Premise:

Most proteins that share a common evolutionary past
-- homologs -- do not exhibit statistically significant
amino acid sequence identity.

Therefore, the ability to generate hypotheses of function
by the exclusive use of methods dependent on statistical
significance is limited and restricts assignment of function
by combined computational and experimental strategies.

Premise:

Although high sequence identity and low probability of random matching is good evidence of homology,

it is inappropriate to assert that two proteins are not homologs simply because sequence identity is < 25% and the expectation value exceeds an arbitrary cutoff.

Criteria developed for annotation are not necessarily optimal for recognition of homology or for generation of hypotheses of function.

An opportunity exists to implement homology search algorithms that are based on a "forensic" (clues-based) analysis of genomic data.

Extreme Protein Homology Search

(work in progress)

Ig, Cupredoxins

Psi-BLAST linked pseudoazurin to a chicken antibody light chain.

```
PAZ  19  ....EPAYIKANPGDTVTFI---------------------PVDKGHNVESIKDMI 49
GAL   5  AALTQPSSVSANPGETVKITCSGGSNSYG------VFQQKSPGSAPVTVIYWDDER 55
MCG   1  SALTQPPSASGSLGQSVTISCTGTSSDVGGYNYVSWFQQHA-GKAPKVIIYEVNKR 56

PAZ  50  PEGA-EKFKSKINENYVLTVTQPGA-----YLVKCTPHYAMGMIALIAVGDSPANL  99
GAL  52  PSGIPSRFSGSKSGS-THTLTITGVQAEDEAVYYCGSIDSSSGYAGFGAGTTLTVL 110
MCG  53  PSGVPDRFSGSKSGN-TASLTVSGLQAEDEADYYCSSYEGSDNFV-FGTGTKVTVL 109
```

PAZ	1PAZ	Pseudoazurin
GAL		Chicken [*Gallus gallus*] immunoglobulin lambda light chain
MCG	2MCG	Human immunoglobulin lambda light chain

Ig, Cupredoxins

Psi-BLAST linked pseudoazurin to a chicken antibody light chain.

```
PAZ  19  ....EPAYIKANPGDTVTFI---------------------PVDKGHNVESIKDMI 49
GAL   5  AALTQPSSVSANPGETVKITCSGGSNSYG------VFQQKSPGSAPVTVIYWDDER 55
MCG   1  SALTQPPSASGSLGQSVTISCTGTSSDVGGYNYVSWFQQHA-GKAPKVIIYEVNKR 56

PAZ  50  PEGA-EKFKSKINENYVLTVTQPGA-----YLVKCTPHYAMGMIALIAVGDSPANL  99
GAL  52  PSGIPSRFSGSKSGS-THTLTITGVQAEDEAVYYCGSIDSSSGYAGFGAGTTLTVL 110
MCG  53  PSGVPDRFSGSKSGN-TASLTVSGLQAEDEADYYCSSYEGSDNFV-FGTGTKVTVL 109
```

PAZ 1PAZ Pseudoazurin
GAL Chicken [*Gallus gallus*] immunoglobulin lambda light chain
MCG 2MCG Human immunoglobulin lambda light chain

Between Twilight and Midnight:
The Ephemeral Zone of Apparent Homology

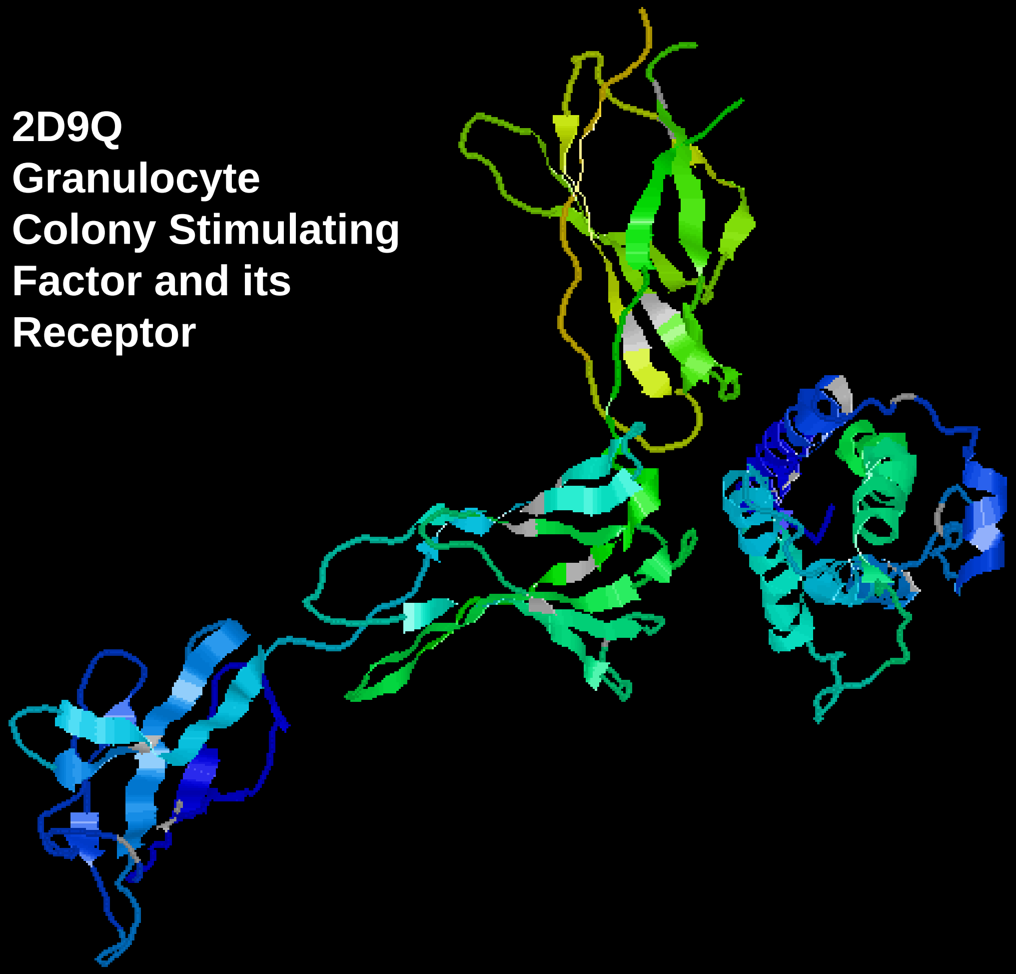
The match between pseudoazurin and the chicken light chain was obtained in the 3rd round of Psi-BLAST: 20% identity, E = 1.3. These values are not conventionally considered to be consistent with homology.

Current searches do not find this match, presumably due to changes in the composition of the database, which has doubled in size.

However, matches are found to another member of the immunoglobulin superfamily

```
gi|266418
Macrophage colony-stimulating factor 1 receptor precursor (CSF-1-R)
[Rattus rattus] Length=978 Expect = 0.29,
Identities = 14/76 (18%), Positives = 19/76 (25%),


Query 19 EPAYIKANPGDTVTFIPVDKGHNVESIKDMIPEGAEKFKSKINENYVLT-VTQPGAYLVK 77
Sbjct 26 SGPELVVEPGETVTLRCVSNG-SVEWDG-PISPYWTLDPESPGSTLTTRNATFKNTGTYR 83

Query 78 CTPHYAMGMIALIAVG 93
Sbjct 84 CTE-LEDPMAGSTTIH 98
```

2D9Q
Granulocyte Colony Stimulating Factor and its Receptor



Laccase exhibits potential homology to heavy chain variable domains.

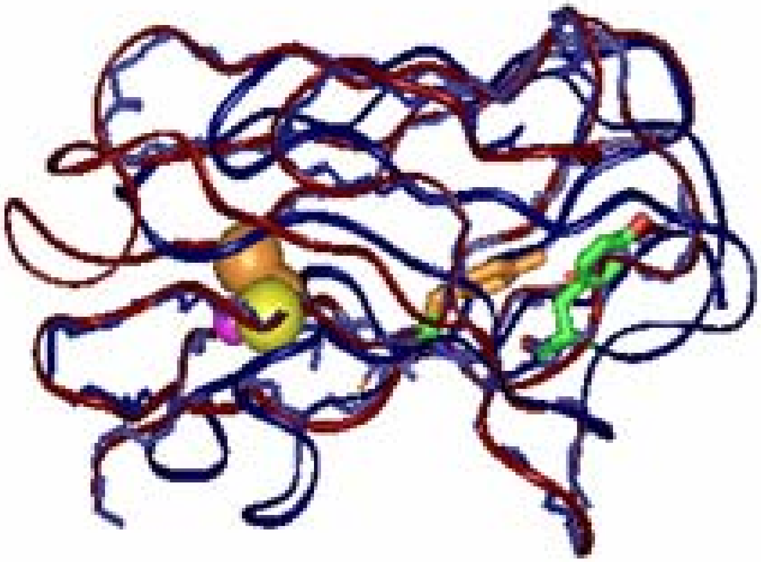
```
LAC  355 IISGAQNA-QDLLPSGSVYVLPSNADIEISFPATAAAPGAPHPFHLHGHAFAVVRSAGSTV-------YNYDNPIFRDV 425
PSP   15 DLMADEPP-SDLSKVTIAANVKNATYRNFVEIIFENREKTIQTYHLDGYSFFAVAIEPGKWSPEKRKNYNLVDAVSRHS  82
VH1   63 ...........................................................................RKVT  65
VH2   42 ...............................................GNKLEYMGYISFSGNTFYHPSLKSRISITRDT  73
VL     2 ...................................................................SYELTQPP---S  10

LAC  426 VSTGTPAAGDNVTIRF----RTDNPGPWFLHCHIDFHLE------AGFAVVFAEDIPDVA--SANPVPQAWSDLCPTYDARD  495
PSP   83 IQVYP---NSWAAVMT----TLDNAGMWNLRSDMWEKFY------LGQQLYFSVLSPSGSLRDEYNLPDNHP-LCGIVKGMP  160
VH    66 LTVDK---SSSTAYMQLSRLTSEDSA--VYYCARTNWERNYAMDYWGQGTSVTVSSAKTTAPSVYPL---AP-VCGGTTG..  137
VH2   74 SKNQH---YLQLSSVT----TEDTA---TYYCANWDGTY------WGEGTLVTVSAAKTTAPSVYPL---AP-VCGDTTG..  133
VL    11 VSVAP---GQTARIIC----GADNIGDKSVH---WYQQK------PGQAPVLVVYDDRDR-PSGIP................   59
```

```
2PLT   3 ..TVKLG--ADSGALEFVPKTLTIKSGETV--NFVNNAGFPHNI---------------VFDEDAIPSGVNADAISRDD  61
1KCW 970 ....................................NEIDLHTVHFHGHSF---------QYKHRGVY-------SSDV 996
VH1   63 ...........................................................................RKVT  65
VH2   42 ...............................................GNKLEYMGYISFSGNTFYHPSLKSRISITRDT  73
VL     2 ...................................................................SYELTQPP---S  10

2PLT  62 YLNAP---GETYSVKL----TA--GAEYGYYCE--PHQG----AGMVGKIIV..........................   97
1KCW 997 FDIFP---GTYQTLEM----FPRTPGIWLLHCHVTDHIH----AGMETTYTVLQNE...................... 1041
VH    66 LTVDK---SSSTAYMQLSRLTSEDSA--VYYCARTNWERNYAMDYWGQGTSVTVSSAKTTAPSVYPLAPVCGGTTG..  137
VH2   74 SKNQH---YLQLSSVT-----TEDTA--TYYCANWDGTY------WGEGTLVTVSAAKTTAPSVYPLAPVCGDTTG..  133
VL    11 VSVAP---GQTARIIC----GADNIGDKSVH---WYQQK------PGQAPVLVVYDDRDR-PSGIP............   59
```

LAC 20270770 Laccase 2 [*Trametes pubescens* (mushroom)] *520aa*
PSP 4105800 Pollen-specific protein [Petunia] 167aa 16%, 0.047
VH1 41352204 Heavy chain variable domain [Mouse] Q: PSP 20%, 1.2
VH2 31615669 Heavy chain variable domain [Mouse] 1NDG Q: PSP 16%, 0.45
VL 62860980 Lambda light chain variable domain [Human] Q: PSP (questionable)
2PLT 443425 2PLT Plastocyanin [*Chlamydomonas reinhardtii* (green algae)] 98a Q:PSP
1KCW 1942284 Ceruloplasmin [Human] 1046aa 18%, 7e-04 Q:PSP

Evidence for an evolutionary link between cupredoxins and immunoglobuliins

1. Suggestive sequence similarity.
2. Structural similarity: superposition of 1PAZ (25-33) with 2MCG (12-20)
   results in a r.m.s. deviation of 1.19 A



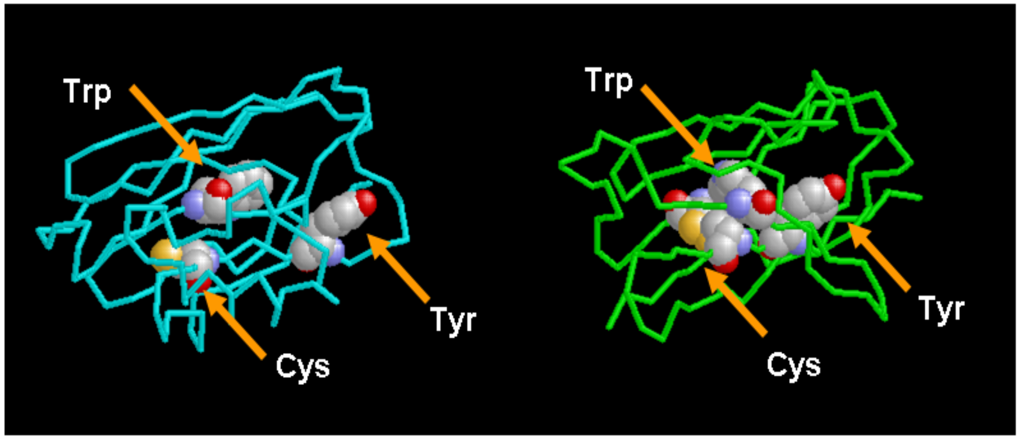
3. Structural similarity: superposition juxtaposes Cys78 (1PAZ) and Cys90 (2MCG).
4. Structural similarity: both proteins have a tyrosine corner near the Cys residues.

Hypothesis:

Immunoglobulin and Cupredoxin Superfamilies are Evolutionarily Linked

CuXn

Ig



When the structure of Superoxide Dismutase was determined, its striking similarity to the structure of immunoglobulins was considered to be due to convergent evolution.

SOD → Cupredoxins → SOD

```
SOD1       3   ..AICVMS---GDVSGQVYFKQEGPQQPVSISGFLLNLPRGLHGFHVHEFG---------DTSNGCTSAG-EHFNPTNQD-HGAPDAAER   76
SOD2      25   ..AKATLKNAEGTEIGTATLTESSKG--VTIKLALKGLPPGEHAFHIHAVG---------KCEPPFTSAG-GHFNPENKK-HGKMAEGGA  103
LAC      278  ANPPYPLITIDRDSW-DENQFSLSTGSKPVWIDFIVNNLDEGPHPFHLHGHT-------FFILSLFESTIGWGSYNPHQPH-LNPSPYPPY  360
1KCW     741   .....MHAINGRMFGNLQGLTMHVGDEVNWYLMGMGNEIDL-HTVHFHGHSFQ-----------------------------------YK  797
LAC9     413   ........................................GPHPFHLHGHAFSVVR-----------SAGSSTYNYENP-----------  430
LAC11    417   ..............................ISG......SGPHPFHLHGHAFSVVR-----------SAGNSSYNYVNPV-KRDVVSMG-  457
SOD3     107   ....................VLTETPSGVEIVAQVQGLTAGLHGFHIHANGQCDPGPDAATGKTVPFGAAGGHFDPGMSHQHGQPGAPGA  173
SOD4      24   ..TAPIYTTGPKPV-AIGKVTFTQTPYGVLITPDLTNLPEGPHGFHLHKN----------ADCGNHGMDAEGHYDPQNTNSHQGP-YGNG   99

SOD1      77   HVG--DLGNVRSV---GCT-ALT-PIEMTDNVISLFG-PLSI----LGRSLVVHTDRDDLGLTDNPLSKITGNSGGRLACGIIAVCK  151
SOD2     100   HAG--DMPNLDVP---ASG-ALS-IDVVNDAVTLAKGKPNSVFK-DGGTALVIHAKADDY------KSDPAGNAGDRIACGVIEEAK  172
LAC      357   DFS--KALERDTVQIPRRG-HAV-LRLRADNPGVWLFHCHILWHLASGMAMLLEVM...............................  408
1KCW     794   HRG---VYSSDVFDIFPGT-YQ--LEMFPRTPGIWLLHCHVTDHIHAGMETTYTVL...............................  840
LAC9     431   -------VRRDVVDVGGAS-DNVTIRFTTDNPGPWFFHCHIEFHLVLGLAMVFM.................................  485
LAC11    458   --GDSDLVTIRFV---TDNPGPW-FFH-----------CHIEPHLVGGLAIVFAEAMEDTAAAH.......................  504
SOD3     174   PIDKAHAGELPNISVGADG-RGT-VRYLNTN........................................................  202
SOD4     100   HLG--DLPVLYVTSNGKAM-IPT-LA.............................................................  121
```

SOD1	**86355642**	SOD (Cu/Zn) [*Hyphantria cunea* nucleopolyhedrovirus]
SOD2	**39933302**	putative superoxide dismutase (Cu/Zn) [Rhodopseudomonas palustris] aa=172
LAC	**76008508**	laccase-like protein [*Coccidiodes posadasii*] aa=408 (domain 3)
1KCW	**180249**	ceruloplasmin **1KCW** [Human] {query 76008508} 19%;7e-06 1046aa (12/12/06)
LAC9	**115371531**	laccase 9 [*Coprinopsis okayama7*] {query 76008508} 525 aa
LAC11	**115371535**	laccase 11 [*Coprinopsis cinerea okayama*] Rnd 3 15%; 4.5 {query 86355642} (01/04/07)
SOD3	**111618878**	superoxide dismutase [*Acidovorax avenae*] Rnd 2 22%; 7.2 {query 76008508} 260aa (01/04/07)
SOD4	**54298239**	superoxde dismutase [*Legionella pneumophila*] Rnd 4 19%; 6.1 {query 76008508} 162 aa (01/04/07)

Notes:

SOD1	**86355642**	SOD (Cu/Zn) [*Hyphantria cunea* nucleopolyhedrovirus]
SOD2	**39933302**	putative superoxide dismutase (Cu/Zn) [Rhodopseudomonas palustris] 172aa
LAC	**76008508**	laccase-like protein [*Coccidiodes posadasii*] 408aa (domain 3)
1KCW	**180249**	ceruloplasmin **1KCW** [Human] {query 76008508} 19%;7e-06 1046aa
LAC9	**115371531**	laccase 9 [*Coprinopsis okayama7*] {query 76008508} 525 aa
LAC11	**115371535**	laccase 11 [*Coprinopsis cinerea okayama*] 15%; 4.5 {query 86355642}
SOD3	**111618878**	superoxide dismutase [*Acidovorax avenae*] 22%; 7.2 {query 76008508} 260aa
SOD4	**54298239**	superoxde dismutase [*Legionella pneumophila*] 19%; 6.1 {query 76008508} 162aa

gi number ↑ **protein name** ↑ **source** ↑ **% identity & expect value** ↑ **query if not original** ↑ **size** ↑

In the alignments shown, the following residues were considered equivalent:

S = T
R = K
L = I = M

Psi-BLAST was used to identify potential homologs -- iterations were continued until highest scoring sequence had less than 30% identity to query sequence.

Percent identity and expect value were disregarded; >50% alignment was required.

Alignments shown are as provided by Psi-BLAST with minor adjustments when appropriate.

A recent search with the regular expression:

[HG].{1,2}H.{0,1}F[HYDE][GV]H.{1}F

based on the metal binding motif found in

SOD and ceruloplasmin returned 255 hits

... only 7 were false.

The rest were SOD, ceruloplasmin or laccase, a cupredoxin also known as multicopper oxidase.

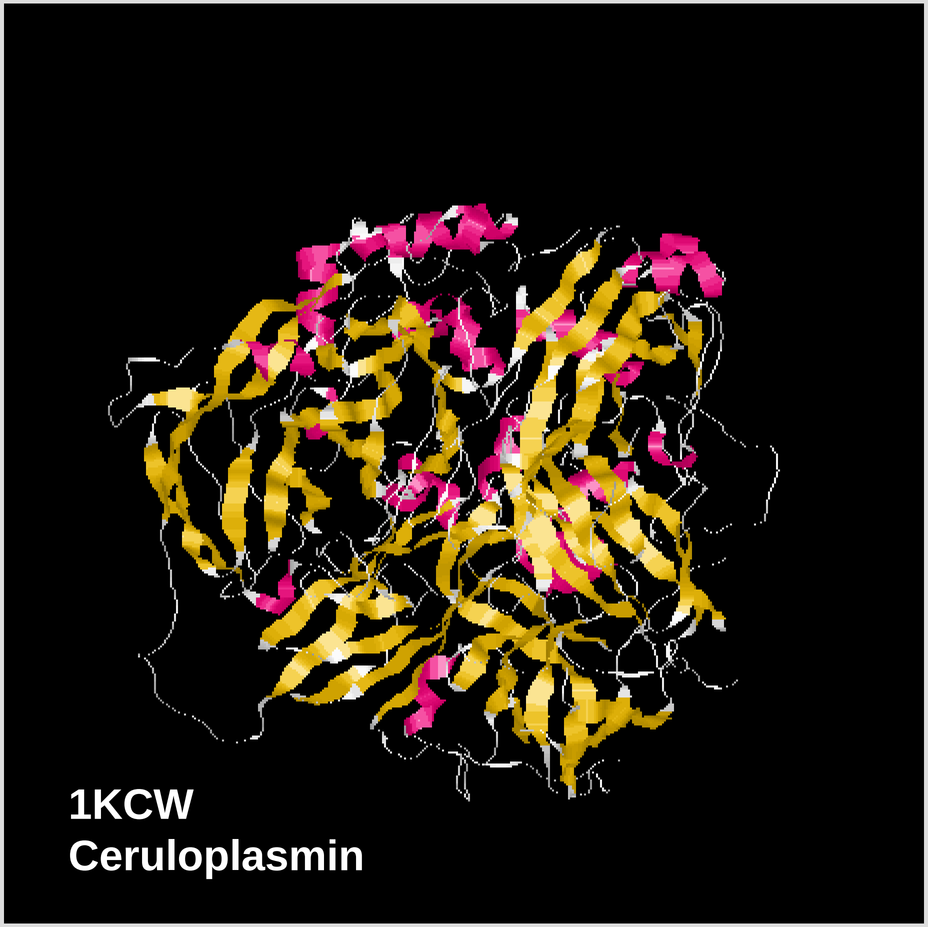
Therefore, the

[HG].{1,2}H.{0,1}F[HYDE][GV]H.{1}F

motif is almost exlusively found in

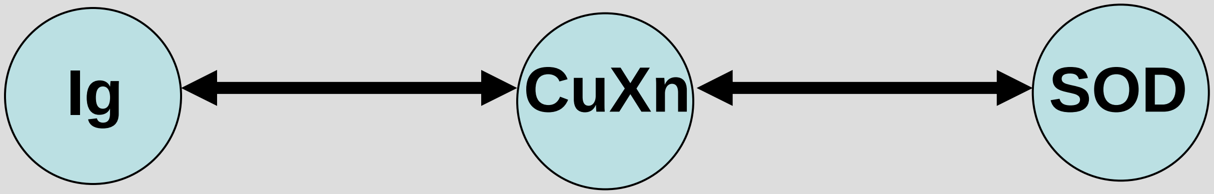
SOD and cupredoxins.

