NATIONAL HUMAN GENOME RESEARCH INSTITUTE Division of Intramural Research

Current Topics in Genome Analysis 2016

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 2016

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*

Profiles, Patterns, Motifs, and Domains

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
```

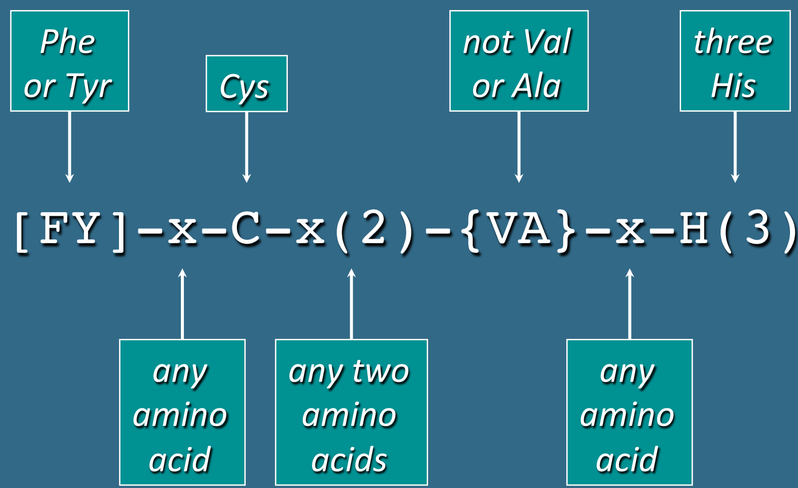
- ***Which residues are seen at each position?***
- ***What is the frequency of observed residues?***
- ***Which positions are conserved?***
- ***Where can gaps be introduced?***

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~~	[illegible]	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



Pfam

- Collection of multiple alignments of protein domains and conserved protein regions that probably have structural, functional, or evolutionary 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. 44: D279-D285, 2016

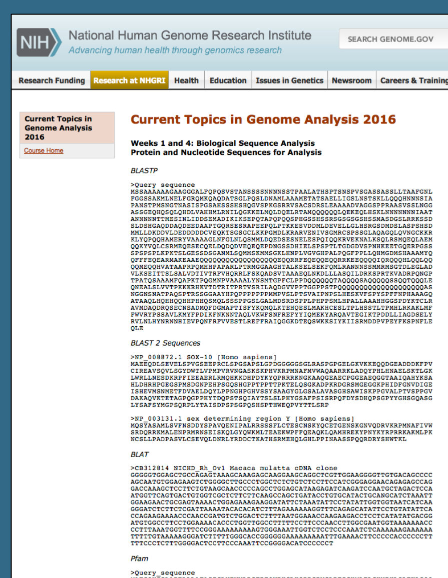
Pfam A

- Based on *curated* multiple alignments of known members of a protein family ('seed alignment')
  - *Pfam definition of 'family': a collection of related protein regions*
  - *Based on reference proteomes (UniProtKB)*
- HMMER used to find all detectable protein sequences belonging to the family
- New 'true members' of the family are then used to generate the 'full alignment' for the protein family
- Given the method used to construct the alignments, hits are highly likely to be true positives

NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

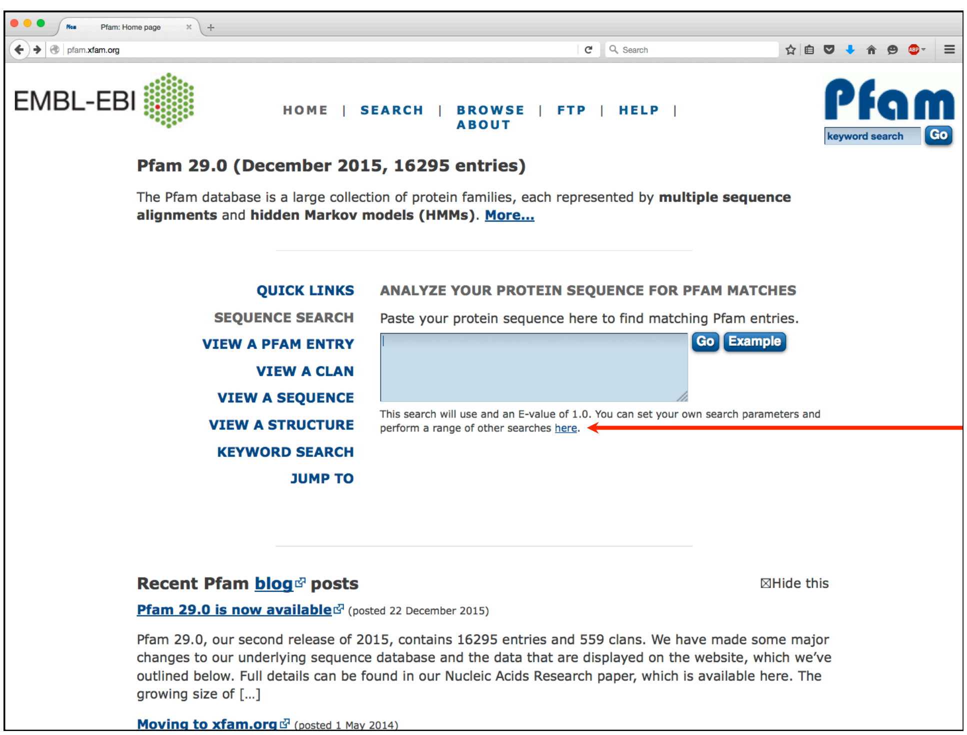
Sequences Used in Examples

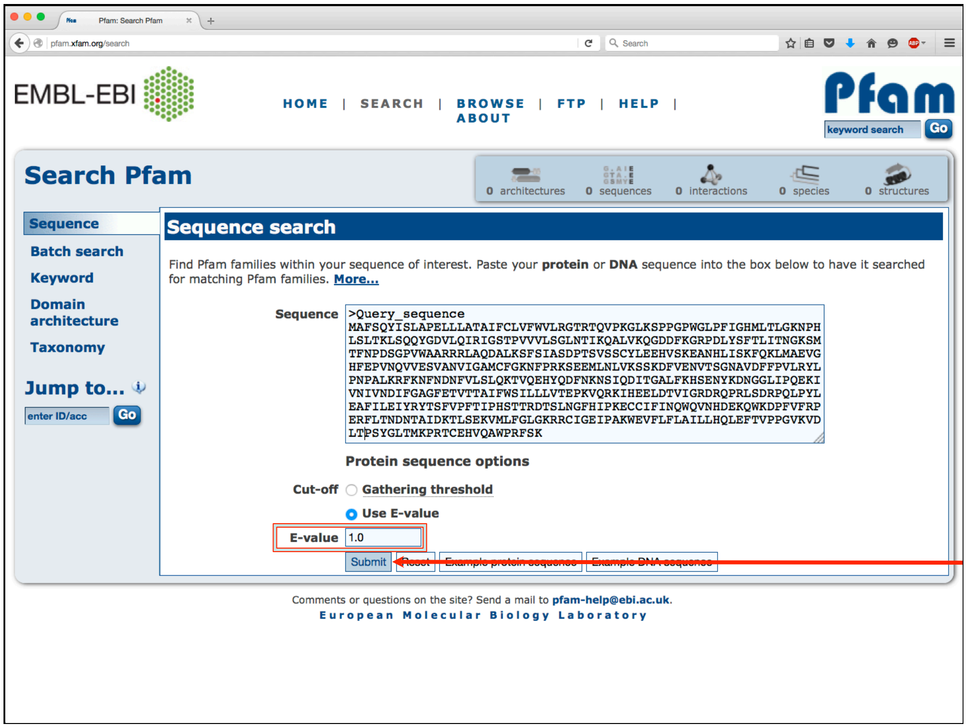
http://research.nhgri.nih.gov/teaching/seq_analysis.shtml

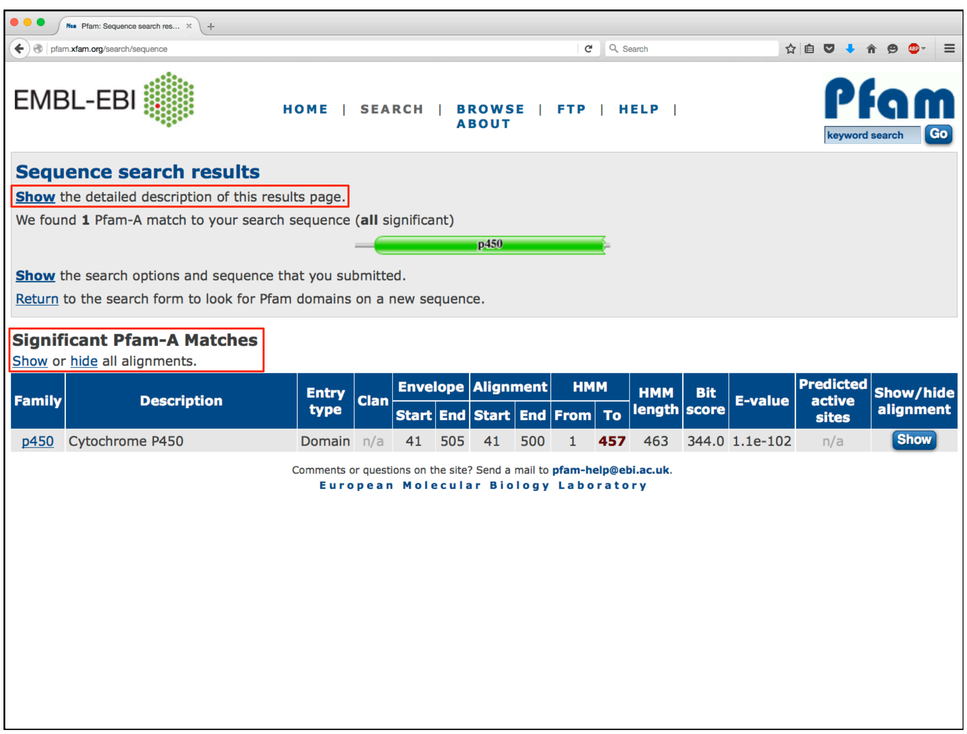


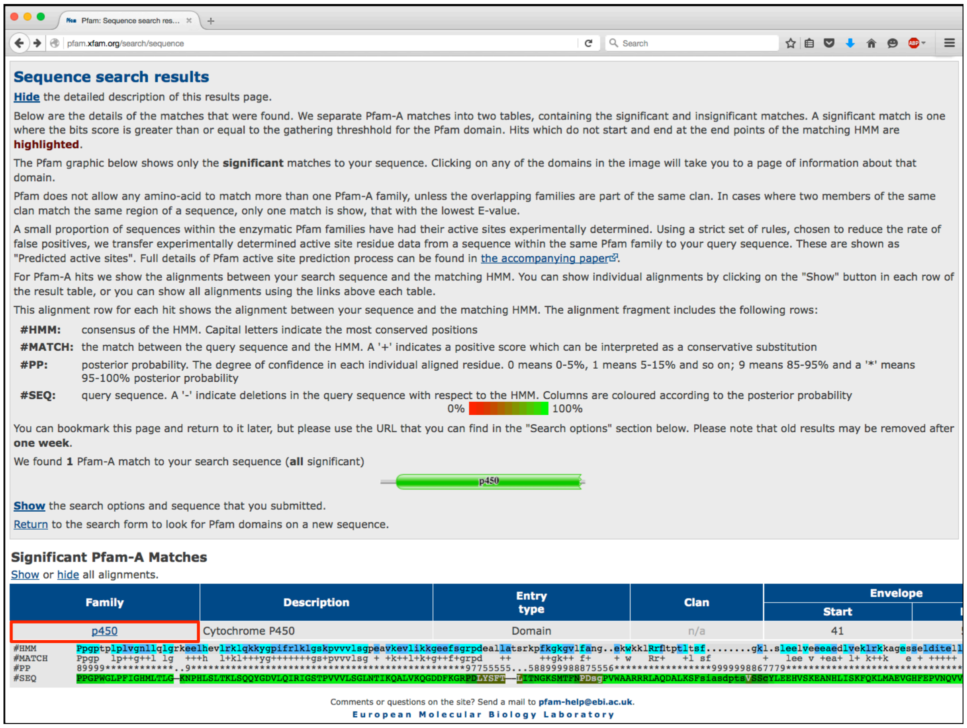
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

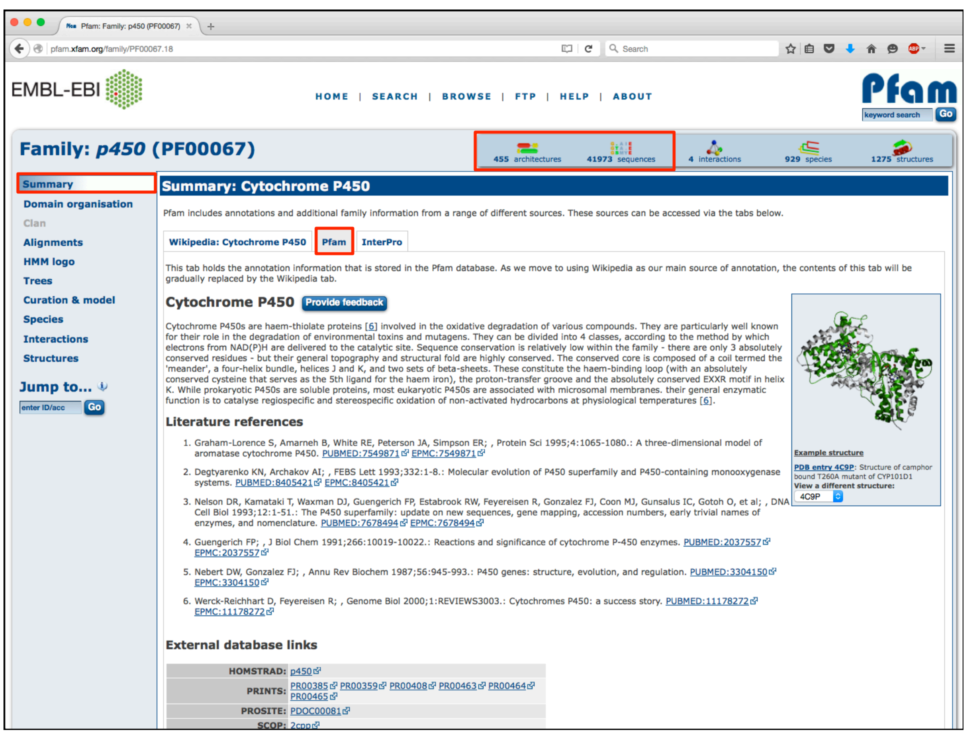

http://pfam.xfam.org
EMBL-EBI
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search Go
Pfam 29.0 (December 2015, 16295 entries)
The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...
