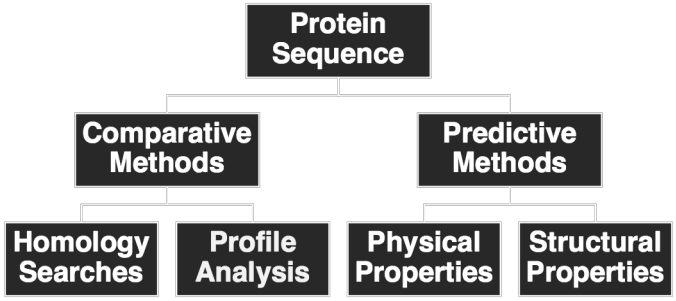
Overview

- Week 2: Comparative methods and concepts
  - Similarity *vs*. Homology
  - Global *vs*. Local Alignments
  - Scoring Matrices
  - BLAST
  - BLAT
- Week 3: 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" *Pfam, InterPro, 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

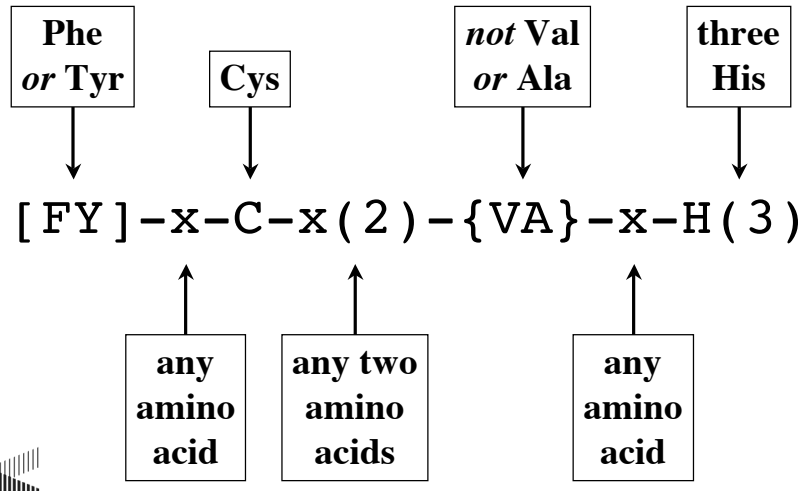
```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
```

- *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~~	~~-6~~	~~-15~~	~~-6~~	~~-6~~	~~-12~~	~~-13~~	~~-6~~	~~-6~~	~~-5~~	~~-5~~	~~-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~~	~~11~~	~~19~~	~~11~~	~~9~~	~~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

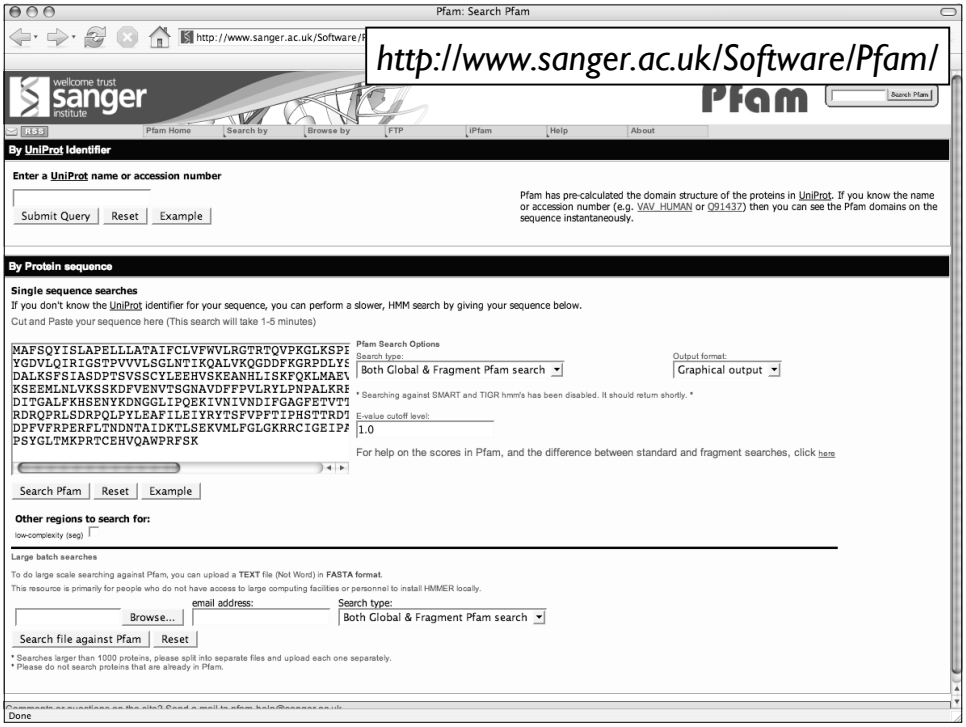


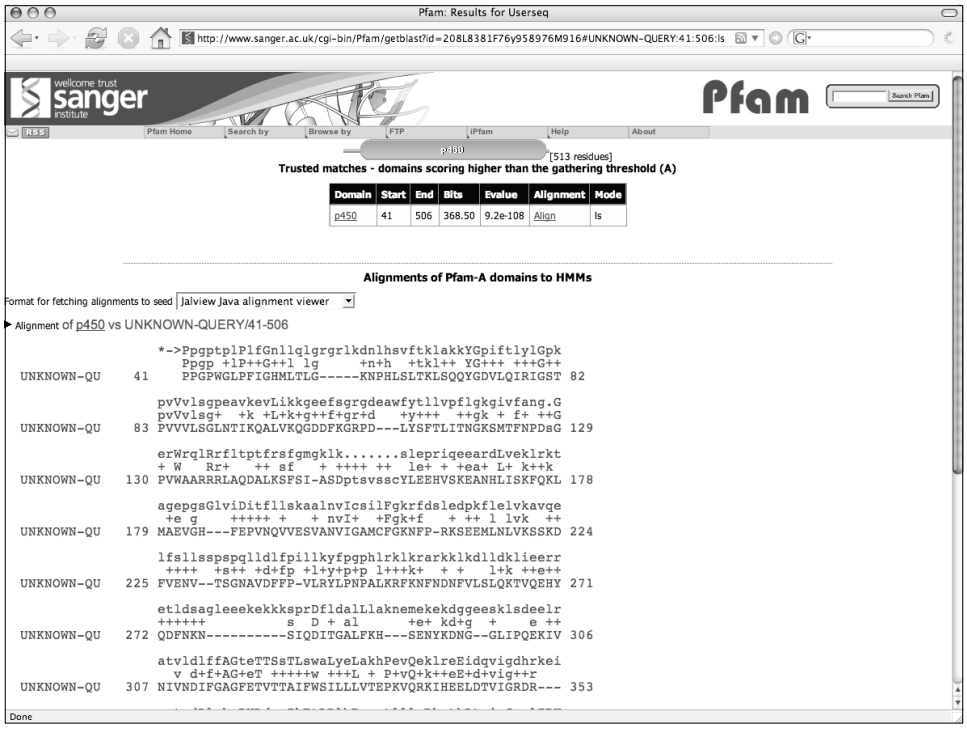
Pfam

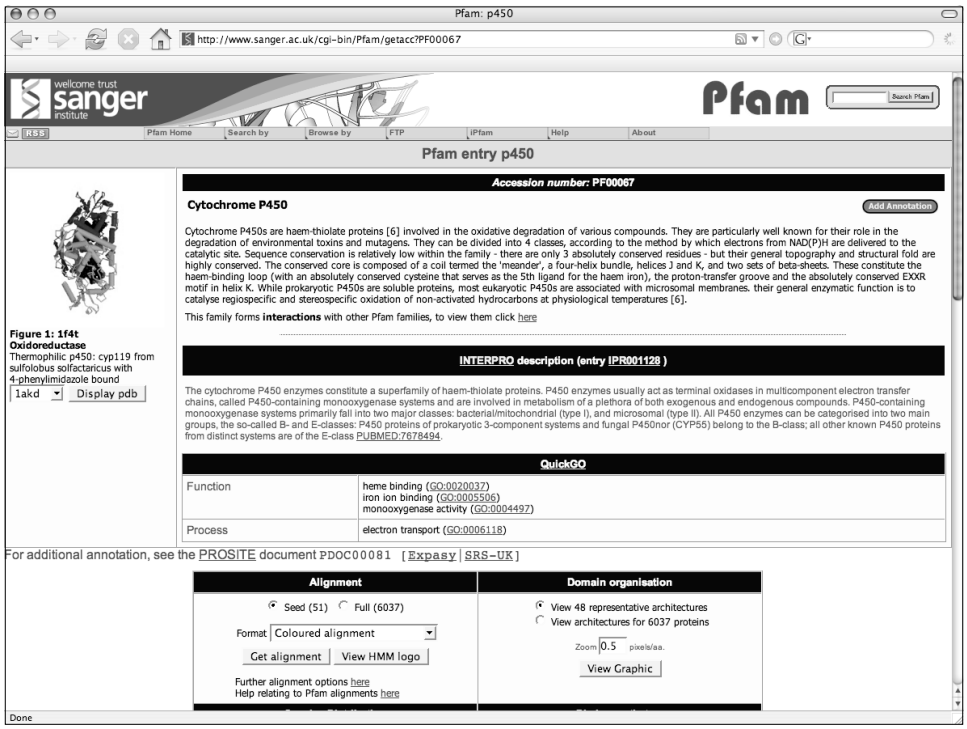
- Collection of multiple alignments of protein domains and conserved protein regions (regions which probably have structural or functional importance)
- Each Pfam entry contains:
  - Multiple sequence alignment of family members
  - Protein domain architectures
  - Species distribution of family members
  - Information on known protein structures
  - Links to other protein family databases

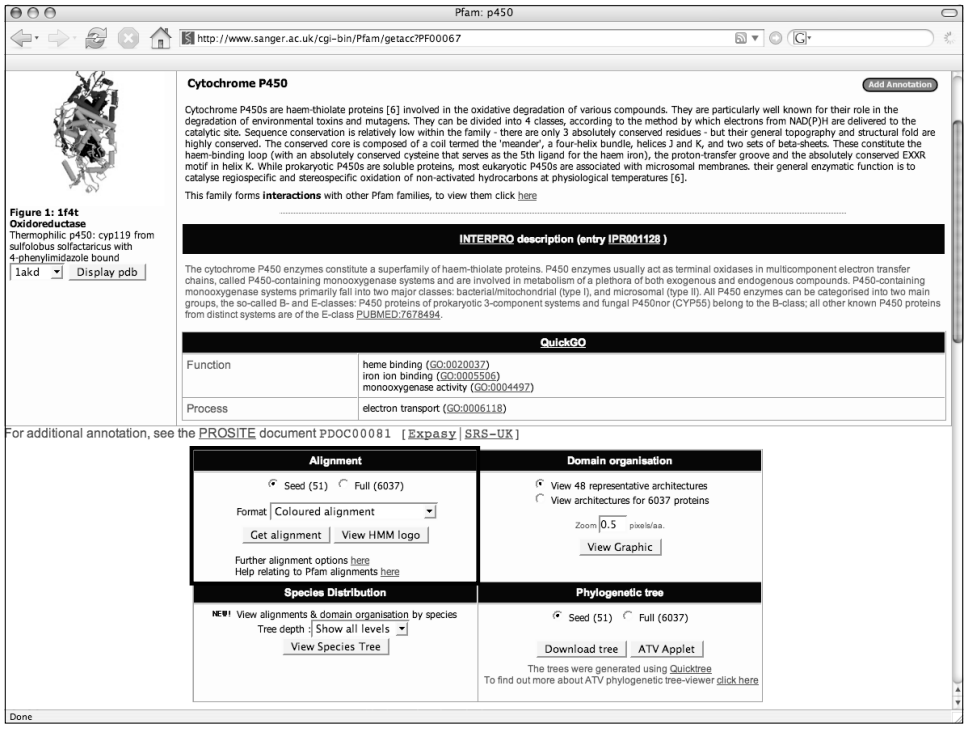
Pfam

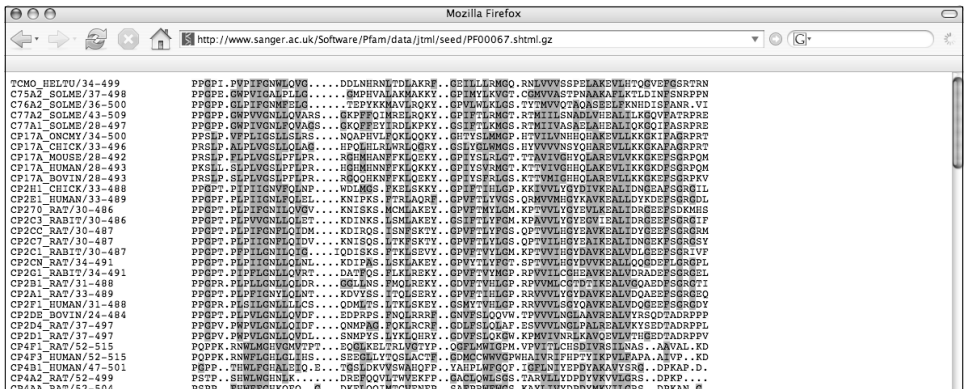
- Pfam A
  - Based on curated multiple alignments
  - Given the method used to construct the alignments, hits are highly likely to be true positives
  - > 74% of all known protein sequences have at least one match to Pfam
- Pfam B
  - Large number of small families taken from the PRODOM database; these families do not overlap with PfamA
  - Deemed "lower quality", but can be useful when no Pfam A family is identified

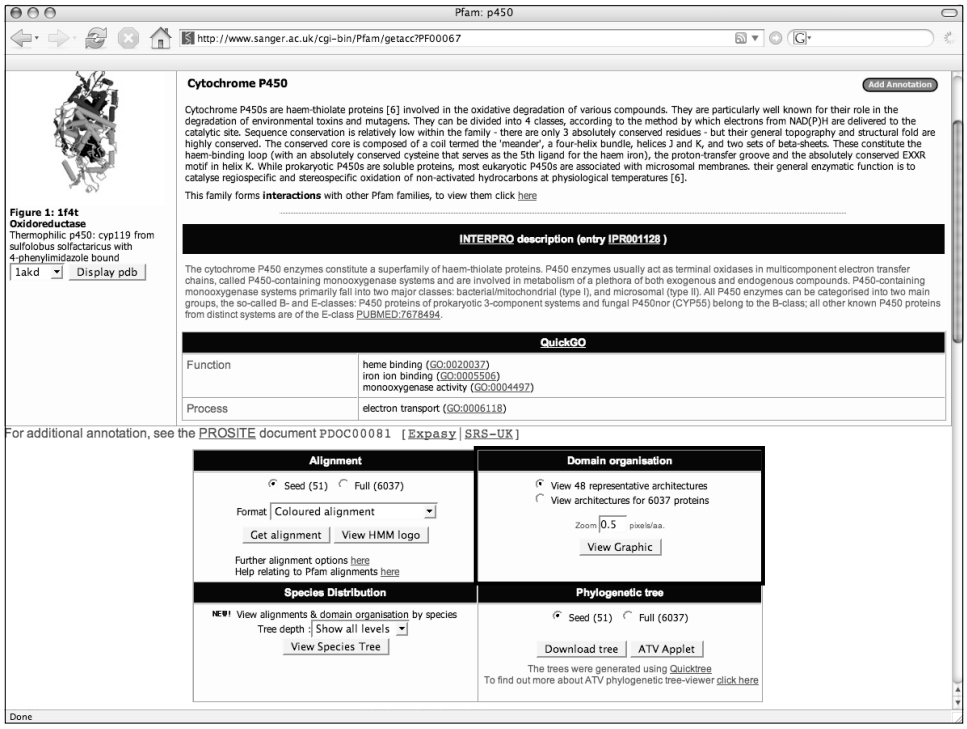


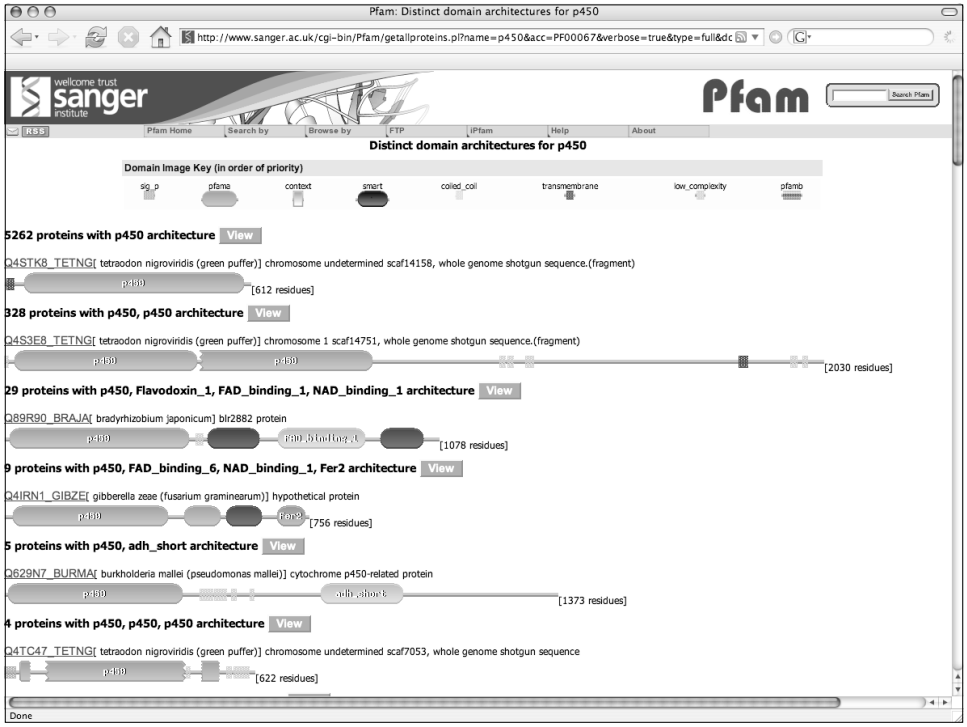

Pfam: Results for Userseq
http://www.sanger.ac.uk/cgi-bin/Pfam/getblast?id=208L8381F76y958976M916#UNKNOWN-QUERY:41:506:ls
wellcome trust sanger institute
Pfam
Pfam Home
Search by
Browse by
FTP
iPfam
Help
About
p450
[513 residues]
Trusted matches - domains scoring higher than the gathering threshold (A)
Domain
Start
End
Bits
Evalue
Alignment
Mode
p450
41
506
368.50
9.2e-108
Align
ls
Alignments of Pfam-A domains to HMMs
Format for fetching alignments to seed
Jalview Java alignment viewer
Alignment of p450 vs UNKNOWN-QUERY/41-506
UNKNOWN-QU 41 PPGPWGLPFIGHMLTLG-----KNPHLSLTKLSQQYGDVLQIRIGST 82
UNKNOWN-QU 83 PVVVLSGLNTIKQALVKQGDDFKGRPD---LYSFTLITNGKSMTFNPDsG 129
UNKNOWN-QU 130 PVWAARRRLAQDALKSFSI-ASDptsvsscYLEEHVSKEANHLISKFQKL 178
UNKNOWN-QU 179 MAEVGH---FEPVNQVVESVANVIGAMCFGKNFP-RKSEEMLNLVKSSKD 224
UNKNOWN-QU 225 FVENV--TSGNAVDFFP-VLRYLPNPALKRFKNFNDNFVLSLQKTVQEHY 271
UNKNOWN-QU 272 QDFNKN----------SIQDITGALFKH---SENYKDNG--GLIPQEKIV 306
UNKNOWN-QU 307 NIVNDIFGAGFETVTTAIFWSILLLVTEPKVQRKIHEELDTVIGRDR--- 353
Done