Multiple Cupredoxin Domains in Ceruloplasmin



Hypothesis:

Cu,Zn-Superoxide Dismutase and Cupredoxin Superfamilies are Evolutionarily Linked



Therefore, Ig, CuXn, and SOD superfamilies have a shared evolutionary origin.

TNF, C1q, C2, IgA constant domain, MHC-II, Multimerin, (Laccase ?)

TNF is a homolog of the head domain of C1q and a C1q-related protein.

TNF appears to be a homolog of a C2 constant domain seen in *C. elegans* and a MHC class-II protein from pig.

TNF may be a homolog of laccase, a cupredoxin. Cupredoxin is related to immunoglobulins and SOD

C1q is a homolog of a domain in multimerin -- a coagulation factor V binding protein. Coagulation factor V includes cupredoxin domains.

C1q interacts with antibody antigen complexes and some bacteria.

C1q interacts with IgG1, C-reactive protein, and pentraxin-3. Roumenina et al, Biochemistry 45, 4093 (2006)

Laccase: A copper-containing enzyme, 1,4-benzenediol oxidase (EC 1.10.3.2), found in higher plants and microorganisms. Laccases are multicopper oxidases of wide specificity that carry out one-electron oxidation of phenolic and related compounds, and reduce O2 to water. The enzymes are polymeric and generally \contain one each of type 1, type 2, type 3 copper centers per subunit, where the type 2 and type 3 are close together forming a trinuclear copper cluster.

TNF, C1q, C2, IgA constant domain, MHC-II, Multimerin, (Laccase ?)

a)

```
TNF   85  SDKPVAHVVAN-------PQAEGQLQWLNRRANALLANGVELRDNQL--------VVPSEGLYLIYSQVLFKGQGCPSTHVLLTHT  155
C1q  181  ..............................QGAFLRGSGLSLASGRF--------TAPVSGIFQFSASLHVDHSELQGKARLRA--  228
 C2   15  SSEPAGYVVIACLVRDFFPSEPLTVTWSPSREGVIVRNFPPAQAG---------------GLYTMSSQLTLPVEQCPADQILKCQV   85
IgA  363  .......................DIKWV--GPNGFKQNGSRLTITSISKAQSGNYTCMATNFLTVYGH---SGSQQRMGTGTTIVD  420

TNF  156  ISRIA-----VSYQTKVNLLSAIKSPCQRETPEGAEAKPWYEP----IYLGGVFQLEKGDRLSAEINRPD-YLDFAESGQVYFGI   219
C1q  229  -------------RDVVCVLICIESLCQRHTCLEAVSGLESNSRVFTLQVQGLLQLQAGQYASVFVDNGSGAVLTIQAGSSFSGL   298
 C2   86  QHLSK-----SSQSVNVPCKVLPSDPCPQCCKPSLSLQP..............................................   119
IgA  421  VKRKPGQAQIVSARQNVDVGETIKLMCQAEDAGNPSASYTWAS----PSSGGIFGLEGHTEKSFEVRNAQ-LSD...........   489
```

b)

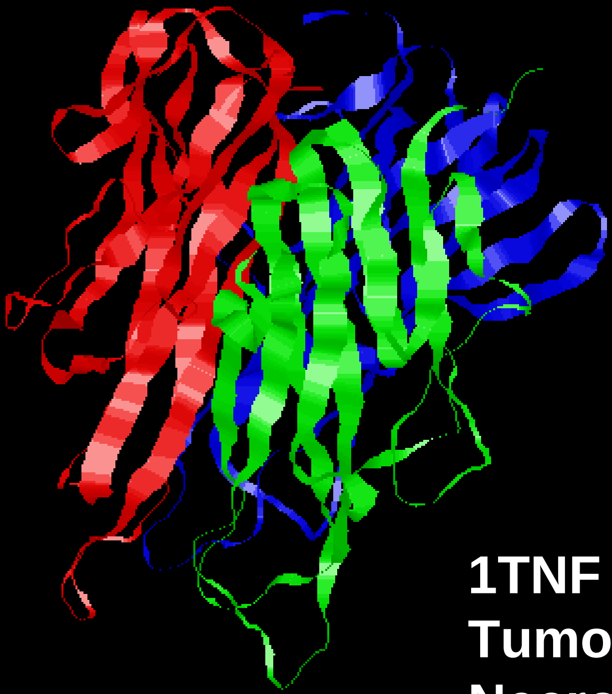
```
C1q    1  TQKIAFSAT--RT---INVPLRRDQ---TIRFDH-VITNNMN--NYEPRSGKFTCKVP-CLYYFTYHAS-----SRGNLC----   63
TNF  102  ............................QWLNRR-ANALLAN--GVELRDNQLVVPSE-GLYLIYSQVL-----FKGQGCPSTH  148
MHC   27  TRPSAFFYLSHRS---QCQYLNGTE---RVRYTQRYIYNRQQYTHYDSDLGKFVADTPLG----EFQAE-----HYNSQT----   96
 MM  824  .SPVAFYAS--FS---EG--TAALQ---TVKFNT-TYINIGS--SYFPEHGYFRAPER-QVYLFAVSVEFGPGPGTGQLV----  888
Lac  423  ....AFSVV--RS---AG--SSTYNYENPVRRD--V--VDVG--ASDNVTIRFTTDNP-GPQFFHCHIEFHLVLGLAMVF----  465

C1q   55  ----------------VNLMKGRERAQKVVTFCDYAYNTFQVTTGGMVLKLEQGENVFLQATDKNSLLGMEGANSIFSGFLLFP  131
TNF  149  VLLTHTISRIAVSQTKVNLSAIKSPCQRETAEGAEAKPWYEPIYLGGVFQLEKNDRKSAEINRPDYLDFAESGQVYF.......  225
MHC   97  ----------------EILEVKRAEV---DTFCRHNYGVFE----SITVQRSVQPKVRVSALQSGSLGESDRLACYVTGF--YP  150
 MM  889  ----------------FGGHH-RTPV------CTTGQGSGSTATVFAMAELQKGERVWFELTQ-GSITKRSLSGTAFGGFLMFK  948
```

----

a) TNF, human TNF-alpha gi|25952111; human C1q, gi|59807841; C2, *C. elegans* C2 constant domain in gi|2136551; IgA, pig heavy chain alpha constant domain gi|17567577

b) C1q, human C1q globular head domain gi|38492827 -- positions 1 to 131 in PDB entry 1PK6 correspond to positions 119-250 in the cognate protein, gi|87298828; TNF human TNF-alpha gi|25952111; MHC, *Coturnix japonica* (Japanese quail) MHC class 2 gi|46091121; MM, human multimerin gi|63101264; Lac, *Coprinopsis cinerea okayama* laccase gi|115371531.

TNF Trimer

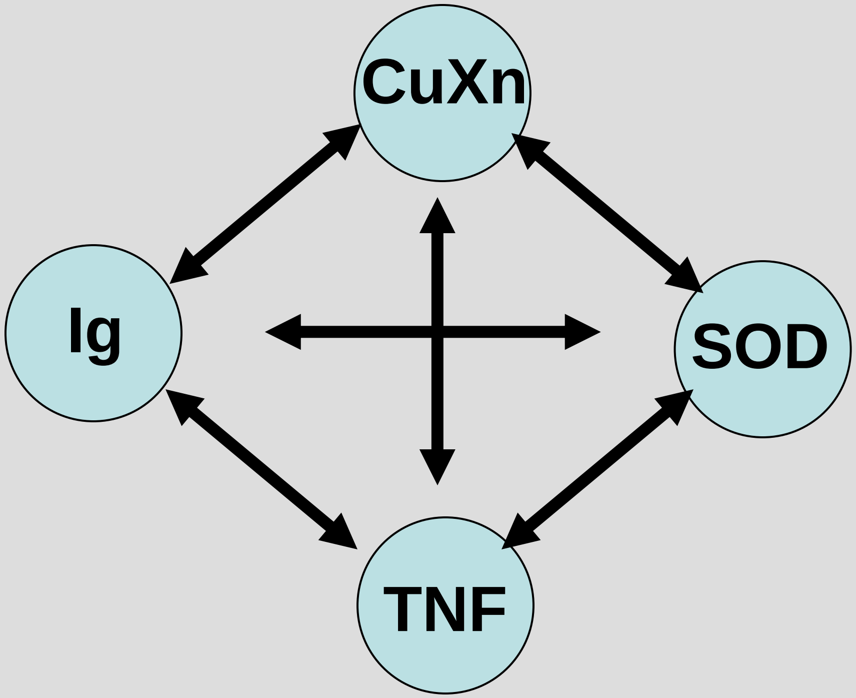


1TNF
Tumor
Necrosis
Factor

Trimeric "Head" of Complement C1q

1PK6
C1q

Hypothesis:
Tumor Necrosis Factor and Immunoglobulin Superfamilies are Evolutionarily Linked



Therefore, Ig, CuXn, SOD, and TNF superfamilies have a shared evolutionary origin.

The relationship between TNF and Ig links the innate and adaptive immune systems.

Interleukin-12 is evolutionarily related to immunoglobulins

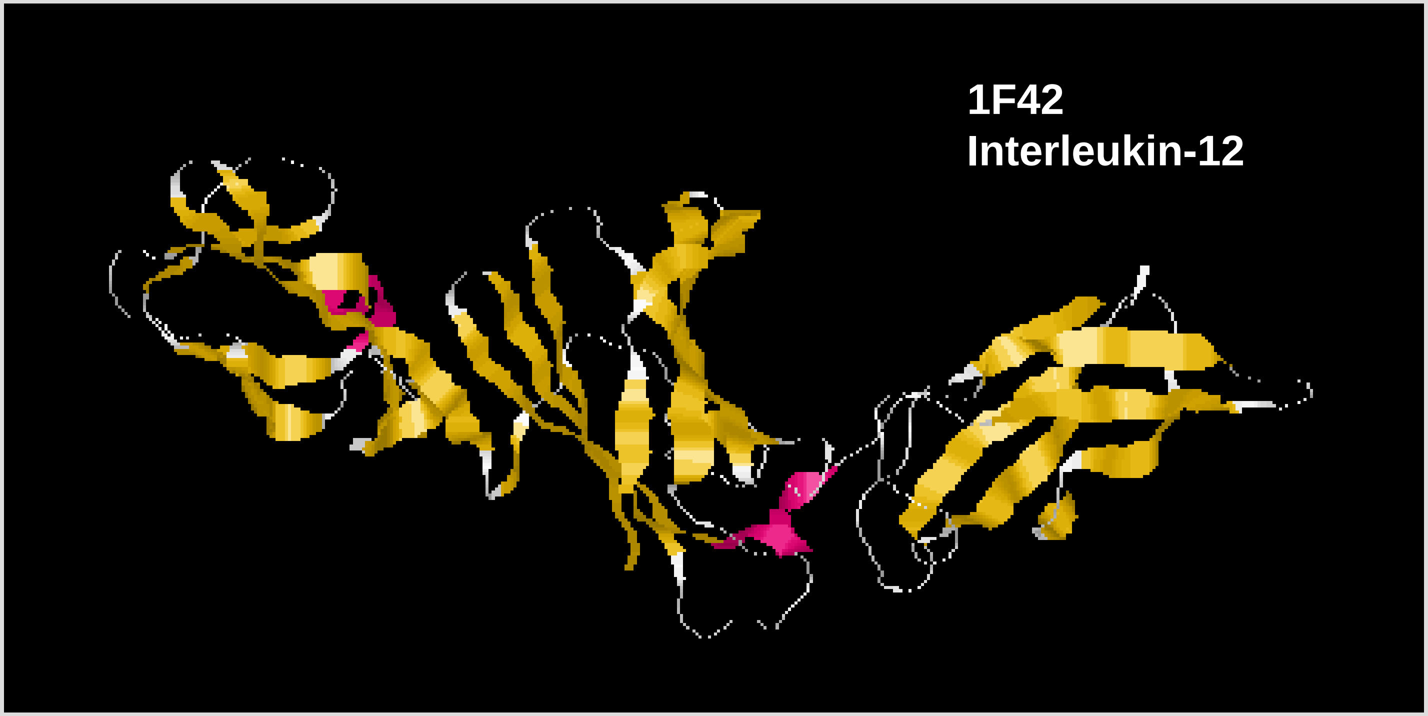
```
IL12   6 KDVYVVELDWYPDAPGE--MVVLTCD--T---PEEDG--ITWTL-------------------------DQSSEVLGSG  50
IL6R  20 ..............PGD--SVTLTCPGVE---PEDNAT-VHWVL-------------------RKPAAGSHPSRWAGMG  59
LC1   35 .........WYQQKPGQAPVLVVYSD--T---DRPSG--IPE---------------------------RFSGSFSGNT  69
LC2   33 .........WYQQKPGQDPVLVIYSD--S---NRPSG--IPE---------------------------RVSGSNPGNT  68
LC3   49 ..VYNNGLAWYQQKPGEAPKLL-----------------IYFA------------------TLQSGIPSRFSGSGSG-T  91
LC4   34 .............TPGE--KVTITCQASW---EGIGNY-LYW-QQKPDQAPKLLIKYASQS--ISGVPSRFSGSGSG-T  90
LC5    9 ........DFQSVTPKE--KVTITCRASQ----SIGSS-LHWYQQKPDQSPKLLIKYASQSQSFSGVPSRFSGSGSG-R  70
LC6   20 GQISVTQSPSTAAQPGE--TVKISCK--TSTDVYDGDS-LFWYLQKPGEAPKLLIYLA--NTLESGTPSRFSGSGSN-S  91
HC    20 .............TAQQ--SVTLTCL--VKDFAPKEIF-VQWTVDDKEIDVSN----------------YKNTELMADS  64
FGFR  72 .............APGE--SLEVRCL--LKDAAV-----ISWTKDGVHL--------------------GPNNRTVLIG 108
RAGE 253 EVQLVVEPEGGAVAPGG--TVTLTCE--V---PAQPSPQIHWMK-------------------------DGVPLPLPPS 299

IL12  51 --------K-TLTIQVKEFGDAGQYTCHKGGEVLSHSLLLLHKKEDGIWSTD  93
IL6R  60 --------R-RLLLRSVQLHDSGNYSCYRAGRPAGTVHLLV...........  91
LC1   70 --------A-TLTISRVEAGDEADYYCH........................  79
LC2   69 --------A-TLTISWXIEADEADYYCRCGTVVV..................  93
LC3   92 --------DF-LTISGVQAEDAGDYYCQSYHEPSSRYVFM............ 122
LC4   91 --------DFTFTISSLEAEDAATYYCQQGNK.................... 113
LC5   71 --------DFTLTISGVQTEDAATYYCHQ.......................  90
LC6   92 -------------------EDAGHYYCQSAHYISS................. 117
HC    65 ADRNYSMYS-MLTISAGEWGRGFSYSCIVGHET...................  96
FGFR 109 --------E-YLQIKGATPRDSGLYACTASRTVDSETWYFMVNVTDAISSGD 151
RAGE 300 --------P-VLILPEIGPQDQGTYSCVATH..................... 321
```

IL12	14719640	1F42 Interleukin 12 [Human] (1-100)/306
IL6R	27574042	1N26 Interleukin-6 Receptor [Human] 325aa 27%, 0.008
LC1	13235110	Light chain variable region [Human] 125aa 36%, 3.0
LC2	4761433	Immunoglobulin lambda light chain variable [Human] 95aa 33%, 5.5
LC3	5042056	Immunoglobulin light chain [*Acipenser baerii* (sturgeon)] 136aa 30%, 9.8
LC4	87867	Ig kappa chain precursor [Human] 115aa 30%, 0.22
LC5	109240688	Ig kappa chain variable region [Human] 106aa 29%, 7.7
LC6	20269263	Ig light chain [*Cyprinus carpio*] 203aa 24%, 6.0
HC	2895077	Ig heavy chain constant region [*Hydrolagus colliei* (Ratfish)] 130aa 24%, 4.4
FGFR	119569745	Fibroblast growth factor receptor [Human] 385aa 27%, 3.8
RAGE	59799500	Receptor for advanced glycation endproducts [RAGE] 363aa 29%, 3.0

Interleukin-12 is evolutionarily related to immunoglobulins

```
IL12   6 KDVYVVELDWYPDAPGE--MVVLTCD--T---PEEDG--ITWTL-------------------------DQSSEVLGSG  50
IL6R  20 ..............PGD--SVTLTCPGVE---PEDNAT-VHWVL---------------RKPAAGSHPSRWAGMG----  59
LC1   35 .........................................WYQQKPGQAPVLVVYSD--TDRPSGIPERFSGSFSGNT  69
LC2   33 .........................................WYQQKPGQDPVLVIYSD--SNRPSGIPERVSGSNPGNT  68
LC3   49 .................................VYNNG-LAWYQQKPGEAPKLL---IYFATLQSGIPSRFSGSGSG-T  91
LC4   34 .............TPGE--KVTITCQASW---EGIGNY-LYW-QQKPDQAPKLLIKYASQS--ISGVPSRFSGSGSG-T  90
LC5    9 ........DFQSVTPKE--KVTITCRASQ----SIGSS-LHWYQQKPDQSPKLLIKYASQSQSFSGVPSRFSGSGSG-R  70
LC6   20 GQISVTQSPSTAAQPGE--TVKISCK--TSTDVYDGDS-LFWYLQKPGEAPKLLIYLA--NTLESGTPSRFSGSGSN-S  91
HC    20 .............TAQQ--SVTLTCL--VKDFAPKEIF-VQWTVDDKEIDVSN----------------YKNTELMADS  64
FGFR  72 .............APGE--SLEVRCL--LKDAAV-----ISWTKDGVHL--------------------GPNNRTVLIG 108
RAGE 253 EVQLVVEPEGGAVAPGG--TVTLTCE--V---PAQPSPQIHWMK-------------------------DGVPLPLPPS 299

IL12   51 --------K-TLTIQVKEFGDAGQYTCHKGGEVLSHSLLLLHKKEDGIWSTD  93
IL6R   60 --------R-RLLLRSVQLHDSGNYSCYRAGRPAGTVHLLV...........  91
LC1    70 --------A-TLTISRVEAGDEADYYCH........................  79
LC2    69 --------A-TLTISWXIEADEADYYCRCGTVVV..................  93
LC3    92 --------DF-LTISGVQAEDAGDYYCQSYHEPSSRYVFM............ 122
LC4    91 --------DFTFTISSLEAEDAATYYCQQGNK.................... 113
LC5    71 --------DFTLTISGVQTEDAATYYCHQ.......................  90
LC6    92 -------------------EDAGHYYCQSAHYISS................. 117
HC     65 ADRNYSMYS-MLTISAGEWGRGFSYSCIVGHET...................  96
FGFR  109 --------E-YLQIKGATPRDSGLYACTASRTVDSETWYFMVNVTDAISSGD 151
RAGE  300 --------P-VLILPEIGPQDQGTYSCVATH..................... 321
```

```
IL12  14719640 1F42 Interleukin 12 [Human] (1-100)/306
IL6R  27574042 1N26 Interleukin-6 Receptor [Human] 325aa 27%, 0.008
LC1   13235110 Light chain variable region [Human] 125aa 36%, 3.0
LC2    4761433 Immunoglobulin lambda light chain variable [Human] 95aa 33%, 5.5
LC3    5042056 Immunoglobulin light chain [Acipenser baerii (sturgeon)] 136aa 30%, 9.8
LC4      87867 Ig kappa chain precursor [Human] 115aa  30%, 0.22
LC5  109240688 Ig kappa chain variable region [Human] 106aa 29%, 7.7
LC6   20269263 Ig light chain [Cyprinus carpio] 203aa  24%, 6.0
HC     2895077 Ig heavy chain constant region [Hydrolagus colliei (Ratfish)] 130aa 24%, 4.4
FGFR 119569745 Fibroblast growth factor receptor [Human] 385aa 27%, 3.8
RAGE  59799500 Receptor for advanced glycation endproducts [RAGE] 363aa 29%, 3.0
```


1F42
Interleukin-12

Fibronectin-III

A structural module in a multitude of proteins -- some very large

```
FnIII    14   ........................SITLSWTASTDNVGVTGYDVY-NGTALATTVT--------------   49
Aph      64   ..................ADSN-ASVPLTWQVATDSS-FAN-IVS-SGSVEALPAN--------------   97
Mph     187   FANP---KKPLYGHISSIDSTA-TSMRLTW-VSGDKE-PQQ-IQYGNGKTVTSAVTTFSQEDMCSSVVPS  249
KBPAP    18   FRVPPGYNAPQQVHITQGDLVG-RAMIISW-VTMDEPGSSA-VRY------------WSEKNGRKRIAKG   72
Fn-a      6   ....SGQAAPSQVVVIRQERAGQTSVSLLW-QEPEQPNGII-LEY------------EIKYYEKDKEMQS   57
Fn-b     12   LRVRQLPHAPEHPVATLSTVER-RAINLTW-TKPFDGNSPL-IRY----------ILEMSENNAPWTV--   66

FnIII    50   ------------GTTATI--SGLAADTSYTF---TVKAKDAAGNVSAASNAVS...........   85
Aph      98   ------------GFTAKVDATGLSAGASYFY-----RFRDAAGTTSTVGATRTLPAASVASVKF  144
Mph     250   --PAKDFGWHDPGYIHSALMTGLKPSSAYSY-----RYGSNSADWSEQTKFSTPPAGGSDELKF  306
KBPAP    73   KMSTYRFFNYSSGFIHHTTIRKLKYNTKYYY-----EVGLRN--TTRRFSFITPPQTGLDVP..  127
Fn-a     58   YSTLKAVTT-------RATVSGLKPGTRYVF-----QVRART--SAGCGRFSQAMEVETGKP..  105
Fn-b     67   -LLASVDPKATS-----VTVKGLPARSYQFRLCAVNDVGKGQ--FSKDTERVSLPESGP.....  118
```

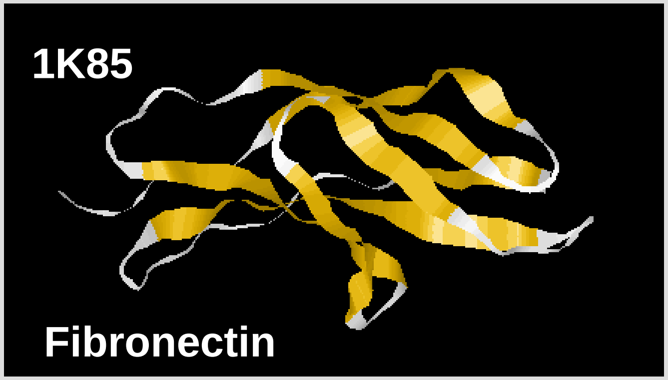
FnIII 27573694 (1K85) Fibronectin-III domain
Aph 121604116 Alkaline phosphatase* [*Polaromonas naphthalenivorans* CJ2] 600aa 30%; 3.4
Mph 92867261 Metallophophoesterase; Fibronectin III [*Medicago truncatula*] 602aa 19%; 0.13 {Query: 121604116}
KBPAP 1827635 (1KBP) Kidney bean purple acid phophatase 432aa 22%; 2e-13 {Query 92867261}
Fn-a 83753623 (1X5L) Second Fn3 domain of Ephrin Receptor 111aa 16%; 3.3{Query: 1827635}
Fn-b 71041818 (1WF5) First Fn3 domain of sidekick-2 121aa 17%; 0.31 {Query: 71041818}

