




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



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*



*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

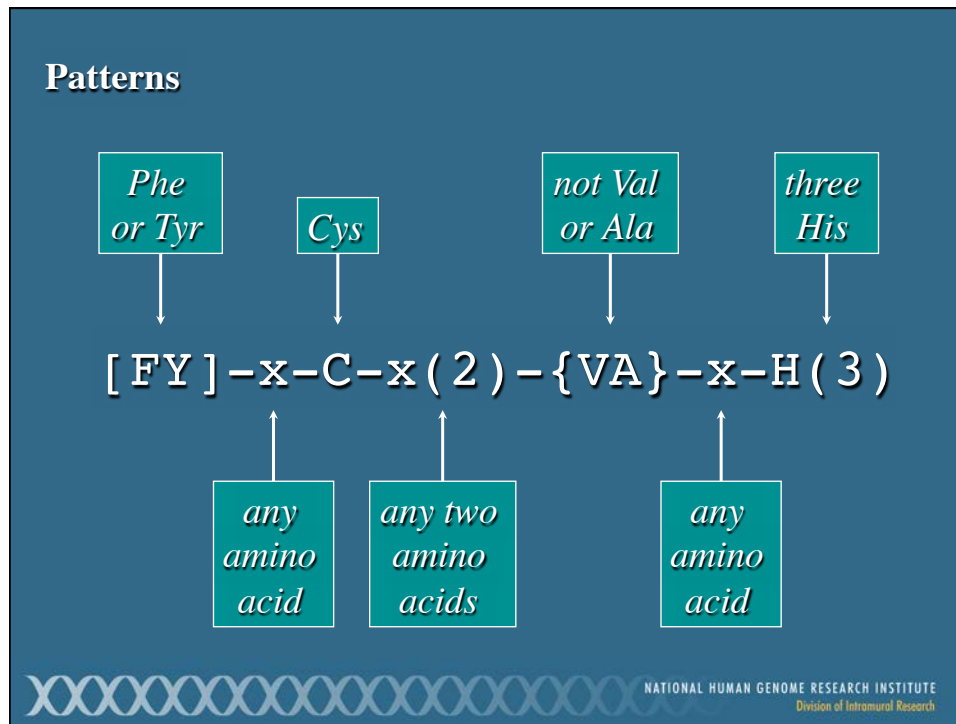
Profile Construction

APHIIVATPG
 GCEIIVATPG
 GVEICIAATPG
 GVDILIGTTC
 RPHIIVATPG
 KPHIIVATPG
 KVQLIIATPG
 RPDIVIAATPG
 APHIIVGTPG
 APHIIVGTPG
 GCHVVIATPG
 NQDIVVATTC

- 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	10	0	13	0	0	-12	13	0	0	0	0	0	0	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	-4	6	50	-19	2	-8
A	5	-9	9	-9	19	-1	-13	57	-9	35	26	-13	-2	-2	-11	-13	-4	9	58	-29	0	-9
T	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
R	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	70	60	20	70	30	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30



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

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

Sequences Used in Examples

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

The screenshot shows a web browser window displaying the NHGRI website. The page title is "Current Topics in Genome Analysis 2014". The main content area is titled "Current Topics in Genome Analysis 2014" and "Weeks 2 and 3: Biological Sequence Analysis Protein and Nucleotide Sequences for Analysis". The page contains several sections of text, including "ELAST7" with a long sequence of amino acids, "ELAST 2 Sequences" with a sequence of nucleotides, and "ELAST" with a sequence of amino acids. The page also includes a navigation menu with links for "Research Funding", "Research at NHGRI", "Health", "Education", "Issues in Genetics", "Newsroom", and "Careers & Training".

http://pfam.sanger.ac.uk

welcome 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

- SEQUENCE SEARCH**
- VIEW A PFAM FAMILY**
- VIEW A CLAN**
- VIEW A SEQUENCE**
- VIEW A STRUCTURE**
- KEYWORD SEARCH**
- JUMP TO**

YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

- Analyze your protein sequence for Pfam matches
- View Pfam family annotation and alignments
- See groups of related families
- Look at the domain organisation of a protein sequence
- Find the domains on a PDB structure
- Query Pfam by keywords

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequences, 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)

Pfam Home page

http://pfam.sanger.ac.uk

welcome 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

- SEQUENCE SEARCH**
- VIEW A PFAM FAMILY**
- VIEW A CLAN**
- VIEW A SEQUENCE**
- VIEW A STRUCTURE**
- KEYWORD SEARCH**
- JUMP TO**

ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES

Paste your protein sequence here to find matching Pfam families.

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

Recent Pfam blog posts Hide this

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

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

Protein sequence options

Cut-off Gathering threshold Use E-value

E-value

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

[Example protein sequence](#) [Example DNA sequence](#)

Sequence search results

[Show](#) the detailed description of this results page.

We found **3** Pfam-A matches to your search sequence (**1** significant and **2** insignificant). You did not choose to search for Pfam-B matches.

[Show](#) the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

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

Insignificant Pfam-A Matches

[Show](#) or [hide](#) all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
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 _{AAA}	Domain	CL0295	246	286	249	277	42	70	142	13.3	0.047	n/a	Show

PFAM Sequence search results

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 shown, 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 paper⁰.
 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 **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.

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	End
				Start	End		
p450	Cytochrome P450	Domain	n/a	41	505		

Insignificant Pfam-A Matches
 Show or hide all alignments.

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

PFAM Family: p450 (PF00067)

wellcome trust sanger | HOME | SEARCH | BROWSE | FTP | HELP | ABOUT | Pfam

Family: p450 (PF00067)
 282 structures | 36592 sequences | 2 interactions | 2977 species | 873 structures

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 topology and structural fold are highly conserved. The conserved core is composed of a coil termed the 'manicor', 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 EXOR motif in helix K. While prokaryotic P450s are soluble proteins, most eukaryotic P450s are associated with microsomal membranes, their general enzymatic function is to catalyse regioselective and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].

Literature references

- Graham-Lorence S, Amarnath B, White RE, Peterson JA, Simpson ER; . Protein Sci 1995;4:1065-1060.: A three-dimensional model of aromatase cytochrome P450. PUBMED:7549871 | EPMC:7549871 |
- Deptyarenko KN, Archakov AI; . FEBS Lett 1993;332:1-6.: Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. PUBMED:8405421 | EPMC:8405421 |
- Neison DR, Kamatani T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, Gonzalez FJ, Coon MJ, Gunsalus IC, Gotoh O, et al; . DNA Cell Biol 1993;12:1-51.: The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. R. (PMED:7578494) | EPMC:7578494 |
- Guengerich FP; . J Biol Chem 1991;266:10019-10022.: Reactions and significance of cytochrome P-450 enzymes. PUBMED:2037557 | EPMC:2037557 |
- Nebert DW, Gonzalez FJ; . Annu Rev Biochem 1987;56:945-993.: P450 genes: structure, evolution, and regulation. PUBMED:3304150 | EPMC:3304150 |
- Wörck-Reichhart D, Feyereisen R; . Genome Biol 2000;1:REVIEWS3003.: Cytochromes P450: a success story. PUBMED:11178272 | EPMC:11178272 |

External database links

HOMSTRAD:	p450
PANDBIT:	PF00067
PRINTS:	PR00385 PR00359 PR00408 PR00463 PR00464 PR00465

Family: p450 (PF00067)

Domain organisation

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

There are **33674** sequences with the following architecture: **p450**
 Q99N18_MOUSE [Mus musculus (Mouse)] Protein Cyp4115 (534 residues)
 Show all sequences with this architecture.

There are **2329** sequences with the following architecture: **p450 x 2**
 Q27YK6_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**
 Q8KUJ0_ACTYA [Actinosynnema pretiosum subsp. aurumicum] Cytochrome P450 (1005 residues)
 Show all sequences with this architecture.

There are **88** sequences with the following architecture: **p450 x 3**
 Q331R6_SACTO [Scopelomyces sp. KCTC 00418P] Cytochrome P-450 (401 residues)
 Show all sequences with this architecture.

