSPDLQWL
FOSB_MOUSE/1-338 -MFQAFPGDYDSGSRCSSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWL
FOSB_HUMAN/1-338 -MFQAFPGDYDSGSRCSSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWL

    Conservation
                 -*98+*6+9*986*657**7*----8598*2*78***8*7-78877875*75*8467+9*9******9**7*****
         Quality

       Consensus
                 MMFQGF+GDYEA+SSRCSSASPAGDSLSYY+SPADSFSSMGSPVN+QDFCADL+VSSANFVPTVTAISTSPDLQWL
Sequence position 1  M 60%
```

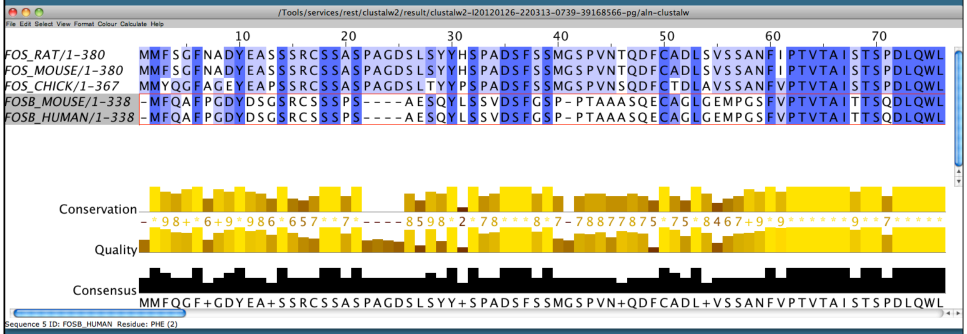
Conservation Conservation of total alignment (indication of percent identity)

Quality Alignment quality, based on BLOSUM scores

Consensus Based on percent identity

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Colour → Percentage Identity

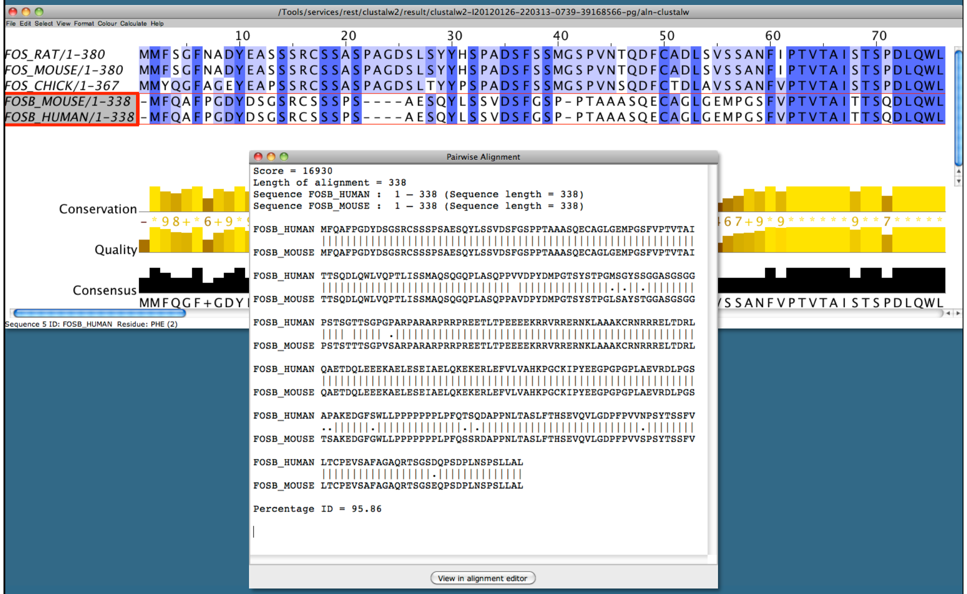

/Tools/services/rest/clustalw2/result/clustalw2-I20120126-220313-0739-39168566-pg/aln-clustalw
File Edit Select View Format Colour Calculate Help
10
20
30
40
50
60
70
Conservation
Quality
Consensus