*formerly annotated as "Twin arginine translocation pathway" (probably contains a cupredoxin domain)

Fibronectin-III

```
Aph    9  EFLLKTAAAVVAGSSVVACGGSDDSTPVPPAQFTYGVASGDPLSDRVILWTYAKVADSNA 68
VH1    6  RFLFVVAAATGVQSQVQLVQSGAEVKKPGSSVKVSCKASGGTFSSYAISWVRQAPGQGPE 63
VH2   26  .....................................ASGDIFSTYVISWVRQAPGQG-- 46
VH3   14  .....................................ASGFTFSSYVIHWVRQAPGKG-- 34
VH4   24  .....................................ASGGTFSNYAISWVRQAPGQG-- 44
VH5   24  .....................................ASGDTKSSFVISWVRQAPGQG-- 44
VH6   24  .....................................ASGNSFGDNAVTWVRQAPGQG-- 44
VH7   16  .....................................ASGFTFSSSAMSWVRQAPGKG-- 38
VH8   24  .....................................ASGFSFSNYAMYWVRQAPGKG-- 44
VH9    6  RFLFVVAAATGVQSQMQVVQSGAEVKKPGSSVTVSCKASGGTFSNYAISWVRQAPGQG-- 63

Aph   69  SVPLTWQVATDSSFANIVSS----GSVEALPANG--FTAKVDATGLSA-GASYFYRFRDAAGTTSTVGAT...  131
VH1   64  ---PEWLGGITPIFGTANYAQNFQGRVTIT-ADESTSTAYMELSSLKSEDTAIYYCARDETGGHGTPGAF...  129
VH2   47  ---LEWMGGIVPTVGITKFAQKFQGRVTIT-ADTSTSTAYMDLNGLRPEDTAVYY..................   97
VH3   35  ---LEWVAVIWYNGSDKYYADSVKGRLTIS-RDNSKNTVYLELNRLRAEDTAVYYCARDQAMVRGVIGR....   99
VH4   45  ---LEWMGGTTPVFGTAHYAQKFQGRVTII-ADTSTSTVYMDLSRLRSEDTAIYYCVRV..............   99
VH5   45  ---LEWMGGINPIFGTPNYAQKFQGRVTIT-ADESTSTAYMELSSLRSEDTAVYYCARPQTTVTTP.......  106
VH6   39  ---LEWMGGIIPIDST-HYAQNFQGRVTMT-ADGSTGTAYMELSSLKPDDTAIYYCARVIRGTSGWIA.....  108
VH7   45  ---LEWVAWKYENGNDKHYADSVNGRFTIS-RNDSKNTLYLIMNSLQAEDTALYYCARDAGPYVSPT......   99
VH8   45  ---LEWVAVIWYDASAEYYADSVKGRFTVS-RDNSNNTLYLQMHNLSAEDTAIYYCARDRGHVS.........  104
VH9   64  ---LEWMGGITPLFGTPTYSQNFQGRVTIT-ADKSTSTAHMELISLRSEDTAVYYCATDRYRQANFDRARVGWF 133
```

Aph	121604116	Alkaline phosphatase [*Polaromonas naphthalenivorans* CJ2] 600aa
VH1	1813670	Immunoglobulin heavy chain [human] 18%; 3.1
VH2	106897341	B cell antibody heavy chain variable region [human] 127aa 22%; 2.6
VH3	4837985	immunoglobulin heavy chain variable region [human] 125aa 23%; 0.97
VH4	4456510	immunoglobulin heavy chian variable region [human] 116aa 20%; 3.4
VH5	118405991	immunoglobulin heavy chain variable region [human] 122aa 21%; 3.4
VH6	80975571	immunoglobulin heavy chain variable region [human] 119aa 18%; 6.2
VH7	484836	Ig heavy chain V region (clone POM) [human] 114aa 20%; 7.6
VH8	33319476	Ig heavy chain variable region [human] 126aa 20%; 8.4
VH9	185362	IgG 476aa 20%; 9.5



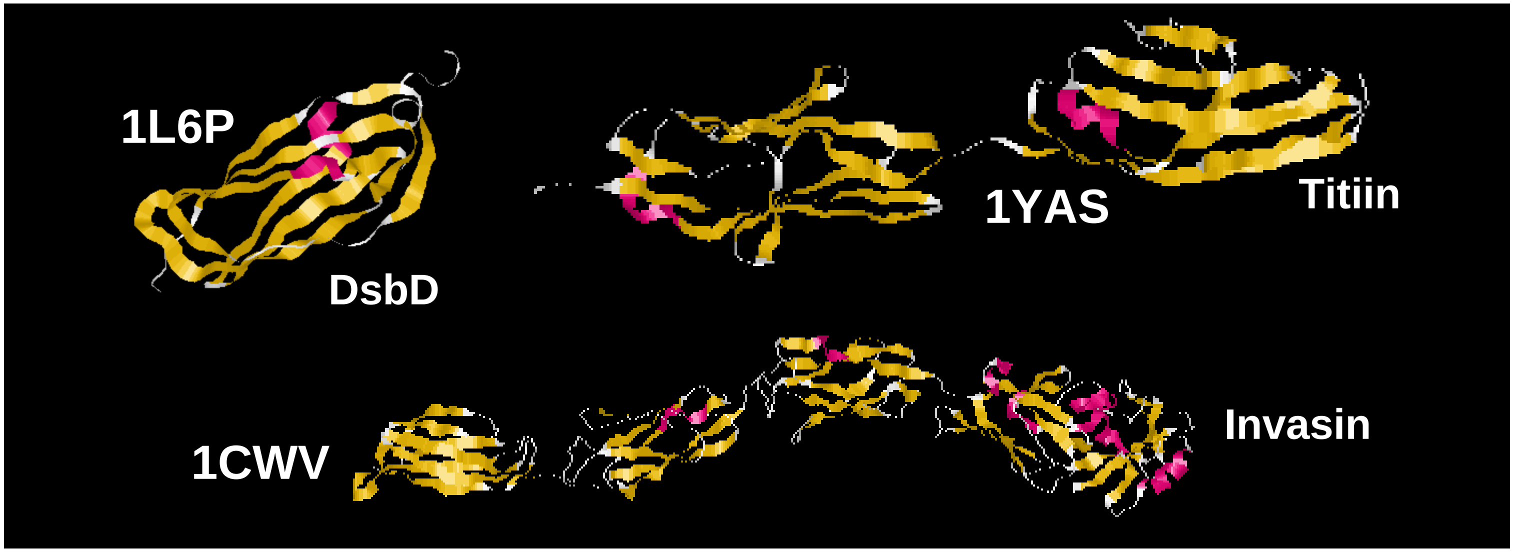
Thiol-Disulfide Interchange Protein

Facilitates the formation of correct disulfide bonds in some periplasmic proteins and for the assembly of the periplasmic c-type cytochromes. Acts by transferring electrons from cytoplasmic thioredoxin to the periplasm. This transfer involves a cascade of disulfide bond formation and reduction steps

Thiol-Disulfide Interchange Protein

```
1L6P         13   PADQAFAFDFQQ------NQHDLNLTWWQIKDG------Y-------YLYR-KQIRITPEHAKIAD   57
BAP        1255   PVSDAFSFSTTQ------NEKEVKLKFKIPAGTSVGGKRYNLIATDKYGYKSEQVSFTITVVSKPK 1314
1YA5         40   .FRDGQVISTSTLPGVQISFSDGRAKLTIPAVTKANSGRYSLKATNGSGQATSTAELLVKAETAPP  104
INVQ        419   ..................................VNTYTLSATAIDNHGNSSNPATLTVIVQQPQF  450
1CWV          1   .........................................................SVTVQQPQL    9

1L6P         58   VQLPQGVWHEDEFY-GK.........................................   73
BAP        1315   IVF-KNKDKEDYNA-GENVTMT--FTNENTDTTNKKFTVYVKVGDNEPRAIPVEVRPE 1329
1YA5        105   NFV-QRLQSMTVRQ-GSQVRLQ--VRVTGIPTPVVKF------RDGAEIQSSLDFQIS  153
INVQ        451   VIT-SEVTDDGALADGRTP-ITVKFTVTNIDGTPVAEQEGVITTSNGALPSKV.....  481
1CWV         10   TLT-AAVIGDGAPANGKTA-ITVEFTVADFEGKPLAGQEVVITTNNGALPNKI.....
```

1L6P	21730662	DsbD Thiol-disulfide interchange [*Escherichia coli*]
BAP	123503827	Bap-like [*Trichomonas baginalis*] 22%, 2.7
1YA5	85543964	Titin domains Q:123503827 17%, 4.2
INVQ	793895	Invasin Q; 1YA5 22%, 0.020
1CWV	6435735	Invasin Q: 793895 49%, 1e-18



Carbohydrate binding modules may be related to immunoglobulins

```
ENDO      1355  ENQAP---KAIFTFSPEDPV--TDENVVFNASNSIDED-GTIAYYAWDFGDGYEGTSTTPTIT-- 1411
1WGO        26  ..TPPRGLQVSIQGEAVAVR--PGEDVLFVVRQ-RQGD-VLTTKYQVDLGDGFKAMYVNLTLTGE   84
PKD        305  ....P---SADFKSNITSGYIFLSEPVQFT-----DLS-KDATSWKWDFGDGSSSKKQNP--T--  352
Hoc1       104  .............ASPAAGV--IGTPVQFTAALASQPD-GASATYQWYVDDSQVGGETNSTFS--  150
Hoc2         4  ........TASISPLDPSVL--EGSNVDFTVTFSGDETVKEVVGYEWLVDDVAQSGETSTTFR--   56
Hoc3a      197  ...........ITPESPTTV--FGVPITLTANVSGAPS-GATTSFQWSMDDSNILDATSATYK--  245
Hoc3b       94  ENNST---VAVTPASPAAVE--IGTATTFTANVSNQPS-GAAIAYTWKVDGVAVDGQKQSTFE--  150
Hoc4       105  ..............SPAAGV--IGTAVEFTAALASQPS-GATATYQWYVDDSPVGEATSATFN--  150

ENDO      1412  ---YKYKNPGTYKVKL----IVTD---NQGAS-SS---FTATIKVTSATGDNSKFNFEDGT 1458
1WGO        85  PIRHRYESPGIYRVSV----RAEN---TAGH--DE---AVLFVQVSGPS............  121
PKD        353  ---HTYSETGIYTVRL----TVSN---SNGTD-SQ---ISTVNVVLKGSPTPS........  391
Hoc1       151  ---YTPTTSGVKRIKCVAQVTATD---YDALSVTS---NEVSLTVNKKTMNPQV.......  195
Hoc2        57  ---KTFDSAGSFTVKC----NVTYALADDGAE-PVVLAAESVVTVEEVPA...........   98
Hoc3a      246  ---FTPTEVGSKTLKC----TVSV---SATNY-VT---KEISAEATVVTNNAT........  284
Hoc3b      151  ---YTPTSEGTKSITC----SVTV-------T-AT---DYVDKTVESSAVSLTVNKKANSS  193
Hoc4       151  ---YTPTTSGVKKIKCVAQVTAEN--YNEKEV-TS---NEVSLTVNKKTMNPQV.......  195
```

ENDO	BAA12070	endoglucanase (gi|1663519) [*Clostridium thermocellum*] 1601aa
1WGO	56966842	(1WGO) PKD domain Sorcs2 [human] 123aa 19%; 2.8
PKD	24987493	PKD domain in surface layer protein [*Methanosarcina mazei*] 391aa 26%; 2e-05
Hoc1	19632750	Hoc head outer capsid protein [Enterobacteria phage T4] 376aa 26%, 0.086
Hoc2	109290149	large outer capsid protein [*Aeromonas salmonicida bacteriophage*] 177aa 21%; 0.19
Hoc3	32453675	Hoc head outer capsid protein [*Enterobacteria phage RB69*] 471aa 15%, 5.5; 18%; 6.5
Hoc4	116326400	large head outer capsid protein [Bacteriophage RB32] 376aa 24%; 7.0

Cellulose-binding domain of beta-glucanase may be a homolog of domains in endonucleases and bacterial endotoxins.

```
1ULO     2 SPIGEGTFD-------DGP-EGWV----------AYGTD---------GPLDT-----STG-ALCVAVPAGSAQYGVG   46
ENDO    22 .QLLNGDFETWSG---NSP-QGWS----------TIDSG---------IAVSSATTPLKTG-KLAAAIAVNTGTQGNT   72
1GUI     2 ..INNGTFDEIVNDQANNP-DEWF----------IWQAGDYGISGARVSDYGV-----RDGYAYITIADPGTDTWHIQ   62
1DLC   481 .......................................................................TENGSAA  487
DTOX  1027 NAVQNGDFN-------SGL-DSW------------NATT---------DATVQ-----QDG-NMYFLVLS---HWDAQ 1066
FnIII 1161 ..IADPGFD-------SQTFDKWNKESTAENTDHITIEN---------ESVQK-----RLG-NDVLKISGNEGA-DAK 1213

1ULO    47 VVLNGVAIEEGTTYTLRYTATASTDVTVRALVGQNGAPYGTVLDT--SPALTSEPRQVTETFTASA   91
ENDO    73 DFLQMVNVEQGKTYQFSVSVYHTEGKVMARLIADG---YQGYSNN----GLTNQWQELTFSYTATS  133
1GUI    63 FNQ-WIGLYRGKTYTISFKAKADTPRPINVKILQNHDPWTNYFAQ--TVNLTADWQTFTFTYT---  122
1DLC   488 TIYVTPDVSYSQKYRARIHYASTS--QITFTLSLDGAPFNQYYFD--KTINKGD------TLTYNS  543
DTOX  1067 VSQQ-FRVQPNCKYVLRVTAKKVGNGDGYVTIQDGAHHRETLTFN--ACDYDVNGTHVNDNSYITK 1129
FnIII 1214 ISQSISGLEEGVTYSVSAWVKNDNNREVTLGVNVGGKDFTNVITSSGGKVRQGEGVKYIDDTFVRM 1278

1ULO   110 TYPATPAADDPEGQIAFQLGGFSADAWTL-CLDDVALDSEVE  148
ENDO   134 T.........................................  134
1GUI   123 ----HPDDADEVVQISFELG--EGTATTI-YFDDVTVSPQ..  155
1DLC   544 FNLASFSTPFELSGNNLQIGVTGLSAGDKVYIDKIEFIPV..  583
DTOX  1131 ELEFYPKTEHMWVEV-------SETEGTF-YIDSIELIETQE 1163
FnIII 1279 EVEFTVPKGVNSADVYLKASEGDADSV-V-LVDDFRI..... 1313
```

1ULO	2098446	N-Terminal Cellulose-Binding Domain From *Cellulomonas Fimi* Beta-1,4-Glucanase C
ENDO	27367770	Endonuclease I [*Vibrio vulnificus*] 543aa 20%, 0.004
1GUI	24158618	Carbohydrate Binding Module 4 155aa 22%, 2e-29
1DLC	640362	Delta-endotoxin CryIIIa 584aa 10%, 2.2
DTOX	54112021	Cry9Bb delta-endotoxin [*Bacillus thuringensis*] 1163aa 14%, 6e-18
FnIII	110799287	Fibronectin type III domain protein [*Clostridium perfringens*] 1686aa 12%, 0.02

Possible homology to carbohydrate binding domain and fibronectin-III suggests that these domains may be homologs of immunoglobulins

Botox, tetanus toxin, and pentraxins may be homologs of immunoglobulins

```
TetX   904   SSVITYPDAQLVPGINGKAIHLVNNESSEVIVHKAMDIEYNDMFNNFTVSFWLRVPKVSASHLEQYGTNE 973
BOT    883   ANVEVYDGVELN---DKNQFKLTSSTNSEIRVTQNQNIIFNSMFLDFSVSFWIRIPKYKNDGIQNYIHNE 949
Lacc    27   ADLTLTNAAVSPDGFSREAVVVNGQTPGPLIAGQKGDRFQLNVIDNLTNHTMLKTTSIHWHGFFQHGTNW  96
PreL   205   ............................................ESFSACIWVKATDVLNK--------- 221

TetX   974   YSIISSMKKHSLSIGSGWSVSLKGN-NLIWTLKDSAGEVRQITFRDLPDKFNAYLANKWVF 1033
BOT    950   YTIINCIKNN-----SGWKISIRGN-RIIWTLTDINGKTKSVFFEYSIRKDVSEYINRWFF 1004
1GNH    40   YTELSSTRGYSI---FSYATKRQDN-EILIFWSKDIGYSFTVGGS-EILFEVPEVTVAPVH   95
2A3W    69   ....................................GEYSLYIGRHKVTSKVIEKFPAPVH   93
Lacc   117   YDFQVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPNDPQASLY--DIDNDDTVITLVDWYH  174
Prel   222   -TILFSYG--TKRNPYEIQLYLSYQ-SIVFVV---GGEENKLVAETMVSLG------RWTH  269

TetX  1034   ITITNDRL-SSANLYINGVLMGSAEITGLGA-IREDNNITLKLDRC-------NNNN--QY--VS 1085
BOT   1005   VTITNNS--DNAKIYINGKLESNIDIKDIGE-VIANGEIIFKLDGD-------IDRT--QF--IW 1055
1GNH    96   ICTSWESASGIVEFWVDGKPRVRKSLKK-GYVGAEASITILGQEQDSFGGNFEGSQS--LV--GD  155
2A3W    94   ICVSWESSSGIAEFWINGTPLVKKGLRQ-GYFVEAQPKIVLGQEQDSYGGKFDRSQS--FV--GE  153
Lacc   175   VAA..............................................................  177
Prel   270   LCSTWNSEKGLTSLWVNGELVATTVEMATGHTVPEGGILQIGQEKNGCCVGGGFDETLAFS--GR  332
sCRP   223   .............LWVDGKPMVRASLRR-GYTVGSGASIVLGQEQDSF--GGGFDKN--QSLVGD  269

TetX  1086   IDKFRIFCKALNPKEIEKLYTS-YLSITFLRDFWGNPLRYDTE-YYLI 1131
BOT   1056   MKYFSIFNTELSQSNIKEIYKI-QSYSEYLKDFWGNPLMYNKE-YYMF 1101
1GNH   156   IGNVNMWDFVLSPDEINTIYLG-GP.......................  179
2A3W   154   IGDLYMWDSVLPPENILSAYQG-TPLPANILD-WQA-LNYEIRGYVII  198
Prel   333   LTGFNIWDSVLSNEEIRE..............................  350
sCRP   277   IEDVNMWDFVLSPSQILTLYTTRALSPNVLN--WRN-LRYETRGEVF.  311
```

TetX	28373188	Tetanus toxin [*Clostridium tetani*] 1315aa
BOT	123229147	Neurotoxin B [*Clostridium botulinum*] 31%, e-119
1GNH	1942435	C-reactive protein [Human] 11%, 2e-05
2A3W	576259	Serum amyloid P component [Human] 14%, 1e-04
Lac	16041067	Laccase [*Pycnoporus coccineus*] 11%, 0.13
Prel	109048575	Pentraxin-related [*Macaca mulata*] 12%, 1e-04
sCRP	74006329	Similar to CRP [*Canis familiaris*] 25%, 0.16

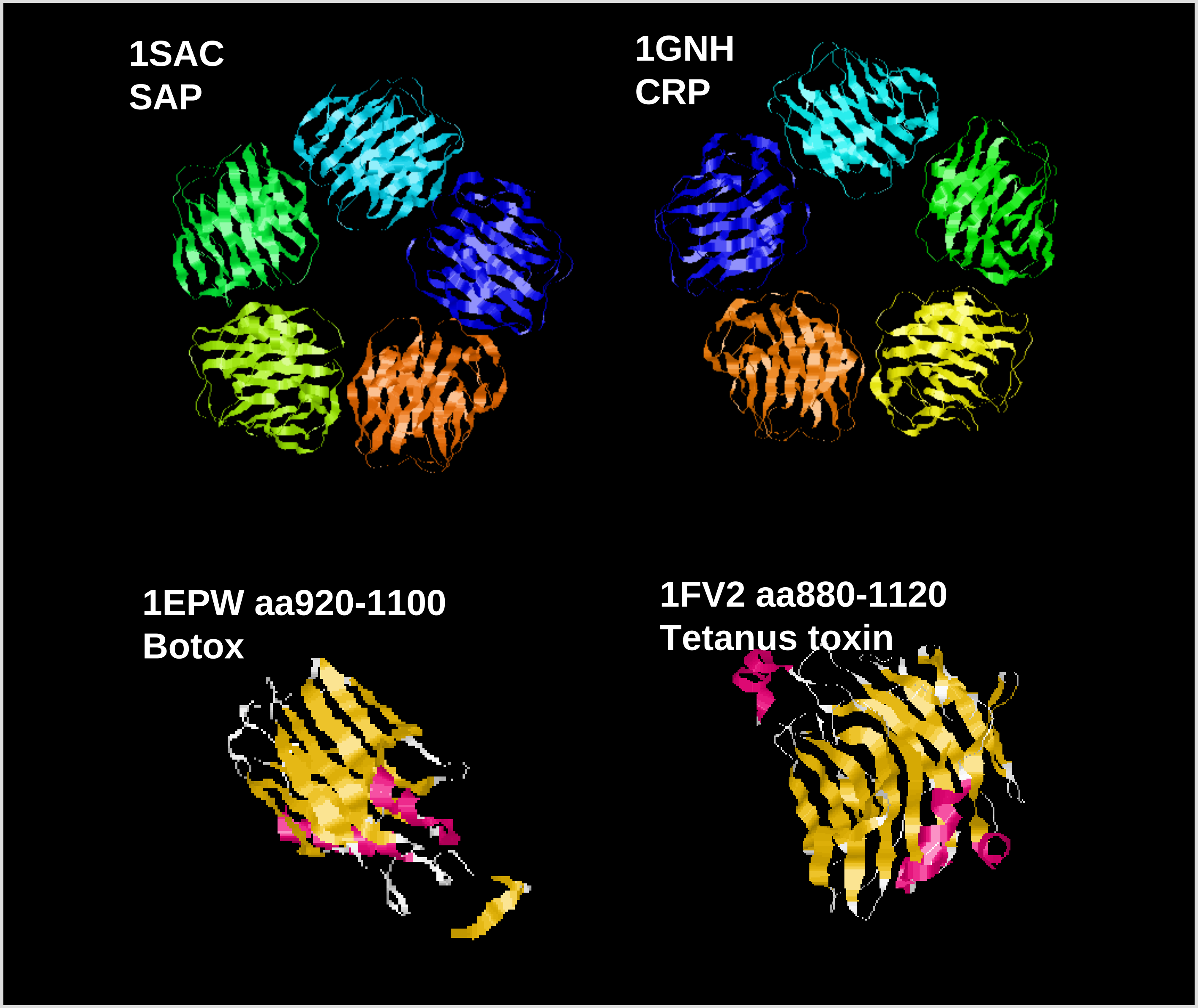
Identity of Laccasse is Confirmed

gi|21730578|pdb|1KYA|A _ Chain A, Active Laccase From
Trametes Versicolor Complexed With 2,5- Xylidine
Length=499 Score = 246 bits (629), Expect = 3e-64
Identities = 145/159 (91%)

```
Query 22 AIGPMADLTLTNAAVSPDGFSREAVVVNGQTPGPLIAGQKGDRFQLNVIDNLTNHTMLKT 81
          IGP+ADLT+TNAAVSPDGFSR+AVVVNG TPGPLI G  GDRFQLNVIDNLTNHTMLK+
Sbjct 1  GIGPVADLTITNAAVSPDGFSRQAVVVNGGTPGPLITGNMGDRFQLNVIDNLTNHTMLKS 60

Query 82 TSIHWHGFFQHGTNWADGPAFINQCPIASGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGL 141
         TSIHWHGFFQ GTNWADGPAFINQCPI+SGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGL
Sbjct 61 TSIHWHGFFQKGTNWADGPAFINQCPISSGHSFLYDFQVPDQAGTFWYHSHLSTQYCDGL 120

Query 142 RGPFVVYDPNDPQASLYDIDNDDTVITLVDWYHVAAKLG 180
          RGPFVVYDPNDP A LYD+DNDDTVITLVDWYHVAAKLG
Sbjct 121 RGPFVVYDPNDPAADLYDVDNDDTVITLVDWYHVAAKLG 159
```

CRP, other pentraxins, Botox, Tetanus, and other toxins are related to immunoglobulins

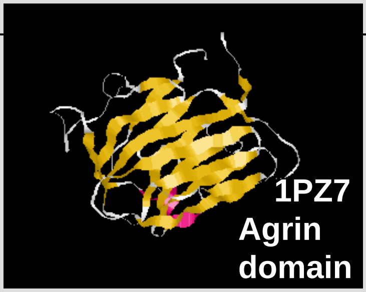
Extra beta strands found in immunoglobulin-related SAP, CRP, and toxins.