There are **86** sequences with the following architecture: **An_peroxidase, p450**
 Q2TW67_ASPOP [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**
 A1J298_BURPM [Burkholderia mallei (strain SA911)] Cytochrome P450 (794 residues)
 Show all sequences with this architecture.

There are **36** sequences with the following architecture: **An_peroxidase x 2, p450**
 Q0C299_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**
 Q3X568_BDAPI [Paraburkholderia denitrificans (Florida lineage) (6-methyloligul)] Putative uncharacterized protein (652 residues)

Family: p450 (PF00067)

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				NCBI (39456)	Meta (2723)
			RP15 (5443)	RP35 (11134)	RP55 (14825)	RP75 (20669)		
Jobview	✓	✓	✓	✓	✓	✓	✓	✓
HTML	✓	✓	—	—	—	—	✗	✗
PP/heatmap	✗	✗	—	—	—	—	✗	✗
Pfam viewer	✓	✓	✗	✗	✗	✗	✗	✗

✗ Cannot generate PP/heatmap alignments for seeds; no PP data available.

Key: ✓ available, ✗ not generated, — not available.

Format an alignment

	Seed (50)	Full (39592)	Representative proteomes				NCBI (39456)	Meta (2723)
			RP15 (5443)	RP35 (11134)	RP55 (14825)	RP75 (20669)		
Alignment:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Format:	Select							
Order:	<input type="radio"/> Tree <input type="radio"/> Alphabetical							
Sequence:	<input type="radio"/> Inserts lower case <input type="radio"/> All upper case							
Gaps:	Caps as "." or "-" (mixed)							
Download/View:	<input type="radio"/> Download <input type="radio"/> 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.

	Seed (50)	Full (39592)	Representative proteomes				NCBI (39456)	Meta (2723)
			RP15 (5443)	RP35 (11134)	RP55 (14825)	RP75 (20669)		
Alignment:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

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

Literature references

- Graham-Lorence S, Amarnath B, White RE, Peterson JA, Simpson ER; *Protein Sci* 1995;4:1065-1060. A three-dimensional model of aromatase cytochrome P450. [PubMed:7549871](#) [EMBL:7549871](#)
- DeGyarekeno KN, Archakov AJ; *FEBS Lett* 1993;332:1-6. Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. [PubMed:10405421](#) [EMBL:10405421](#)
- Neison DR, Kamatani T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, Gonzalez FJ, Coon MJ, Gunsalus IC, Gotoh O, et al; *DNA Cell Biol* 1993;12:1-31. The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. [PubMed:7578494](#) [EMBL:7578494](#)
- Guengerich FP; *J Biol Chem* 1991;266:10019-10022. Reactions and significance of cytochrome P-450 enzymes. [PubMed:2037557](#) [EMBL:2037557](#)
- Neibert DW, Gonzalez FJ; *Annu Rev Biochem* 1987;56:945-993. P450 genes: structure, evolution, and regulation. [PubMed:3304150](#) [EMBL:3304150](#)
- Wreck-Reichhart D, Feyereisen R; *Genome Biol* 2000;1:REVIEWS3003. Cytochromes P450: a success story. [PubMed:11178272](#) [EMBL:11178272](#)

External database links

HOMSTRAD:	p450
PANDIT:	PF00067
PRINTS:	PRD0385 PRD0359 PRD0408 PRD0463 PRD0464 PRD0465
Pseudofam:	PF00067
SCOP:	2cnp
SYSTEMS:	p450

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

Literature references

- Graham-Lorence S, Amarnath B, White RE, Peterson JA, Simpson ER; *Protein Sci* 1995;4:1065-1060. A three-dimensional model of aromatase cytochrome P450. [PubMed:7549871](#) [EMBL:7549871](#)
- DeGyarekeno KN, Archakov AJ; *FEBS Lett* 1993;332:1-6. Molecular evolution of P450 superfamily and P450-containing monooxygenase systems. [PubMed:10405421](#) [EMBL:10405421](#)
- Neison DR, Kamatani T, Waxman DJ, Guengerich FP, Estabrook RW, Feyereisen R, Gonzalez FJ, Coon MJ, Gunsalus IC, Gotoh O, et al; *DNA Cell Biol* 1993;12:1-31. The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. [PubMed:7578494](#) [EMBL:7578494](#)
- Guengerich FP; *J Biol Chem* 1991;266:10019-10022. Reactions and significance of cytochrome P-450 enzymes. [PubMed:2037557](#) [EMBL:2037557](#)
- Neibert DW, Gonzalez FJ; *Annu Rev Biochem* 1987;56:945-993. P450 genes: structure, evolution, and regulation. [PubMed:3304150](#) [EMBL:3304150](#)
- Wreck-Reichhart D, Feyereisen R; *Genome Biol* 2000;1:REVIEWS3003. Cytochromes P450: a success story. [PubMed:11178272](#) [EMBL:11178272](#)

External database links

HOMSTRAD:	p450
PANDIT:	PF00067
PRINTS:	PRD0385 PRD0359 PRD0408 PRD0463 PRD0464 PRD0465
PROSITE:	PDOC00081
Pseudofam:	PF00067
SCOP:	2cnp
SYSTEMS:	p450

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk. Our cookie policy.

The Wellcome Trust

PROSITE documentation PDOC00081
Cytochrome P450 cysteine heme-iron ligand signature

Description

Cytochrome P450s [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, P450s have been classified into about forty different families [4,5]. P450s are proteins of 400 to 530 amino acids; the only exception is *Bacillus BM-3* (CYP1102) which is a protein of 1048 residues that contains a N-terminal P450 domain followed by a reductase domain. P450s are heme proteins. A conserved cysteine residue in the C-terminal part of P450s 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 P450s.

Note:
 The term 'cytochrome' P450, while commonly used, is incorrect as P450s 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-[BGNH]-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 'outliers')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS00086: 48 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 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 1BVV 1CBJ ... [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

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



<http://ncbi.nlm.nih.gov/Structure>

NCBI Conserved Domain Search

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 EASIA format

OPTIONS

Search against database (D):

Expect Value (E) threshold:

Apply low-complexity filter

Force live search

Maximum number of hits (D):

Result mode (D): Concise Standard Full

Retrieve previous CD-search result

Request ID:

References:

- Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", *Nucleic Acids Res.* **39**(D):225-9.
- Marchler-Bauer A et al. (2009), "CDD: specific functional annotation with the Conserved Domain Database.", *Nucleic Acids Res.* **37**(D):205-10.
- Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", *Nucleic Acids Res.* **32**(W):327-331.

Help | Disclaimer | Write to the Help Desk
 NCBI | NLM | NIH

Conserved domains on [cd05722]

List of domain hits

Accession	Description	Psacid	Multi-dom	E-value
Ig1_Neogenin[cd05722]	First immunoglobulin (Ig)-like domain in neogenin and similar proteins; Ig1_Neogenin: first immunoglobulin (Ig)-like domain ...	143199	no	6.80e-50
H1FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	3.76e-16
H2FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	1.17e-17
H3FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	2.22e-16
H4FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	2.04e-13
Ig[cd00666]	Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig superfamily is a ...	143165	yes	1.03e-10
H1FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	3.63e-10
H2FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	9.45e-09
Ig super family[cd11960]	Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig superfamily is a ...	264487	no	8.84e-05
Ig[cd00666]	Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig superfamily is a ...	264487	no	1.40e-14
Ig1_Neogenin_C[cd05722]	Neogenin C-terminus; This family represents the C-terminus of eukaryotic neogenin precursor proteins, which ...	253838	yes	5.62e-143
H1set[sm07670]	Immunoglobulin I-set domain;	254352	yes	3.24e-19
H2set[sm07670]	Immunoglobulin I-set domain;	254352	yes	5.04e-19
H1C2[sm00408]	Immunoglobulin C-2 Type;	197706	yes	7.30e-16
H2C2[sm00408]	Immunoglobulin C-2 Type;	197706	yes	1.04e-08

Ig1_Neogenin [cd05722]

Description: 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.

