NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

Current Topics in Genome Analysis 2014

Week 3: Biological Sequence Analysis II

Andy Baxevanis, Ph.D.

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

JOHNS HOPKINS
MEDICINE
CONTINUING MEDICAL EDUCATION

Current Topics in Genome Analysis 2014

Andy Baxevanis, Ph.D.

No Relevant Financial Relationships with Commercial Interests

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, CDD*
    or "many-against-one": *PSI-BLAST, DELTA-BLAST*

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Profiles, Patterns, Motifs, and Domains

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
```

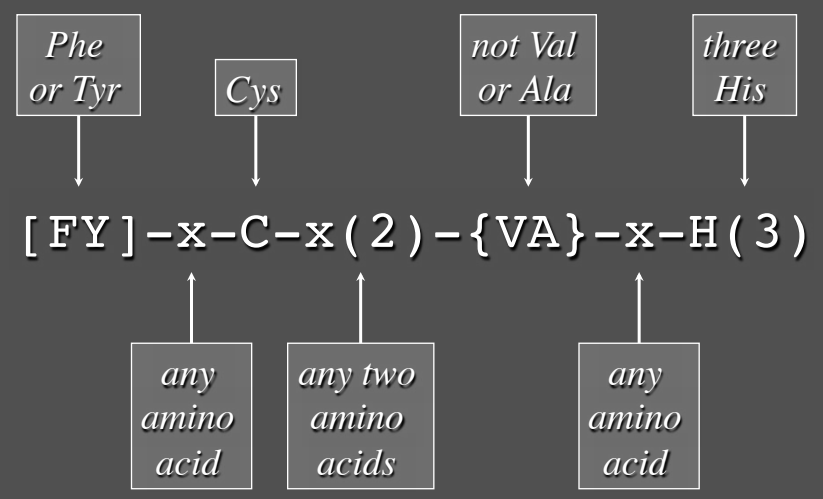
- ***Which residues are seen at each position?***
- ***What is the frequency of observed residues?***
- ***Which positions are conserved?***
- ***Where can gaps be introduced?***

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	16	6	15	6	6	-12	15	0	8	-5	-5	-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	-11	13	11	-9	6	-16	-11	8	89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Patterns



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Pfam

- Collection of multiple alignments of protein domains and conserved protein regions that 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

Finn et al., Nucleic Acids Res. 42: D222-D230, 2014

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

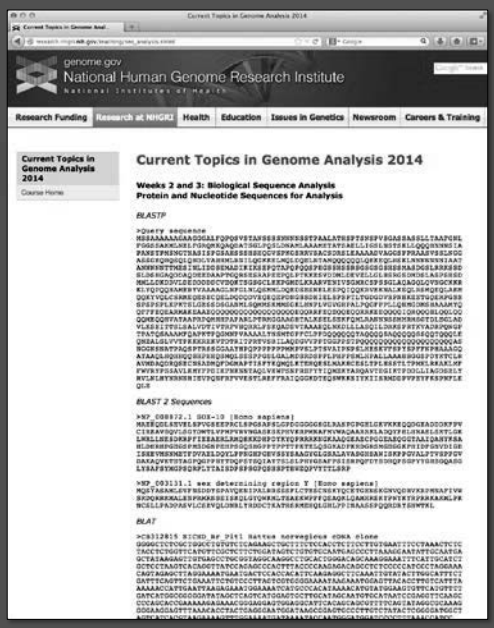
Pfam

- Pfam A
  - Based on *curated* multiple alignments ("seed alignment")
  - HMMER used to find all detectable protein sequences belonging to the family (*Eddy, 2011*)
  - 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

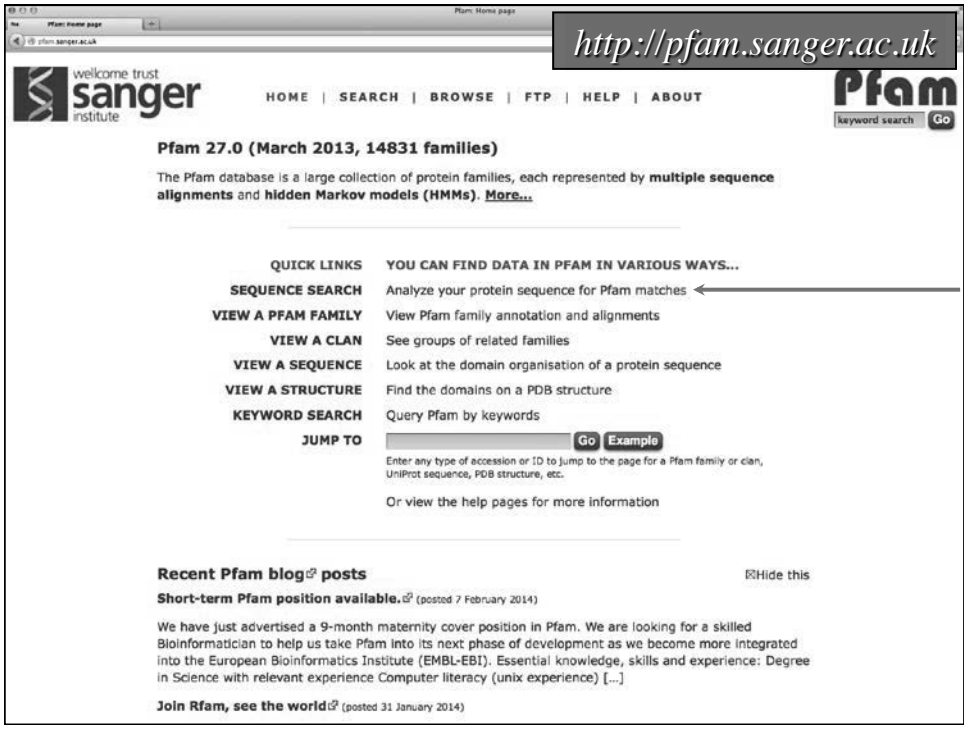
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

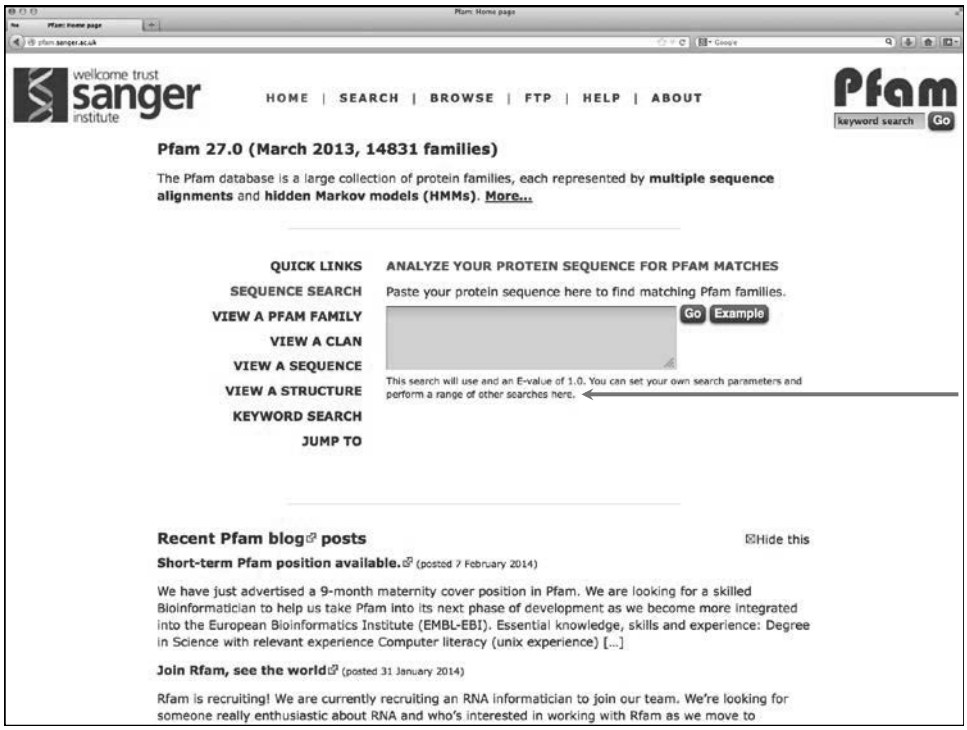
Sequences Used in Examples

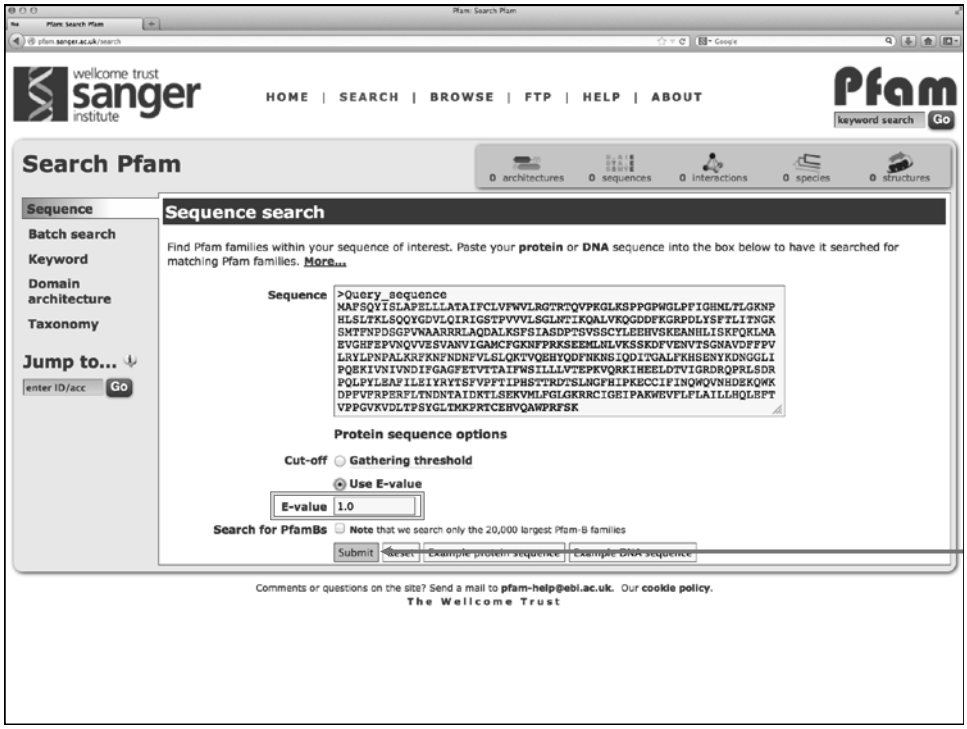
http://research.nhgri.nih.gov/teaching/seq_analysis.shtml

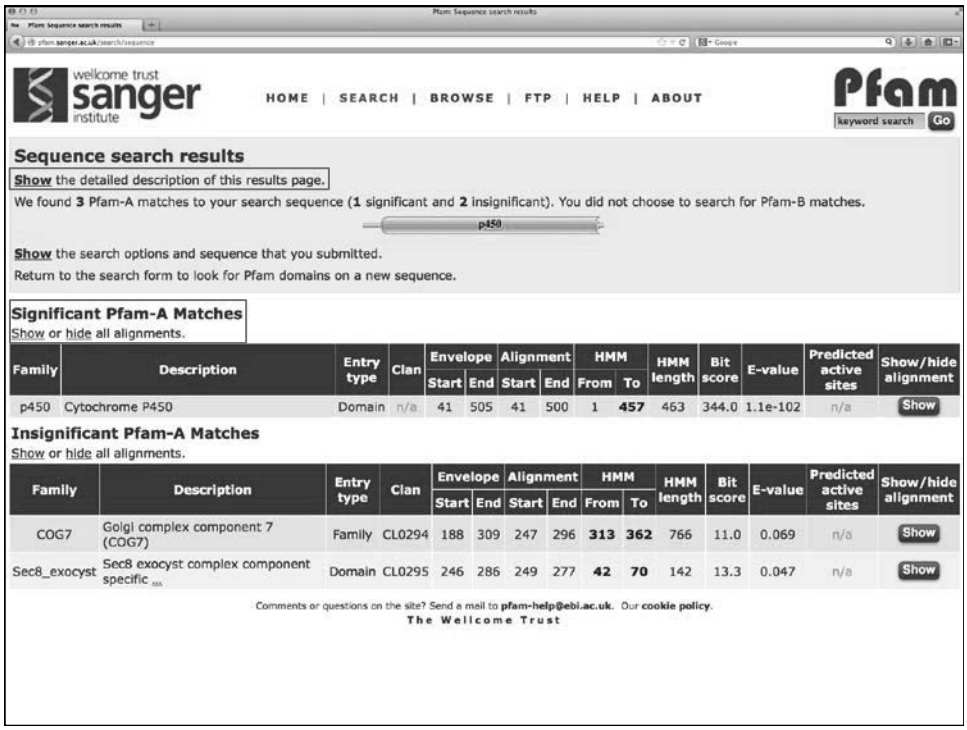


NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


http://pfam.sanger.ac.uk
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search Go
Pfam 27.0 (March 2013, 14831 families)
The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...
QUICK LINKS
YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH
Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY
View Pfam family annotation and alignments
VIEW A CLAN
See groups of related families
VIEW A SEQUENCE
Look at the domain organisation of a protein sequence
VIEW A STRUCTURE
Find the domains on a PDB structure
KEYWORD SEARCH
Query Pfam by keywords
JUMP TO
Go Example
Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.
Or view the help pages for more information
Recent Pfam blog posts
Hide this
Short-term Pfam position available. (posted 7 February 2014)
We have just advertised a 9-month maternity cover position in Pfam. We are looking for a skilled Bioinformatician to help us take Pfam into its next phase of development as we become more integrated into the European Bioinformatics Institute (EMBL-EBI). Essential knowledge, skills and experience: Degree in Science with relevant experience Computer literacy (unix experience) [...]
Join Rfam, see the world (posted 31 January 2014)

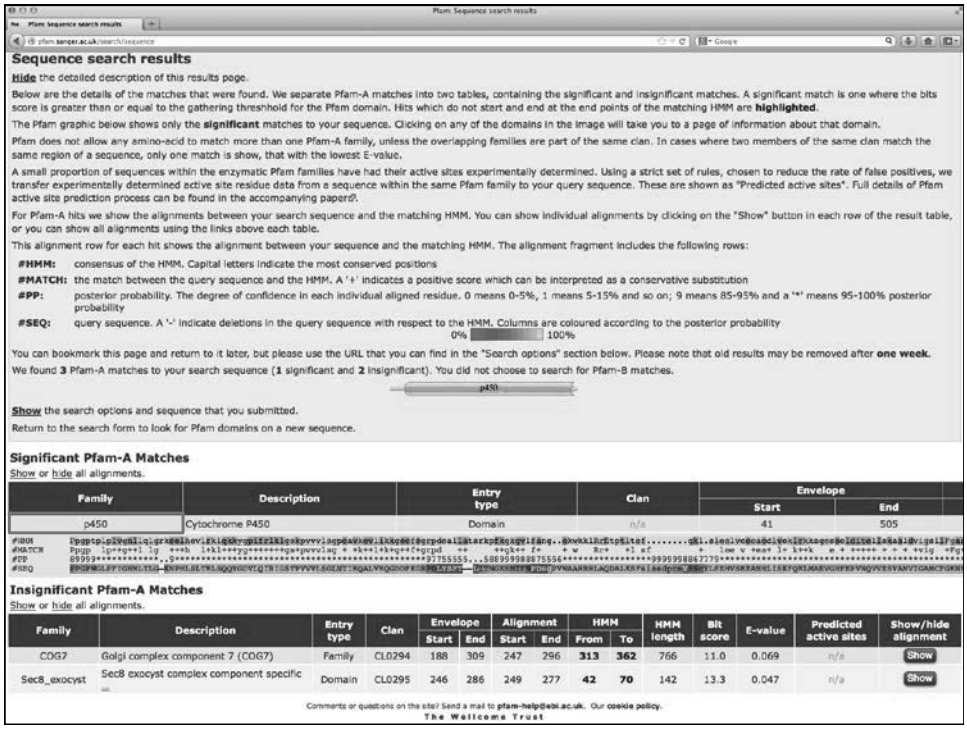

wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search Go
Pfam 27.0 (March 2013, 14831 families)
The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...
QUICK LINKS
ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES
SEQUENCE SEARCH
Paste your protein sequence here to find matching Pfam families.
VIEW A PFAM FAMILY
Go Example
VIEW A CLAN
VIEW A SEQUENCE
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.
VIEW A STRUCTURE
KEYWORD SEARCH
JUMP TO
Recent Pfam blog posts
Hide this
Short-term Pfam position available. (posted 7 February 2014)
We have just advertised a 9-month maternity cover position in Pfam. We are looking for a skilled Bioinformatician to help us take Pfam into its next phase of development as we become more integrated into the European Bioinformatics Institute (EMBL-EBI). Essential knowledge, skills and experience: Degree in Science with relevant experience Computer literacy (unix experience) [...]
Join Rfam, see the world (posted 31 January 2014)