QUICK LINKS
YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH
Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY
View Pfam annotation and alignments
VIEW A CLAN
See groups of related entries
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 entry or clan, UniProt sequence, PDB structure, etc.
Or view the help pages for more information
Recent Pfam blog posts
Hide this
Pfam 29.0 is now available (posted 22 December 2015)
Pfam 29.0, our second release of 2015, contains 16295 entries and 559 clans. We have made some major changes to our underlying sequence database and the data that are displayed on the website, which we've outlined below. Full details can be found in our Nucleic Acids Research paper, which is available here. The

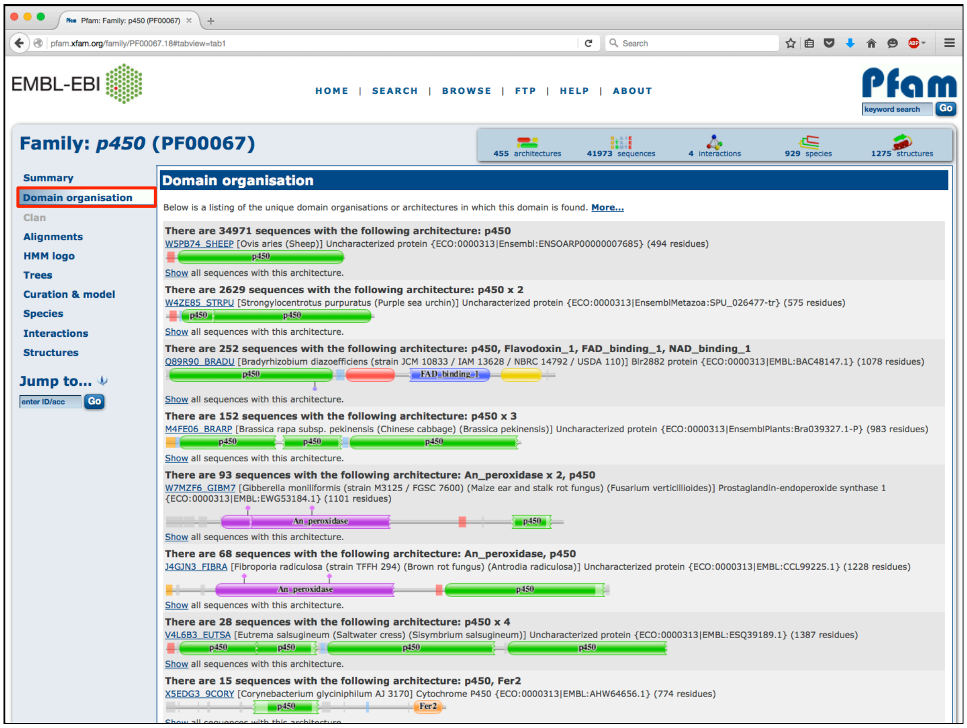

EMBL-EBI
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search Go
Pfam 29.0 (December 2015, 16295 entries)
The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...
QUICK LINKS
ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES
SEQUENCE SEARCH
Paste your protein sequence here to find matching Pfam entries.
VIEW A PFAM ENTRY
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
Pfam 29.0 is now available (posted 22 December 2015)
Pfam 29.0, our second release of 2015, contains 16295 entries and 559 clans. We have made some major changes to our underlying sequence database and the data that are displayed on the website, which we've outlined below. Full details can be found in our Nucleic Acids Research paper, which is available here. The growing size of [...]
Moving to xfam.org (posted 1 May 2014)

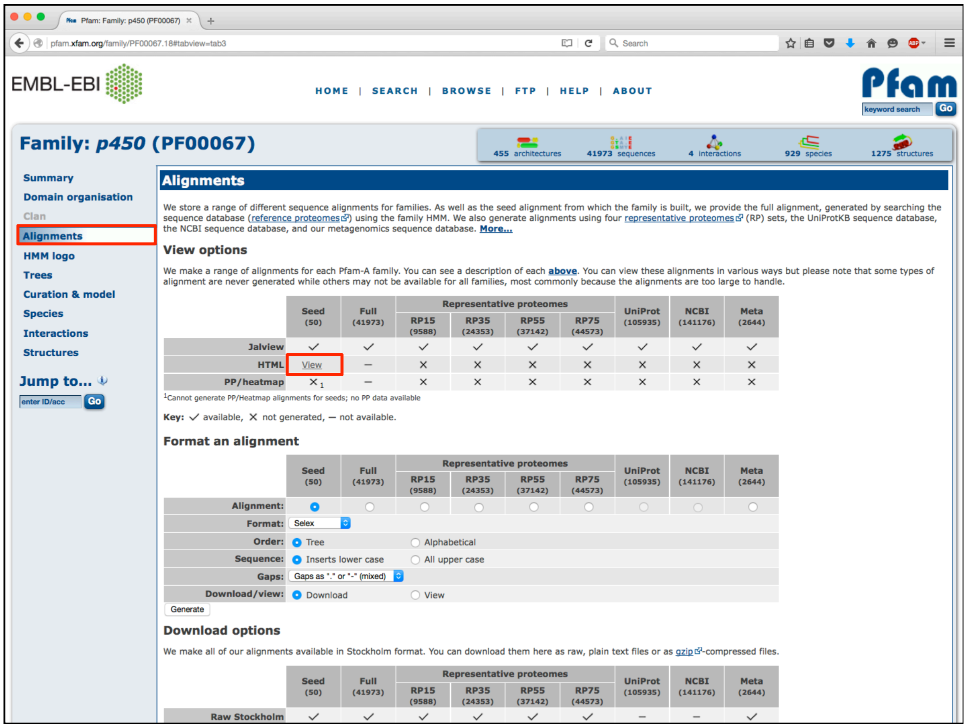

EMBL-EBI
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
Search Pfam
0 architectures 0 sequences 0 interactions 0 species 0 structures
Sequence
Batch search
Keyword
Domain architecture
Taxonomy
Jump to...
Sequence search
Find Pfam families within your sequence of interest. Paste your protein or DNA sequence into the box below to have it searched for matching Pfam families. More...
Sequence
>Query_sequence
MAFSQYISLAPELLLATAIFCLVFWVLRGTRTQVPKGLKSPPGPWGLPFIGHMLTLGKNPH
LSLTKLSQQYGDVLQIRIGSTPVVVLSGLNTIKQALVKQGDDFKGRPDLYSFTLITNGKSM
TFNPDSGPVWAARRRLAQDALKSFSIASDPTSVSSCYLEEHVSKEANHLISKFQKLMAEVG
HFEPVNQVVESVANVIGAMCFGKNFPRKSEEMLNLVKSSKDFVENVTSGNAVDFFPVLRYL
PNPALKRFKNFNDNFVLSLQKTVQEHYQDFNKNSIQDITGALFKHSENYKDNGGLIPQEKI
VNIVNDIFGAGFETVTTAIFWSILLLVTEPKVQRKIHEELDTVIGRDRQPRLSDRPQLPYL
EAFILEIYRYTSFVPFTIPHSTTRDTSLNGFHIPKECCIFINQWQVNHDEKQWKDPFVFRP
ERFLTNDNTAIDKTLSEKVMLFGLGKRRCIGEIPAKWEVFLFLAILLHQLEFTVPPGVKVD
LTPSYGLTMKPRTCEHVQAWPRFSK
Protein sequence options
Cut-off
Gathering threshold
Use E-value
E-value 1.0
Submit
Reset
Example protein sequence
Example DNA sequence
Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

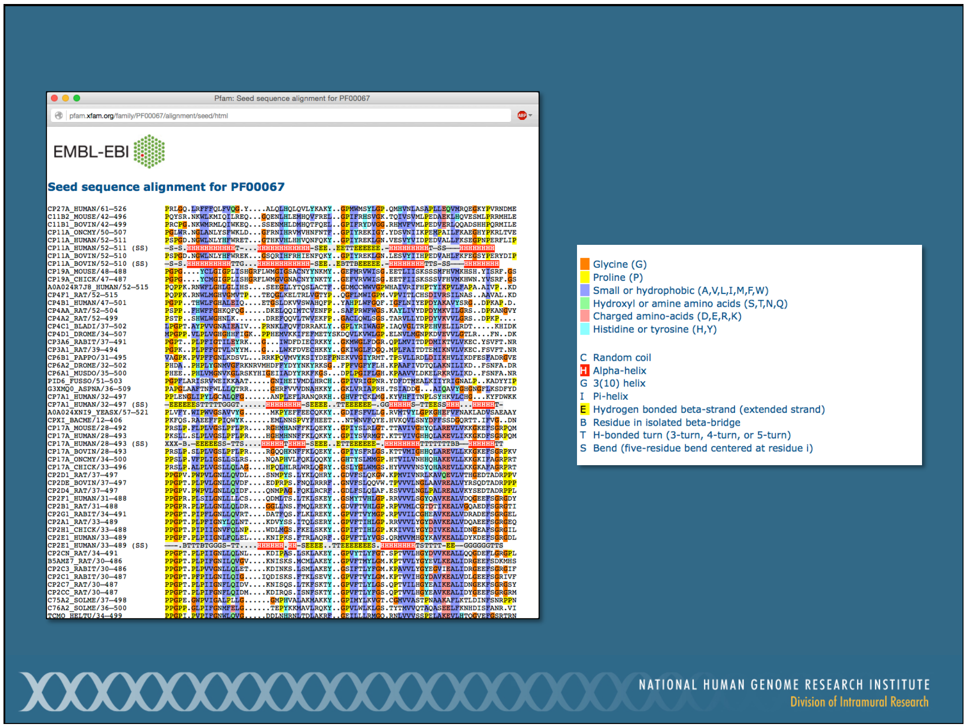

EMBL-EBI
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
Sequence search results
Show the detailed description of this results page.
We found 1 Pfam-A match to your search sequence (all significant)
p450
Show the search options and sequence that you submitted.
Return to the search form to look for Pfam domains on a new sequence.
Significant Pfam-A Matches
Show or hide all alignments.
Family | Description | Entry type | Clan | Envelope Start | Envelope End | 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
Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

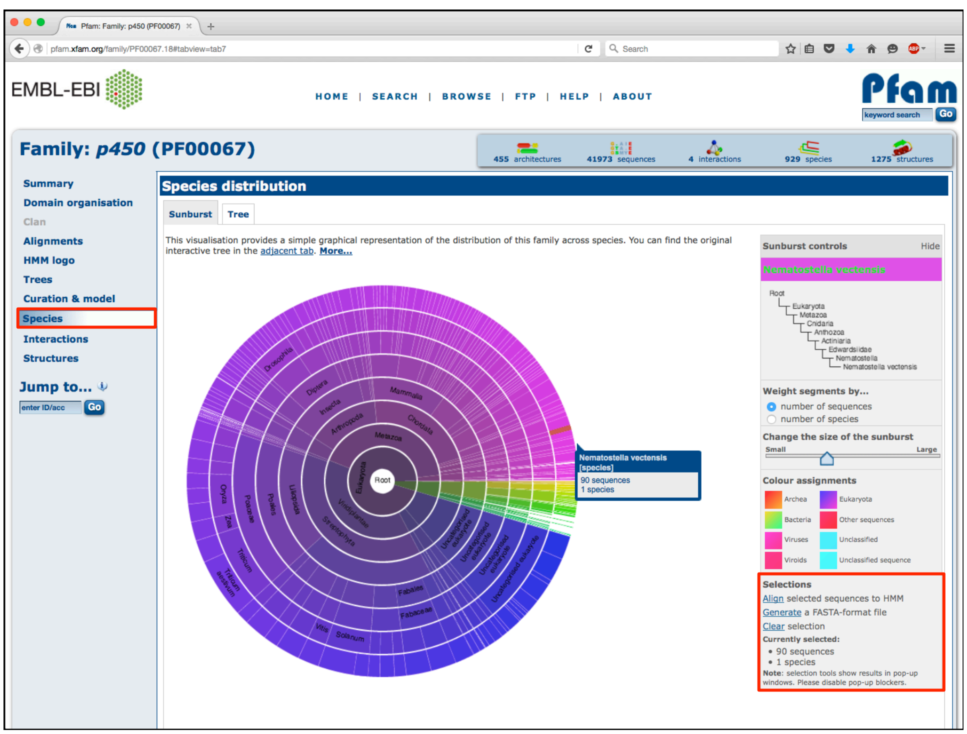

Pfam: Sequence search res...
pfam.xfam.org/search/sequence
Sequence search results
Hide the detailed description of this results page.