Conserved Domain Architecture: FN3 FN3 FN3 FN3 FN3 FN3

Sequence Alignment:

```

seqsig_646d6fc2d397e2ad4c43c4e6ae54c1b1 41  ... 90
cd1cd05722 1  MFLKRPD1YAVTGGPVLN:SHGEP-PYI:KFKGVLLNVSDRRQLPHGSLIITGVVSRKPKPSPGPFQCVIAQ 79
                             90
seqsig_646d6fc2d397e2ad4c43c4e6ae54c1b1 121  LDRSLISLITVAWAV 136
cd1cd05722 80  NDSLGLIVIRKRLDVTY 95
    
```

List of domain hits

Accession	Description	Psacid	Multi-dom	E-value
Ig1_Neogenin[cd05722]	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.	143199	no	6.80e-50
H1FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	3.76e-16
H2FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	1.17e-17
H3FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	2.22e-16
H4FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	2.04e-13
Ig[cd00666]	Immunoglobulin domain; Ig: immunoglobulin (Ig) domain found in the Ig superfamily. The Ig superfamily is a ...	143165	yes	1.03e-10
H1FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	3.63e-10
H2FN3[cd00663]	Fibronectin type 3 domain; One of three types of internal repeats found in the plasma ...	238020	no	9.45e-09

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: Representative, Specific Protein, Related Protein, Related Structure, Architectures
 Superfamily: cl11960
 BioSystems: 310 links

Statistics
 PSSM-Td: 143199
 View PSSM: cd05722
 Aligned: 7 rows
 ThresholdBitScore: 148.395
 ThresholdSettingG: 148277558
 Created: 27-Sep-2007
 Updated: 17-Jan-2013

Structure
 Interactive View
 Aligned Rows: All 7 rows
 Download Cn3D

PubMed References
 Neogenin: one receptor; many functions. *Int J Biochem Cell Biol* 2007; 39(5):674-676
 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(5):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* 1987 May 1; 4(3):414-421
 The immunoglobulin fold. Structural classification, sequence patterns and common core. *J Mol Biol* 1994 Sep 30; 242(4):308-320
 The immunoglobulin superfamily: an insight on its topology, species, and functional diversity. *J Mol Evol* 1999 Apr; 49(4):395-402
 Evolution of antigen binding receptors. *Annu Rev Immunol* 1996; 17:109-147

cd05722 is part of a hierarchy of related CD models. Use the graphical representation to navigate this hierarchy. **cd05722** is a member of the superfamily cl11960.

cd05722 Sequence Cluster
 Dendrogram showing sequence relationships with bootstrap values.

Sub-family Hierarchy
 Interactive Display with CDTree

- cd05722 Ig1_Neogenin
- cd05723 Ig4_Neogenin
- cd05724 Ig2_Robe
- cd05725 Ig3_Robe
- cd05726 Ig4_Robe
- cd05727 Ig2_ConkactLin-2-like
- cd05728 Ig4_ConkactLin-2-like
- cd05729 Ig2_FGFRL1-like
- cd05856 Ig2_FIFRL1-like
- cd05857 Ig2_FGFR
- cd05730 Ig3_MCHN-1-like
- cd05731 Ig3_L1-CRH-like
- cd05876 Ig3_L1-CRH
- cd05732 Ig5_MCHN-1-like
- cd05863 Ig5_MCHN-1
- cd05870 Ig5_MCHN-2
- cd05733 Ig6_L1-CRH_Like
- cd05874 T.E. M-FRM

Structure
 Interactive View
 Aligned Rows: All 7 rows
 Download Cn3D

Hierarchy
 Interactive Display
 Display: cd05722 branch
 Download CDTree

LinkOut - more resources

Sequence Alignment
 Include consensus sequence
 Reformat: Format: Compact Hypertext Row Display: All 7 rows Color Bits: 2.0 bit Type Selection: top listed sequences

g1_62204258	35	WFSLEPSDZLA	[5]	VLLNCVNS	[3]	AKIENKEDGFLSL	[8]	LADGSLISSVNSK	[1]	HKPDEGYQCV	111
g1_110645196	48	YFLEPVDVTE	[5]	AVLMCSAVA	[3]	PKIENKEDGFLSL	[8]	LPSGSLISSVNSK	[1]	HKPDEGYQCV	124
g1_113675978	28	FFLEPVDVTA	[5]	VVLDCCARG	[3]	IGIENKEDGFLSL	[8]	LSNGSLISSVNSK	[1]	DKSDGDFYQCL	101
g1_148277558	30	WFLSEPSDIA	[5]	LVKRCVNS	[3]	IKIENKEDGFLSL	[8]	PTSGSLISSVNSK	[1]	GSDEGDFYQCL	108
g1_1169233	43	WFLSEPSDAVT	[5]	VLLDCSARS	[4]	PVIENKEDGFLSL	[8]	LSNGSLISSVNSK	[1]	HKPDEGLYQCR	118
g1_10720134	20	WFLSEPSDLS	[5]	VIMRCSTVC	[3]	PKIENKEDGFLSL	[8]	LPSGSLISSVNSK	[1]	HKPDEGYQCV	96
g1_147903889	43	WFLSEPSDAVT	[5]	VVLMCSAQS	[4]	PKIENKEDGFLSL	[8]	LPSGSLISSVNSK	[1]	HKPDEGYQCV	118

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*



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



The screenshot shows the NCBI BLAST homepage. At the top, there is a navigation bar with 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. A URL bar at the top right displays 'http://ncbi.nlm.nih.gov/BLAST'. The main content area is titled 'NCBI BLAST Home' and includes a search box, a 'New DELTA-BLAST' button, and a section for 'BLAST Assembled RefSeq Genomes' with a list of species like Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera. Below this is the 'Basic BLAST' section with links for 'nucleotide blast', 'protein blast', 'blastx', 'tblastn', and 'tblastx'. A red arrow points to the 'protein blast' link. The 'Specialized BLAST' section is also visible at the bottom.

The screenshot shows the 'Standard Protein BLAST' search interface. It includes a 'Query Subrange' section with 'From' and 'To' input fields. The 'Choose Search Set' section is highlighted with a red box and contains a 'Database' dropdown menu set to 'UniProtKB/Swiss-Prot[swissprot]', an 'Organism' dropdown set to 'Vertebrata (6467742)', and an 'Exclude' checkbox. The 'Program Selection' section shows 'blastp (protein-protein BLAST)' selected. At the bottom, there is a 'BLAST' button and a 'General Parameters' section with a 'Max target sequences' dropdown set to '500'.

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



Standard Protein BLAST

Enter Query Sequence

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

Or, upload file

Job Title

Choose Search Set

Database: UniProtKB/Swiss-Prot[swissprot]

Organism: Vertebrata (taxid:7742)

Exclude: Models (XM/XP) [] Uncultured/environmental sample sequences []

Entrez Query

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)

Algorithm parameters

General Parameters

Max target sequences: 500

The screenshot shows the Protein BLAST search page. Key parameters are highlighted with red boxes and arrows pointing to callouts:

- Expect threshold:** Set to 0.001. A callout box indicates "Default = 10".
- PSI-BLAST Threshold:** Set to 0.001. A callout box indicates "Default = 0.005".
- Filters and Masking:** The "Filter" section is highlighted with a red box, showing "Low complexity regions" selected.

At the bottom, a red starburst icon labeled "BLAST" is next to the search button.

The screenshot shows the BLAST results page for the query "NP_002119.1 high-mobility group box 1 [Homo sapiens]".

Query Information:

- Query ID: ic156894
- Description: NP_002119.1 high-mobility group box 1 [Homo sapiens]
- Molecule type: amino acid
- Query Length: 215

Database Information:

- Database Name: swissprot
- Description: Non-redundant UniProtKB/SwissProt sequences
- Program: BLASTP 2.2.29+

Graphic Summary:

- Putative conserved domains have been detected. Two domains are shown: "HMG-box superfamily" (residues 1-100 and 120-200).
- Distribution of 77 Blast Hits on the Query Sequence is shown as a horizontal bar chart.
- Color key for alignment scores: <40 (black), 40-80 (blue), 80-85 (green), 85-200 (red), >=200 (magenta).