Rfam is recruiting! We are currently recruiting an RNA informatician to join our team. We're looking for someone really enthusiastic about RNA and who's interested in working with Rfam as we move to

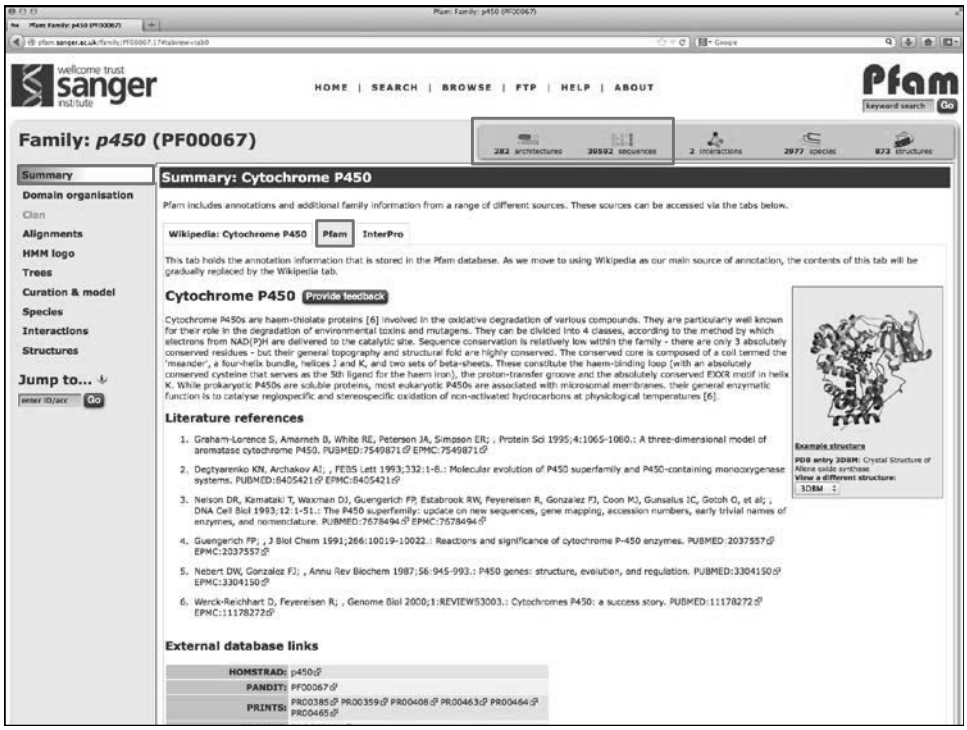


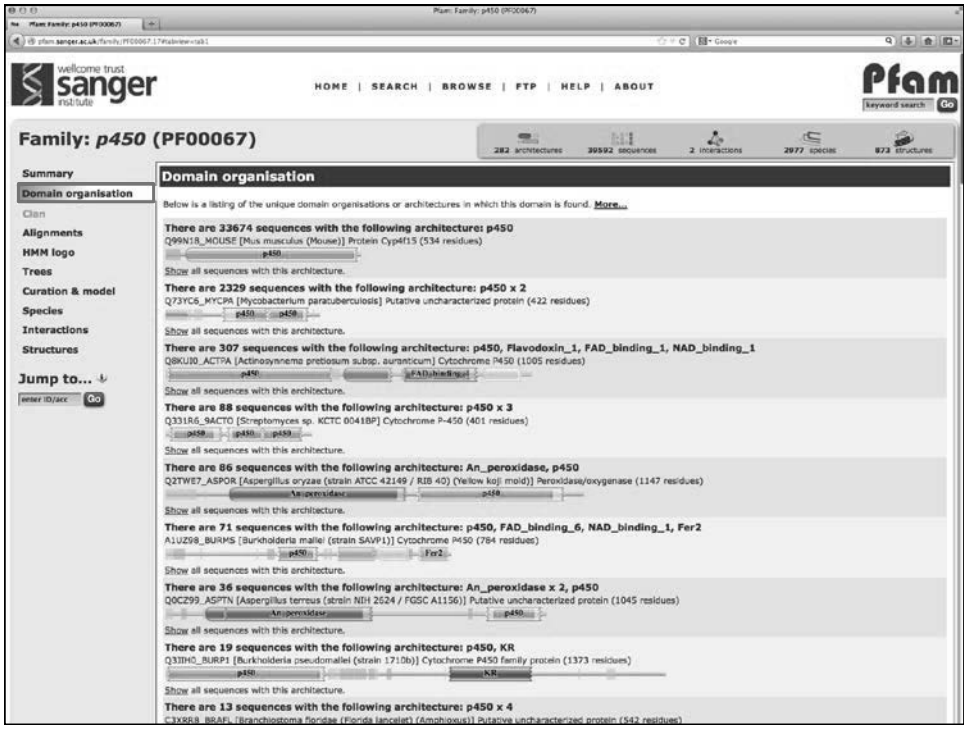


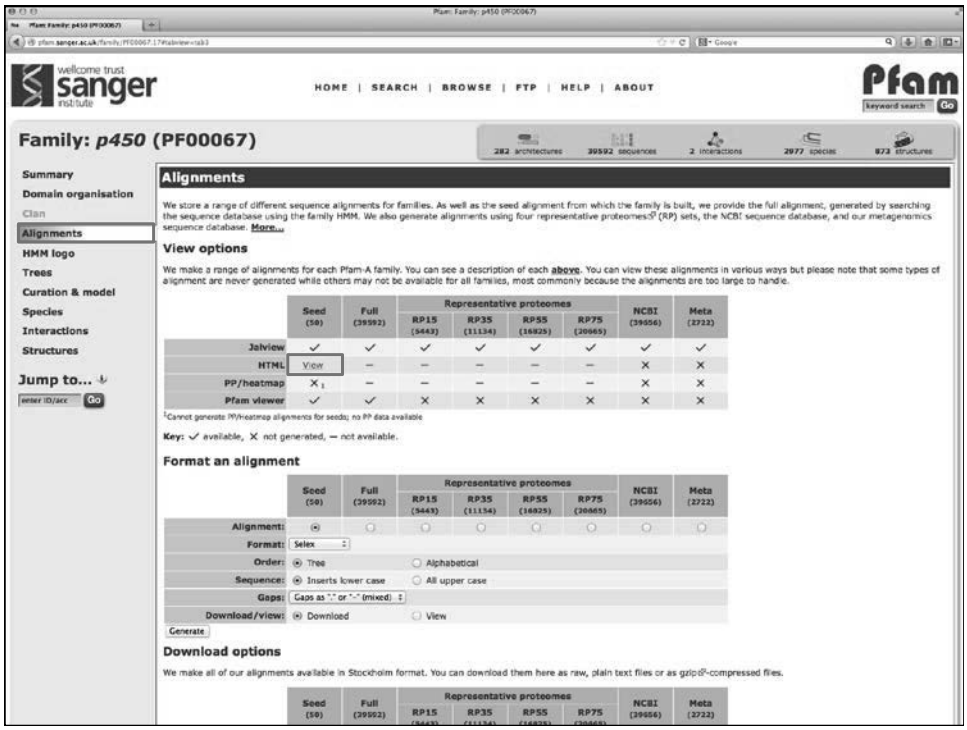
Family	Description	Entry type	Clan	Envelope Start	Envelope End	Alignment Start	Alignment End	HMM From	HMM To	HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
p450	Cytochrome P450	Domain	n/a	41	505	41	500	1	**457**	463	344.0	1.1e-102	n/a	Show

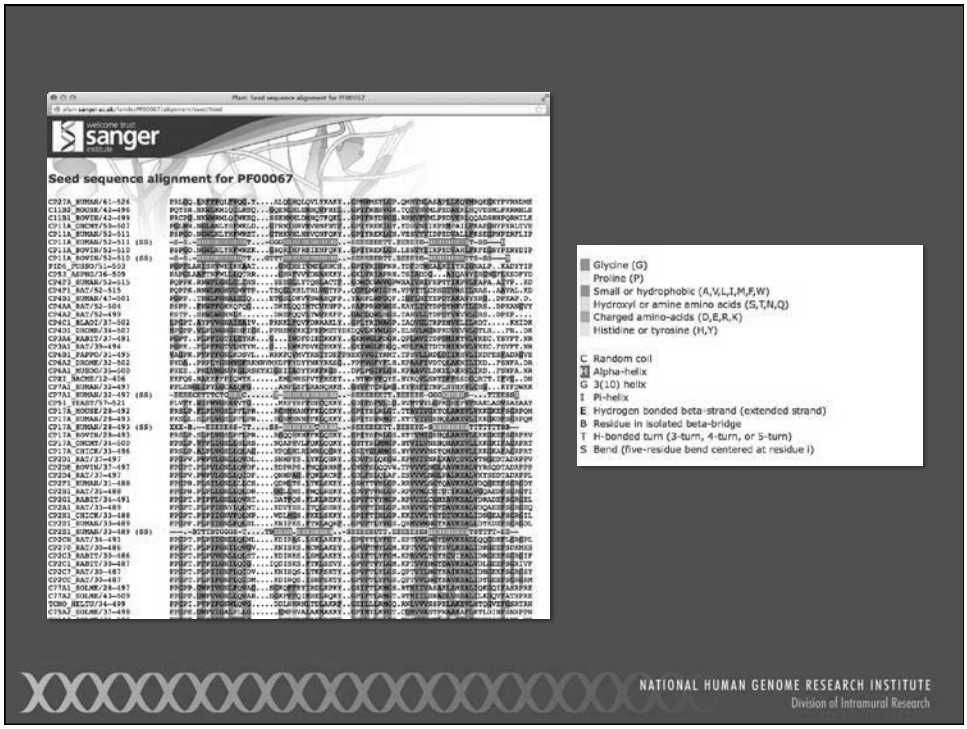
Family	Description	Entry type	Clan	Envelope Start	Envelope End	Alignment Start	Alignment End	HMM From	HMM To	HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
COG7	Golgi complex component 7 (COG7)	Family	CL0294	188	309	247	296	**313**	**362**	766	11.0	0.069	n/a	Show
Sec8_exocyst	Sec8 exocyst complex component specific	Domain	CL0295	246	286	249	277	**42**	**70**	142	13.3	0.047	n/a	Show

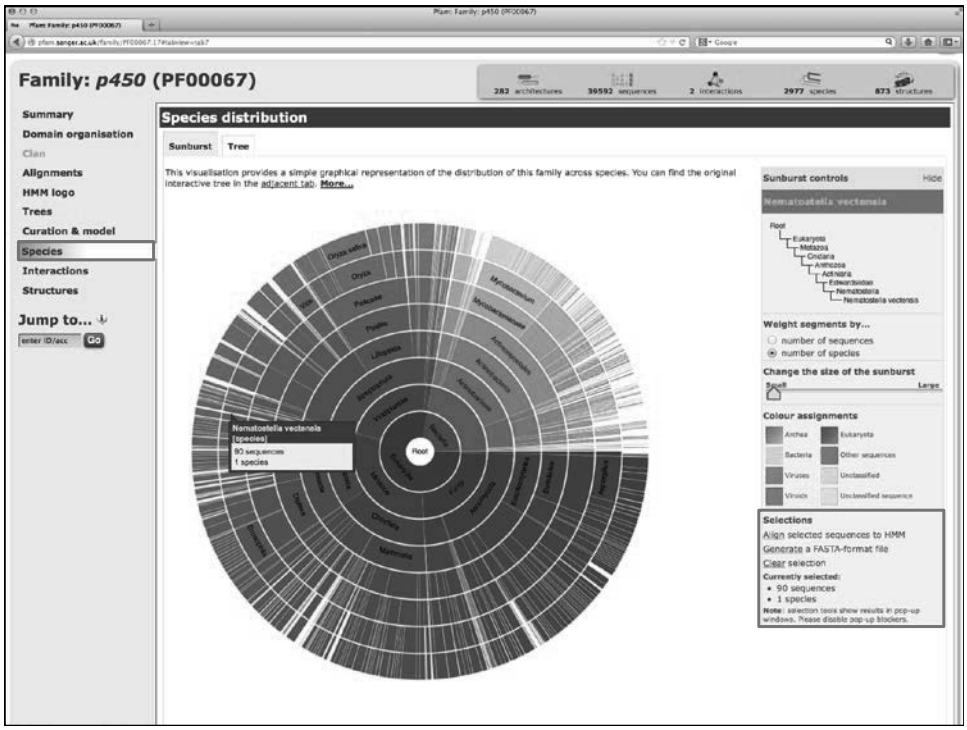

Sequence search results
Hide the detailed description of this results page.
Below are the details of the matches that were found. We separate Pfam-A matches into two tables, containing the significant and insignificant matches. A significant match is one where the bits score is greater than or equal 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.
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. 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). You did not choose to search for Pfam-B matches.
p450
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 Start | Envelope End
p450 | Cytochrome P450 | Domain | n/a | 41 | 505
Insignificant Pfam-A Matches
Show or hide all alignments.
Family | Description | Entry type | Clan | Envelope Start | Envelope End | Alignment Start | Alignment End | HMM From | HMM To | HMM length | Bit score | E-value | Predicted active sites | Show/hide alignment
COG7 | Golgi complex component 7 (COG7) | Family | CL0294 | 188 | 309 | 247 | 296 | 313 | 362 | 766 | 11.0 | 0.069 | n/a | Show
Sec8_exocyst | Sec8 exocyst complex component specific | Domain | CL0295 | 246 | 286 | 249 | 277 | 42 | 70 | 142 | 13.3 | 0.047 | n/a | Show
Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk. Our cookie policy.
The Wellcome Trust

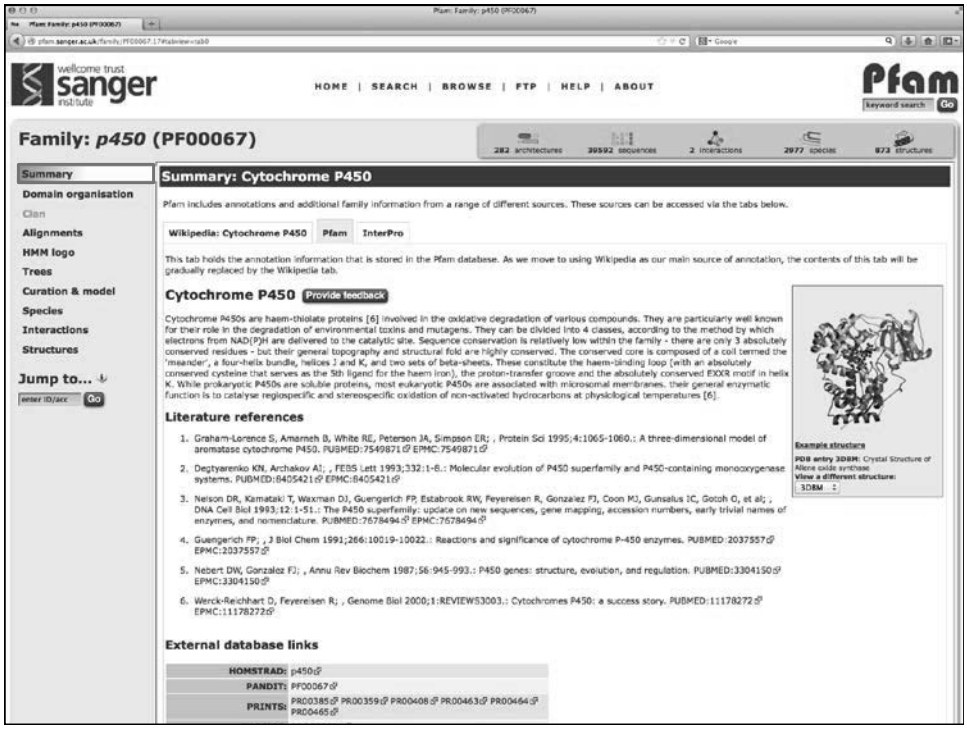

wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
Family: p450 (PF00067)
282 architectures
30592 sequences
2 interactions
2977 species
873 structures
Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to...
Summary: Cytochrome P450
Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.
Wikipedia: Cytochrome P450
Pfam
InterPro
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
Provide feedback
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].
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 EPMC:7549871
2. Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-6.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. PUBMED:8405421 EPMC: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 EPMC:7678494
4. Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. PUBMED:2037557 EPMC:2037557
5. Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. PUBMED:3304150 EPMC:3304150
6. Werck-Reichhart D, Feyereisen R; , Genome Biol 2000;1:REVIEW5003.: Cytochromes P450: a success story. PUBMED:11178272 EPMC:11178272
External database links
HOMSTRAD: p450
PANDIT: PF00067
PRINTS: PR00385 PR00359 PR00408 PR00463 PR00464 PR00465

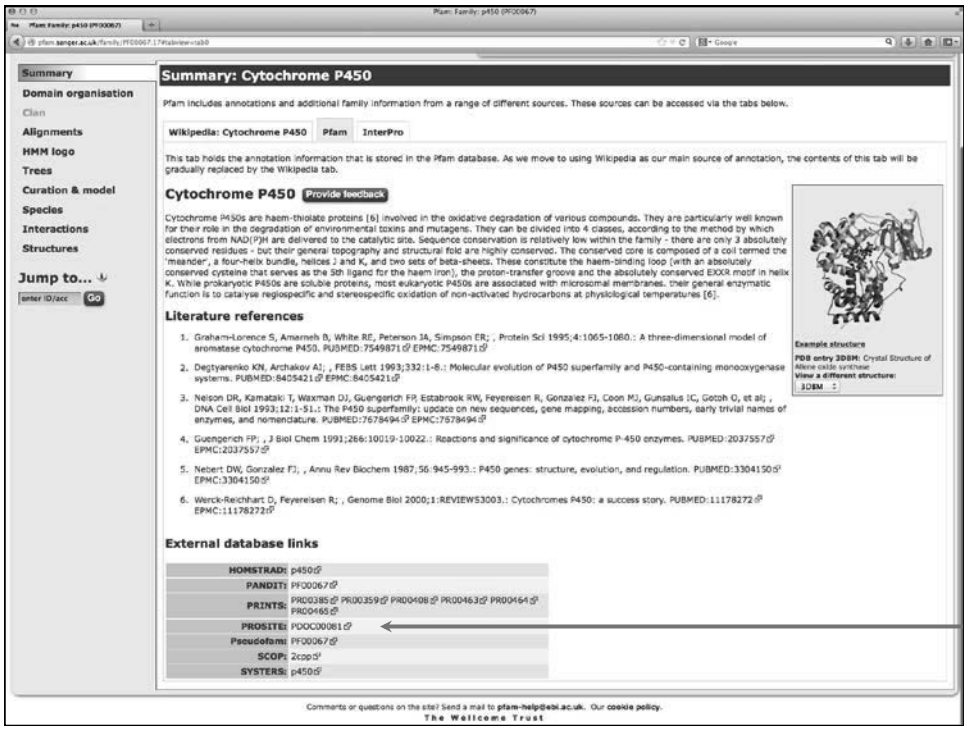

Pfam: Family: p450 (PF00067)
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
Family: p450 (PF00067)
282 architectures
39592 sequences
2 interactions
2977 species
873 structures
Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to...
Domain organisation
Below is a listing of the unique domain organisations or architectures in which this domain is found. More...
There are 33674 sequences with the following architecture: p450
Q99N18_MOUSE [Mus musculus (Mouse)] Protein Cyp4f15 (534 residues)
Show all sequences with this architecture.
There are 2329 sequences with the following architecture: p450 x 2
Q73YC6_MYCPA [Mycobacterium paratuberculosis] Putative uncharacterized protein (422 residues)
Show all sequences with this architecture.
There are 307 sequences with the following architecture: p450, Flavodoxin_1, FAD_binding_1, NAD_binding_1