Below are the details of the matches that were found. We separate Pfam-A matches into two tables, containing the significant and insignificant matches. A significant match is one where the bits score is greater than or equal to the gathering threshold 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 1 Pfam-A match to your search sequence (all significant)
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
p450 | Cytochrome P450 | Domain | n/a | 41
#HMM
#MATCH
#PP
#SEQ
Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

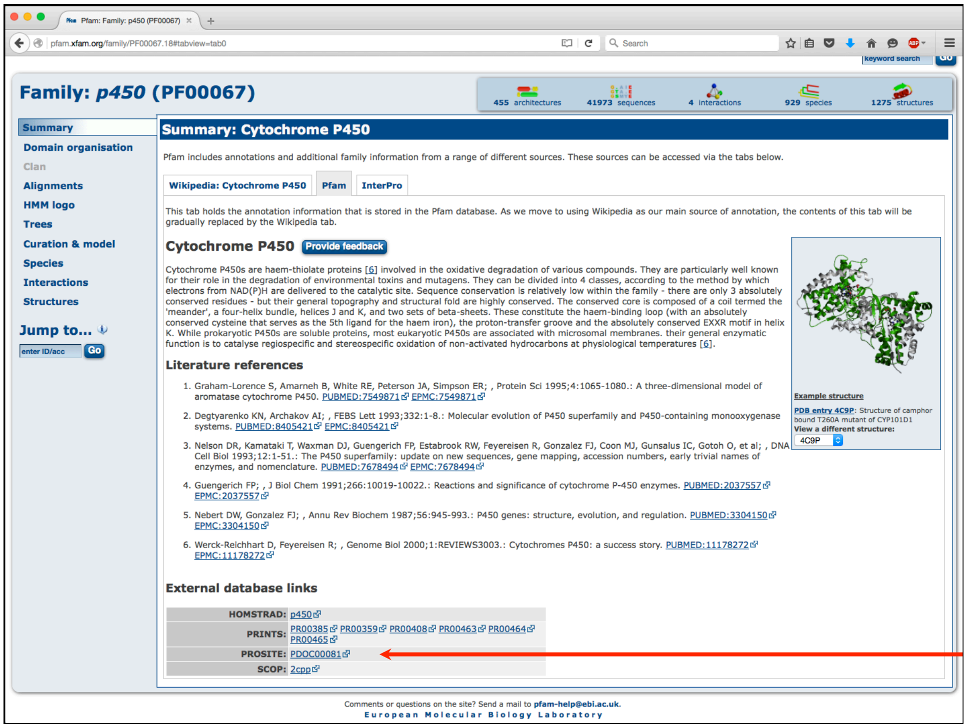

Pfam: Family: p450 (PF00067)
pfam.xfam.org/family/PF00067.18
EMBL-EBI
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search Go
Family: p450 (PF00067)
455 architectures
41973 sequences
4 interactions
929 species
1275 structures
Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & 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 4C9P: Structure of camphor bound T260A mutant of CYP101D1
View a different structure:
4C9P
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
PRINTS: PR00385 PR00359 PR00408 PR00463 PR00464 PR00465
PROSITE: PDOC00081
SCOP: 2cpp

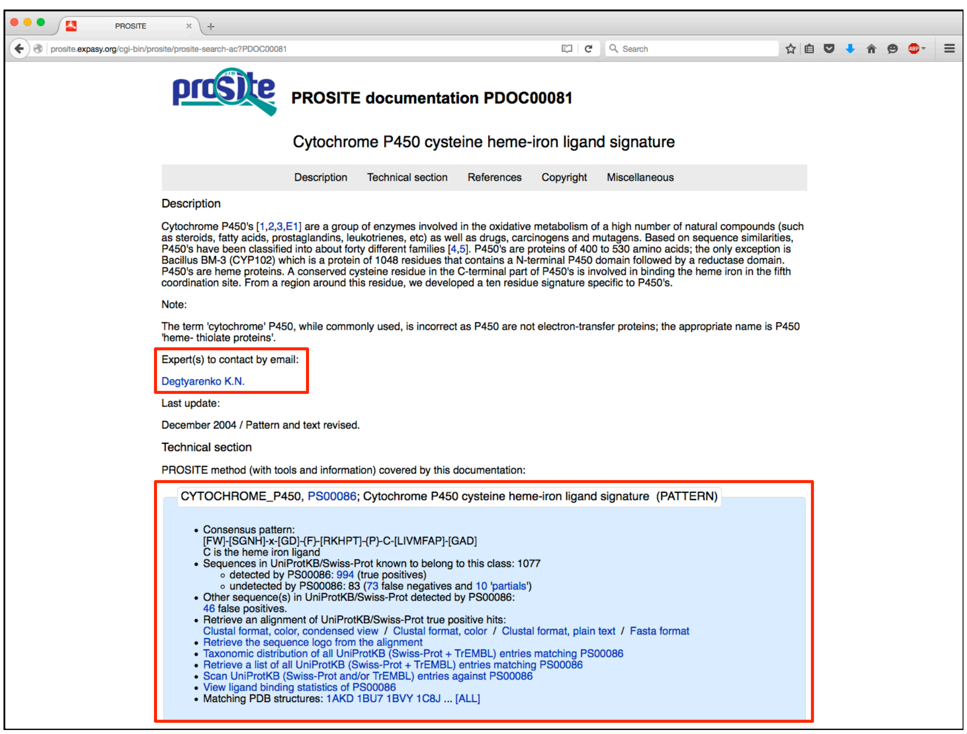





EMBL-EBI
Seed sequence alignment for PF00067
Glycine (G)
Proline (P)
Small or hydrophobic (A,V,L,I,M,F,W)
Hydroxyl or amine amino acids (S,T,N,Q)
Charged amino-acids (D,E,R,K)
Histidine or tyrosine (H,Y)
C Random coil
H Alpha-helix
G 3(10) helix
I Pi-helix
E Hydrogen bonded beta-strand (extended strand)
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


EMBL-EBI
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
Family: p450 (PF00067)
455 architectures
41973 sequences
4 interactions
929 species
1275 structures
Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures
Jump to...
Species distribution
Sunburst
Tree
This visualisation provides a simple graphical representation of the distribution of this family across species. You can find the original interactive tree in the adjacent tab. More...
Sunburst controls
Nematostella vectensis [species]
90 sequences
1 species
Weight segments by...
number of sequences
number of species
Change the size of the sunburst
Colour assignments
Selections
Align selected sequences to HMM
Generate a FASTA-format file
Clear selection
Currently selected:
• 90 sequences
• 1 species
Note: selection tools show results in pop-up windows. Please disable pop-up blockers.


Family: p450 (PF00067)
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
Cytochrome P450
Provide feedback
Literature references
External database links
HOMSTRAD: p450
PRINTS: PR00385 PR00359 PR00408 PR00463 PR00464 PR00465
PROSITE: PDOC00081
SCOP: 2cpp


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: 1077
detected by PS00086: 994 (true positives)
undetected by PS00086: 83 (73 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, color / Clustal format, plain text / Fasta format
Retrieve the sequence logo from the alignment
Taxonomic distribution 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
  - Simple Modular Architecture Research Tool (SMART)
  - COG (orthologous prokaryotic protein families)
  - PRK ('protein clusters' of related protein RefSeq entries)
  - TIGRFAM

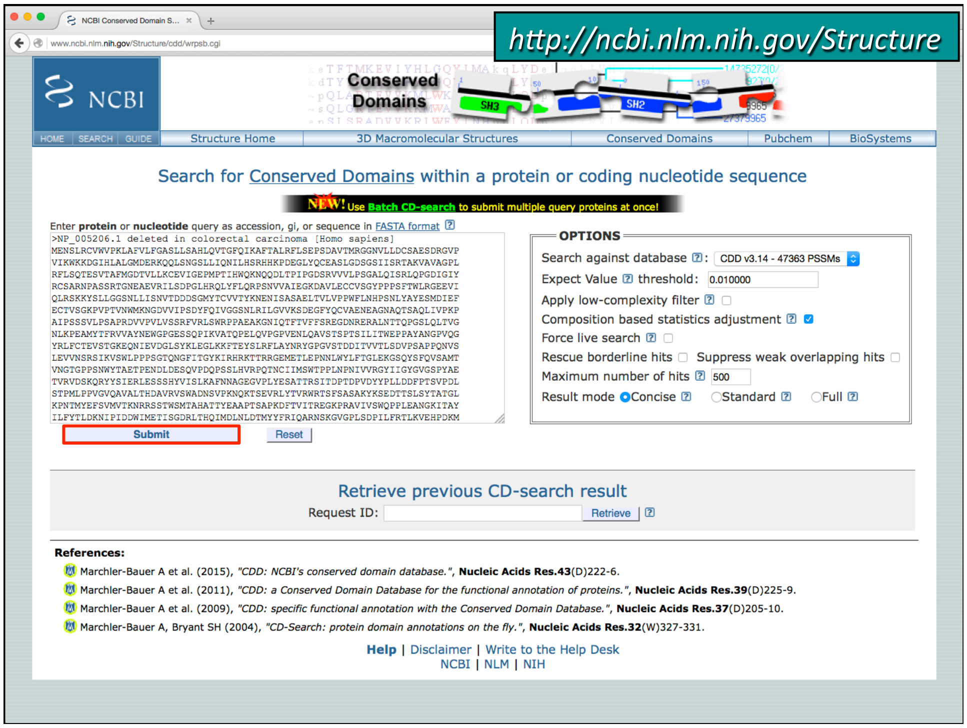
Marchler-Bauer et al., Nucleic Acids Res. 43: D222-D226, 2015

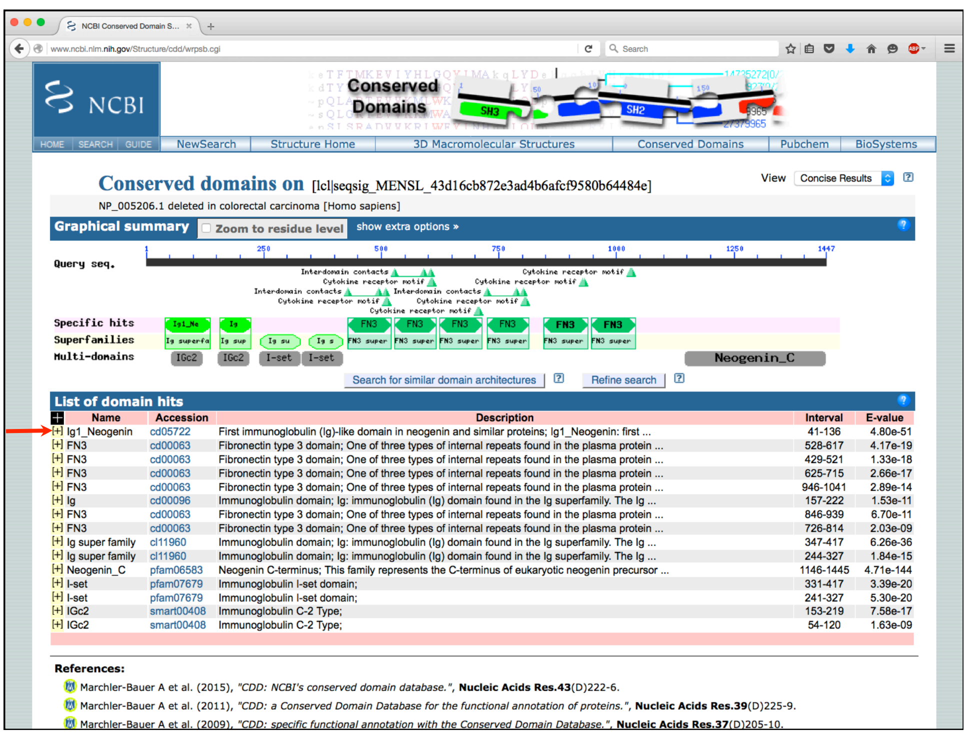
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

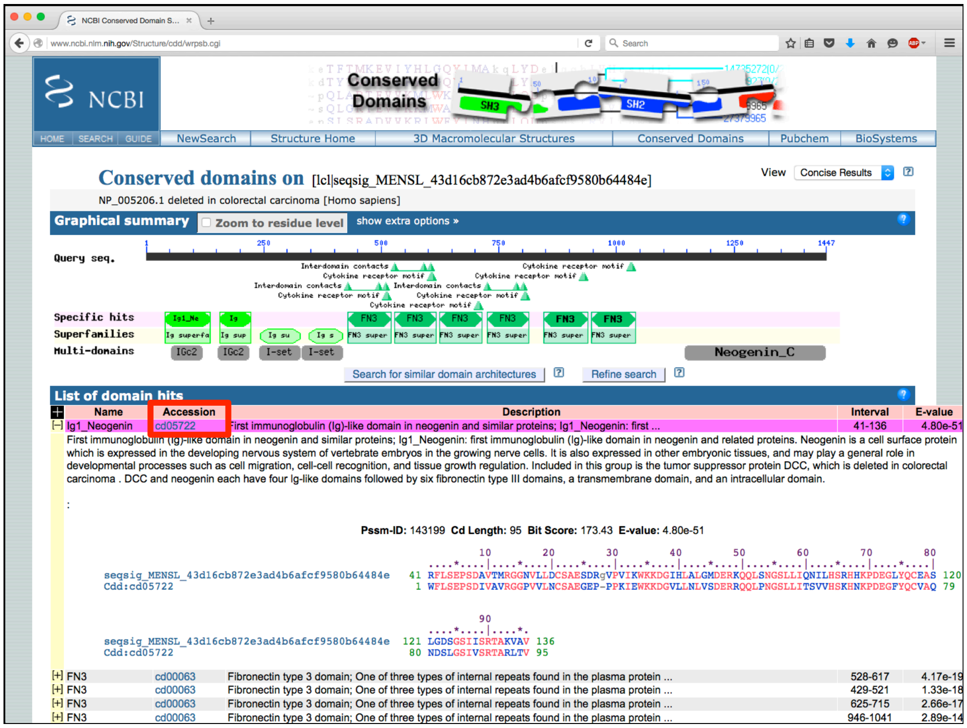
Conserved Domain Database (CDD)

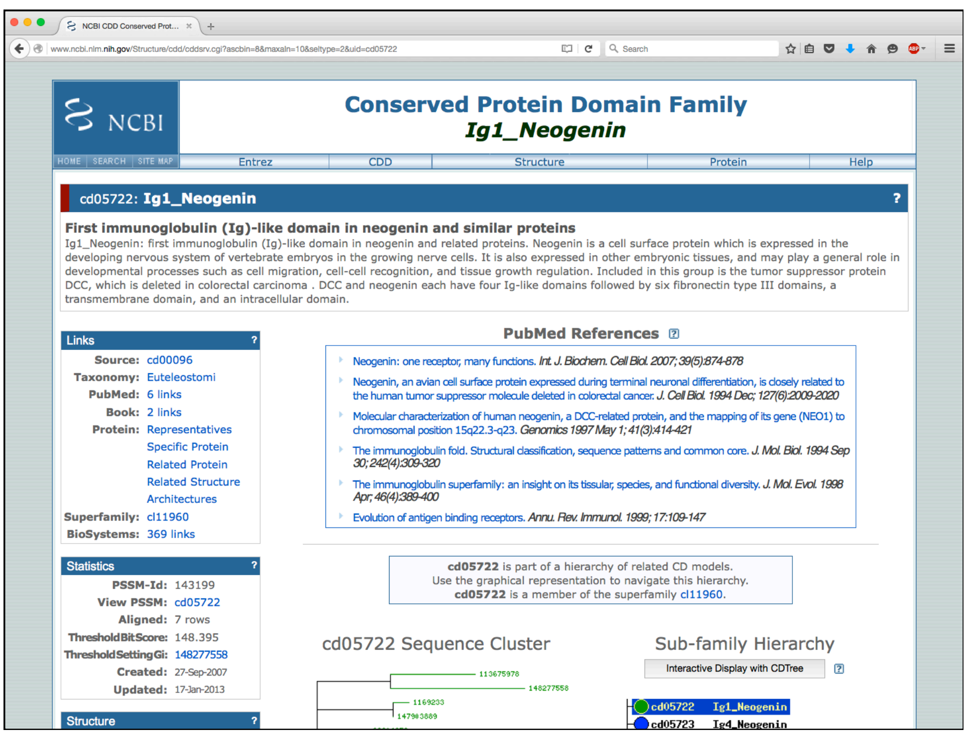
- CD-Search performed using RPS-BLAST
- Query sequence is used to search a database of pre-calculated position-specific scoring matrices
- *Not* the same method used by Pfam

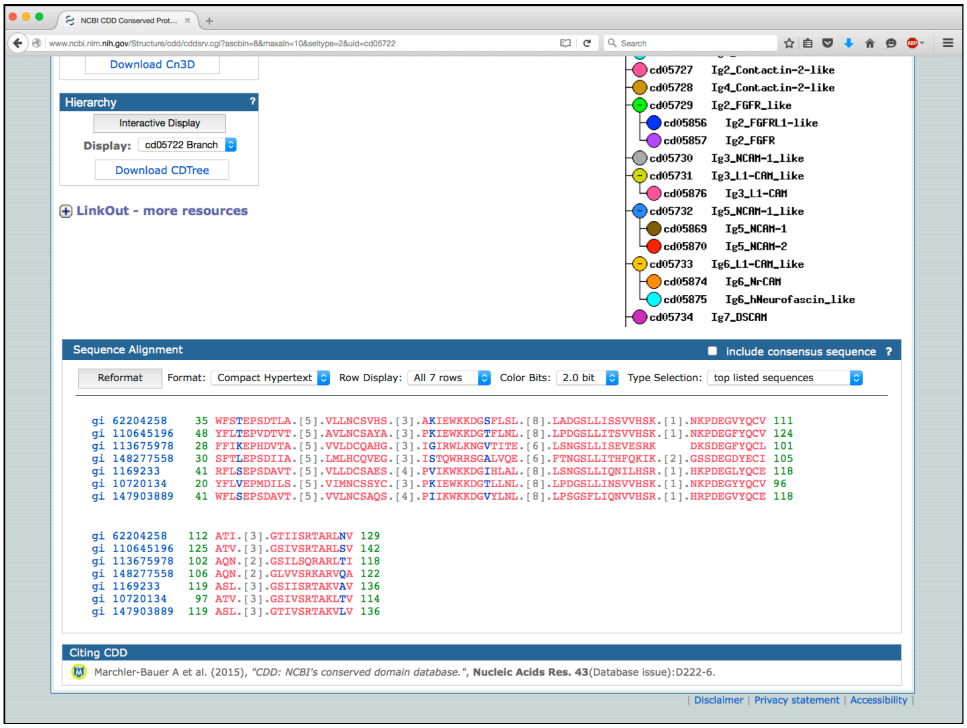
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

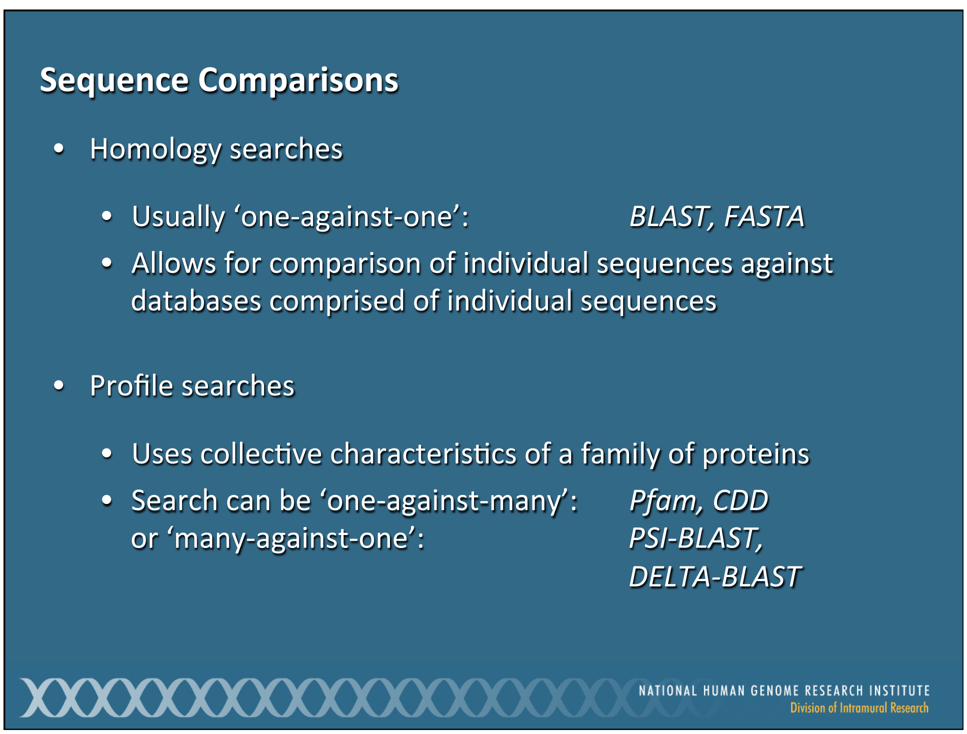

http://ncbi.nlm.nih.gov/Structure
NCBI
Conserved Domains
HOME | SEARCH | GUIDE
Structure Home
3D Macromolecular Structures
Conserved Domains
Pubchem
BioSystems
Search for Conserved Domains within a protein or coding nucleotide sequence
NEW! Use Batch CD-search to submit multiple query proteins at once!
Enter protein or nucleotide query as accession, gi, or sequence in FASTA format
>NP_005206.1 deleted in colorectal carcinoma [Homo sapiens]
OPTIONS
Search against database: CDD v3.14 - 47363 PSSMs
Expect Value threshold: 0.010000
Apply low-complexity filter
Composition based statistics adjustment
Force live search
Rescue borderline hits
Suppress weak overlapping hits
Maximum number of hits 500
Result mode Concise Standard Full
Submit
Reset
Retrieve previous CD-search result
Request ID:
Retrieve
References:
Marchler-Bauer A et al. (2015), "CDD: NCBI's conserved domain database.", Nucleic Acids Res.43(D)222-6.
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.
Help | Disclaimer | Write to the Help Desk
NCBI | NLM | NIH


NCBI
Conserved Domains
HOME | SEARCH | GUIDE
NewSearch
Structure Home
3D Macromolecular Structures
Conserved Domains
Pubchem
BioSystems
Conserved domains on [lcl|seqsig_MENSL_43d16cb872e3ad4b6afcf9580b64484e]
NP_005206.1 deleted in colorectal carcinoma [Homo sapiens]
View Concise Results