NCBI BLAST search results - NCBINBLAST_002119.1 high-mobility group box 3 (HMG3)

Run PSI-Blast iteration 2 with max. 500

Sequences producing significant alignments with E-value BETTER than threshold

Description	Max score	Total score	Query cover	E value	Ident	Accession	Search to PSI blast	Used to build PSSM
RefName: Full=high mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1	310	310	78%	2e-106	100%	P10103.3	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein B1; AltName: Full=Angiostatin; AltName: Full=Heparin-binding protein p30; AltName: Full=Hhg	310	310	78%	2e-106	100%	P93199.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1; espQ6YKAA; Sp=HGB1_CANE	310	310	78%	2e-106	100%	P36429.3	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1	308	308	78%	1e-105	99%	P12682.3	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein B1; AltName: Full=High mobility group protein 1; Short=HMG-1	299	299	78%	4e-102	95%	Q0Y106.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Putative high mobility group protein B1-like 1; AltName: Full=High mobility group protein B1 pseudogene 1; AltName: Full=	297	297	78%	2e-101	95%	B29202.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Putative high mobility group protein 1 like 10; Short=HMG-1; 10	290	290	78%	1e-98	95%	Q05026.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein S2; AltName: Full=High mobility group protein 2; Short=HMG-2	257	257	78%	1e-85	85%	P26564.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group-T protein; Short=HMG-T; AltName: Full=HMG-T1; Short=HMG-1	257	257	77%	1e-85	83%	P07148.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein S2; AltName: Full=High mobility group protein 2; Short=HMG-2; espP46873; Sp=HMG2_B2UV	252	252	78%	5e-84	86%	P26563.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein S2; AltName: Full=High mobility group protein 2; Short=HMG-2	251	251	78%	2e-83	86%	P32925.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein S2; AltName: Full=High mobility group protein 2; Short=HMG-2	249	249	78%	2e-82	86%	P30881.3	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein S2; AltName: Full=High mobility group protein 2; Short=HMG-2	245	245	75%	5e-81	87%	P17241.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein S1; AltName: Full=High mobility group protein 1; Short=HMG-1	239	239	62%	3e-79	100%	P07186.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein S3; AltName: Full=High mobility group protein 2a; Short=HMG-2a; AltName: Full=High mobility	210	210	73%	2e-67	75%	O56375.3	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein S3; AltName: Full=High mobility group protein 2a; Short=HMG-2a; AltName: Full=High mobility	209	209	73%	5e-67	75%	P40619.3	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein S3	209	209	73%	5e-67	75%	Q32131.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein S3; AltName: Full=High mobility group protein 2a; Short=HMG-2a; AltName: Full=High mobility	208	208	73%	1e-66	75%	O15347.4	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Nuclear autoantigen Sp-100; AltName: Full=Nuclear dot-associated Sp100 protein; AltName: Full=Speckled 100 kDa	207	207	70%	6e-66	85%	Q0R106.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Nuclear autoantigen Sp-100; AltName: Full=Nuclear dot-associated Sp100 protein; AltName: Full=Speckled 100 kDa	201	201	70%	2e-63	85%	Q2N102.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Nuclear autoantigen Sp-100; AltName: Full=Nuclear dot-associated Sp100 protein; AltName: Full=Speckled 100 kDa	201	201	68%	2e-63	87%	Q0R106.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Nuclear autoantigen Sp-100; AltName: Full=Nuclear dot-associated Sp100 protein; AltName: Full=Speckled 100 kDa	211	211	75%	1e-62	82%	P33497.3	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Putative high mobility group protein S3-like protein	197	197	73%	2e-62	69%	P30865.1	<input type="checkbox"/>	<input type="checkbox"/>

NCBI BLAST search results - NCBINBLAST_002119.1 high-mobility group box 3 (HMG3)

Run PSI-Blast iteration 2 with max. 500

Sequences producing significant alignments with E-value BETTER than threshold

Description	Max score	Total score	Query cover	E value	Ident	Accession	Search to PSI blast	Used to build PSSM
RefName: Full=Thymocyte selection-associated high mobility group box protein TOX; AltName: Full=Thymus high mobility group box prot	49.7	49.7	34%	1e-06	41%	Q54900.3	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 3; AltName: Full=CAG trinucleotide repeat-containing gene F9 protein; AltName	49.7	49.7	34%	1e-06	40%	O15456.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 3; AltName: Full=Trinucleotide repeat-containing gene 9 protein	49.7	49.7	34%	1e-06	40%	Q00700.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 3; AltName: Full=Trinucleotide repeat-containing gene 9 protein	49.7	49.7	34%	1e-06	40%	B78902.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 2; AltName: Full=Granulosa cell HMG box protein 1; Short=GCCL1	48.9	48.9	24%	2e-06	40%	Q66384.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 4	47.8	47.8	24%	5e-06	38%	Q07964.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 4	47.8	47.8	24%	5e-06	38%	O586A8.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 4; AltName: Full=Epidermal Langerhans cell protein LCP1	47.8	47.8	24%	5e-06	38%	O54842.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 4; AltName: Full=Epidermal Langerhans cell protein LCP1	47.8	47.8	24%	5e-06	38%	Q099M1.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 4; AltName: Full=Epidermal Langerhans cell protein LCP1	47.8	47.8	24%	5e-06	38%	Q08111.3	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=HMG domain-containing protein 4; AltName: Full=HMG box-containing protein 4; AltName: Full=High mobility group prote	47.8	47.8	21%	5e-06	43%	Q04025.2	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 4A	47.4	47.4	24%	7e-06	38%	Q6038.0	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 4	47.4	47.4	24%	7e-06	38%	A529P0.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=TOX high mobility group box family member 4B	47.4	47.4	24%	8e-06	38%	Q07880.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=SWY5NF-related matrix-associated actin-dependent regulator of chromatin: subfamily E member 1-related; Short=SMARC	45.4	89.7	62%	2e-05	37%	Q05902.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=HMG box-containing protein 4; AltName: Full=High mobility group protein 2 like 1; Short=HMG2L1	45.4	45.4	19%	3e-06	45%	Q6YK06.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=SWY5NF-related matrix-associated actin-dependent regulator of chromatin: subfamily E member 1-related; Short=SMARC	45.1	89.7	62%	3e-05	37%	Q32136.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=SWY5NF-related matrix-associated actin-dependent regulator of chromatin: subfamily E member 1-related; Short=SMARC	45.1	87.4	62%	3e-05	37%	Q52104.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein 20A; AltName: Full=HMG box-containing protein 20A	45.1	45.1	34%	4e-05	31%	Q60505.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=HMG box-containing protein 4; AltName: Full=High mobility group protein 2 like 1	45.1	45.1	19%	4e-05	45%	Q68156.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Protein polypyrone-1	45.1	45.1	32%	6e-05	36%	Q09911.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein 20A; AltName: Full=HMG box-containing protein 20A	44.3	44.3	34%	7e-05	31%	Q6A278.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Transcription factor A, mitochondrial; Short=TF1A; Flaga; Precursor	43.5	86.6	64%	1e-04	29%	Q61207.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Transcription factor A, mitochondrial; Short=TF1A; Flaga; Precursor	43.1	43.1	64%	1e-04	29%	Q50164.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein 20A; AltName: Full=HMG box-containing protein 20A	43.5	43.5	34%	1e-04	31%	Q52074.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=Transcription factor A, mitochondrial; Short=TF1A; Flaga; Precursor	41.6	41.6	64%	5e-04	29%	Q0187.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein 20A; AltName: Full=HMG box-containing protein 20A; AltName: Full=HMG domain-containing	41.6	41.6	34%	5e-04	30%	Q50333.1	<input type="checkbox"/>	<input type="checkbox"/>
RefName: Full=high mobility group protein 20A; AltName: Full=HMG box-containing protein 20A; AltName: Full=HMG domain-containing	41.2	41.2	34%	6e-04	30%	Q0R166.1	<input type="checkbox"/>	<input type="checkbox"/>