Q8KUI0_ACTPA [Actinosynnema pretiosum subsp. auranticum] Cytochrome P450 (1005 residues)
Show all sequences with this architecture.
There are 88 sequences with the following architecture: p450 x 3
Q331R6_9ACTO [Streptomyces sp. KCTC 0041BP] Cytochrome P-450 (401 residues)
Show all sequences with this architecture.
There are 86 sequences with the following architecture: An_peroxidase, p450
Q2TWE7_ASPOR [Aspergillus oryzae (strain ATCC 42149 / RIB 40) (Yellow koji mold)] Peroxidase/oxygenase (1147 residues)
Show all sequences with this architecture.
There are 71 sequences with the following architecture: p450, FAD_binding_6, NAD_binding_1, Fer2
A1UZ98_BURMS [Burkholderia mallei (strain SAVP1)] Cytochrome P450 (784 residues)
Show all sequences with this architecture.
There are 36 sequences with the following architecture: An_peroxidase x 2, p450
Q0CZ99_ASPTN [Aspergillus terreus (strain NIH 2624 / FGSC A1156)] Putative uncharacterized protein (1045 residues)
Show all sequences with this architecture.
There are 19 sequences with the following architecture: p450, KR
Q33IH0_BURP1 [Burkholderia pseudomallei (strain 1710b)] Cytochrome P450 family protein (1373 residues)
Show all sequences with this architecture.
There are 13 sequences with the following architecture: p450 x 4


Pfam: Family: p450 (PF00067)
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
Family: p450 (PF00067)
282 architectures
39592 sequences
2 interactions
2977 species
873 structures
Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to...
Alignments
We store a range of different sequence alignments for families. As well as the seed alignment from which the family is built, we provide the full alignment, generated by searching the sequence database using the family HMM. We also generate alignments using four representative proteomes (RP) sets, the NCBI sequence database, and our metagenomics sequence database. More...
View options
We make a range of alignments for each Pfam-A family. You can see a description of each above. You can view these alignments in various ways but please note that some types of alignment are never generated while others may not be available for all families, most commonly because the alignments are too large to handle.
Seed (50) | Full (39592) | Representative proteomes: RP15 (5443) | RP35 (11134) | RP55 (16825) | RP75 (20665) | NCBI (39656) | Meta (2722)
Jalview | HTML | PP/heatmap | Pfam viewer
View
Key: ✓ available, ✗ not generated, – not available.
Format an alignment
Alignment:
Format: Selex
Order: Tree / Alphabetical
Sequence: Inserts lower case / All upper case
Gaps: Gaps as "." or "-" (mixed)
Download/view: Download / View
Generate
Download options
We make all of our alignments available in Stockholm format. You can download them here as raw, plain text files or as gzip-compressed files.


sanger
Seed sequence alignment for PF00067
Glycine (G)
Proline (P)
Small or hydrophobic (A,V,L,I,M,F,W)
Hydroxyl or amine amino acids (S,T,N,Q)
Charged amino-acids (D,E,R,K)
Histidine or tyrosine (H,Y)
C Random coil
H Alpha-helix
G 3(10) helix
I Pi-helix
E Hydrogen bonded beta-strand (extended strand)
B Residue in isolated beta-bridge
T H-bonded turn (3-turn, 4-turn, or 5-turn)
S Bend (five-residue bend centered at residue i)
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research


Family: p450 (PF00067)
Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to...
Species distribution
Sunburst
Tree
Sunburst controls
Weight segments by...
number of sequences
number of species
Change the size of the sunburst
Colour assignments
Selections
Nematostella vectensis (species)
90 sequences
1 species
Root


Pfam: Family: p450 (PF00067)
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search Go
Family: p450 (PF00067)
282 architectures 39592 sequences 2 interactions 2977 species 873 structures
Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to...
enter ID/acc Go
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.
Wikipedia: Cytochrome P450 Pfam InterPro
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 Provide feedback
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 3D8M: Crystal Structure of Allene oxide synthase
View a different structure:
3D8M
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 EPMC:7549871
2. Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. PUBMED:8405421 EPMC: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 EPMC:7678494
4. Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. PUBMED:2037557 EPMC:2037557
5. Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. PUBMED:3304150 EPMC:3304150
6. Werck-Reichhart D, Feyereisen R; , Genome Biol 2000;1:REVIEWS3003.: Cytochromes P450: a success story. PUBMED:11178272 EPMC:11178272
External database links
HOMSTRAD: p450
PANDIT: PF00067
PRINTS: PR00385 PR00359 PR00408 PR00463 PR00464 PR00465


Pfam: Family: p450 (PF00067)
Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to...
enter ID/acc Go
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.
Wikipedia: Cytochrome P450 Pfam InterPro
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 Provide feedback
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 3D8M: Crystal Structure of Allene oxide synthase
View a different structure:
3D8M
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 EPMC:7549871
2. Degtyarenko KN, Archakov AI; , FEBS Lett 1993;332:1-8.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. PUBMED:8405421 EPMC: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 EPMC:7678494
4. Guengerich FP; , J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. PUBMED:2037557 EPMC:2037557
5. Nebert DW, Gonzalez FJ; , Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. PUBMED:3304150 EPMC:3304150
6. Werck-Reichhart D, Feyereisen R; , Genome Biol 2000;1:REVIEWS3003.: Cytochromes P450: a success story. PUBMED:11178272 EPMC: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@ebi.ac.uk. Our cookie policy.
The Wellcome Trust

PROSITE

Home | ScanProsite | ProRule | Documents | Downloads | Links | Funding

PROSITE documentation PDOC00081

Cytochrome P450 cysteine heme-iron ligand signature

Description | Technical section | References | Copyright | Miscellaneous

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 in UniProtKB/Swiss-Prot known to belong to this class: 992
  - detected by PS00086: 922 (true positives)
  - undetected by PS00086: 70 (60 false negatives and 10 'partials')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS00086: 46 false positives.
- Retrieve an alignment of UniProtKB/Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, plain text / Fasta format
- Retrieve the sequence logo from the alignment
- Taxonomic tree view of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS00086
- Retrieve a list of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS00086
- Scan UniProtKB (Swiss-Prot and/or TrEMBL) entries against PS00086
- View ligand binding statistics of PS00086
- Matching PDB structures: 1AKD 1BU7 1BVY 1C8J ... [ALL]

Conserved Domain Database (CDD)

- Identify conserved domains in a protein sequence
- Incorporates three-dimensional structural information to define domain boundaries and refine alignments
- Source data derived from:
  - 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

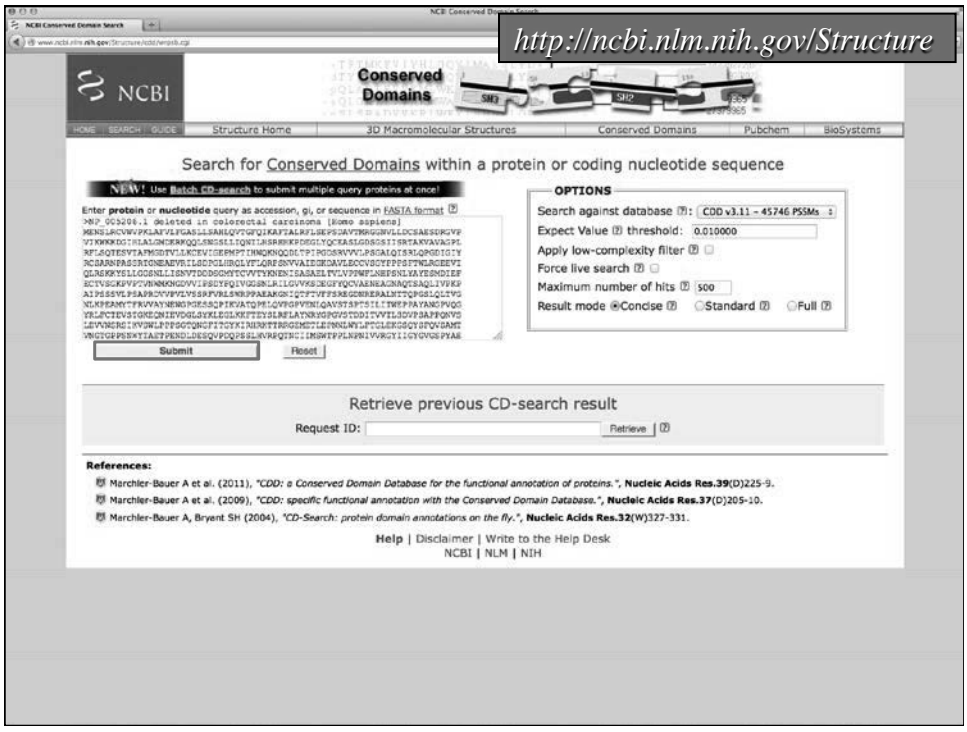
Marchler-Bauer et al., Nucleic Acids Res. 41: D348-D352, 2013

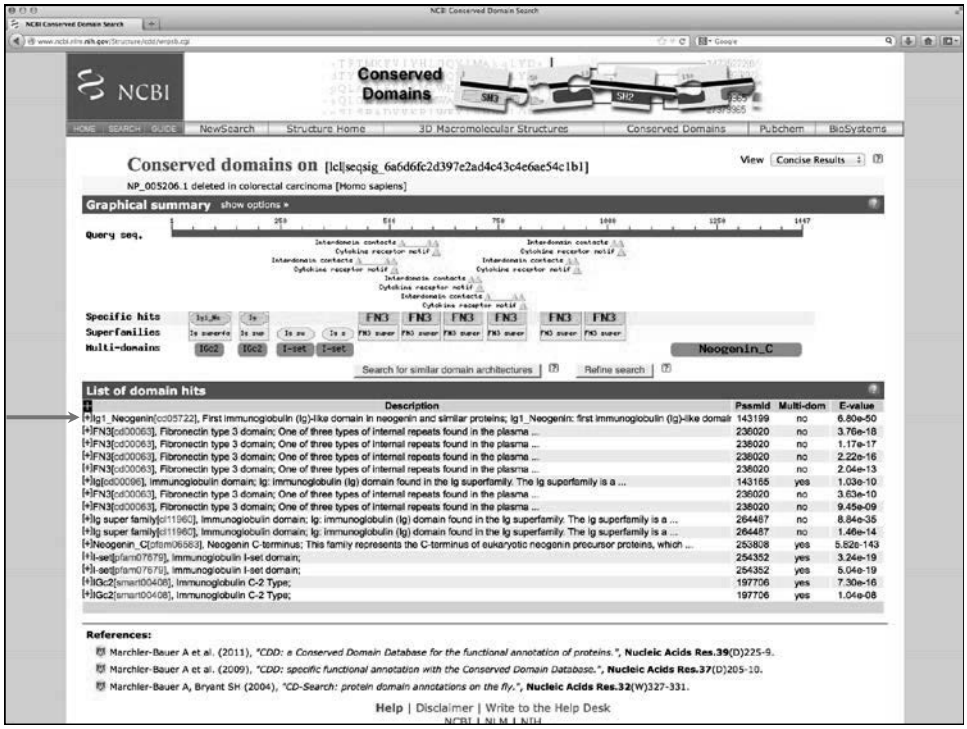
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

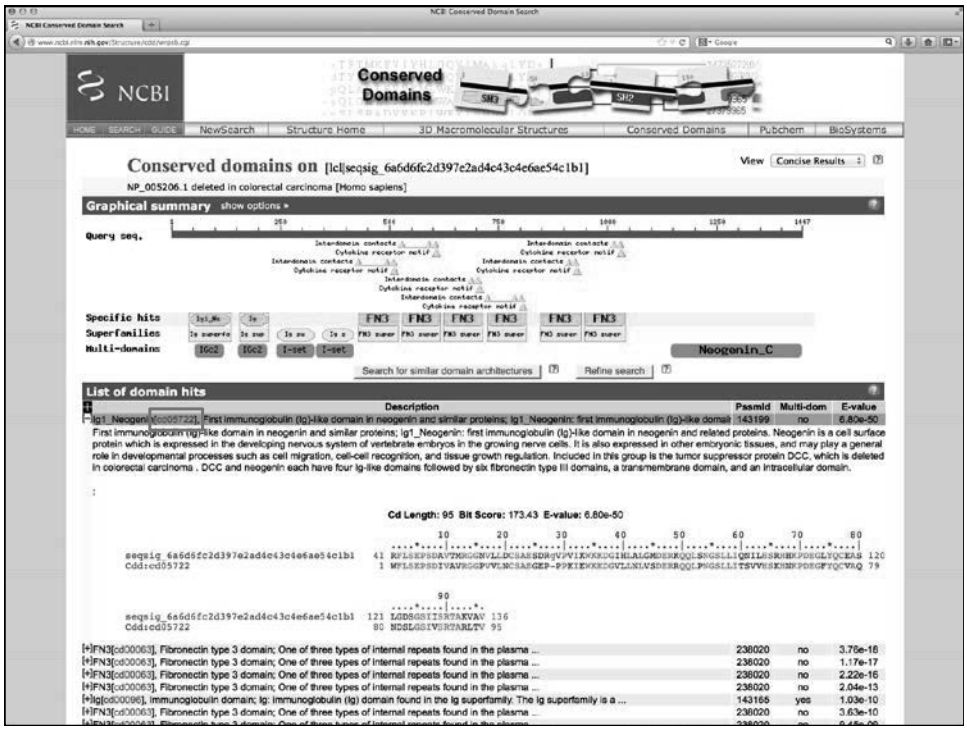
Conserved Domain Database (CDD)

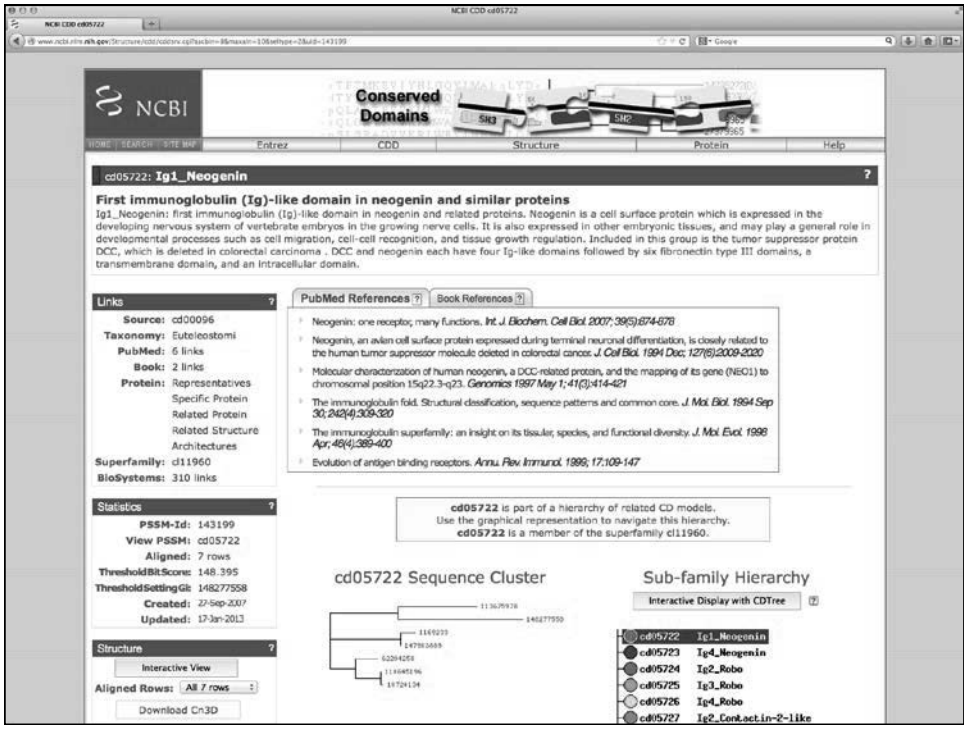
- CD-Search performed using RPS-BLAST
- Query sequence is used to search a database of precalculated position-specific scoring matrices
- *Not* the same method used by Pfam

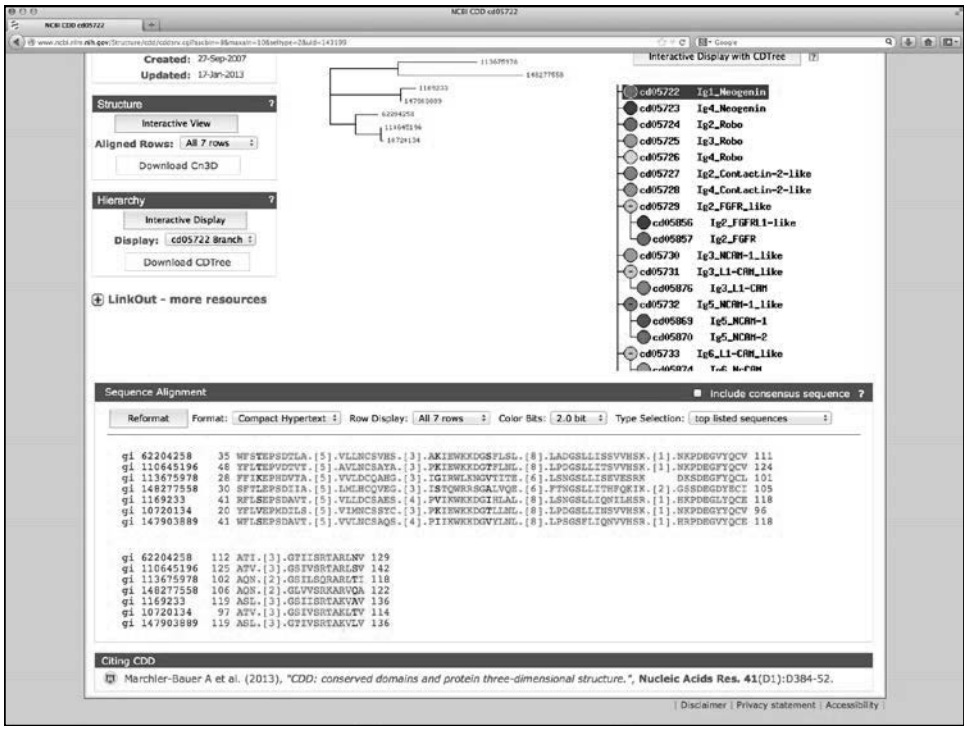
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research




NCBI
Conserved Domains
Conserved domains on [lcl|seqsig_6a6d6fc2d397e2ad4c43c4e6ae54c1b1]
NP_005206.1 deleted in colorectal carcinoma [Homo sapiens]
Graphical summary
Query seq.
Specific hits
Superfamilies
Multi-domains
Neogenin_C
List of domain hits
Description
Pssmid
Multi-dom
E-value
References:
Help | Disclaimer | Write to the Help Desk


NCBI
Conserved Domains
Conserved domains on [lcl|seqsig_6a6d6fc2d397e2ad4c43c4e6ae54c1b1]
NP_005206.1 deleted in colorectal carcinoma [Homo sapiens]
Graphical summary
Query seq.
Specific hits
Superfamilies
Multi-domains
Neogenin_C
List of domain hits
Description
Pssmid
Multi-dom
E-value
Cd Length: 95 Bit Score: 173.43 E-value: 6.80e-50


NCBI
Conserved Domains
Entrez
CDD
Structure
Protein
Help
cd05722: Ig1_Neogenin
First immunoglobulin (Ig)-like domain in neogenin and similar proteins
