Current Topics in Genome Analysis
Spring 2005

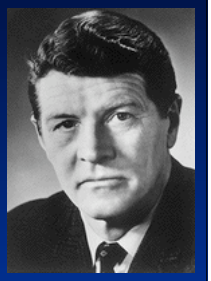
Week 5
Biological Sequence Analysis II

Andy Baxevanis, Ph.D.

Overview

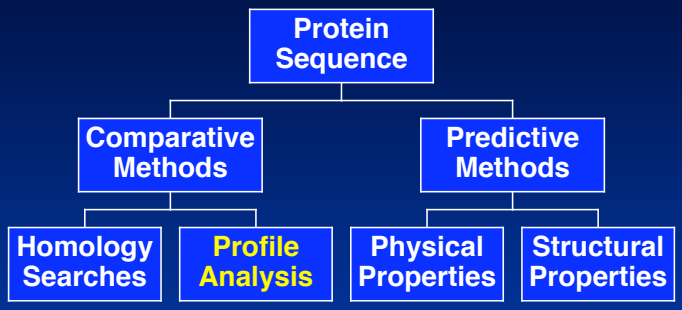
- Week 4: Comparative methods and concepts
  - Similarity *vs.* Homology
  - Global *vs.* Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 5: Predictive methods and concepts
  - Profiles, patterns, motifs, and domains
  - Secondary structure prediction
  - Structures: VAST, Cn3D, and *de novo* prediction

Protein Conformation



- Christian Anfinsen
  Studies on reversible denaturation →
  "Sequence specifies conformation"
- Chaperones and disulfide interchange enzymes:
  involved but not controlling final state
- "Starting with a newly-determined sequence, what can be determined computationally about its possible function and structure?"

Protein Sequence Analysis



- *Common structure?*
- *Common function?*
- *Evolutionary relationship?*
- *Global or local similarity?*

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" — *ProfileScan*, *CDD*

    or "many-against-one" — *PSI-BLAST*

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
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	10	0	12	0	0	12	13	0	0	3	3	1		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	10	11	0	6	16	11		89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	50	0	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

Patterns

Phe *or* Tyr → [FY]

Cys → C

any amino acid → x

any two amino acids → x(2)

Val *or* Ala → {VA}

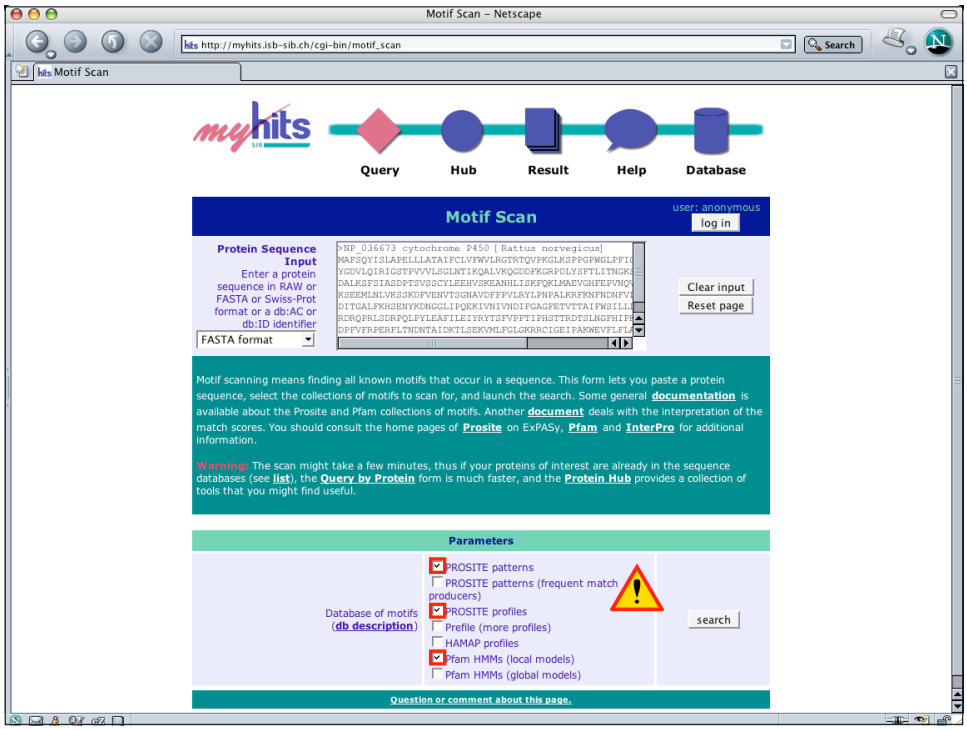
any amino acid → x

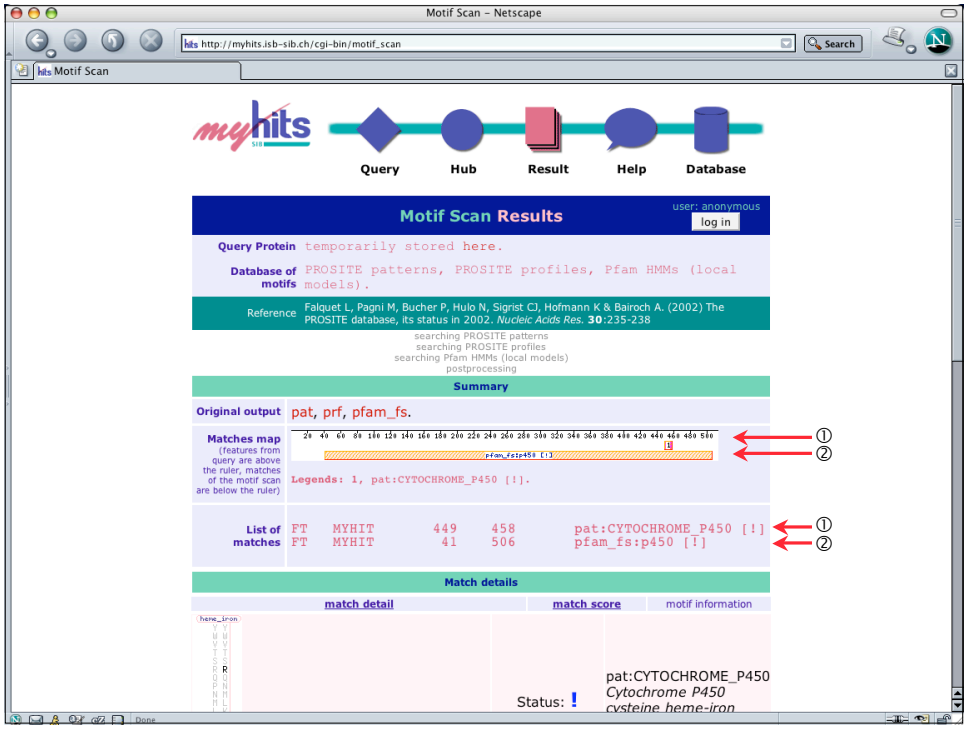
three His → H(3)

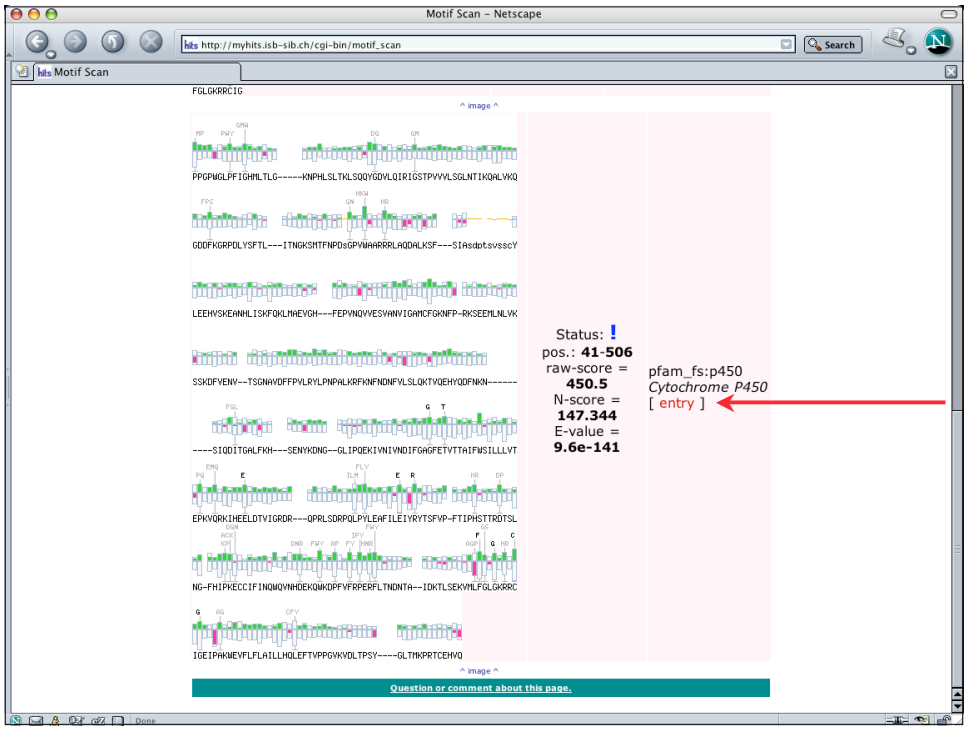
[FY]-x-C-x(2)-{VA}-x-H(3)

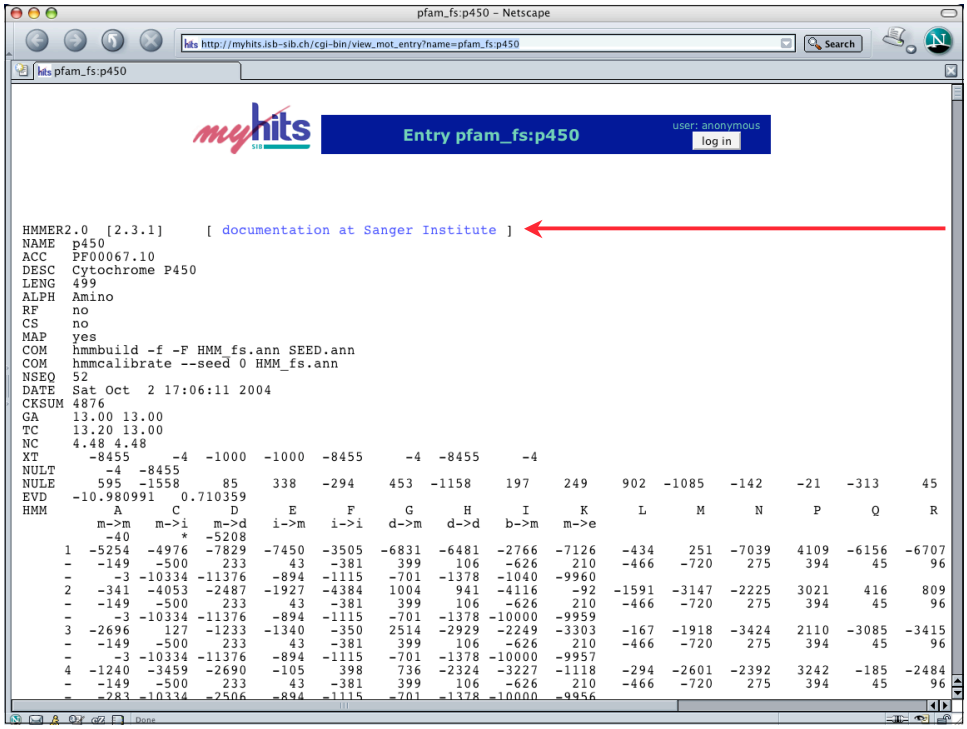
ProfileScan

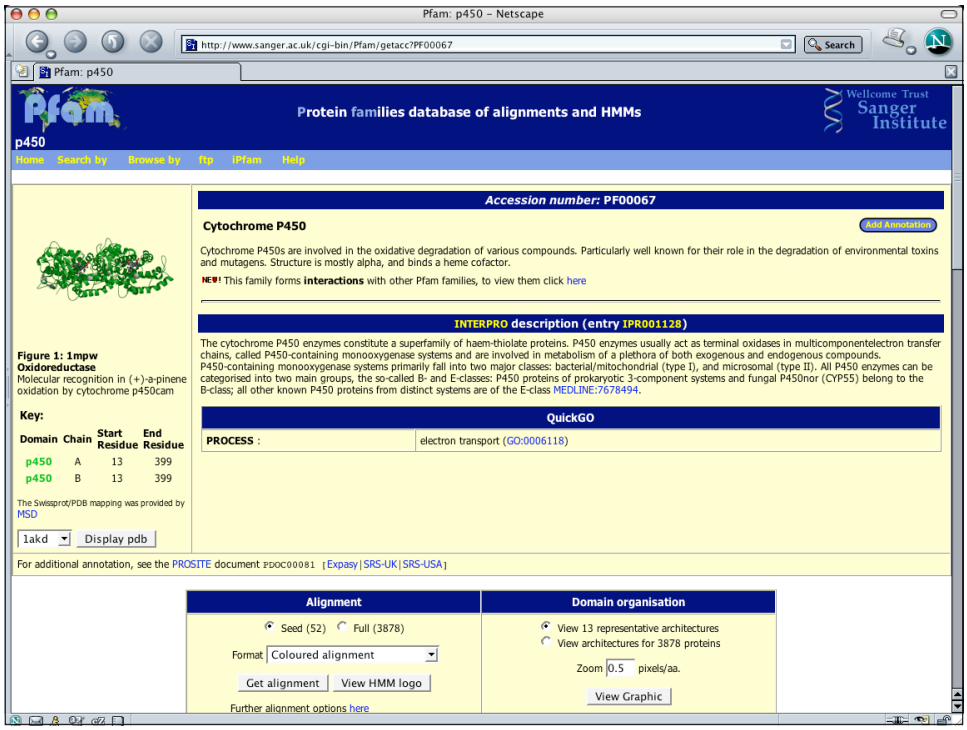
- Search sequence against a collection of profiles and patterns
- Databases available
  - PROSITE profiles
  - PROSITE patterns
  - PfamA
  - PfamB
  - InterPro families
  - HAMAP profiles (microbial)
  - TIGRfam protein families
- *http://hits.isb-sib.ch/cgi-bin/PFSCAN*