Change cutoffs to show hits "below the line"

NCBI Blast Search Results - high-mobility group box 1 (HMG1)

Run PSI-Blast iteration 3 with max: 500

Sequences producing significant alignments with E-value BETTER than threshold

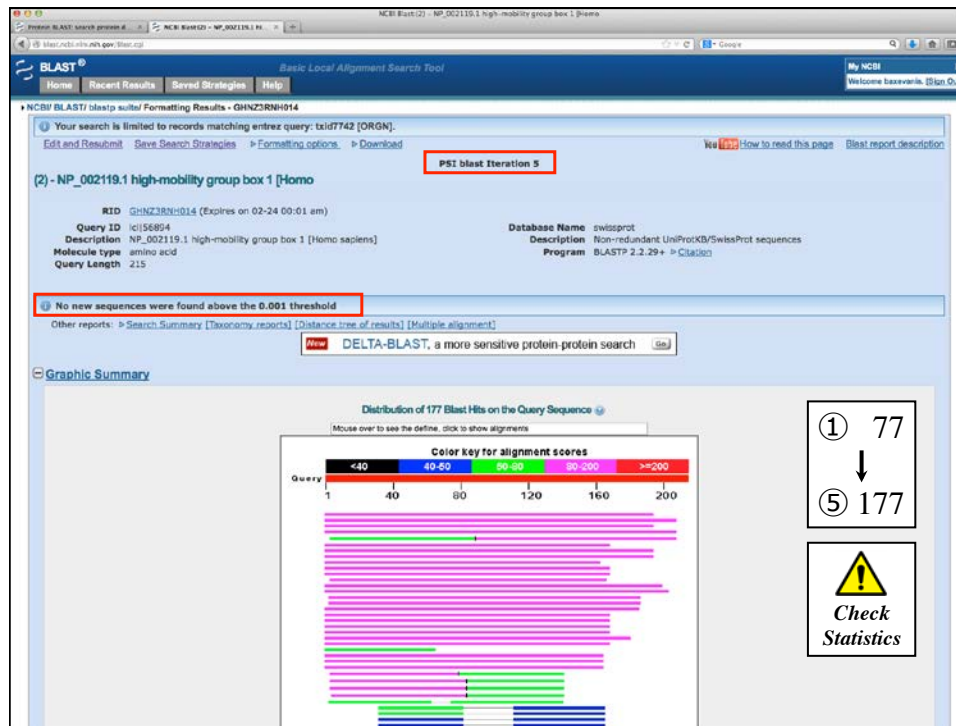
Select: All None Selected0 Yellow: sequences scoring below threshold on previous iteration

Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI	Used to build PDB
ReName: Full=High mobility group protein B1, AName: Full=High mobility group protein 1, Short=HMG-1	242	242	78%	9e-80	99%	P12862.3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein B1, AName: Full=High mobility group protein 1, Short=HMG-1	242	242	78%	9e-80	100%	P10103.3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein B1, AName: Full=Ankolein, AName: Full=Leptin-binding protein p30, AName: Full=High	242	242	78%	9e-80	100%	P93189.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein B1, AName: Full=High mobility group protein 1, Short=HMG-1, sp=QBYYAM,SH=HGB1_CANFA	242	242	78%	9e-80	100%	P39429.3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein B1, AName: Full=High mobility group protein 1, Short=HMG-1	238	238	78%	3e-78	96%	Q0Y105.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Putative high mobility group protein B1-like 1, AName: Full=High mobility group protein B1 pseudogene 1, AName: Full=H	236	236	78%	2e-77	95%	B9Z950.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Putative high mobility group protein 1-like 10, Short=HMG-1.10	230	230	78%	7e-75	95%	Q5U0V6.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group-T protein, Short=HMG-T, AName: Full=HMG-T1, Short=HMG-1	211	211	77%	6e-68	83%	P37745.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein S2, AName: Full=High mobility group protein 2, Short=HMG-2	211	211	78%	1e-67	85%	P28568.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein S2, AName: Full=High mobility group protein 2, Short=HMG-2	206	206	78%	2e-65	86%	P52925.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein S2, AName: Full=High mobility group protein 2, Short=HMG-2, sp=P46873,SH=HMG2_BOVIN1	204	204	78%	4e-65	86%	P28583.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein S2, AName: Full=High mobility group protein 2, Short=HMG-2	203	203	78%	2e-64	86%	P30581.3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Nuclear autoantigen Sp-100, AName: Full=Nuclear dot-associated Sp100 protein, AName: Full=Sp-speckled 100 kDa	202	257	76%	8e-64	81%	Q0H106.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Nuclear autoantigen Sp-100, AName: Full=Nuclear dot-associated Sp100 protein, AName: Full=Sp-speckled 100 kDa	214	273	76%	1e-63	82%	P23497.3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein S3, AName: Full=High mobility group protein 2a, Short=HMG-2a, AName: Full=High mobility	200	200	78%	2e-63	76%	F40818.3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein S2, AName: Full=High mobility group protein 2, Short=HMG-2	200	200	75%	2e-63	87%	P17741.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein S3	198	198	78%	8e-63	76%	Q32131.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein S3, AName: Full=High mobility group protein 2a, Short=HMG-2a, AName: Full=High mobility	198	198	78%	1e-62	76%	O15347.6	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein S3, AName: Full=High mobility group protein 2a, Short=HMG-2a, AName: Full=High mobility	198	198	78%	1e-62	76%	O54795.3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Nuclear autoantigen Sp-100, AName: Full=Nuclear dot-associated Sp100 protein, AName: Full=Sp-speckled 100 kDa	196	249	76%	6e-62	80%	Q0H107.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=High mobility group protein S4	195	196	78%	2e-61	44%	O32134.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Putative high mobility group protein B3-like protein	193	193	78%	8e-61	70%	P30858.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Nuclear autoantigen Sp-100, AName: Full=Nuclear dot-associated Sp100 protein, AName: Full=Sp-speckled 100 kDa	190	244	76%	4e-59	80%	Q0H108.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

NCBI Blast Search Results - NP_002119.1

2... 3... 4...