Links
Source: cd00096
Taxonomy: Euteleostomi
PubMed: 6 links
Book: 2 links
Protein: Representatives
Specific Protein
Related Protein
Related Structure
Architectures
Superfamily: cl11960
BioSystems: 310 links
PubMed References
Book References
Statistics
PSSM-Id: 143199
View PSSM: cd05722
Aligned: 7 rows
ThresholdBitScore: 148.395
ThresholdSettingGi: 148277558
Created: 27-Sep-2007
Updated: 17-Jan-2013
Structure
Interactive View
Aligned Rows: All 7 rows
Download Cn3D
cd05722 is part of a hierarchy of related CD models.
Use the graphical representation to navigate this hierarchy.
cd05722 is a member of the superfamily cl11960.
cd05722 Sequence Cluster
Sub-family Hierarchy
Interactive Display with CDTree
cd05722 Ig1_Neogenin
cd05723 Ig4_Neogenin
cd05724 Ig2_Robo
cd05725 Ig3_Robo
cd05726 Ig4_Robo
cd05727 Ig2_Contactin-2-like


Created: 27-Sep-2007
Updated: 17-Jan-2013
Structure
Interactive View
Aligned Rows: All 7 rows
Download Cn3D
Hierarchy
Interactive Display
Display: cd05722 Branch
Download CDTree
LinkOut - more resources
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
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
Sequence Alignment
Include consensus sequence
Reformat
Format: Compact Hypertext
Row Display: All 7 rows
Color Bits: 2.0 bit
Type Selection: top listed sequences
Citing CDD
Marchler-Bauer A et al. (2013), "CDD: conserved domains and protein three-dimensional structure.", Nucleic Acids Res. 41(D1):D384-52.
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, CDD*
    or "many-against-one": *PSI-BLAST, DELTA-BLAST*

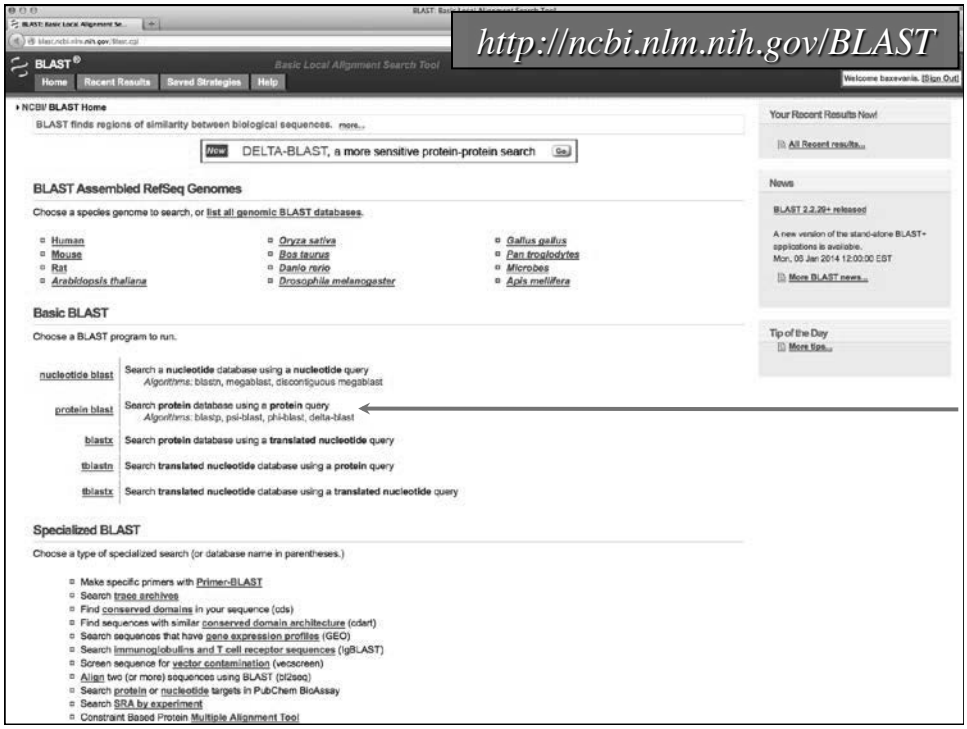
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

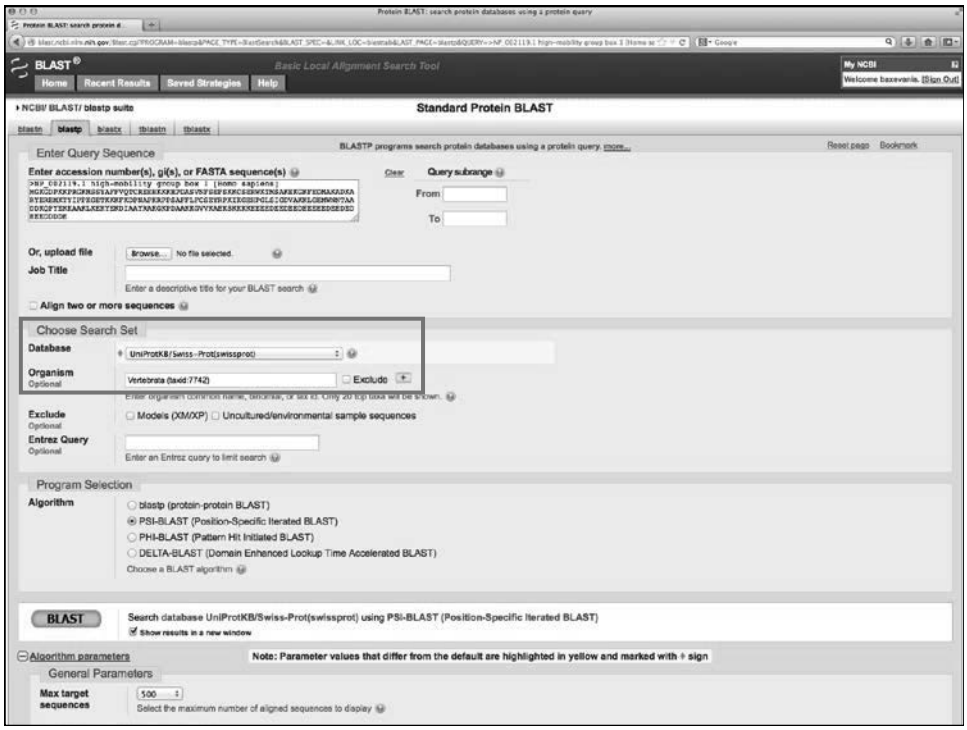
PSI-BLAST

- Position-Specific Iterated BLAST search
- Used to identify distantly related sequences that are possibly missed during a standard 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

Altschul et al., Nucleic Acids Res. 25: 3389-3402, 1997

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

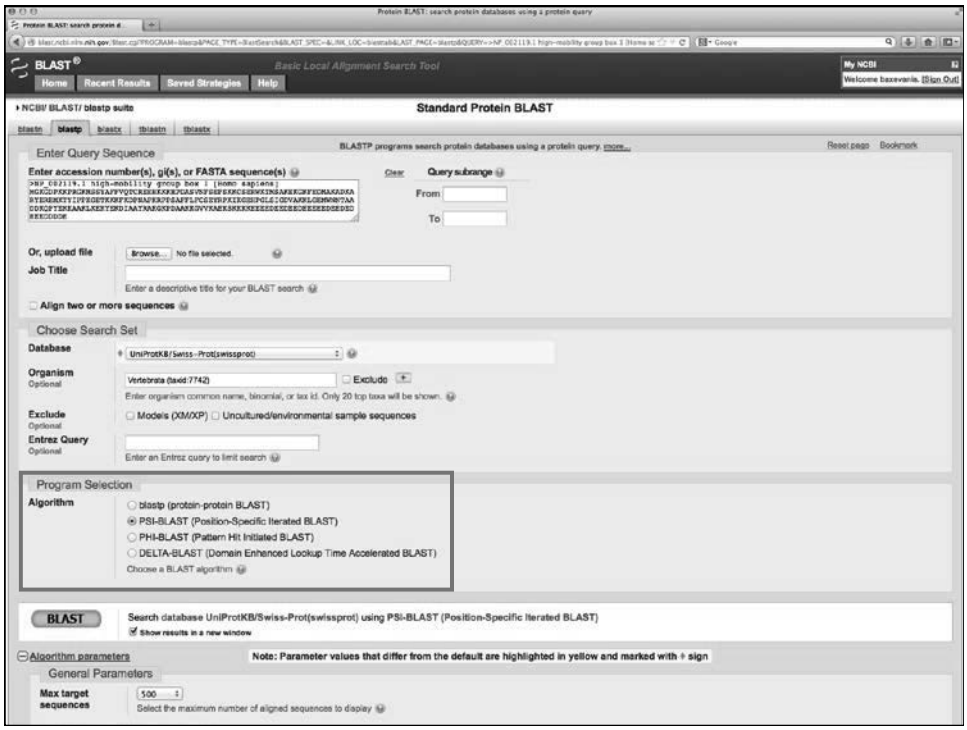

http://ncbi.nlm.nih.gov/BLAST
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI BLAST Home
BLAST finds regions of similarity between biological sequences. more...
DELTA-BLAST, a more sensitive protein-protein search
BLAST Assembled RefSeq Genomes
Choose a species 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, delta-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 and T cell receptor sequences (IgBLAST)
Screen sequence for vector contamination (vecscreen)
Align two (or more) sequences using BLAST (bl2seq)
Search protein or nucleotide targets in PubChem BioAssay
Search SRA by experiment
Constraint Based Protein Multiple Alignment Tool
Your Recent Results Now!
All Recent results...
News
BLAST 2.2.29+ released
A new version of the stand-alone BLAST+ applications is available.
Mon, 06 Jan 2014 12:00:00 EST
More BLAST news...
Tip of the Day
More tips...

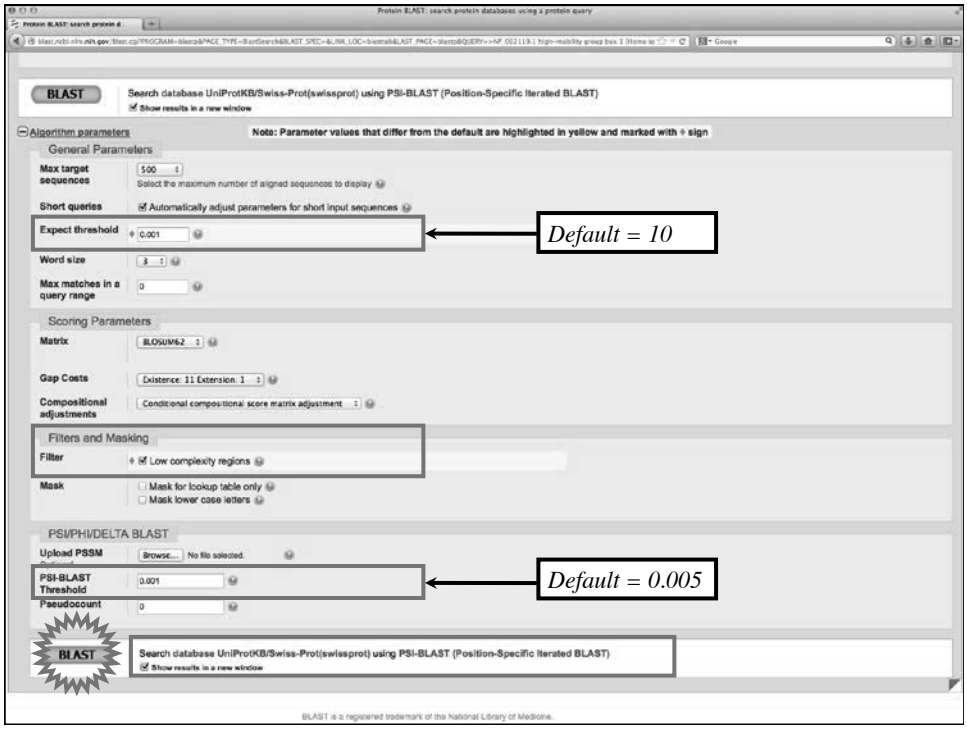

BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI BLAST/ blastp suite
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein query. more...
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Browse...
No file selected.
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
UniProtKB/Swiss-Prot(swissprot)
Organism
Optional
Vertebrata (taxid:7742)
Exclude
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)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database UniProtKB/Swiss-Prot(swissprot) using PSI-BLAST (Position-Specific Iterated BLAST)
Show results in a new window
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign
General Parameters
Max target sequences
500
Select the maximum number of aligned sequences to display

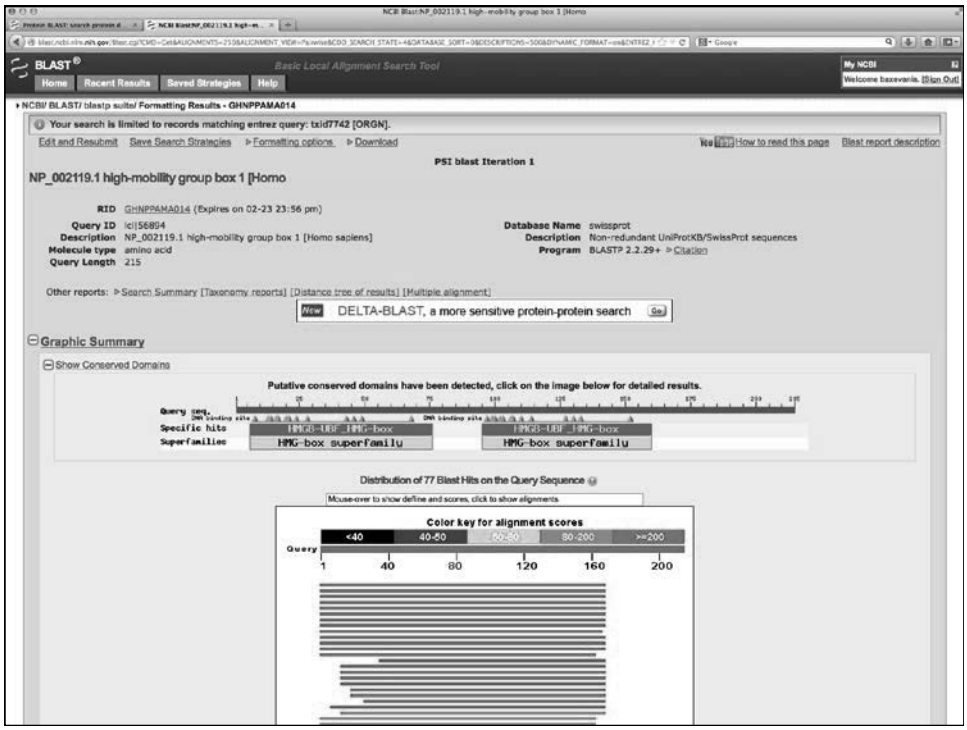
Swiss-Prot

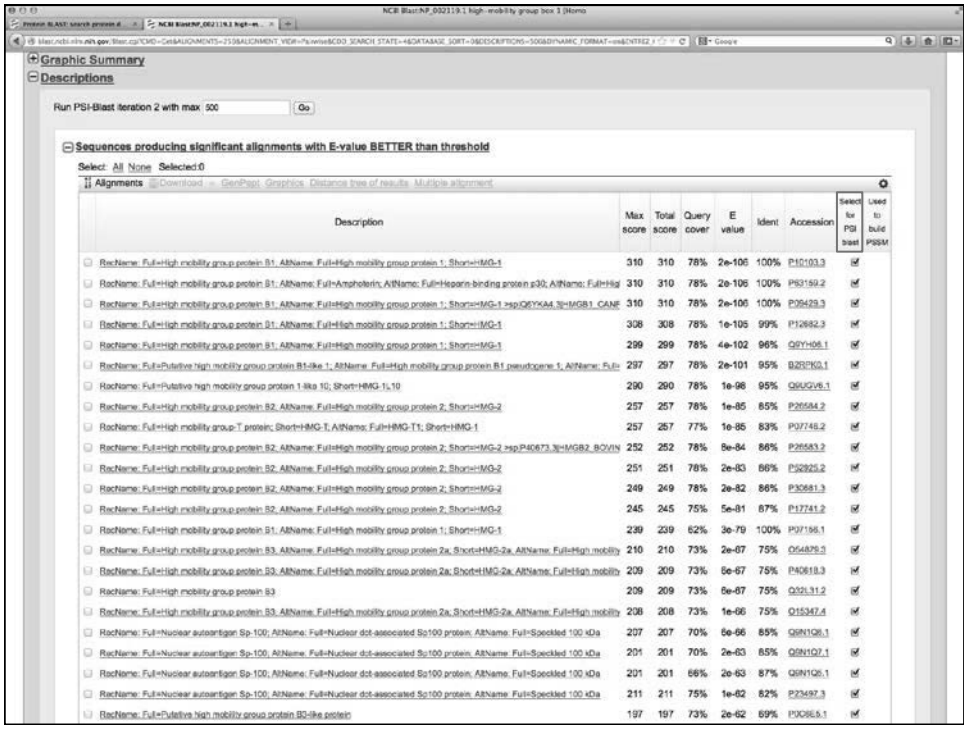
- *Goal:* Provide a single reference sequence for each protein sequence
- Distinguishing Features
  - Non-redundancy
  - Ongoing curation by EBI staff and *external experts*
  - Expert annotation includes editing/updates of
    - `KW` Keyword lines
    - `CC` Comment lines (the "executive summary")
    - `FT` Feature table
  - Distinct accession series
    `[OPQ]12345`

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

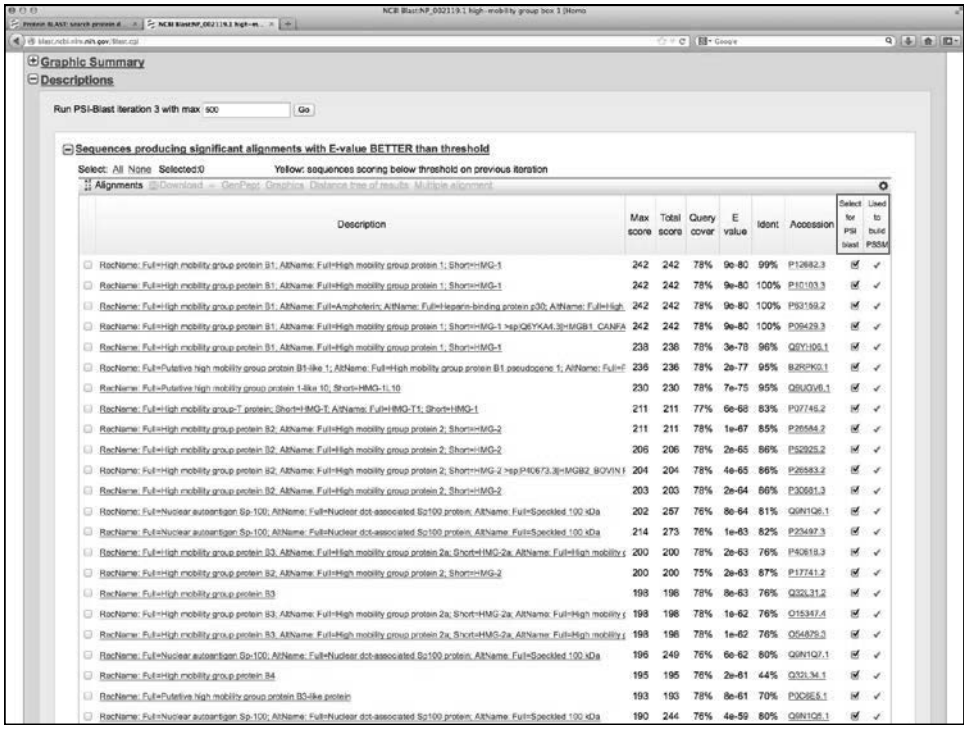


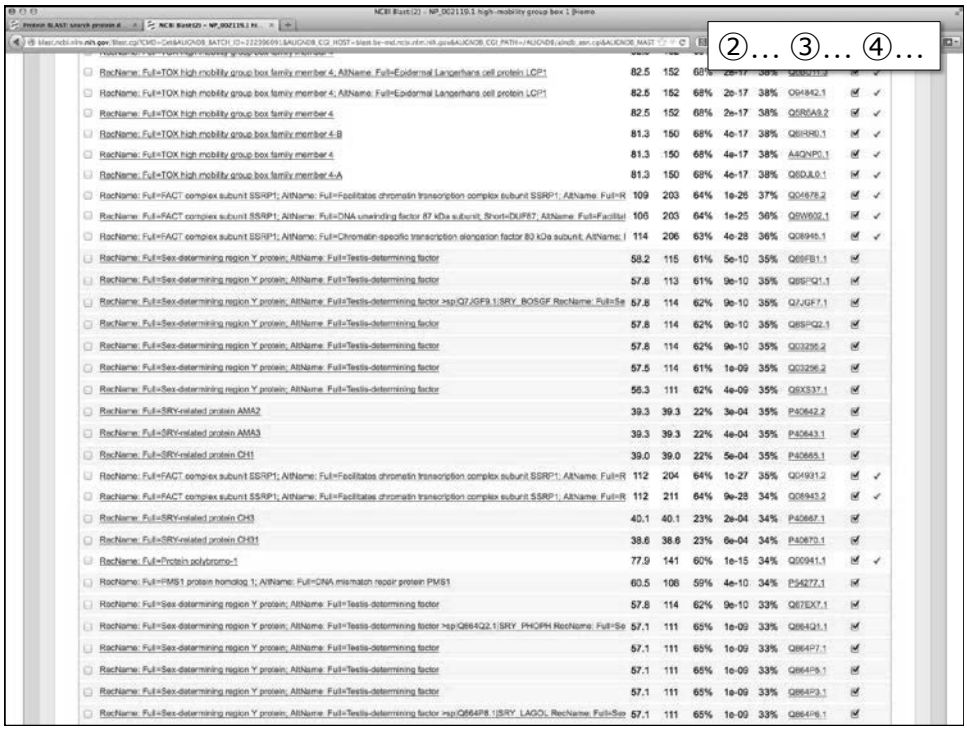

BLAST