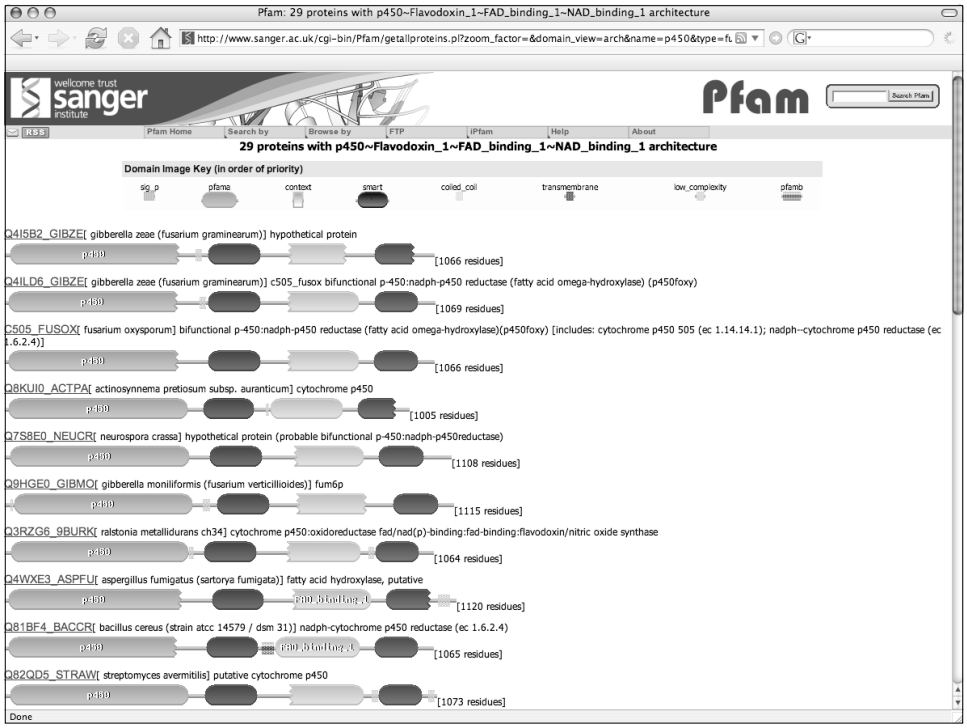

Pfam: p450
http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067
wellcome trust sanger institute
Pfam
Pfam Home
Search by
Browse by
FTP
iPfam
Help
About
Pfam entry p450
Accession number: PF00067
Cytochrome P450
Add Annotation
Cytochrome P450s are haem-thiolate proteins [6] involved in the oxidative degradation of various compounds. They are particularly well known for their role in the degradation of environmental toxins and mutagens. They can be divided into 4 classes, according to the method by which electrons from NAD(P)H are delivered to the catalytic site. Sequence conservation is relatively low within the family - there are only 3 absolutely conserved residues - but their general topography and structural fold are highly conserved. The conserved core is composed of a coil termed the 'meander', a four-helix bundle, helices J and K, and two sets of beta-sheets. These constitute the haem-binding loop (with an absolutely conserved cysteine that serves as the 5th ligand for the haem iron), the proton-transfer groove and the absolutely conserved EXXR motif in helix K. While prokaryotic P450s are soluble proteins, most eukaryotic P450s are associated with microsomal membranes. their general enzymatic function is to catalyse regiospecific and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].
This family forms interactions with other Pfam families, to view them click here
Figure 1: 1f4t
Oxidoreductase
Thermophilic p450: cyp119 from sulfolobus solfactaricus with 4-phenylimidazole bound
1akd
Display pdb
INTERPRO description (entry IPR001128)
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 PUBMED:7678494.
QuickGO
Function
heme binding (GO:0020037)
iron ion binding (GO:0005506)
monooxygenase activity (GO:0004497)
Process
electron transport (GO:0006118)
For additional annotation, see the PROSITE document PDOC00081 [Expasy | SRS-UK]
Alignment
Seed (51)
Full (6037)
Format
Coloured alignment
Get alignment
View HMM logo
Further alignment options here
Help relating to Pfam alignments here
Domain organisation
View 48 representative architectures
View architectures for 6037 proteins
Zoom 0.5 pixels/aa.
View Graphic
Done

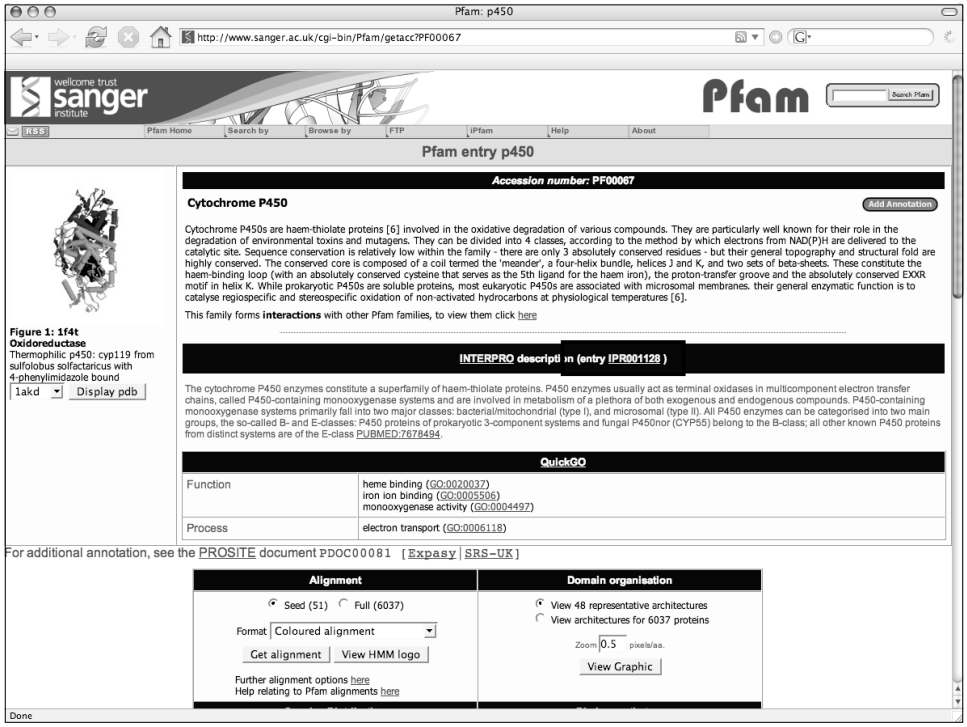

Pfam: p450
http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067
Cytochrome P450
Add Annotation
Cytochrome P450s are haem-thiolate proteins [6] involved in the oxidative degradation of various compounds. They are particularly well known for their role in the degradation of environmental toxins and mutagens. They can be divided into 4 classes, according to the method by which electrons from NAD(P)H are delivered to the catalytic site. Sequence conservation is relatively low within the family - there are only 3 absolutely conserved residues - but their general topography and structural fold are highly conserved. The conserved core is composed of a coil termed the 'meander', a four-helix bundle, helices J and K, and two sets of beta-sheets. These constitute the haem-binding loop (with an absolutely conserved cysteine that serves as the 5th ligand for the haem iron), the proton-transfer groove and the absolutely conserved EXXR motif in helix K. While prokaryotic P450s are soluble proteins, most eukaryotic P450s are associated with microsomal membranes. their general enzymatic function is to catalyse regiospecific and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].
This family forms interactions with other Pfam families, to view them click here
Figure 1: 1f4t
Oxidoreductase
Thermophilic p450: cyp119 from sulfolobus solfactaricus with 4-phenylimidazole bound
1akd
Display pdb
INTERPRO description (entry IPR001128)
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 PUBMED:7678494.
QuickGO
Function
heme binding (GO:0020037)
iron ion binding (GO:0005506)
monooxygenase activity (GO:0004497)
Process
electron transport (GO:0006118)
For additional annotation, see the PROSITE document PDOC00081 [Expasy | SRS-UK]
Alignment
Seed (51) Full (6037)
Format Coloured alignment
Get alignment View HMM logo
Further alignment options here
Help relating to Pfam alignments here
Domain organisation
View 48 representative architectures
View architectures for 6037 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 (51) Full (6037)
Download tree ATV Applet
The trees were generated using Quicktree
To find out more about ATV phylogenetic tree-viewer click here
Done

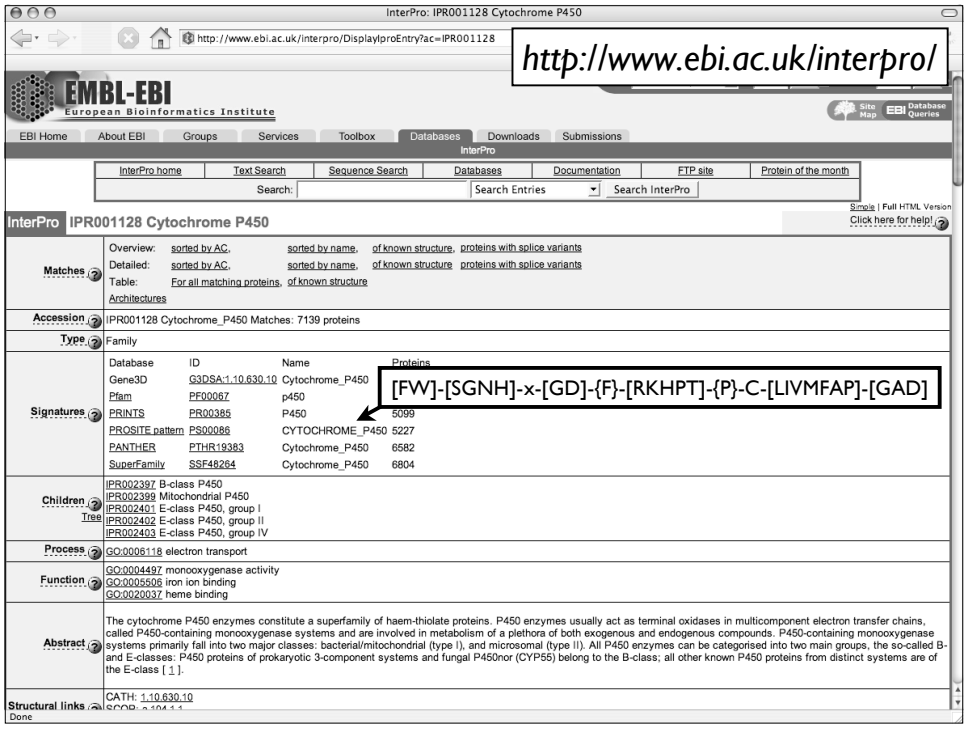

Mozilla Firefox
http://www.sanger.ac.uk/Software/Pfam/data/jtml/seed/PF00067.shtml.gz
TCMO_HELTU/34-499 PPGPI.PWPIFGNWLQVG.....DDLNHRNLTDLAKRF..GSILLLRMGQ.RNLVVVSSPELAKEVLHTQDVEFGSRTRN
C75A2_SOLME/37-498 PPGPP.GWPVIGALPLLG.....EMPHVALAKMAKKY..GPIMYLKVGT.CSMVVASTPNAAKAFLKTLDINFSNRPPN
C76A2_SOLME/36-500 PPGPP.GLPIFGNMFELG.....TEPYKKMAVLRQKY..GPVLWLKLGS.TYTMVVQTAQASEELFKNHDISFANR.VI
C77A2_SOLME/43-509 PPGPP.GWPVVGNLLQVARS...GKPFFQIMRELRQKY..GPIFTLRMGT.RTMIILSNADLVHEALILKEQVFATRPRE
C77A1_SOLME/28-497 PPGPP.GWPIVGNLPQVAGS...GKQFFSYIRDLKPKY..GSIFTLKMGS.RTMIIVASAELAHEALIQKGQIFASRPRE
CP17A_ONCMY/34-500 PPSEP.VFPLIGSLFSRS....NQAPHVLFQKLQQKY..GHTYSLWMGP.HTVILVNHHQAAEVLLKKGKIFAGRPRT
CP17A_CHICK/33-496 PRSLP.ALPLVGSLLQLAG....HPQLHLRLWRLQQRY..GSLYGLWMGS.HYVVVVNSYQHAREVLLKKGKAFAGRPRT
CP17A_MOUSE/28-492 PRSLP.FLPLVGSLPFLPR....RGHMHANFFKLQEKY..GPIYSLRLGT.TTAVIVGHYQLAREVLVKKGKEFSGRPQM
CP17A_HUMAN/28-493 PKSLL.SLPLVGSLPFLPR....HGHMHNNFFKLQKKY..GPIYSVRMGT.KTTVIVGHHQLAKEVLIKKGKDFSGRPQM
CP17A_BOVIN/28-493 PRSLP.SLPLVGSLPFLPR....RGQGHKNFFKLQEKY..GPIYSFRLGS.KTTVMIGHHQLAREVLLKKGKEFSGRPKV
CP2H1_CHICK/33-488 PPGPT.PIPIIGNVFQLNP....WDLMGS.FKELSKKY..GPIFTIHLGP.KKIVVLYGYDIVKEALIDNGEAFSGRGIL
CP2E1_HUMAN/33-489 PPGPF.PLPIIGNLFQLEL....KNIPKS.FTRLAQRF..GPVFTLYVGS.QRMVVMHGYKAVKEALLDYKDEFSGRGDL
CP270_RAT/30-486 PPGPT.PLPIFGNILQVGV....KNISKS.MCMLAKEY..GPVFTMYLGM.KPTVVLYGYEVLKEALIDRGEEFSDKMHS
CP2C3_RABIT/30-486 PPGPT.PLPVVGNLLQLET....KDINKS.LSMLAKEY..GSIFTLYFGM.KPAVVLYGYEVIREALIDRGEEFSGRGIF
CP2CC_RAT/30-487 PPGPT.PLPIFGNFLQIDM....KDIRQS.ISNFSKTY..GPVFTLYFGS.QPTVVLHGYEAVKEALIDYGEEFSGRGRM
CP2C7_RAT/30-487 PPGPT.PLPIIGNFLQIDV....KNISQS.LTKFSKTY..GPVFTLYLGS.QPTVILHGYEAIKEALIDNGEKFSGRGSY
CP2C1_RABIT/30-487 PPGPT.PFPIIGNILQIG....IQDISKS.FTKLSEVV..GPVFTYVLGM.KPTVVIHGYDAVKEALVDLGEEFSGRIVF
CP2CN_RAT/34-491 PPGPT.PLPIIGNLLQLNL....KDIPAS.LSKLAKEY..GPVFTLYFGT.SPTVVLHGYDVVKEALLDQGDEFLGRGPL
CP2C1_RABIT/34-491 PPGPT.PIPFLGNLLQVRT....DATFQS.PLKLREKY..GPVFTVYMGP.RPVVILCGHEAVKEALVDRADEFSGRGEL
CP2B1_RAT/31-488 PPGPR.PLPLLGNLLQLDR....GGLLNS.FMQLREKY..GDVFTVHLGP.RPVVMLCGTDTIKEALVQQAEDFSGRGTI
CP2A1_RAT/33-489 PPGPT.PLPFIGNYLQLNT....KDVYSS.TQLSSRV..GPVFTIHLGP.RNVVVLYGYDAVKEALVDQAEEFSGRGEQ
CP2F1_HUMAN/31-488 PPGPR.PLSLGNLLLLCS....QDMETS.LTKLSKEY..GSMYTVHLGP.RNVVVLSGYQAVKEALVDQGEEFSGRGDY
CP2DE_BOVIN/24-484 PPGPT.PLPVLGNLLQVDF....EDPRPS.FNQLRRRF..GNVFSLQVVW.TPVVVLNGLAAVREALVRSQDTADRPPP
CP2D4_RAT/37-497 PPGPV.PWPVLGNLLQIDP....QNMPAG.FQKLRCRF..GDLFSLQLAF.DSVVVLNGLPALREALVKYSEDTADRPPL
CP2D1_RAT/37-497 PPGPV.PWPVLGNLLQVDL....SNMPYS.LYKLQHRY..GDVFSLQKGW.KPMVIVNRLKAVQEVLVTHGEDTADRPPV
CP4F1_RAT/52-515 PQPPK.RNWLMGHVMVTPT....EQEKKLTRLMQTYP..QSFDWFGPM.VPVITICHSDIVRSILNAS..ANVAL.KD
CP4F3_HUMAN/52-515 PQPPK.RNWFLGLLGLHHS....SEEGLLVTQSLACTF..GDMCCWWVGPWHAIVRIFHPTYIKPVLFAPA.AIVP..KD
CP4B1_HUMAN/47-501 PGPP..THWLFGHALEIQ.E...TGSLDKVVSWAHQFP..YAHPLWFGQF.IGFLNIYEPDYAKAVYSRG..DPKAP.D.
CP4A2_RAT/52-499 PSTP..SHWLWGHNLK.......DREFQQVLTWVEKFP..GACLQWLSGS.TARVLLYDPDYVKVVLGRS..DPKP....
CP4AA_RAT/52-504 PSPP..EHWFPGHKQV..G...DKELQQINTCVENFP..SAFPHWFWGS.KAYLLVYDPDYMKVILGRS..DPKAN.G.
CP4C1_BLADI/37-502 LPGPT.AYPVVGNAIEAIVP..RNKLPQVFDRRAKLY..GPLYRIWAGP.IAQVGLTRPEHVELLIRDT.....KHIDK
CP4D1_DROME/34-507 MPGPP.VLPLVGHHFIQK..PPHEMVKKIFEFMETYSKDQVLKVWLGP.ELNVLMGNPKDVEVVLGTLR...FN..DK
CP6B1_PAPPO/31-495 VAGPK.PVPFFGNLKDSVLR...KKPQVMVYKSIYDEFPNSKVVGIYRMT.TPSVLLRDLDLIKHVLIKDFDSFADRGVE
CP6A1_MUSDO/35-500 PHES..PHLVWGNVKQRGSKVHIGSIIADYYSKEKQS...DPLDGIFLGH.KPAAVVLDKELRRVLLKD..FSNFA.NR
CP6A2_DROME/32-502 PHDA..PHPLYGNMVGFRKNRVMHDFFYDYYNKYRKSG..PPFVGFYFLH.KPAAPIVDTQLAKNLIKD..FSNFA.DR
CP3A3_HUMAN/38-493 PGPT..PLPFLGNILSYHKG......FCMFDMECHKKY..GKVWGFYDGQ.QPVLAITDPDMIKLVLVKEC.YSVFT.NR
CP3A6_RABIT/37-491 PGPT..PLPFIGTILSYRKG......IWDFDIECRKKY..GKMWGLFDGR.QPLMVITDPDMIKTVLVKEC.YSVFT.NR
CP3A1_RAT/39-494 PGPX..PLPFPGTVNYVWG.....LWKPDVDFGHKFY..GKIWGLFDGQ.MPLAITDTDMIKNVLVKEC.FSVFT.NR
CP19A_MOUSE/48-488 PGPG....YCLGIGPLISHGRFLWMGIGSACNYYNKMY..GSFMAVWISG.EETLIISKSSSMFHVMKHSH.YISRF.GS
CP19A_CHICK/47-487 PGPG....YCMGIGPLISHGRFLWMGVGNACNYYNKTY..GSFVAVWISG.EETFIISKSSSVFHVMKHWN.YVSRF.GS
C11B2_MOUSE/42-496 PQYSR.NKWLKMIQLLREQG...QENLHLEMHQVFREL..GPIFRHSVGK.TQIVSVMLPEDAEKLHQVDSMLPRMHLE
C11B1_BOVIN/42-499 PRCPQ.NKWNMQLQEWKSQS....SENNHLDMHQFFQEL..GPIFPYSVGG.RHMVFVMLPEDVERLQQADSHIPQRMILE
CP11A_ONCMY/50-507 PGLWR.NGLANLYSFWKLDG...FRNIHRVMVHNFNTF..GPIYREKIGY.YDSVNIIKPEMKPALLFKAEGHYPKRLTVE
CP11A_HUMAN/52-511 PSPGD.NGWLNLYHFWRETG...THKVHLHHVQNPQKY..GPIYREKLGN.VESVYVIDPEDVALLFKSEGPNPERFLIP
CP11A_BOVIN/52-510 PSPGD.NGWLNLYHFWREKG...SQRIHFRHIENPQKY..GPIYREKLGN.LESVYIIHPEDVAHLFKFEGSYPERYDIP
CP27A_HUMAN/61-526 PRLGQ.LPFPQLFVQG.Y.....ALQLHQLQVDKFKAKY..GPMWMSYLGP.QMHVNLASPELFQVMNQEGKYPVRNDMS
PID6_FUSSO/51-503 PGPFLARISRVWEIKKAAT.....GNIHEIVMDLHRCH..GPIVRICPNR.YDFDTMEALKIIYRIGNALP..KADYYIP
CP53_ASPNG/36-509 PAPGLAAFTNFWLLQTRR.....GHREVVVDNAHKKY..GKLVRIAPRH.TSIADDG...AIQAVYGHGNGFLKSDFYD
CP7A1_HUMAN/32-497 PPLENGLIPYLGCALQFG......ANPSEFLRANQRKH..GHVFTCKLMG.KYVHFITNPLSYHKVLCHG...KYFDWKK
CP51_YEAST/57-521 PLVEV.WDPWGSLKSVKG......MKPSEFSELGKYY..GDIFSFELG.RVMTYVLGPKGHEFVFNAKLANVSAEAAY
CPXI_BACME/12-406 PKFQS.RAEEFFIQWYK....EMLNNSPVYEHEET..NTWNVFQYE.HVKQVLSNYDFFSSDGQRTT.IFVG..DN
TCMO_HELTU/34-499 VVFDITT..GKGQDMVFTVY.GCHWRKMRRIMTVPFFTNKVVQQ..YRYGWEAEAAAVVDDVKKNPAAATEG...IVIRR
C75A2_SOLME/37-498 AGKTH..MAYNAQDMVFAPY.GPRWKLLRKLSNLHMLGG....R.AENWANVRAMELGHMLKSMFDASHVG.ERIVMAD
C76A2_SOLME/36-500 PDVNQAHSYYQGSLAIAPYG..PFWRFORRICTEMFVHKK.IS.ETEPVRRKCVDNMLKWIEKEANSAEKG.SGIDVTR
C77A2_SOLME/43-509 NPTRT..VFSCDKFTVNAAVYGPVWRSLRKNMVQNGLSSIRLKE..FRAVRKSAMDKMIEKIRAEADAN.....EGVV
C77A1_SOLME/28-497 NPTRT...IFSCNKFSVNAAVYGPVWRSLRRNMVQNMLSPSRLKE..FREFREIAMDKLISRIRVDAKEN.....NDVV
CP17A_ONCMY/34-500 VFTDLLT..RDGKDIAEADV.GACWRFDKTVHEALCMFGSGSA.SEKKICREALSLCDTFRSFSAS.....LDESP
CP17A_CHICK/33-496 VTTDLLS..RGGKDIAFASY.GPLWKFQRKLVHAALSMPGDGSV.ALEKIICREAASLCETLGAAQDMA.....LDMAP
CP17A_MOUSE/28-492 VTLGLLS..DQGKGVAFADS.SSSWQLFRKLVFSTFSLFRD.DQ.KLEKMICQEANSLCDLILTYDGES.....RDLST