ReName: Full=TOX high mobility group box family member 4, AName: Full=Epidermal Langerhans cell protein LCP1	82.5	152	60%	2e-17	38%	Q9H42.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=TOX high mobility group box family member 4, AName: Full=Epidermal Langerhans cell protein LCP1	82.5	152	68%	2e-17	38%	Q9H42.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=TOX high mobility group box family member 4	82.5	152	68%	2e-17	38%	Q9H42.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=TOX high mobility group box family member 4 B	81.3	150	68%	4e-17	38%	Q9H40.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=TOX high mobility group box family member 4	81.3	150	68%	4e-17	38%	A62NP5.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=TOX high mobility group box family member 4 A	81.3	150	68%	4e-17	38%	Q9D40.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=FACT complex subunit SSRP1, AName: Full=Facilitates chromatin transcription complex subunit SSRP1, AName: Full=R	109	203	64%	1e-25	37%	Q0578.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=FACT complex subunit SSRP1, AName: Full=DNA unwinding factor 87 kDa subunit, Short=DURF7, AName: Full=Facilit	106	203	64%	1e-25	36%	Q09022.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=FACT complex subunit SSRP1, AName: Full=Chromatin-specific transcription elongation factor 87 kDa subunit, AName: I	114	206	63%	4e-28	36%	Q28945.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	58.2	115	61%	5e-10	35%	Q09F11.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.8	113	61%	9e-10	35%	Q95F12.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.8	114	62%	9e-10	35%	Q7JGF7.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.8	114	62%	9e-10	35%	Q85F02.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.8	114	62%	9e-10	35%	Q03256.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.5	114	61%	1e-09	35%	Q03256.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	56.3	111	62%	4e-09	35%	Q0X37.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=SRY-related protein AMS2	39.3	39.3	22%	3e-04	35%	P42642.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=SRY-related protein AMS3	39.3	39.3	22%	4e-04	35%	P42643.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=SRY-related protein CH1	39.0	39.0	22%	5e-04	35%	P42665.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=FACT complex subunit SSRP1, AName: Full=Facilitates chromatin transcription complex subunit SSRP1, AName: Full=R	112	204	64%	1e-27	35%	Q04912.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=FACT complex subunit SSRP1, AName: Full=Facilitates chromatin transcription complex subunit SSRP1, AName: Full=R	112	211	64%	9e-28	34%	Q08452.2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=SRY-related protein CH3	40.1	40.1	23%	2e-04	34%	P42667.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=SRY-related protein CH1	38.6	38.6	23%	6e-04	34%	P42670.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Protein polyoma-1	77.9	141	60%	1e-15	34%	Q00941.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=MMS1 protein homolog 1, AName: Full=DNA mismatch repair protein PMS1	60.5	106	59%	4e-10	34%	P54272.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.8	114	62%	9e-10	33%	Q87DX7.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.1	111	65%	1e-09	33%	Q86421.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.1	111	65%	1e-09	33%	Q86477.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.1	111	65%	1e-09	33%	Q86478.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.1	111	65%	1e-09	33%	Q86479.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.1	111	65%	1e-09	33%	Q86480.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
ReName: Full=Sex-determining region Y protein, AName: Full=Testis-determining factor	57.1	111	65%	1e-09	33%	Q86481.1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>



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

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

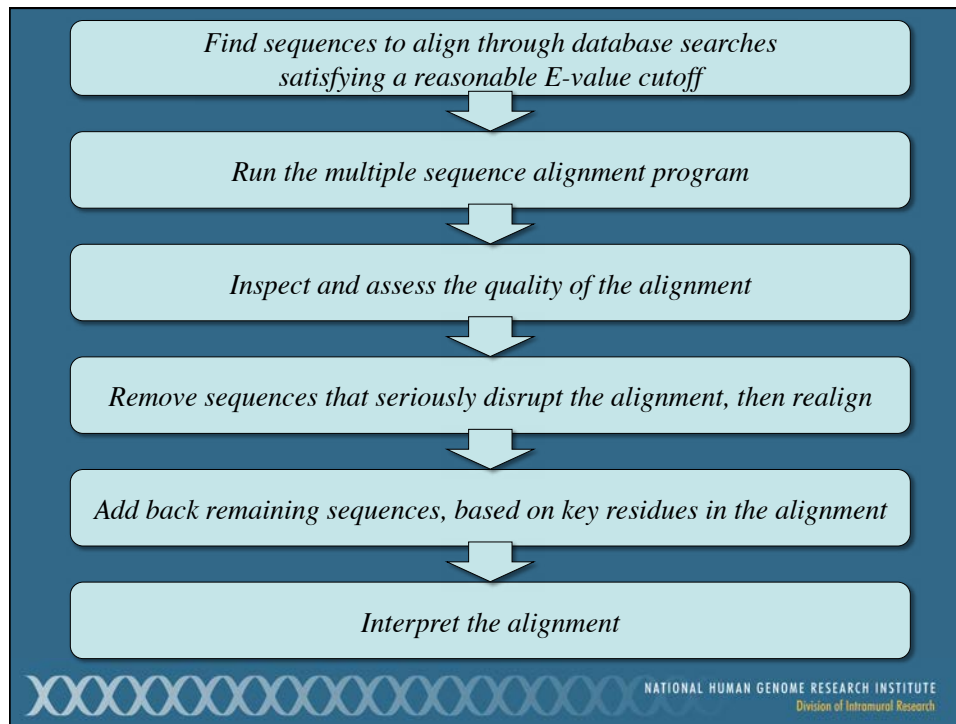
- Absolute sequence similarity
Create the alignment by lining up as many common characters as possible
- Conservation
Take into account residues that can substitute for one another and not adversely affect the function of the protein
- Structural similarity
Knowledge of the secondary or tertiary structure of the proteins being aligned can be used to fine-tune the alignment



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





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

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)



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” (*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



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



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

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



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



Progressive Alignment

```
>sequence A
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST
>sequence B
VQLSGE EKAAVLALWDKVN EEVGGEALGRLLVVYPWTQRFFDSFGDSL N
>sequence C
VLSPADKTNVKA AWGKVG AHAGEYGA EALERMFLSFPTTKTYFPHFDLSH
>sequence D
VLSAADKTNVKA AWSKVG GHAGEYGA EALERMFLGFPTTKTYFPHFDLSH
```



Progressive Alignment

1. Calculate a similarity score (percent identity) between every pair of sequences to drive the alignment

For N sequences, this requires the calculation of $[N \times (N - 1)] / 2$ pairwise alignments

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



Progressive Alignment

```
>sequence A
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST
>sequence B
VQLSGEEKAAVLALWDKVNEEVGGEALGRLLVVYPWTQRFFDSFGDSLN
>sequence C
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFSLFPTTKTYFPHFDLSH
>sequence D
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSH
```

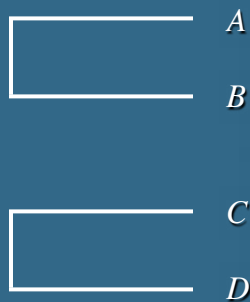
%ID	A	B	C	D
A	100			
B	80	100		
C	44	40	100	
D	40	40	92	100



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



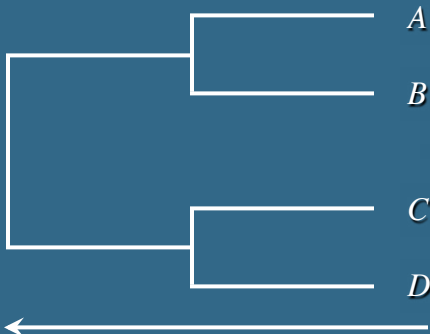
Progressive Alignment

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



Progressive Alignment

- Align “sequence” AB with “sequence” CD
- Continue following the branching order of the tree, from the tips to the root, merging each new pair of “sequences”



Progressive Alignment: Advantages

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



Progressive Alignment: Disadvantages

- If initial alignments are made on distantly related sequences, there may be errors in the initial alignments
- Once an alignment is “fixed”, it is not reconsidered, so any errors in the early alignments may propagate through subsequent alignments
- 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)



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



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



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)



<https://www.ebi.ac.uk/Tools/msa>

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:

```
>FO5B_MOUSE Protein fosB
MFQAFPGDYDSGRCSSSPSAESCYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVPQTLISSMAQSGQPLASQPPAVDYPDMPGTSYTPGLSAYSTGASGS
GGPSTSTITSGPVSARPARARPRRPHREELTPEEEKRRVRRERKLAALKCHRRRREL
DRLQAEITDQLEEKAELESEAELOKEKRELFVLAHKPGCKPYEEGPGPLAEVRD
LPGSTSAEKEDFGWLLPPPPPPPLPFQSSRDAPPLTASLFTHSEVGLGDPPFVSPSY
TSSPVLTCPEVSFAFAGADRTSQSEQPSDPLNSPLLAL
```

Or, upload a file: [Browse...](#) No file selected.

STEP 2 - Set your parameters

OUTPUT FORMAT [Clustal w/o numbers](#)

The default settings will fulfil 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.

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:

```
>FO5B_MOUSE Protein fosB
MFQAFPGDYDSGRCSSSPSAESCYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVPQTLISSMAQSGQPLASQPPAVDYPDMPGTSYTPGLSAYSTGASGS
GGPSTSTITSGPVSARPARARPRRPHREELTPEEEKRRVRRERKLAALKCHRRRREL
DRLQAEITDQLEEKAELESEAELOKEKRELFVLAHKPGCKPYEEGPGPLAEVRD
LPGSTSAEKEDFGWLLPPPPPPPLPFQSSRDAPPLTASLFTHSEVGLGDPPFVSPSY
TSSPVLTCPEVSFAFAGADRTSQSEQPSDPLNSPLLAL
```

Or, upload a file: [Browse...](#) No file selected.