Search database UniProtKB/Swiss-Prot(swissprot) using PSI-BLAST (Position-Specific Iterated BLAST)
Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign
General Parameters
Max target sequences
Short queries
Expect threshold
Default = 10
Word size
Max matches in a query range
Scoring Parameters
Matrix
Gap Costs
Compositional adjustments
Filters and Masking
Filter
Low complexity regions
Mask
PSI/PHI/DELTA BLAST
Upload PSSM
PSI-BLAST Threshold
Default = 0.005
Pseudocount
BLAST
Search database UniProtKB/Swiss-Prot(swissprot) using PSI-BLAST (Position-Specific Iterated BLAST)

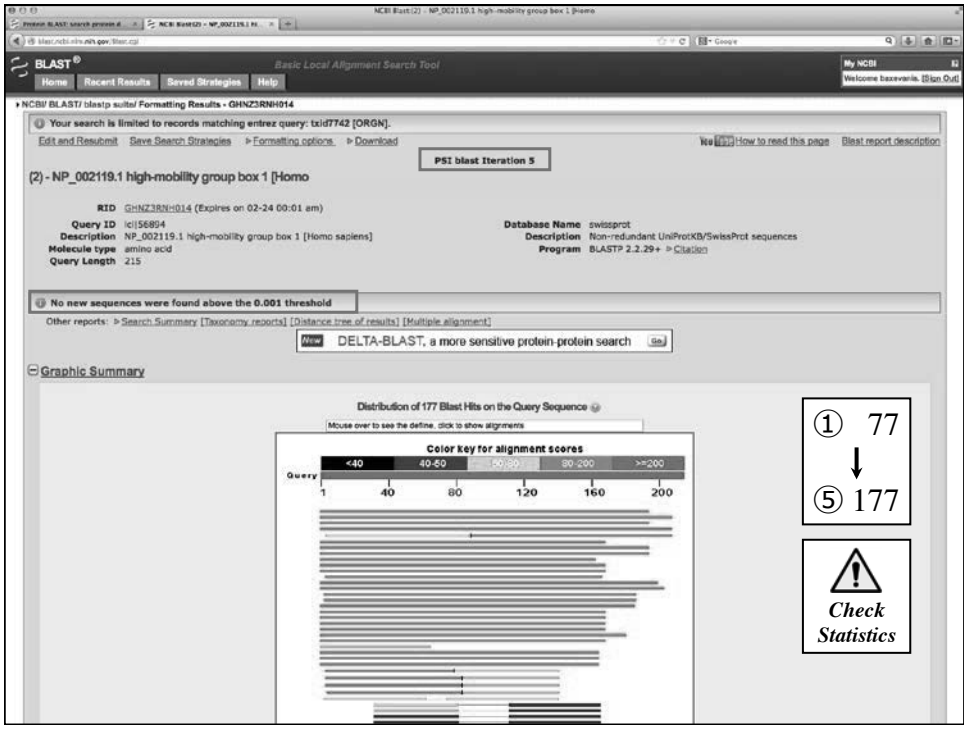

BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
Your search is limited to records matching entrez query: txid7742 [ORGN].
Edit and Resubmit
Save Search Strategies
Formatting options
Download
How to read this page
Blast report description
PSI blast Iteration 1
NP_002119.1 high-mobility group box 1 [Homo
RID
Query ID
Description
Molecule type
Query Length
Database Name
Program
Other reports:
DELTA-BLAST, a more sensitive protein-protein search
Graphic Summary
Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
Specific hits
Superfamilies
HMG-box superfamily
Distribution of 77 Blast Hits on the Query Sequence
Color key for alignment scores
<40
40-80
80-200
>=200
Query
1
40
80
120
160
200


Graphic Summary
Descriptions
Run PSI-Blast iteration 2 with max 500
Go
Sequences producing significant alignments with E-value BETTER than threshold
Select: All None Selected:0
Alignments
Description
Max score
Total score
Query cover
E value
Ident
Accession
Select for PSI blast
Used to build PSSM


Run PSI-Blast iteration 2 with max 500
Go
Change cutoffs to show hits "below the line"


Graphic Summary
Descriptions
Run PSI-Blast iteration 3 with max 500
Go
Sequences producing significant alignments with E-value BETTER than threshold
Select: All None Selected:0
Yellow: sequences scoring below threshold on previous iteration
Alignments
Description
Max score
Total score
Query cover
E value
Ident
Accession
Select for PSI blast
Used to build PSSM


②… ③… ④…



DELTA-BLAST

- Method different from that used by PSI-BLAST

Step 1:	Align the query against conserved domains derived from CDD
Step 2:	Compute PSSM
Step 3:	Search sequence databases using PSSM as the query

- Intended to improve homology detection
- Produces high-quality alignments, even at low levels of sequence similarity
- Dependent on homologous relationships captured within CDD

Boratyn et al., Biology Direct 7: 12, 2012

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Multiple Sequence Alignment: A Quick Primer

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, thereby determining evolutionary relationships (inferring homology)
  - 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

Protein vs. Nucleotide Multiple Sequence Alignments

- Concentrate on the protein level rather than on the nucleotide level
- Protein alignments tend to be 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

1. 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: 50-100 sequences

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Selecting the Sequences

2. Sequences should be of about the same length
3. 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

4. Consider the degree of similarity in the sequence set, depending on what question is being asked
   - Use closely-related sequences to determine "required" (highly conserved) 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 is good, can add new sequences to data set, then realign
- If alignment is 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
- Gap-free blocks probably correspond to regions of secondary structure, while gap-rich blocks probably correspond to unstructured or loop regions

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Clustal Omega

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

Sievers et al., Mol. Syst. Biol. 7: 539, 2011

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Progressive Alignment

- Align two sequences at a time, starting with the two most related sequences
- Gradually build up the multiple sequence alignment by adding additional (less-related) sequences to the alignment
- 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 $[N \times (N - 1)] / 2$ 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
```