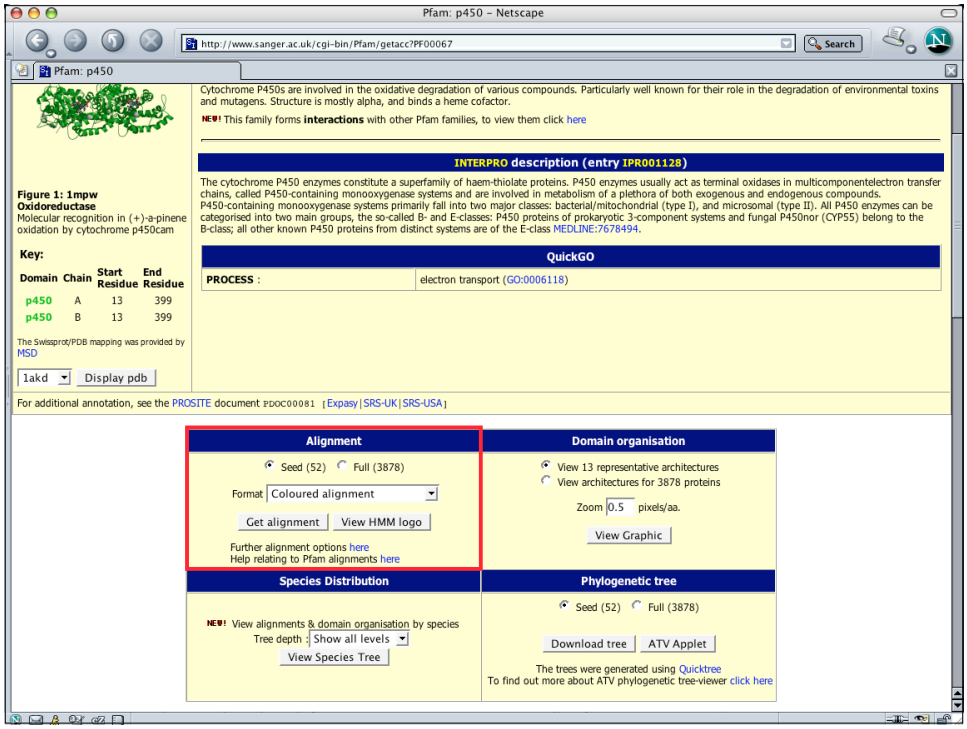


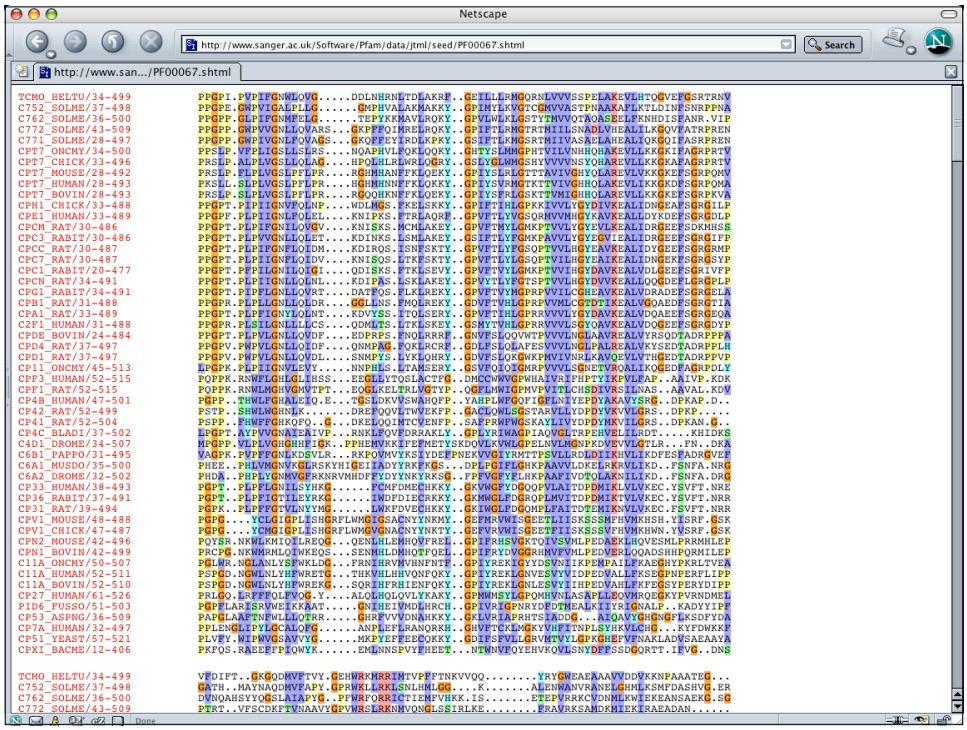

Motif Scan - Netscape
http://myhits.isb-sib.ch/cgi-bin/motif_scan
Motif Scan
Query
Hub
Result
Help
Database
Motif Scan Results
user: anonymous
log in
Query Protein temporarily stored here.
Database of motifs PROSITE patterns, PROSITE profiles, Pfam HMMs (local models).
Reference Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K & Bairoch A. (2002) The PROSITE database, its status in 2002. Nucleic Acids Res. 30:235-238
searching PROSITE patterns
searching PROSITE profiles
searching Pfam HMMs (local models)
postprocessing
Summary
Original output pat, prf, pfam_fs.
Matches map (features from query are above the ruler, matches of the motif scan are below the ruler)
Legends: 1, pat:CYTOCHROME_P450 [!].
List of matches
FT MYHIT 449 458 pat:CYTOCHROME_P450 [!]
FT MYHIT 41 506 pfam_fs:p450 [!]
Match details
match detail
match score
motif information
Status: !
pat:CYTOCHROME_P450
Cytochrome P450
cysteine heme-iron

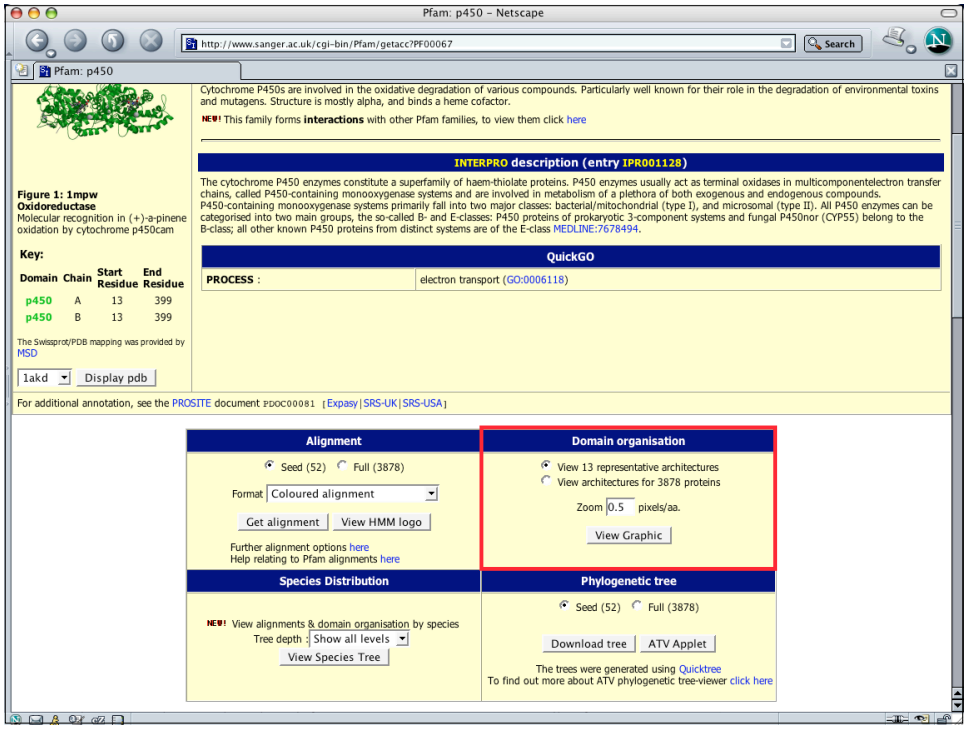

Motif Scan - Netscape
http://myhits.isb-sib.ch/cgi-bin/motif_scan
Motif Scan
Status: !
pos.: 41-506
raw-score = 450.5
N-score = 147.344
E-value = 9.6e-141
pfam_fs:p450
Cytochrome P450
[entry]
Question or comment about this page.

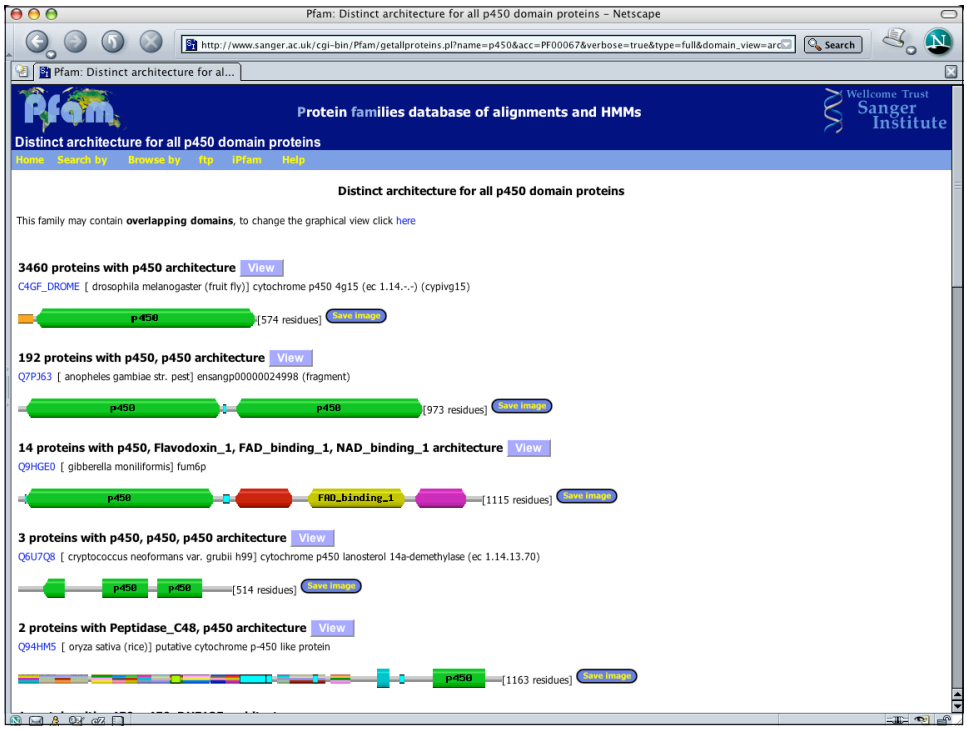

pfam_fs:p450 - Netscape
http://myhits.isb-sib.ch/cgi-bin/view_mot_entry?name=pfam_fs:p450
Entry pfam_fs:p450
user: anonymous
log in
HMMER2.0 [2.3.1] [documentation at Sanger Institute]
NAME p450
ACC PF00067.10
DESC Cytochrome P450
LENG 499
ALPH Amino
RF no
CS no
MAP yes
COM hmmbuild -f -F HMM_fs.ann SEED.ann
COM hmmcalibrate --seed 0 HMM_fs.ann
NSEQ 52
DATE Sat Oct 2 17:06:11 2004
CKSUM 4876
GA 13.00 13.00
TC 13.20 13.00
NC 4.48 4.48
XT -8455 -4 -1000 -1000 -8455 -4 -8455 -4
NULT -4 -8455
NULE 595 -1558 85 338 -294 453 -1158 197 249 902 -1085 -142 -21 -313 45
EVD -10.980991 0.710359
HMM A C D E F G H I K L M N P Q R
m->m m->i m->d i->m i->i d->m d->d b->m m->e
-40 * -5208
1 -5254 -4976 -7829 -7450 -3505 -6831 -6481 -2766 -7126 -434 251 -7039 4109 -6156 -6707
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96
- -3 -10334 -11376 -894 -1115 -701 -1378 -1040 -9960
2 -341 -4053 -2487 -1927 -4384 1004 941 -4116 -92 -1591 -3147 -2225 3021 416 809
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96
- -3 -10334 -11376 -894 -1115 -701 -1378 -10000 -9959
3 -2696 127 -1233 -1340 -350 2514 -2929 -2249 -3303 -167 -1918 -3424 2110 -3085 -3415
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96
- -3 -10334 -11376 -894 -1115 -701 -1378 -10000 -9957
4 -1240 -3459 -2690 -105 398 736 -2324 -3227 -1118 -294 -2601 -2392 3242 -185 -2484
- -149 -500 233 43 -381 399 106 -626 210 -466 -720 275 394 45 96


Pfam: p450 - Netscape
http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067
Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute
p450
Home Search by Browse by ftp iPfam Help
Accession number: PF00067
Cytochrome P450
Add Annotation
Cytochrome P450s are involved in the oxidative degradation of various compounds. Particularly well known for their role in the degradation of environmental toxins and mutagens. Structure is mostly alpha, and binds a heme cofactor.
NEW! This family forms interactions with other Pfam families, to view them click here
INTERPRO description (entry IPR001128)
The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponentelectron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds.
P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class MEDLINE:7678494.
QuickGO
PROCESS : electron transport (GO:0006118)
Figure 1: 1mpw
Oxidoreductase
Molecular recognition in (+)-a-pinene oxidation by cytochrome p450cam
Key:
Domain Chain Start Residue End Residue
p450 A 13 399
p450 B 13 399
The Swissprot/PDB mapping was provided by MSD
1akd
Display pdb
For additional annotation, see the PROSITE document PDOC00081 [Expasy | SRS-UK | SRS-USA]
Alignment
Seed (52) Full (3878)
Format Coloured alignment
Get alignment
View HMM logo
Further alignment options here
Domain organisation
View 13 representative architectures
View architectures for 3878 proteins
Zoom 0.5 pixels/aa.
View Graphic


Pfam: p450 - Netscape
http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067
Pfam: p450
Cytochrome P450s are involved in the oxidative degradation of various compounds. Particularly well known for their role in the degradation of environmental toxins and mutagens. Structure is mostly alpha, and binds a heme cofactor.
NEW! This family forms interactions with other Pfam families, to view them click here
INTERPRO description (entry IPR001128)
The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponentelectron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds.
P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class MEDLINE:7678494.
QuickGO
PROCESS : electron transport (GO:0006118)
Figure 1: 1mpw Oxidoreductase
Molecular recognition in (+)-a-pinene oxidation by cytochrome p450cam
Key:
Domain Chain Start Residue End Residue
p450 A 13 399
p450 B 13 399
The Swissprot/PDB mapping was provided by MSD
1akd
Display pdb
For additional annotation, see the PROSITE document PDOC00081 [Expasy | SRS-UK | SRS-USA]
Alignment
Seed (52) Full (3878)
Format Coloured alignment
Get alignment
View HMM logo
Further alignment options here
Help relating to Pfam alignments here
Domain organisation
View 13 representative architectures
View architectures for 3878 proteins
Zoom 0.5 pixels/aa.
View Graphic
Species Distribution
NEW! View alignments & domain organisation by species
Tree depth : Show all levels
View Species Tree
Phylogenetic tree
Seed (52) Full (3878)
Download tree
ATV Applet
The trees were generated using Quicktree
To find out more about ATV phylogenetic tree-viewer click here


Netscape
http://www.sanger.ac.uk/Software/Pfam/data/jtml/seed/PF00067.shtml
http://www.san.../PF00067.shtml


Pfam: p450 – Netscape
http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067
Pfam: p450
Cytochrome P450s are involved in the oxidative degradation of various compounds. Particularly well known for their role in the degradation of environmental toxins and mutagens. Structure is mostly alpha, and binds a heme cofactor.
NEW! This family forms interactions with other Pfam families, to view them click here
INTERPRO description (entry IPR001128)
The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponentelectron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds.
P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class MEDLINE:7678494.
QuickGO
PROCESS : electron transport (GO:0006118)
Figure 1: 1mpw Oxidoreductase
Molecular recognition in (+)-a-pinene oxidation by cytochrome p450cam
Key:
Domain Chain Start Residue End Residue
p450 A 13 399
p450 B 13 399
The Swissprot/PDB mapping was provided by MSD
1akd
Display pdb
For additional annotation, see the PROSITE document PDOC00081 [Expasy|SRS-UK|SRS-USA]
Alignment
Seed (52) Full (3878)
Format Coloured alignment
Get alignment
View HMM logo
Further alignment options here
Help relating to Pfam alignments here
Domain organisation
View 13 representative architectures
View architectures for 3878 proteins
Zoom 0.5 pixels/aa.
View Graphic
Species Distribution
NEW! View alignments & domain organisation by species
Tree depth : Show all levels
View Species Tree
Phylogenetic tree
Seed (52) Full (3878)
Download tree
ATV Applet
The trees were generated using Quicktree
To find out more about ATV phylogenetic tree-viewer click here


Pfam: Distinct architecture for all p450 domain proteins – Netscape
http://www.sanger.ac.uk/cgi-bin/Pfam/getallproteins.pl?name=p450&acc=PF00067&verbose=true&type=full&domain_view=arc
Pfam: Distinct architecture for al...
Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute
Distinct architecture for all p450 domain proteins
Home Search by Browse by ftp iPfam Help
Distinct architecture for all p450 domain proteins
This family may contain overlapping domains, to change the graphical view click here
3460 proteins with p450 architecture View
C4GF_DROME [drosophila melanogaster (fruit fly)] cytochrome p450 4g15 (ec 1.14.-.-) (cypivg15)
p450 [574 residues] Save image
192 proteins with p450, p450 architecture View
Q7PJ63 [anopheles gambiae str. pest] ensangp00000024998 (fragment)
p450 p450 [973 residues] Save image
14 proteins with p450, Flavodoxin_1, FAD_binding_1, NAD_binding_1 architecture View
Q9HGE0 [gibberella moniliformis] fum6p
p450 FAD_binding_1 [1115 residues] Save image
3 proteins with p450, p450, p450 architecture View
Q6U7Q8 [cryptococcus neoformans var. grubii h99] cytochrome p450 lanosterol 14a-demethylase (ec 1.14.13.70)
p450 p450 [514 residues] Save image
2 proteins with Peptidase_C48, p450 architecture View
Q94HM5 [oryza sativa (rice)] putative cytochrome p-450 like protein
p450 [1163 residues] Save image

Pfam: p450 – Netscape

http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067

Protein families database of alignments and HMMs

Wellcome Trust Sanger Institute

p450

Home Search by Browse by ftp iPfam Help

Accession number: PF00067

Cytochrome P450

Add Annotation

Cytochrome P450s are involved in the oxidative degradation of various compounds. Particularly well known for their role in the degradation of environmental toxins and mutagens. Structure is mostly alpha, and binds a heme cofactor.