Pentraxin/CRP Domains are Found in Proteglycans/Glucosaminoglycans

```
MCSP   26 LASAASFFGENHLEV----------------PVATALTDIDLQLQFSTSQPEALLLLAA----GPADHLLL-QLYSGRLQVRLVLGQEELRLQTPA 100
1PZ7   11 DAEAIAFDGRTYMEYHNAVTKSPDALDYPAEPSEKALQSNHFELSIKTEATQGLILWSGKG-LERSDYIAL-AIVDGFVQMMYDLGSKPVVLR--S 102
CRP    24 ............VMV----------------GTLPDLQEITLCYWFKVNCLKGTLHMFSYA-TAKKDNELL-TFLDEQGDFLFNVHGAP-QLKVQC  88
NP2s  300 ...............................KTLPELYAFTVCLWLRSSASPGIGTPFSYAVPGQANEIVLIEWGNNPIELLINDK------VAQL 358

MCSP  101 ETLLSDSIPHTV--VLTVVEGWATLSVDGFL---NASS-AVP-GAPLEVPYGLFVGGT--GTLGLPYLRGTSRPLRGCLHAATLNGRSLLRPLTPDVHE 190
1PZ7  103 TVPINTNHWTHI--KAYRVQREGSLQVGNEA---PITGSSPLGATQLDTDGALWLGGMERLSVAHKLPKAYSTGFIGCIRDVIVDRQELHL........ 188
CRP    89 PNKIHIGKWHHVCHTWSSWEGEATIAVDGFHCKGNATGIAV--GRTL-SQGGLVVLGQ--DQDSVGGKFDATQSLEGELSELNLWNTVLNHEQ...... 176
NP2s  359 PLFISDGKWHHICITWTTRDGMWEAFQDGEK---LGTGENLAPWHPI-KPGGVLILGQ--EQDTVGGRFDATQAFVGEMSQFNIWDR............ 439
NP2   192 ....................................TV-ALG-LTTAPNPTQLAQRGP--GSLQLWRDRQVAKSSPQHRSSPHDVTVHVQEMQKFQTPS 250

MCSP  191 GCAEEFSASDDVALGFSGPHSLAAFPAWGTQDEGTLE/FTLTTQ/SRQAP---LAFQAGG-RRGDFIYV/DIFEGHLRAVVEKGQGTVLLHNSVPVADGQP 283
1PZ7   42 .........................PSEKALQSNHFE/LSIKTE/ATQGL---ILWSGKGLERSDYIAL/AIVDGFVQMMYDLGSKPVVLRSTVPINTNHW 111
CRP    16 ......................VMVGTLPDLHEITLC/YWFKLH/RLNGAPH-IFSYATS-ETDNEILT/SLNENGDFLFNIHGNTQLNVQCNNKILAGRW  89
NP2s  308 .................................FTVC/LWLRSS/ASPGIGTPFSYAVPG-QANEIVLI/EWGNNPIELLIN----DKVAQLPLFISDGKW 367
NP2   251 SHQAAPPRTYQGPGNICNTDPVLIFPNTSTENVIFLS∠LSFCSW∠SHLGT---LLSYATK-DNDNKLVL∠HLVPGSIHFVIGDPDFRELSL--KPLLDGQW 360

MCSP  284 HEVSVHINAHRL--EISVDQY-PTHTSNRGVLSYLEPR-GSLLLGGLDAEASRHLQEHRLGLTPEATNASLLGCMEDL-SVNGQRRGLREALLTRNMA 377
1PZ7  112 THIKAYRVQREG--SLQVGNEAPITGSSPLGATQLDTD-GALWLGGMERLSVAKL--------PKAYSTGFIGCIRDV-IVDRQELHLVEDALNNPTI 198
CRP    90 HHVCLTWSSWEGEATIAVDGF-HCKGNATGKATGVTFRQGGLVVIGQDQDSVGGGFDEKQSLVGELSELN----LWDMVLNHEQI............. 169
NP2s  368 HHICITWTTRDG--MWEAFQDGEKLGENLAPWHPIKPG-GVLILGQEQDTVGGRFDA----------TQAFVGEMSQFNIWDRVL............. 441
NP2   361 HHICIIWTSVEGKYWLHIDRRLVATGSRFREGYEIPPG-GSLVLGQEQDTVGGEF----------DSSEAFVGSISGLAIWDRAL............. 434

∠ : inserts in NP2::    PGFLMPLRA;  VRMAT;  GRNS
```

MCSP	1617314	Melanoma-associated chondroitin sulfate proteoglycan [Human] 2322aa
1PZ7	48425233	Agrin [*Gallus gallus*] 204aa (fragment) 21% 3e-25; 17%, 6e-16
CRP	6091973	C-reactive protein [*Tachypleus tridentatus*] 202aa 19%, 2.7; 14%, 0.23
NP2s	18097729	Similar to neuronal pentraxin-II [*Gallus gallus*] 489aa 10%, 0.12; 11%; 0.75
NP2	38082085	Neuronal pentraxin-II [*Mus musculus*] 482aa 12%, 0.003


1PZ7
Agrin domain

LamininG in Perlecan may be a homolog of pentraxins; Laminin B apparently is not.

```
PERL 3531 LPGNSFSRSLPE-VPETIEFEVRTSTADGLLLWQGVVREASRSKDFISLGLQDGHLVFSY 3589
LAMG  543 .................LNLRFKTHSPNGLILWTGR-HSALEGDDFLSLGVENGFLHLRY  584
NP1   236 YMYAKVKKSLPEMYAFTVCMWLKSSATPGV------------GTPFSYAVPGQANELVLI  283
CRP    60 ................TLCYWFKVNRLKGTL---HMFSYATAKKDNELLTLIDEQGDFLF  100
NP2   332 .................................................SLVPGSIHFVI  342

PERL 3590 Q-LGSGEARL-VSGDPI-------NDGEWHRITALREGQRG--SIQVDGEDLVTGRSPGPNV 3640
LAMG  585 N-LGSGEVNIKYNSTKV-------SDGLWHRVRALRNSQDG--TLKVDGGKSITRRSPGKLR  636
NP1   284 E-WGNNPMEI-LINDKVAKLPFVINDGKWHHICVTWTTRDGVWEAYQDGTQGGSGENLAPYH  343
CRP   101 NVHGAPQLKV-QCPNKI-------HIGKWHHVCHTWSSWEGEATIAVDGFHCKGNATGIAVG  154
NP2   343 ---GDPDFRE-LSLKPL-------LDGQWHHICIIWTSVEGKYWLHIDRRLVATGSRFREGY  393
PD   1493 ....NGKEKI-TNCPSV-------NDGIWHHIAITWTSTGGAWRVYIDGELSDSGTGLSVGK 1541

PERL 3641 AVNTKDIIYIGGAPDVATLTRGKFSSG--ITGCIKNLVLHTARP.................. 3682
LAMG  637 QLNTDTGLYVGGLPAASFYTRQRYSSG--IVGCISELIL.......................  673
NP1   344 PIKPQGVLVLGQEQDTLG-GGFDATQA--FVGELAHFNIWDRKL..................  384
CRP   155 RTLSQGGLVVLGQDQDSVGGKFDATQS--LEGELSELNL.......................  191
NP2   394 EIPPGGSLVLG---QEQDTVGGEFDSSEAFVGSISGLAI.......................  429
PD   1542 AIPGGGALVLGQEQDKKG-EGFNPAES--FVGSISQLNL....................... 1578
```

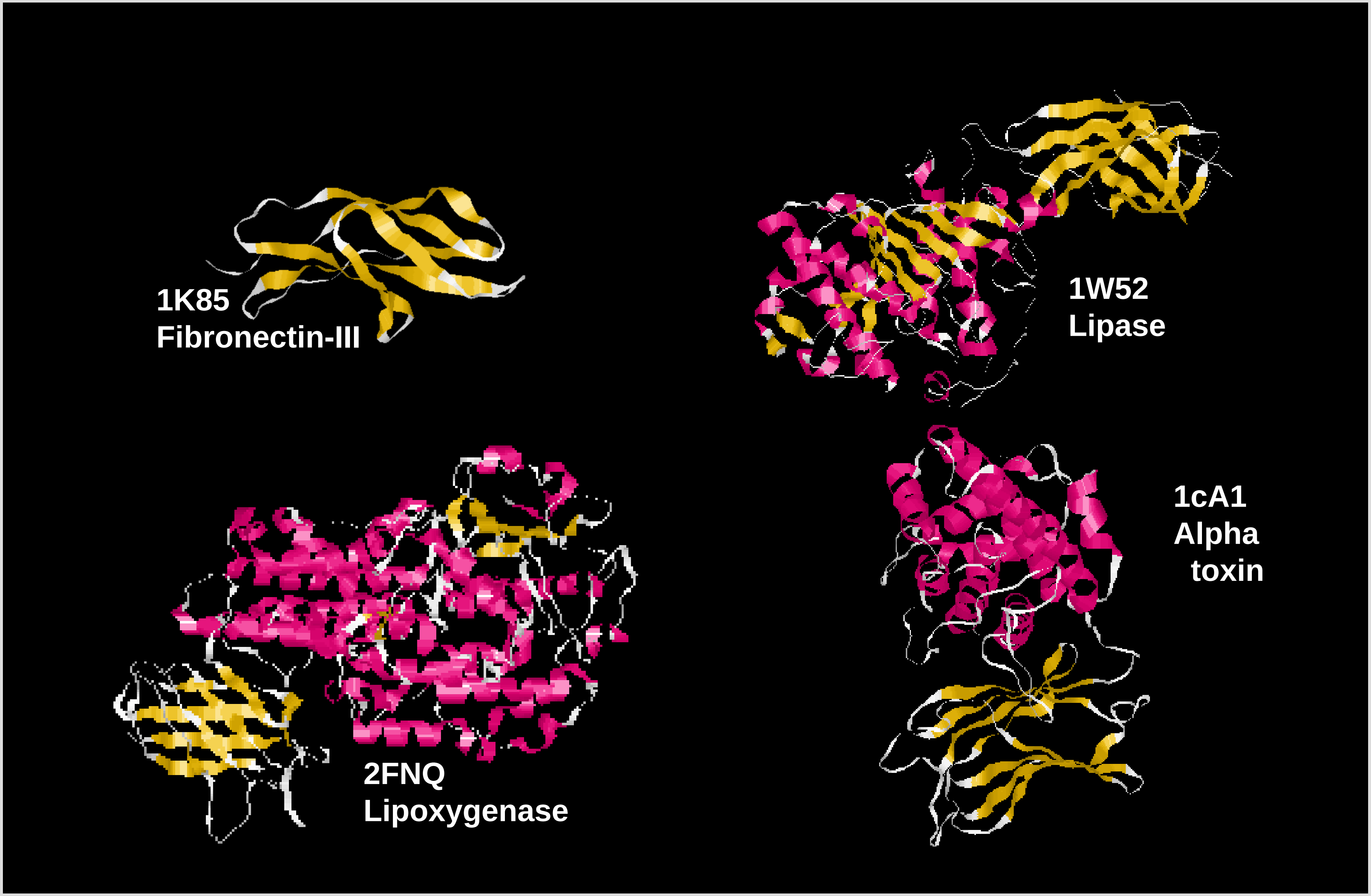
PERL	200296	Perlecan/Heparan Sulfate proteoglycan [Mouse] 3707aa
LAMG	108869902	Laminin Gamma-3 [*Aedes aegypti*] 697aa 38%, 2e-22
NP1	1438954	Neuronal pentraxin 1 [Human] 430aa 14%, 0.013
CRP	117481	C-reactive protein [Limulus polyphemus] 242aa 17%, 5.3
NP2	38082085	Neuronal pentraxin 2-like [Mouse] 482aa 22%, 0.32
PD	109476547	Similar to sushi, von Willebrand factor type A, EGF and pentraxin domain containing 1 [Rat] 3578aa 21%, 1.4

Lipases, Lipoxygenases, and Alpha-Toxins may be related to immunoglobulins

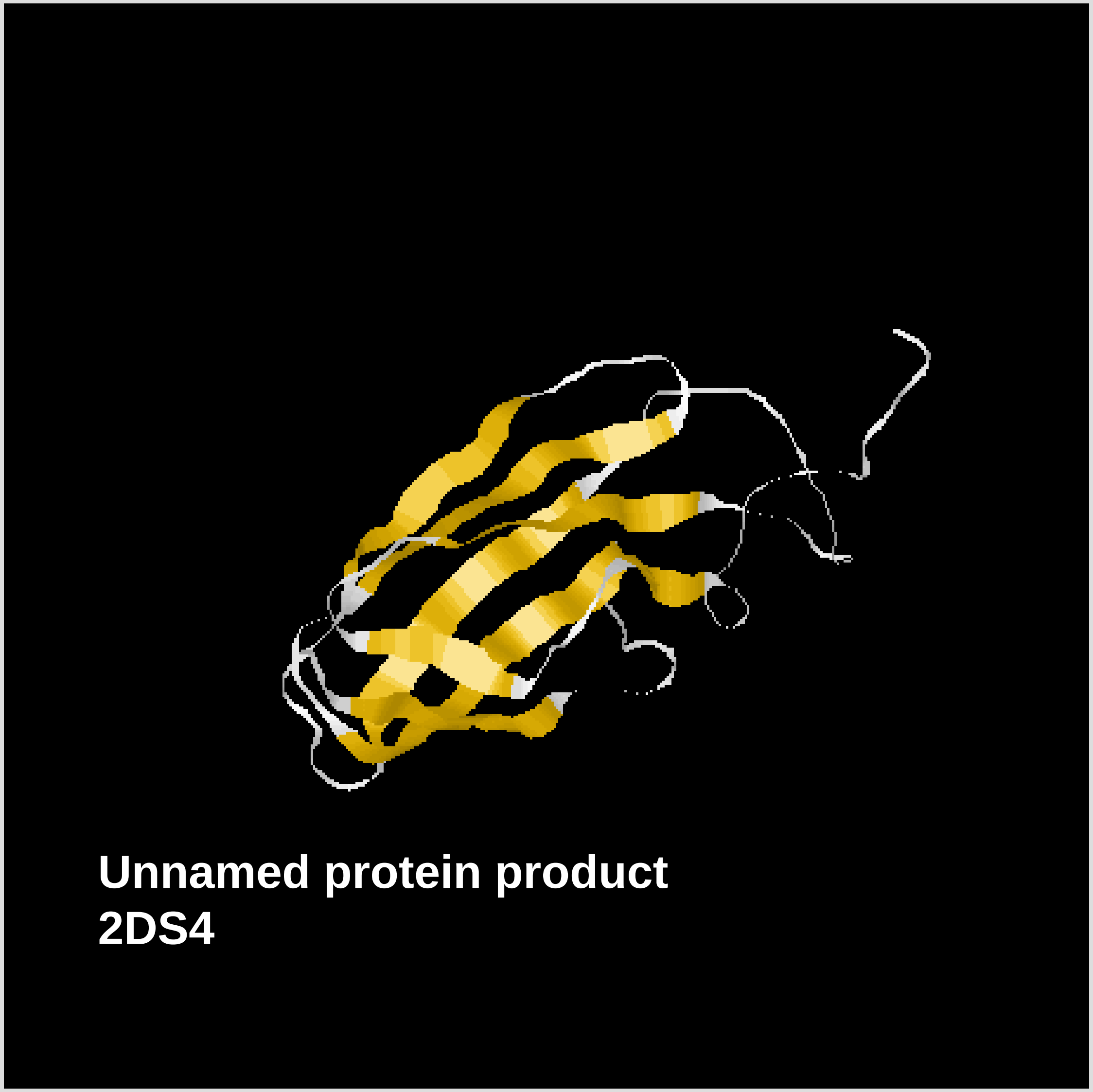
```
FnIII  1157 YGSEIADPGFDSQTFDKWNKESTAENTDHITIENES--VQKRLGNDVLKISGNEGADAKISQ 1216
1W52    324 EQTFFLNTG-ESGDYTSWRYR------VSITLAGSG--KANGYLKVTLRGSNGNSKQYEIFK  376
2FNQ      9 ...................YN------VEVETGDREHAGTDATITIRITGAKGRTDYLKLDK   45
1CA1    261 ............................YISTSGEKDAGTDDYMYFGIKTKDGKTQEWEMDN  294

FnIII  1217 -SISGLEEGVTYSVSAWVKND----NNREVTLGVNVGGKDFTNVITSGGKVRQGEGVKYIDDTFV 1276
1W52    377 ---GSLQPDSSYTLDVDVNFI----IGKIQEVKF-VWNKTVLNLSKPQLGASRITVQSGADGTEY  433
2FNQ     46 WFHNDFEAGSKEQYTVQ-GFD----VGDIQLIEL-HSDGGGYWSGDPDWFVNRVIIISSTQDRVY  104
1CA1    295 P-GNDFMTGSKDTYTFK-LKDENLKIDDIQNMWI-RKRKYTAFP--DAYKPENIKNIA-NGKVVV  353

FnIII  1277 RM-EVEFTVPKGVNSA 1291
1W52    434 KF-CGSGTVQDNVEQS  448
2FNQ    105 SFPCFRWVIKDMVL..  118
1CA1    354 DKDINEWISGNSTY..  367
```

FnIII 110799287 Fibronectin-III domain protein (search 1155-1320)
1W52 112489604 Pancreaic Lipase related protein 2 [Horse] 14%, 1.0
2FNQ 90109545 8r-Lipoxygenase [*Plexaura homalia* Coral] 699aa 17%, 1e-08 Q:1W52
1CA1 4929954 Alpha-toxin [*Clostridium perifringens*] 370aa 21%,4e-10 Q:2FNQ

Lipases, lipoxygenases, and alphatoxins have immunoglobulin-related adaptor domains.