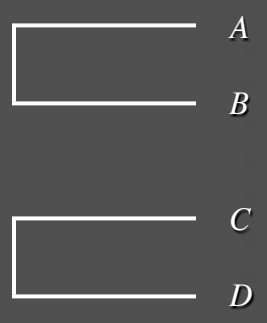
%ID	A	B	C	D
A	100			
B	80	100		
C	44	40	100	
D	40	40	92	100

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Progressive Alignment

2. Derive a guide tree based on the pairwise comparisons

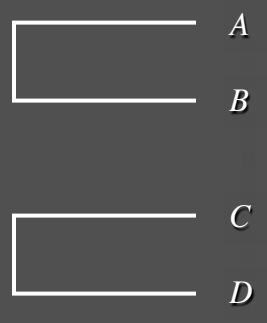
   Can infer from tree that A and B share greater similarity with each other than with C or D



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Progressive Alignment

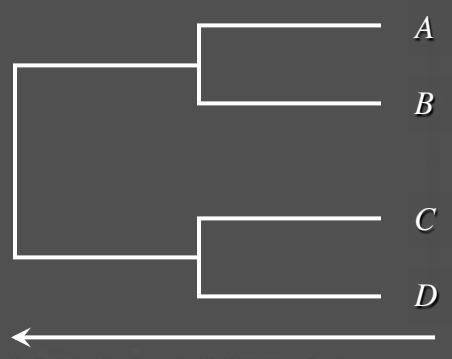
- Align A with B → alignment AB (fixed)
- Align C with D → alignment CD (fixed)
- Represent alignments AB and CD as *single sequences*



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"



Progressive Alignment: Advantages

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

Progressive Alignment: Disadvantages

- If initial alignments are made on distantly related sequences, there may be errors in the initial alignments
- Once an alignment is "fixed", it is not reconsidered, so any errors in the early alignments may propagate through subsequent alignments
- Clustal Omega does allow for guide tree iterations to hedge against errors introduced early in the alignment process (at the cost of increased compute time)

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Clustal Omega Output

- Pairwise alignment scores
- Multiple sequence alignment
- 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

Clustal Omega 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

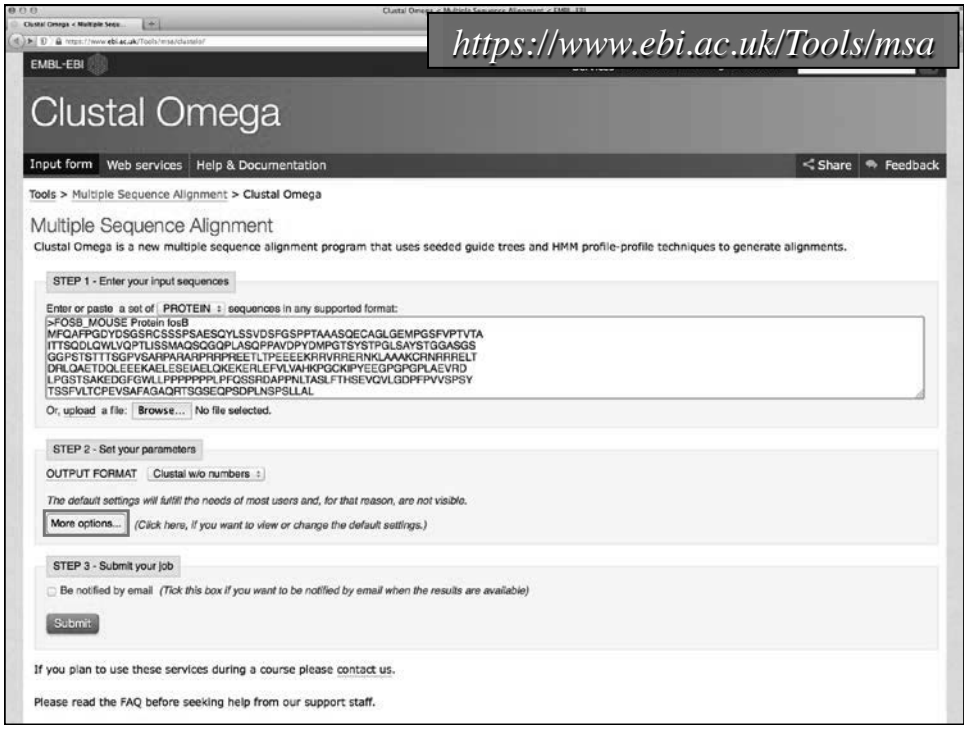
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

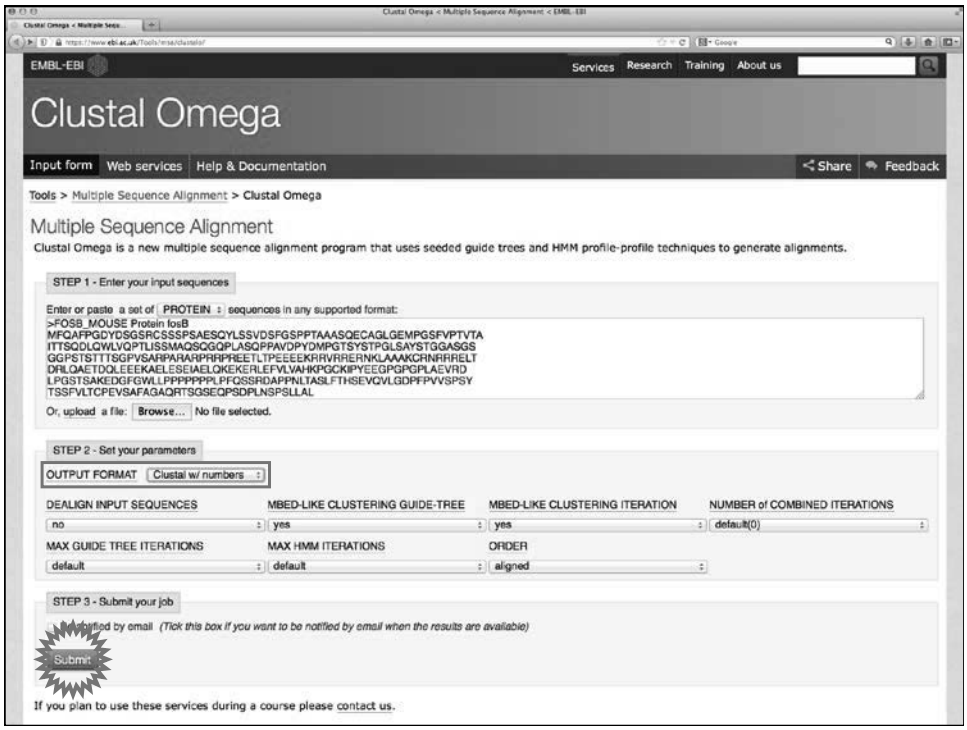
Clustal Omega Conservation Patterns

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

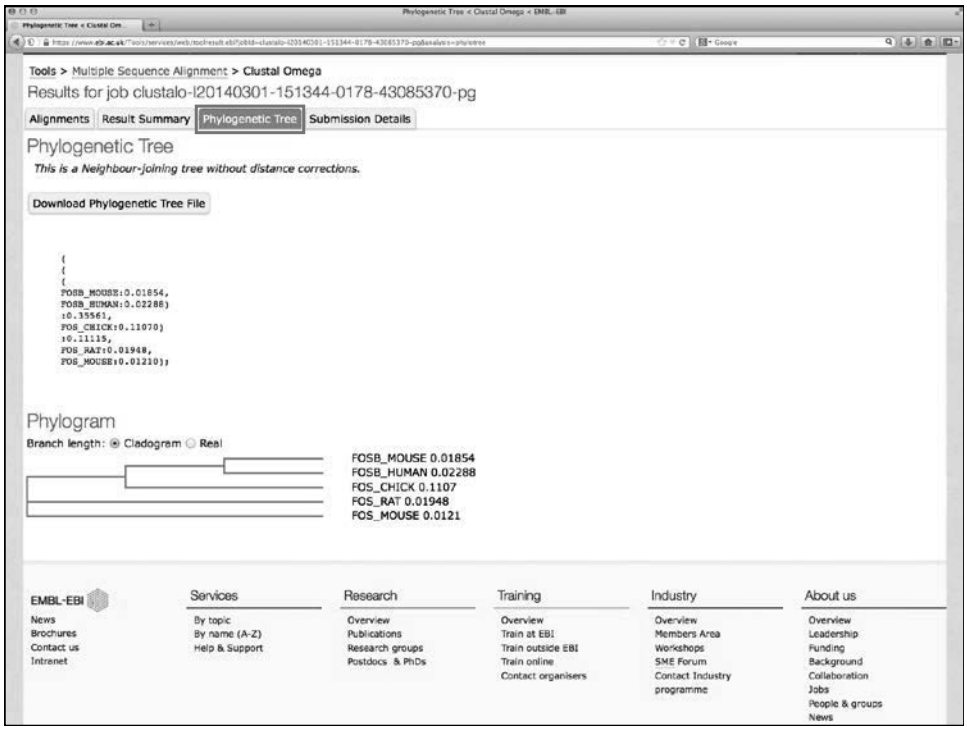
*	entirely conserved column (want in at least 10% of positions)
:	"conserved" (strongly similar properties)
.	"semi-conserved" (weakly similar properties)

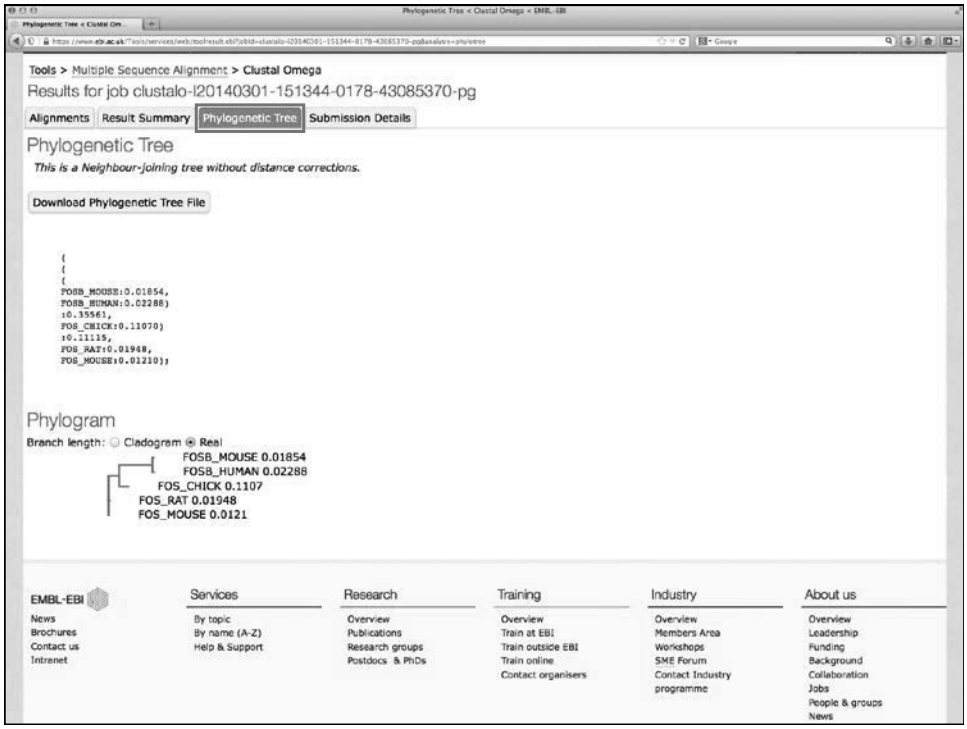
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

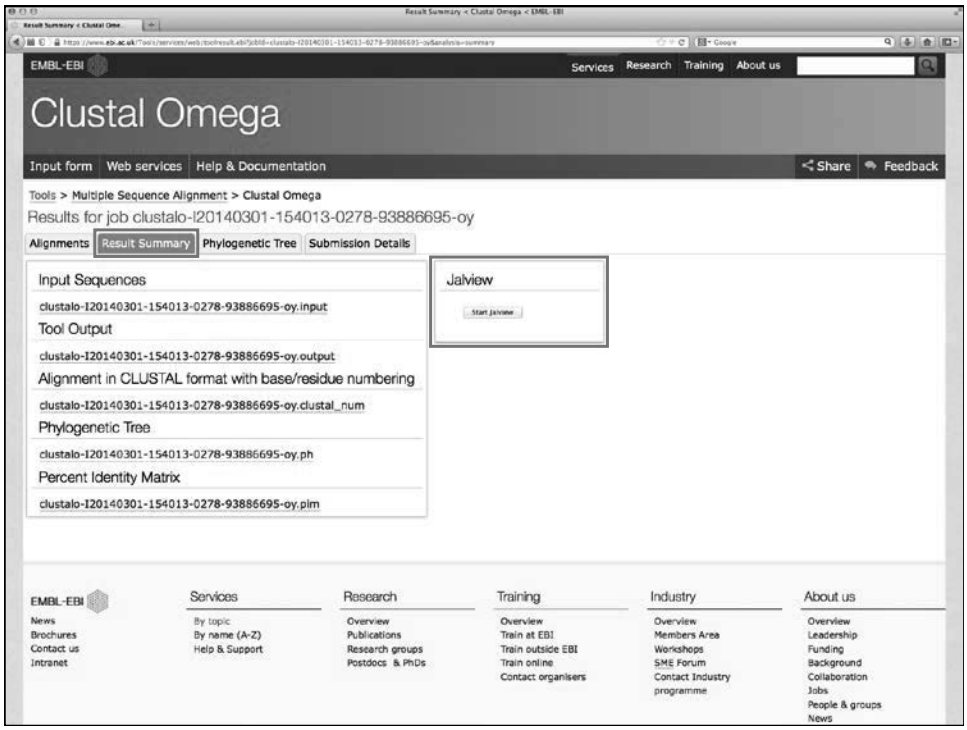

https://www.ebi.ac.uk/Tools/msa
EMBL-EBI
Clustal Omega
Input form
Web services
Help & Documentation
Share
Feedback
Tools > Multiple Sequence Alignment > Clustal Omega
Multiple Sequence Alignment
Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.
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: Browse... No file selected.
STEP 2 - Set your parameters
OUTPUT FORMAT Clustal w/o numbers
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 - 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.