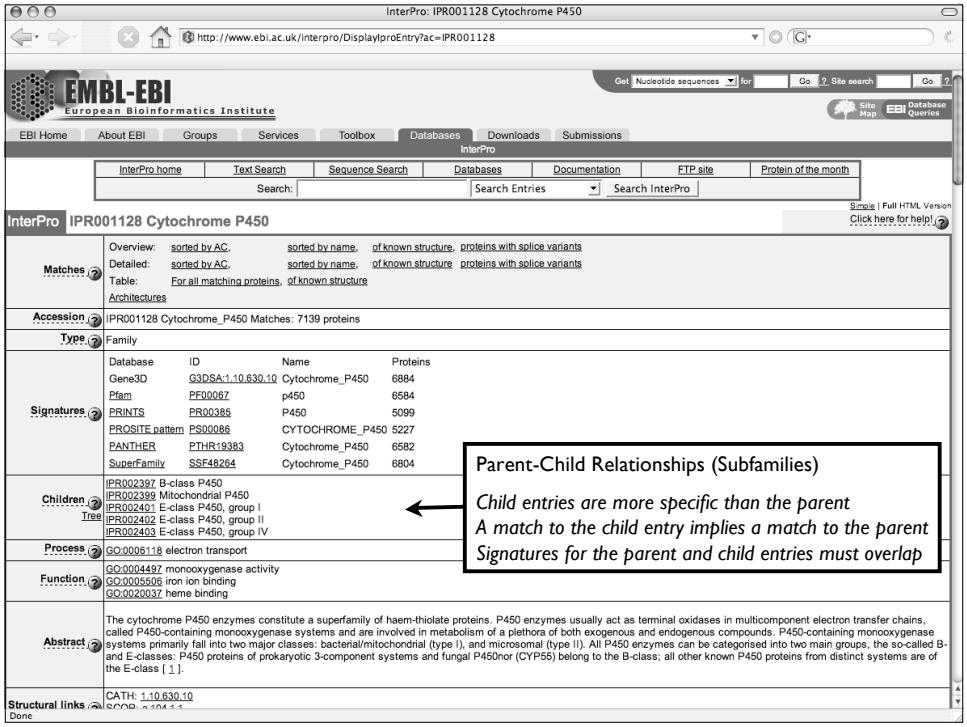

Pfam: p450
http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067
Cytochrome P450
Add Annotation
Cytochrome P450s are haem-thiolate proteins [6] involved in the oxidative degradation of various compounds. They are particularly well known for their role in the degradation of environmental toxins and mutagens. They can be divided into 4 classes, according to the method by which electrons from NAD(P)H are delivered to the catalytic site. Sequence conservation is relatively low within the family - there are only 3 absolutely conserved residues - but their general topography and structural fold are highly conserved. The conserved core is composed of a coil termed the 'meander', a four-helix bundle, helices J and K, and two sets of beta-sheets. These constitute the haem-binding loop (with an absolutely conserved cysteine that serves as the 5th ligand for the haem iron), the proton-transfer groove and the absolutely conserved EXXR motif in helix K. While prokaryotic P450s are soluble proteins, most eukaryotic P450s are associated with microsomal membranes. their general enzymatic function is to catalyse regiospecific and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].
This family forms interactions with other Pfam families, to view them click here
Figure 1: 1f4t
Oxidoreductase
Thermophilic p450: cyp119 from sulfolobus solfactaricus with 4-phenylimidazole bound
1akd
Display pdb
INTERPRO description (entry IPR001128)
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 PUBMED:7678494.
QuickGO
Function
heme binding (GO:0020037)
iron ion binding (GO:0005506)
monooxygenase activity (GO:0004497)
Process
electron transport (GO:0006118)
For additional annotation, see the PROSITE document PDOC00081 [Expasy | SRS-UK]
Alignment
Seed (51) Full (6037)
Format Coloured alignment
Get alignment View HMM logo
Further alignment options here
Help relating to Pfam alignments here
Domain organisation
View 48 representative architectures
View architectures for 6037 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 (51) Full (6037)
Download tree ATV Applet
The trees were generated using Quicktree
To find out more about ATV phylogenetic tree-viewer click here
Done

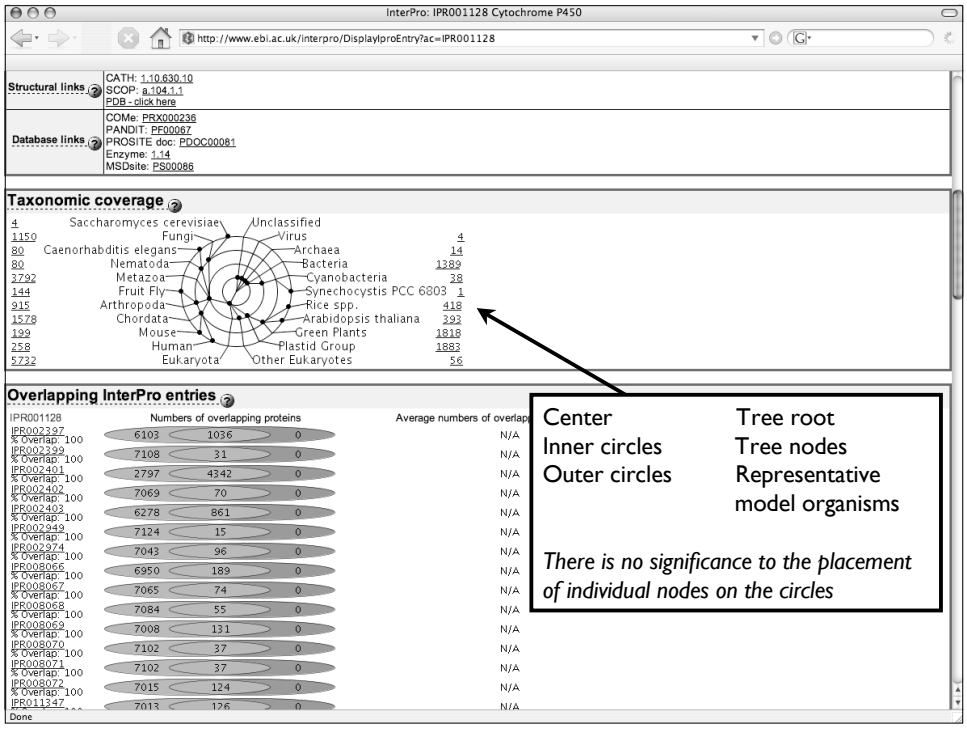

Pfam: Distinct domain architectures for p450
http://www.sanger.ac.uk/cgi-bin/Pfam/getallproteins.pl?name=p450&acc=PF00067&verbose=true&type=full&dc
wellcome trust sanger institute
Pfam
Search Pfam
RSS
Pfam Home
Search by
Browse by
FTP
iPfam
Help
About
Distinct domain architectures for p450
Domain Image Key (in order of priority)
sig_p
pfama
context
smart
coiled_coil
transmembrane
low_complexity
pfamb
5262 proteins with p450 architecture View
Q4STK8_TETNG[tetraodon nigroviridis (green puffer)] chromosome undetermined scaf14158, whole genome shotgun sequence.(fragment)
p450
[612 residues]
328 proteins with p450, p450 architecture View
Q4S3E8_TETNG[tetraodon nigroviridis (green puffer)] chromosome 1 scaf14751, whole genome shotgun sequence.(fragment)
p450
p450
[2030 residues]
29 proteins with p450, Flavodoxin_1, FAD_binding_1, NAD_binding_1 architecture View
Q89R90_BRAJA[bradyrhizobium japonicum] blr2882 protein
p450
FAD_binding_1
[1078 residues]
9 proteins with p450, FAD_binding_6, NAD_binding_1, Fer2 architecture View
Q4IRN1_GIBZE[gibberella zeae (fusarium graminearum)] hypothetical protein
p450
Fer2
[756 residues]
5 proteins with p450, adh_short architecture View
Q629N7_BURMA[burkholderia mallei (pseudomonas mallei)] cytochrome p450-related protein
p450
adh_short
[1373 residues]
4 proteins with p450, p450, p450 architecture View
Q4TC47_TETNG[tetraodon nigroviridis (green puffer)] chromosome undetermined scaf7053, whole genome shotgun sequence
p450
[622 residues]
Done

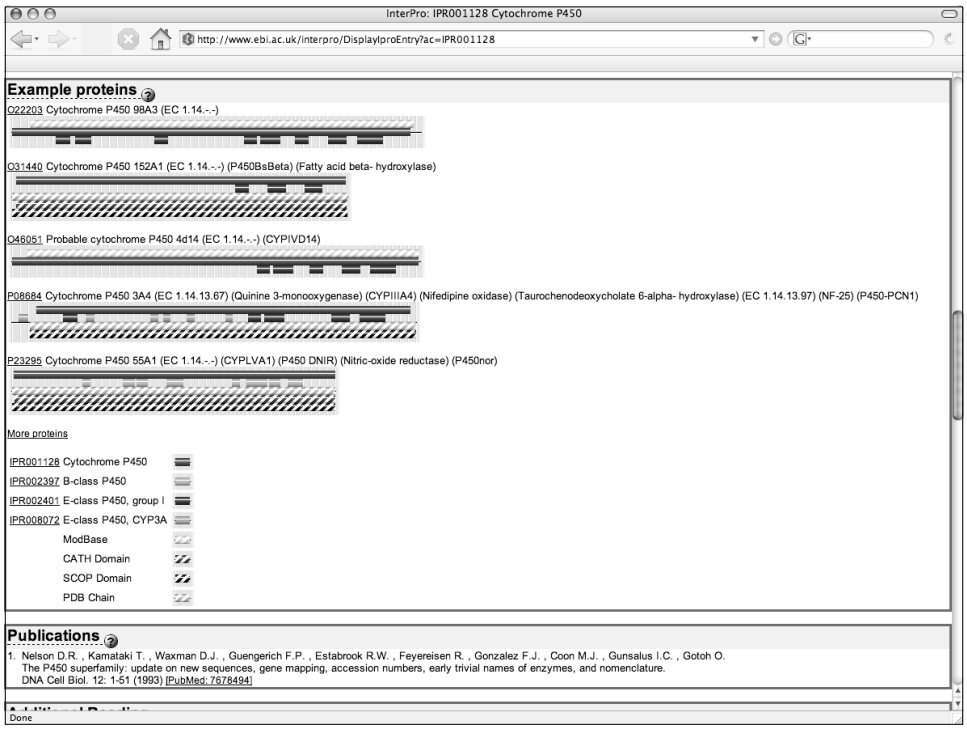

Pfam: 29 proteins with p450~Flavodoxin_1~FAD_binding_1~NAD_binding_1 architecture
http://www.sanger.ac.uk/cgi-bin/Pfam/getallproteins.pl?zoom_factor=&domain_view=arch&name=p450&type=fl
wellcome trust sanger institute
Pfam
Search Pfam
Pfam Home
Search by
Browse by
FTP
iPfam
Help
About
29 proteins with p450~Flavodoxin_1~FAD_binding_1~NAD_binding_1 architecture
Domain Image Key (in order of priority)
sig_p
pfama
context
smart
coiled_coil
transmembrane
low_complexity
pfamb
Q4I5B2_GIBZE[gibberella zeae (fusarium graminearum)] hypothetical protein
[1066 residues]
Q4ILD6_GIBZE[gibberella zeae (fusarium graminearum)] c505_fusox bifunctional p-450:nadph-p450 reductase (fatty acid omega-hydroxylase) (p450foxy)
[1069 residues]
C505_FUSOX[fusarium oxysporum] bifunctional p-450:nadph-p450 reductase (fatty acid omega-hydroxylase)(p450foxy) [includes: cytochrome p450 505 (ec 1.14.14.1); nadph--cytochrome p450 reductase (ec 1.6.2.4)]
[1066 residues]
Q8KUI0_ACTPA[actinosynnema pretiosum subsp. auranticum] cytochrome p450
[1005 residues]
Q7S8E0_NEUCR[neurospora crassa] hypothetical protein (probable bifunctional p-450:nadph-p450reductase)
[1108 residues]
Q9HGE0_GIBMO[gibberella moniliformis (fusarium verticillioides)] fum6p
[1115 residues]
Q3RZG6_9BURK[ralstonia metallidurans ch34] cytochrome p450:oxidoreductase fad/nad(p)-binding:fad-binding:flavodoxin/nitric oxide synthase
[1064 residues]
Q4WXE3_ASPFU[aspergillus fumigatus (sartorya fumigata)] fatty acid hydroxylase, putative
[1120 residues]
Q81BF4_BACCR[bacillus cereus (strain atcc 14579 / dsm 31)] nadph-cytochrome p450 reductase (ec 1.6.2.4)
[1065 residues]
Q82QD5_STRAW[streptomyces avermitilis] putative cytochrome p450
[1073 residues]
Done


Pfam: p450
http://www.sanger.ac.uk/cgi-bin/Pfam/getacc?PF00067
wellcome trust sanger institute
Pfam
Search Pfam
Pfam Home
Search by
Browse by
FTP
iPfam
Help
About
Pfam entry p450
Accession number: PF00067
Cytochrome P450
Add Annotation
Cytochrome P450s are haem-thiolate proteins [6] involved in the oxidative degradation of various compounds. They are particularly well known for their role in the degradation of environmental toxins and mutagens. They can be divided into 4 classes, according to the method by which electrons from NAD(P)H are delivered to the catalytic site. Sequence conservation is relatively low within the family - there are only 3 absolutely conserved residues - but their general topography and structural fold are highly conserved. The conserved core is composed of a coil termed the 'meander', a four-helix bundle, helices J and K, and two sets of beta-sheets. These constitute the haem-binding loop (with an absolutely conserved cysteine that serves as the 5th ligand for the haem iron), the proton-transfer groove and the absolutely conserved EXXR motif in helix K. While prokaryotic P450s are soluble proteins, most eukaryotic P450s are associated with microsomal membranes. their general enzymatic function is to catalyse regiospecific and stereospecific oxidation of non-activated hydrocarbons at physiological temperatures [6].
This family forms interactions with other Pfam families, to view them click here
Figure 1: 1f4t
Oxidoreductase
Thermophilic p450: cyp119 from sulfolobus solfactaricus with 4-phenylimidazole bound
1akd
Display pdb
INTERPRO descripti on (entry IPR001128)
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 PUBMED:7678494.
QuickGO
Function
heme binding (GO:0020037)
iron ion binding (GO:0005506)
monooxygenase activity (GO:0004497)
Process
electron transport (GO:0006118)
For additional annotation, see the PROSITE document PDOC00081 [Expasy | SRS-UK]
Alignment
Seed (51) Full (6037)
Format Coloured alignment
Get alignment
View HMM logo
Further alignment options here
Help relating to Pfam alignments here
Domain organisation
View 48 representative architectures
View architectures for 6037 proteins
Zoom 0.5 pixels/aa.
View Graphic
Done


InterPro: IPR001128 Cytochrome P450
http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128
http://www.ebi.ac.uk/interpro/
EMBL-EBI
European Bioinformatics Institute
EBI Home
About EBI
Groups
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 IPR001128 Cytochrome P450
Matches
Accession
IPR001128 Cytochrome_P450 Matches: 7139 proteins
Type
Family
Signatures
Database ID Name Proteins
Gene3D G3DSA:1.10.630.10 Cytochrome_P450
Pfam PF00067 p450
PRINTS PR00385 P450 5099
PROSITE pattern PS00086 CYTOCHROME_P450 5227
PANTHER PTHR19383 Cytochrome_P450 6582
SuperFamily SSF48264 Cytochrome_P450 6804
[FW]-[SGNH]-x-[GD]-{F}-[RKHPT]-{P}-C-[LIVMFAP]-[GAD]
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
GO:0006118 electron transport
Function
GO:0004497 monooxygenase activity
GO:0005506 iron ion binding
GO:0020037 heme binding
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
CATH: 1.10.630.10
Done

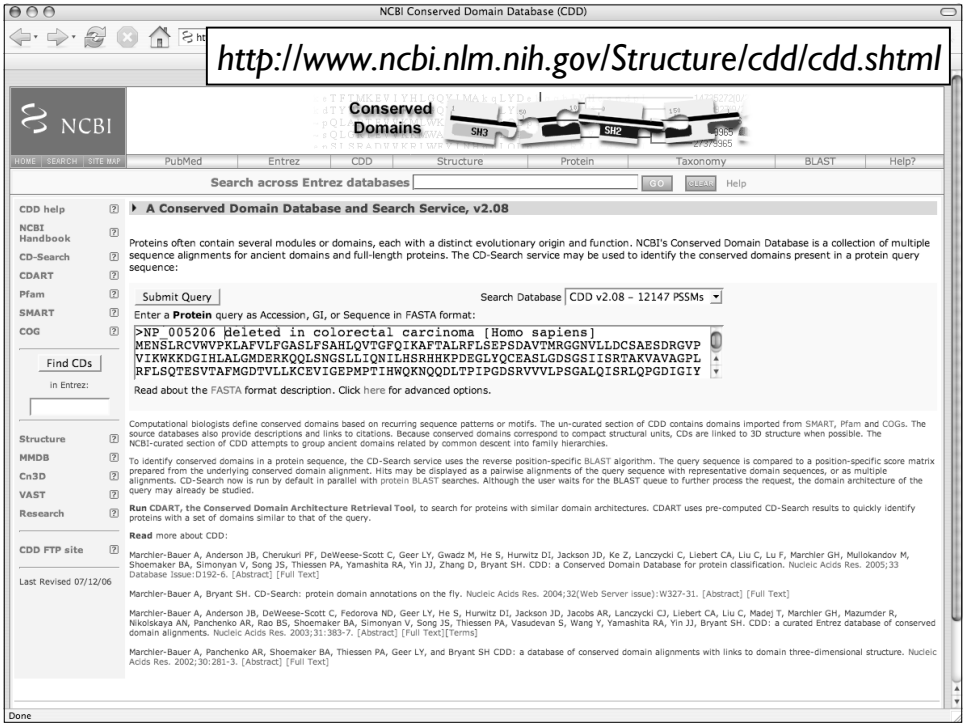

InterPro: IPR001128 Cytochrome P450
http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128
EMBL-EBI
European Bioinformatics Institute
EBI Home
About EBI
Groups
Services
Toolbox
Databases
Downloads
Submissions
InterPro
InterPro home
Text Search
Sequence Search
Databases
Documentation
FTP site
Protein of the month
Search:
Search Entries
Search InterPro
InterPro IPR001128 Cytochrome P450
Matches
Accession
IPR001128 Cytochrome_P450 Matches: 7139 proteins
Type
Family
Signatures
Database ID Name Proteins
Gene3D G3DSA:1.10.630.10 Cytochrome_P450 6884
Pfam PF00067 p450 6584
PRINTS PR00385 P450 5099
PROSITE pattern PS00086 CYTOCHROME_P450 5227
PANTHER PTHR19383 Cytochrome_P450 6582
SuperFamily SSF48264 Cytochrome_P450 6804
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
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
Process
GO:0006118 electron transport
Function
GO:0004497 monooxygenase activity
GO:0005506 iron ion binding
GO:0020037 heme binding
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
CATH: 1.10.630.10
Done