Graphical summary
Zoom to residue level
show extra options »
Query seq.
Specific hits
Superfamilies
Multi-domains
Neogenin_C
Search for similar domain architectures
Refine search
List of domain hits
Name | Accession | Description | Interval | E-value
Ig1_Neogenin | cd05722 | First immunoglobulin (Ig)-like domain in neogenin and similar proteins; Ig1_Neogenin: first ... | 41-136 | 4.80e-51
FN3 | cd00063 | Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein ... | 528-617 | 4.17e-19
FN3 | cd00063 | Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein ... | 429-521 | 1.33e-18
FN3 | cd00063 | Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein ... | 625-715 | 2.66e-17
FN3 | cd00063 | Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein ... | 946-1041 | 2.89e-14
Ig | cd00096 | Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig ... | 157-222 | 1.53e-11
FN3 | cd00063 | Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein ... | 846-939 | 6.70e-11
FN3 | cd00063 | Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein ... | 726-814 | 2.03e-09
Ig super family | cl11960 | Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig ... | 347-417 | 6.26e-36
Ig super family | cl11960 | Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig ... | 244-327 | 1.84e-15
Neogenin_C | pfam06583 | Neogenin C-terminus; This family represents the C-terminus of eukaryotic neogenin precursor ... | 1146-1445 | 4.71e-144
I-set | pfam07679 | Immunoglobulin I-set domain; | 331-417 | 3.39e-20
I-set | pfam07679 | Immunoglobulin I-set domain; | 241-327 | 5.30e-20
IGc2 | smart00408 | Immunoglobulin C-2 Type; | 153-219 | 7.58e-17
IGc2 | smart00408 | Immunoglobulin C-2 Type; | 54-120 | 1.63e-09
References:
Marchler-Bauer A et al. (2015), "CDD: NCBI's conserved domain database.", Nucleic Acids Res.43(D)222-6.
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.


NCBI Conserved Domain S...
www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
NCBI
Conserved Domains
HOME | SEARCH | GUIDE
NewSearch
Structure Home
3D Macromolecular Structures
Conserved Domains
Pubchem
BioSystems
Conserved domains on [lcl|seqsig_MENSL_43d16cb872e3ad4b6afcf9580b64484e]
View Concise Results
NP_005206.1 deleted in colorectal carcinoma [Homo sapiens]
Graphical summary
Zoom to residue level
show extra options »
Query seq.
Specific hits
Superfamilies
Multi-domains
Neogenin_C
Search for similar domain architectures
Refine search
List of domain hits
Name
Accession
Description
Interval
E-value
Ig1_Neogenin
cd05722
First immunoglobulin (Ig)-like domain in neogenin and similar proteins; Ig1_Neogenin: first ...
41-136
4.80e-51
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.
Pssm-ID: 143199 Cd Length: 95 Bit Score: 173.43 E-value: 4.80e-51
seqsig_MENSL_43d16cb872e3ad4b6afcf9580b64484e 41 RFLSEPSDAVTMRGGNVLLDCSAESDRgVPVIKWKKDGIHLALGMDERKQQLSNGSLLIQNILHSRHHKPDEGLYQCEAS 120
Cdd:cd05722 1 WFLSEPSDIVAVRGGPVVLNCSAEGEP-PPKIEWKKDGVLLNLVSDERRQQLPNGSLLITSVVHSKHNKPDEGFYQCVAQ 79
seqsig_MENSL_43d16cb872e3ad4b6afcf9580b64484e 121 LGDSGSIISRTAKVAV 136
Cdd:cd05722 80 NDSLGSIVSRTARLTV 95
FN3 cd00063 Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein ... 528-617 4.17e-19
FN3 cd00063 Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein ... 429-521 1.33e-18
FN3 cd00063 Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein ... 625-715 2.66e-17
FN3 cd00063 Fibronectin type 3 domain; One of three types of internal repeats found in the plasma protein ... 946-1041 2.89e-14


NCBI CDD Conserved Prot...
www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?ascbin=8&maxaln=10&seltype=2&uid=cd05722
NCBI
Conserved Protein Domain Family
Ig1_Neogenin
Entrez
CDD
Structure
Protein
Help
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: 369 links
PubMed References
Neogenin: one receptor, many functions. Int J Biochem Cell Biol. 2007; 39(5):874-878
Neogenin, an avian cell surface protein expressed during terminal neuronal differentiation, is closely related to the human tumor suppressor molecule deleted in colorectal cancer. J Cell Biol. 1994 Dec; 127(6):2009-2020
Molecular characterization of human neogenin, a DCC-related protein, and the mapping of its gene (NEO1) to chromosomal position 15q22.3-q23. Genomics 1997 May 1; 41(3):414-421
The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol. 1994 Sep 30; 242(4):309-320
The immunoglobulin superfamily: an insight on its tissular, species, and functional diversity. J Mol Evol. 1998 Apr; 46(4):389-400
Evolution of antigen binding receptors. Annu Rev Immunol. 1999; 17:109-147
Statistics
PSSM-Id: 143199
View PSSM: cd05722
Aligned: 7 rows
Threshold Bit Score: 148.395
Threshold Setting Gi: 148277558
Created: 27-Sep-2007
Updated: 17-Jan-2013
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
Structure


Download Cn3D
Hierarchy
Interactive Display
Display: cd05722 Branch
Download CDTree
LinkOut - more resources
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
cd05875 Ig6_hNeurofascin_like
cd05734 Ig7_DSCAM
Sequence Alignment
include consensus sequence
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. (2015), "CDD: NCBI's conserved domain database.", Nucleic Acids Res. 43(Database issue):D222-6.