NEW! This family forms **interactions** with other Pfam families, to view them click here

INTERPRO description (entry IPR001128)

The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponentelectron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds. P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class MEDLINE:7678494.

Figure 1: 1mpw Oxidoreductase
Molecular recognition in (+)-a-pinene oxidation by cytochrome p450cam

Key:

Domain	Chain	Start Residue	End Residue
p450	A	13	399
p450	B	13	399

The Swissprot/PDB mapping was provided by MSD

1akd Display pdb

QuickGO	
PROCESS :	electron transport (GO:0006118)

For additional annotation, see the PROSITE document PDOC00081 [Expasy | SRS-UK | SRS-USA]

Alignment	Domain organisation
Seed (52) Full (3878)	View 13 representative architectures
Format Coloured alignment	View architectures for 3878 proteins
Get alignment View HMM logo	Zoom 0.5 pixels/aa.
Further alignment options here	View Graphic

InterPro: Cytochrome P450 – Netscape

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

EMBL-EBI European Bioinformatics Institute

EBI Home | About EBI | Research | Services | Toolbox | Databases | Downloads | Submissions

InterPro home | Text Search | Sequence Search | Databases | Documentation | FTP site | Protein of the month

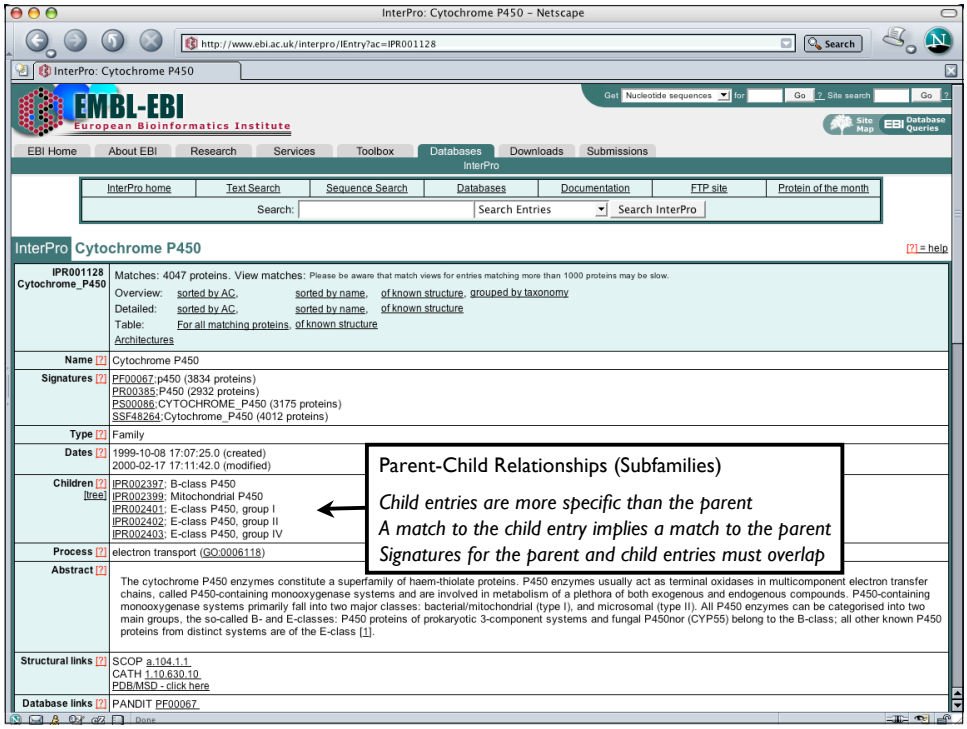
Search: Search Entries Search InterPro

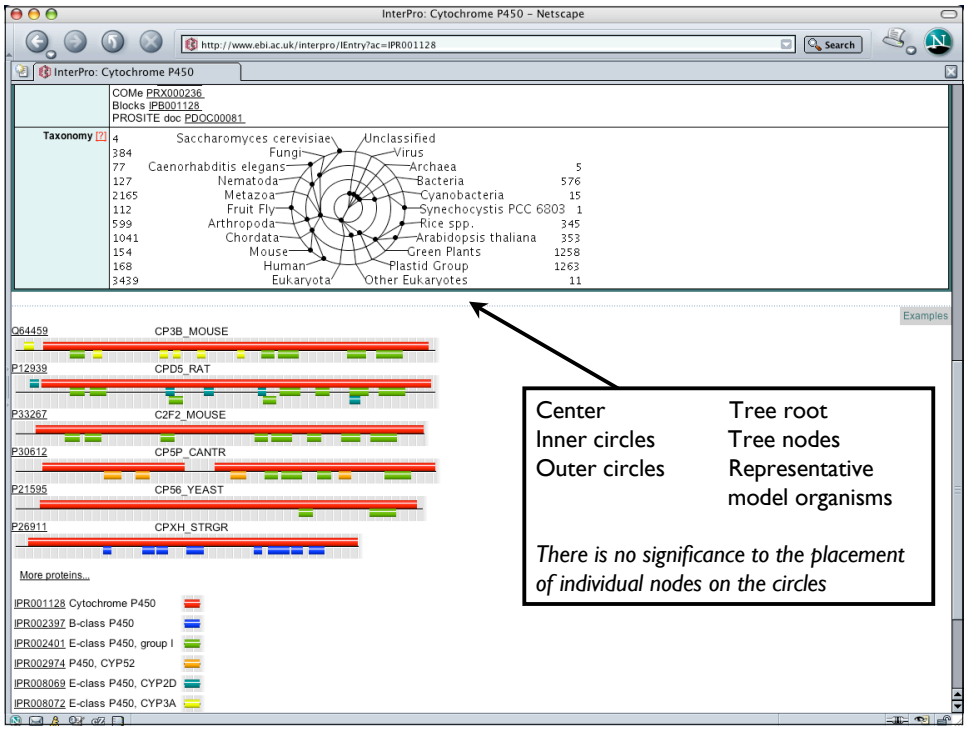
InterPro Cytochrome P450

[?] = help

IPR001128 Cytochrome_P450	Matches: 4047 proteins. View matches: Please be aware that match views for entries matching more than 1000 proteins may be slow. Overview: sorted by AC, sorted by name, of known structure, grouped by taxonomy. Detailed: sorted by AC, sorted by name, of known structure. Table: For all matching proteins, of known structure. Architectures
Name [?]	Cytochrome P450
Signatures [?]	PF00067:p450 (3834 proteins) PR00385:P450 (2932 proteins) PS00086:CYTOCHROME_P450 (3175 proteins) SSF48264:Cytochrome_P450 (4012 proteins)
Type [?]	Family
Dates [?]	1999-10-08 17:07:25.0 (created) 2000-02-17 17:11:42.0 (modified)
Children [?] [tree]	IPR002397: B-class P450 IPR002399: Mitochondrial P450 IPR002401: E-class P450, group I IPR002402: E-class P450, group II IPR002403: E-class P450, group IV
Process [?]	electron transport (GO:0006118)
Abstract [?]	The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponent electron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds. P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class [1].
Structural links [?]	SCOP a.104.1.1 CATH 1.10.630.10 PDB/MSD - click here
Database links [?]	PANDIT PF00067

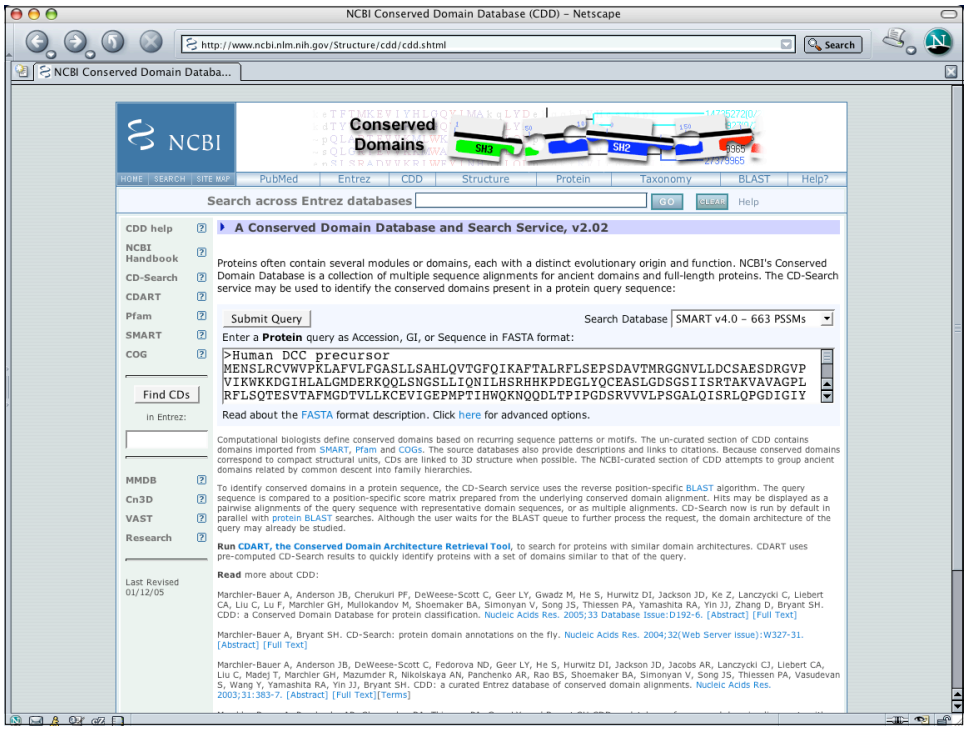
[FW]-[SGNH]-x-[GD]-{F}-[RKHPT]-{P}-C-[LIVMFAP]-[GAD]


InterPro: Cytochrome P450 – Netscape
http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001128
EMBL-EBI
European Bioinformatics Institute
EBI Home
About EBI
Research
Services
Toolbox
Databases
Downloads
Submissions
InterPro
InterPro home
Text Search
Sequence Search
Databases
Documentation
FTP site
Protein of the month
Search:
Search Entries
Search InterPro
InterPro Cytochrome P450
IPR001128 Cytochrome_P450
Matches: 4047 proteins. View matches:
Name
Cytochrome P450
Signatures
PF00067:p450 (3834 proteins)
PR00385:P450 (2932 proteins)
PS00086:CYTOCHROME_P450 (3175 proteins)
SSF48264:Cytochrome_P450 (4012 proteins)
Type
Family
Dates
1999-10-08 17:07:25.0 (created)
2000-02-17 17:11:42.0 (modified)
Children
IPR002397: B-class P450
IPR002399: Mitochondrial P450
IPR002401: E-class P450, group I
IPR002402: E-class P450, group II
IPR002403: E-class P450, group IV
Process
electron transport (GO:0006118)
Abstract
The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponent electron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds. P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class [1].
Structural links
SCOP a.104.1.1
CATH 1.10.630.10
PDB/MSD - click here
Database links
PANDIT PF00067
Parent-Child Relationships (Subfamilies)
Child entries are more specific than the parent
A match to the child entry implies a match to the parent
Signatures for the parent and child entries must overlap


InterPro: Cytochrome P450 – Netscape
http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001128
COMe PRX000236
Blocks IPB001128
PROSITE doc PDOC00081
Taxonomy
Saccharomyces cerevisiae
Unclassified
Fungi
Virus
Caenorhabditis elegans
Archaea
Nematoda
Bacteria
Metazoa
Cyanobacteria
Fruit Fly
Synechocystis PCC 6803
Arthropoda
Rice spp.
Chordata
Arabidopsis thaliana
Mouse
Green Plants
Human
Plastid Group
Eukaryota
Other Eukaryotes
Examples
Q64459 CP3B_MOUSE
P12939 CPD5_RAT
P33267 C2F2_MOUSE
P30612 CP5P_CANTR
P21595 CP56_YEAST
P26911 CPXH_STRGR
More proteins...
IPR001128 Cytochrome P450
IPR002397 B-class P450
IPR002401 E-class P450, group I
IPR002974 P450, CYP52
IPR008069 E-class P450, CYP2D
IPR008072 E-class P450, CYP3A
Center
Inner circles
Outer circles
Tree root
Tree nodes
Representative model organisms
There is no significance to the placement of individual nodes on the circles

Conserved Domain Database (CDD)

- Identify conserved domains in a protein sequence
- "Secondary database"
  - Pfam A and B
  - Simple Modular Architecture Research Tool (SMART)
  - Clusters of Orthologous Groups
- Search performed using RPS-BLAST
  - Query sequence is used to search a database of precalculated position-specific scoring tables
  - *Not* the same method used by ProfileScan
- *http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml*



NCBI CD-Search – Netscape

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

NCBI CD-Search

NCBI

NCBI Conserved Domain Search

New Search | PubMed | Nucleotide | Protein | Structure | CDD | Taxonomy | Help?

```
RPS-BLAST 2.2.9 [May-01-2004]
Query= local sequence: lcl|tmpseq_0 Human DCC precursor
       (1447 letters)

Database: oasis_smart.v2.02
```

Light blue — Low-complexity region

Click on boxes for multiple alignments

Show | Domain Relatives

.. This CD alignment includes 3D structure. To display structure, download **Cn3D**!

PSSMs producing significant alignments:		Score (bits)	E value
gnl|CDD|25322	smart00409, IG, Immunoglobulin;	67.1	3e-13
gnl|CDD|25322	smart00409, IG, Immunoglobulin;	61.7	1e-11
gnl|CDD|25322	smart00409, IG, Immunoglobulin;	59.0	8e-11
gnl|CDD|25322	smart00409, IG, Immunoglobulin;	42.8	6e-06
gnl|CDD|365	smart00408, IGc2, Immunoglobulin C-2 Type;	63.5	3e-12
gnl|CDD|365	smart00408, IGc2, Immunoglobulin C-2 Type;	59.3	6e-11
gnl|CDD|365	smart00408, IGc2, Immunoglobulin C-2 Type;	48.1	1e-07
gnl|CDD|365	smart00408, IGc2, Immunoglobulin C-2 Type;	45.4	8e-07
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	56.9	3e-10
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	56.5	5e-10
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	55.3	8e-10
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	48.8	8e-08
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	47.2	2e-07
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	37.2	2e-04

NCBI CD-Search – Netscape

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

NCBI CD-Search

gnl|CDD|25322, smart00409, IG, Immunoglobulin;

```
            CD-Length = 86 residues,  98.8% aligned
            Score = 67.1 bits (163), Expect = 3e-13

Query:  337   PSNLYAYESMDIEFECTVSGKPVPTVNWMKN-GDVVIPSDYFQIVGGSN---LRILGVVK  392
Sbjct:  1     PPSVTVKEGESVTLSCEASGNPPPEVTWYKQGGKLLAYSGRFSVSRSGGNSTLTISNVTP  60

Query:  393   SDEGFYQCVAENEAGNAQTSAQLIV  417
Sbjct:  61    EDSGTYTCAATNSSGSASSGTTLTV  85
```

gnl|CDD|25322, smart00409, IG, Immunoglobulin;

```
            CD-Length = 86 residues,  91.9% aligned
            Score = 61.7 bits (149), Expect = 1e-11

Query:  147   ESVTAFMGDTVLLKCEVIGEPMPTIHWQKNQQDLTPIPGDSRVVVLPSG---ALQISRLQ  203
Sbjct:  2     PSVTVKEGESVTLSCEASGNPPPEVTWYK--QGGKLLAYSGRFSVSRSGGNSTLTISNVT  59

Query:  204   PGDIGIYRCSARNPASSRTGN  224
Sbjct:  60    PEDSGTYTCAATNSSGSASSG  80
```

gnl|CDD|25322, smart00409, IG, Immunoglobulin;

```
            CD-Length = 86 residues, 100.0% aligned
            Score = 59.0 bits (142), Expect = 8e-11

Query:  246   PSNVVAIEGKDAVLECCVSGYPPPSFTWLRGEEVIQLRSKKY---SLLGGSNLLISNVTD  302
Sbjct:  1     PPSVTVKEGESVTLSCEASGNPPPEVTWYKQGGKLLAYSGRFSVSRSGGNSTLTISNVTP  60

Query:  303   DDSGMYTCVVTYKNENISASAELTVL  328
Sbjct:  61    EDSGTYTCAATNSSGSASSGTTLTVL  86
```

gnl|CDD|25322, smart00409, IG, Immunoglobulin;

```
            CD-Length = 86 residues,  98.8% aligned
            Score = 42.8 bits (100), Expect = 6e-06
```

NCBI CD-Search - Netscape

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

NCBI CD-Search

NCBI

NCBI Conserved Domain Search

New Search | PubMed | Nucleotide | Protein | Structure | CDD | Taxonomy | Help?

```
RPS-BLAST 2.2.9 [May-01-2004]
Query= local sequence: lcl|tmpseq_0 Human DCC precursor
        (1447 letters)

Database: oasis_smart.v2.02
```

Click on boxes for multiple alignments

Show | Domain Relatives

.. This CD alignment includes 3D structure. To display structure, download **Cn3D**!

PSSMs producing significant alignments:		Score (bits)	E value
gnl|CDD|25322	smart00409, IG, Immunoglobulin;	67.1	3e-13
gnl|CDD|25322	smart00409, IG, Immunoglobulin;	61.7	1e-11
gnl|CDD|25322	smart00409, IG, Immunoglobulin;	59.0	8e-11
gnl|CDD|25322	smart00409, IG, Immunoglobulin;	42.8	6e-06
gnl|CDD|365	smart00408, IGc2, Immunoglobulin C-2 Type;	63.5	3e-12
gnl|CDD|365	smart00408, IGc2, Immunoglobulin C-2 Type;	59.3	6e-11
gnl|CDD|365	smart00408, IGc2, Immunoglobulin C-2 Type;	48.1	1e-07
gnl|CDD|365	smart00408, IGc2, Immunoglobulin C-2 Type;	45.4	8e-07
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	56.9	3e-10
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	56.5	5e-10
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	55.3	8e-10
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	48.8	8e-08
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	47.2	2e-07
gnl|CDD|25286	smart00060, FN3, Fibronectin type 3 domain; One of three types...	37.2	2e-04

NCBI CDD smart00409 - Netscape

http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=smart00409&version=v2.02

NCBI CDD smart00409

NCBI

Conserved Domains

HOME | SEARCH | SITE MAP | Entrez | CDD | Structure | Protein | Help

smart00409.10

IG

Immunoglobulin;

Links:
Source: Smart
Taxonomy: root
Proteins: smart00409 related
Related CD: 7 links

Statistics:
PSSM-Id: 25322
Aligned: 472 rows
PSSM: 86 columns
Status: Alignment from source
Created: 12-Dec-2003
Updated: 12-Dec-2003

Structure:
Show Structure
Program: Cn3D
Drawing: Virtual Bonds
(**download Cn3D**)

This domain model appears to be related to other CDs:

[mouse over cd tag to display the number of PSSM pairs and cd name]

Show Alignment | Format: Compact Hypertext | Row Display: up to 10 | Color Bits: 2.0 bits | Type Selection: the most diverse members

```
consensus     1 PPSVTVKEGESVTLSCEASG.[1].PPPEVTW     YK.[2].GKLL.[ 6].SVSR.[3].NSTLTISNVTPE.[2]. 63
1MCP_H        7 SGGGLVQPGGSLRLSCATSG.[3].SDFYMEW     VR.[6].LEWI.[22].IVSR.[5].ILYLQMNALRAE.[2]. 93
1GYA          6 ALETWGALGQDINLDIPSFQ.[3].DIDDIKW     EK.[3].KKKI.[16].KLFK     NGTLKIKHLKTD.[2]. 78
1ZXQ          9 PKKLAVEPKGSLEVNCSTTC.[1].QPEVGGL     ET.[1].LNKI      LLDE.[3].WKHYLVSNISHD      62
gi 399208    25 QRPLLIVANRTATLVCNYTY.[4].KEFRASL     HK.[4].AVEV.[20].RGIH.[3].KVIFNLWNMSAS.[2]. 106
```

NCBI CDD smart00409 - Netscape

http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=smart00409&version=v2.02

NCBI CDD smart00409

Show Structure

Program: Cn3D

Drawing: Virtual Bonds

(download Cn3D)

[mouse over cd tag to display the number of PSSM pairs and cd name]

Show Alignment

Format: Compact Hypertext Row Display: up to 10 Color Bits: 2.0 bits

Type Selection: the most diverse members

```
consensus      1 PPSVTVKEGESVTLSCEASG.[1].PPPEVTW     YK.[2].GKLL.[ 6].SVSR.[3].NSTLTISNVTPE.[2]. 63
1MCP_H         7 SGGGLVQPGGSLRLSCATSG.[3].SDFYMEW     VR.[6].LEWI.[22].IVSR.[5].ILYLQMNALRAE.[2]. 93
1GYA           6 ALETWGALGQDINLDIPSFQ.[3].DIDDIKW     EK.[3].KKKI.[16].KLFK     NGTLKIKHLKTD.[2]. 78
1ZXQ           9 PKKLAVEPKGSLEVNCSTTC.[1].QPEVGGL     ET.[1].LNKI      LLDE.[3].WKHYLVSNISHD     62
gi 399208     25 QRPLLIVANRTATLVCNYTY.[4].KEFRASL     HK.[4].AVEV.[20].RGIH.[3].KVIFNLWNMSAS.[2]. 106
gi 461714    216 SNTFYAREGDQVEFSFPLSF.[2].ENLVGEL     RW.[9].LWIS.[19].QMKE.[2].PLRFTLPQVLSR.[2]. 298
gi 6166597    64 PQGGTVKVGEDITFIAKVKA.[6].PTIKWFK.[1].KW.[6].AGKH.[ 7].ERHS.[3].TFEMQIIKAKDN.[2]. 137
gi 729801    435 QRTQYGLVGDTARIECFASS.[3].ARHVSWT     FN.[1].QEIS.[ 7].SILV.[7].KSTLIIRDSQAY.[2]. 503
gi 124310    243 LKTISASLGSRLTIPCKVFL.[4].PLTTMLW     WT.[1].NDTH.[18].SENN.[4].EVPLIFDPVTRE.[3]. 321
gi 1709202   281 REGETMSLGCRVVITPEIKH     FQPEIRW     YR.[1].GVPL.[ 6].QTLW.[3].RATLTFSHLNKE.[2]. 341

consensus     64 GTYTCAAT.[2].SGSASS     GTTLTVL 86
1MCP_H        94 AIYYCARN.[6].YFDVWG.[1].GTTVTVS 121
1GYA          79 DIYKVSIY.[4].KNVLEK     IFDLKIQ 103
1ZXQ          63 TVLQCHFT.[2].GKQESM     NSNVSVY 85
gi 399208    107 DIYFCKIE.[7].VYNEKS.[1].GTVIHVR 135
gi 461714    299 GSGILTLN.[2].KGTLYQ     EVNLVVM 321
gi 6166597   138 GNYRCEVT.[2].DKFDSC     SFDLEVH 160
gi 729801    504 GKYNCTVV.[2].YGNDVA     EIQLQAK 526
gi 124310    322 MDFKCVVH     NTLSFQ     TLRTTVK 342
gi 1709202   342 GLYTIRVR     MGEYYE     QYSAYVF 362
```

Citing CDD: Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2005), *"CDD: a Conserved Domain Database for protein classification."*, **Nucleic Acids Res. 33**: D192-6

| Disclaimer | Privacy statement | Accessibility |

NCBI CD-Search - Netscape

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

NCBI CD-Search

NCBI

NCBI Conserved Domain Search

New Search | PubMed | Nucleotide | Protein | Structure | CDD | Taxonomy | Help?

```
RPS-BLAST 2.2.9 [May-01-2004]
Query= local sequence: lcl|tmpseq_0 Human DCC precursor
       (1447 letters)

Database: oasis_smart.v2.02
```

Click on boxes for multiple alignments

1 200 400 600 800 1000 1200 1447

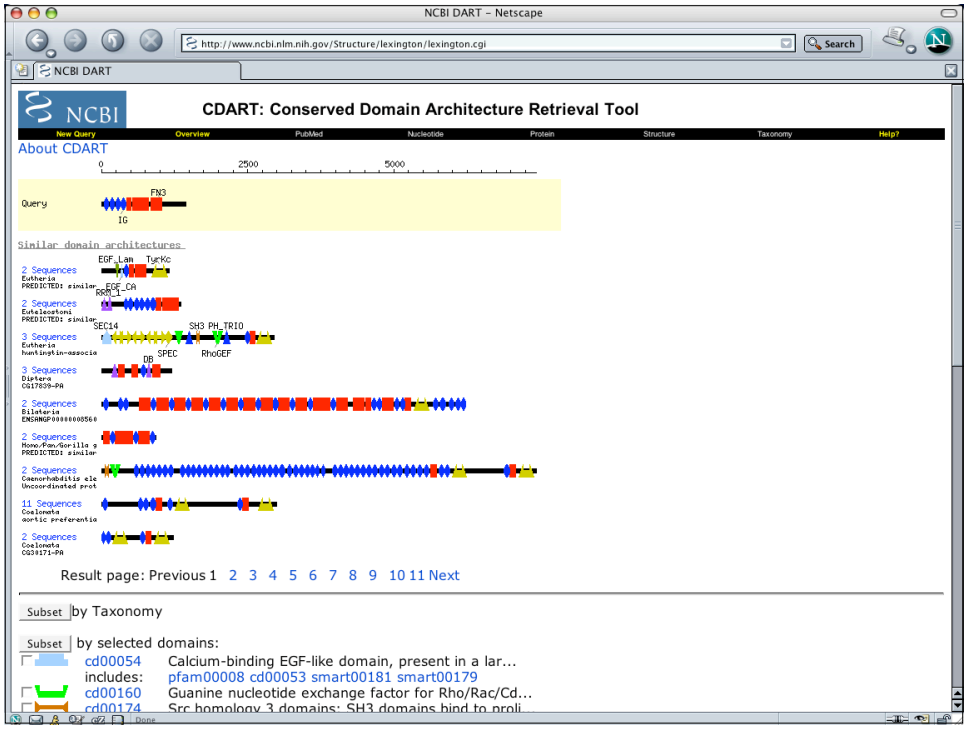
IG IG IG IG FN3 FN3 FN3 FN3 FN3 FN3

IGc2 IGc2 IGc2 IGc2

Show | Domain Relatives

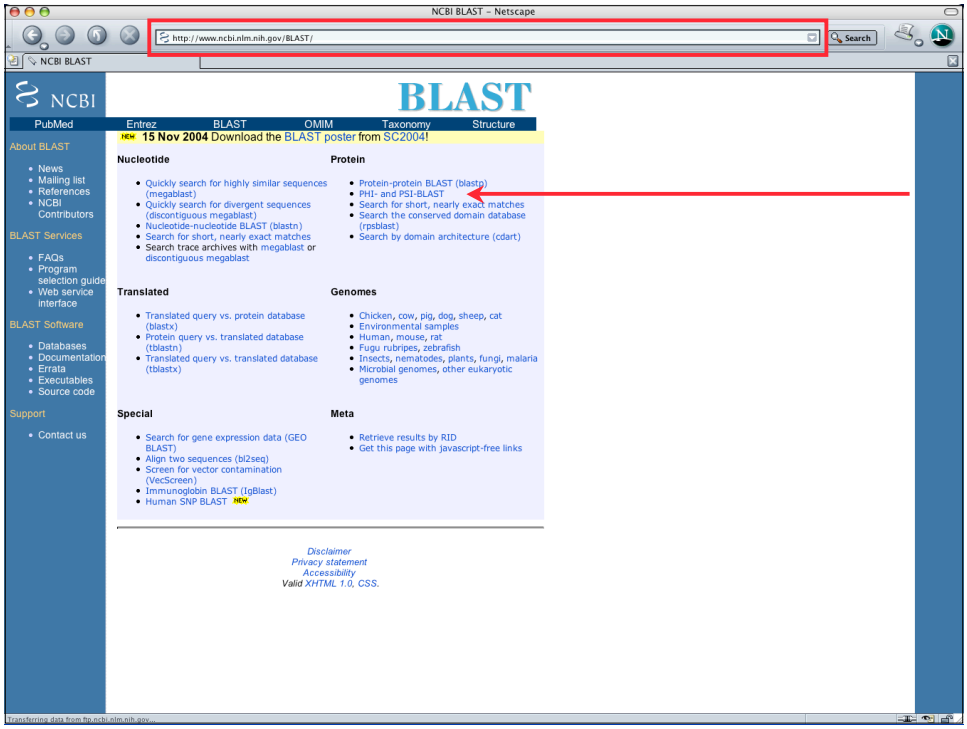
.. This CD alignment includes 3D structure. To display structure, download **Cn3D**!

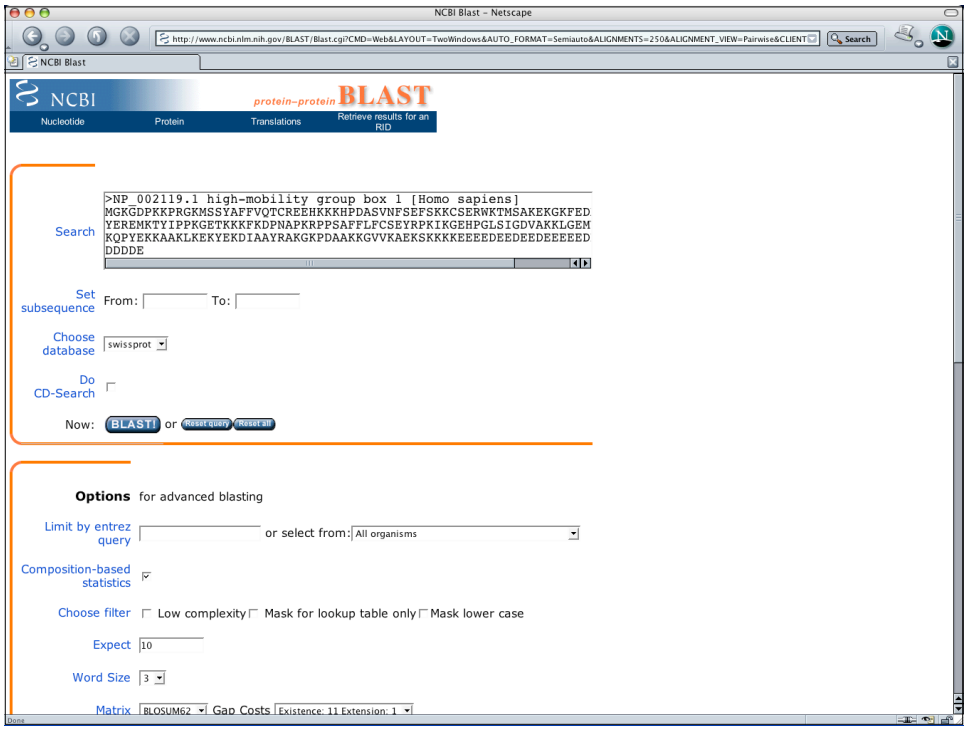
PSSMs producing significant alignments:	Score (bits)	E value
gnl|CDD|25322 smart00409, IG, Immunoglobulin;	67.1	3e-13
gnl|CDD|25322 smart00409, IG, Immunoglobulin;	61.7	1e-11
gnl|CDD|25322 smart00409, IG, Immunoglobulin;	59.0	8e-11
gnl|CDD|25322 smart00409, IG, Immunoglobulin;	42.8	6e-06
gnl|CDD|365 smart00408, IGc2, Immunoglobulin C-2 Type;	63.5	3e-12
gnl|CDD|365 smart00408, IGc2, Immunoglobulin C-2 Type;	59.3	6e-11
gnl|CDD|365 smart00408, IGc2, Immunoglobulin C-2 Type;	48.1	1e-07
gnl|CDD|365 smart00408, IGc2, Immunoglobulin C-2 Type;	45.4	8e-07
gnl|CDD|25286 smart00060, FN3, Fibronectin type 3 domain; One of three types...	56.9	3e-10
gnl|CDD|25286 smart00060, FN3, Fibronectin type 3 domain; One of three types...	56.5	5e-10
gnl|CDD|25286 smart00060, FN3, Fibronectin type 3 domain; One of three types...	55.3	8e-10
gnl|CDD|25286 smart00060, FN3, Fibronectin type 3 domain; One of three types...	48.8	8e-08
gnl|CDD|25286 smart00060, FN3, Fibronectin type 3 domain; One of three types...	47.2	2e-07
gnl|CDD|25286 smart00060, FN3, Fibronectin type 3 domain; One of three types...	37.2	2e-04

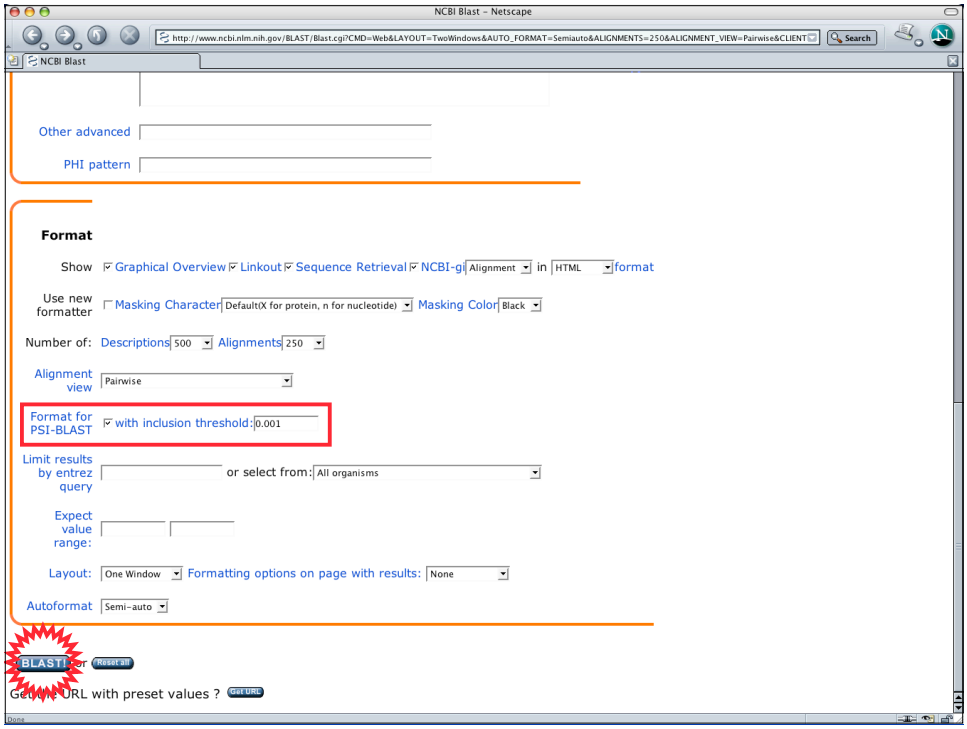


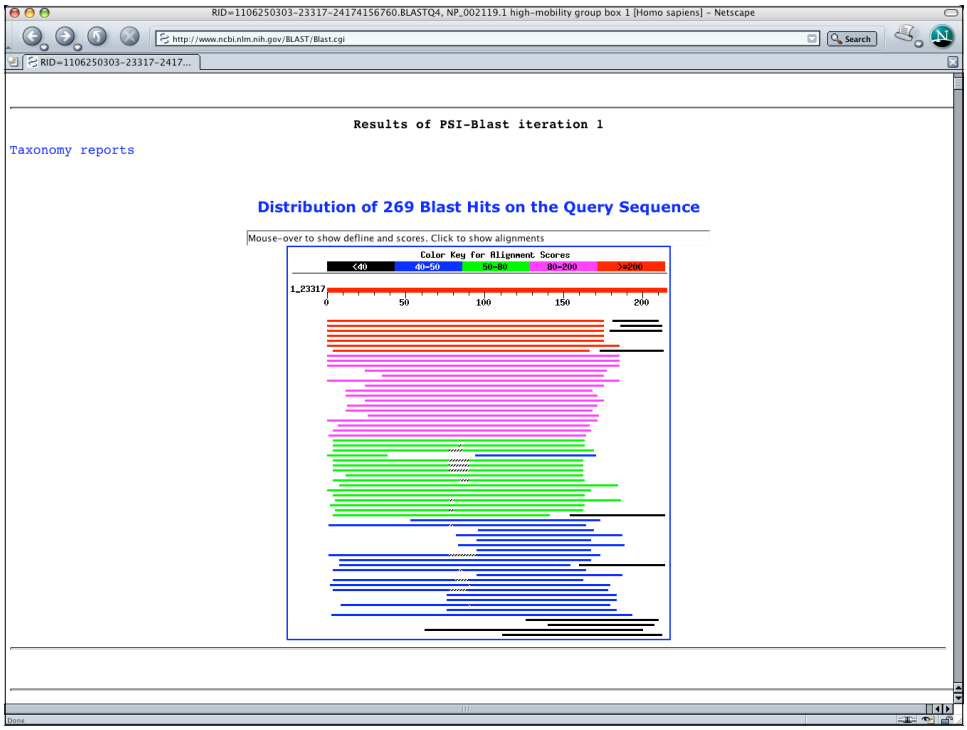
PSI-BLAST

- Position-Specific Iterated BLAST search
- Easy-to-use version of a profile-based search
  - Perform BLAST search against protein database
  - Use results to calculate a position-specific scoring matrix
  - PSSM replaces query for next round of searches
  - May be iterated until no new significant alignments are found
    - Convergence – all related sequences deemed found
    - Divergence – query is too broad, make cutoffs more stringent


NCBI BLAST – Netscape
http://www.ncbi.nlm.nih.gov/BLAST/
Search
NCBI BLAST
NCBI
BLAST
PubMed
Entrez
BLAST
OMIM
Taxonomy
Structure
15 Nov 2004 Download the BLAST poster from SC2004!
About BLAST
News
Mailing list
References
NCBI Contributors
BLAST Services
FAQs
Program selection guide
Web service interface
BLAST Software
Databases
Documentation
Errata
Executables
Source code
Support
Contact us
Nucleotide
Quickly search for highly similar sequences (megablast)
Quickly search for divergent sequences (discontiguous megablast)
Nucleotide-nucleotide BLAST (blastn)
Search for short, nearly exact matches
Search trace archives with megablast or discontiguous megablast
Protein
Protein-protein BLAST (blastp)
PHI- and PSI-BLAST
Search for short, nearly exact matches
Search the conserved domain database (rpsblast)
Search by domain architecture (cdart)
Translated
Translated query vs. protein database (blastx)
Protein query vs. translated database (tblastn)
Translated query vs. translated database (tblastx)
Genomes
Chicken, cow, pig, dog, sheep, cat
Environmental samples
Human, mouse, rat
Fugu rubripes, zebrafish
Insects, nematodes, plants, fungi, malaria
Microbial genomes, other eukaryotic genomes
Special
Search for gene expression data (GEO BLAST)
Align two sequences (bl2seq)
Screen for vector contamination (VecScreen)
Immunoglobin BLAST (IgBlast)
Human SNP BLAST
Meta
Retrieve results by RID
Get this page with javascript-free links
Disclaimer
Privacy statement
Accessibility
Valid XHTML 1.0, CSS.


NCBI Blast – Netscape
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CLIENT
Search
NCBI Blast
NCBI
protein-protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Search
>NP_002119.1 high-mobility group box 1 [Homo sapiens]
MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKFED
YEREMKTYIPPKGETKKKFKDPNAPKRPPSAFFLFCSEYRPKIKGEHPGLSIGDVAKKLGEM
KQPYEKKAAKLKEKYEKDIAAYRAKGKPDAAKKGVVKAEKSKKKKEEEEDEEDEEDEEEEED
DDDDE
Set subsequence
From:
To:
Choose database
swissprot
Do CD-Search
Now:
BLAST!
or
Reset query
Reset all
Options
for advanced blasting
Limit by entrez query
or select from:
All organisms
Composition-based statistics
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
10
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1


NCBI Blast - Netscape
NCBI Blast
Other advanced
PHI pattern
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
NCBI-gi
Alignment
in
HTML
format
Use new formatter
Masking Character
Masking Color
Number of:
Descriptions
Alignments
Alignment view
Pairwise
Format for PSI-BLAST
with inclusion threshold:
0.001
Limit results by entrez query
or select from:
All organisms
Expect value range:
Layout:
One Window
Formatting options on page with results:
None
Autoformat
Semi-auto
BLAST!
Get URL with preset values ?
Get URL


RID=1106250303-23317-24174156760.BLASTQ4, NP_002119.1 high-mobility group box 1 [Homo sapiens] - Netscape
Results of PSI-Blast iteration 1
Taxonomy reports
Distribution of 269 Blast Hits on the Query Sequence
Mouse-over to show defline and scores. Click to show alignments
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1_23317
0
50
100
150
200

RID=1106250303-23317-24174156760.BLASTQ4, NP_002119.1 high-mobility group box 1 [Homo sapiens] - Netscape

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

Legend:

- means that the alignment score was below the threshold on the previous iteration

- means that the alignment was checked on the previous iteration

Run PSI-Blast iteration 2

Hit list size 500

Sequences with E-value BETTER than threshold

Sequences producing significant alignments:	Score (bits)	E Value
gi|123371|sp|P12682|HMG1_PIG High mobility group protein 1 (HMG-1)	238	7e-63
gi|52783747|sp|P63158|HMG1_MOUSE High mobility group protein 1 (...	238	1e-62
gi|123367|sp|P10103|HMG1_BOVIN High mobility group protein 1 (HM...	238	1e-62
gi|123369|sp|P09429|HMG1_HUMAN High mobility group protein 1 (HM...	238	1e-62
gi|20138433|sp|Q9UGV6|HM1X_HUMAN High mobility group protein 1-l...	229	5e-60
gi|123373|sp|P26584|HMG2_CHICK High mobility group protein 2 (HM...	202	6e-52
gi|123382|sp|P07746|HMGT_ONCMY High mobility group-T protein (HM...	201	1e-51
gi|1708260|sp|P52925|HMG2_RAT High mobility group protein 2 (HMG-2)	194	2e-49
gi|123374|sp|P26583|HMG2_HUMAN High mobility group protein 2 (HM...	193	2e-49
gi|1708259|sp|P30681|HMG2_MOUSE High mobility group protein 2 (H...	193	2e-49
gi|13878931|sp|P23497|SP10_HUMAN Nuclear autoantigen Sp-100 (Spe...	191	1e-48
gi|123368|sp|P07156|HMG1_CRIGR High mobility group protein 1 (HM...	188	9e-48

RID=1106250303-23317-24174156760.BLASTQ4, NP_002119.1 high-mobility group box 1 [Homo sapiens] - Netscape

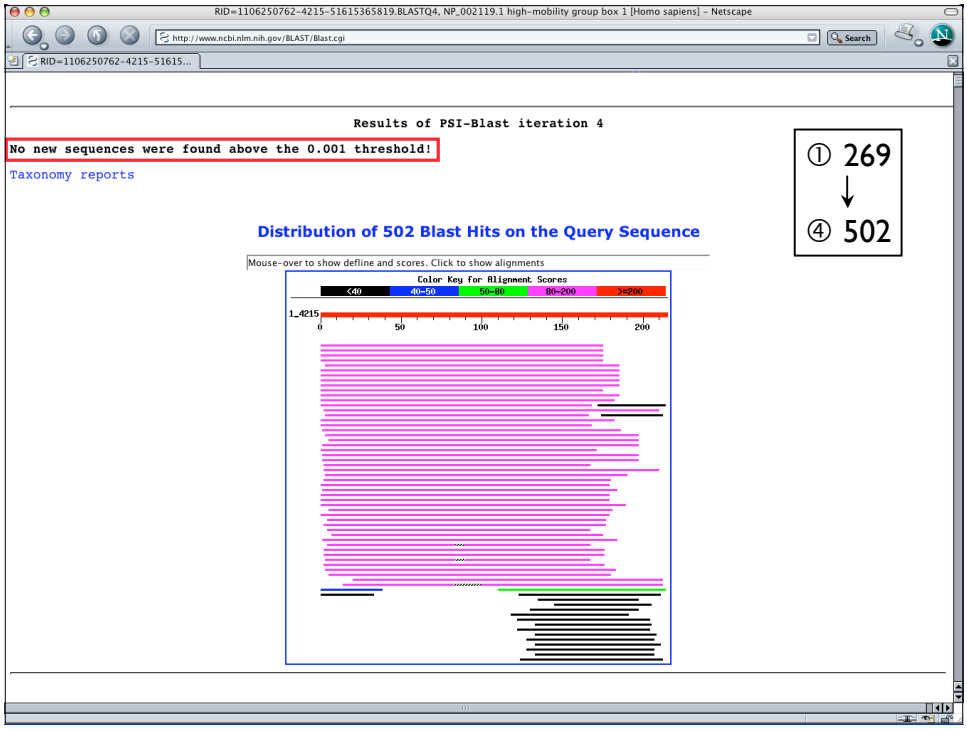
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi

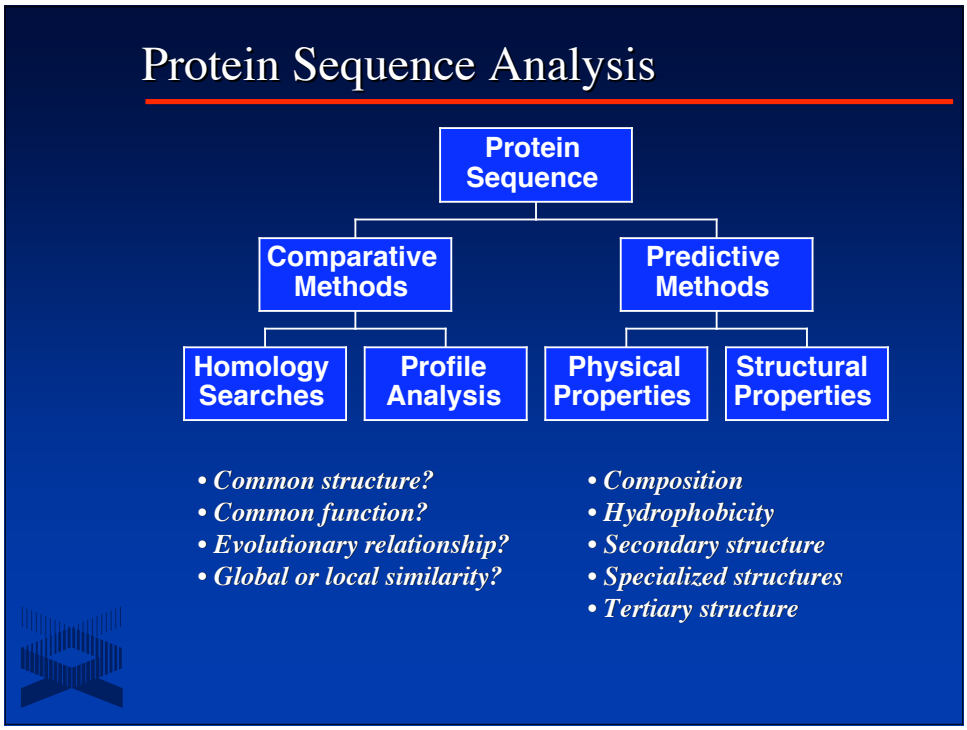
gi|2495255|sp|Q03435|NH10_YEAST Non-histone protein 10 (High mob...	42	8e-04
gi|6175054|sp|P36389|SRY_HORSE Sex-determining region Y protein ...	42	8e-04
gi|22654148|sp|Q91ZW1|TFAM_RAT Transcription factor A, mitochond...	42	8e-04
gi|6175076|sp|Q04888|SX10_MOUSE Transcription factor SOX-10 (SOX...	42	8e-04
gi|6094380|sp|O55170|SX10_RAT Transcription factor SOX-10	42	9e-04

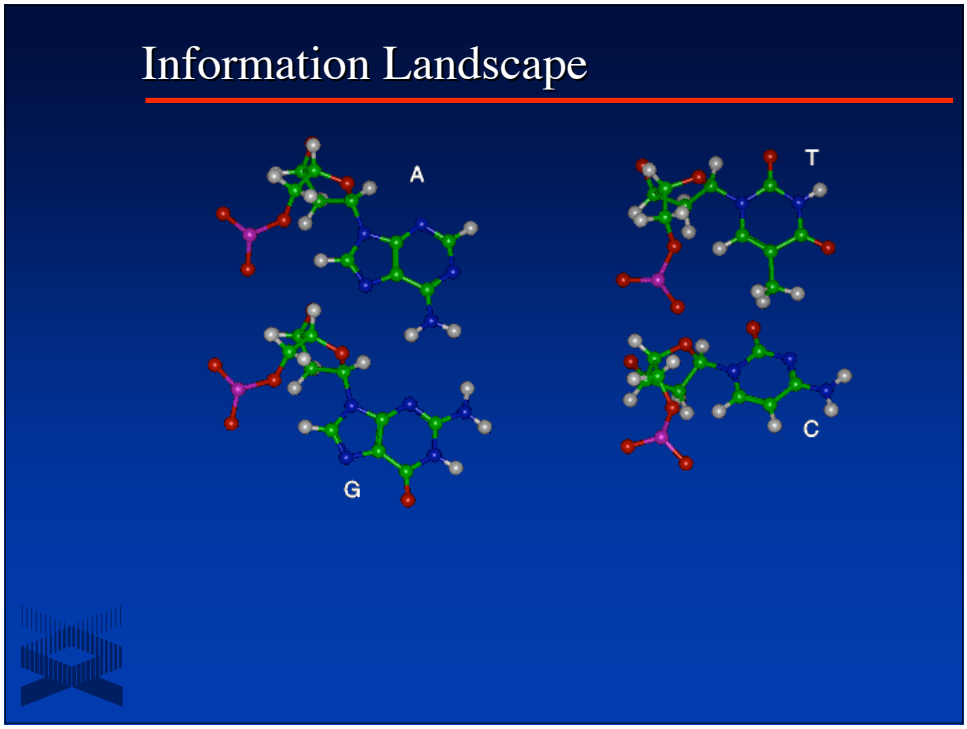
Run PSI-Blast iteration 2 ← ② … ③ … ④

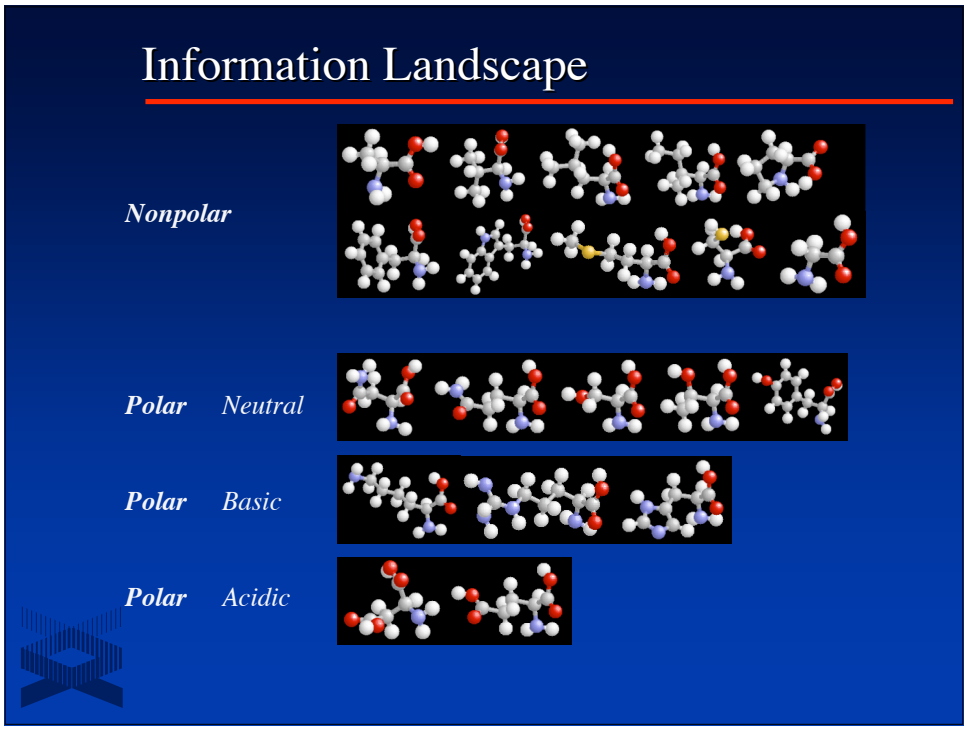
Sequences with E-value WORSE than threshold

gi|6831689|sp|O95416|SX14_HUMAN Transcription factor SOX-14 >gi|...	42	0.001
gi|2506521|sp|P48434|SOX9_CHICK Transcription factor SOX-9	42	0.001
gi|24638225|sp|Q9W7R6|SX14_CHICK Transcription factor SOX-14	42	0.001
gi|19862533|sp|Q04892|SX14_MOUSE Transcription factor SOX-14	42	0.001
gi|1711465|sp|P54231|SOX2_SHEEP Transcription factor SOX-2	42	0.001
gi|1351091|sp|P48431|SOX2_HUMAN Transcription factor SOX-2	42	0.001
gi|3913481|sp|Q24533|DICH_DROME SOX-domain protein dichaete (Fis...	42	0.001
gi|12644266|sp|P43267|SX15_MOUSE SOX-15 protein	42	0.001
gi|1723428|sp|Q10241|CMB1_SCHPO Mismatch-binding protein cmb1	42	0.001
gi|6094324|sp|P48432|SOX2_MOUSE Transcription factor SOX-2	42	0.001
gi|136654|sp|P25977|UBF1_RAT Nucleolar transcription factor 1 (U...	42	0.001
gi|136652|sp|P17480|UBF1_HUMAN Nucleolar transcription factor 1 ...	42	0.002
gi|729738|sp|P40621|HMGL_WHEAT HMG1/2-like protein	41	0.002
gi|730136|sp|P40632|NHP1_BABBO High mobility group protein homol...	41	0.002


RID=1106250762-4215-51615365819.BLASTQ4, NP_002119.1 high-mobility group box 1 [Homo sapiens] - Netscape
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Search
RID=1106250762-4215-51615...
Results of PSI-Blast iteration 4
No new sequences were found above the 0.001 threshold!
Taxonomy reports
① 269
④ 502
Distribution of 502 Blast Hits on the Query Sequence
Mouse-over to show define and scores. Click to show alignments
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1_4215
0
50
100
150
200


Protein Sequence Analysis
Protein Sequence
Comparative Methods
Predictive Methods
Homology Searches
Profile Analysis
Physical Properties
Structural Properties
• Common structure?
• Common function?
• Evolutionary relationship?
• Global or local similarity?
• Composition
• Hydrophobicity
• Secondary structure
• Specialized structures
• Tertiary structure


Information Landscape
A
T
G
C


Information Landscape
Nonpolar
Polar Neutral
Polar Basic
Polar Acidic

ProtParam

- Computes physicochemical parameters
  - Molecular weight
  - Theoretical pI
  - Amino acid composition
  - Extinction coefficient
- Simple query
  - SWISS-PROT accession number
  - User-entered sequence, in single-letter format
- *http://www.expasy.ch/tools/protparam.html*

ProtParam Query

```
MNGEADCPTDLEMAAPKGQDRWSQEDMLTLLECMKNNLPSNDSSKFKTTESHMDWEKVAFKDFSGDMCKL
KWVEISNEVRKFRTLTELILDAQEHVKNPYKGKKLKKHPDFPKKPLTPYFRFFMEKRAKYAKLHPEM...
```

Compute parameters

```
Number of amino acids: 727
Molecular weight: 84936.8
Theoretical pI: 5.44

Amino acid composition:

Ala (A)  35       4.8%         Leu (L)  57       7.8%
Arg (R)  39       5.4%         Lys (K)  97      13.3%
Asn (N)  28       3.9%         Met (M)  25       3.4%
Asp (D)  58       8.0%         Phe (F)  18       2.5%
Cys (C)   6       0.8%         Pro (P)  39       5.4%
Gln (Q)  36       5.0%         Ser (S)  67       9.2%
Glu (E)  98      13.5%         Thr (T)  22       3.0%
Gly (G)  26       3.6%         Trp (W)  11       1.5%
His (H)  11       1.5%         Tyr (Y)  20       2.8%
Ile (I)  18       2.5%         Val (V)  16       2.2%

Asx (B)   0       0.0%
Glx (Z)   0       0.0%
Xaa (X)   0       0.0%

Total number of negatively charged residues (Asp + Glu): 156
Total number of positively charged residues (Arg + Lys): 136
```

Expert Protein Analysis System (ExPASy)

- All tools available through a single Web front-end, at *http://us.expasy.org/tools*
- Primary sequence analysis tools include:

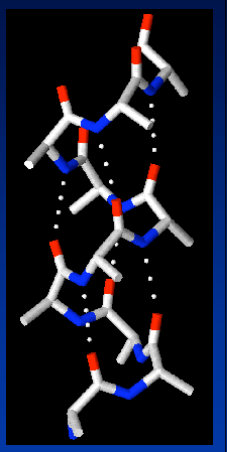
ProtParam
Compute pI/Mw
Titration Curve
ProtScale
Plot any measurable (e.g., hydrophobicity) by sequence position
HelixWheel/HelixDraw
Display protein sequence as a helical wheel

Secondary Structure Prediction

- Deduce the most likely position of alpha-helices and beta-strands
- Confirm structural or functional relationships when sequence similarity is weak
- Determine guidelines for rational selection of specific mutants for further laboratory study
- Basis for further structure-based studies

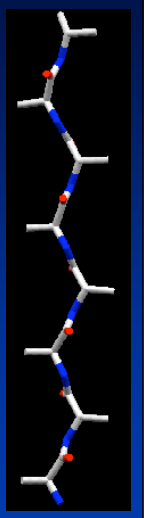
Alpha-helix

- Corkscrew
- Main chain forms backbone, side chains project out
- Hydrogen bonds between CO group at *n* and NH group at *n+4*
- Helix-formers: Ala, Glu, Leu, Met
- Helix-breaker: Pro



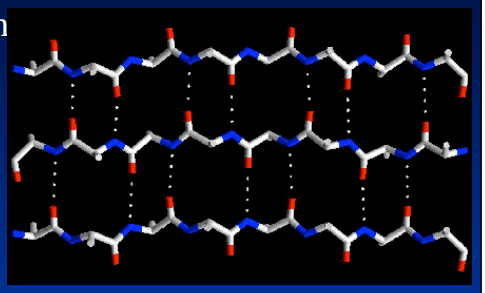
Beta-strand

- Extended structure ("pleated")
- Peptide bonds point in opposite directions
- Side chains point in opposite directions
- No hydrogen bonding *within* strand



Beta-sheet

- Stabilization through hydrogen bonding
- Parallel or antiparallel
- Variant: beta-turn



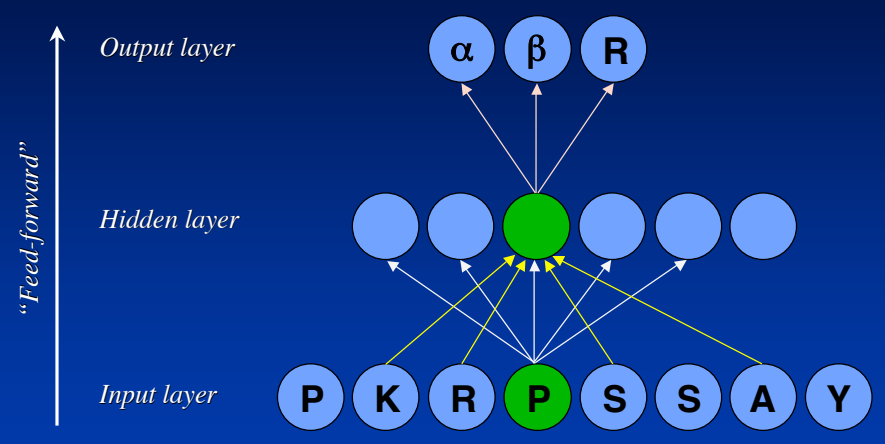
Folding Classes

α	β	α+β	α/β
Cyt c	*CD4*	*Staph nuclease*	*Triose phosphate isomerase*
Globins Orthogonal EF-hand Up-Down Cytochrome	Orthogonal Super-barrel Greek key Sandwich Jelly roll	Split sandwich Meander Metal-rich Open roll OB/UB roll	TIM barrel Doubly-wound

Neural Networks

- Used when direct cause-and-effect rules between the beginning and end states are not known
  - Beginning and end states must be related
  - Neural networks attempt to deduce the relationship between the beginning and end states
- Supervised learning approach
  - Involves use of "training sets" where relationship is known
  - Based on data in training sets, network attempts to "learn" the relationship between input and output layers

Neural Networks



nnpredict

- Neural network approach to making predictions
  (Kneller et al., 1990)
- Best-case accuracy > 65%
- Search engines
  - E-mail *nnpredict@celeste.ucsf.edu*
  - Web *http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html*

nnpredict Query

```
option: a/b
>flavodoxin - Anacystis nidulans
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVGELQSDWEGIY
DDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNNQFVG
LAIDEDNQPDLTKNRIKTWVSQLKSEFGL
```

↓ *α/β folding class*

```
Tertiary structure class: alpha/beta

Sequence:
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVG
ELQSDWEGIYDDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYW
PIEGYDFNESKAVRNNQFVGLAIDEDNQPDLTKNRIKTWVSQLKSEFGL

Secondary structure prediction (H = helix, E = strand, - = no prediction):
----EEE------EEEHHHHHHH------EEEH---------------EEEE--------
---------------HHHH---EEEE------------H--HHHHHHHH------E--E-
-E-------------HH--E---------------EHHHHH------
```

PredictProtein

- Multi-step predictive algorithm *(Rost et al., 1994)*
  - Protein sequence queried against SWISS-PROT
  - MaxHom used to generate iterative, profile-based multiple sequence alignment *(Sander and Schneider, 1991)*
  - Multiple alignment fed into neural network (PROFsec)
- Accuracy
  - Average > 70%
  - Best-case > 90%
- Search engines

  http://www.embl-heidelberg.de/predictprotein/

  http://cubic.bioc.columbia.edu/predictprotein/

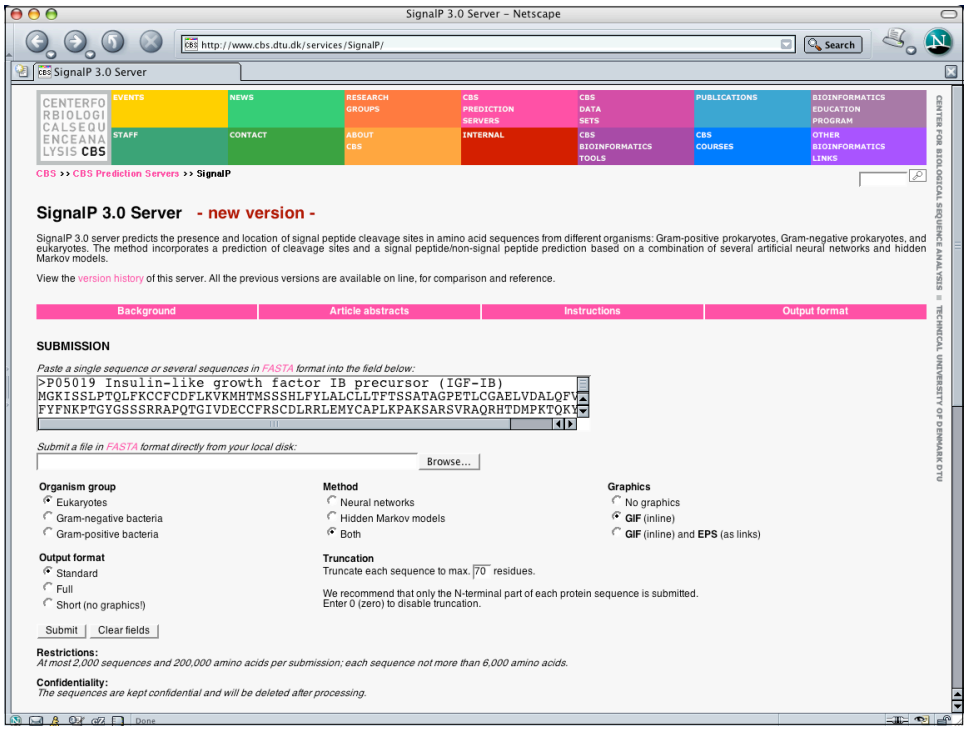
```
>flavodoxin - Anacystis nidulans
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVGELQSDWEGIY
DDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNNQFVG
LAIDEDNQPDLTKNRIKTWVSQLKSEFGL
```

```
PROF results (normal)

                ....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8....,....9....,....1
AA              AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVGELQSDWEGIYDDLDSVNFQGKKVAYFGAGDQVGYSDNFQD
OBS_sec
PROF_sec         EEEEEEE     HHHHHHHHHHHHH    EEEEE             EEEEEE          HHHHHHHHHH       EEEEEEE         HHHH
Rel_sec         9278998426762688889888731068425442235670001225537887203476665421678888863045688637888414664433114 46

                0.1.,....11.1.,....12.1.,....13.1.,....14.1.,....15.1.,....16.1.,....
AA              AMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNNQFVGLAIDEDNQPDLTKNRIKTWVSQLKSEFGL
OBS_sec
PROF_sec        HHHHHHHHHHH     EEEEE                 EEEEEE         HHHHHHHHHHHHH
Rel_sec         7888998887407825422024565445331001782688764266642111178899999888754289
```

- *SWISS-PROT hits*
- *Multiple alignment*
- *PDB homologues*

SignalP

- Neural network trained based on phylogeny
  - Gram-negative prokaryotic
  - Gram-positive prokaryotic
  - Eukaryotic
- Predicts secretory signal peptides (*not* those involved in intracellular signal transduction)
- *http://www.cbs.dtu.dk/services/SignalP/*



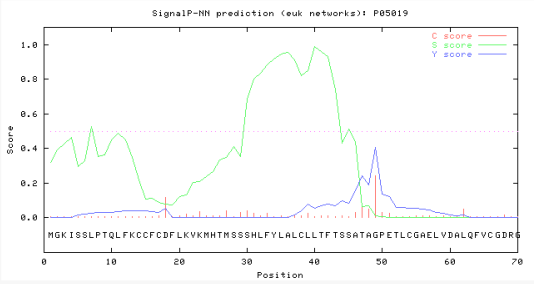
SignalP 3.0 Server - prediction results - Netscape

http://www.cbs.dtu.dk/cgi-bin/nph-webface?jobid=signalp,41F3C35406CB0AF4&opt=none

SignalP 3.0 Server - prediction re...

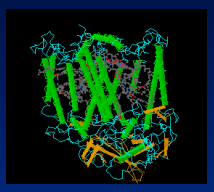
>P05019 Insulin-like growth factor IB precursor _IGF-IB_

SignalP-NN result:



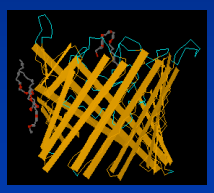

```
# data

>P05019                 length = 70
# Measure  Position  Value  Cutoff  signal peptide?
  max. C     49       0.241   0.32   NO
  max. Y     49       0.406   0.33   YES
  max. S     40       0.989   0.87   YES
  mean S     1-48     0.464   0.48   NO
       D     1-48     0.435   0.43   YES
# Most likely cleavage site between pos. 48 and 49: ATA-GP
```

Transmembrane Classes



- Helix bundles
  - Long stretches of apolar amino acids
  - Fold into transmembrane alpha-helices
  - "Positive-inside rule"

Cell surface receptors
Ion channels
Active and passive transporters



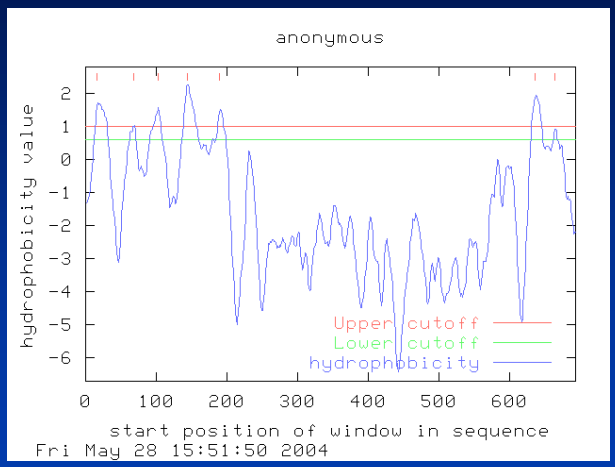
- Beta-barrel
  - Anti-parallel sheets rolled into cylinder

Outer membrane of Gram-negative bacteria
Porins (passive, selective diffusion)

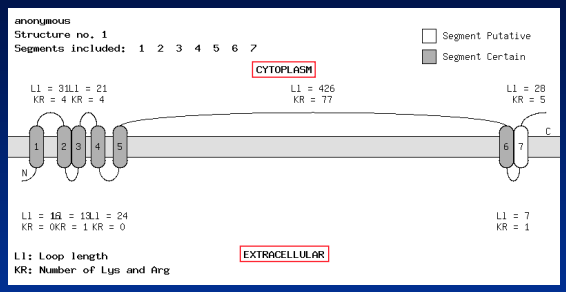
TopPred

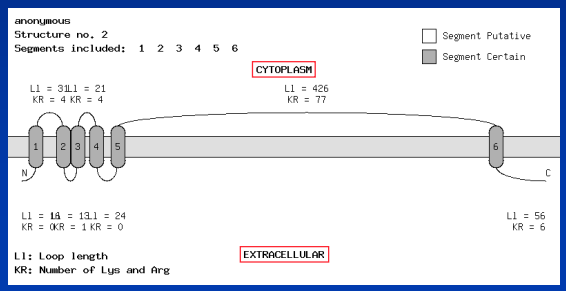
- Combines hydrophobicity analysis with the analysis of electrical charges
  - Calculates hydrophobicity profile
  - Hydrophobic-rich regions marked as "transmembrane"
  - Hydrophobic regions that fail to exceed a predefined cutoff are considered "putative transmembrane"
  - Topology prediction with and without putative helices
- Web-based search
  - *http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html*

TopPred



TopPred



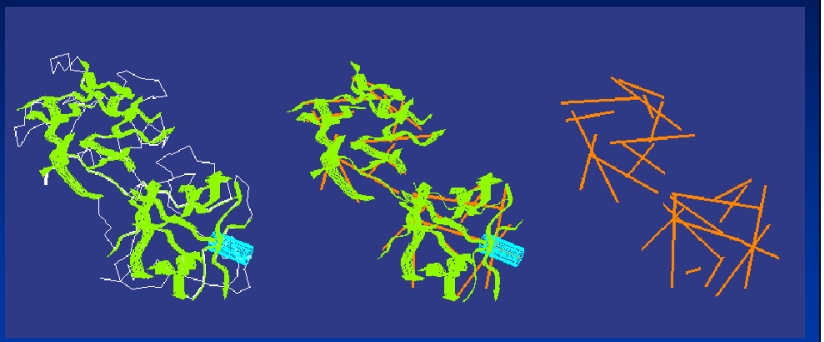


Predicting Tertiary Structure

- Sequence specifies conformation, *but* conformation does *not* specify sequence
- Structure is conserved to a much greater extent than sequence
- Similarities between proteins may not necessarily be detected through “traditional” methods

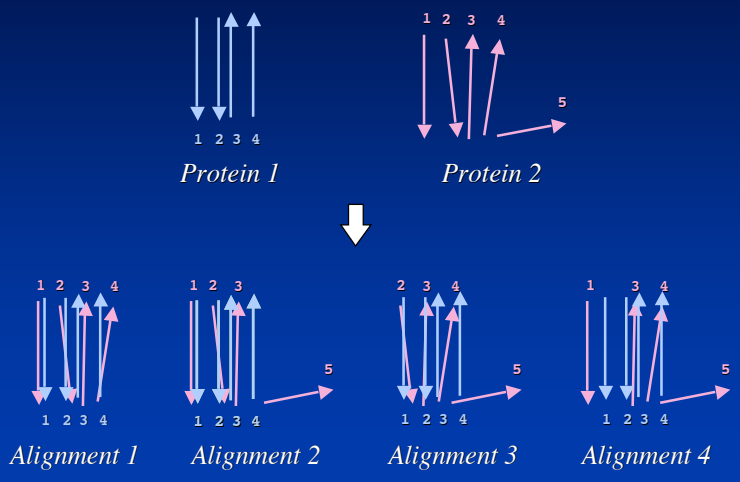
VAST Structure Comparison

Step 1: Construct vectors for secondary structure elements



VAST Structure Comparison

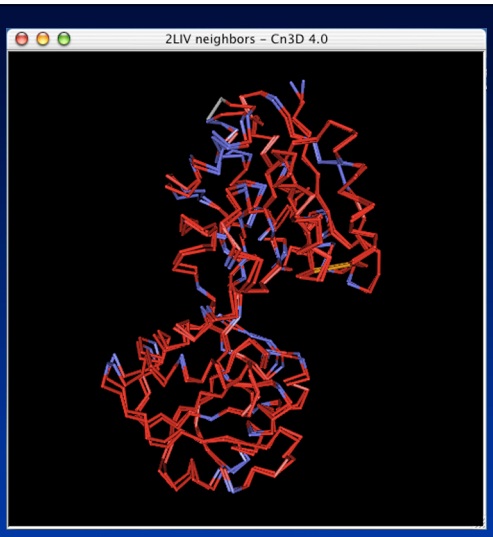
Step 2: Optimally align structure element vectors

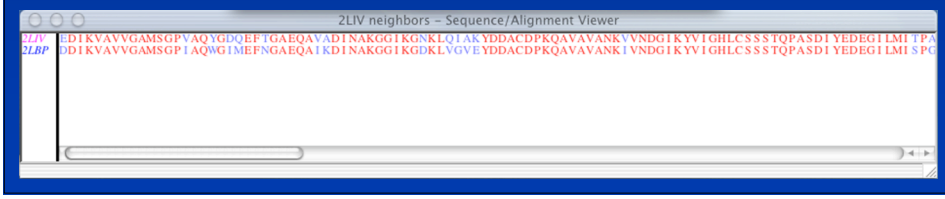


Cn3D Viewer

Rendering: Tubes

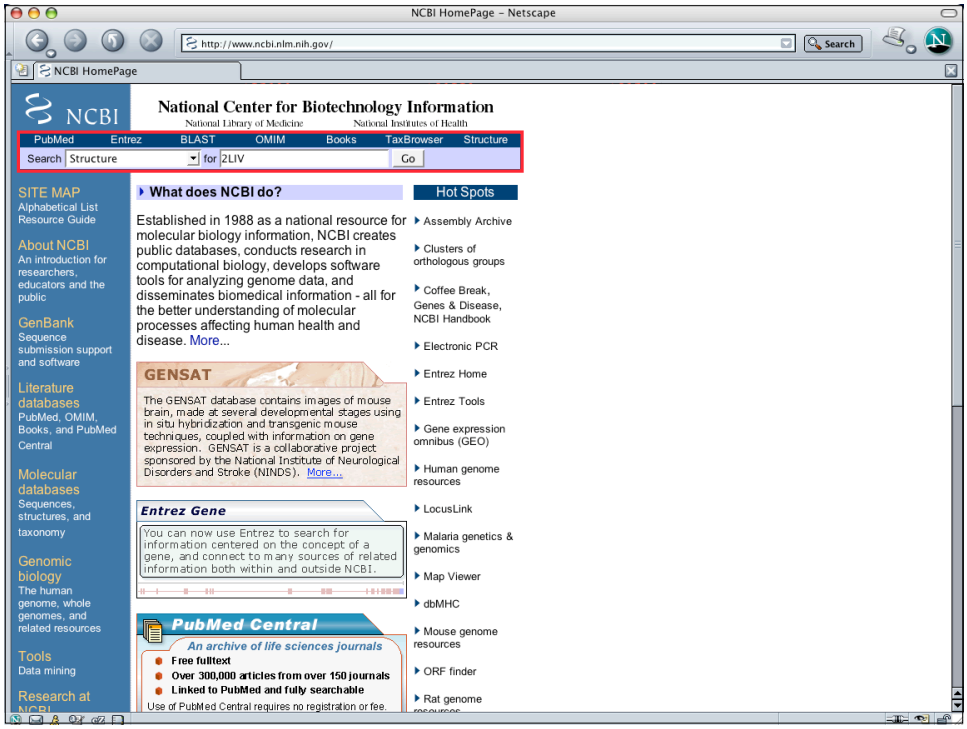
Coloring: Identity
Red – matches
Blue – mismatches

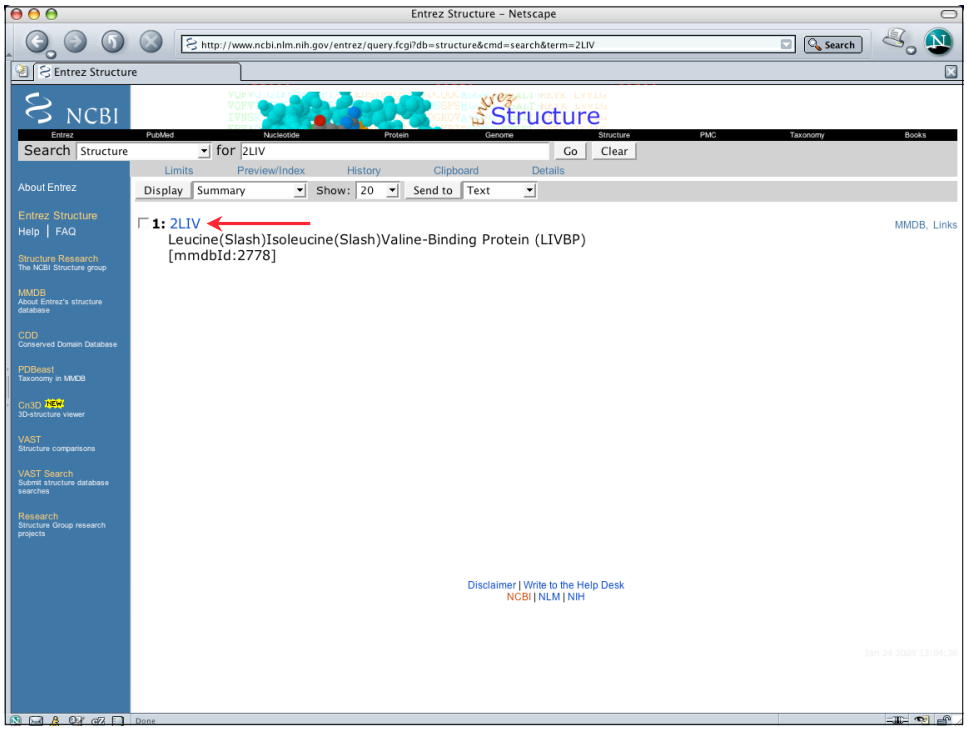


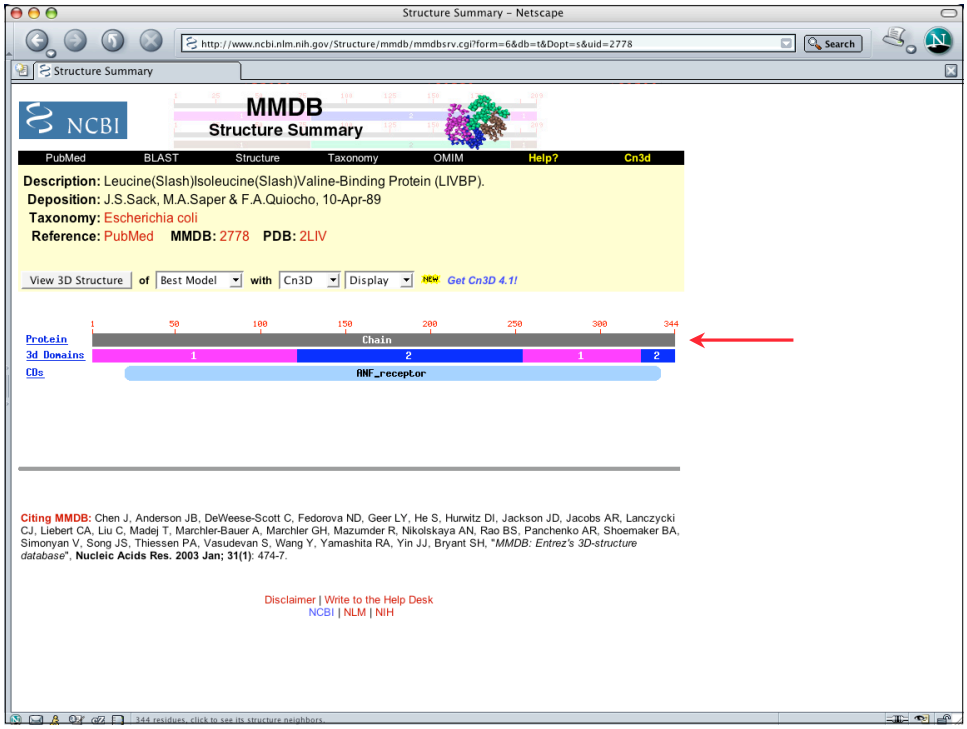


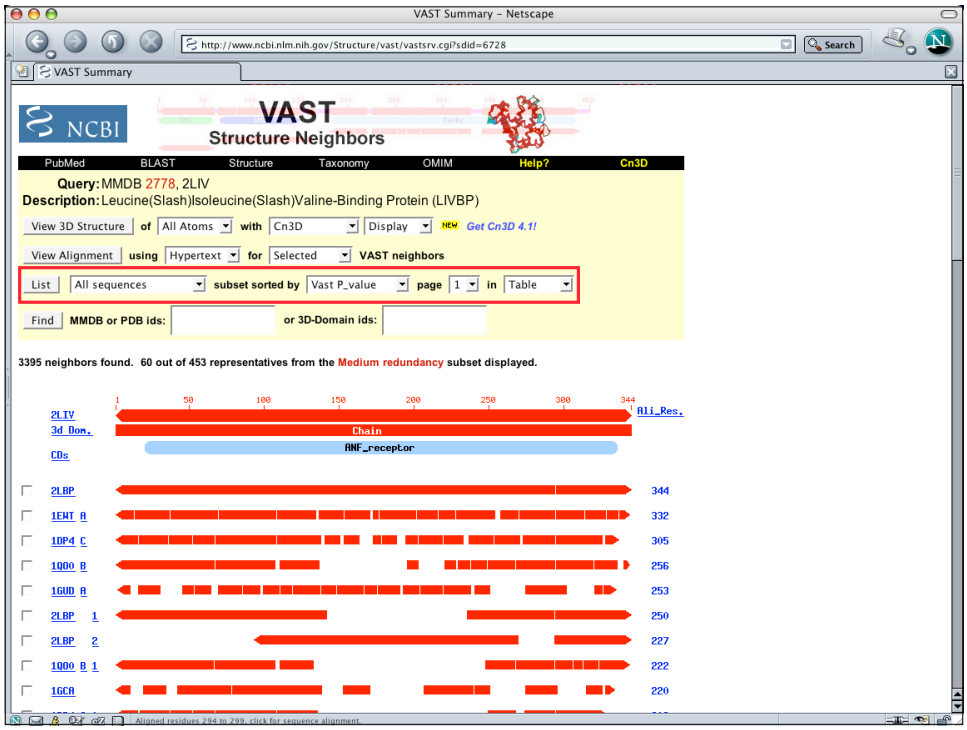
VAST Shortcomings

- Not the best method for determining structural similarities
- Reducing a structure to a series of vectors necessarily results in a loss of information (less confidence in prediction)
- Regardless of the "simplicity" of the method, provides a simple and fast first answer to the question of structural similarity