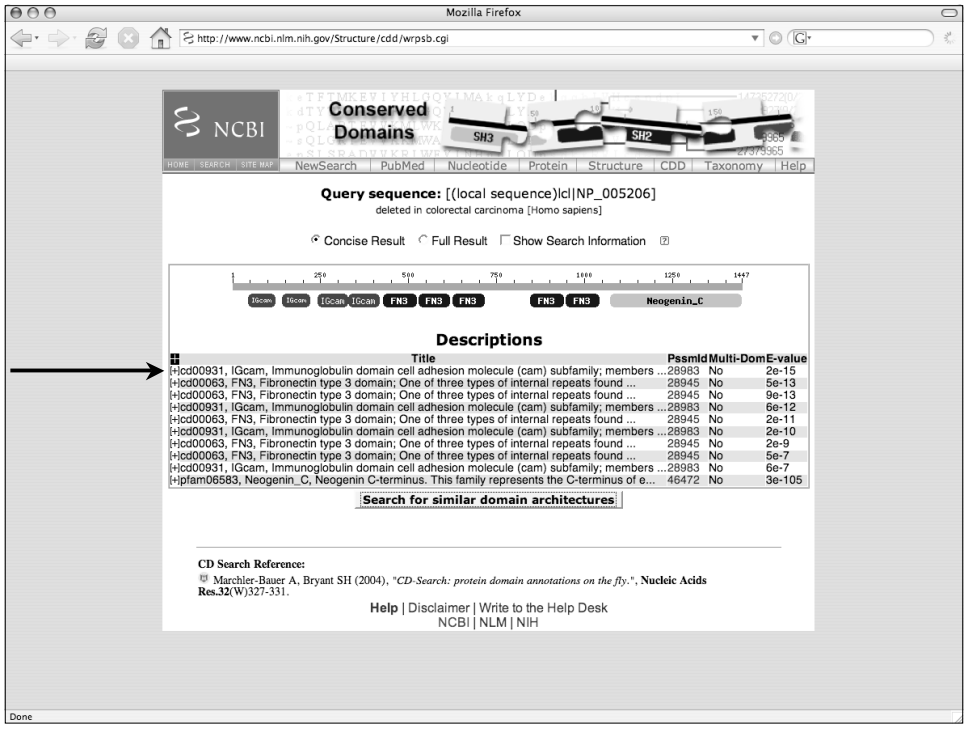

InterPro: IPR001128 Cytochrome P450
http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128
Structural links
CATH: 1.10.630.10
SCOP: a.104.1.1
PDB - click here
Database links
COMe: PRX000236
PANDIT: PF00067
PROSITE doc: PDOC00081
Enzyme: 1.14
MSDsite: PS00086
Taxonomic coverage
Saccharomyces cerevisiae
Fungi
Caenorhabditis elegans
Nematoda
Metazoa
Fruit Fly
Arthropoda
Chordata
Mouse
Human
Eukaryota
Unclassified
Virus
Archaea
Bacteria
Cyanobacteria
Synechocystis PCC 6803
Rice spp.
Arabidopsis thaliana
Green Plants
Plastid Group
Other Eukaryotes
Overlapping InterPro entries
Numbers of overlapping proteins
Average numbers of overlap
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
Done

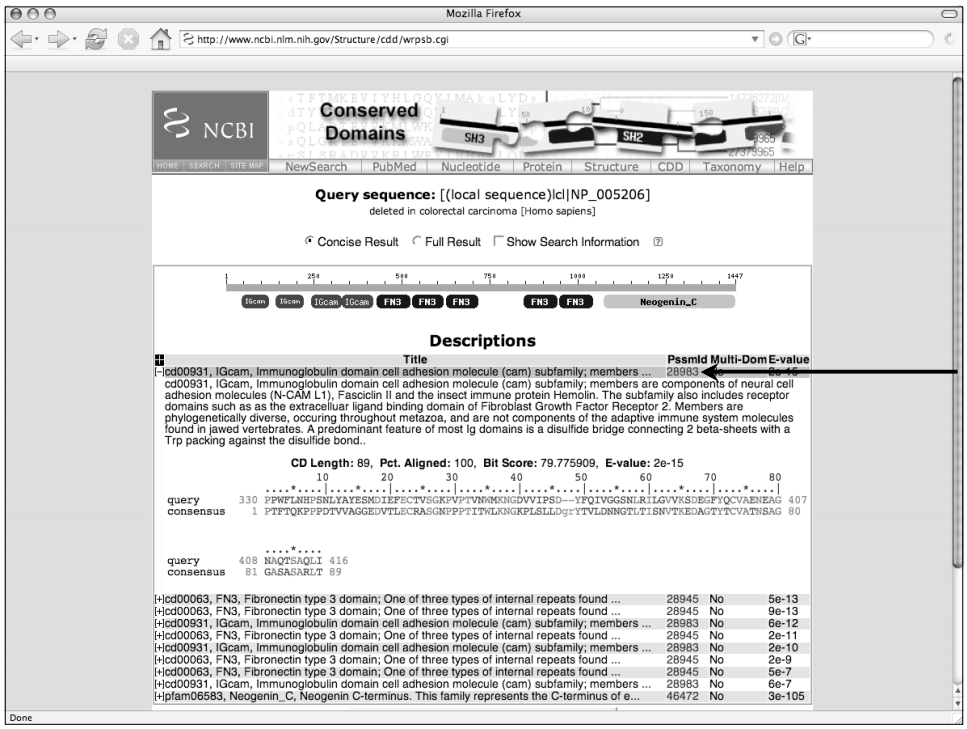

InterPro: IPR001128 Cytochrome P450
http://www.ebi.ac.uk/interpro/DisplayIproEntry?ac=IPR001128
Example proteins
O22203 Cytochrome P450 98A3 (EC 1.14.-.-)
O31440 Cytochrome P450 152A1 (EC 1.14.-.-) (P450BsBeta) (Fatty acid beta- hydroxylase)
O46051 Probable cytochrome P450 4d14 (EC 1.14.-.-) (CYPIVD14)
P08684 Cytochrome P450 3A4 (EC 1.14.13.67) (Quinine 3-monooxygenase) (CYPIIIA4) (Nifedipine oxidase) (Taurochenodeoxycholate 6-alpha- hydroxylase) (EC 1.14.13.97) (NF-25) (P450-PCN1)
P23295 Cytochrome P450 55A1 (EC 1.14.-.-) (CYPLVA1) (P450 DNIR) (Nitric-oxide reductase) (P450nor)
More proteins
IPR001128 Cytochrome P450
IPR002397 B-class P450
IPR002401 E-class P450, group I
IPR008072 E-class P450, CYP3A
ModBase
CATH Domain
SCOP Domain
PDB Chain
Publications
1. Nelson D.R. , Kamataki T. , Waxman D.J. , Guengerich F.P. , Estabrook R.W. , Feyereisen R. , Gonzalez F.J. , Coon M.J. , Gunsalus I.C. , Gotoh O.
The P450 superfamily: update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature.
DNA Cell Biol. 12: 1-51 (1993) [PubMed: 7678494]
Done

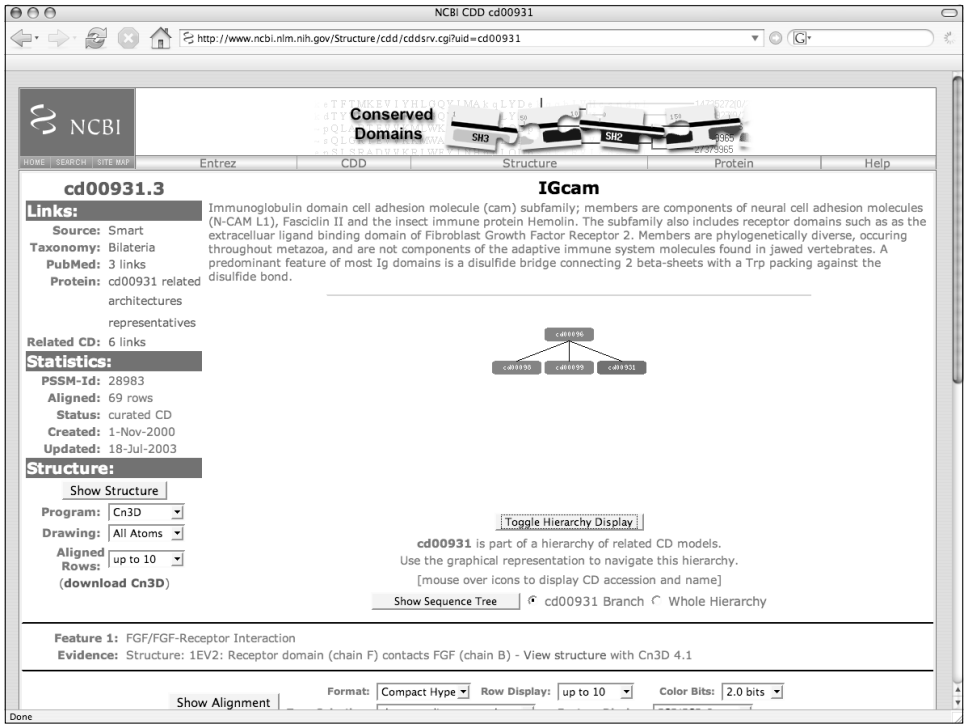
Conserved Domain Database (CDD)

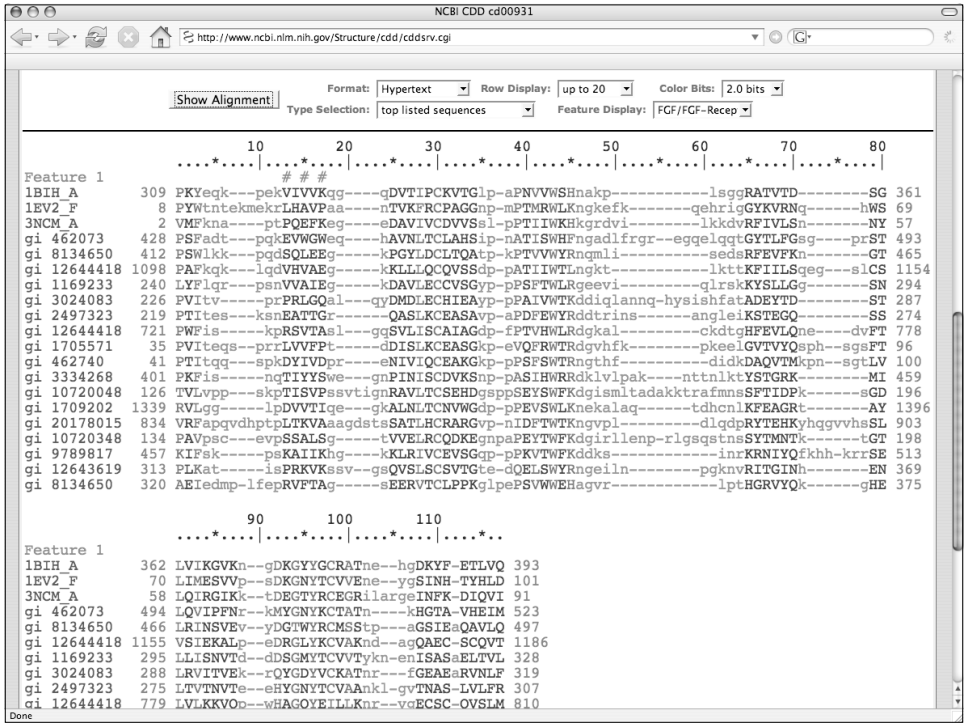
- Identify conserved domains in a protein sequence
- "Secondary database"
  - Pfam A and B
  - Simple Modular Architecture Research Tool (SMART)
  - Clusters of Orthologous Groups
- Search performed using RPS-BLAST
  - Query sequence is used to search a database of precalculated position-specific scoring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



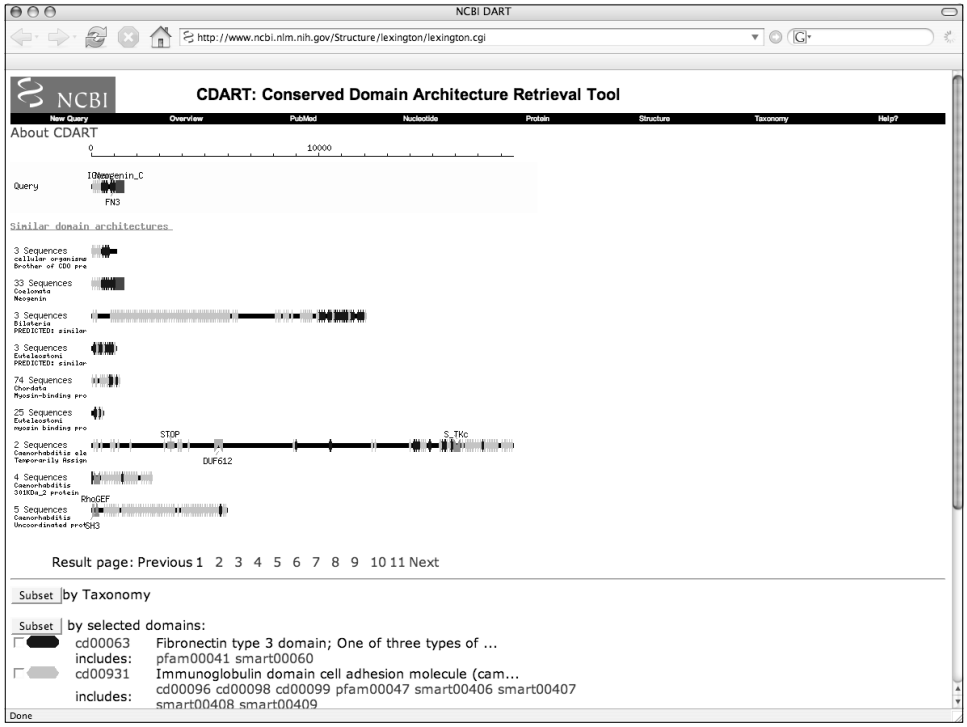

Mozilla Firefox
http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
NCBI
Conserved Domains
SH3
SH2
HOME | SEARCH | SITE MAP
NewSearch | PubMed | Nucleotide | Protein | Structure | CDD | Taxonomy | Help
Query sequence: [(local sequence)lcl|NP_005206]
deleted in colorectal carcinoma [Homo sapiens]
Concise Result
Full Result
Show Search Information
IGcam IGcam IGcam IGcam FN3 FN3 FN3 FN3 FN3 Neogenin_C
Descriptions
Title Pssmid Multi-Dom E-value
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 2e-15
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 5e-13
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 9e-13
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 6e-12
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 2e-11
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 2e-10
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 2e-9
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 5e-7
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 6e-7
[+]pfam06583, Neogenin_C, Neogenin C-terminus. This family represents the C-terminus of e... 46472 No 3e-105
Search for similar domain architectures
CD Search Reference:
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
Done


Mozilla Firefox
http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
NCBI
Conserved Domains
SH3
SH2
HOME | SEARCH | SITE MAP
NewSearch | PubMed | Nucleotide | Protein | Structure | CDD | Taxonomy | Help
Query sequence: [(local sequence)lcl|NP_005206]
deleted in colorectal carcinoma [Homo sapiens]
Concise Result
Full Result
Show Search Information
IGcam IGcam IGcam IGcam FN3 FN3 FN3 FN3 FN3 Neogenin_C
Descriptions
Title Pssmid Multi-Dom E-value
[-]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 2e-15
cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members are components of neural cell adhesion molecules (N-CAM L1), Fasciclin II and the insect immune protein Hemolin. The subfamily also includes receptor domains such as as the extracelluar ligand binding domain of Fibroblast Growth Factor Receptor 2. Members are phylogenetically diverse, occuring throughout metazoa, and are not components of the adaptive immune system molecules found in jawed vertebrates. A predominant feature of most Ig domains is a disulfide bridge connecting 2 beta-sheets with a Trp packing against the disulfide bond..
CD Length: 89, Pct. Aligned: 100, Bit Score: 79.775909, E-value: 2e-15
query 330 PPWFLNHPSNLYAYESMDIEFSCTVSGKPVPTVNWMKNGDVVIPSD--YFQIVGGSNLRILGVVKSDEGFYQCVAENEAG 407
consensus 1 PTFTQKPPPDTVVAGGEDVTLECRASGNPPPTITWLKNGKPLSLLDgrYTVLDNNGTLTISNVTKEDAGTYTCVATNSAG 80
query 408 NAQTSAQLI 416
consensus 81 GASASARLT 89
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 5e-13
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 9e-13
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 6e-12
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 2e-11
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 2e-10
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 2e-9
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 5e-7
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 6e-7
[+]pfam06583, Neogenin_C, Neogenin C-terminus. This family represents the C-terminus of e... 46472 No 3e-105
Done


NCBI CDD cd00931
http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=cd00931
NCBI
Conserved Domains
SH3
SH2
HOME SEARCH SITE MAP
Entrez CDD Structure Protein Help
cd00931.3
IGcam
Links:
Source: Smart
Taxonomy: Bilateria
PubMed: 3 links
Protein: cd00931 related architectures representatives
Related CD: 6 links
Statistics:
PSSM-Id: 28983
Aligned: 69 rows
Status: curated CD
Created: 1-Nov-2000
Updated: 18-Jul-2003
Structure:
Show Structure
Program: Cn3D
Drawing: All Atoms
Aligned Rows: up to 10
(download Cn3D)
Immunoglobulin domain cell adhesion molecule (cam) subfamily; members are components of neural cell adhesion molecules (N-CAM L1), Fasciclin II and the insect immune protein Hemolin. The subfamily also includes receptor domains such as as the extracelluar ligand binding domain of Fibroblast Growth Factor Receptor 2. Members are phylogenetically diverse, occuring throughout metazoa, and are not components of the adaptive immune system molecules found in jawed vertebrates. A predominant feature of most Ig domains is a disulfide bridge connecting 2 beta-sheets with a Trp packing against the disulfide bond.
Toggle Hierarchy Display
cd00931 is part of a hierarchy of related CD models.
Use the graphical representation to navigate this hierarchy.
[mouse over icons to display CD accession and name]
Show Sequence Tree
cd00931 Branch
Whole Hierarchy
Feature 1: FGF/FGF-Receptor Interaction
Evidence: Structure: 1EV2: Receptor domain (chain F) contacts FGF (chain B) - View structure with Cn3D 4.1
Show Alignment
Format: Compact Hype
Row Display: up to 10
Color Bits: 2.0 bits
Done