EMBL-EBI
Services
Research
Training
About us
Clustal Omega
Input form
Web services
Help & Documentation
Share
Feedback
Tools > Multiple Sequence Alignment > Clustal Omega
Multiple Sequence Alignment
Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.
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: Browse... No file selected.
STEP 2 - Set your parameters
OUTPUT FORMAT Clustal w/ numbers
DEALIGN INPUT SEQUENCES
no
MBED-LIKE CLUSTERING GUIDE-TREE
yes
MBED-LIKE CLUSTERING ITERATION
yes
NUMBER of COMBINED ITERATIONS
default(0)
MAX GUIDE TREE ITERATIONS
default
MAX HMM ITERATIONS
default
ORDER
aligned
STEP 3 - 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.


Alignments
Result Summary
Phylogenetic Tree
Submission Details
Download Alignment File
Hide Colors
Send to ClustalW2_Phylogeny
Residue
Colour
Property
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


Tools > Multiple Sequence Alignment > Clustal Omega
Results for job clustalo-I20140301-151344-0178-43085370-pg
Alignments
Result Summary
Phylogenetic Tree
Submission Details
Phylogenetic Tree
This is a Neighbour-joining tree without distance corrections.
Download Phylogenetic Tree File
Phylogram
Branch length: Cladogram Real
FOSB_MOUSE 0.01854
FOSB_HUMAN 0.02288
FOS_CHICK 0.1107
FOS_RAT 0.01948
FOS_MOUSE 0.0121
EMBL-EBI
News
Brochures
Contact us
Intranet
Services
By topic
By name (A-Z)
Help & Support
Research
Overview
Publications
Research groups
Postdocs & PhDs
Training
Overview
Train at EBI
Train outside EBI
Train online
Contact organisers
Industry
Overview
Members Area
Workshops
SME Forum
Contact Industry programme
About us
Overview
Leadership
Funding
Background
Collaboration
Jobs
People & groups
News


Tools > Multiple Sequence Alignment > Clustal Omega
Results for job clustalo-I20140301-151344-0178-43085370-pg
Alignments
Result Summary
Phylogenetic Tree
Submission Details
Phylogenetic Tree
This is a Neighbour-joining tree without distance corrections.
Download Phylogenetic Tree File
(
(
(
FOSB_MOUSE:0.01854,
FOSB_HUMAN:0.02288)
:0.35561,
FOS_CHICK:0.11070)
:0.11115,
FOS_RAT:0.01948,
FOS_MOUSE:0.01210);
Phylogram
Branch length: Cladogram Real
FOSB_MOUSE 0.01854
FOSB_HUMAN 0.02288
FOS_CHICK 0.1107
FOS_RAT 0.01948
FOS_MOUSE 0.0121
EMBL-EBI
News
Brochures
Contact us
Intranet
Services
By topic
By name (A-Z)
Help & Support
Research
Overview
Publications
Research groups
Postdocs & PhDs
Training
Overview
Train at EBI
Train outside EBI
Train online
Contact organisers
Industry
Overview
Members Area
Workshops
SME Forum
Contact Industry programme
About us
Overview
Leadership
Funding
Background
Collaboration
Jobs
People & groups
News

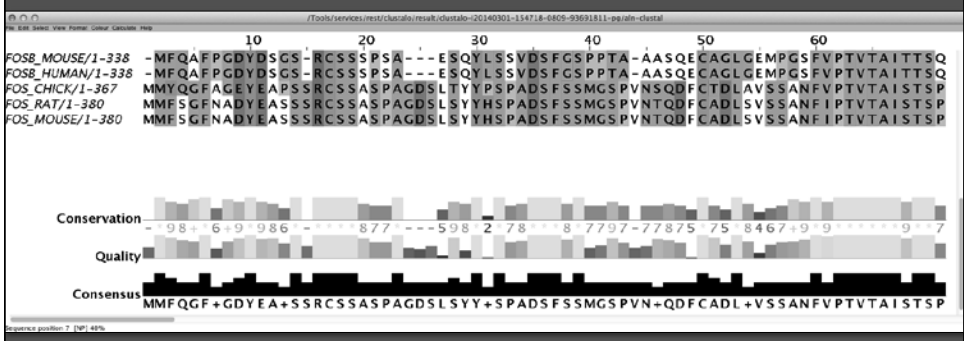

EMBL-EBI
Services
Research
Training
About us
Clustal Omega
Input form
Web services
Help & Documentation
Share
Feedback
Tools > Multiple Sequence Alignment > Clustal Omega
Results for job clustalo-I20140301-154013-0278-93886695-oy
Alignments
Result Summary
Phylogenetic Tree
Submission Details
Input Sequences
clustalo-I20140301-154013-0278-93886695-oy.input
Tool Output
clustalo-I20140301-154013-0278-93886695-oy.output
Alignment in CLUSTAL format with base/residue numbering
clustalo-I20140301-154013-0278-93886695-oy.clustal_num
Phylogenetic Tree
clustalo-I20140301-154013-0278-93886695-oy.ph
Percent Identity Matrix
clustalo-I20140301-154013-0278-93886695-oy.pim
Jalview
Start Jalview
EMBL-EBI
News
Brochures
Contact us
Intranet
Services
By topic
By name (A-Z)
Help & Support
Research
Overview
Publications
Research groups
Postdocs & PhDs
Training
Overview
Train at EBI
Train outside EBI
Train online
Contact organisers
Industry
Overview
Members Area
Workshops
SME Forum
Contact Industry programme
About us
Overview
Leadership
Funding
Background
Collaboration
Jobs
People & groups
News

Jalview

- Java applet available within Clustal Omega results
- Used to manually edit Clustal Omega 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

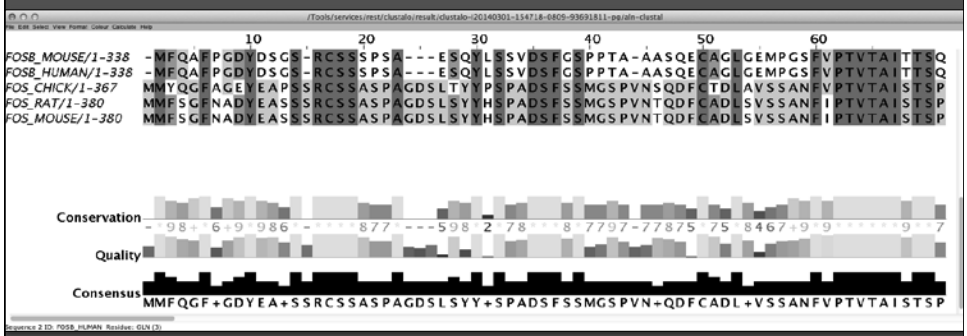
Default view



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

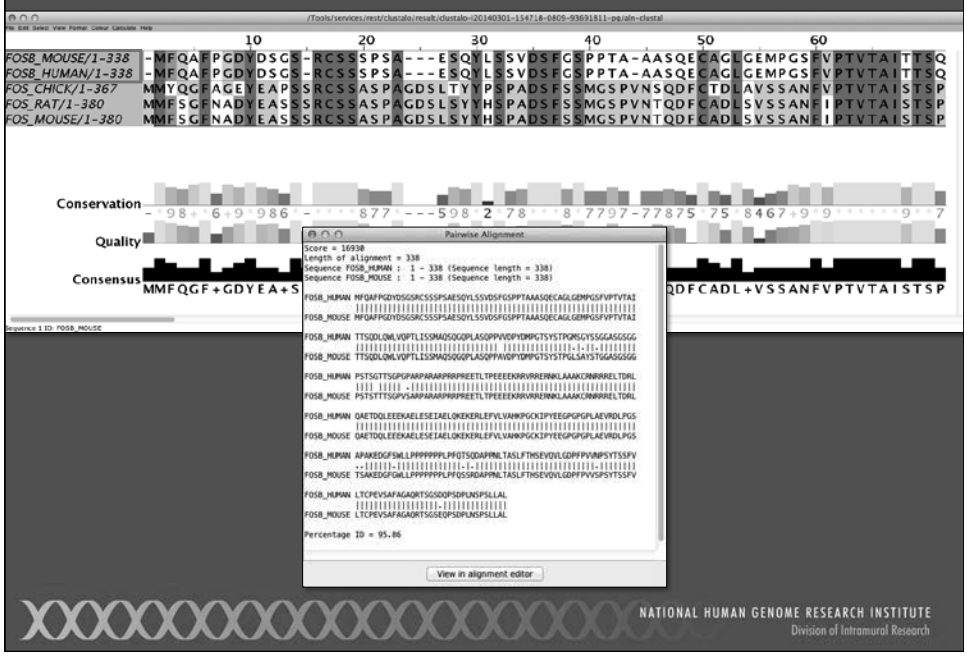
Colour → Percentage Identity



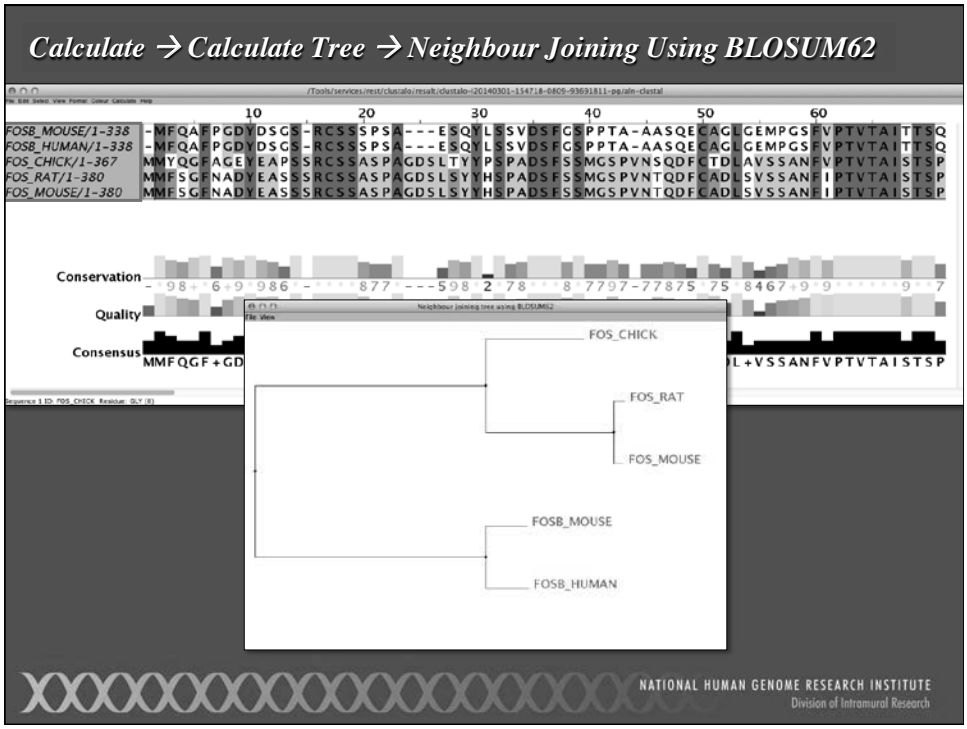
Agreement	*Background Color*
81 - 100%	Dark blue
61 - 80%	Medium blue
41 - 60%	Light blue
≤ 40%	White

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Calculate → Pairwise Alignments



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research



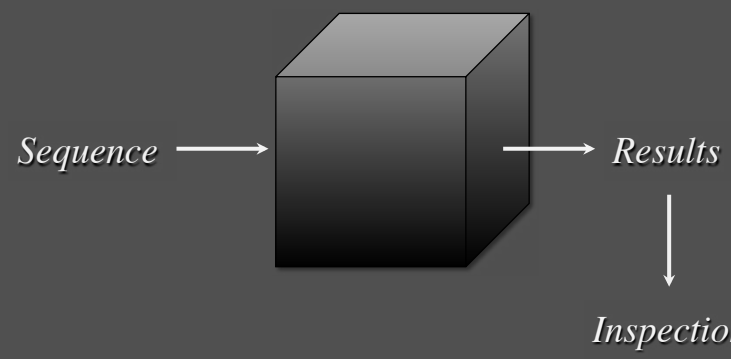
T-COFFEE

- Combines sequence, profile, and structural information
  - Protein structures
  - RNA secondary structures
- Specialized algorithm for aligning transmembrane proteins, non-coding RNAs, and homologous promoter regions
- Can combine output from other methods into a single "master alignment"
- Freely available at *http://tcoffee.org*

Magis et al., Methods Mol. Biol. 1079: 117-129 (2014)

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Understanding Analyses



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Online Training Resources

Suggested curriculum tracks tailored to individual needs

- Bioinformatic Analysis
- Data Mining
- Bioinformatics Tools
- Bioinformatics Systems
- Computational Biology

Perspective

An Online Bioinformatics Curriculum

David B. Searls*

Searls, PLoS Comput. Biol. 8: e1002632, 2012

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Current Topics in Genome Analysis 2014

Next Lecture
March 19, 2014

Genome-Scale Sequence Analysis

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

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research