A structurally characterized "unnamed protein" is a filamin -- actin binding protein --


Unnamed protein product
2DS4

A structurally characterized "unnamed protein" is a filamin

```
2DS4   8   EVDPAKCVLQGEDLHRA-REKQT--ASFTLLCKDAAGEIMGRGGDNVQVAVVPKDK  60
2D7O   8   AINSRHVSAYGPGLSHG-MVNKP--ATFTIVTKDA-----GEGGLSLAV------E  49
2DJ4   8   VVDPSKVKIAGPGLGSGVRARVL--QSFTVDSSKA-------GLAPLEVRVLGPRG  54
1WLH   4   ..DPEKSYAEGPGLDGG-ECFQP--SKFKIHAVDPDGVHRTDGGDGFVVTI----E  50
1WLH 216   ..........................TFTVAAKNKKGEVKTYGGDKFEVSITGPAE 245
1WLH 111   .........EGEGLVKV-FDNAP--AEFTIFAVDTKGVARTDGGDPFEVAINGPD- 153
2AAV  14   ..........GPGLTHG-VVNKP--ATFTVNTKDA-----GEGGLSLAI------E  45
1KSR   2   ..DPEKSYAEGPGLDGG-ECFQP--SKFKIHAVDPDGVHRTDGGDGFVVTI----E  48
2DIA   8   PFDPSKVVASGPGLEHG-KVGEA--GLLSVDCSEAGPGALG---------LEAVSD  51
2D7M   8   AHDASKVRASGPGLNAS-GIPASLPVEFTIDARDA-----GEGLLTVQIL----DP  53
2DI8  10   ..DARRAKVYGRGLSEG-RTFEM--SDFIVDTRDA-----GYGGISLAV------E  49
1QFH 117   ..........................TFTVAAKNKKGEVKTYGGDKFEVSITGPAE 146
1QFH  12   .........EGEGLVKV-FDNAP--AEFTIFAVDTKGVARTDGGDPFEVAINGPD-  54

2DS4  61   KDSPVRTMVQDNKDGTYYISYTPKEPGVYTVWVCIKEQHVQGSPFTVTVR     110
2D7O  50   GPSKAEITCKDNKDGTCTVSYLPTAPGDYSIIVRFDDKHIPGSPFTAKI.      98
2DJ4  55   LVEPVN--VVDNGDGTHTVTYTPSQEGPYMVSVKYADEEIPRSPFKVKV.     101
1WLH  51   GPAPVDPVMVDNGDGTYDVEFEPKEAGDYVINLTLDGDNVNGFPKTVTVK     100
1WHL 246   E---ITLDAIDNQDGTYTAAYSLVGNGRFSTGVKLNGKHIEGSPF.....     287
1WHL 154   -GLVVDAKVTDNNDGTYGVVYDAPVEGNYNVNVTLRGNPIKNMPIDV...     199
2AAV  46   GPSKAEISCTDNQDGTCSVSYLPVLPGDYSILVKYNEQHVPGSPFTARV.      93
1KSR  49   GPAPVDPVMVDNGDGTYDVEFEPKEAGDYVINLTLDGDNVNGFPKTVTVK      98
2DIA  52   SGTKAEVSIQNNKDGTYAVTYVPLTAGMYTL...................      82
2D7M  54   EGKPKKANIRDNGDGTYTVSYLPDMSGRYTITIKYGGDEIPYSPFRI...     100
2DI8  50   GPSKVDIQTEDLEDGTCKVSYFPTVPGVYIVSTKFADEHVPGSPFTVKI.      98
1QFH 147   E---ITLDAIDNQDGTYTAAYSLVGNGRFSTGVKLNGKHIEGSPF.....     188
1QFH  55   -GLVVDAKVTDNNDGTYGVVYDAPVEGNYNVNVTLRGNPIKNMPIDV...     100
```

```
2DS4 122919976 Unnamed protein product [human]
2D7O 109157447 17th filamin domain filamin C [human] 35%; 3e-08
2DJ4 116667073 13th filamin domain filamin B [human]  37%; 7e-07
1WHL 55670676  Rod domain of Dictoystellium filamin 32%, 7e-05; 30%, 0.006; 31%, 0.03
2AAV 99031796  Filamin A domain 17  36%; 1e-04
1KSR 2392410   F-actin cross-linking gelation factor (ABP-120)  32%; 1e-04
2DIA 116667052 10th filamin domain filamin B [human] 30%; 3e-04
2D7M 109157445 14th filamin domain filamin C [human] 33%; 5e-04
2DI8 116667050 19th filamin domain Filamin B [human] 335; 7e-04
1QFH 4930107   Gelation factor from Dictyostelium discoideum domains 5 & 6  30%, 0.004; 31%, 0.026
```

Filamins may be homologs of Hsp70 chaperones

```
2DS4        8 EVDPAKCVLQGEDLHRAREKQTASFT----LLCKDA---AGEIMGRGGDNV--------QVAVVPK-DK  60
2D70        8 AINSRHVSAYGPGLSHGMVNKPATFT----IVTKDA--------GEGGLSL--------AV-------E  49
HSP1      181 ..........GKKVTHAVVTVPAYFN----DAQRQATKDAG---TIAGLNVLRIVNEPTAAAIAYGLDK 232
DnaK      163 ..........GTTVKNAVVTVPAYFNDAQRQSTKDAGAIAG-------LDVQRILNEPTAAAIAYGLDK 214
BiP       167 ..........GKEVKHAVVTVPAYFN----VAQRQALKYAG---TIVGLNVVRI-NEPTAAAIAYGLDK 218
HSP2      161 ..........GKEVKHAVITVPAYFNDAQRQATKDAGKTAG-------LNVLRILNEPTAAVMAYGLNK 212
GRP       183 ..........GKKVTHAVVTVPAYFNDAQRQATKDAGAIAG-------LNILRIVNEPTAAAIAYGLDK 234
HSP3      167 ..........GKPVTHAVVTVPAYFNDAQRQATKDAGTIAG-------LNVIRIVNEPTAAAIAYGLDK 218
HSP4      129 ..........GEKVTEAVITVPAYFNDAQRQATKDAGRTAG-------LEVKRIINEPTAAALAYGLDK 180

2DS4       61 KDSPVRTMVQDNKDGTYYISYTPKEPGVYTVWVCIKEQHVQGSPFTVTV--------RRKH 113
2D70       50 GPSKAEITCKDNKDGTCTVSYLPTAPGDYSIIVRFDDKHIPGSPFTAKI............  98
HSP1      233 GDQEKQIIVYDLGGGTFDVSLLSIEGGVFEVLATAGDTHLGGEDFDFKIVRYLAKQFKKKH 293
DnaK      215 KQGEKNILVFDLGGGTFDVSILTIDEGVFEVIATSGDTHLGGADFDQRVMDYFIKLTKKKH 275
BiP       219 KDGERNILVFDLGGGTFDVSMLTIDNGVFEVLATNGDTHLGGEDFDQRV............ 267
HSP2      213 VGGEKNILVFDLGGGTFDVSLLNIEDNVFDVISTSGNTHLGGSDFDQKV............ 261
GRP       235 TEKEHQIIVYDLGGGTFDVSLLSIENGVFEVQATAGDTHLGGEDFDYKL--------VR.. 285
HSP3      219 TDTEKHIVVYDLGGGTFDVSLLSIDNGVFEVLATSGDTHLGGEDFDNRV............ 267
HSP4      181 KNASEKVAVFDLGGGTFDISILELGEGVFEVKSTDGDTHLGGDDFDQKI............ 229
```

2DS4	122919976	Unnamed protein product [human]
2D70	109157447	17th filamin domain filamin C [human] 35%; 3e-08
HSP1	68472794	Putative HSP70/BiP chaperone [*Candida albicans*] 23%, 0.060D
DnaK	118350929	DnaK protein BiP [*Tetrahymena thermophila*] 22%, 1.6
BiP	156348	BiP (heat shock protein 3) 25%, 0.39
HSP2	3746803	Hsp70-BiP precursor [*Entamoeba histolytica*]
GRP	121568	GRP78/BiP [*Kluyveromyces lactis*] 22%, 4.2
HSP3	19114157	BiP [*Schizosaccharomyces pombe*] 22%, 0.12
HSP4	78189848	Hsp70 [*Chlorobium chlorochromatii*] 22%, 5.8

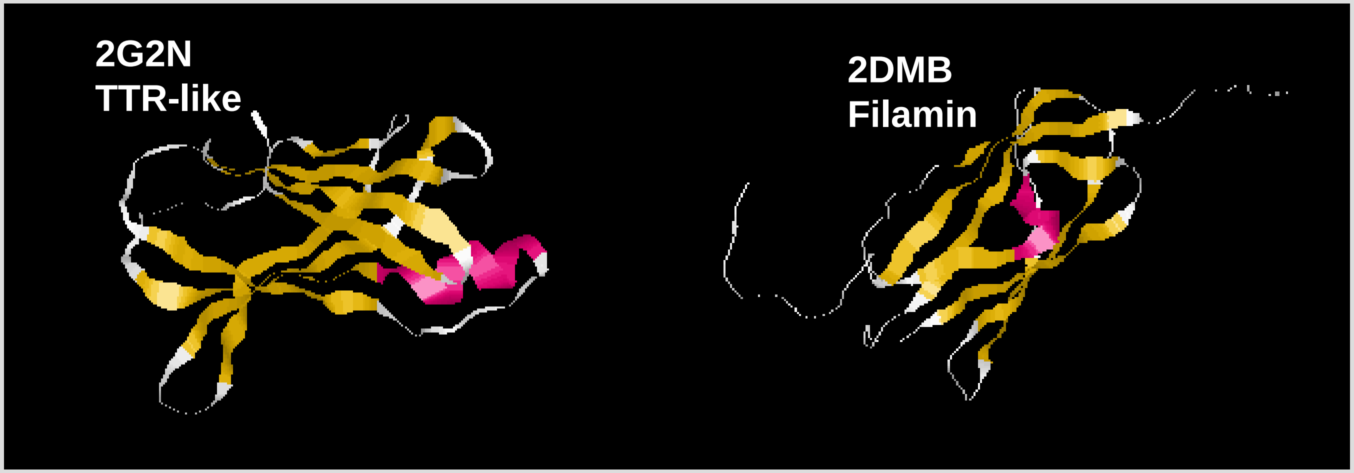
Transthyretins may be homologs of filamins, which may be homologs of immunoglobulins.

```
TTR-like   25 ........QQNILSVHILNQQTGKPAADVTVTL-EKKADNGWLQLNTAK---TDKDG--RIKALWPEQTATT   82
TonB1      40 ...........TLSVQVLDAAGNAPLEGVQVAI--------RECRCGGI---TDRDG--RYSRVLPQ-----   82
TonB2      90 LYGSAAFAQSSTIIGTVIDAQSRQPAADVVVTA-TSPNLQGEQ---TVV---TDAQGNYRIP------QLPP  148
PKD        21 LLSTHAQA-SSTISGTVY--GGPAVLEGASVGL-LDENQEAIE---SITADITDSQGLYHFP------MLAD   80
FIL      1634 GLGIAPTVRTGEEVGFVVDAKSAG-KGKVTCTVLTPDGTEAEA---DVV---ENEDGTYDIFYT----AAKP 1694
2DMB       19 GPGIASTVKTGEEVGFVVDAKTAG-KGKVTCTVLTPDGTEAEA---DVI---ENEDGTYDIFYT----AAKP   78

TTR-like   83 GDYRVVFK---TGDYFKKQNLESFFPEIPVE...........  110
TonB1      83 GSYSVEFF------YLGFQGVQR---QVDLR...........  104
TonB2     149 GDYTLRFE---KEQFKPYARSAIQLRLNRTIRVNVELLPEAL  187
PKD        81 GTYYLTVTPPQSSGFPSSSAEQIVI-AGNDVQHNVVLLGSAV  120
FIL      1695 GTYVIYVR---FGGVDIPN....................... 1710
2DMB       79 GTYVIYVR---FGGVDIPN.......................   94
```

RED indicates identity to TTR-like

BLUE indicates identity to TonB1 or TonB2

TTR-l	15802403	Transthyretin-like [*E. coli*] (Homolog of vertebrate transthyretin (prealubumin))
TonB1	88803938	Putative TonB-dependent receptor [*Robiginitalea biformata*] 22%, 1.9
TonB2	108763648	TonB-dependent receptor [*Myxococcus xanthus*] 29%, 0.82
PKD	119944034	PKD domain containing protein [*Psychromonas ingrahami*] Query: TonB2 27%, 1E-06
FIL	71896431	Filamin B [*Gallus gallus*] 2567aa Query: TonB2 26%, 0.39
2DMB	118137567	15th Filamin B domain [Human] Query: FIL 93%, 1e-29



Lactadherin may be evolutionarily related to pentraxins and immunoglobulins

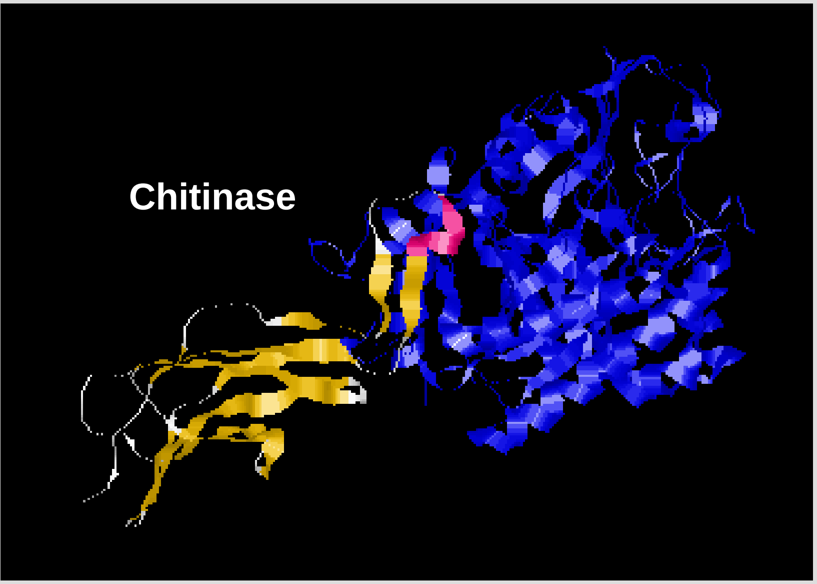
```
LACT     45  vrgdvfpsytctclkgyagnhcetkcveplgmengniansq 85
PENT      5  LLFLKSQVFGLTVETLNGERNGDFEQQKENHGAKNVAPQGI 45
1W8N    429  TTPGRYRVGATLRTSAGNASTTFTVTVGLLDQARMSIAD-- 477

LACT     86  IAASSVRVTFLGLQHWVPELARLNRAGMVNAWTPSSNDDNPWIQVNLLRRMWVTGVVTQG  145
2J1A     14  ITASSEETSGENAPASFASDGDMNTF-WHSKWSSPAHEGPHHLTLELDNVYEINKVKYAP  72
PENT     46  PYQSSYY--GQKEQAKRVIDGSLASNYMEGDCCHTEKQMHPWWQLDMKSKMRVHSVAITN  103
1W8N    478  VD--SEETAREDGRASNVIDGNPSTFWHTEWSRADAPGYPHRISLDLGGTHTISGLQYTR  525

LACT    146  ASRLASHEYLKAFKV--AYSLN--GHEFDFIHDVNKKHKEFVGNWNKNAVHVNLFETPVEAQYV  205
2J1A     73  RQD-SKNGRITGYKV--SVSLD--GENFTEVKTGTLE---------DNAAIKFIEFDSVDAKYV  122
PENT    104  RGD-CCRERINGAEIRIGNSKKEGGLNSTRCGVVFKM---------NYEETLSFNCK-ELEGRYV  157
1W8N    526  RQN-SANEQ--VADYEIYTSLNGTTWDGP-VASGRFT---------TSLAPQRAVFP-ARDARYI  576

LACT    206  RLyptschtact---lrfellgcelngcan  232
2J1A    123  RLDVTDSVSDQANGRGKFATA-AEVNVHG.  150
PENT    158  TVTIPDR--------IEYLTL-CEVQVFAD  178
1W8N    577  RLVALSEQTG-----HKYAAV-AELEVEGQ  600
```

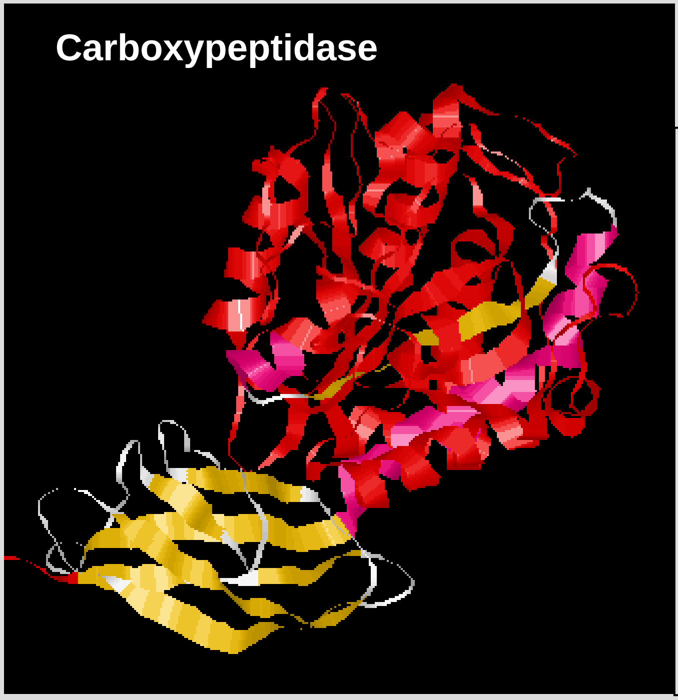
LACT   5174557 Lactadherin [Human] 387aa Q: (50-230)
2J1A 114794892 Carbohydrate binding module 32 [*Clostridum perifringens*] 150aa 18%, 2e-07
PENT   738985 Pentraxin [*Xenopus laevis*] 311aa 16%, 4e-08 Q: 2J1A
1W8N   55670665 Neuraminidase [*Micromonospora viridifaciens*] 601aa 11%, 7e-13 Q: PENT(1-200)

Some immunoglobulin domains serve as accessory homing modules.

```
PKD   27 QASSTISGTVYGGPAVLEGASVGLLDENQEAI
1EIB 22 QAATAYNNLVKVKNAADVSVSWNLWNGDTGTT

PKD   59 --------ESITADITDSQGLYHFPMLADGTY
1EIB 43 AKVLLNGKEAWSGPSTGSSGTANFKVNKGGRY

PKD   83 YLTVTPPQSSGFPSSSAEQIVIAGNDVQH 111
1EIB 86 QMQVALCNADGCTASDATEIVVADTDGSH 114
```

PKD 119944034 PKD domain containing protein
1EIB 13399645 Chitinase   23%, 3.9





```
TonB1   41  LSVQVLDAAGNAPLEGVQVAIRECRCGG--
1QMU   305  ..GFVLDATDGRGILNATISVADINHPV--
1UWY   298  VKGQVFDQNGN-PLPNVIVEVQDRKHICPY

TonB1   69  ITDRDGRYSRVLPQGSYSVEFFYLGFQGVQ
1QMU   330  TTYKDGDYWRLLVQGTYKVTASARGYDPVT
1UWY   326  RTNKYGEYYLLLLPGSYIINVTVPGHDPHI

TonB1   99  RQVDL--RESVDLVVRMQEQA 114
1QMU   361  KTVEVDSKGGVQVNFTLS... 378
1UWY   357  TKVII--PEKSQNFSALKKDI 375
```

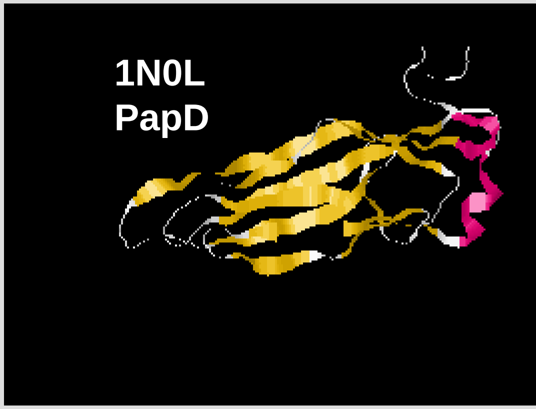
TonB1 88803938   Putative TonB-dependent receptor [*Robiginitalea biformata*]
1QMU 57012643   Carboxypeptidase [Duck] 380aa 24%, 8e-11
1UWY 48425844   Carboxypeptidase [Human] 426aa 24%, 5e-04

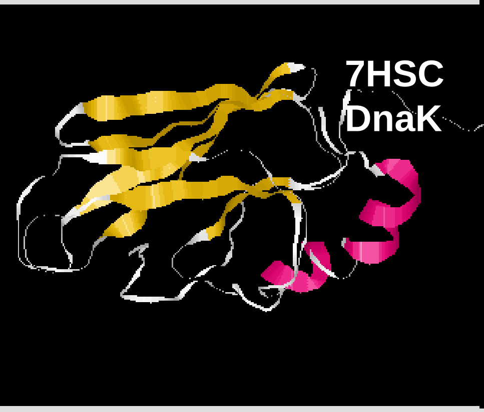
PapD may be a homolog of DnaK (Hsp70) and the "plug" domain of Ton-B dependent receptor and other porins.

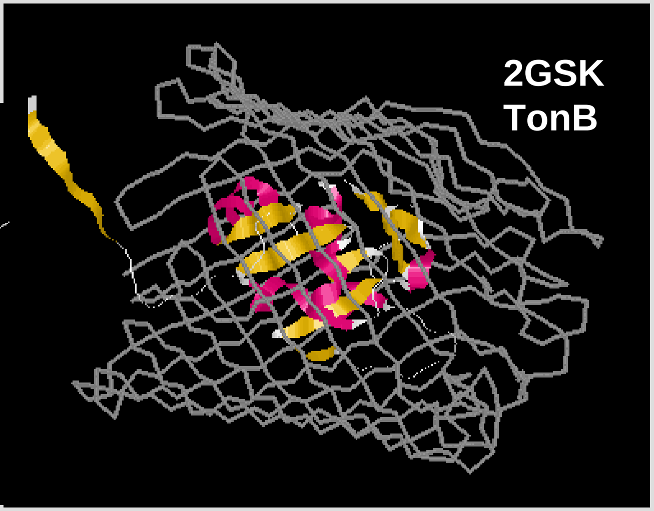
```
PapD    4  LSMIRKKILMAAIPLFVISGADAA----------VSLDRTRAVFDGSEKSMTLDISN----DNKQLPYLAQ   60
DnaK1 365  ..............VAMGAAIQAG----------VLMGEVRDVVLLDVTPLSLGVET----KGGVMTVLIP  407
DnaK2 345  .............CVAMGAAIQAG----------VLAGEVKDILLLDVTPLSLGVET----LGGIMTRLIE  387
DnaK3 366  ..............VAIGAAVQAG----------ILTGELRDLLLNDVTPLSLGLET----VGGLMKVLIP  408
TonB    1  MKMMTKKPVVVALTMAFSSTAYAQQVEQVTEFDEVLVTATRIAEKASQSSRSVAVVDEEELQEAQPASVAE   71

PapD   61  AWIENENQEKIIAGPVIATPPVQRLEPGAKSMVRLSTTPDISKLPQDRESLFYFNLREIPPRSEKANVLQIAL 133
DnaK1 408  ---------RNTTIPTRKCEIFTTAEH-NQTAVEIHVLQGERPMAQDNKSLGRFRLEGIPPMPAGVPQIEV.. 468
DnaK2 388  ---------RNTTIPTRKSQIFTTAAD-NQTSVEIHVLQGERPLAKDNISLGRFTLVGIPPAPRGIPQIEVTF 451
DnaK3 409  ---------RNTPIPVRQSDVFSTSEP-NQSSVEIHVWQGERQMAADNKSLGRFRLSGIPPAPRGVPQIQVAF 471
TonB   72  A-LQNEANVTVSNGPRASSQGVEIRGLGGQRVL........................................ 103
```

PapD	26250993	PapD [*E. coli*] 254aa Q:1-150
DnaK1	55981460	DnaK [*Thermus thermophilus*] 615aa 14%, 0.50
DnaK2	116754043	DnaK [*Methanosaeta thermophil*] 615aa 11%, 2.1
DnaK3	87124445	DnaK [*Synechococcus*] 664aa 16%, 6.4
TonB	118072687	TonB [*Shewanella woodyi*] 681aa 20%, 1.9

PapD and the plug domain may be evolutionarily related to Igs







Ig light and heavy chain domains, TTR, and lactadherin generate amyloid.

Amyloid fibrils formed physiologically, by any protein, are multi-protein complexes that include SAP and glycosaminoglycans, both of which are apparently related to immunoglobulins.

BiP, an hsp70 chaperone, is related to immunoglobulins and may increase risk of AL by "rescuing" unstable light chains that may otherwise not have been able to form functional antibodies, thus causing the B-cell line to be deleted.

BiP has been shown to mitigate TTR amyloidosis by trapping aggregates within the ER.

Immunoglobulins or immunoglobulin-related proteins are significantly involved in <u>all</u> forms of amyloidosis.

Filamins may be homologs of cytochrome c oxidase, a cupredoxin

```
2DS4        8 EVDPAKCVLQGEDLHRA---REKQTASFTLL----CKDAAGEIMG-RGGDN-VQVAVVPKD----K  60
2D7O        8 AINSRHVSAYGPGLSHG---MVNKPATFTIV----TKDA-----G-EGGLS-LAV----------E  49
2DJ4        8 VVDPSKVKIAGPGLGSG--VRARVLQSFTVDSSK-------------AGLAPLEVRVLGPR-----  53
2AAV       14 ..........GPGLTHG--VVNKP-ATFTVNTKDA---------G-EGGLS-LAI----------E  45
67462210   19 RVDPTTVRQEGPWADPAQAVVQTGPNQYTVYVLAFAFGYQPNPIEVPQGAE-IVFKITSPDVIHGF  83

2DS4       61 KDSPVRTMVQDNKDGTYY--ISYTPKEPGVYTVWVCIKEQHVQGSPFTVTV--------RRKH 113
2D7O       50 GPSKAEITCKDNKDGTCT--VSYLPTAPGDYSIIVRFDDKHIPGSPFTAKI............  98
2DJ4       54 ---GLVEPVNVVDNGDGTHTVTYTPSQEGPYMVSVKYADEEIPRSPFKVKV............ 101
2AAV       46 GPSKAEISCTDNQDGTCS--VSYLPVLPGDYSILVKYNEQHVPGSPFTARV............  93
67462210   84 HVEGTNINVEVLPGEVST--VRYTFKRPGEYRI.............................. 147
```

Red indicates identity of residue in cytochrome oxidase (67462210) to residue in a representative filamin.

