QuickTree: Neighbor-Joining for Proteins

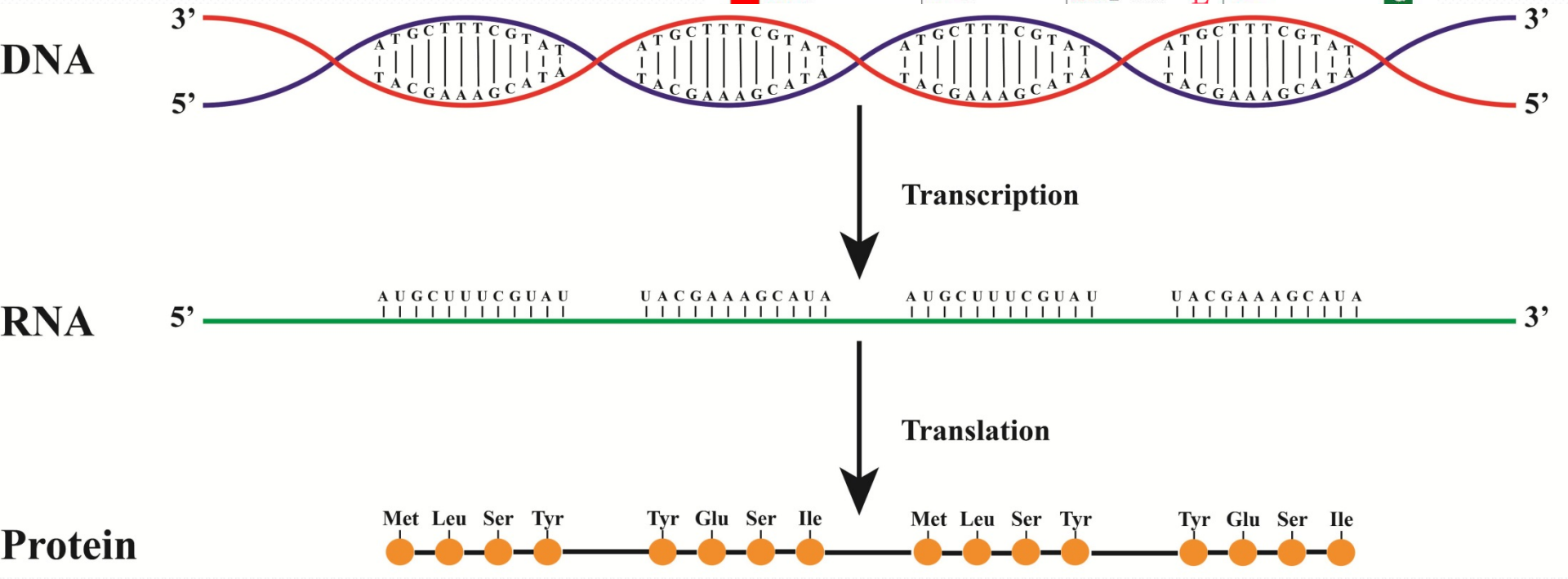
Arcadia Kratkiewicz
CSC 334

Outline

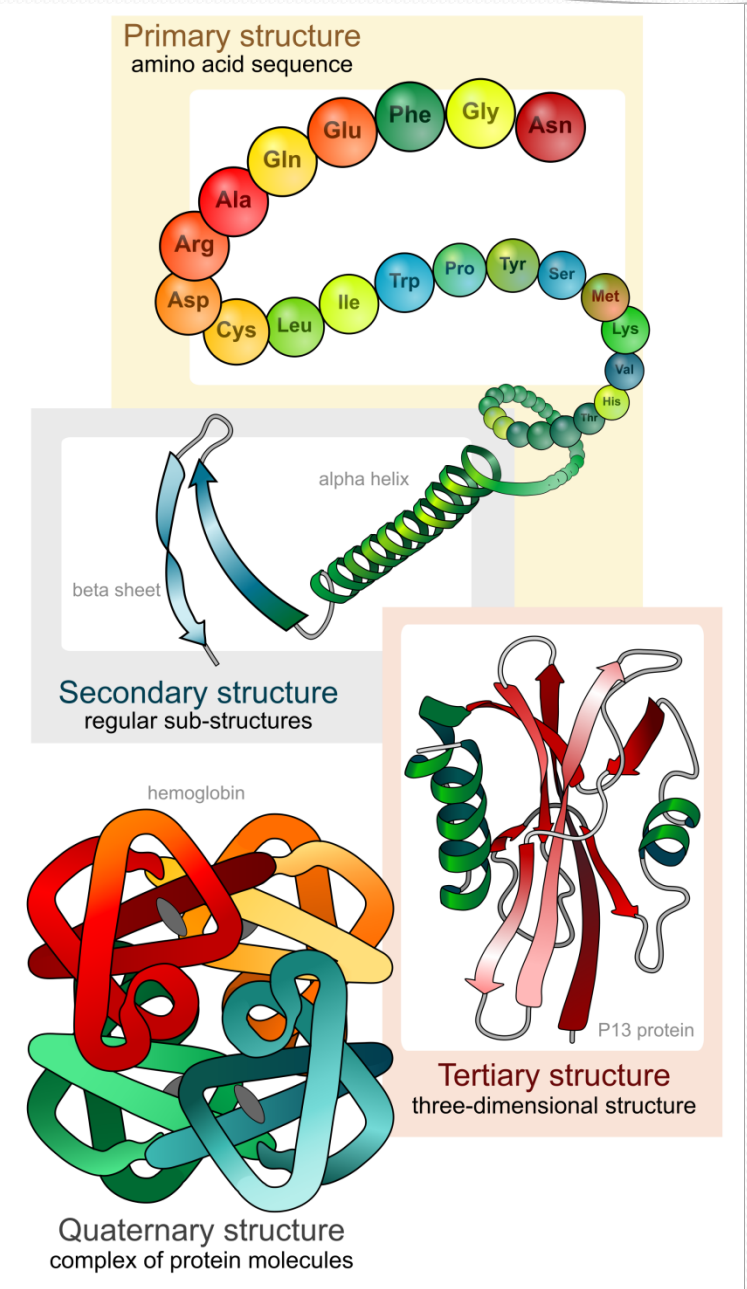
- Background
  - Introduction to proteins
- Motivation
- Methods
  - Creation of Distance Matrix
  - Neighbor-Joining
- Results
- Applications

Proteins

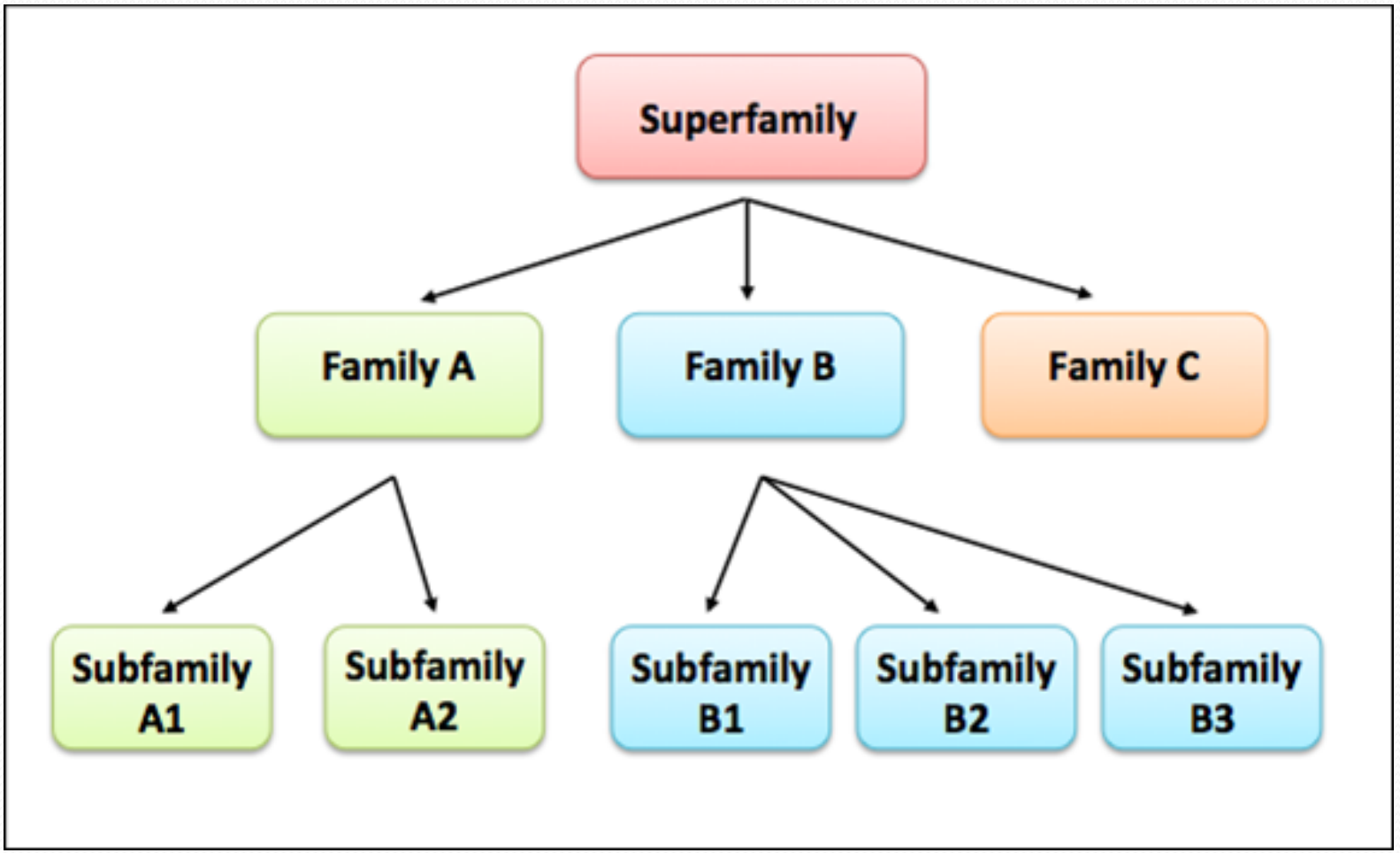
First base	Second base: U	Second base: C	Second base: A	Second base: G	Third base
U	UUU, UUC — Phenylalanine (F)	UCU, UCC, UCA, UCG — Serine (S)	UAU, UAC — Tyrosine (Y)	UGU, UGC — Cysteine (C)	U C A G
	UUA, UUG — Leucine (L)		UAA, UAG — Stop codon	UGA — Stop codon; UGG — Tryptophan (W)	
C	CUU, CUC, CUA, CUG — Leucine (L)	CCU, CCC, CCA, CCG — Proline (P)	CAU, CAC — Histidine (H)	CGU, CGC, CGA, CGG — Arginine (R)	U C A G
			CAA, CAG — Glutamine (Q)		
A	AUU, AUC, AUA — Isoleucine (I)	ACU, ACC, ACA, ACG — Threonine (T)	AAU, AAC — Asparagine (N)	AGU, AGC — Serine (S)	U C A G
	AUG — Methionine (M) start codon		AAA, AAG — Lysine (K)	AGA, AGG — Arginine (R)	
G	GUU, GUC, GUA, GUG — Valine (V)	GCU, GCC, GCA, GCG — Alanine (A)	GAU, GAC — Aspartic acid (D)	GGU, GGC, GGA, GGG — Glycine (G)	U C A G
			GAA, GAG — Glutamic acid (E)		