Disclaimer | Privacy statement | Accessibility

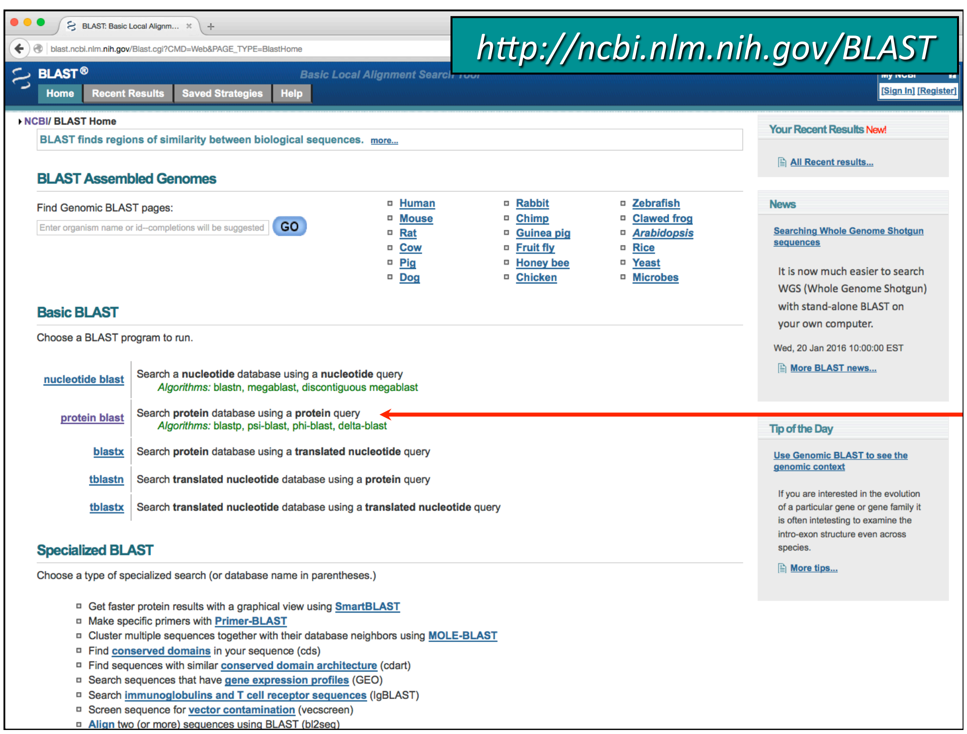

Sequence Comparisons
• Homology searches
• Usually 'one-against-one': BLAST, FASTA
• Allows for comparison of individual sequences against databases comprised of individual sequences
• Profile searches
• Uses collective characteristics of a family of proteins
• Search can be 'one-against-many': Pfam, CDD
or 'many-against-one': PSI-BLAST, DELTA-BLAST
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

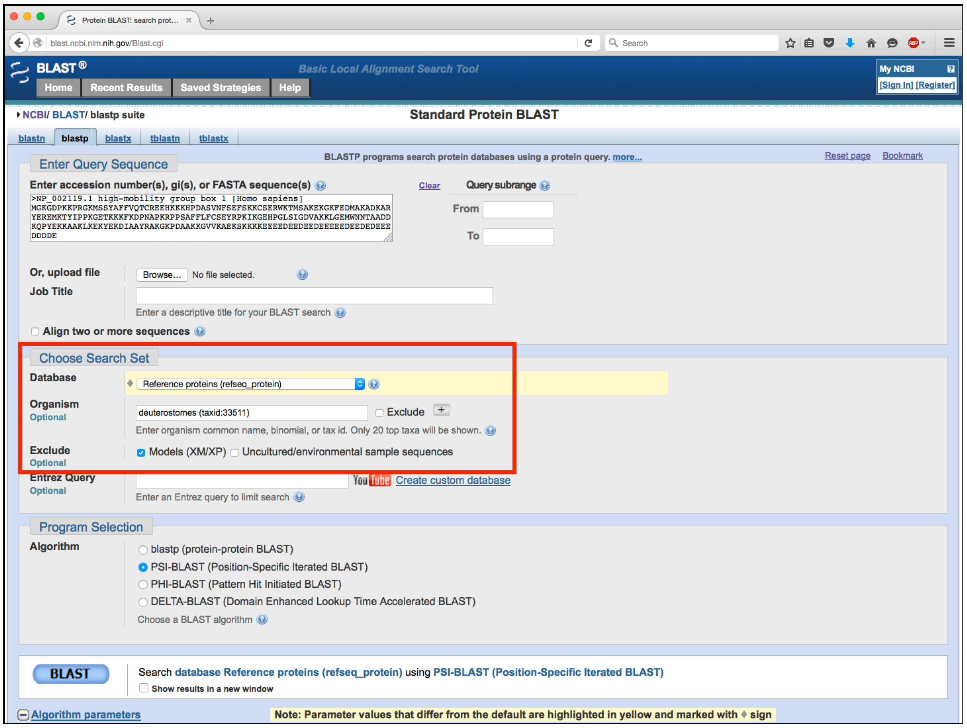
PSI-BLAST

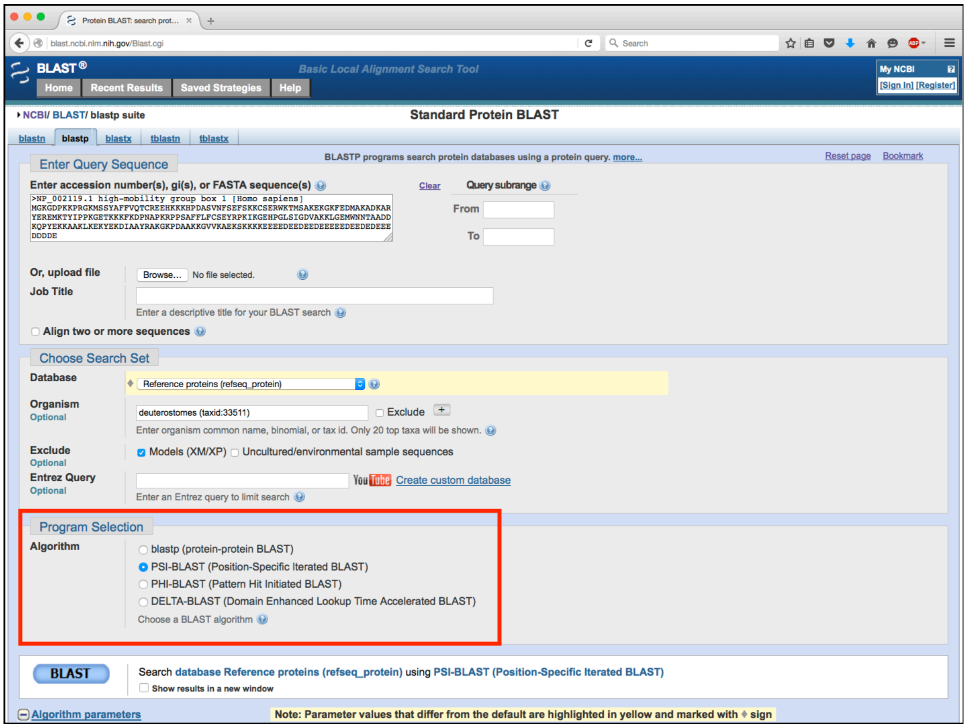
- Position-Specific Iterated BLAST search
- Used to identify distantly related sequences that are possibly missed during a standard BLAST search
- Easy-to-use version of a profile-based search
  - Perform BLAST search against protein database
  - Use results to calculate a position-specific scoring matrix
  - PSSM replaces query for next round of searches
  - May be iterated until no new significant alignments are found

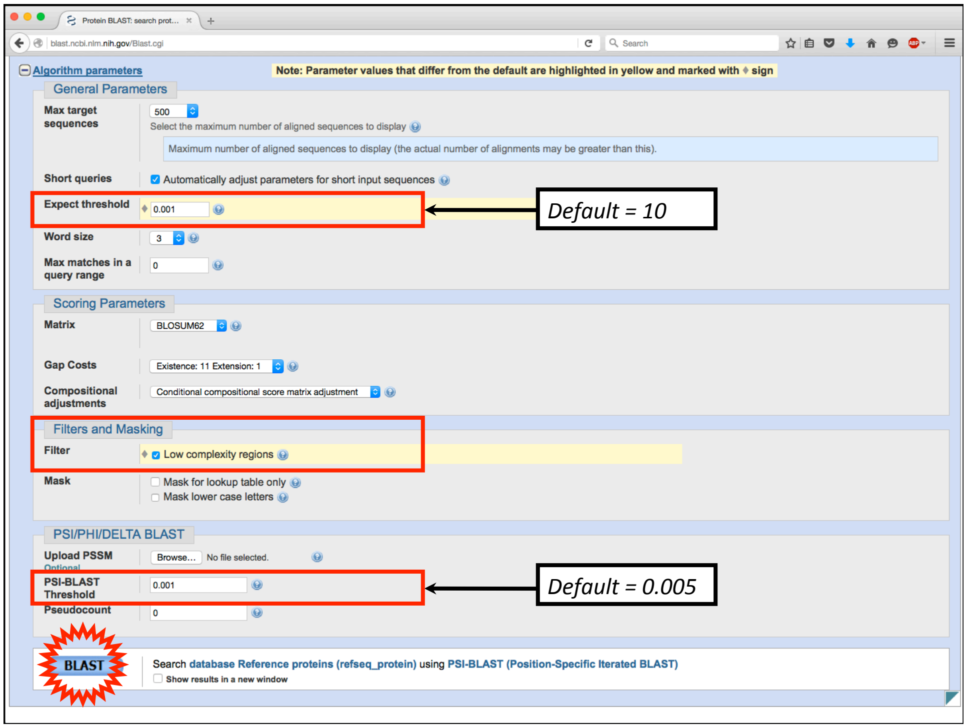
Altschul et al., Nucleic Acids Res. 25: 3389-3402, 1997

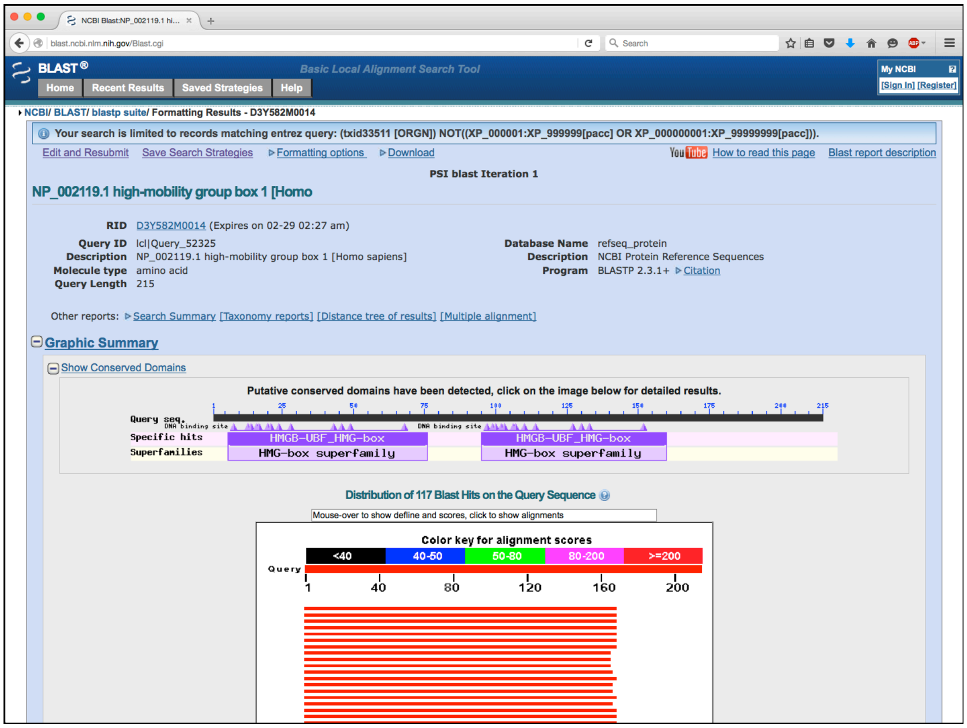
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

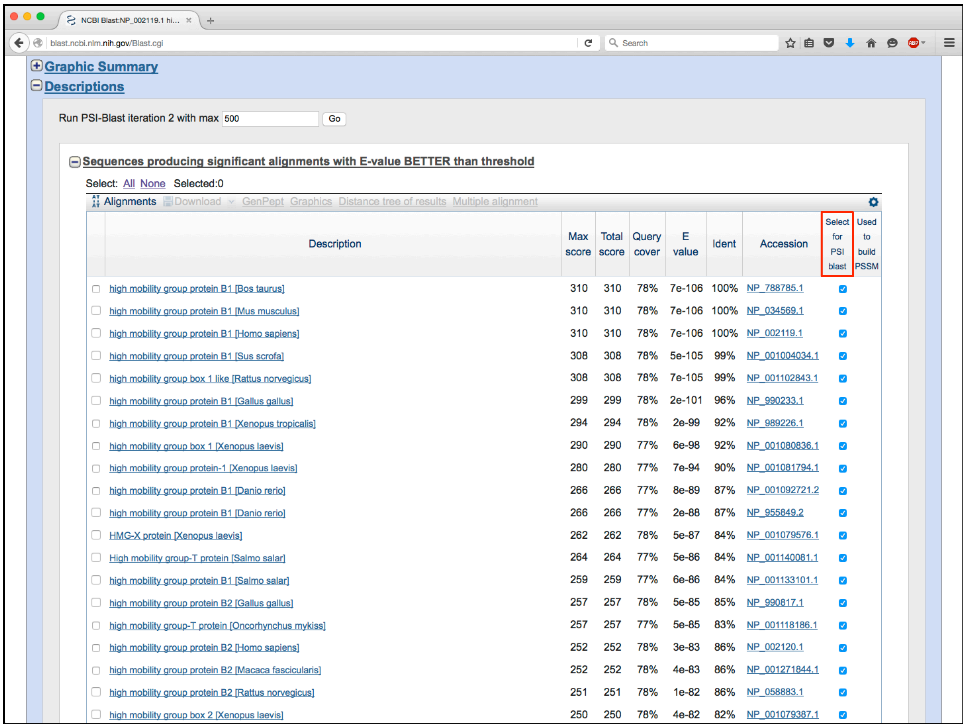




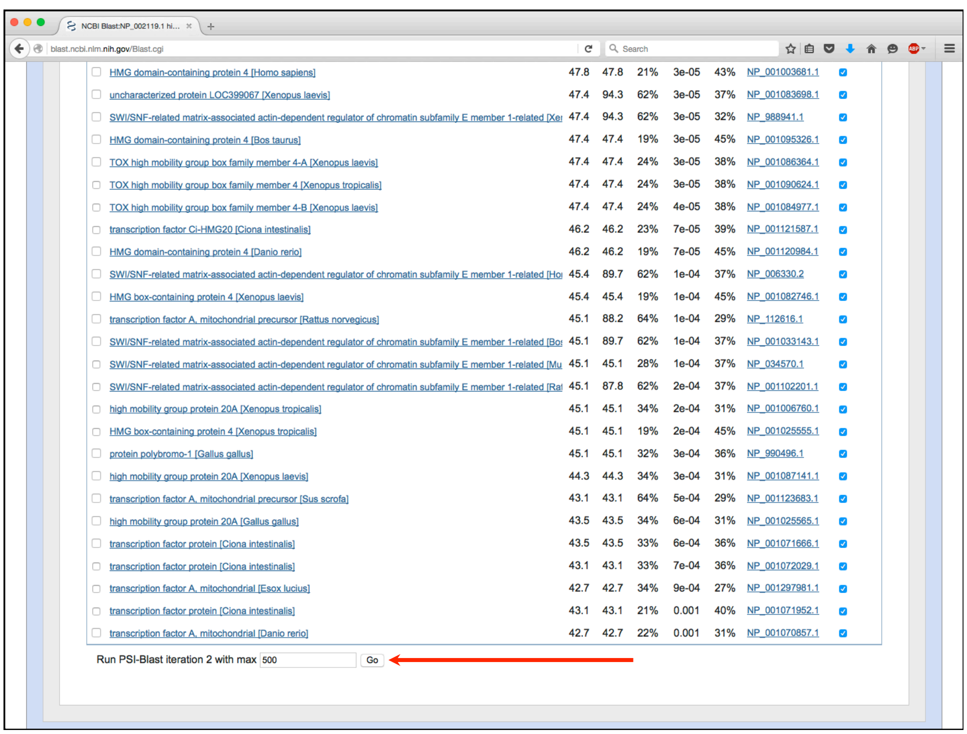



Algorithm parameters
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign
General Parameters
Max target sequences
Short queries
Expect threshold
Default = 10
Word size
Max matches in a query range
Scoring Parameters
Matrix
Gap Costs
Compositional adjustments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

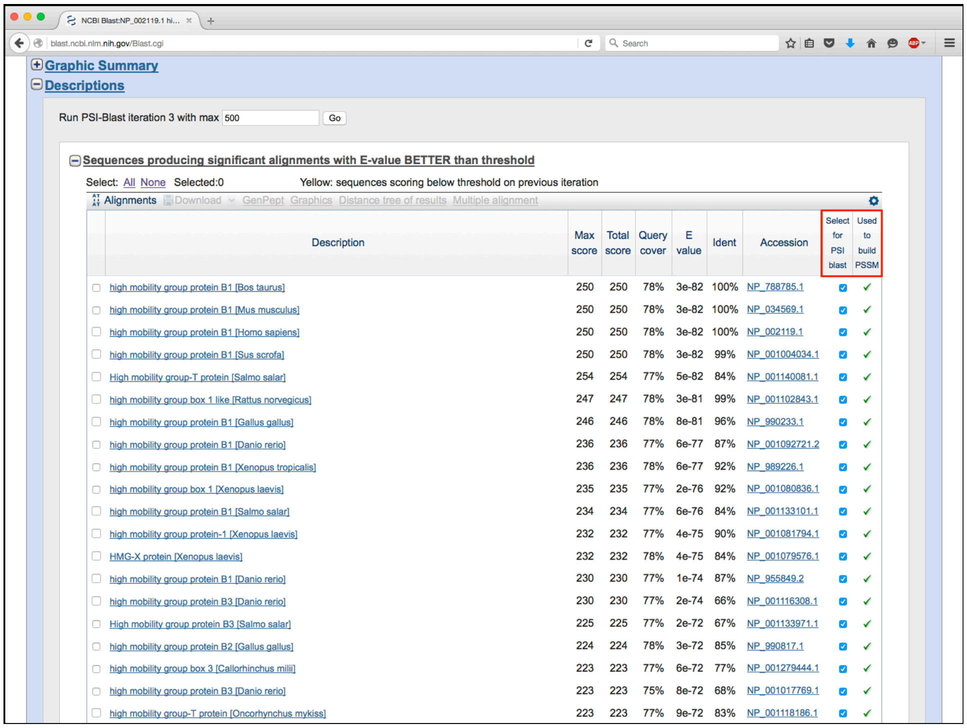

BLAST
Basic Local Alignment Search Tool
PSI blast Iteration 1
NP_002119.1 high-mobility group box 1 [Homo
Graphic Summary
Show Conserved Domains
Putative conserved domains have been detected, click on the image below for detailed results.