2FWL: Cytochrome Oxidase

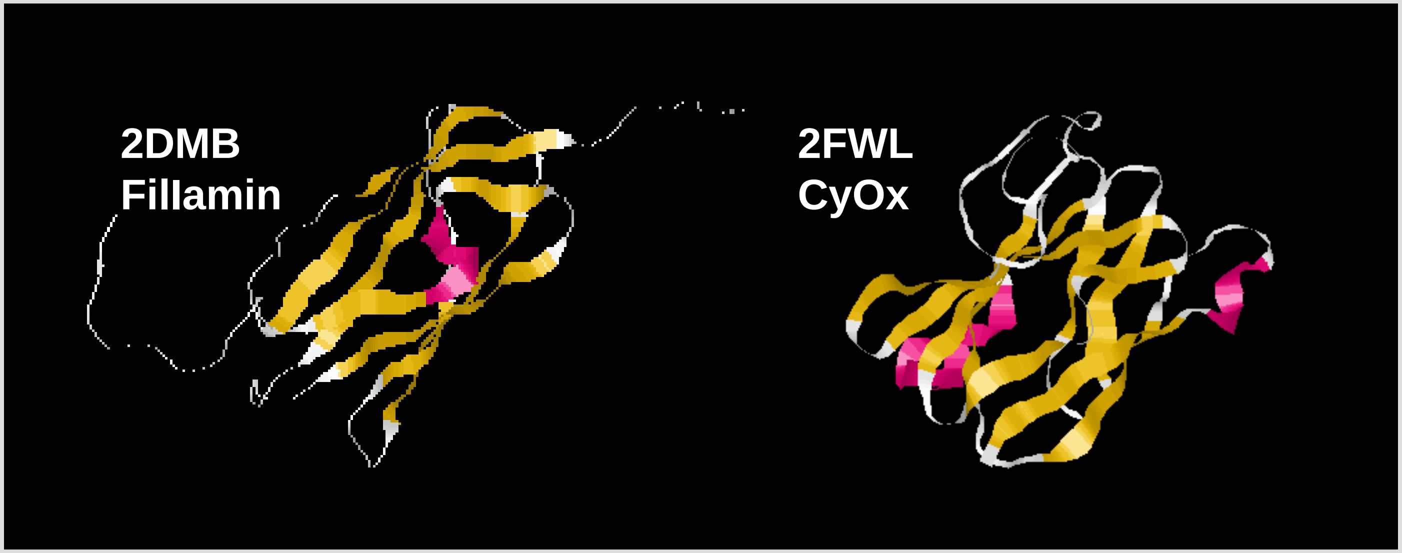
Homology to cupredoxins implies homology to immunoglobulins

and implies that Hsp70 chaperones, including BiP are homologs of immunoglobulins

Additional Evidence

```
2DMB SGSSGTGDASKCLATGPGIASTVKTGE--EVGFVVD----AKTAGKGKVTC--------T  48
2FWL AGKLERVDPTTVRQEGPWADPAQAVVQTGPNQYTVYVLAFAFGYQPNPIEVPQGAEIVFK  74

2DMB VLTPD--------GTEAEADVIENEDGTYDIFYTAAKPGTYVI.................  83
2FWL ITSPDVIHGFHVEGTNINVEVLPGEVST--VRYTFKRPGEYRI................. 115
```

```
2DMB 118137567 15th Filamin domain [Human] 124aa
2FWL  93279785 Cytochrome oxidase [Thermus thermophilus] 136aa  18%, 3.2
```



Additional Evidence

```
IG   906 .............MTTELTFTVKDAYGNP-VTGLKPDA----PVFSGAASTGS-ERPSAGNWT  35
1W8N 364 .........LEPGQQVTVPVAVTNQSGIA-VP--KPSL----QLDASPDWQVQ-GSVEPLMPG 414
SAX  277 DNRGLRIERVQPSDEGEYVCYARNPAGTL-EA--SAHL----RVQAPPSFQTK-PADQSVPAG 331
2FDB  68 ....LIMESVVPSDKGNYTCVVENEYGSI-NH--TYHLDVVERSRHRPILQAGLPANASTVVG 123
1EPF  60 DSSTLTIYNANIDDAGIYKCVVTAEDGTQSEA--TVNV----KIFQKLMFKNA-PTPQEFKEG 115
2ID5 397 ...........................................................VDEG 400
1TLK  25 ........................NEDAFL-EE--VAEE----KPHVKPYFTKT-ILDMDVVEG  56

Fil   36 DKGDGSCDVR-----------------YWP-TEPGEYAVHVIC--DDE-DIRDSPFI--AHILP--APPDCFPDKVKAFGPGLE   94
IG   950 EKGNGVYVAT-----------------LTLGSAAGQLSVMPRV--NGQ-NAVAQPLV--LNVAG--DASKAEIRDMTVKVNNQL 1009
1W8N 415 RQAKGQVTIT-----------------VPAGTTPGRYRVGATL--RTS-AGNASTTF--TVTVGLLDQARMSIADVD-SEETAR  475
SAX  332 GTATFECTLVGQPSPAYFW--------SKEGQQDLLFPSYVSA--D-G-RTKVSPTG--TLTIEEVRQVDEGAYVCA-GMNSAG  400
2FDB 124 GDVEFVCKVYSDAQPHIQWIKHVEKNGSKYGPDGLPYLKVLKA--A-GVNTTDKEIE--VLYIRNVTFEDAGEYTCL-AGNSIG  201
1EPF 116 EDAVIVCDVVSSLPPTIIW--------KHKGRDVIL-----KK--D-V-RFIVLSNN--YLQIRGIKKTDEGTRYCE-G.....  174
2ID5 401 HTVQFVCRADGDPPPAILW--------LSPRKH------LVSAKSN-G-RLTVFPDG--TLEVRYAQVQDNGTYLCI-AANAGG  465
1TLK  57 SAARFDCKVEGYPDPEVMW--------FKDDN-----PVKESR--H-F-QIDYDEEGNCSLTISEVCGDDDAKYTCK-AVNSLG  122

Fil   95 PT-GCIVDKPAEFTIDARAAGKGDLKLYAQDADGCPIDI--KVIPNGNGTFRCSYVPTKPI  152
IG  1010 AN-GQSANQITLTVVDSYGNPLQGQEVTLTLPQGVTSKTGNTVTTNAAGKVDIELMSTVAG 1069
1W8N 476 ED-GRASNVI-------DGNPST......................................  490
SAX  401 SSLSKAALKV-------TTKAVT......................................  416
2FDB 202 ISFHSAWLTV...................................................  211
2ID5 466 NDSMPAHLHV-------RS..........................................  477
1TLK 123 EATCTAELLV-------ET..........................................  134

Fil  153 KHTIIISWQGVNVPKSPFRVNVGEGSHPERVKVYGPGVEKTASRPMSPPTFTVDCSEAGQGDVSIGI  219
IG  1070 EHSITASVNN---AQKTVTVKFKADFSTGQATLEVDGSTPKVANDNDAFTLTATVKDQYGNLLPGAV 1133
```

Fil 109942357 Filamin [*Tania solium* (pork tapeworm)] 697aa
IG   124526630 Ig domain protein [*E. coli*] 1418aa 10%, 0.20
1W8N 55670665 Neuraminidase [*Micromonospora viridifaciens*] 601aa 16%, 3.2 Q:IG
SAX   72003712 Sensory AXon guidance family member [*C. elegans*] 1273aa 14%, 0.62 Q: 1W8N
2FDB 90109330 Fibroblast growth factor receptor [ ] 220aa 31%. 5e-06 Q: SAX
1EPF 11514267 Neural cell adhsion molecule - NCAM [ ] 191 aa   27%, 2e-04 Q: SAX
2ID5 116668111 Lingo-1 ectodomain [ ] 477aa 33%, 0.001 Q:SAX
1TLK   3024078 Telokin [ ] 154aa 30%, 0.002 Q:SAX

Red indicates identity to Fil or IG

Additional Evidence (2)

```
IG    906 ...............MTTELTF-TVKD-----AYGNPVTGLKPDAPVFSGAASTGSERPSAGNWTEKGNG  40
1W8N  364 ...........LEPGQQVTVPV-AVTN-----QSGIAVP--KPSLQLDASPDWQVQGSVEPLMPGRQAKG 419
SAX   277 AKDNRGLRIERVQPSDEGEYVC-YARN-----PAGTLEA--SAHLRVQAPPSFQTKPADQSVPAGGTATF 336
KAP1   89 SGTDFTLTISSLQPEDSAAYYCQQVYNAPPGASIGSNTR--TSVSHTKGAVLMTQSPSSLSASVGDRVTI 156
KAP2   13 ....................SC-ERIN-----HGGSVFN--FSIVDARCDIQMTQSPASLSASVGETVTI  54
LAM1    2 ...................................................YVLIQPPSVSVAPGQTASL  20
LAM2   24 ...................................................YVLTQPPSVSVSPGQTARI  42
XL3    22 .....................................................ITQPVSESVKLGETVRI  38

Fil    41 SCDVR-------------YWP-TEPGEYAVHVIC--------DDEDIRDSPFI-----AHILP--APPDCFPDKVKAFGPGLE   94
IG    955 VYVAT-------------LTLGSAAGQLSVMPRV--------NGQNAVAQPLV-----LNVAG--DASKAEIRDMTVKVNNQL 1009
1W8N  420 QVTIT-------------VPAGTTPGRYRVGATL--------RTSAGNASTTF-----TVTVGLLDQARMSIADVD-SEETAR  475
SAX   337 ECTLVGQPS--PAYFW--SKEGQQDLLFPSYVSA---------DGRTKVSPTG-----TLTIEEVRQVDEGAYVCA-GMNSAG  400
KAP1  157 TCQASQGIT--NDLAWYQQKPGETPKLLIYEASSLQSGIP----SRFSGSGSGTDF--TLTISSLQSEDFATYYCQ.......  224
KAP2   55 TCRASGNIH--NYLAWYQQKQGKSPQLLVYNAKTLAEGVP----SRFSGSGSGTQY--SLKINSLQPEDFGSYYC........  121
LAM1   21 TCGGDNIGS--TNVHWYQQKPGQAPILVVYDDKDRPSGIP----ERFSGSNSGHTA--TLTISRVEAGDEADYFCQVWHSNS.   94
LAM2   43 TCSADALPK--QYAYWYQQKPGQAPVLVIYKDSERPSGIP----ERFSGSSSGTTV--TLTISGVQAEDEADYYCQSADSSGT  117
XL3    39 SCTLSGASISGYHVNWYQQKAGNRPRYLLRFYSDSNKHQGDGVPDRFSGSKDSPNNIGYLTIKGALLEDDADYYCATWHASSY  121

Fil    95 PT-GCIVDKPAEFTIDARAAGKGDLKLYAQDADGCPIDI--KVIPNGNGTFRCSYVPTKPI  152
IG   1010 AN-GQSANQITLTVVDSYGNPLQGQEVTLTLPQGVTSKTGNTVTTNAAGKVDIELMSTVAG 1069
1W8N  476 ED-GRASNVI-------DGNPST......................................  490
SAX   401 SSLSKAALKV-------TTKAVT......................................  416
```

```
Fil  109942357 Filamin [Tania solium (pork tapeworm)] 697aa
IG   124526630 Ig domain protein [E. coli] 1418aa 10%, 0.20
1W8N  55670665 Neuraminidase [Micromonospora viridifaciens] 601aa  16%, 3.2  Q:IG
SAX   72003712 Sensory AXon guidance family member [C. elegans] 1273aa  14%, 0.62 Q: 1W8N
KAP1 109104346 Immunoglobulin kappa constant [Macaca mulatta] 239aa  21%, 1e-06
KAP2    967179 Vk12-13 Ig kappa chain variable region [ ] 128aa 25%, 3e-06 Q: SAX
LAM1  22095209 Immunoglobulin light chain variable region [Human] 113aa  29%, 0.037 Q: SAX
LAM2  21669541 Immunoglobulin lambda light chain VLJ region [Human] 271aa 26%, 6e-06 Q: SAX
XL3   17467119 Immunoglobulin light chain type III [Xenopus laevis] 154aa 21%, 0.037 Q: SAX
```

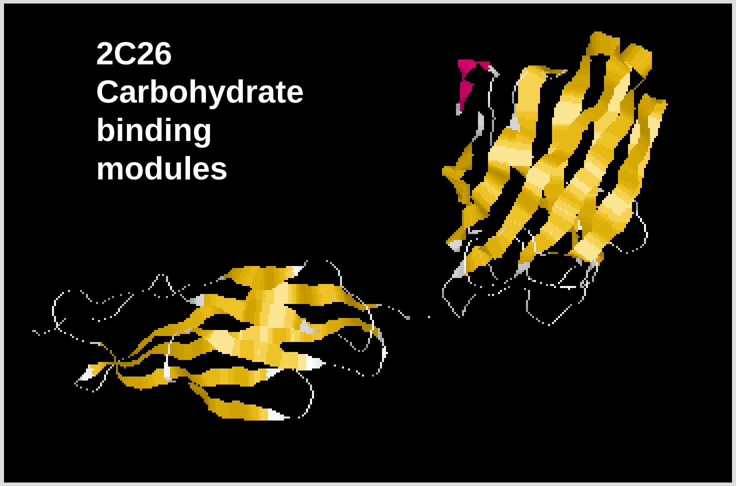
Red indicates identity to Fil or IG

Additional Evidence (3)

```
IG    1009 LANGQSANQITLTVVDSY-GNPLQGQEVTLTLPQGVTSKTGNTVT--------------TNAAGK---VDIEL---MSTV 1067
TTR1    20 LVNAAQQNILSVHILNQQTGKPAADVTVTLEKKADNGWLQLNTAK--------------TDKDGR---IKALW---PEQT   79
2G2N    16 ...................GKPAADVTVTLEKKADNGWLQLNTAK--------------TDKDGR---IKALW---PEQT   56
TTR2    25 ......SNILSVHILDQQTGKPAPGVEVVLEQKKDNGWTQLNTGH--------------TDQDGR---IKALWPE-KAAA   81
TTR3     9 .........LTTHVLDTAAGRPAAGMEIALYRFNGDMRTHLKTVR--------------TNADGR---CDAPLLEGRSFT   62
MCO    530 ....LPPAVPTRITFDINPKEPVAGTPVTFTAQLAGLKKDAAPASFVEFVIDGGGHPVFTIEDGV---ATYTT---TFRK  599
IgK     16 ....................PPSPAELATGTATIVCVANKYFPDGTVTWKVDGITQSSGINNSRTPQDPTYCT---YNLS   72
2C26     4 ....VPENQAPKAIFTFSPEDPVTDENVVFNA-----SNSIDEDGTIAYYVWDFGDGYEGTST-T---PTITY---KYKN   67

IG    1068 A--GEHSITASVNNAQKTVTVK.......... 1087
TTR1    80 ATTGDYRVVFKTGDYFKKQNLE..........  101
2G2N    57 ATTGDYRVVFKTGDYFKKQNLE..........   78
TTR2    82 P--GDYRVIFKTG...................   91
TTR3    63 P--GRYEI........................   68
MCO    600 P--GEHKLSVRYNGDDVYS----DSSSDAVQS  611
IgK     73 S--TLTLSSDEYNSHNEYTCQVADSGSPVVQ.  102
2C26    68 P--GTYK--VKLIVTDNQGASSSFTATIKVTS   95
```

Red indicates identity to IG

```
IG   124526630 Ig domain protein [E. coli] 1418aa 10%, 0.20
TTR1   3915454 Transthyretin-like [E. coli] 137aa  18%, 1.8
2G2N 122920292 Transthyretin-related [E. coli] 114aa 19%, 3.3
TTR2  20455396 Transthyretin-like [Salmonella enterica] 136aa 19%, 0.55
TTR3 110635165 Transthyretin [Mesorhizobium] 121aa  33%, 0.17
MCO   94968524 Multicopper oxidase [Acidobacteria bacterium] 631aa, 18%, 7.8
IgK    1552361 Kap light chain [Orcytolagus cuniculus] 109aa, 22%, 0.75  Q: MCO
```



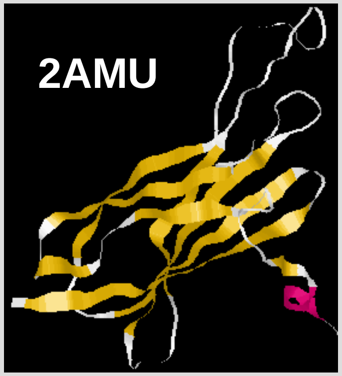
Domain on left is a standard Ig-domain.
Domain on right is Ig-related pentraxin-like domain

Superoxide reductase may be evolutionarily linked to immunoglobulins

```
TAT     1 ..MNR--RTFLKTAALGAVAA--GITREAAAAAEKYFPV-KADQSLFATINRAKDPAKKTPL  55
SOY     1 MYVTR--RKLFALSAGAATAGMFAFAPGRAFATVEATEKALADFTGG------KTPETGK--  52
TONB1   1 ..MSF---RRLLASTFAGALVASAAQAQNAPA------------GSG------ASQGVGD--  38
TONB2   5 ..........VLLAGGAAPLALIAPGAFAAEAPVEAVPTATAPVDAP------ADAQSGD--  58
TONB3   1 ..MTR--KTVLSILAGVSLAALANGA--SAQTVPDQADDQRQGL--------------EE--  40
TONB4   5 .................LAAYAAGLMACSSFTAPAFAQDNETQAEPS------PAANDRV--  42
TONB5   1 MDARRMKRAFL---ATASALLVLPAVPALAQAAEQDA--------GY------NPASLGD--  43

SOR    13 GEKHVPVI---EYEREGELVKVKVQV-GKEIPHPNTTEHHIRYI-ELYF-L-PEGENFVYQVGR   69
TAT    56 EQKHAPVIKAPHAVKAGEPFTVEVTV-GEQV-HPMGPTHWIEYI-ELNV-GN-EP------AGR 108
SOY    53 -----ITLTAPEIAENGNTVPISVDV-ES----PMTDDSYVESVTIFAE-GNPNPE-----VAT 100
TONB1  39 -----IVVTARRRAESLQNTPVAVSAITSAALE-QKGATNIAAVA...................  77
TONB2  59 -----IVVTARRRAETAQDVPLAISV......................................  79
TONB3  41 -----IVVTARRQAENLQTTPVSVSA-VS---EKMLARANVTQI....................  75
TONB4  43 -----IIVTATRRAQDIQDVPIAVTA---------ATQEQLDRQGVVNV-QN-ITQV----SPS  85
TONB5  44 -----IVVTARKREESVQTTPLSISA-FG----AQALQD--RNVQSSADIANFVPNV------Q  89

SOR     70 VE--FTAHGESVNGPNTSDVYTEPIAYFVLKTKKKG------KLYALSYCNIHGLWENEVTL..... 123
TAT    109 IA--MQPRG-----------FLHPKVTFTVVIPKEAAPAGKITLVAHQRCNLHGYWEGSLDVAVT.. 160
SOY    101 FH--FTPMS-----------GAAA-ATTRIRL------AKTQNVIAVAKMSDGSTYSDRKEVKVTIG 145
TONB4   86 FS--TSQAQ-----------IASG--TVVLRIR---GVGTTSN........................ 110
TONB5   90 FDSAASESG-----------GGAS-SQISIRG------IGQTDYVITVEPAVG-LYLDGVYVGKSVG 137
```

```
SOR    18977653 Superoxide reductase  (2AMU)
TAT    78223810 Twin arginine translocation signal [Geobacter]  28%, 2e-20
SOY    90417856 Sulfur oxidation Y protein [Aurantimonas]_Q:TAT 15%, 4e-22
TONB1  22417099 Prob. TonB-dep. receptor [Sphingobium]  Q:SOY  20%, 0.019
TONB2  87199725 TonB-dependent receptor [Novosphingobium] 28%, 0.094
TONB3 118759028 TonB-dependent receptor, plug [Sphingomonas] 23%, 0.64
TONB4  85710277 TonB-dependent receptor [ErythrobacterI] 15%, 0.90
TONB5 118760864 TonB-dependent receptor [Sphingomonas] 21%, 0.85
```

Plug domain of TonB may be a homolog of immunoblobulins (see below)

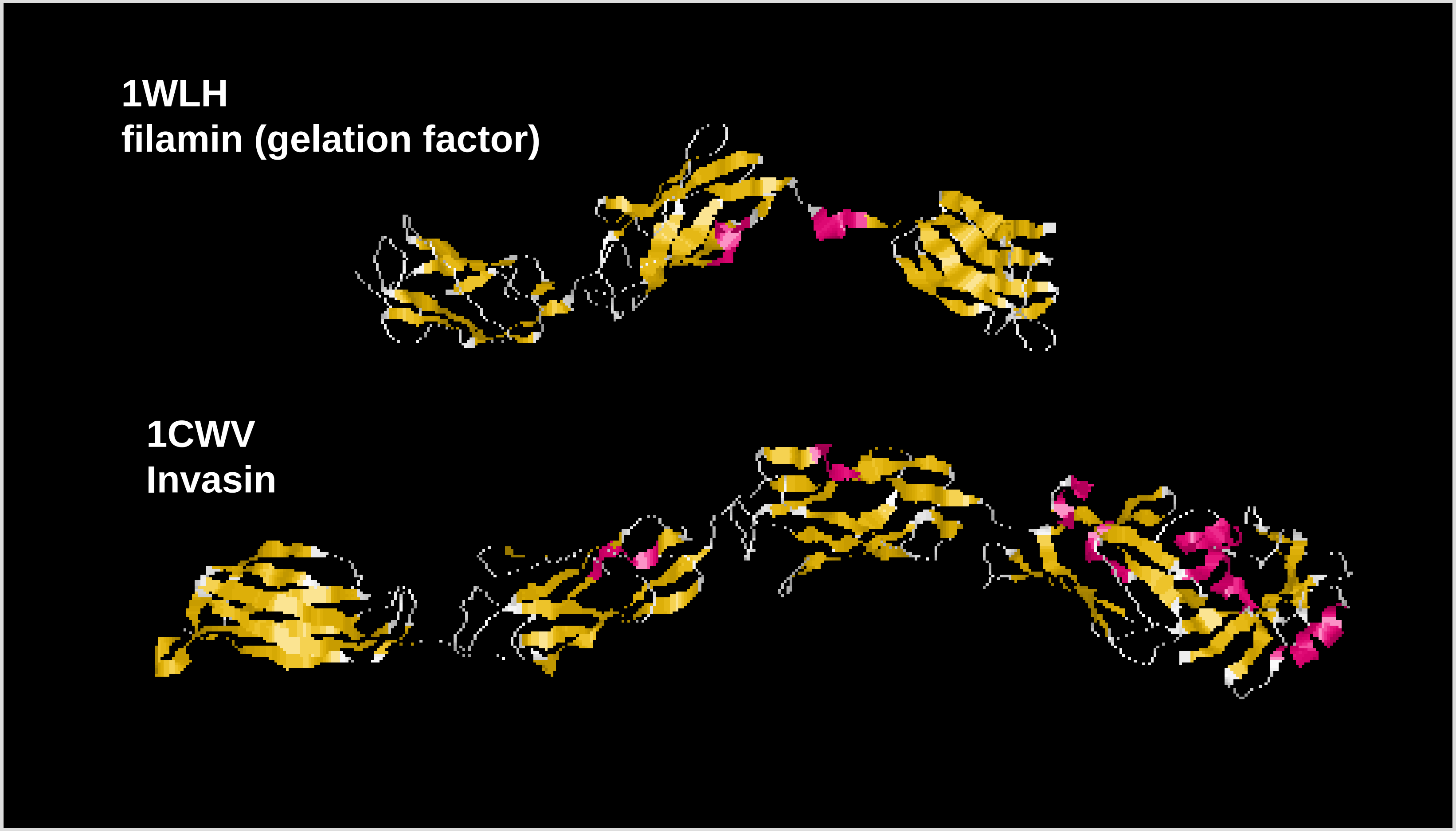


Invasins and intimins may be homologs of filamins

```
1WLH       16  DGGECFQPSKFKIHAVDPDGVHRTDGGDGFVVTI-EGPAPV--DPVM---VD-NGDGTYDVEFEP-KEAGDYVINL   83
Int       527  ..........VTARAYDRNGN--SSNNVLLTITV-LSNGQVVDQVGV---TD-FTADKTSAKADG-TEAITYTATV  584
IntR      518  ...VQGGSNIYKVTARAYDRNGNSSNNVQLTITV-LSNGQVVDQVGV---TD-FTADKTSAKADN-ADTITYTATV  584
Inv       798  DHVKAGESTTVTLVAKDAHGNAIS--GLSLSASL-TGTAS--EGATVSSWTE-KGDGSYVATLTTGGKTGELRVMP  867
1CWV       18  DGAPANGKTAITVEFTVADFEGKPLAGQEVVITTNNGALP-NKITEK---TDANGVARIA--LTN-TTDGVTVVTA   86

1WLH       84  TLDGD---NVNGFPKTV--TVKPAPSAEHSYAEGEGLVKVFDNAPAE-FTIFAVDTKGVAR...  138
Int       585  KKNGVAQANVPVSENIVSGTAVLSANSANTNSSGKATVTLTSNKPDQ-VVVSAKTAEMTSA...  644
IntR      585  KKNGVAQANAPVTFSIVSGTATLGANSAKTDGNGKATVTLKSGTPGQ-VVVSAKTAEMTSA...  644
Inv       868  LFNGQ---PAATEAAQL--TVIAGEMSSANSTLVADNKTPTVKTTTE-LTFTVKDAYGNPVTGL  925
1CWV       86  EVEGQ-----RQSVDTH--FVKGTIAADKSTLAAVPTSIIADGLMASTITLELKDTYGPDQ...  140

1WLH      139  TDGGDPFEVAINGPDGLVVDAKVTDN-NDGTYGVVYDA-PVEGNYNVNVTLRG--------------NPIKNMPID  198
Int       645  LNANAVIFVD-------QTKASITEIKADKTTAVANNQ-DA-ITYTVKVMENG--------------QPLSGE--E  695
IntR      645  INAGSVIFID-------QTKASITEITNDKSTAIANDK-DA-ITYTVKVMKND--------------QPVPNHLVT  697
Inv       926  KPDAPVFSGAASTGSERPSAGNWTEKGN-GVYVSTLTLGSAAGQLSVMPRVNG--------------QNAVAQPLV  986
1CWV      141  AGANVAFDTTL-GNMGV-----ITDH-NDGTYSAPLTS-TTLGVATVTVKVDGAAFSVPSVTVNFTADPIPDAGRS  208

1WLH      199  VKCIEGANGEDSSFGS-------------FTFTVAAKNKKGEVKT-YGGDKFEVSITGPAEEITLDAID 253
Int       696  VTFFTDFGALDKTKVT-------------TDQSGYATVKNLSSST-SGKAIVRAKVSDVDTEVKAAAVE 750
IntR      698  FTTTFGKFNGKQSSET-------------VTTGNDGRAIVTLTSGLAGKAIVSAKVNEVNTEVKAKTVE 753
Inv       997  LNVAGDASKAEIRDMTVKVNNQLANGQSANQITLTVVDSYGN-PL-QGQEVTLTLPQGVSTKTGNTVTT 1053
1CWV      209  SFTVSTPDILADGTMS-------------STLSFVPVDKNGHFIS--GMQGLSFTQNGVPVSISP-ITE 261

1WLH       254  NQDGTYTAAYSLVGNGRFSTGVKLNG  274
Inv       1054  NAAGKVDIELMSTVAGEHNISASVNG 1079
1CWV       260  QPD-SYTATVVGNSVGDVTIT.....  281
```

1WHL	55670676	Rod domain of *Dictoystellium* filamin
Int	1947048	Intimin 15%, 0.038
IntR	48714777	Intimin rho [*Escherichia coli*] 13%, 0.004
Inv	110640572	Putative adhesin/invasin [*Escherichia coli*] 13%, 4e-15
1CWV	6435735	Invasin 18%, 1e-06

Filamins may be homologs of invasins and intimins



Ig family protein (*Solibacter usitatus*) may be a homolog of PKD and cadherin domains.

```
IgF        809 SGSGFVYTVASGSALPPG--LTLNAGTGVIS-GTPT--------TPGTYMVRTVVTDSVGGTDDVTCT  865
1E07       137 TQDATYLWWVNNQSLPVSPRLQLSNGNRTLTLFNVT--------RNDTASYKCETQNPVSARRSDSVL  196
1L3W       374 .......SYFIGNDPARW--LTVNKDNGIVT-GNGNLDRESEYVKNNTYTVIMLVTDDGVSVGTGTGT  431

IgF        866 IIVAGPPLNLVCGTCGN---SKATVGSAYSSTLAVQGGTASFTF----SIVSGSLPP-G-LTLNPTTGA  925
1E07       197 LNVLYGPDAPTISPLNT---SYRSGENLNLSCHAASNPPAQYSW-----FV--------NGTFQQSTQE  249
2C26         5 ....PENQAPKAIFTFS---PEDPVTDENVVFNASNSIDEDGTI----AYYVWDFGD-G-YEGTSTTPT   60
1L3W       432 LILHVLDVNDNGPVPSPRVFTMCDQNPEPQVLTISDADIPPNTYPYKVSLSHGSDLT-WKAELDSKGTS  499
1L3W       373 ...........................................L----SYFIGNDPARW-LTVNKDNGI  393

IgF        926 I---TGTPTA--------TGTYTFTSKVVDANGTSDTAQCGIV--VVASPVNLDCGSCGSNRATLGTAYT  982
1E07       250 LFIPNITVNN--------SGSYTCQAHNSD-TGLNRTTVTTIT--VYAEPPKPFITSNNSNPVEDEDAVA  308
2C26        61 I---TYKYKN--------PGTYKVKLIVTDNQGASSSFTATIK--VTSATGDNSKFNFEDGTLGGFTTSG  117
1L3W       500 M---LLSPTQQLK-----KGDYSIYVLLSDAQNNPQLTVVNAT--V........................  535
1L3W       394 V---TGNGNLDRESEYVKNNTYTVIMLVTDDGVSVGTGTGTLILHVLDVNDNGPVPSPRVFTMCDQNPEP  459

IgF        983 SKLTVSGGK-----ASYAYSIISGALPAG--ITLKSDG----TISGTPT---ATGTFTFTSKVVDA 1034
1E07       309 LTCEPEIQN-----TTYLWWVNNQSLPVSPRLQLSNDNRTLTLLSVTRN---DVGPYE........ 358
2C26       118 TNATGVVVN-----TTEKAFKGERGLKWT--VTSEGEG----T....................... 149
1L3W       460 QVLTISDADIPPNTYPYKVSLSHGSDLTWK-AELDSKG---TSMLLSPTQQLKKGDYSIYVLLSDA 522
```

IgF	116624946	Ig family protein [*Solibacter usitatus*]
1E07	82407267	Carcinoembryonic antigen [human] 16%, 6.1
2C26	82408215	Carbohydrate binding module (PKD domain) 15%, 2.5
1L3W	20664275	Cadherin ectodomain 17%, 0.056; 20%, 0.14

Red: identity to IgF
Blue: identity to 1E07

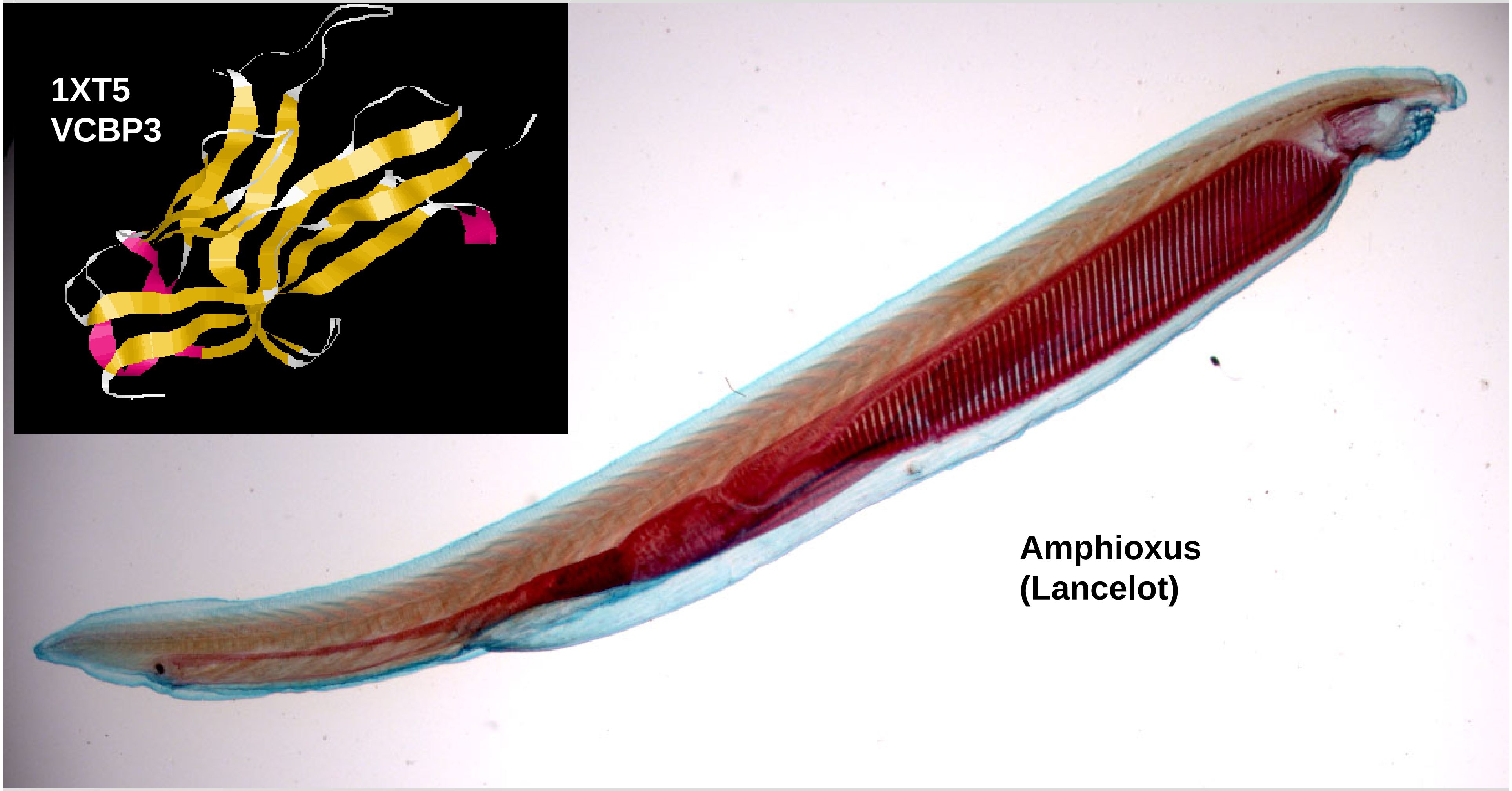
Cadherins are related to immunoglobulins

The "immune-type" receptor in the non-vertebrate Amphioxus may be an immunoglobulin.

```
VCBP3    1 ....IMTVRTTHTEVEVHAGGTVELPCSYQLANDTQ-PPV-ISWLKGASP----DRSTK-VFKGNYNW  60
2AVG    25 ...............EVTVGGSITFSARVAGASLLK-PPV-VKWFKGKWV----DLSSK-V-------  63
2J12     8 .....LSITTPEEMIEKAKGETAYLPCKFTLSPEDQ-GPLDIEWLISPADNQKVDQVIILY-SGDKI-  66
BUTY    32 .......VTAPQEPVLALVGSDAELTCGFSPNASSE-YME-LLWFR---Q----TRSTA-V--LLYRD  80
1PKO    12 ..........PGHPIRALVGDEAELPCRISPGKNAT-GXE-VGWYRSPFS-----RVVHLYRNGK---  59
PVR      1 ............PEVRGQLGGTVELPCHLLSPLPGLFVSL-VTWERSDVP----VKQQN-VAAFHPKL  50
SIRP    42 ................VGAGGSATLNCTVTSLLPVG-P---IRWFKGVGQ----SRLLI-YSFTGERF  84
LAM      2 ......TVVTQEPSLTVSPGGTVTLTCASSTGAVTS-GSY-ANWFQQKPG----QAPRALIYSTSNR-  56
KAP      3 ....ALVMTQTPASVEAAVGGTVTIKCQASQSIS----NL-LAWYQQKPG----QPPKLLIYYASNL-  56

VCBP3   61 QGEGLGFVESDSYKESFGDFLGRASVAN------LAAPTLRLTHVHPQDGGRYWCQVAQWSIRTEFGLDAKSVVLKV 131
2AVG    64 -GQHLQL--HDSYD--------RASKVY------LF--ELHITDAQPAFTGGYRCEVS................... 102
2J12    67 ------------YDDYYPDLKGRVHFTS--NDLKSGDASINVTNLQLSDIGTYQCKV.................... 110
BUTY    81 GQEQEGQQMTE--------YRGRATLAT--AGLLDGRATLLIRDVRVSDQGEYRCLFKD---NDDF--EEAAVYLKV 142
1PKO    60 ----------DQDAEQAPEYRGRTELLK--ESIGEGKVALRIQNVRFSDEGGYTC...................... 102
PVR     51 GAS---FPSPEPGSERL-SFVSAKQSTGQDTEAELQDATLALQGLTVEDEGNYTCEFA................... 104
SIRP    85 PRITNV---SDVTKRSNLDF------------------SIRISNVTPADSGTYYCVKFQRGPS.............. 126
LAM     57 --------HSWTPARFSGSLLG-------------GKAALTLSGVRPEDEADYYC......................  90
KAP     57---------ASGVPSRFKG-------SRS------GTEFTLTISDLECADAATYYCQCTYSSSTGTFGGGTKVVV... 109
```

VCBP3	78100871	1XT5 Variable region-containing chitin-binding protein [*Branchiostoma floridae*]
2AVG	114793452	CC1 comain from human cardiac myosin binding protein C 31%, 2.4
2J12	114794891	Ad37 fibre head 21%, 2e-06
BUTY	7304935	Butyrophilin [*Mus musculus*] 23%, 4e-04
1PKO	34810546	Myelin oligodendrocyte glycoprotein 17%, 5e-04
PVR	26106008	Poliovirus receptor related [*Cebus apellai*] (capuchin monkey) 25%, 0.004
SIRP	68303969	SIRP beta 1 cell surface protein [*Mus musculus*] 25%, 0.014
LAM	95007553	Lambda light chain variable region [Human] 21%, 0.028
KAP	6502877	Kappa light chain variable region [*Oryctolagus cuniculus*] (domestic rabbit) 19%, 3e-07

The "immune-type" receptor in the non-vertebrate Amphioxus may be an immunoglobulin.



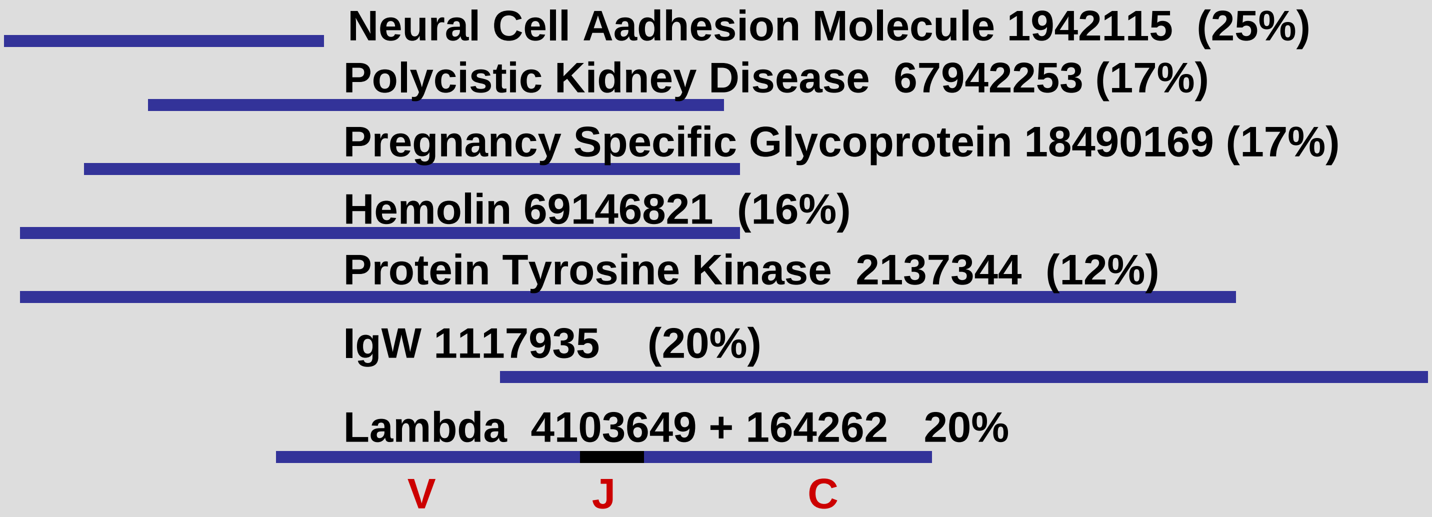
Bacteriophage Head Outer Capsid Protein (Hoc)

```
Hoc         94  ENNSTVAVTPASPAAV-EI---------GTATTFT---ANVSNQPS-GAAYIATWKVDGVA----VD 143
lambda     130  .SQPTLTLMPPSPEEV-KAK--------GTATLV---CLADHFYPD--EV-GVEWKKDGAA----IS 178
kappa      144  ...........QPSEE-QLQT-------GSAS---VVCFVNNFYPK--AA-TVQWKVDNVV----RS 182
mu         129  ..GTMLTVTNAAPSAP-SPFILFTCEDQGSSGSFTYGCLALGYSPA-GAS--VSWKKDDIKLETGVK 189
IgA2       251  CPSVSVSLHPPSLESL-FLDK-------G--ANLT--C-ELTGVSN-VKGVNFSW--------SPLS 195
IgD        140  ASTTAPSVFPLAPSCGSTS---------GSTVALA--C-LVSGYFP--EPVTVSWNSG--------- 183
FcR         20  ............................DSVT-LT--C-RGTHSPE-SDS--IQWFHNGNL----IP  47
SIGLEC     146  .TALTNTPQILLPETL-EA---------GHPSNLT--C-SVPWDCGWTAPPIFSWT--GTS----VS 192

Hoc        144  -----G-----------QKQSTFEYTP-TSEGTKSITCSVTVTATDYVDKTVESSAVS  183
lambda     179  -----A-----------GVQT--SNLR-ASDSTYSCSSLLTLSGSDWESNARFSCALT  216
kappa      183  -----S-----------GVVTSFTEQD-SQDSTYSLSSTLALTASDYNAYETYACEVT  222
mu         190  -----GYPAVFNKLGTYTRSSELTITR-AAAAGGDIFCVVQHNHNEYKVKVQLPDRV-  240
IgA2       196  -----G-----------TARP---VDG-PAVKDDKGKYTITSTLEVCTDEWMRGDKYT  233
IgD        184  -----S-----------LTSGVHTFPS-VLQSSGLYSLSSTVTVPSSRWPSET.....  219
FcR         48  -----T-----------HTQPSYRFKA-NNNDSGEYTCQ--------TGQTSLSDPVH   80
SIGLEC     193  FLSTNT-----------TGSSVLTITPQPQDHGTNLTCQVTLPGTNVSTRMTT----R  235

Hoc        184  LTVN----KKANSSTLKITPE  200
mu         241  --VH----HPTVTITITALDE  255
IgA2       234  CTVSHPELPKPVTKTIT....  350
FcR         81  LTV..................   83
SIGLEC     236  LNVS----YAPKNLTVTI...  249
```

Hoc	32453675	Hoc head outer capsid protein [*Enterobacteria phage RB69*] 471aa
lambda	476635	Ig lambda-like chain, V-C region [Nurse shark] 234aa 22%, 5.6
kappa	16306468	Ig kappa laight chain [*Trichosurus vulpecula* (bushtail possum)] 240aa 21%, 1.4
mu	103715	Ig mu chain [Little skate] 573aa 20%, 5.3
IgA2	17223801	IgA2 [*Ornithorhynchus anatinus* (duckbill platypus)] 484aa 12%, 4.3
IgD	17066530	Ig gamma heavy chain [*Canis familiaris*] 470aa 12%, 4.6
FcR	7245548	(2FCB) Fc gamma receptor (CD32) [Human] 20%, 2.1

Possible Structural Organization of Hoc 33620536

1 404



Neural Cell Aadhesion Molecule 1942115 (25%)

Polycistic Kidney Disease 67942253 (17%)

Pregnancy Specific Glycoprotein 18490169 (17%)

Hemolin 69146821 (16%)

Protein Tyrosine Kinase 2137344 (12%)

IgW 1117935 (20%)

Lambda 4103649 + 164262 20%

Beta Domain Protein Families that may be Evolutionarily Linked to Immunoglobulins Include:

Cupredoxin / Ephrin
Cu,Zn-Superoxide Dismutase
Superoxide Reductase
Purple Acid Phosphatase
Thiol-Disulfide Interchange Protein
Receptor/Protein Tyrosine Kinase
Lipoxygenase
Tumor Necrosis Factor
Complement C1q
Killer-cell Ig-like receptor
Leukocyte Ig-like receptor
Sialic-acid-binding Ig-like lectin (SIGLEC)
Signal regulatory proteins
CD200 receptors
Paired Ig-like receptors
CMRF35 receptors
CRP / pentraxins
Hsp70 Chaperone
PapD / Fimbrial Chaperone

TonB-dependent receptor (plug domain)
Transthyretin
Lipocalin /Retinol binding protein
Plexins / NFκB / Notch
Signal Axon Guidance proteins
Filamin
Fibronectin-III
Cadherin
LamininG
Polycystic Kidney Disease Domain
Invasin / Intimin
Hemolin
Toxins: botulinum, tetanus, alpha, delta, RTX,...
Actinoxanthin
Carbohydrate binding modules
Lipid/protein/nucleic acid binding modules
Bacteriophage Hoc Protein
Variable region containing chitin-binding proteins (VCBPS) [amphioxus]

Extreme homology searches may be an effective means for generation of experimentally testable hypotheses of protein function

-- for those proteins for which functionally characterized homologs have been identified.

Evolution has not been restricted by a requirement to maintain statistically significant sequence identity.

The hypothetical evolutionary linkage between the immunoglobulin superfamily and a diverse array of other beta-domain proteins previously thought to be "immunoglobulin-like" may be principally academic.

However, the hypothesis is testable *in silico* by identifying a chain of homologs in which each link exhibits statistical significance.

This is the computational challenge.

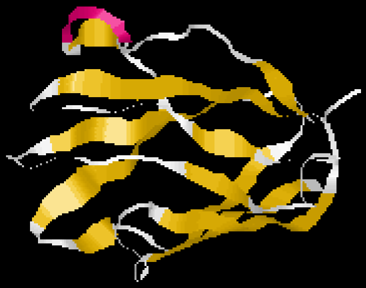
Validation of Distant Homology

Pseudoazurin
1PAZ

No significant

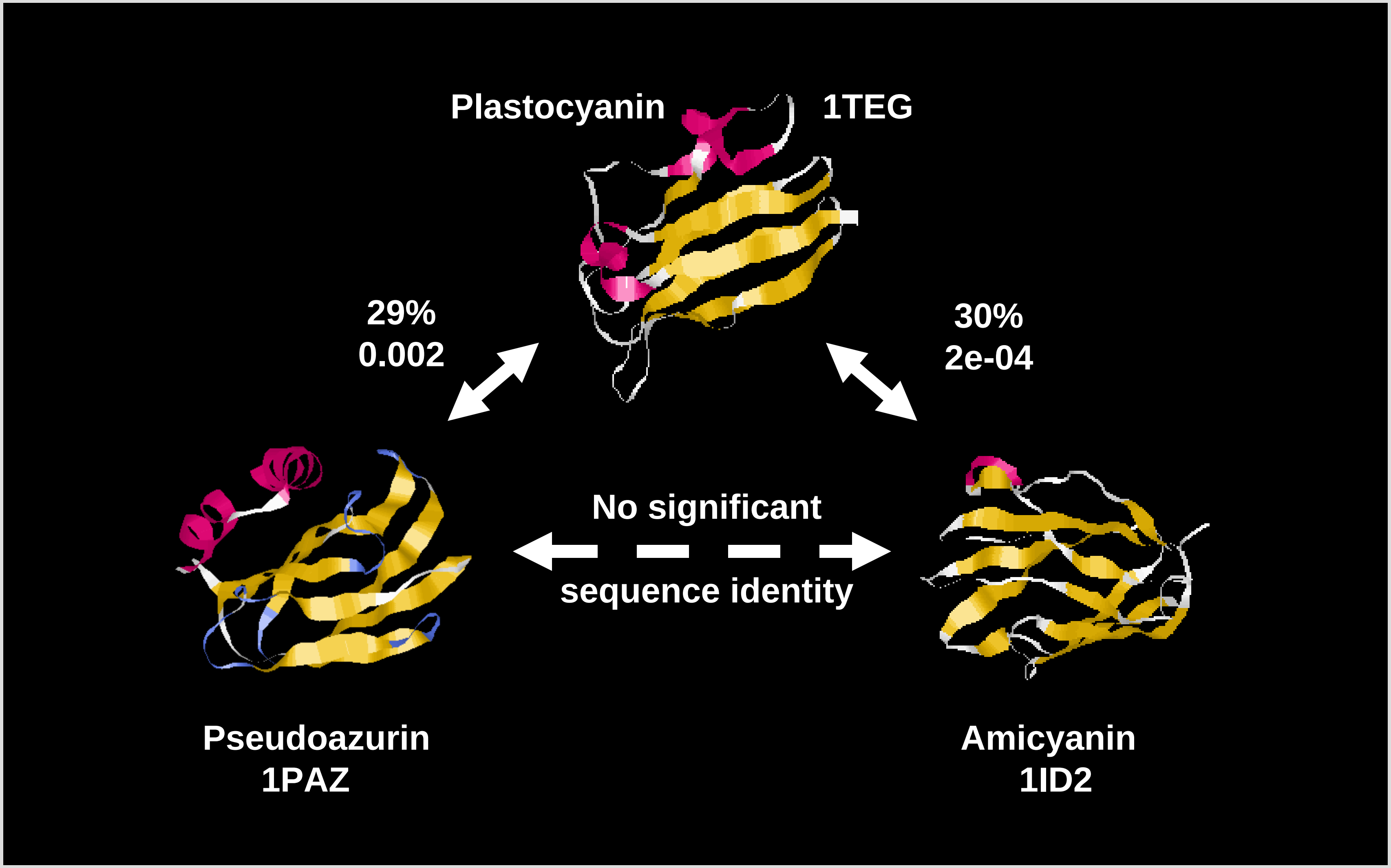
⟵ – – ⟶

sequence identity



Amicyanin
1ID2

Validation of Distant Homology



Plastocyanin 1TEG

29%
0.002

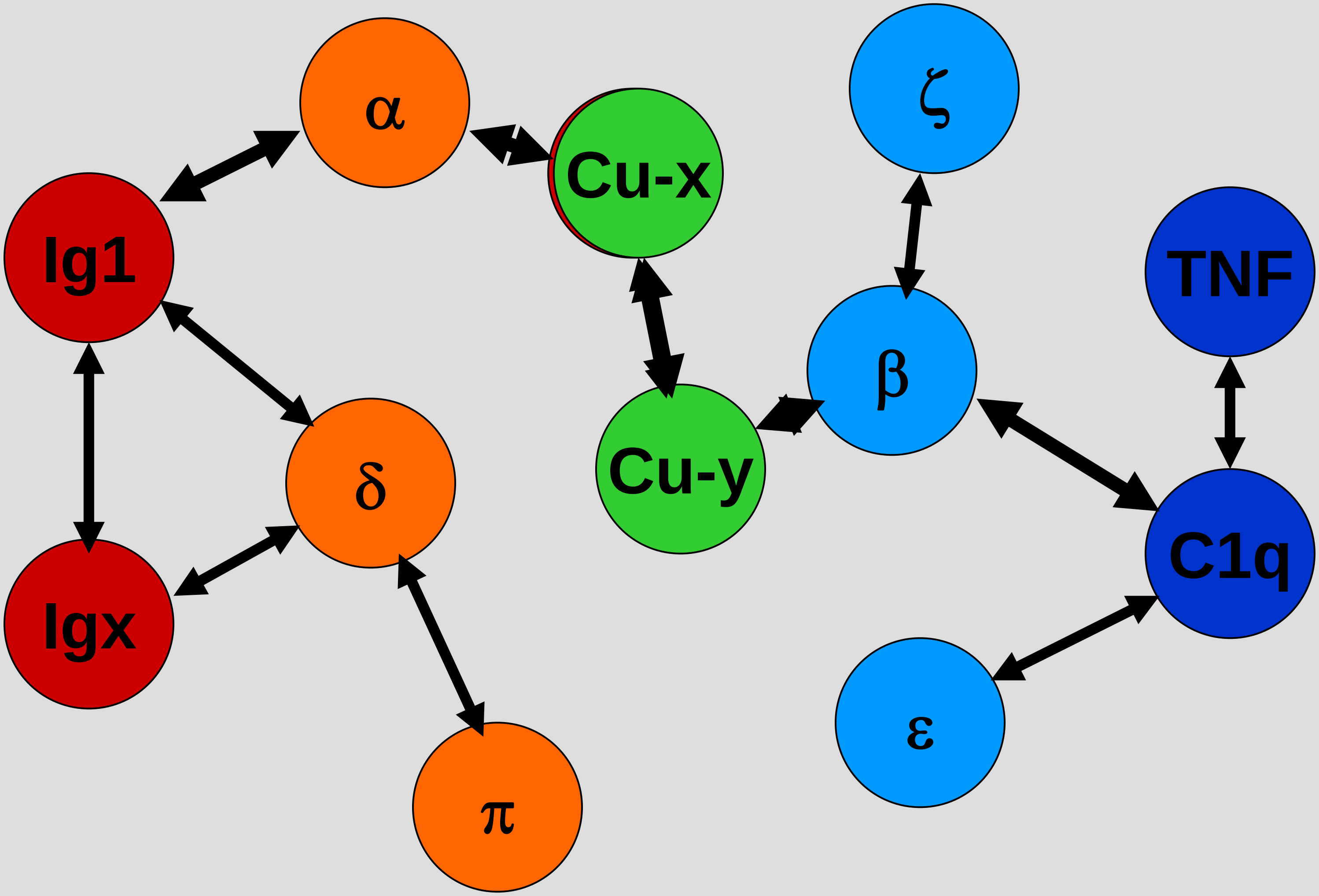
30%
2e-04

No significant
sequence identity

Pseudoazurin
1PAZ

Amicyanin
1ID2

Can an efficient algorithm be developed to find a pathway of statistical significance?

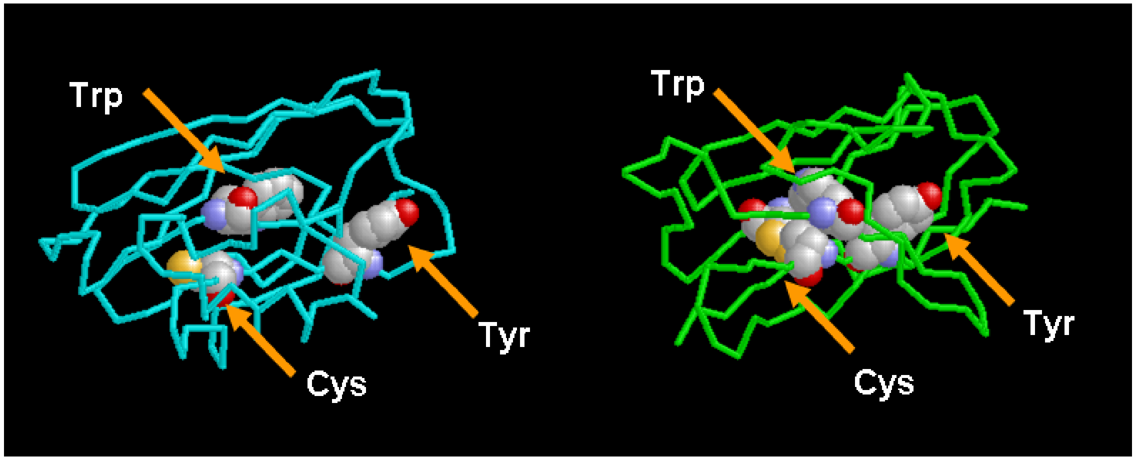


Evolutionary linkage of proteins with significantly different functions may not be only academic.

In some cases, it may be possible to merge functions by engineering new proteins that do not exist in nature.

For instance,

Cuprebodies



Step 1: Engineer hyperstable antibody (we have demonstrated feasibility)

Step 2: Remove one Cys residue and intoduce others to create a cupredoxin-like Cu-binding site.

Result: Antibodies with a built-in signalling capability

For instance,

Immunodismutases

Step 1: Engineer hyperstable MPZ or MOG (we have demonstrated feasibility with antibodies)

Step 2: Introduce functionally active amino acids found in SOD

Result: Nerve coat proteins with a built-in anti-free radical activity.

For instance,

Transcriptoglobulins

Step 1: Engineer phage-display scaffolds of hyperstable dimerizing immunoglobulin domains.

Step 2: Randomly mutate amino acids in binding loops to create diversity.

Step 3: Screen for constructs that bind to predetermined chromosomal targets.

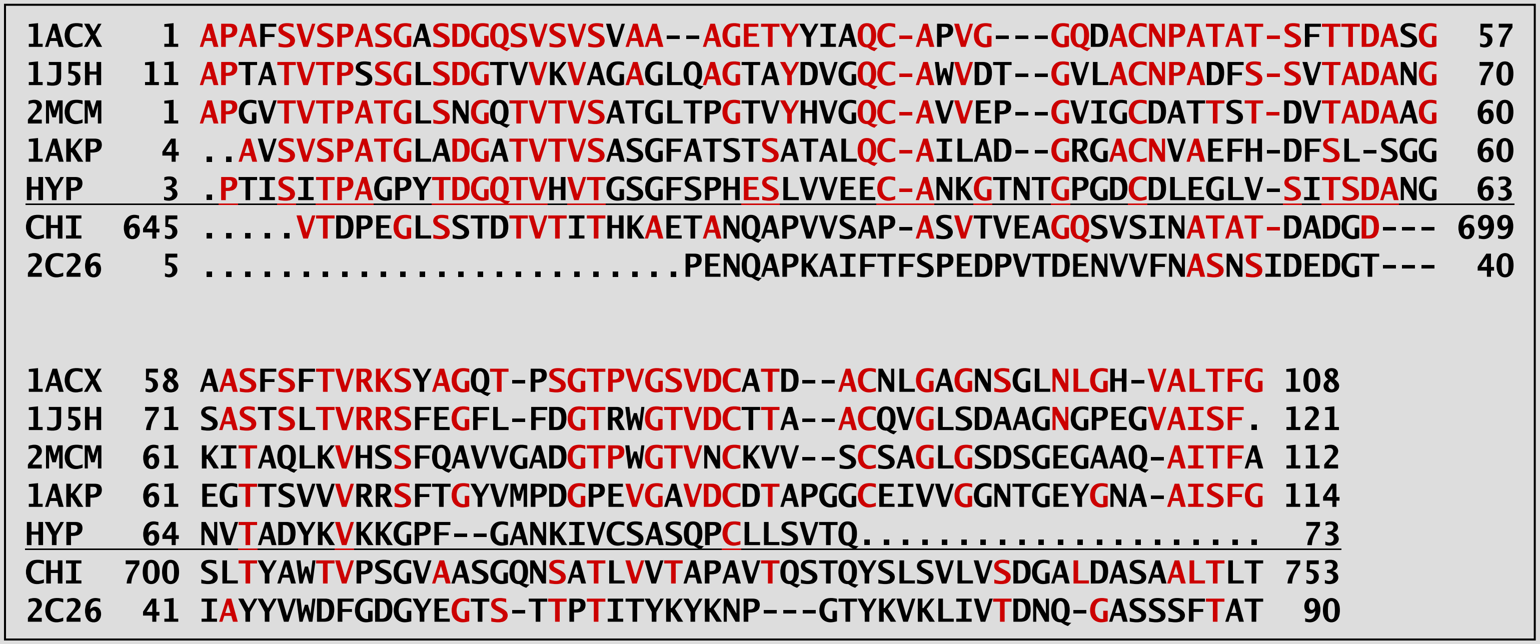
Result: Proteins that modify cellular metabolism for basic research applications, biotechnological uses, and possible future therapeutics.

query: sod1 > sod2 (39933302) > tonB (78048066) (1-150)

```
Query 23 CLLLGAAPAFAQSTAATIRGQVTVD--AAPAAQAQVTATNLATGLTRTVQVSNGGYSVGG--LP 82
Sbjct  5 IIFIWTLTAFAQECRGQISVTQSPS--TAAQPGETVKISCKTSSDVYRWSDGNEGLAWYL--QK 64
TTR    2 ................TLSTHVLDATTGRPAANVAVTLTAADTPVADGLTDADGRITGLGGELA 49

Query 83 PGSYRIDVTANGQTSSQNVTVQVGQTATLNLGVGGEPATAAGGNATTLDA 132
Sbjct 65 PG------EAPKLLIYAANTLQSGTPSRFSGSGSNSDFTLTISGVQTEDA 108
TTR   50 SGIYRLHFDTGAYFAARHVA..............................  69
```

LC   20269271 Immunoglobulin light chain (kappa) [*Cyprinus carpio*] 208aa 17%, 2.3
TTR 120406225 Transthyretin [*Mycobacterium vanbaalenii*] 105aa 23%, 1.3

Actinoxanthin superfamily may be evolutionarily related to immunoglobulins



```
1ACX    1 APAFSVSPASGASDGQSVSVSVAA--AGETYYIAQC-APVG---GQDACNPATAT-SFTTDASG   57
1J5H   11 APTATVTPSSGLSDGTVVKVAGAGLQAGTAYDVGQC-AWVDT--GVLACNPADFS-SVTADANG   70
2MCM    1 APGVTVTPATGLSNGQTVTVSATGLTPGTVYHVGQC-AVVEP--GVIGCDATTST-DVTADAAG   60
1AKP    4 ..AVSVSPATGLADGATVTVSASGFATSTSATALQC-AILAD--GRGACNVAEFH-DFSL-SGG   60
HYP     3 .PTISITPAGPYTDGQTVHVTGSGFSPHESLVVEEC-ANKGTNTGPGDCDLEGLV-SITSDANG   63
CHI   645 .....VTDPEGLSSTDTVTITHKAETANQAPVVSAP-ASVTVEAGQSVSINATAT-DADGD--- 699
2C26    5 .........................PENQAPKAIFTFSPEDPVTDENVVFNASNSIDEDGT---   40


1ACX   58 AASFSFTVRKSYAGQT-PSGTPVGSVDCATD--ACNLGAGNSGLNLGH-VALTFG 108
1J5H   71 SASTSLTVRRSFEGFL-FDGTRWGTVDCTTA--ACQVGLSDAAGNGPEGVAISF. 121
2MCM   61 KITAQLKVHSSFQAVVGADGTPWGTVNCKVV--SCSAGLGSDSGEGAAQ-AITFA 112
1AKP   61 EGTTSVVVRRSFTGYVMPDGPEVGAVDCDTAPGGCEIVVGGNTGEYGNA-AISFG 114
HYP    64 NVTADYKVKKGPF--GANKIVCSASQPCLLSVTQ.....................  73
CHI   700 SLTYAWTVPSGVAASGQNSATLVVTAPAVTQSTQYSLSVLVSDGALDASAALTLT 753
2C26   41 IAYYVWDFGDGYEGTS-TTPTITYKYKNP---GTYKVKLIVTDNQ-GASSSFTAT  90
```

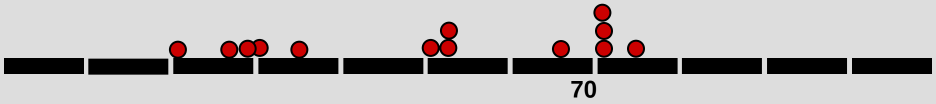
1ACX 229668 Actinoxanthin [*Streptomyces globisporus*] 108aa
1J5H 24158731 Apo-neocarzinostatin [ ] 122aa 46%, 2e-26
2MCM 230625 Macromycin [ ] 112aa 35%, 9e-19
1AKP 729892 Apokedarcidin [ ] 114aa 36%, 5e-15
HYP 117927356 Hypothetical protein [*Acidothermus cullulolyticus*] 111aa 30%, 0.049
CHI 114048379 Chitinase [*Shewanella sp*] 868aa 18%, 1.9 Q: HYP
2C26 82408215 Carbohydrate binding module [ ] 260aa 18%, 1.3 Q:CHI

Red indicates identity to 1ACX

1QHO Five domain alpha-amylase from Bacillus stearothermophilus

gi 8569360 plexin - related

```
LIPO   31 PPLSKVPLQQNFQDNQFQ-G--KWYVVGLAGNAILR-EDK------------DPQKMYATIYELKEDKS-----YNVTSVL 90
1I4U   27 ............NRNSYA-G--VWYQFALTNNPYQL-IEKCVRNEYSF-----------------DGKQ-----FVIESTG 69
1SY0    8 .......AQTGFNKDKYFNG-DVWYVTDYLDLEPDD-VPKRYCAALAAGTASGKLKEALYHYDPKTQDT-----FYDVSEL 74
1AVG    6 ..CSIEKAMGDFKPEEFFNG--TWY-LAHGP------GVT------------SPAVCQKFTTSGSKGFT-----QIVEIGY 58
1QAB    1 CAVSSFRVKENFDKARFS-G--TWYAMA----KKDP-EGL------------FLQDNIVAEFSVDETGQ-----MSATAKG 55
1E5P    2 ............DFAELQ-G--KWYTIVIAADNLEKIEEG------------GPLRFYFRHIDCYKNCS----EXEITFYV 51
ICYN   36 GYCPDVKPVDDFDLSAFA-G--TWHEIA-----KLPLENE------------NIGKCTIAEYTVNGGKASVYNSFVVNGVK 84
1IW2   11 SPISTIQPKANFDAQQFA-G--TWLLVAVGSAGRFLQEQG------------HRAEATTLHVAPQGTAMAVSTFRKLDGIC 76

LIPO   91 F-----RKKKCDYWIR-TFVPGCQPGEFTLGNIK-------SYPGLTSYL--VRVVSTNYNQHA-MVF-FK---KVSQNREYFK 154
1I4U   70 I-----AYDGNLLKRNGKLYPNPF-GEPHLSIDY-------ENSFAAPLV----ILETDSSNYACLYS-CI---DYNFGYHSDF 132
1SY0   75 Q-----VES-LG-KYTANFKKVDKNGNVKVAVT-----------AGNYYT--FTVMYADDSSAL-IHV-CL---HKGNKDLGDL 133
1AVG   59 N-----KFESNVKFQC-NQVDNKN-GEQYSFKCK-------SSDN-TEFEADFTFISVSYDNFA-LV--CRSITFTSQPKEDRY 124
1QAB   56 RVRLLNNWDVCADMVG-TFTDTEDPAKFKMKYWGVA-----SFLQKGNDD--HWIVDTDYDTYA-VQYSCR-LLNLDGTCADSY 130
1E5P   52 I-----TNNQCS--------KTTVIGYLK-GNGTY-------ETQFEGNNI-FQPLYITSDKIF-FTNKN----XDRAGQETNX 108
ICYN   85 E-----YMEGDL--------EIAPDAKYT-KQGKY--VMTFKFGERVVKLV-PWVLATDYKNYA-INYNC----NTHPDKKAHS 146
1IW2   77 W-----QVRQLY-------------GDTG-VLGRF--LLQARGAR--GAVH-VVVAETDYQSFA-VLYLE----------RAGQ 125

LIPO 155 I--TLYGRTK-ELTSELKENFIRFSKSLGLP-ENHIVFPVPIDQCID 197
1I4U 133 S--FIFSRSA-NLADQYVKKCEAAFKNINVD-TTRFVKTVQGSSCP. 174
1SY0 134 Y--AVLNRNK---DAAAGDKVKSAVSAATLE-FSKFIST-KENNC.. 171
1AVG 125 L-VFERTKS--DTDPDAKE............................ 140
1QAB 131 S--FVFSRDPNGLPPEAQKIVAQRQEELCLA-AQYRLIV-HNGYC.. 171
1E5P 109 I--VVAGKGN-ALTPEENEILVQFAHEKKIPVEN-ILNILATDTCPE 151
ICYN 147 VHAWVLSKNK-VLEGNVKEVVDNVLKTFSHL-ID............. 178
1IW2 126 LSVKLYARSL-PVSDSVLSGFEQRVQEAHLT-ED............. 157
```


70