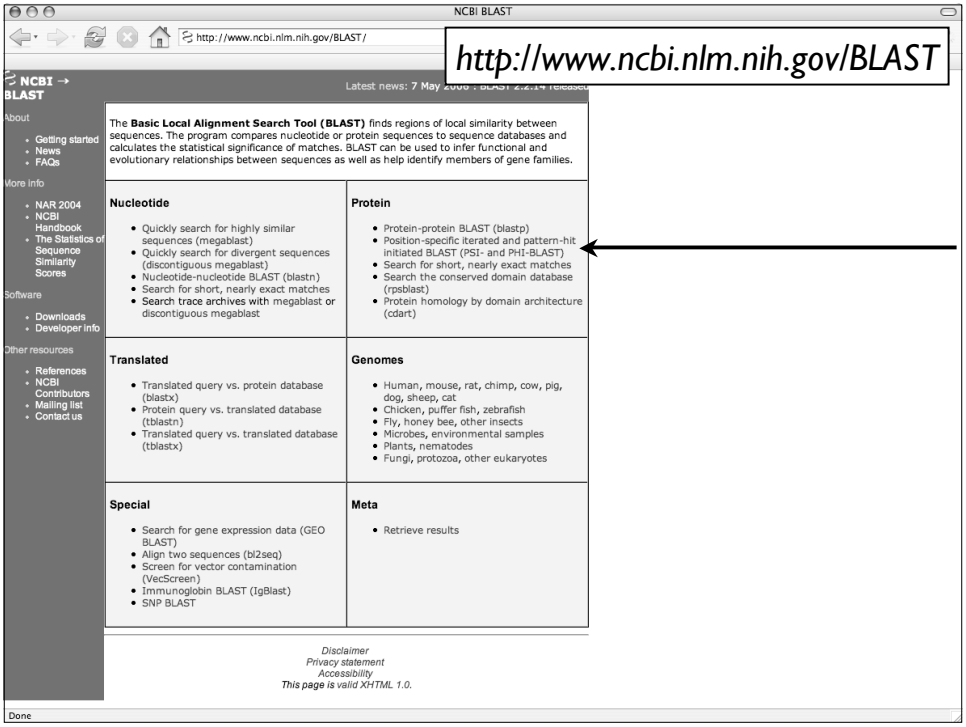

NCBI CDD cd00931
http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi
Show Alignment
Format: Hypertext
Row Display: up to 20
Color Bits: 2.0 bits
Type Selection: top listed sequences
Feature Display: FGF/FGF-Recep
Feature 1
1BIH_A 309 PKYeqk---pekVIVVKqg----qDVTIPCKVTGlp-aPNVVWSHnakp----------lsggRATVTD-------SG 361
1EV2_F 8 PYWtntekmekrLHAVPaa----nTVKFRCPAGGnp-mPTMRWLKngkefk-------qehrigGYKVRNq------hWS 69
3NCM_A 2 VMFkna----ptPQEFKeg----eDAVIVCDVVSsl-pPTIIWKHkgrdvi--------lkkdvRFIVLSn-------NY 57
gi 462073 428 PSFadt---pgkEVWGWeq----hAVNLTCLAHSip-nATISWHFngadlfrgr--egqelqqtGYTLFGsg----prST 493
gi 8134650 412 PSWlkk---pqdSQLEEg-----kPGYLDCLTQAtp-kPTVVWYRngmli---------sedsRFEVFKn-------GT 465
gi 12644418 1098 PAFkqk---lqdVHVAEg-----kKLLLQCQVSSdp-pATIIWTLngkt----------lkttKFIILSqeg---slCS 1154
gi 1169233 240 LYFlqr---psnVVAIEg-----kDAVLECCVSGyp-pPSFTWLRgeevi--------qlrskKYSLLGg-------SN 294
gi 3024083 226 PVItv-----prPRLGQal---qyDMDLECHIEAyp-pPAIVWTKddiqlannq-hysishfatADEYTD--------ST 287
gi 2497323 219 PTItes---ksnEATTGr------QASLKCEASAvp-aPDFEWYRddtrins------angleiKSTEGQ--------SS 274
gi 12644418 721 PWFis-----kpRSVTAsl---gqSVLISCAIAGdp-fPTVHWLRdgkal--------ckdtgHFEVLQne----dvFT 778
gi 1705571 35 PVItegs--prrLVVFPt-----dDISLKCEASGkp-eVQFRWTRdgvhfk-------pkeelGVTVYQsph--sgsFT 96
gi 462740 41 PTItqq---spkDYIVDpr----eNIVIQCEAKGkp-pPSFSWTRngthf---------didkDAQVTMkpn--sgtLV 100
gi 3334268 401 PKFis-----nqTIYYSwe---gnPINISCDVKSnp-pASIHWRRdklvlpak----nttnlktYSTGRK--------MI 459
gi 10720048 126 TVLvpp---skpTISVPssvtignRAVLTCSEHDgsppSEYSWFKdgismltadakktrafmnsSFTIDPk------sGD 196
gi 1709202 1339 RVLgg-----lpDVVTIqe---gkALNLTCNVWGdp-pPEVSWLKnekalaq------tdhcnlKFEAGRt-------AY 1396
gi 20178015 834 VRFapqvdhptpLTKVAaagdstsSATLHCRARGvp-nIDFTWTKngvpl--------dlqdpRYTEHKyhqgvvhsSL 903
gi 10720348 134 PAVpsc---evpSSALSg-----tVVELRCQDKGnpaPEYTWFKdgirllenp-rlgsqstnsSYTMNTk------tGT 198
gi 9789817 457 KIFsk-----psKAIIKhg----kKLRIVCEVSGqp-pPKVTWFKddks----------inrKRNIYQfkhh-krrSE 513
gi 12643619 313 PLKat-----isPRKVKssv--gsQVSLSCSVTGte-dQELSWYRngeiln--------pgknvRITGINh-------EN 369
gi 8134650 320 AEIedmp-lfepRVFTAg-----sEERVTCLPPKglpePSVWWEHagvr----------lptHGRVYQk------gHE 375
Feature 1
1BIH_A 362 LVIKGVKn--gDKGYYGCRATne--hgDKYF-ETLVQ 393
1EV2_F 70 LIMESVVp--sDKGNYTCVVEne--ygSINH-TYHLD 101
3NCM_A 58 LQIRGIKk--tDEGTYRCEGRilargeINFK-DIQVI 91
gi 462073 494 LQVIPFNr--kMYGNYKCTATn----kHGTA-VHEIM 523
gi 8134650 466 LRINSVEv--yDGTWYRCMSStp---aGSIEaQAVLQ 497
gi 12644418 1155 VSIEKALp--eDRGLYKCVAKnd--agQAEC-SCQVT 1186
gi 1169233 295 LLISNVTd--dDSGMYTCVVTykn-enISASaELTVL 328
gi 3024083 288 LRVITVEk--rQYGDYVCKATnr---fGEAEaRVNLF 319
gi 2497323 275 LTVTNVTe--eHYGNYTCVAAnkl-gvTNAS-LVLFR 307
gi 12644418 779 LVLKKVOp--wHAGOYEILLKnr--vgECSC-OVSLM 810
Done

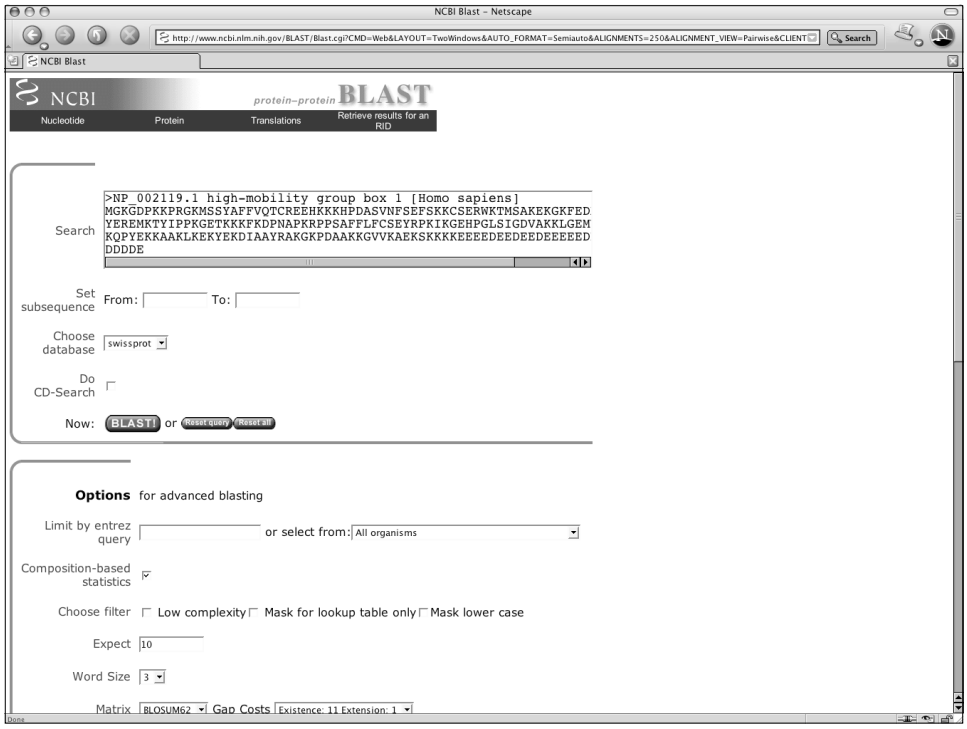

Mozilla Firefox
http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
NCBI
Conserved Domains
SH3
SH2
HOME | SEARCH | SITE MAP
NewSearch | PubMed | Nucleotide | Protein | Structure | CDD | Taxonomy | Help
Query sequence: [(local sequence)lcl|NP_005206]
deleted in colorectal carcinoma [Homo sapiens]
Concise Result
Full Result
Show Search Information
IGcam
FN3
Neogenin_C
Descriptions
Title
PssmId
Multi-Dom
E-value
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 2e-15
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 5e-13
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 9e-13
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 6e-12
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 2e-11
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 2e-10
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 2e-9
[+]cd00063, FN3, Fibronectin type 3 domain; One of three types of internal repeats found ... 28945 No 5e-7
[+]cd00931, IGcam, Immunoglobulin domain cell adhesion molecule (cam) subfamily; members ... 28983 No 6e-7
[+]pfam05583, Neogenin_C, Neogenin C-terminus. This family represents the C-terminus of e... 46472 No 3e-105
Search for similar domain architectures
CD Search Reference:
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
Done

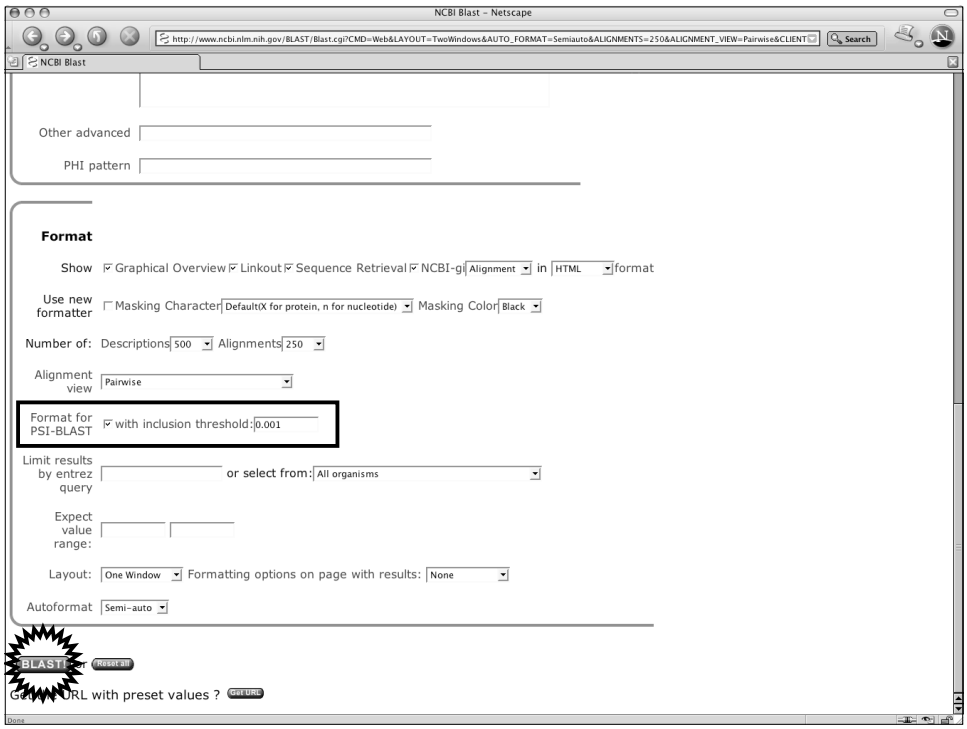

NCBI DART
http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi
NCBI
CDART: Conserved Domain Architecture Retrieval Tool
New Query | Overview | PubMed | Nucleotide | Protein | Structure | Taxonomy | Help?
About CDART
Query
Neogenin_C
FN3
Similar domain architectures
3 Sequences
33 Sequences
3 Sequences
3 Sequences
74 Sequences
25 Sequences
2 Sequences
4 Sequences
5 Sequences
STOP
S_TKc
DUF612
RhoGEF
SH3
Result page: Previous 1 2 3 4 5 6 7 8 9 10 11 Next
Subset by Taxonomy
Subset by selected domains:
cd00063 Fibronectin type 3 domain; One of three types of ...
includes: pfam00041 smart00060
cd00931 Immunoglobulin domain cell adhesion molecule (cam...
includes: cd00096 cd00098 cd00099 pfam00047 smart00406 smart00407 smart00408 smart00409
Done

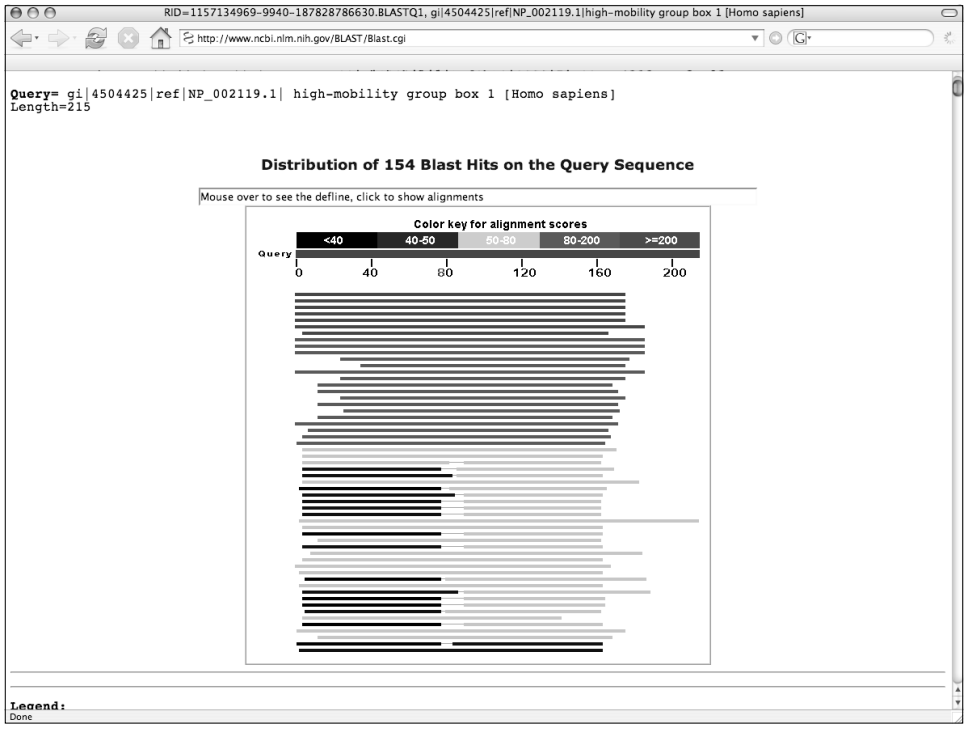
PSI-BLAST

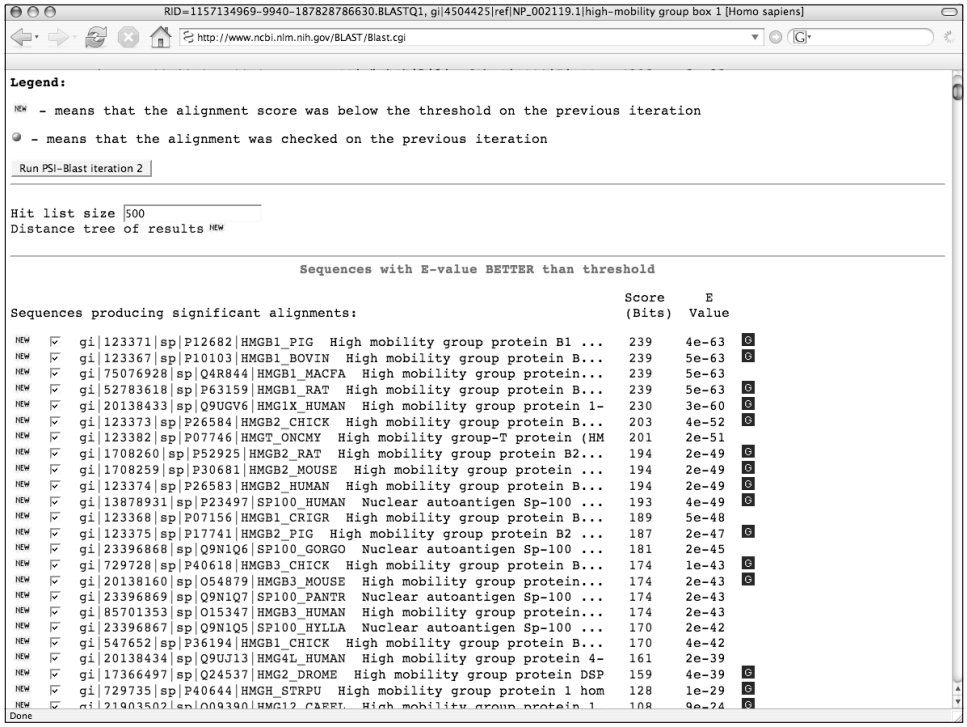
- Position-Specific Iterated BLAST search
- Easy-to-use version of a profile-based search
  - Perform BLAST search against protein database
  - Use results to calculate a position-specific scoring matrix
  - PSSM replaces query for next round of searches
  - May be iterated until no new significant alignments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

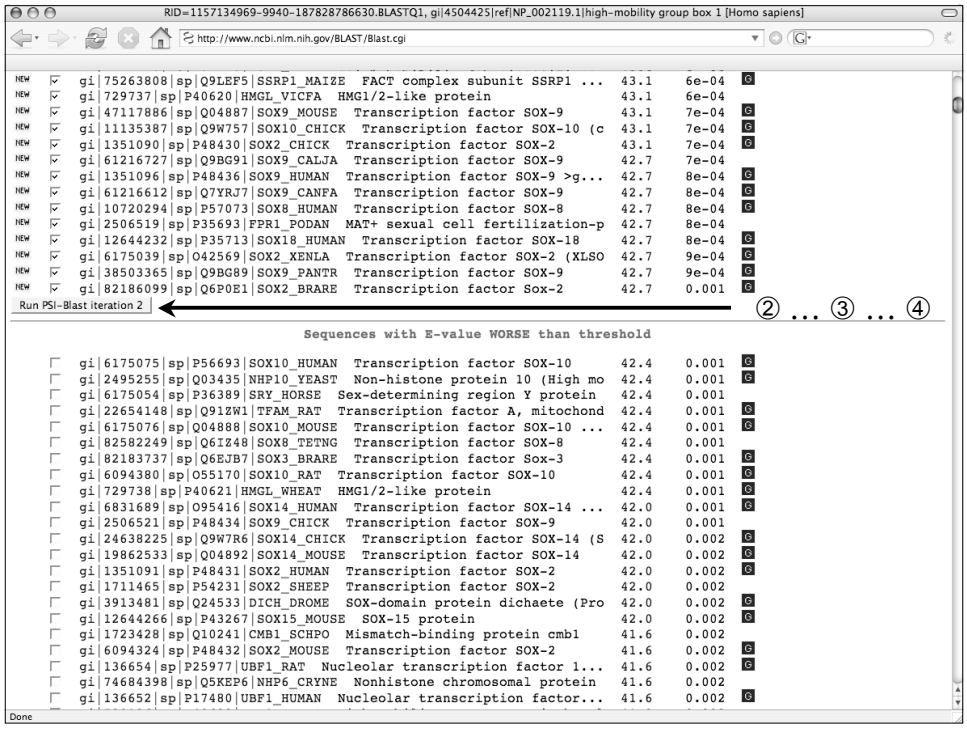


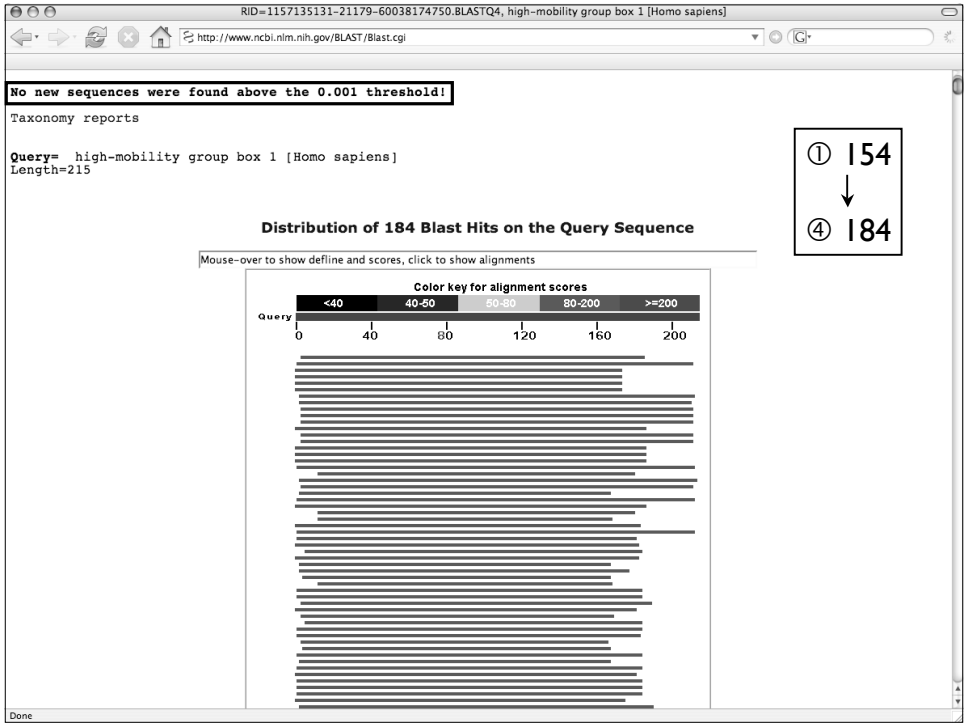

NCBI Blast - Netscape
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CLIENT
Search
NCBI Blast
NCBI
protein-protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Search
>NP_002119.1 high-mobility group box 1 [Homo sapiens]
MGKGDPKKPRGKMSSYAFFVQTCREEHKKKHPDASVNFSEFSKKCSERWKTMSAKEKGKFED
YEREMKTYIPPKGETKKKFKDPNAPKRPPSAFFLFCSEYRPKIKGEHPGLSIGDVAKKLGEM
KQPYEKKAAKLKEKYEKDIAAYRAKGKPDAAKKGVVKAEKSKKKKEEEEDEEDEEDEEEEED
DDDDE
Set subsequence
From:
To:
Choose database
swissprot
Do CD-Search
Now:
BLAST!
or
Reset 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
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
Other advanced
PHI pattern
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
NCBI-gi
Alignment
in
HTML
format
Use new formatter
Masking Character
Default(X for protein, n for nucleotide)
Masking Color
Black
Number of:
Descriptions
500
Alignments
250
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!
Reset all
with preset values ?
Get URL

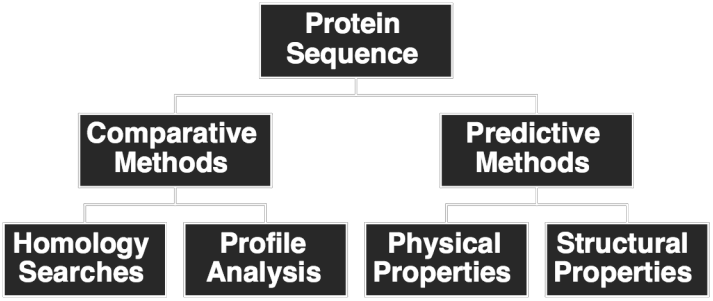

RID=1157134969-9940-187828786630.BLASTQ1, gi|4504425|ref|NP_002119.1|high-mobility group box 1 [Homo sapiens]
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Query= gi|4504425|ref|NP_002119.1| high-mobility group box 1 [Homo sapiens]
Length=215
Distribution of 154 Blast Hits on the Query Sequence
Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
40
80
120
160
200
Legend:
Done


RID=1157134969-9940-187828786630.BLASTQ1, gi|4504425|ref|NP_002119.1|high-mobility group box 1 [Homo sapiens]
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Legend:
NEW - 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
Distance tree of results NEW
Sequences with E-value BETTER than threshold
Sequences producing significant alignments: Score (Bits) E Value
NEW gi|123371|sp|P12682|HMGB1_PIG High mobility group protein B1 ... 239 4e-63 G
NEW gi|123367|sp|P10103|HMGB1_BOVIN High mobility group protein B... 239 5e-63 G
NEW gi|75076928|sp|Q4R844|HMGB1_MACFA High mobility group protein... 239 5e-63
NEW gi|52783618|sp|P63159|HMGB1_RAT High mobility group protein B... 239 5e-63 G
NEW gi|20138433|sp|Q9UGV6|HMG1X_HUMAN High mobility group protein 1- 230 3e-60 G
NEW gi|123373|sp|P26584|HMGB2_CHICK High mobility group protein B... 203 4e-52 G
NEW gi|123382|sp|P07746|HMGT_ONCMY High mobility group-T protein (HM 201 2e-51
NEW gi|1708260|sp|P52925|HMGB2_RAT High mobility group protein B2... 194 2e-49 G
NEW gi|1708259|sp|P30681|HMGB2_MOUSE High mobility group protein ... 194 2e-49 G
NEW gi|123374|sp|P26583|HMGB2_HUMAN High mobility group protein B... 194 2e-49 G
NEW gi|13878931|sp|P23497|SP100_HUMAN Nuclear autoantigen Sp-100 ... 193 4e-49 G
NEW gi|123368|sp|P07156|HMGB1_CRIGR High mobility group protein B... 189 5e-48
NEW gi|123375|sp|P17741|HMGB2_PIG High mobility group protein B2 ... 187 2e-47 G
NEW gi|23396868|sp|Q9N1Q6|SP100_GORGO Nuclear autoantigen Sp-100 ... 181 2e-45
NEW gi|729728|sp|P40618|HMGB3_CHICK High mobility group protein B... 174 1e-43 G
NEW gi|20138160|sp|O54879|HMGB3_MOUSE High mobility group protein... 174 2e-43 G
NEW gi|23396869|sp|Q9N1Q7|SP100_PANTR Nuclear autoantigen Sp-100 ... 174 2e-43
NEW gi|85701353|sp|O15347|HMGB3_HUMAN High mobility group protein... 174 2e-43
NEW gi|23396867|sp|Q9N1Q5|SP100_HYLLA Nuclear autoantigen Sp-100 ... 170 2e-42
NEW gi|547652|sp|P36194|HMGB1_CHICK High mobility group protein B... 170 4e-42
NEW gi|20138434|sp|Q9UJ13|HMG4L_HUMAN High mobility group protein 4- 161 2e-39
NEW gi|17366497|sp|Q24537|HMG2_DROME High mobility group protein DSP 159 4e-39 G
NEW gi|729735|sp|P40644|HMGH_STRPU High mobility group protein 1 hom 128 1e-29 G
NEW gi|21903502|sp|O09390|HMG12_CAEEL High mobility group protein 1 108 9e-24 G
Done


RID=1157134969-9940-187828786630.BLASTQ1, gi|4504425|ref|NP_002119.1|high-mobility group box 1 [Homo sapiens]
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
gi|75263808|sp|Q9LEF5|SSRP1_MAIZE FACT complex subunit SSRP1 ... 43.1 6e-04
gi|729737|sp|P40620|HMGL_VICFA HMG1/2-like protein 43.1 6e-04
gi|47117886|sp|Q04887|SOX9_MOUSE Transcription factor SOX-9 43.1 7e-04
gi|11135387|sp|Q9W757|SOX10_CHICK Transcription factor SOX-10 (c 43.1 7e-04
gi|1351090|sp|P48430|SOX2_CHICK Transcription factor SOX-2 43.1 7e-04
gi|61216727|sp|Q9BG91|SOX9_CALJA Transcription factor SOX-9 42.7 7e-04
gi|1351096|sp|P48436|SOX9_HUMAN Transcription factor SOX-9 >g... 42.7 8e-04
gi|61216612|sp|Q7YRJ7|SOX9_CANFA Transcription factor SOX-9 42.7 8e-04
gi|10720294|sp|P57073|SOX8_HUMAN Transcription factor SOX-8 42.7 8e-04
gi|2506519|sp|P35693|FPR1_PODAN MAT+ sexual cell fertilization-p 42.7 8e-04
gi|12644232|sp|P35713|SOX18_HUMAN Transcription factor SOX-18 42.7 8e-04
gi|6175039|sp|O42569|SOX2_XENLA Transcription factor SOX-2 (XLSO 42.7 9e-04
gi|38503365|sp|Q9BG89|SOX9_PANTR Transcription factor SOX-9 42.7 9e-04
gi|82186099|sp|Q6P0E1|SOX2_BRARE Transcription factor Sox-2 42.7 0.001
Run PSI-Blast iteration 2
② ... ③ ... ④
Sequences with E-value WORSE than threshold
gi|6175075|sp|P56693|SOX10_HUMAN Transcription factor SOX-10 42.4 0.001
gi|2495255|sp|Q03435|NHP10_YEAST Non-histone protein 10 (High mo 42.4 0.001
gi|6175054|sp|P36389|SRY_HORSE Sex-determining region Y protein 42.4 0.001
gi|22654148|sp|Q91ZW1|TFAM_RAT Transcription factor A, mitochond 42.4 0.001
gi|6175076|sp|Q04888|SOX10_MOUSE Transcription factor SOX-10 ... 42.4 0.001
gi|82582249|sp|Q6IZ48|SOX8_TETNG Transcription factor SOX-8 42.4 0.001
gi|82183737|sp|Q6EJB7|SOX3_BRARE Transcription factor Sox-3 42.4 0.001
gi|6094380|sp|O55170|SOX10_RAT Transcription factor SOX-10 42.4 0.001
gi|729738|sp|P40621|HMGL_WHEAT HMG1/2-like protein 42.4 0.001
gi|6831689|sp|O95416|SOX14_HUMAN Transcription factor SOX-14 ... 42.0 0.001
gi|2506521|sp|P48434|SOX9_CHICK Transcription factor SOX-9 42.0 0.001
gi|24638225|sp|Q9W7R6|SOX14_CHICK Transcription factor SOX-14 (S 42.0 0.002
gi|19862533|sp|Q04892|SOX14_MOUSE Transcription factor SOX-14 42.0 0.002
gi|1351091|sp|P48431|SOX2_HUMAN Transcription factor SOX-2 42.0 0.002
gi|1711465|sp|P54231|SOX2_SHEEP Transcription factor SOX-2 42.0 0.002
gi|3913481|sp|Q24533|DICH_DROME SOX-domain protein dichaete (Pro 42.0 0.002
gi|12644266|sp|P43267|SOX15_MOUSE SOX-15 protein 42.0 0.002
gi|1723428|sp|Q10241|CMB1_SCHPO Mismatch-binding protein cmb1 41.6 0.002
gi|6094324|sp|P48432|SOX2_MOUSE Transcription factor SOX-2 41.6 0.002
gi|136654|sp|P25977|UBF1_RAT Nucleolar transcription factor 1... 41.6 0.002
gi|74684398|sp|Q5KEP6|NHP6_CRYNE Nonhistone chromosomal protein 41.6 0.002
gi|136652|sp|P17480|UBF1_HUMAN Nucleolar transcription factor... 41.6 0.002
Done


RID=1157135131-21179-60038174750.BLASTQ4, high-mobility group box 1 [Homo sapiens]
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
No new sequences were found above the 0.001 threshold!
Taxonomy reports
Query= high-mobility group box 1 [Homo sapiens]
Length=215
① 154
④ 184
Distribution of 184 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
Query
0
40
80
120
160
200
Done

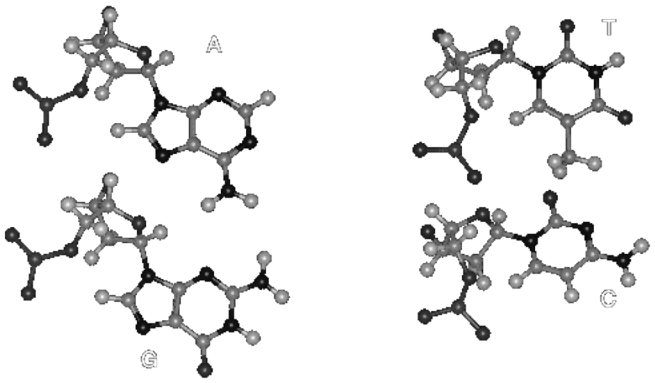
Protein Sequence Analysis



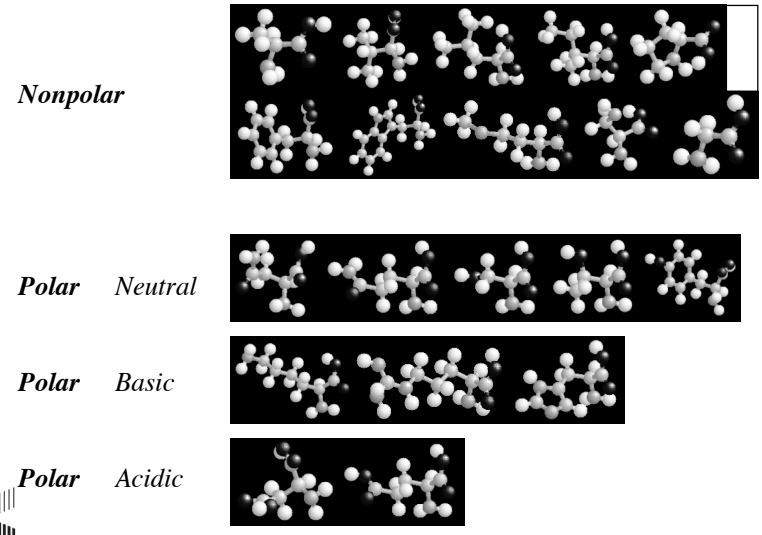
- ***Common structure?***
- ***Common function?***
- ***Evolutionary relationship?***
- ***Global or local similarity?***

- ***Composition***
- ***Hydrophobicity***
- ***Secondary structure***
- ***Specialized structures***
- ***Tertiary structure***

Information Landscape



Information Landscape



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 characteristic (e.g., hydrophobicity) by sequence position

  HelixWheel/HelixDraw

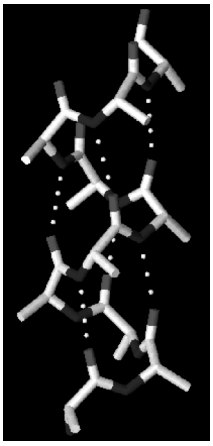
  Display protein sequence as a helical wheel

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

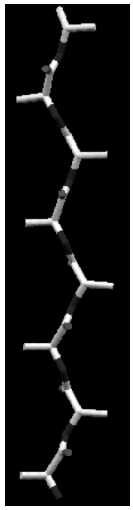
Alpha-helix

- 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



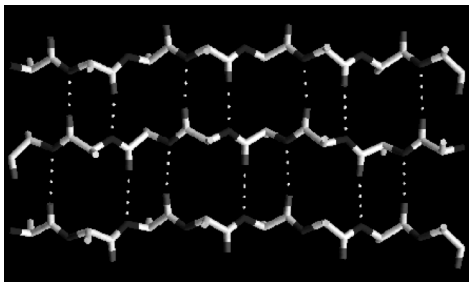
Beta-strand

- 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



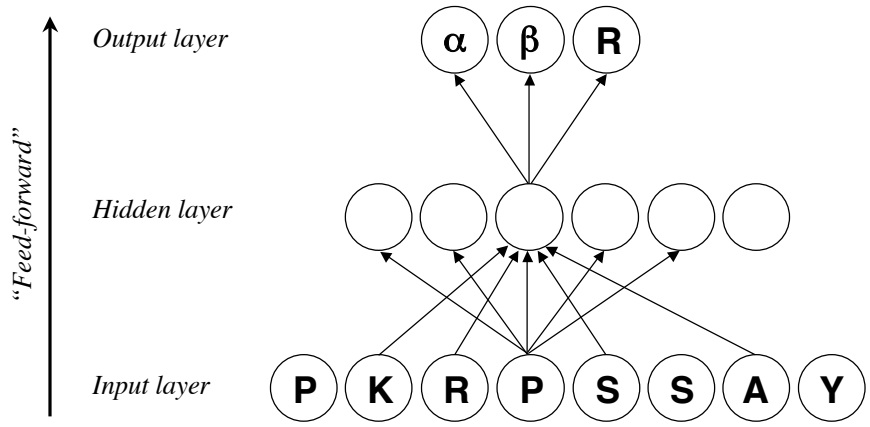
Folding Classes

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

Neural Networks

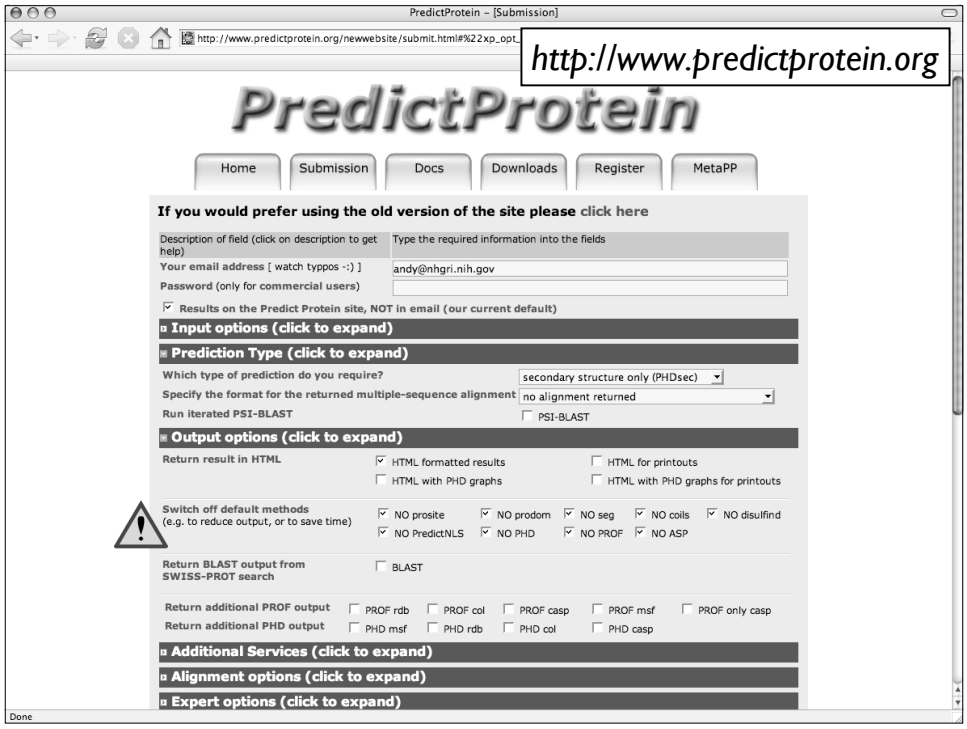
- Used when direct cause-and-effect rules between the beginning and end states are not known
  - Beginning and end states must be related
  - Neural networks attempt to deduce the relationship between the beginning and end states
- Supervised learning approach
  - Involves use of “training sets” where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

Neural Networks



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%


PredictProtein – [Submission]
http://www.predictprotein.org/newwebsite/submit.html#%22xp_opt_
http://www.predictprotein.org
PredictProtein
Home
Submission
Docs
Downloads
Register
MetaPP
If you would prefer using the old version of the site please click here
Description of field (click on description to get help)
Type the required information into the fields
Your email address [watch typpos -:)]
andy@nhgri.nih.gov
Password (only for commercial users)
Results on the Predict Protein site, NOT in email (our current default)
Input options (click to expand)
Prediction Type (click to expand)
Which type of prediction do you require?
secondary structure only (PHDsec)
Specify the format for the returned multiple-sequence alignment
no alignment returned
Run iterated PSI-BLAST
PSI-BLAST
Output options (click to expand)
Return result in HTML
HTML formatted results
HTML for printouts
HTML with PHD graphs
HTML with PHD graphs for printouts
Switch off default methods
(e.g. to reduce output, or to save time)
NO prosite
NO prodom
NO seg
NO coils
NO disulfind
NO PredictNLS
NO PHD
NO PROF
NO ASP
Return BLAST output from SWISS-PROT search
BLAST
Return additional PROF output
PROF rdb
PROF col
PROF casp
PROF msf
PROF only casp
Return additional PHD output
PHD msf
PHD rdb
PHD col
PHD casp
Additional Services (click to expand)
Alignment options (click to expand)
Expert options (click to expand)
Done


PredictProtein – [Submission]
http://www.predictprotein.org/newwebsite/submit.html#%22xp_opt_click
Return result in HTML
HTML formatted results
HTML for printouts
HTML with PHD graphs
HTML with PHD graphs for printouts
Switch off default methods
(e.g. to reduce output, or to save time)
NO prosite
NO prodom
NO seg
NO coils
NO disulfind
NO PredictNLS
NO PHD
NO PROF
NO ASP
Return BLAST output from SWISS-PROT search
BLAST
Return additional PROF output
PROF rdb
PROF col
PROF casp
PROF msf
PROF only casp
Return additional PHD output
PHD msf
PHD rdb
PHD col
PHD casp
Additional Services (click to expand)
Alignment options (click to expand)
Expert options (click to expand)
Paste, or type your sequence
amino acids in one-letter code
any number of white spaces allowed
non-standard amino acids to 'X'
use SRS6 to get your sequence from a public database
>flavodoxin - Anacystis nidulans
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVGEI
DDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKA
LAIDEDNQPDLTKNRIKTWVSQLKSEFGL
Batch or interactive?
interactive
Final action
SUBMIT / RUN PREDICTION
CLEAR PAGE
Search
GO
Done

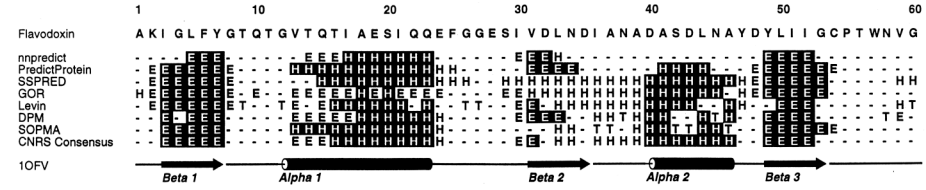
PROF results (normal)

```
            ....,....1....,....2....,....3....,....4....,....5....,....6
AA          AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVG
OBS_sec
PROF_sec     EEEEEEE      HHHHHHHHHHHHH     EEEEE               EEEEEE
Rel_sec     927899843676168888888887202682354422246600023355379873034766

            ....,....7....,....8....,....9....,....10.1.,....11.1.,....1
AA          ELQSDWEGIYDDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYW
OBS_sec
PROF_sec        HHHHHHHHHH        EEEEEEE         HHHHHHHHHHHHHHH     EEEEE
Rel_sec     654126788887412566886378884146654331033678889988874078244111

            ....,....13.1.,....14.1.,....15.1.,....16.1.,....
AA          PIEGYDFNESKAVRNNQFVGLAIDEDNQPDLTKNRIKTWVSQLKSEFGL
OBS_sec
PROF_sec               EEE    EEEEEE          HHHHHHHHHHHHHH
Rel_sec     135532243311115826776524666420236889999988775438 9
```

`Prof_sec` Prediction, where `H` = helix and `E` = strand

`Rel_sec` Reliability of the prediction at each position

Accuracy of Predictions



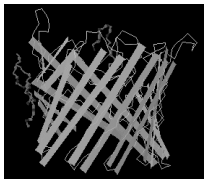
Secondary Structure Prediction Methods

- PredictProtein
  http://www.predictprotein.org
- PSIPRED
  http://bioinf.cs.ucl.ac.uk/psipred/
- SAM-T99
  http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html
- Jpred
  http://www.compbio.dundee.ac.uk/~www-jpred/submit.html

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

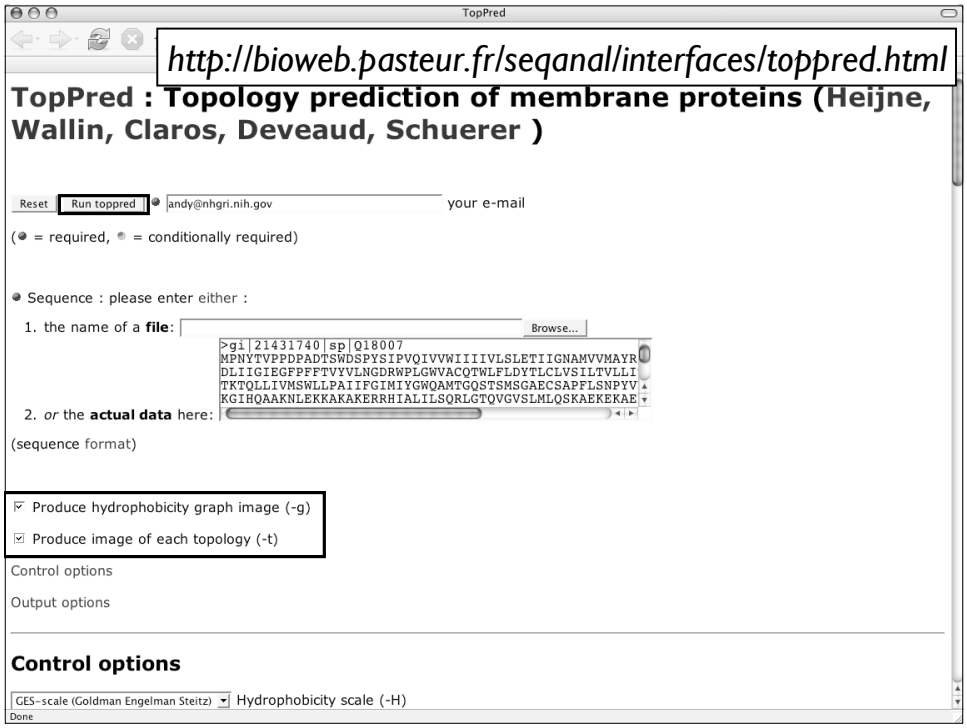


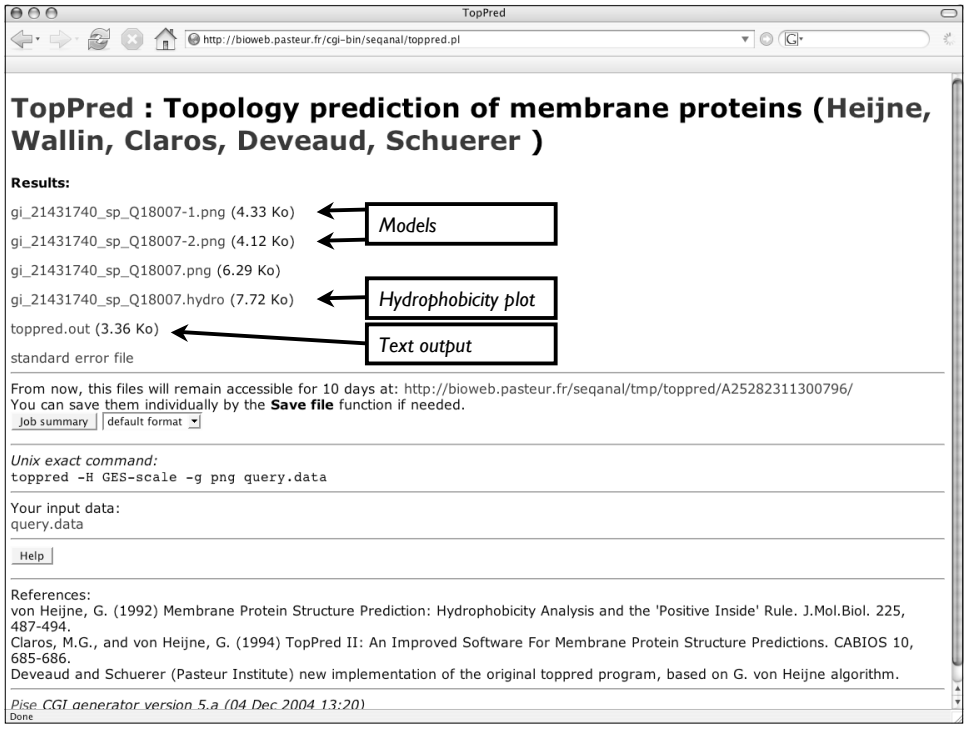
- Beta-barrel
  - Anti-parallel sheets rolled into cylinder

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

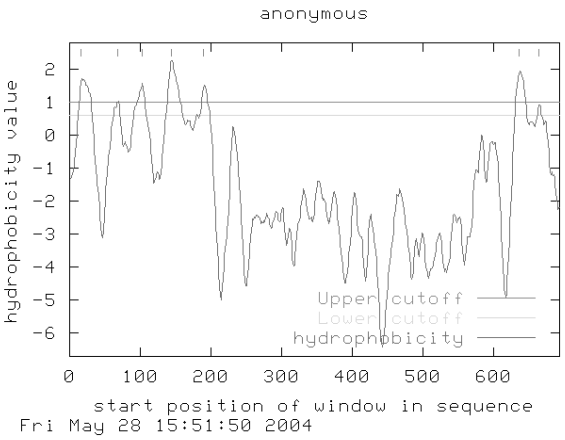




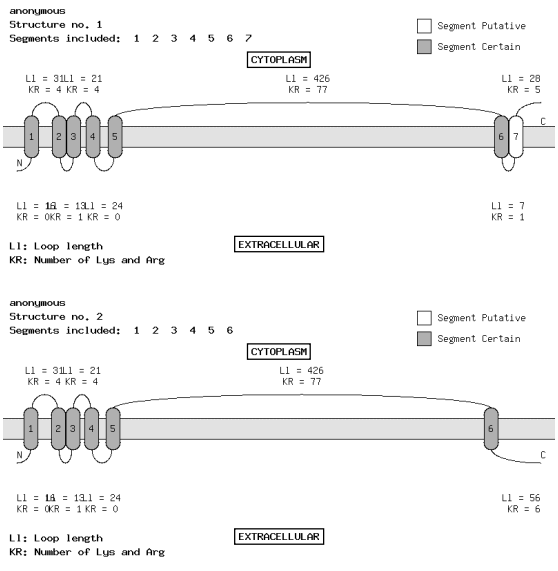
TopPred

```
Algorithm specific parameters:

Full window size : 21
Core window size : 11
Wedge window size: 5
Using hydrophobicity file: GES-scale

Cutoff for certain transmembrane segments: 1.00
Cutoff for putative transmembrane segments: 0.60
Critical distance between 2 transmembrane segments: 2

Critical loop length:  60

Kingdom: procaryote

Using cyt/ext file: CYTEXT-scale


Sequence : gi_21431740_sp_Q18007  (713 res)
MPNYTVPPDPADTSWDSPYSIPVQIVVWIIIIVLSLETIIGNAMVVMAYRIERNISKQVS
NRYIVSLAISDLIIGIEGFPFFTVYVLNGDRWPLGWVACQTWLFLDYTLCLVSILTVLLI
TADRYLSVCHTAKYLKWQSPTKTQLLIVMSWLLPAIIFGIMIYGWQAMTGQSTSMSGAEC
SAPFLSNPYVNMGMYVAYYWTTLVAMLLYKGIHQAAKNLEKKAKAKERRHIALILSQRL
GTQVGVSLMLQSKAEKEKAEEAQKDSGYTSNQACDANNLRRFGFSEPETSQFRVDPNSNN
NLNVEGSLNTENDQNLGVIEEERSGFLSRRESNESYYPGPHPTAANSRRCSEMEKVSLLS
ESDGVPSTRPAKSYGRLSLRSRYSASESITTTHENDEKEVEKADSLQKLFADDELGSVLN
FKEEKLKNTDSNNDSDTTSVILQRSKKYKKNKKPRSSRRSEHSTPRQIAKVKQAEGTAAQ
LIEESVPDDDQTETIEVKRTDRWVVSMKKRIARALIRRRSTTRPERGSSSNSDDSSSEVE
GEEKPEVRNNGLKIPQLTVNNENRGETSSQPGRDRLAPPNKTDTFLSASGVSRKISTIST
VITREKVISSIFAPIAVFNRGRKQTKAEKRAHKAFRTITFIVGFFAILWSPYYIMATVYG
FCKGECIPSFLYTLSYYMCYLNSSGNPFAYALANRQFRSAFMRMFRGNFNKVA

Found: 7 segments

Candidate membrane-spanning segments:

 Helix Begin - End   Score Certainity
     1    17 - 37    1.717 Certain
     2    69 - 89    1.024 Certain
     3   103 - 123   1.555 Certain
     4   145 - 165   2.264 Certain
     5   190 - 210   1.531 Certain
     6   637 - 657   1.931 Certain
     7   665 - 685   0.920 Putative
```

TopPred

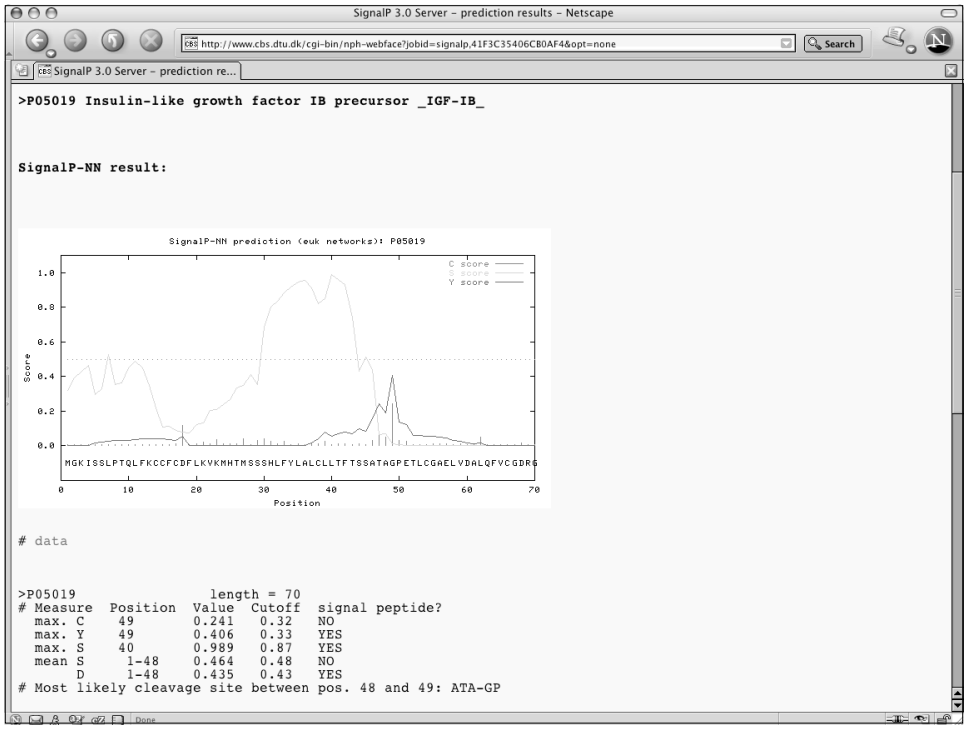


TopPred



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)

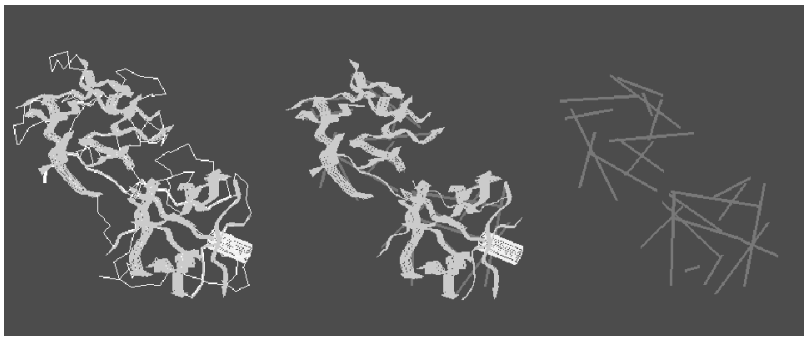


Predicting Tertiary Structure

- 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

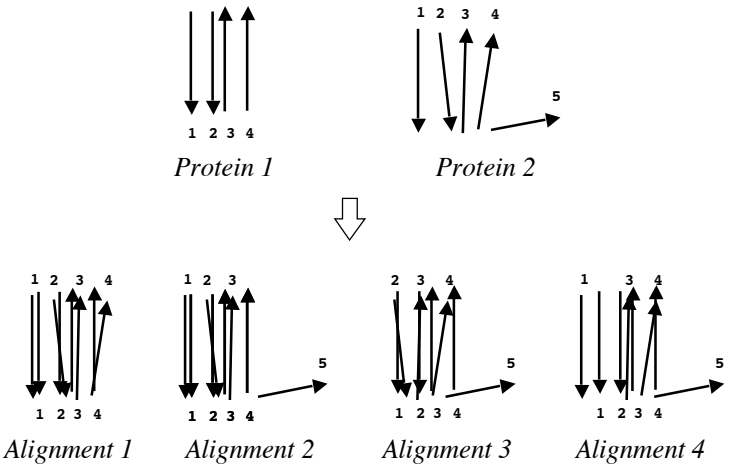
VAST Structure Comparison

Step 1: Construct vectors for secondary structure elements



VAST Structure Comparison

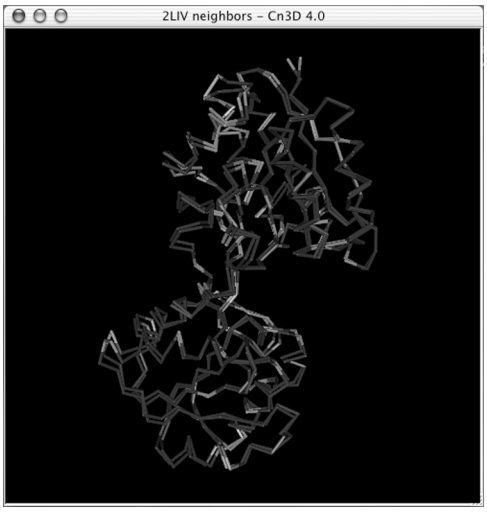
Step 2: Optimally align structure element vectors

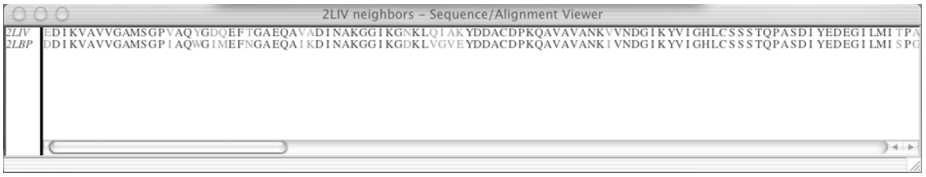


Cn3D Viewer

Rendering: Tubes

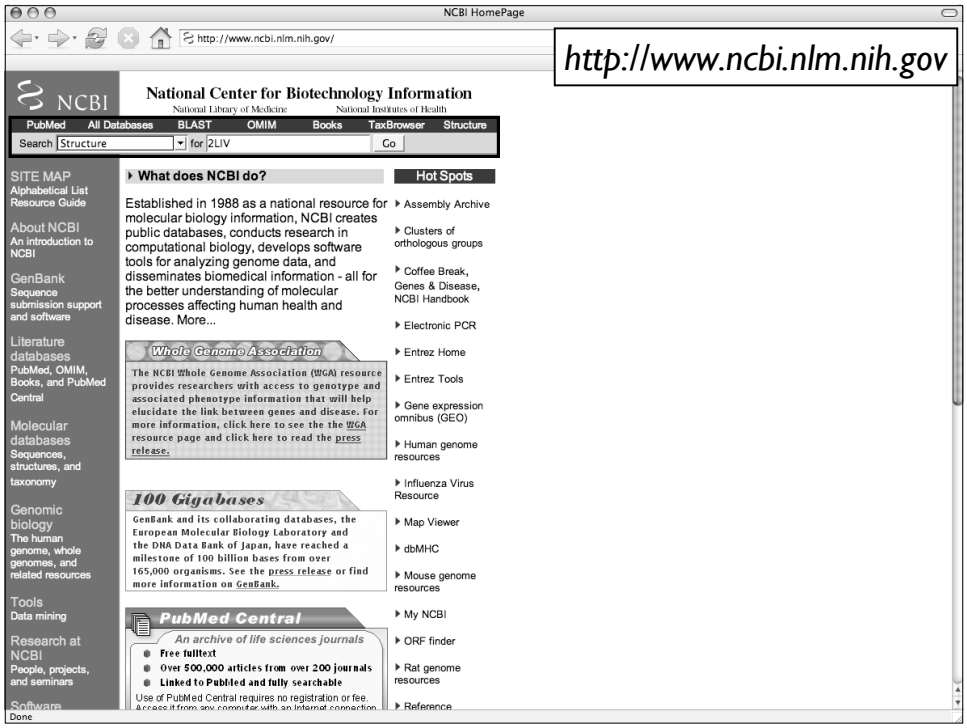
Coloring: Identity
Red – matches
Blue – mismatches