NCBI HomePage - Netscape
http://www.ncbi.nlm.nih.gov/
NCBI HomePage
NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed Entrez BLAST OMIM Books TaxBrowser Structure
Search Structure for 2LIV Go
SITE MAP
Alphabetical List Resource Guide
About NCBI
An introduction for researchers, educators and the public
GenBank
Sequence submission support and software
Literature databases
PubMed, OMIM, Books, and PubMed Central
Molecular databases
Sequences, structures, and taxonomy
Genomic biology
The human genome, whole genomes, and related resources
Tools
Data mining
What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...
GENSAT
The GENSAT database contains images of mouse brain, made at several developmental stages using in situ hybridization and transgenic mouse techniques, coupled with information on gene expression. GENSAT is a collaborative project sponsored by the National Institute of Neurological Disorders and Stroke (NINDS). More...
Entrez Gene
You can now use Entrez to search for information centered on the concept of a gene, and connect to many sources of related information both within and outside NCBI.
PubMed Central
An archive of life sciences journals
Free fulltext
Over 300,000 articles from over 150 journals
Linked to PubMed and fully searchable
Use of PubMed Central requires no registration or fee.
Hot Spots
Assembly Archive
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
Electronic PCR
Entrez Home
Entrez Tools
Gene expression omnibus (GEO)
Human genome resources
LocusLink
Malaria genetics & genomics
Map Viewer
dbMHC
Mouse genome resources
ORF finder
Rat genome


Entrez Structure - Netscape
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=structure&cmd=search&term=2LIV
Entrez Structure
NCBI
Structure
Entrez PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Search Structure for 2LIV Go Clear
Limits Preview/Index History Clipboard Details
Display Summary Show: 20 Send to Text
About Entrez
Entrez Structure
Help | FAQ
Structure Research
The NCBI Structure group
MMDB
About Entrez's structure database
CDD
Conserved Domain Database
PDBeast
Taxonomy in MMDB
Cn3D
3D-structure viewer
VAST
Structure comparisons
VAST Search
Submit structure database searches
Research
Structure Group research projects
1: 2LIV
Leucine(Slash)Isoleucine(Slash)Valine-Binding Protein (LIVBP)
[mmdbId:2778]
MMDB, Links
Disclaimer | Write to the Help Desk
NCBI | NLM | NIH
Done


Structure Summary - Netscape
http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?form=6&db=t&Dopt=s&uid=2778
Structure Summary
NCBI
MMDB
Structure Summary
PubMed BLAST Structure Taxonomy OMIM Help? Cn3d
Description: Leucine(Slash)Isoleucine(Slash)Valine-Binding Protein (LIVBP).
Deposition: J.S.Sack, M.A.Saper & F.A.Quiocho, 10-Apr-89
Taxonomy: Escherichia coli
Reference: PubMed MMDB: 2778 PDB: 2LIV
View 3D Structure of Best Model with Cn3D Display NEW Get Cn3D 4.1!
Protein
3d Domains
CDs
Chain
ANF_receptor
Citing MMDB: Chen J, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler-Bauer A, Marchler GH, Mazumder R, Nikolskaya AN, Rao BS, Panchenko AR, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH, "MMDB: Entrez's 3D-structure database", Nucleic Acids Res. 2003 Jan; 31(1): 474-7.
Disclaimer | Write to the Help Desk
NCBI | NLM | NIH


VAST Summary - Netscape
http://www.ncbi.nlm.nih.gov/Structure/vast/vastsrv.cgi?sdid=6728
VAST Summary
NCBI
VAST
Structure Neighbors
PubMed BLAST Structure Taxonomy OMIM Help? Cn3D
Query: MMDB 2778, 2LIV
Description: Leucine(Slash)Isoleucine(Slash)Valine-Binding Protein (LIVBP)
View 3D Structure of All Atoms with Cn3D Display NEW Get Cn3D 4.1!
View Alignment using Hypertext for Selected VAST neighbors
List All sequences subset sorted by Vast P_value page 1 in Table
Find MMDB or PDB ids: or 3D-Domain ids:
3395 neighbors found. 60 out of 453 representatives from the Medium redundancy subset displayed.
2LIV
3d Dom.
CDs
Chain
ANF_receptor
Ali_Res.
2LBP 344
1EWT A 332
1DP4 C 305
1QOO B 256
1GUD A 253
2LBP 1 250
2LBP 2 227
1QOO B 1 222
1GCA 220

VAST Summary – Netscape

NCBI

VAST
Structure Neighbors

PubMed | BLAST | Structure | Taxonomy | OMIM | Help? | Cn3D

Query: MMDB 2778, 2LIV
Description: Leucine(Slash)Isoleucine(Slash)Valine-Binding Protein (LIVBP)

View 3D Structure of All Atoms with Cn3D Display NEW Get Cn3D 4.1!

View Alignment using Hypertext for Selected VAST neighbors

List All sequences subset sorted by Vast P_value page 1 in Table

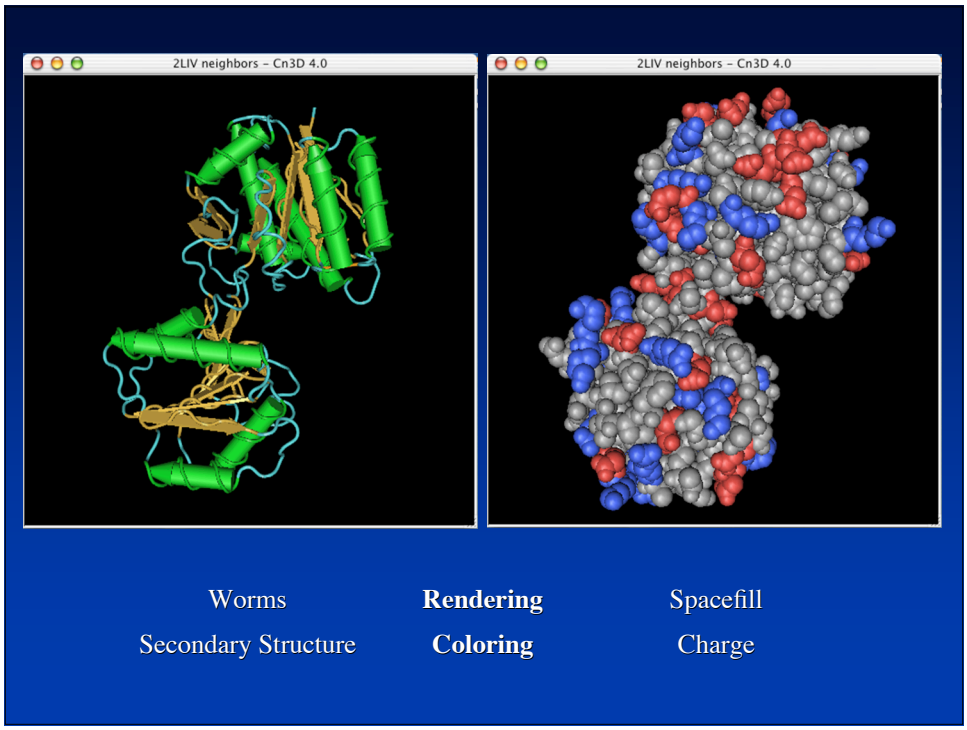
Find MMDB or PDB ids: or 3D-Domain ids:

60 out of 3395 neighbors displayed.

	PDB C D	Ali. Len.	SCORE	P-VAL	RMSD	%Id	MMDB Date	Description
	2LBP	344	39.8	10e-44.6	0.9	79.1	03/2001	Leucine-Binding Protein (LBP)
	1USG A	343	40.1	10e-42.4	2.0	79.0	01/2004	L-Leucine-Binding Protein, Apo Form
	1JDP B	310	29.9	10e-22.6	4.3	14.8	10/2001	Crystal Structure Of HormoneRECEPTOR COMPLEX
	1ISS B	330	30.3	10e-22.5	3.4	15.5	04/2002	Crystal Structure Of Metabotropic Glutamate Receptor Subtype 1 Complexed With An Antagonist
	1JDP A	322	29.8	10e-22.4	4.6	14.6	10/2001	Crystal Structure Of HormoneRECEPTOR COMPLEX

P-value ≤ 0.001
and
% Identity > 25
over at least 20 residues

Read the descriptions!



SWISS-MODEL

- Automated comparative protein modelling server
- Web front-end at
  http://www.expasy.org/swissmod
  Results returned by E-mail

BLAST search to find similarities in PDB *by sequence*

↓

Select templates with sequence identity > 25% and projected model size > 20 amino acids

↓

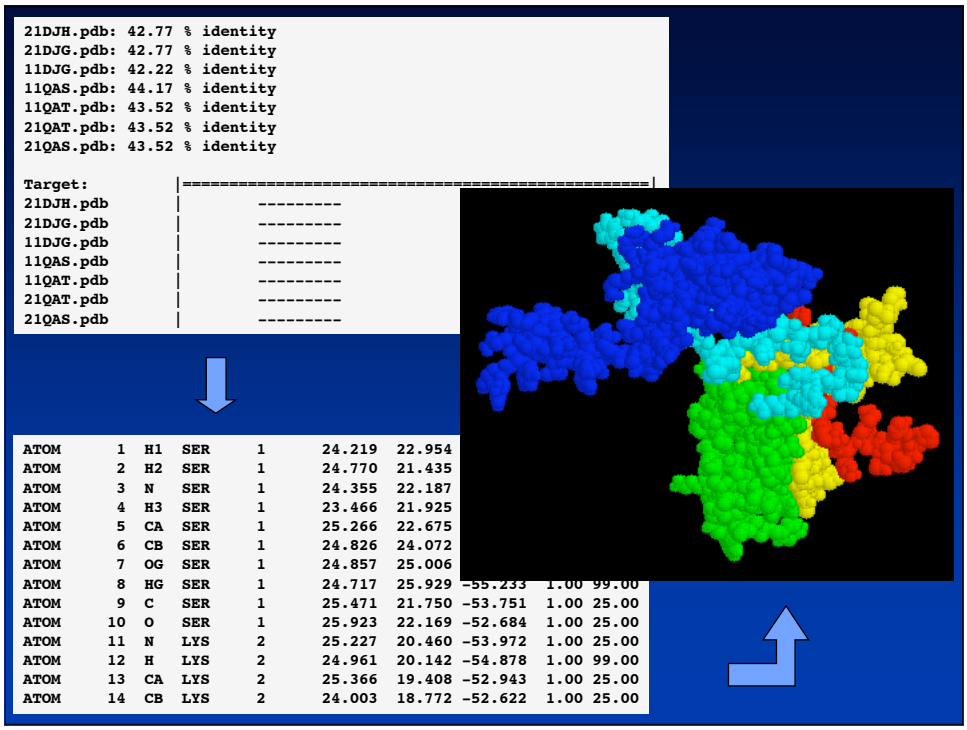
Generate models

↓

Do energy minimization

↓

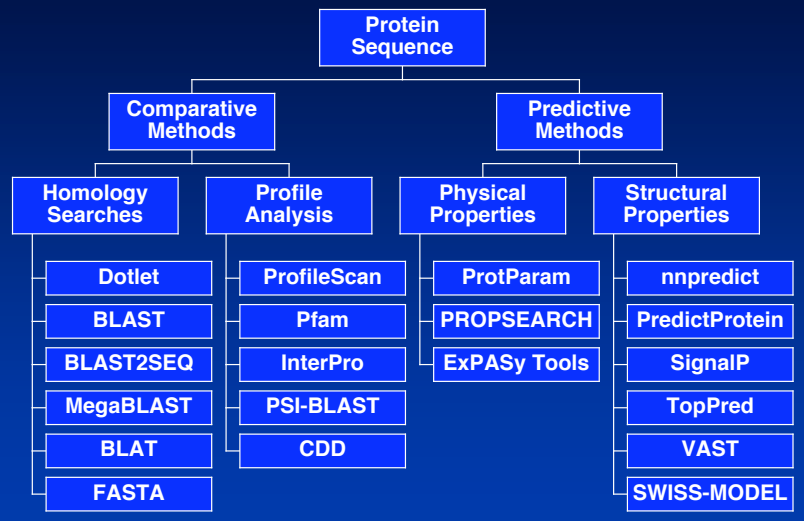
Generate PDB file for new protein model



Structural Modeling Software

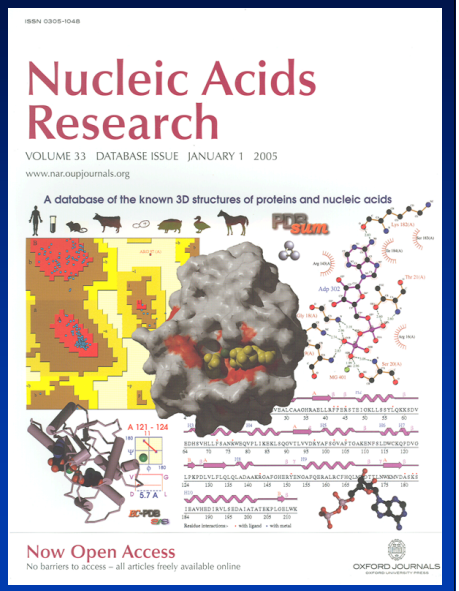
- 3D-JIGSAW
  http://www.bmm.icnet/uk/servers/3djigsaw
- ESyPred3D
  http://www.fundp.ac.be/urbm/bioinfo/esypred
- MODELLER
  http://www.salilab.org/modeller/modeller.html
- Protinfo
  http://protinfo.compbio.washington.edu

Protein Sequence Analysis

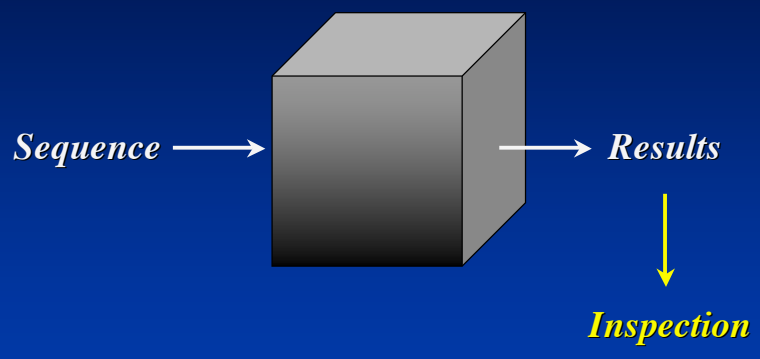


Annual *NAR* Database Issue

http://nar.oupjournals.org



Understanding Analyses



A User's Guide to the Human Genome II

http://www.nature.com/ng/supplements/

Commentary: Keeping Biology in Mind