Distribution of 117 Blast Hits on the Query Sequence
Color key for alignment scores



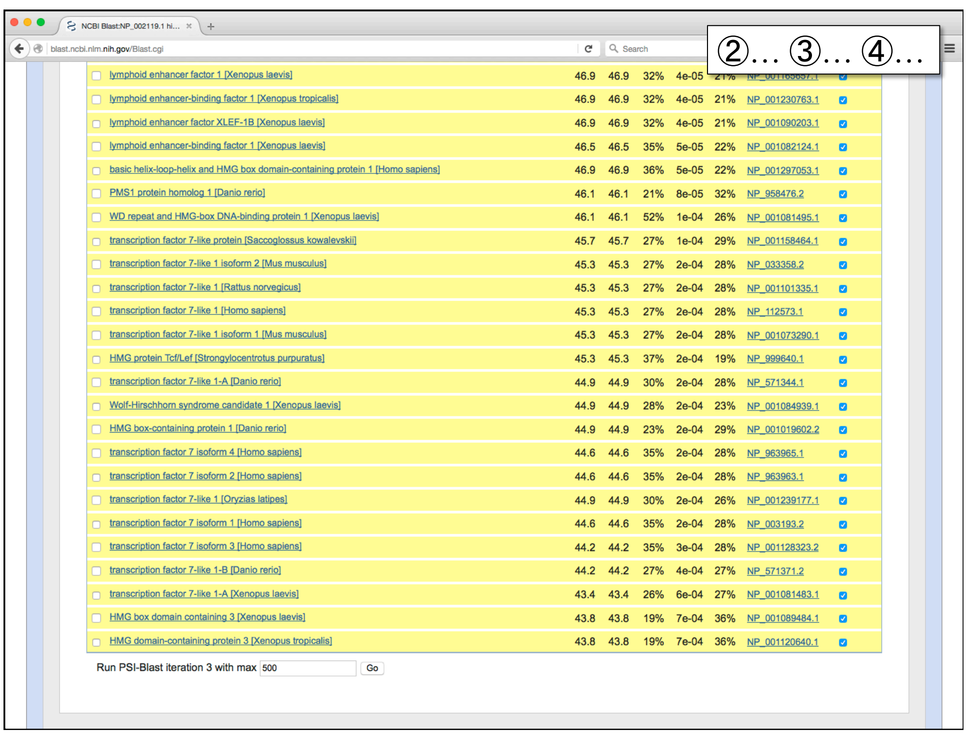
Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
high mobility group protein B1 [Bos taurus]	310	310	78%	7e-106	100%	NP_788785.1	☑	
high mobility group protein B1 [Mus musculus]	310	310	78%	7e-106	100%	NP_034569.1	☑	
high mobility group protein B1 [Homo sapiens]	310	310	78%	7e-106	100%	NP_002119.1	☑	
high mobility group protein B1 [Sus scrofa]	308	308	78%	5e-105	99%	NP_001004034.1	☑	
high mobility group box 1 like [Rattus norvegicus]	308	308	78%	7e-105	99%	NP_001102843.1	☑	
high mobility group protein B1 [Gallus gallus]	299	299	78%	2e-101	96%	NP_990233.1	☑	
high mobility group protein B1 [Xenopus tropicalis]	294	294	78%	2e-99	92%	NP_989226.1	☑	
high mobility group box 1 [Xenopus laevis]	290	290	77%	6e-98	92%	NP_001080836.1	☑	
high mobility group protein-1 [Xenopus laevis]	280	280	77%	7e-94	90%	NP_001081794.1	☑	
high mobility group protein B1 [Danio rerio]	266	266	77%	8e-89	87%	NP_001092721.2	☑	
high mobility group protein B1 [Danio rerio]	266	266	77%	2e-88	87%	NP_955849.2	☑	
HMG-X protein [Xenopus laevis]	262	262	78%	5e-87	84%	NP_001079576.1	☑	
High mobility group-T protein [Salmo salar]	264	264	77%	5e-86	84%	NP_001140081.1	☑	
high mobility group protein B1 [Salmo salar]	259	259	77%	6e-86	84%	NP_001133101.1	☑	
high mobility group protein B2 [Gallus gallus]	257	257	78%	5e-85	85%	NP_990817.1	☑	
high mobility group-T protein [Oncorhynchus mykiss]	257	257	77%	5e-85	83%	NP_001118186.1	☑	
high mobility group protein B2 [Homo sapiens]	252	252	78%	3e-83	86%	NP_002120.1	☑	
high mobility group protein B2 [Macaca fascicularis]	252	252	78%	4e-83	86%	NP_001271844.1	☑	
high mobility group protein B2 [Rattus norvegicus]	251	251	78%	1e-82	86%	NP_058883.1	☑	
high mobility group box 2 [Xenopus laevis]	250	250	78%	4e-82	82%	NP_001079387.1	☑	



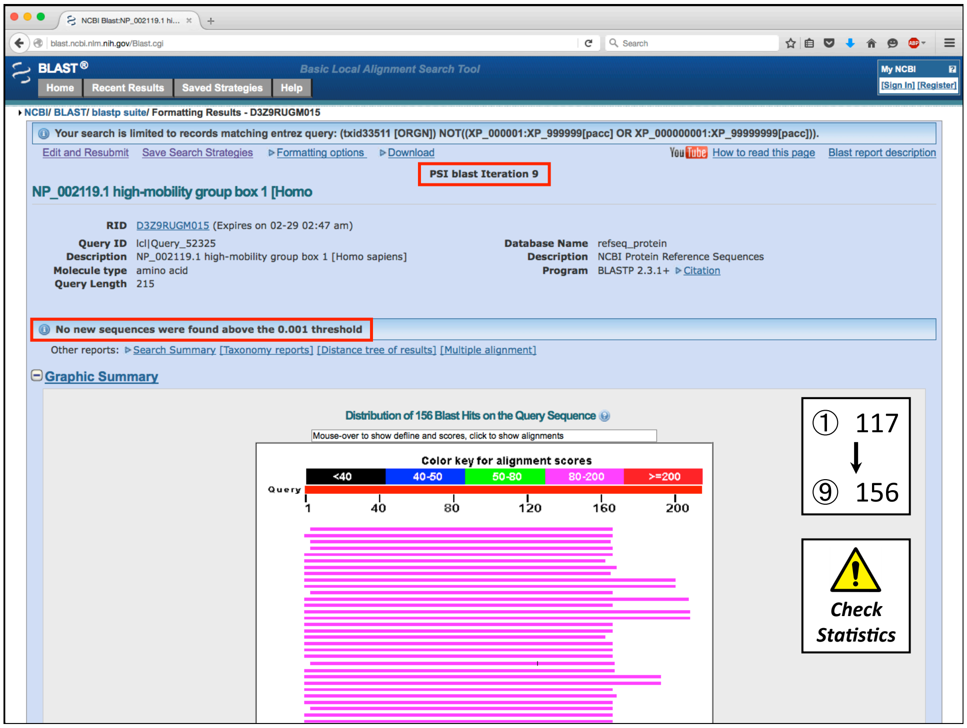
HMG domain-containing protein 4 [Homo sapiens]	47.8	47.8	21%	3e-05	43%	NP_001003681.1	☑
uncharacterized protein LOC399067 [Xenopus laevis]	47.4	94.3	62%	3e-05	37%	NP_001083698.1	☑
SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1-related [Xe	47.4	94.3	62%	3e-05	32%	NP_988941.1	☑
HMG domain-containing protein 4 [Bos taurus]	47.4	47.4	19%	3e-05	45%	NP_001095326.1	☑
TOX high mobility group box family member 4-A [Xenopus laevis]	47.4	47.4	24%	3e-05	38%	NP_001086364.1	☑
TOX high mobility group box family member 4 [Xenopus tropicalis]	47.4	47.4	24%	3e-05	38%	NP_001090624.1	☑
TOX high mobility group box family member 4-B [Xenopus laevis]	47.4	47.4	24%	4e-05	38%	NP_001084977.1	☑
transcription factor Ci-HMG20 [Ciona intestinalis]	46.2	46.2	23%	7e-05	39%	NP_001121587.1	☑
HMG domain-containing protein 4 [Danio rerio]	46.2	46.2	19%	7e-05	45%	NP_001120984.1	☑
SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1-related [Ho	45.4	89.7	62%	1e-04	37%	NP_006330.2	☑
HMG box-containing protein 4 [Xenopus laevis]	45.4	45.4	19%	1e-04	45%	NP_001082746.1	☑
transcription factor A, mitochondrial precursor [Rattus norvegicus]	45.1	88.2	64%	1e-04	29%	NP_112616.1	☑
SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1-related [Bo	45.1	89.7	62%	1e-04	37%	NP_001033143.1	☑
SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1-related [Mu	45.1	45.1	28%	1e-04	37%	NP_034570.1	☑
SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1-related [Ra	45.1	87.8	62%	2e-04	37%	NP_001102201.1	☑
high mobility group protein 20A [Xenopus tropicalis]	45.1	45.1	34%	2e-04	31%	NP_001006760.1	☑
HMG box-containing protein 4 [Xenopus tropicalis]	45.1	45.1	19%	2e-04	45%	NP_001025555.1	☑
protein polybromo-1 [Gallus gallus]	45.1	45.1	32%	3e-04	36%	NP_990496.1	☑
high mobility group protein 20A [Xenopus laevis]	44.3	44.3	34%	3e-04	31%	NP_001087141.1	☑
transcription factor A, mitochondrial precursor [Sus scrofa]	43.1	43.1	64%	5e-04	29%	NP_001123683.1	☑
high mobility group protein 20A [Gallus gallus]	43.5	43.5	34%	6e-04	31%	NP_001025565.1	☑
transcription factor protein [Ciona intestinalis]	43.5	43.5	33%	6e-04	36%	NP_001071666.1	☑
transcription factor protein [Ciona intestinalis]	43.1	43.1	33%	7e-04	36%	NP_001072029.1	☑
transcription factor A, mitochondrial [Esox lucius]	42.7	42.7	34%	9e-04	27%	NP_001297981.1	☑
transcription factor protein [Ciona intestinalis]	43.1	43.1	21%	0.001	40%	NP_001071952.1	☑
transcription factor A, mitochondrial [Danio rerio]	42.7	42.7	22%	0.001	31%	NP_001070857.1	☑



Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
high mobility group protein B1 [Bos taurus]	250	250	78%	3e-82	100%	NP_788785.1	☑	✓
high mobility group protein B1 [Mus musculus]	250	250	78%	3e-82	100%	NP_034569.1	☑	✓
high mobility group protein B1 [Homo sapiens]	250	250	78%	3e-82	100%	NP_002119.1	☑	✓
high mobility group protein B1 [Sus scrofa]	250	250	78%	3e-82	99%	NP_001004034.1	☑	✓
High mobility group-T protein [Salmo salar]	254	254	77%	5e-82	84%	NP_001140081.1	☑	✓
high mobility group box 1 like [Rattus norvegicus]	247	247	78%	3e-81	99%	NP_001102843.1	☑	✓
high mobility group protein B1 [Gallus gallus]	246	246	78%	8e-81	96%	NP_990233.1	☑	✓
high mobility group protein B1 [Danio rerio]	236	236	77%	6e-77	87%	NP_001092721.2	☑	✓
high mobility group protein B1 [Xenopus tropicalis]	236	236	78%	6e-77	92%	NP_989226.1	☑	✓
high mobility group box 1 [Xenopus laevis]	235	235	77%	2e-76	92%	NP_001080836.1	☑	✓
high mobility group protein B1 [Salmo salar]	234	234	77%	6e-76	84%	NP_001133101.1	☑	✓
high mobility group protein-1 [Xenopus laevis]	232	232	77%	4e-75	90%	NP_001081794.1	☑	✓
HMG-X protein [Xenopus laevis]	232	232	78%	4e-75	84%	NP_001079576.1	☑	✓
high mobility group protein B1 [Danio rerio]	230	230	77%	1e-74	87%	NP_955849.2	☑	✓
high mobility group protein B3 [Danio rerio]	230	230	77%	2e-74	66%	NP_001116308.1	☑	✓
High mobility group protein B3 [Salmo salar]	225	225	77%	2e-72	67%	NP_001133971.1	☑	✓
high mobility group protein B2 [Gallus gallus]	224	224	78%	3e-72	85%	NP_990817.1	☑	✓
high mobility group box 3 [Callorhinchus milii]	223	223	77%	6e-72	77%	NP_001279444.1	☑	✓
high mobility group protein B3 [Danio rerio]	223	223	75%	8e-72	68%	NP_001017769.1	☑	✓
high mobility group-T protein [Oncorhynchus mykiss]	223	223	77%	9e-72	83%	NP_001118186.1	☑	✓



Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast
lymphoid enhancer factor 1 [Xenopus laevis]	46.9	46.9	32%	4e-05	21%	[illegible]	☑
lymphoid enhancer-binding factor 1 [Xenopus tropicalis]	46.9	46.9	32%	4e-05	21%	NP_001230763.1	☑
lymphoid enhancer factor XLEF-1B [Xenopus laevis]	46.9	46.9	32%	4e-05	21%	NP_001090203.1	☑
lymphoid enhancer-binding factor 1 [Xenopus laevis]	46.5	46.5	35%	5e-05	22%	NP_001082124.1	☑
basic helix-loop-helix and HMG box domain-containing protein 1 [Homo sapiens]	46.9	46.9	36%	5e-05	22%	NP_001297053.1	☑
PMS1 protein homolog 1 [Danio rerio]	46.1	46.1	21%	8e-05	32%	NP_958476.2	☑
WD repeat and HMG-box DNA-binding protein 1 [Xenopus laevis]	46.1	46.1	52%	1e-04	26%	NP_001081495.1	☑
transcription factor 7-like protein [Saccoglossus kowalevskii]	45.7	45.7	27%	1e-04	29%	NP_001158464.1	☑
transcription factor 7-like 1 isoform 2 [Mus musculus]	45.3	45.3	27%	2e-04	28%	NP_033358.2	☑
transcription factor 7-like 1 [Rattus norvegicus]	45.3	45.3	27%	2e-04	28%	NP_001101335.1	☑
transcription factor 7-like 1 [Homo sapiens]	45.3	45.3	27%	2e-04	28%	NP_112573.1	☑
transcription factor 7-like 1 isoform 1 [Mus musculus]	45.3	45.3	27%	2e-04	28%	NP_001073290.1	☑
HMG protein Tcf/Lef [Strongylocentrotus purpuratus]	45.3	45.3	37%	2e-04	19%	NP_999640.1	☑
transcription factor 7-like 1-A [Danio rerio]	44.9	44.9	30%	2e-04	28%	NP_571344.1	☑
Wolf-Hirschhorn syndrome candidate 1 [Xenopus laevis]	44.9	44.9	28%	2e-04	23%	NP_001084939.1	☑
HMG box-containing protein 1 [Danio rerio]	44.9	44.9	23%	2e-04	29%	NP_001019602.2	☑
transcription factor 7 isoform 4 [Homo sapiens]	44.6	44.6	35%	2e-04	28%	NP_963965.1	☑
transcription factor 7 isoform 2 [Homo sapiens]	44.6	44.6	35%	2e-04	28%	NP_963963.1	☑
transcription factor 7-like 1 [Oryzias latipes]	44.9	44.9	30%	2e-04	26%	NP_001239177.1	☑
transcription factor 7 isoform 1 [Homo sapiens]	44.6	44.6	35%	2e-04	28%	NP_003193.2	☑
transcription factor 7 isoform 3 [Homo sapiens]	44.2	44.2	35%	3e-04	28%	NP_001128323.2	☑
transcription factor 7-like 1-B [Danio rerio]	44.2	44.2	27%	4e-04	27%	NP_571371.2	☑
transcription factor 7-like 1-A [Xenopus laevis]	43.4	43.4	26%	6e-04	27%	NP_001081483.1	☑
HMG box domain containing 3 [Xenopus laevis]	43.8	43.8	19%	7e-04	36%	NP_001089484.1	☑
HMG domain-containing protein 3 [Xenopus tropicalis]	43.8	43.8	19%	7e-04	36%	NP_001120640.1	☑



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

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

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 can sometimes be intractable
   - 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 such as BLAST
   - Profile-based search methods such as PSI-BLAST

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

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

Inspection: An Iterative Process

- Use visualization tools to identify 'key residues' and 'problem regions'
- 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

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
- Aligns data sets quickly and easily
- Can align sequences against a pre-existing alignment (an 'external profile')
- Can bias the location of gaps, based on known structural information
- Works with Jalview, a 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

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 can 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

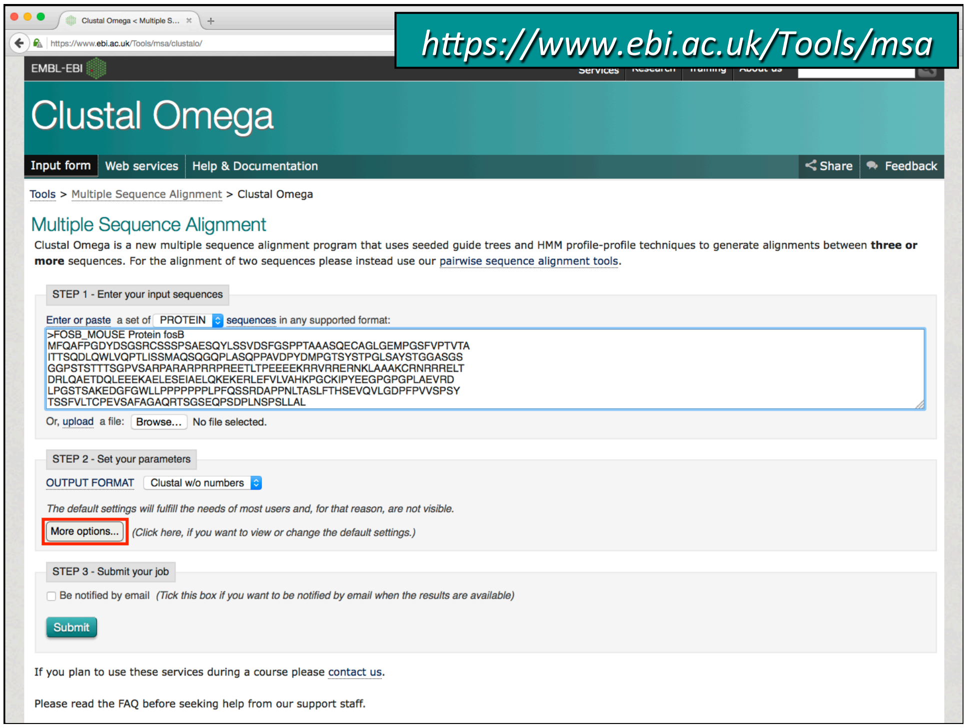
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

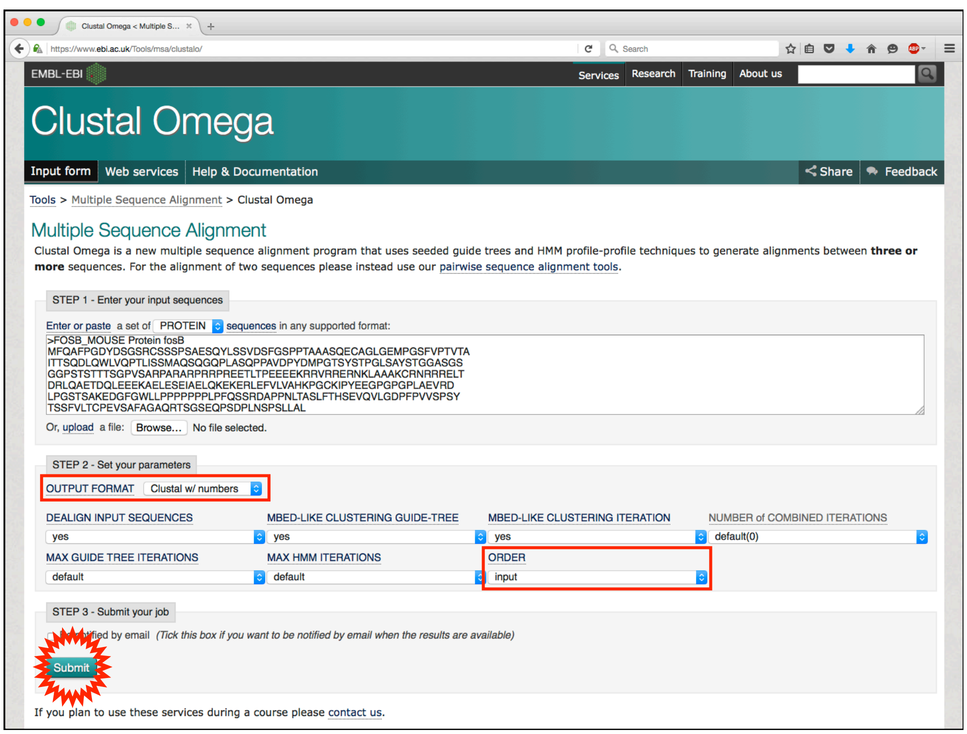
Clustal Omega Conservation Patterns

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

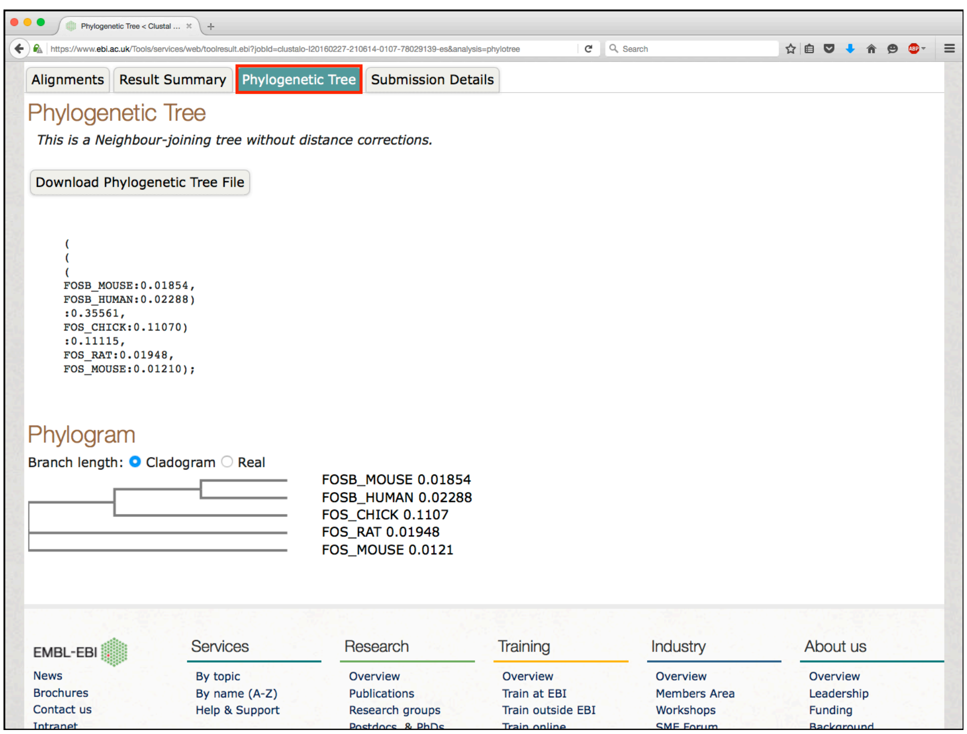
*	entirely conserved column (want in at least 10% of positions)
:	conserved (strongly similar properties)
.	semi-conserved (weakly similar properties)

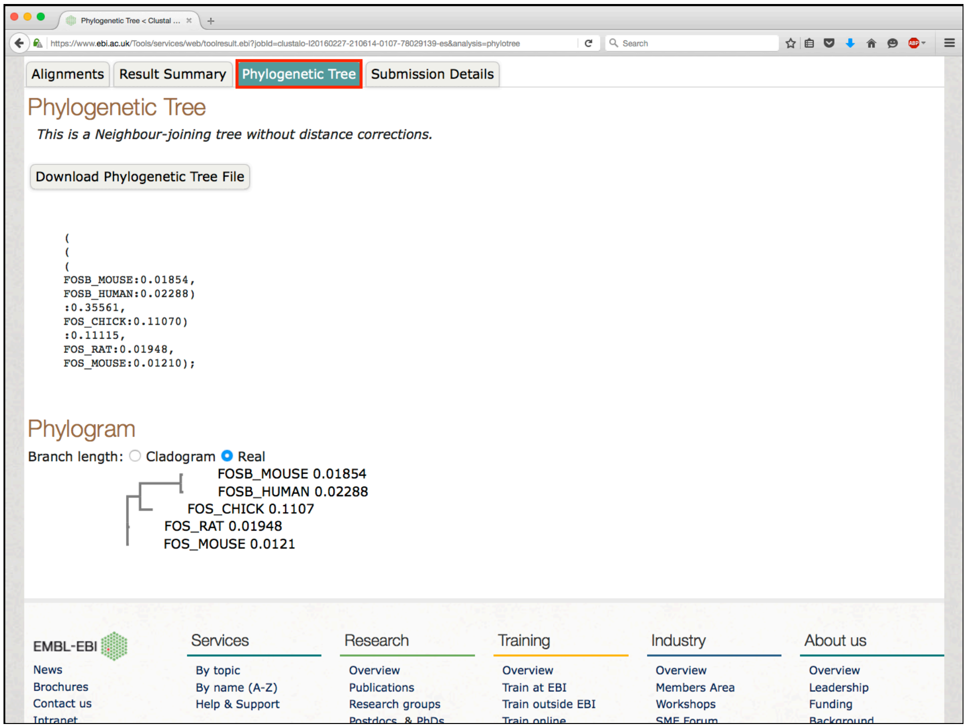
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

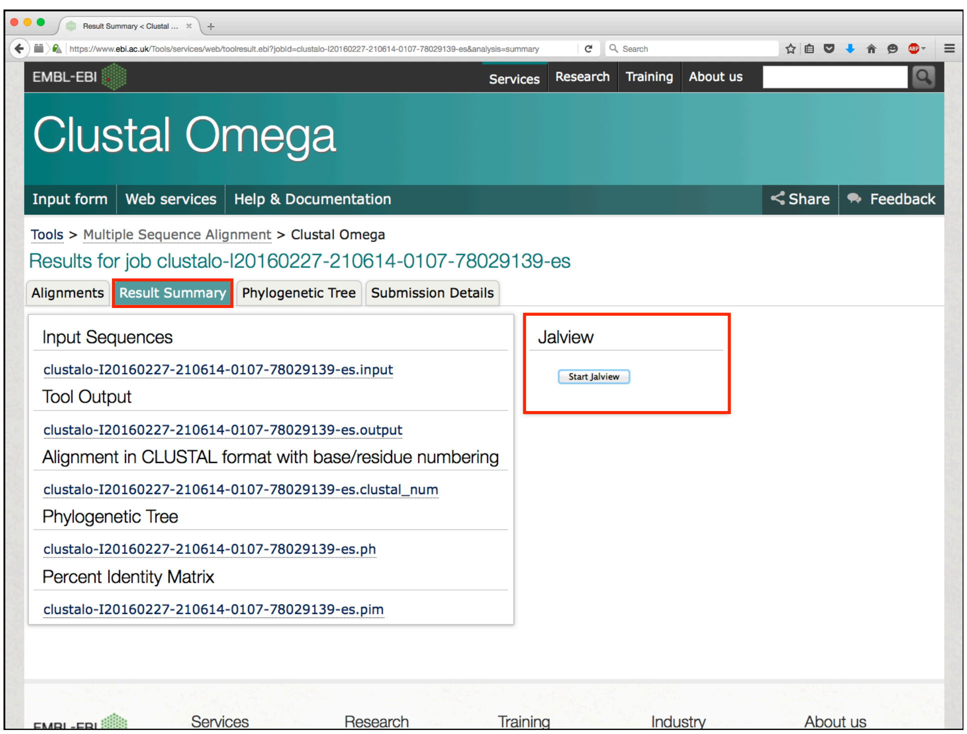

https://www.ebi.ac.uk/Tools/msa
EMBL-EBI
Clustal Omega
Input form
Web services
Help & Documentation
Share
Feedback
Tools > Multiple Sequence Alignment > Clustal Omega
Multiple Sequence Alignment
Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.