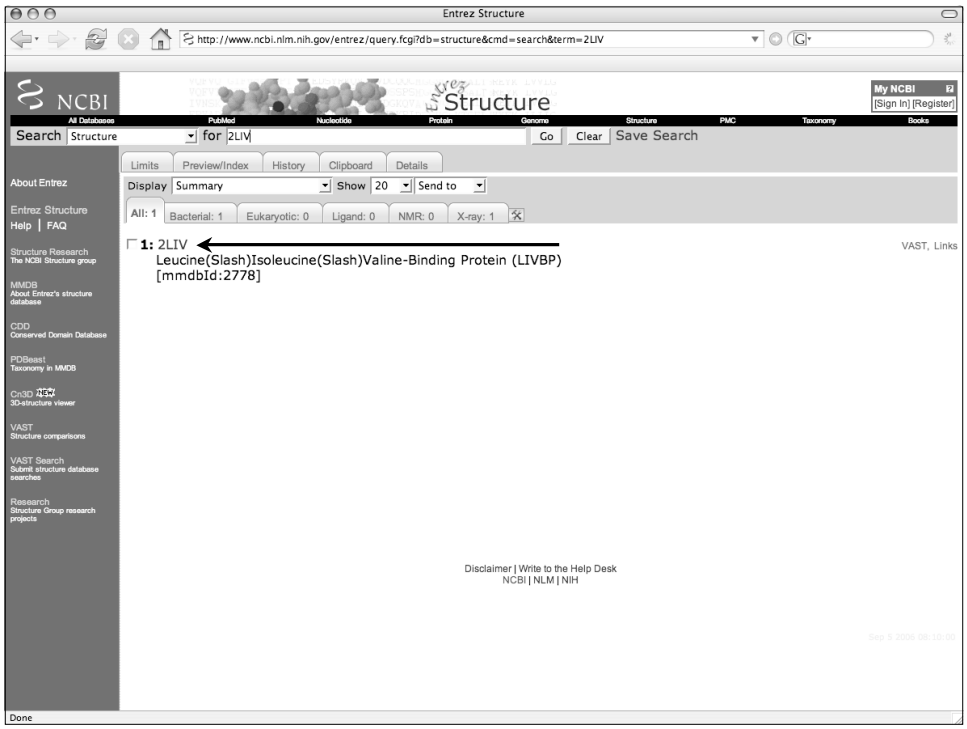


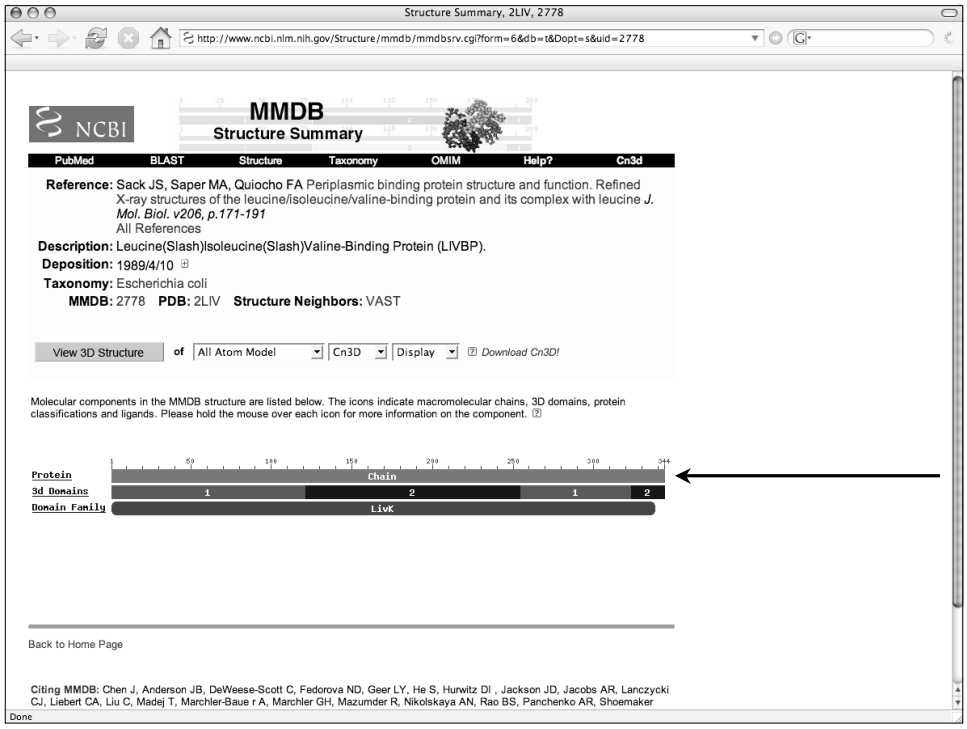


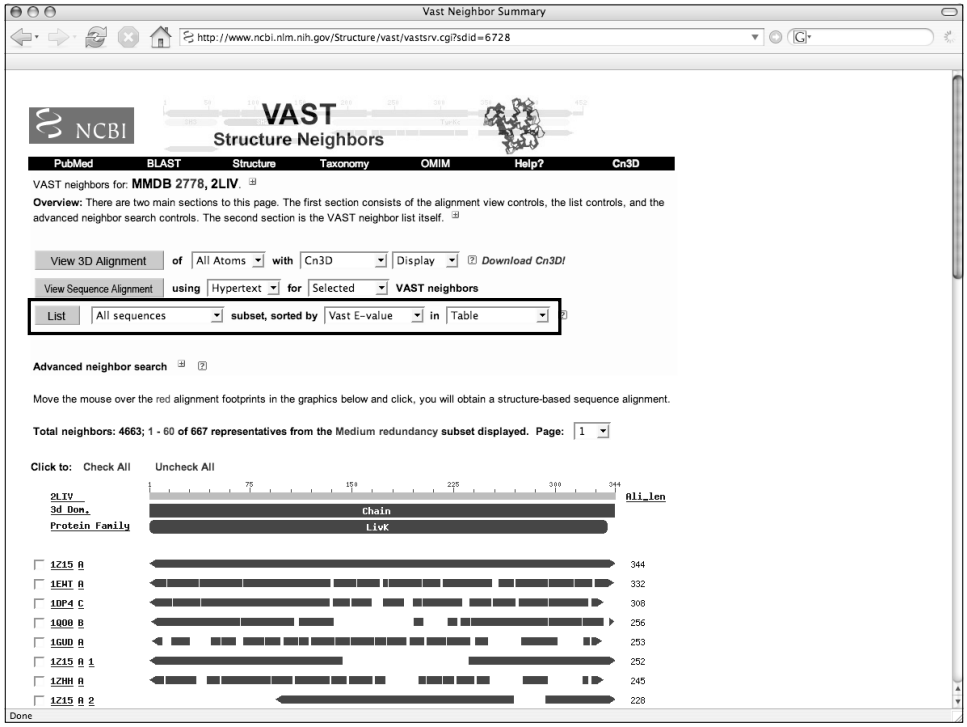
VAST Shortcomings

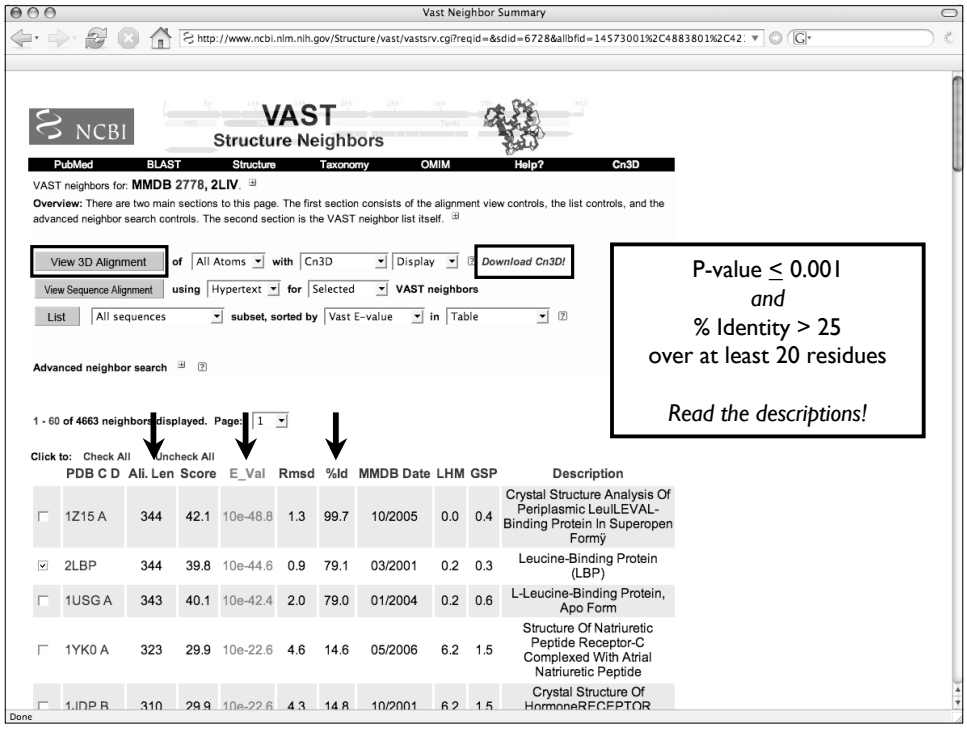
- Not the best method for determining structural similarities
- Reducing a structure to a series of vectors necessarily results in a loss of information (less confidence in prediction)
- Regardless of the "simplicity" of the method, provides a simple and fast first answer to the question of structural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


NCBI HomePage
http://www.ncbi.nlm.nih.gov/
http://www.ncbi.nlm.nih.gov
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed All Databases BLAST OMIM Books TaxBrowser Structure
Search Structure for 2LIV Go
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...
Whole Genome Association
The NCBI Whole Genome Association (WGA) resource provides researchers with access to genotype and associated phenotype information that will help elucidate the link between genes and disease. For more information, click here to see the the WGA resource page and click here to read the press release.
100 Gigabases
GenBank and its collaborating databases, the European Molecular Biology Laboratory and the DNA Data Bank of Japan, have reached a milestone of 100 billion bases from over 165,000 organisms. See the press release or find more information on GenBank.
PubMed Central
An archive of life sciences journals
Free fulltext
Over 500,000 articles from over 200 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
Influenza Virus Resource
Map Viewer
dbMHC
Mouse genome resources
My NCBI
ORF finder
Rat genome resources
Reference
SITE MAP
Alphabetical List Resource Guide
About NCBI
An introduction to NCBI
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
Research at NCBI
People, projects, and seminars
Software
Done


Entrez Structure
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=structure&cmd=search&term=2LIV
NCBI
Structure
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Search Structure for 2LIV Go Clear Save Search
Limits Preview/Index History Clipboard Details
About Entrez
Entrez Structure
Help | FAQ
Display Summary Show 20 Send to
All: 1 Bacterial: 1 Eukaryotic: 0 Ligand: 0 NMR: 0 X-ray: 1
1: 2LIV
Leucine(Slash)Isoleucine(Slash)Valine-Binding Protein (LIVBP)
[mmdbId:2778]
VAST, Links
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
Disclaimer | Write to the Help Desk
NCBI | NLM | NIH
Done


Structure Summary, 2LIV, 2778
http://www.ncbi.nlm.nih.gov/Structure/mmdb/mmdbsrv.cgi?form=6&db=t&Dopt=s&uid=2778
NCBI
MMDB
Structure Summary
PubMed BLAST Structure Taxonomy OMIM Help? Cn3d
Reference: Sack JS, Saper MA, Quiocho FA Periplasmic binding protein structure and function. Refined X-ray structures of the leucine/isoleucine/valine-binding protein and its complex with leucine J. Mol. Biol. v206, p.171-191
All References
Description: Leucine(Slash)Isoleucine(Slash)Valine-Binding Protein (LIVBP).
Deposition: 1989/4/10
Taxonomy: Escherichia coli
MMDB: 2778 PDB: 2LIV Structure Neighbors: VAST
View 3D Structure of All Atom Model Cn3D Display Download Cn3D!
Molecular components in the MMDB structure are listed below. The icons indicate macromolecular chains, 3D domains, protein classifications and ligands. Please hold the mouse over each icon for more information on the component.
Protein
3d Domains
Domain Family
Chain
1 2 1 2
LivK
Back to Home Page
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-Baue r A, Marchler GH, Mazumder R, Nikolskaya AN, Rao BS, Panchenko AR, Shoemaker
Done

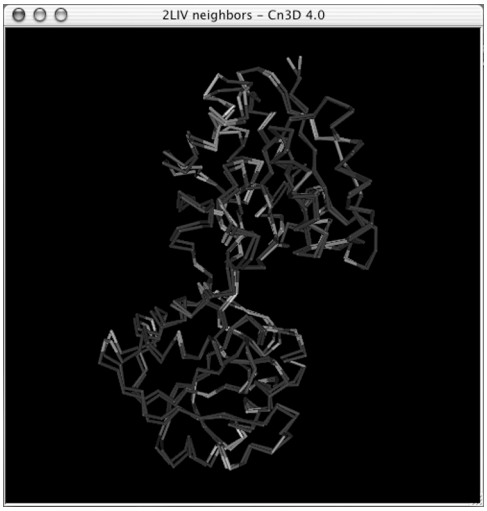

Vast Neighbor Summary
http://www.ncbi.nlm.nih.gov/Structure/vast/vastsrv.cgi?sdid=6728
NCBI
VAST
Structure Neighbors
PubMed BLAST Structure Taxonomy OMIM Help? Cn3D
VAST neighbors for: MMDB 2778, 2LIV.
Overview: There are two main sections to this page. The first section consists of the alignment view controls, the list controls, and the advanced neighbor search controls. The second section is the VAST neighbor list itself.
View 3D Alignment of All Atoms with Cn3D Display Download Cn3D!
View Sequence Alignment using Hypertext for Selected VAST neighbors
List All sequences subset, sorted by Vast E-value in Table
Advanced neighbor search
Move the mouse over the red alignment footprints in the graphics below and click, you will obtain a structure-based sequence alignment.
Total neighbors: 4663; 1 - 60 of 667 representatives from the Medium redundancy subset displayed. Page: 1
Click to: Check All Uncheck All
2LIV
3d Dom.
Protein Family
Chain
LivK
Ali_len
1Z15 A 344
1EWT A 332
1DP4 C 308
1QQ0 B 256
1GUD A 253
1Z15 A 1 252
1ZHH A 245
1Z15 A 2 228
Done

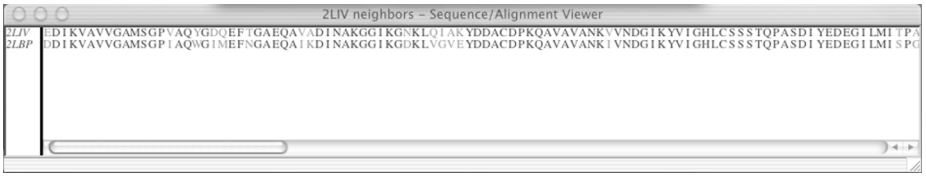


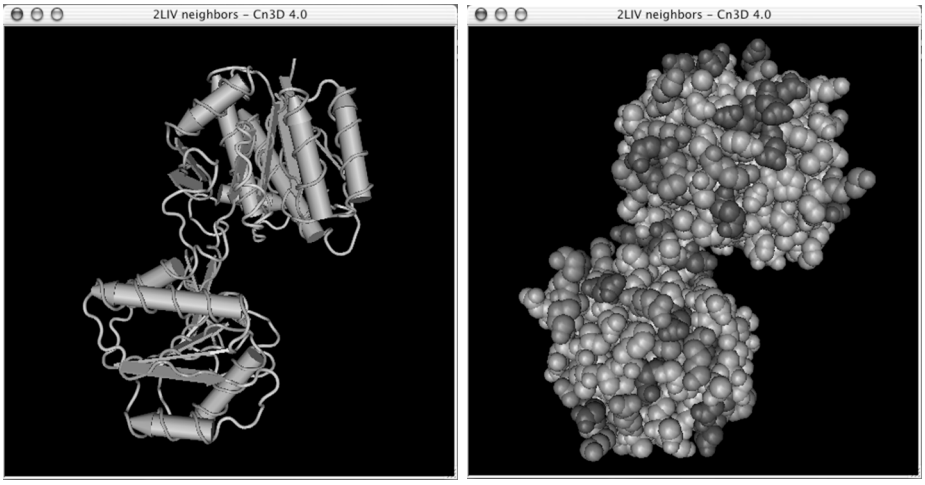
Cn3D Viewer

Rendering: Tubes

Coloring: Identity
Red – matches
Blue – mismatches







Worms	**Rendering**	Spacefill
Secondary Structure	**Coloring**	Charge

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

↓

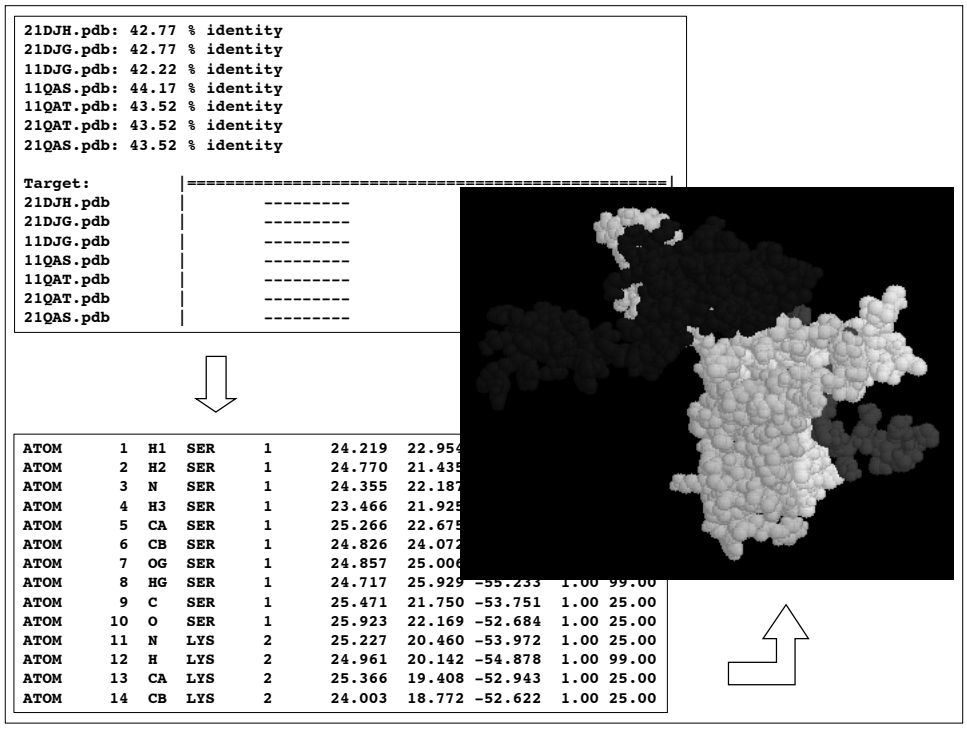
Generate models

↓

Do energy minimization

↓

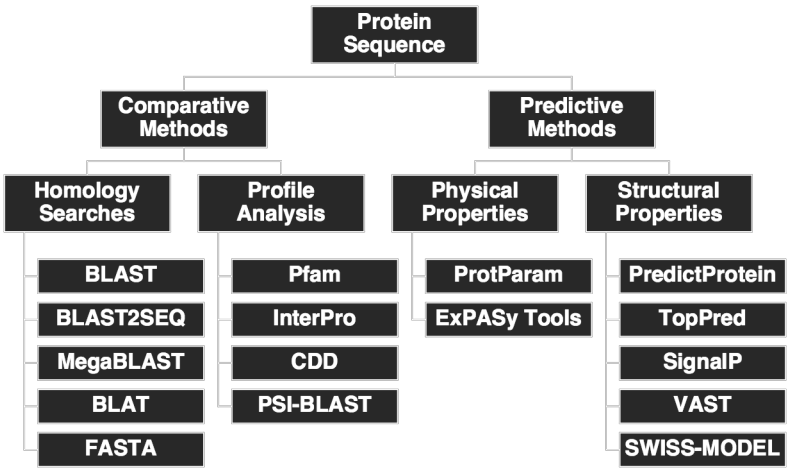
Generate PDB file for new protein model



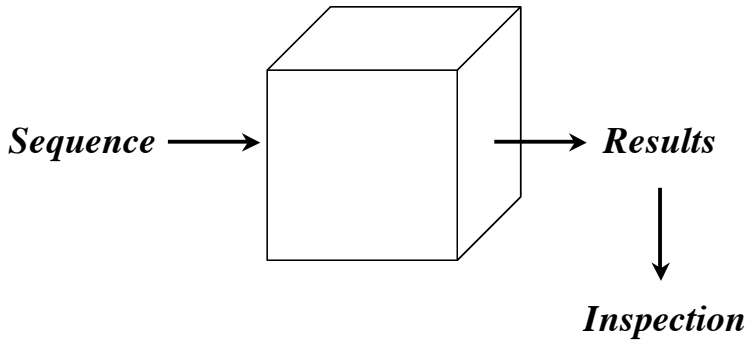
Structural Modeling Software

- Modeller
  http://www.salilab.org/modeller/about_modeller.html
- DeepView
  http://us.expasy.org/spdbv/
- WHAT IF
  http://swift.cmbi.kun.nl/whatif/

Protein Sequence Analysis



Understanding Analyses



A User's Guide to the Human Genome II

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

Commentary: Keeping Biology in Mind