```
FOS_RAT/1-380     MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQWL
FOS_MOUSE/1-380   MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQWL
FOS_CHICK/1-367   MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANFVPTVTAISTSPDLQWL
FOSB_MOUSE/1-338  -MFQAFPGDYDSGSRCSSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWL
FOSB_HUMAN/1-338  -MFQAFPGDYDSGSRCSSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWL

Conservation      -*98+*6+9*986*657**7*----8598*2*78***8*7-78877875*75*8467+9*9******9**7*****
Consensus         MMFQGF+GDYEA+SSRCSSASPAGDSLSYY+SPADSFSSMGSPVN+QDFCADL+VSSANFVPTVTAISTSPDLQWL

Sequence 5 ID: FOSB_HUMAN  Residue: PHE (2)
```

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Calculate → Pairwise Alignments

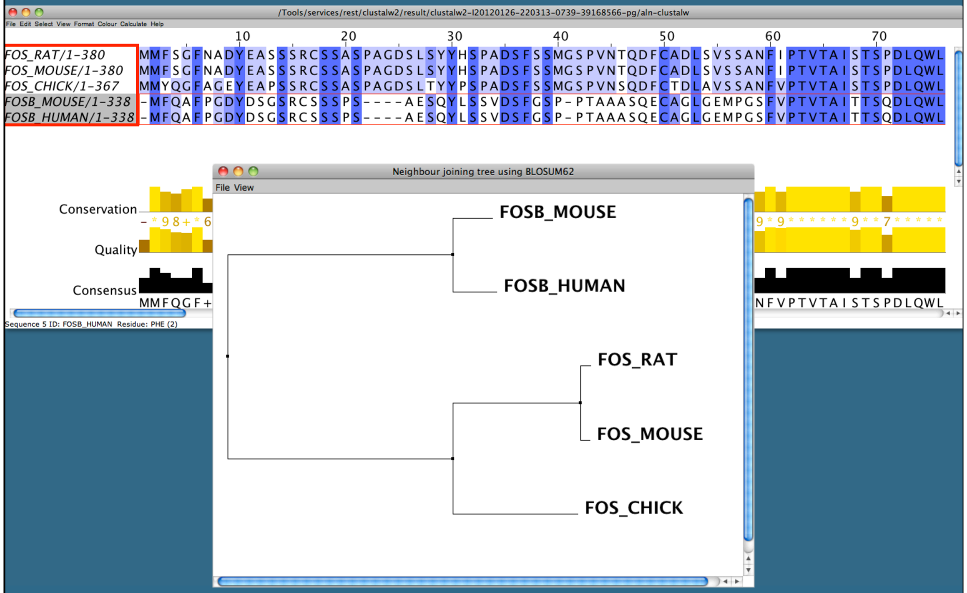

/Tools/services/rest/clustalw2/result/clustalw2-I20120126-220313-0739-39168566-pg/aln-clustalw
File Edit Select View Format Colour Calculate Help
10
20
30
40
50
60
70
Conservation
Quality
Consensus
Pairwise Alignment
View in alignment editor

```
FOS_RAT/1-380     MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQWL
FOS_MOUSE/1-380   MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQWL
FOS_CHICK/1-367   MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANFVPTVTAISTSPDLQWL
FOSB_MOUSE/1-338  -MFQAFPGDYDSGSRCSSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWL
FOSB_HUMAN/1-338  -MFQAFPGDYDSGSRCSSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWL

Sequence 5 ID: FOSB_HUMAN  Residue: PHE (2)
```

```
Score = 16930
Length of alignment = 338
Sequence FOSB_HUMAN :  1 - 338 (Sequence length = 338)
Sequence FOSB_MOUSE :  1 - 338 (Sequence length = 338)

FOSB_HUMAN MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTAI
           |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
FOSB_MOUSE MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTAI

FOSB_HUMAN TTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGSGG
           ||||||||||||||||||||||||||||||| |||||||||||||||.|.||.||||||||
FOSB_MOUSE TTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGSGG

FOSB_HUMAN PSTSGTTSGPGPARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELTDRL
           |||| ||||| .|||||||||||||||||||||||||||||||||||||||||||||||||
FOSB_MOUSE PSTSTTTSGPVSARPARARPRRPREETLTPEEEEKRRVRRERNKLAAAKCRNRRRELTDRL

FOSB_HUMAN QAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRDLPGS
           |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
FOSB_MOUSE QAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRDLPGS

FOSB_HUMAN APAKEDGFSWLLPPPPPPPLPFQTSQDAPPNLTASLFTHSEVQVLGDPFPVVNPSYTSSFV
           ..||||||.||||||||||||||.|.||||||||||||||||||||||||||.||||||||
FOSB_MOUSE TSAKEDGFGWLLPPPPPPPLPFQSSRDAPPNLTASLFTHSEVQVLGDPFPVVSPSYTSSFV

FOSB_HUMAN LTCPEVSAFAGAQRTSGSDQPSDPLNSPSLLAL
           ||||||||||||||||||.||||||||||||||
FOSB_MOUSE LTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL

Percentage ID = 95.86
```

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Calculate → Calculate Tree → Neighbour Joining Using BLOSUM62

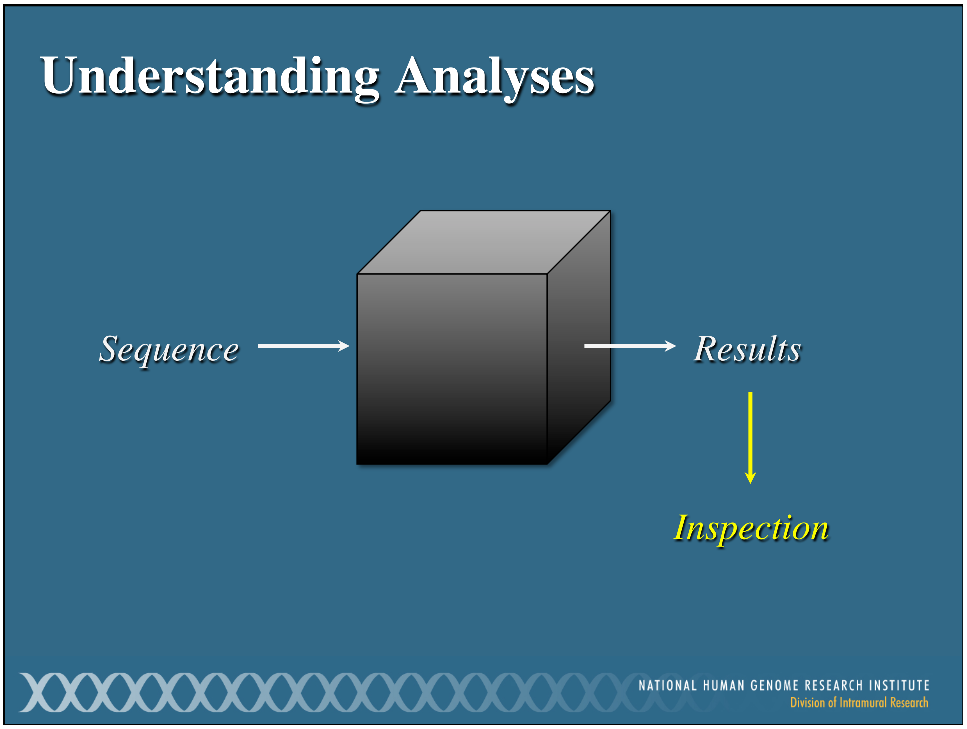

/Tools/services/rest/clustalw2/result/clustalw2-I20120126-220313-0739-39168566-pg/aln-clustalw
File Edit Select View Format Colour Calculate Help
Neighbour joining tree using BLOSUM62
File View
FOSB_MOUSE
FOSB_HUMAN
FOS_RAT
FOS_MOUSE
FOS_CHICK