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.


https://www.ebi.ac.uk/Tools/msa/clustalo/
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 between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.
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 yes
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 input
STEP 3 - Submit your job
(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
Download Alignment File
Hide Colors
Send to ClustalW2
CLUSTAL O(1.2.1) multiple sequence alignment
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
FOSB_MOUSE
FOSB_HUMAN
FOS_CHICK
FOS_RAT
FOS_MOUSE


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
Services
By topic
By name (A-Z)
Help & Support
Research
Overview
Publications
Research groups
Training
Overview
Train at EBI
Train outside EBI
Industry
Overview
Members Area
Workshops
About us
Overview
Leadership
Funding


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
Services
By topic
By name (A-Z)
Help & Support
Research
Overview
Publications
Research groups
Training
Overview
Train at EBI
Train outside EBI
Industry
Overview
Members Area
Workshops
About us
Overview
Leadership
Funding

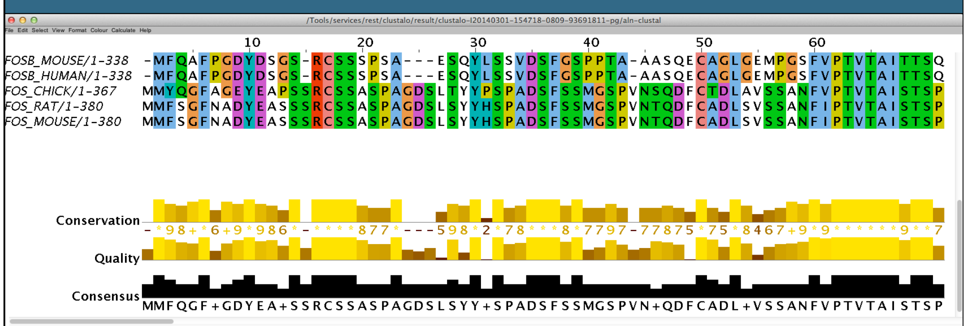

EMBL-EBI
Services
Research
Training
About us
Clustal Omega
Input form
Web services
Help & Documentation
Share
Feedback
Tools > Multiple Sequence Alignment > Clustal Omega
Results for job clustalo-I20160227-210614-0107-78029139-es
Alignments
Result Summary
Phylogenetic Tree
Submission Details
Input Sequences
clustalo-I20160227-210614-0107-78029139-es.input
Tool Output
clustalo-I20160227-210614-0107-78029139-es.output
Alignment in CLUSTAL format with base/residue numbering
clustalo-I20160227-210614-0107-78029139-es.clustal_num
Phylogenetic Tree
clustalo-I20160227-210614-0107-78029139-es.ph
Percent Identity Matrix
clustalo-I20160227-210614-0107-78029139-es.pim
Jalview
Start Jalview

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

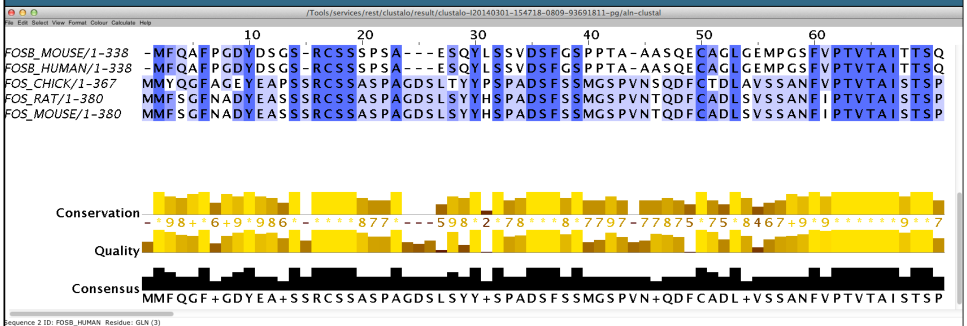
Default view



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

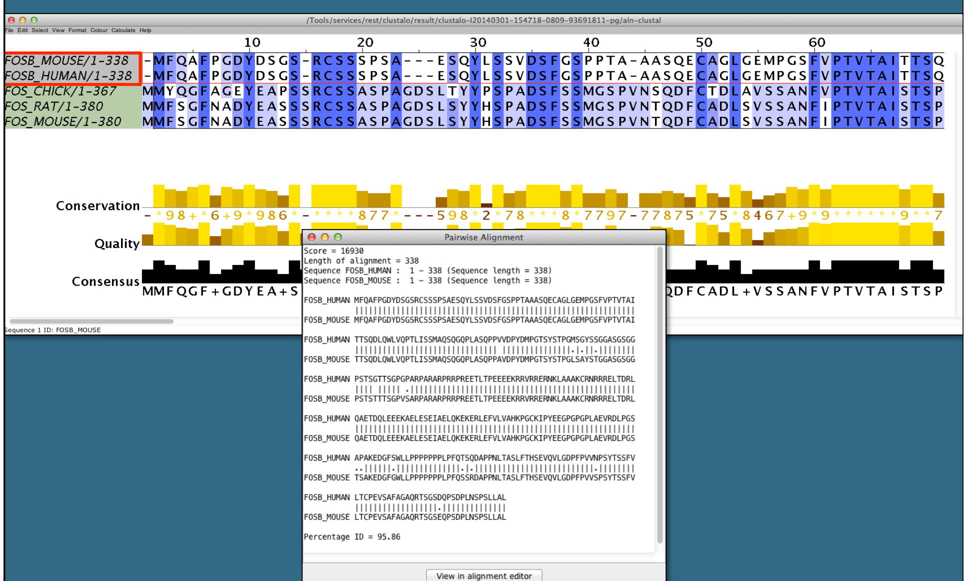
Colour → Percentage Identity



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

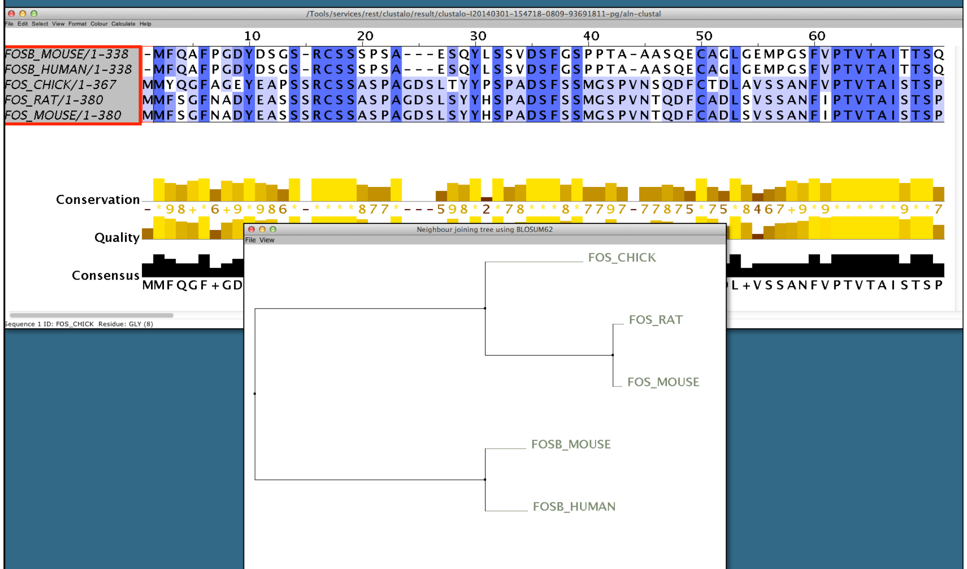
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Calculate → Pairwise Alignments



NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Calculate → Calculate Tree → Neighbour Joining Using BLOSUM62



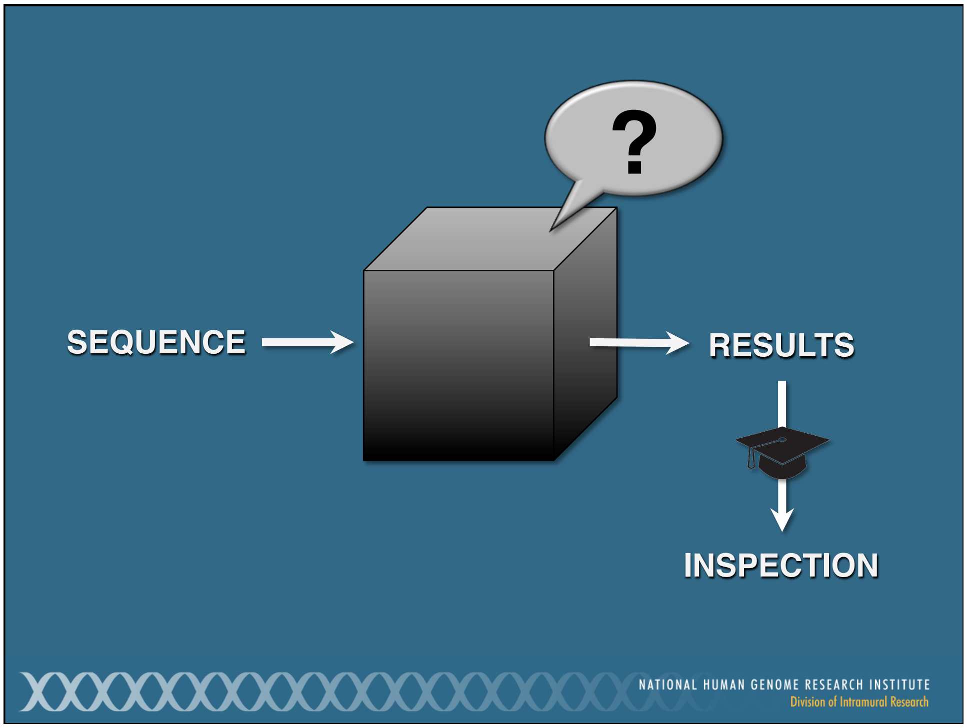
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

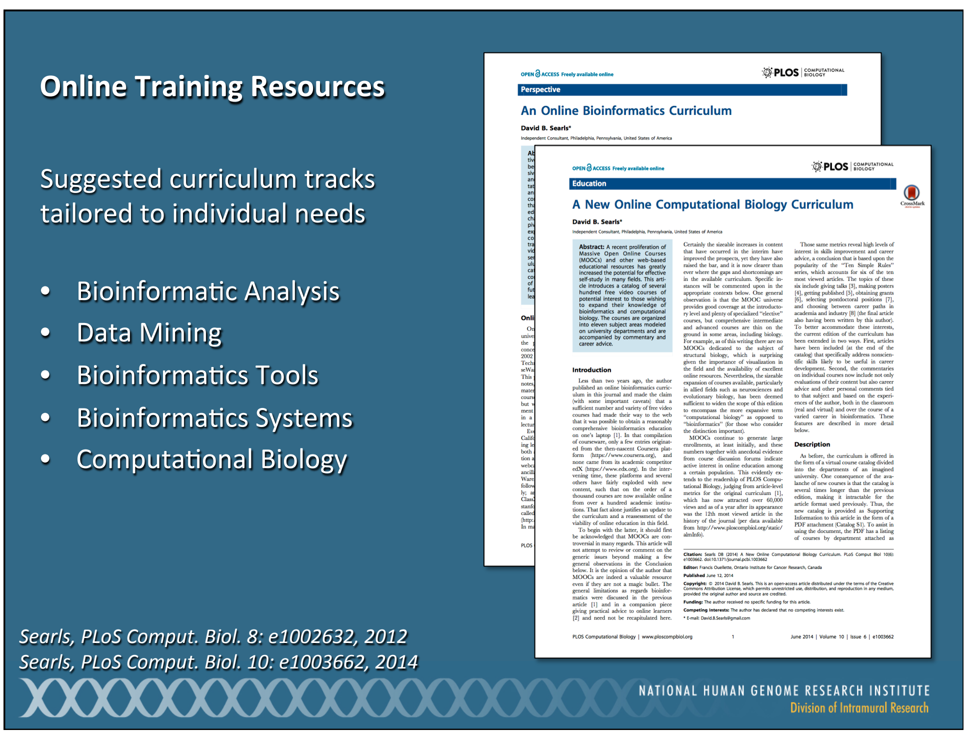
T-COFFEE

- 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


?
SEQUENCE
RESULTS
INSPECTION
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
Education
A New Online Computational Biology Curriculum
David B. Searls
Searls, PLoS Comput. Biol. 8: e1002632, 2012
Searls, PLoS Comput. Biol. 10: e1003662, 2014
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
Division of Intramural Research

Current Topics in Genome Analysis 2016

Next Lecture

March 16, 2014

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

NIH Intramural Research Program
Our Research Changes Lives

one program
many people
infinite possibilities

irp.nih.gov