STEP 2 - Set your parameters

OUTPUT FORMAT [Clustal w/o numbers](#)

DEALIGN INPUT SEQUENCES	MBED-LIKE CLUSTERING GUIDE-TREE	MBED-LIKE CLUSTERING ITERATION	NUMBER OF COMBINED ITERATIONS
no	yes	yes	default(0)
MAX GUIDE TREE ITERATIONS	MAX HMM ITERATIONS	ORDER	
default	default	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](#).

The screenshot shows the Clustal Omega web interface for a multiple sequence alignment. The top navigation bar includes 'Alignments', 'Result Summary', 'Phylogenetic Tree', and 'Submission Details'. Below the navigation, there are options to 'Download Alignment File', 'Hide Colors', and 'Send to ClustalW2_Phylogeny'. The main area displays the CLUSTAL O(1.2.0) multiple sequence alignment for the protein FOSB. The alignment is color-coded by residue properties: RED for small hydrophobic residues, BLUE for acidic residues, MAGENTA for basic residues, GREEN for hydroxyl and sulfhydryl groups, and GREY for unusual amino acids. A legend on the right side of the interface explains these color codes. The alignment shows conserved regions across the five species, with gaps indicated by dashes.

The screenshot shows the Clustal Omega web interface for a phylogenetic tree. The top navigation bar includes 'Alignments', 'Result Summary', 'Phylogenetic Tree', and 'Submission Details'. Below the navigation, there are options to 'Download Phylogenetic Tree File'. The main area displays the phylogenetic tree results for the job clustalo-l20140301-151344-0178-43085370-pg. The tree is a Neighbour-joining tree without distance corrections. The tree shows the relationships between the five species: FOSB_MOUSE (0.01854), FOSB_HUMAN (0.02288), FOSB_CHICK (0.1107), FOS_RAT (0.01948), and FOS_MOUSE (0.0121). A phylogram is also shown below the tree, with branch lengths indicated. The bottom of the page features a footer with navigation links for EMBL-EBI, Services, Research, Training, Industry, and About us.

Phylogenetic Tree < Clustal Omega < EMBL-EBI

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)
10.13561,
FOS_CHICK:0.11070)
10.11115,
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

- 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

Result Summary < Clustal Omega < EMBL-EBI

EMBL-EBI

Services Research Training About us

Clustal Omega

Input form Web services Help & Documentation

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

JaView

start jaView

EMBL-EBI

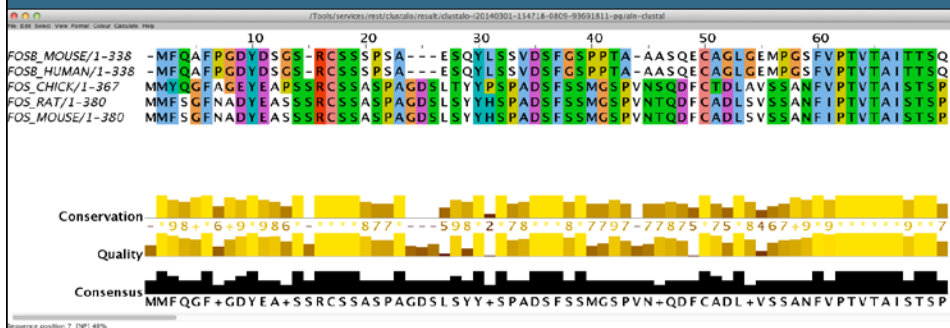
- Services
 - By topic
 - By name (A-Z)
 - Help & Support
- Research
 - Overview
 - Publications
 - Research groups
 - Postdocs & PhDs
- Training
 - Overview
 - Train at EBI
 - Train outside EBI
 - Train online
 - Contact organisers
- Industry
 - Overview
 - Members Area
 - Workshops
 - SME Forum
 - 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



Default view

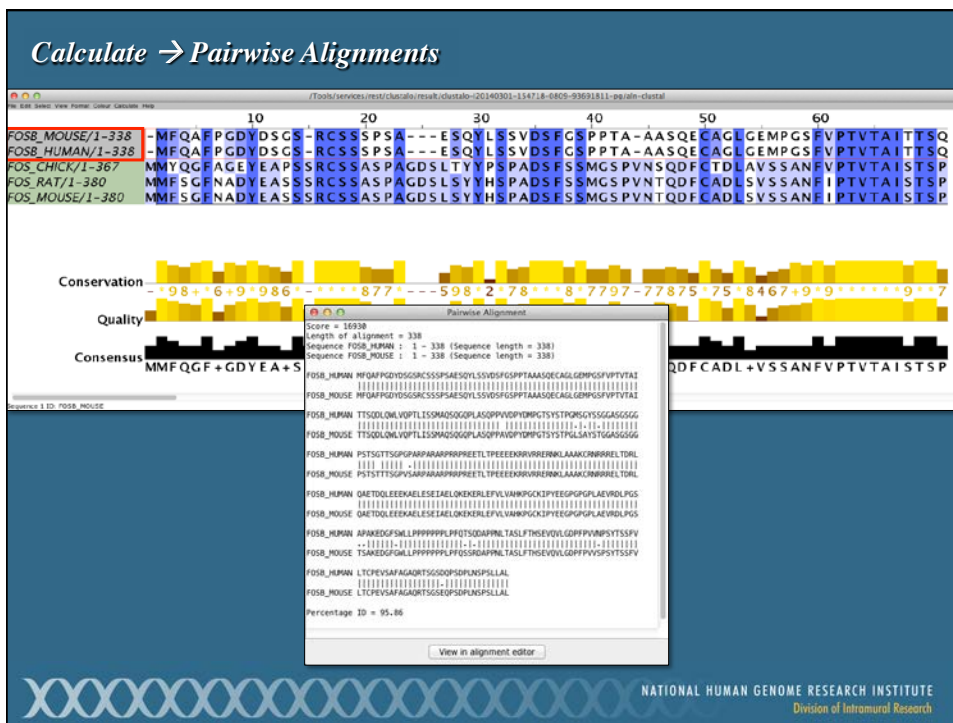
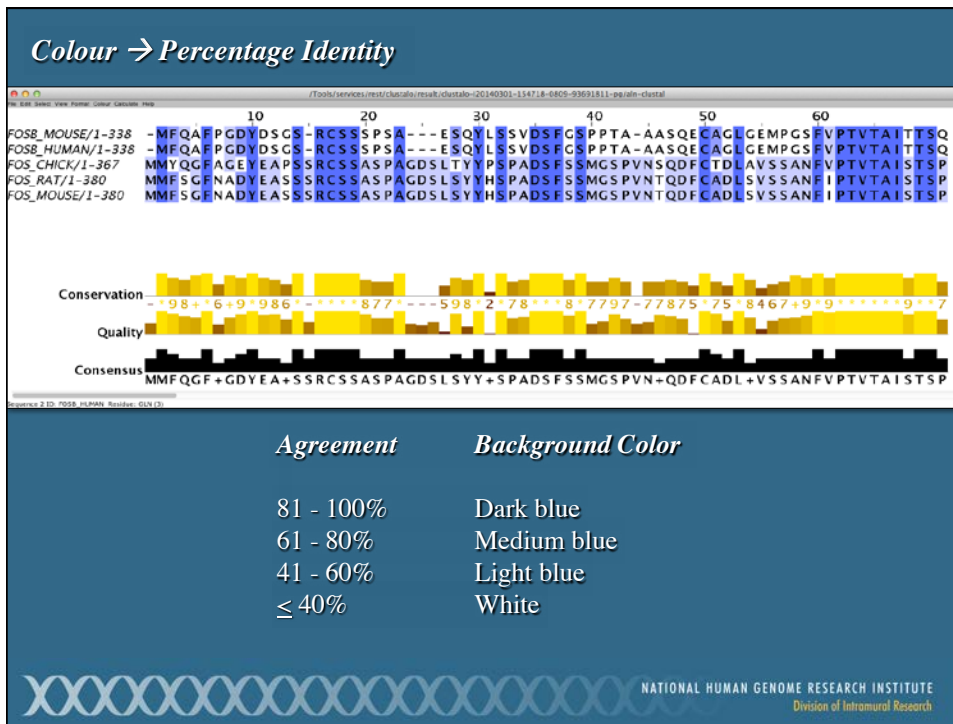


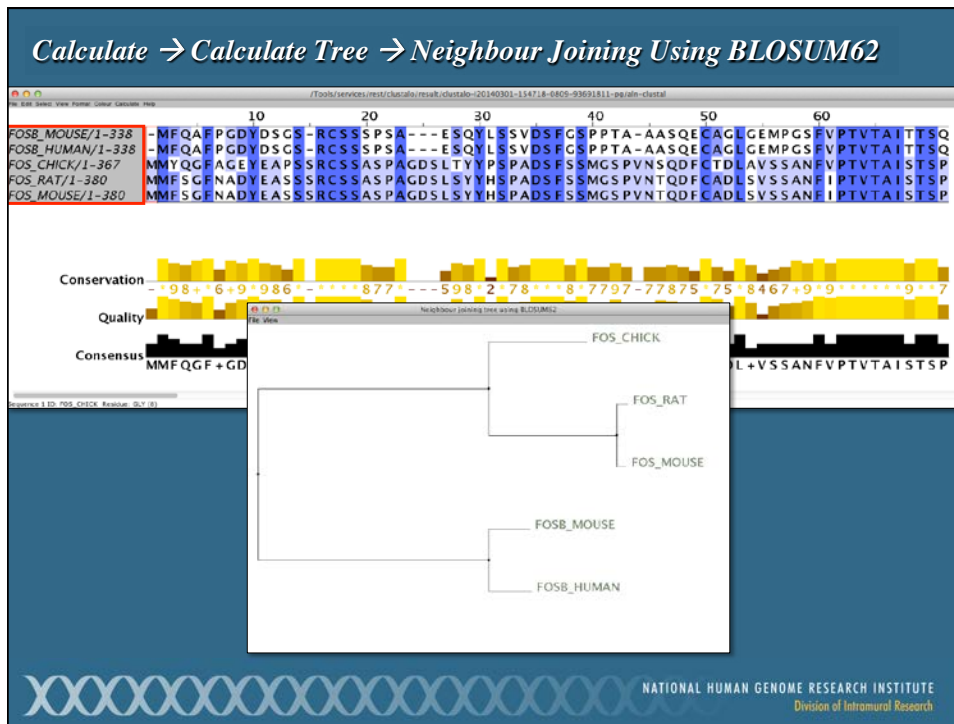
Conservation Conservation of total alignment (indication of percent identity)

Quality Alignment quality, based on BLOSUM scores

Consensus Based on percent identity







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)

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*
 Independent Consultant, Philadelphia, Pennsylvania, United States of America

Abstract: Online learning initiatives over the past decade have become increasingly commonplace in their delivery, content, and participation in their provision. Coupled with the recent announcement of a number of initiatives and their activities that promise to make a university education not just a necessity but a choice, a new possibility. An this pivotal moment, it is appropriate to explore the potential for obtaining comprehensive bioinformatics training with currently existing free online resources. This article presents such a bioinformatics curriculum in the form of a structured catalog, together with associated commentary and an assessment of emerging challenges and likely future directions for open online learning in this field.

Online Learning Comes of Age

Online academic "resources" at the secondary level has now been available to the public for a decade. An online content effort, being organized in 2012 with the Massachusetts Institute of Technology (MIT) and their OpenCourseWare Initiative (<http://ocw.mit.edu>). This project offered up the slides, lecture notes, videos, exams, and/or other study materials for a vast range of courses, at the discretion of professors but with strong support and encouragement from MIT's administration. MIT is a variety of ways with video of lecture points.

Tom Nelson, MIT, The University of California, Berkeley, had several webcasting systems, and eventually began posting both audio and video for public consumption at their Berkeley Webcast site (<http://www.berkeley.edu/webcast>). A number of other universities followed suit, though without as much success. Some were facilitated with the Canvas learning center (<http://Canvas.com>), and others with the Blackboard Learning Environment (<http://www.blackboard.com>). In many cases, individual faculty members took the initiative to post course materials, including slides, to widely viewing formats. Some adopted the use of "Reclaim the online" or similar tools to make the user participatory by the Kahn Academy, which resulted in a wide YouTube archive, and the same success was with Blackboard (<http://www.blackboard.com>).

You also noticed because the distribution of some academic videos, which are now aggregated by institutions under YouTube (<http://www.youtube.com/watch?v=...>), Apple has also put its distribution effort on online learning with iTunes U (<http://www.apple.com/education/itunesu>), also regulated by Blackboard, but with integrated search capability and of course deployment to iPad and iPhone apps. Considerable attention also available collections of video courses, but generally with little value added.

MIT University began in 2007 to release Open Yale Courses (<http://open.yale.edu>) in a more formal and content format than most other efforts, including high-quality video and extensive lecture notes, organized incrementally with just under 50 available to date. Then, in 2013, MIT expanded several of its other courses into a much more structured format called FutureLearn, with training modules in multiple languages, video integrated with self-assessment and other activities. For a somewhat different view, the non-profit FutureLearn emerged as a complementary online capability curriculum computing course that are essentially teaching of video and text resources from many different sources, including a number of

other distributed sites (<http://www.futurelearn.com>). In the fall of 2013, a highly published online course, "Introduction to Artificial Intelligence" (AI), was conducted by Turing University and National Open University's Director of Research, Prof. Tommi, based on the Stanford AI course. It was "free" in the sense that new videos were released and homework assignments collected on a weekly basis, and updates and reissues were given to its users, while discussion help allowed for some degree of interaction. The course attracted 100,000 students from 100 countries, 2,000 of which studied successfully and were granted "certificates of completion" (1). Shortly afterwards, MIT set up a similar agreement in a new platform called MITx, offering a course in thermodynamics that attracted comparable numbers of students (<http://MITx.mit.edu>).

The trend to structured presentation and high production quality class activities, and seek an integrated approach. The AI course was effectively spun off by Prof. Tommi into a Web course called iCOURSE (<http://www.icourse.com>), which is currently live with its content. In April of 2013, MIT added Stanford courses, Prof. Andrew Ng and Daphne Koller, announced a similar course called Coursera (<http://www.coursera.com>), with backing from major Silicon Valley venture capital firms. Coursera, also now live, is being studied with interest from academic partners, Stanford, Princeton, University of the University of Pennsylvania, and the University of Chicago. This has been recently supported with a number of other course releases announced. And in May of 2013, heavily on occasion after MIT had rolled out its

Chapters from the 2013 An Online Bioinformatics Curriculum, MIT Center for the Study of Computational Biology

Author: David B. Searls, Independent Consultant, United States of America

Published: September 18, 2013

Copyright: © 2013 David B. Searls. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The author received no specific funding for this article.

Competing Interests: The author has declared that no competing interests exist.

* David B. Searls is an Associate Editor of PLOS Computational Biology.

PLOS Computational Biology | www.ploscompbiol.org | September 2013 | Volume 8 | Issue 9 | e1002632

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



A promotional banner for the NIH Intramural Research Program. It features a light blue background with a white wave-like pattern. In the center, there are five small, tilted photographs of researchers in various lab settings. Below the photos, on the left, is the NIH logo (a stylized 'NIH' in a purple and grey box) followed by the text "Intramural Research Program" and the tagline "Our Research Changes Lives". On the right, the slogan "one program many people infinite possibilities" is written in green, with the website "irp.nih.gov" in blue below it.