```
                          10        20        30        40        50        60        70
FOS_RAT/1-380     MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQWL
FOS_MOUSE/1-380   MMFSGFNADYEASSSRCSSASPAGDSLSYYHSPADSFSSMGSPVNTQDFCADLSVSSANFIPTVTAISTSPDLQWL
FOS_CHICK/1-367   MMYQGFAGEYEAPSSRCSSASPAGDSLTYYPSPADSFSSMGSPVNSQDFCTDLAVSSANFVPTVTAISTSPDLQWL
FOSB_MOUSE/1-338  -MFQAFPGDYDSGSRCSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWL
FOSB_HUMAN/1-338  -MFQAFPGDYDSGSRCSSPS----AESQYLSSVDSFGSP-PTAAASQECAGLGEMPGSFVPTVTAITTSQDLQWL

Conservation      -*98+*6 ... 9*9*****9**7*****
Quality
Consensus         MMFQGF+ ... NFVPTVTAISTSPDLQWL
Sequence 5 ID: FOSB_HUMAN  Residue: PHE (2)
```

Further Reading

Current Protocols in Bioinformatics
Unit 2.3
ClustalW

Current Protocols in Bioinformatics
Unit 3.8
T-Coffee



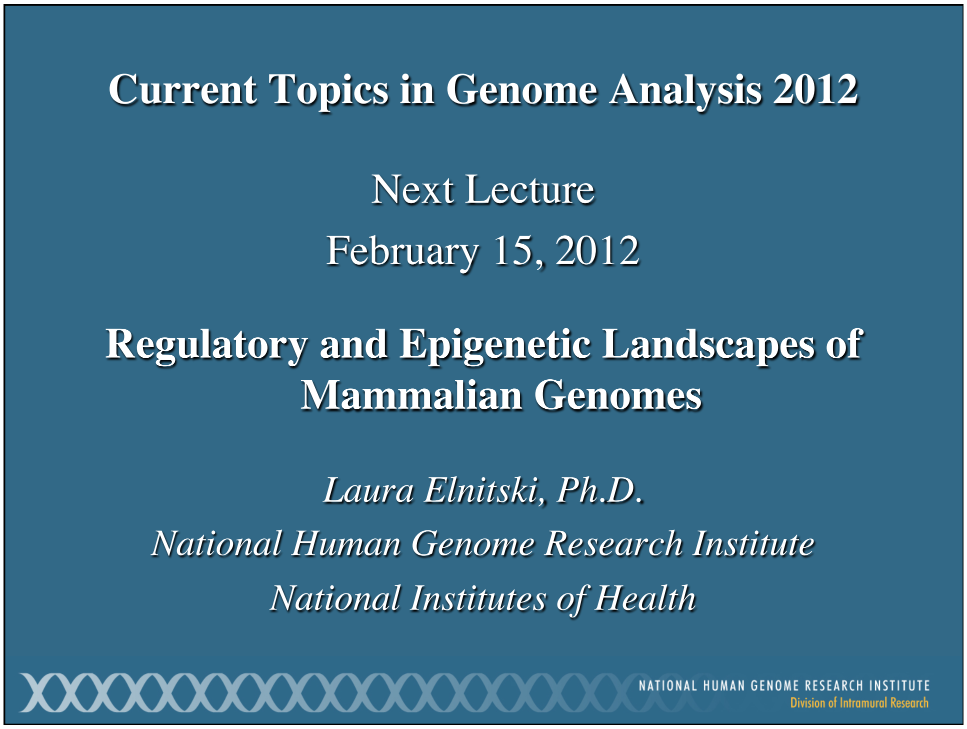
Understanding Analyses

Sequence → → *Results*

↓

Inspection

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



Current Topics in Genome Analysis 2012

Next Lecture

February 15, 2012

Regulatory and Epigenetic Landscapes of Mammalian Genomes

Laura Elnitski, Ph.D.
National Human Genome Research Institute
National Institutes of Health

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Intramural Research Program

Our Research Changes Lives

one program
many people
infinite possibilities

irp.nih.gov