Proteins


Primary structure
amino acid sequence
Phe
Gly
Asn
Glu
Gln
Ala
Arg
Asp
Cys
Leu
Ile
Trp
Pro
Tyr
Ser
Met
Lys
Val
His
Thr
alpha helix
beta sheet
Secondary structure
regular sub-structures
hemoglobin
P13 protein
Tertiary structure
three-dimensional structure
Quaternary structure
complex of protein molecules

Proteins


Superfamily
Family A
Family B
Family C
Subfamily A1
Subfamily A2
Subfamily B1
Subfamily B2
Subfamily B3

Motivation

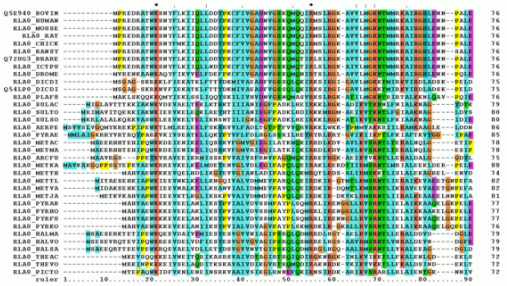
- Faster implementation of neighbor-joining for reconstructing phylogenies of large protein families
- Protein families can contain hundreds to thousands of members
- HIV GP120 over 27,000 sequences

Methods

Input:

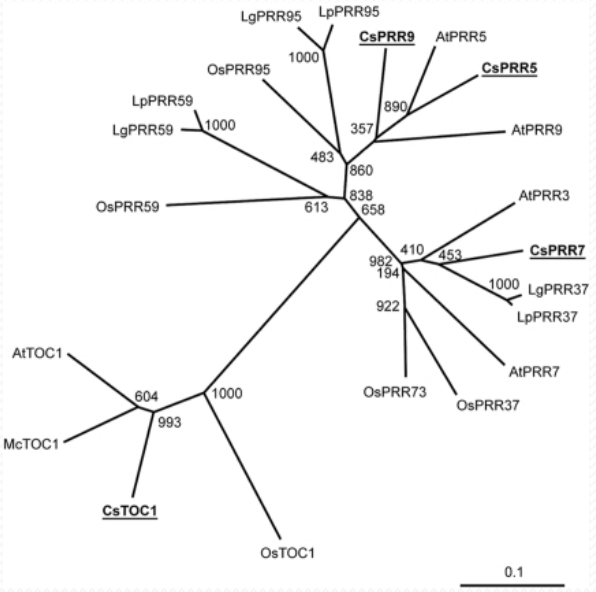
- Distance Matrix
- Multiple Sequence Alignment
  - Create distance matrix with modified CLUSTAL W

	GS	GG	SM	YN	KL	PL	PD	CH	BR	WH
GS	0.000									
GG	0.270	0.000								
SM	0.443	0.382	0.000							
YN	0.515	0.344	0.418	0.000						
KL	0.319	0.272	0.326	0.391	0.000					
PL	0.370	0.470	0.350	0.393	0.394	0.000				
PD	0.245	0.283	0.346	0.357	0.204	0.401	0.000			
CH	0.241	0.237	0.433	0.521	0.392	0.559	0.382	0.000		
BR	0.233	0.283	0.381	0.447	0.267	0.439	0.203	0.279	0.000	
WH	0.390	0.199	0.413	0.483	0.296	0.465	0.344	0.297	0.372	0.000

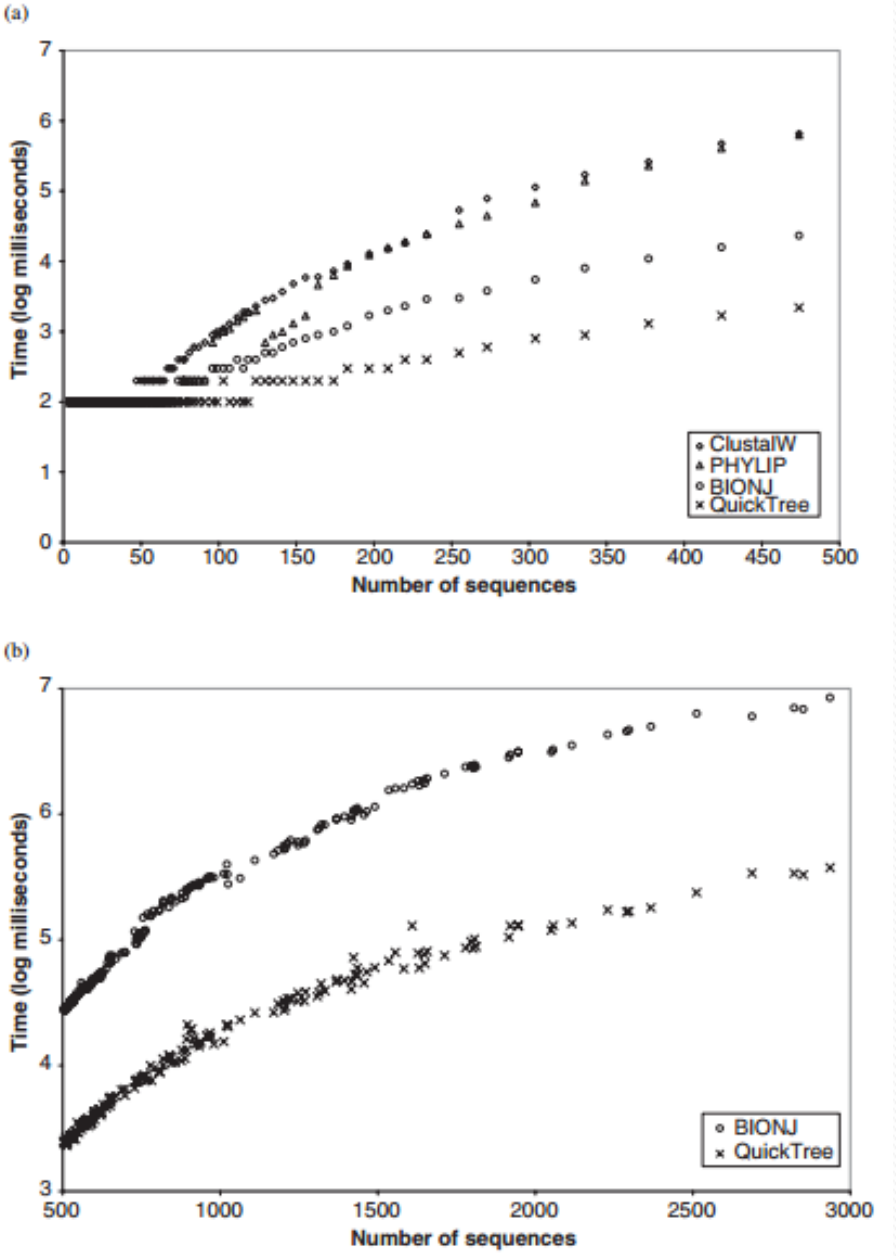


Make Tree:

- Neighbor-Joining
  - (Durbin *et al.*, 1998)

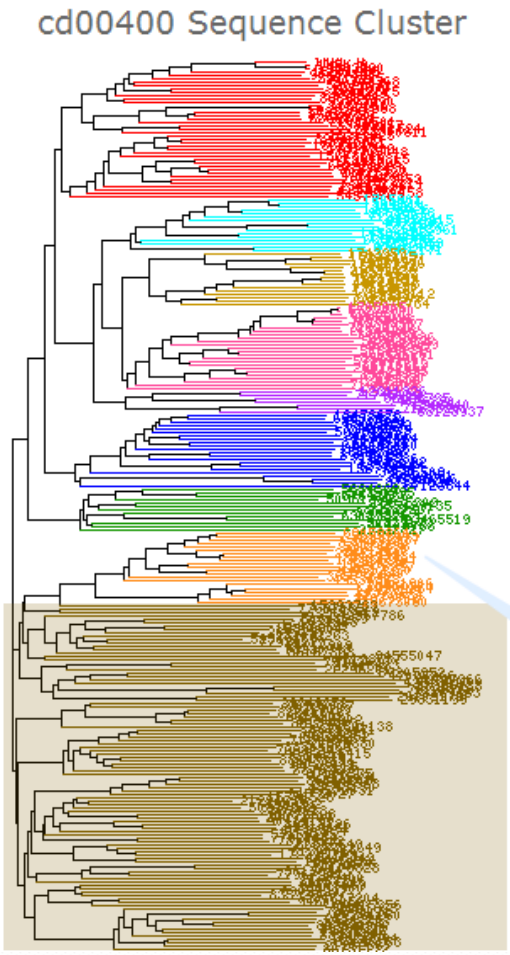


Results


(a)
Time (log milliseconds)
Number of sequences
ClustalW
PHYLIP
BIONJ
QuickTree
(b)
Time (log milliseconds)
Number of sequences
BIONJ
QuickTree

Applications

NCBI curated domain hierarchy for voltage-gated chloride channel



parent node, in this case, encompasses 3 kingdoms of life: archaea, eubacteria, eukaryota. Domain has unique double-barreled architecture and voltage-dependent gating mechanism.

Sub-family Hierarchy

- cd00400 Voltage_gated_ClC
  - cd01031 EriC
  - cd01033 ClC_like
  - cd01034 EriC_like
  - cd01036 ClC_euk
    - cd03683 ClC_1_like
    - cd03684 ClC_3_like
    - cd03685 ClC_6_like
  - cd03682 ClC_sycA_like

child node: domain model specifically found in archaea and eubacteria and associated with extreme acid resistance

putative domain models split out as separate child nodes due to the natural phylogenetic clustering of their member protein sequences, as shown in the sequence cluster tree

intermediate parent: domain model found only in eukaryotes

found in human CLCN1 gene (myotonia), CLCN2 gene (epilepsy), and CLCNKA and CLCNKB (Bartter syndrome)

found in human CLCN3, CLCN4, and CLCN5 (Dent disease, nephrolithiasis, proteinuria, and hypophosphatemic rickets)

found in human CLCN6, CLCN7 (osteoporosis)

found in bacteria; facilitates acid resistance in acidic soil

colors in sequence cluster and subfamily hierarchy correspond to each other

The goal of the NCBI conserved domain curation project is to provide insights into how patterns of residue conservation and divergence in a family relate to functional properties, and to provide useful links to more detailed information that may help to understand those

References

- Howe, K., Bateman, A., and Durbin, R. (2002) QuickTree: building huge Neighbor-Joining trees of protein sequences. Bioinformatics. 18(11): 1546-7.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research. 22(22): 4673-80.
- Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge.
- Saitou, N. and Nei, M. (1987) The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. Mol. Biol. Evol. 4(4): 406-425
- Pfam, EMBL-EBI http://pfam.xfam.org/family/GP120#tabview=tab7
- Protein Classification, NCBI http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml