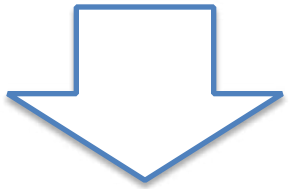
Protein homology and resources for protein families

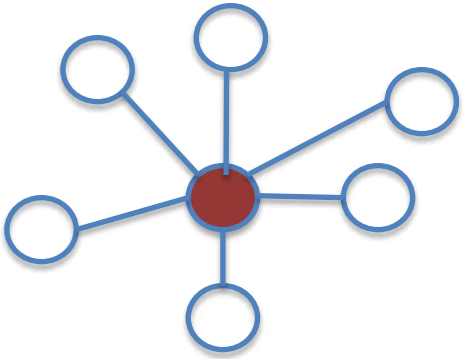
Marco Punta

Centre for Evolution and Cancer
Institute of Cancer Research
London, UK

Protein of interest

Experiments

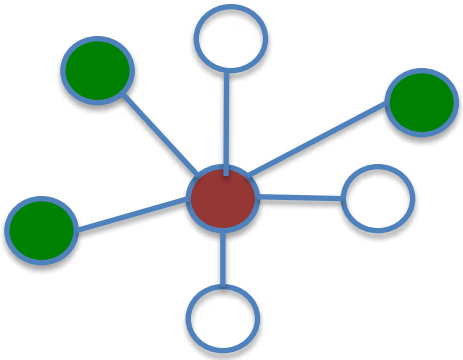




Interaction network

Protein of interest

Experiments



Interaction network

“Known function“

Unknown function

"…uncover homologous relationships between proteins remains the single most powerful tool for function prediction"

Ochoa….Singh PLOS CB (2015)

Outline:

Now:

- Homology (definition, implications, how to detect)
- Protein domains and protein families

To follow:

- Practical: Profile-based sequence searches for protein function annotation

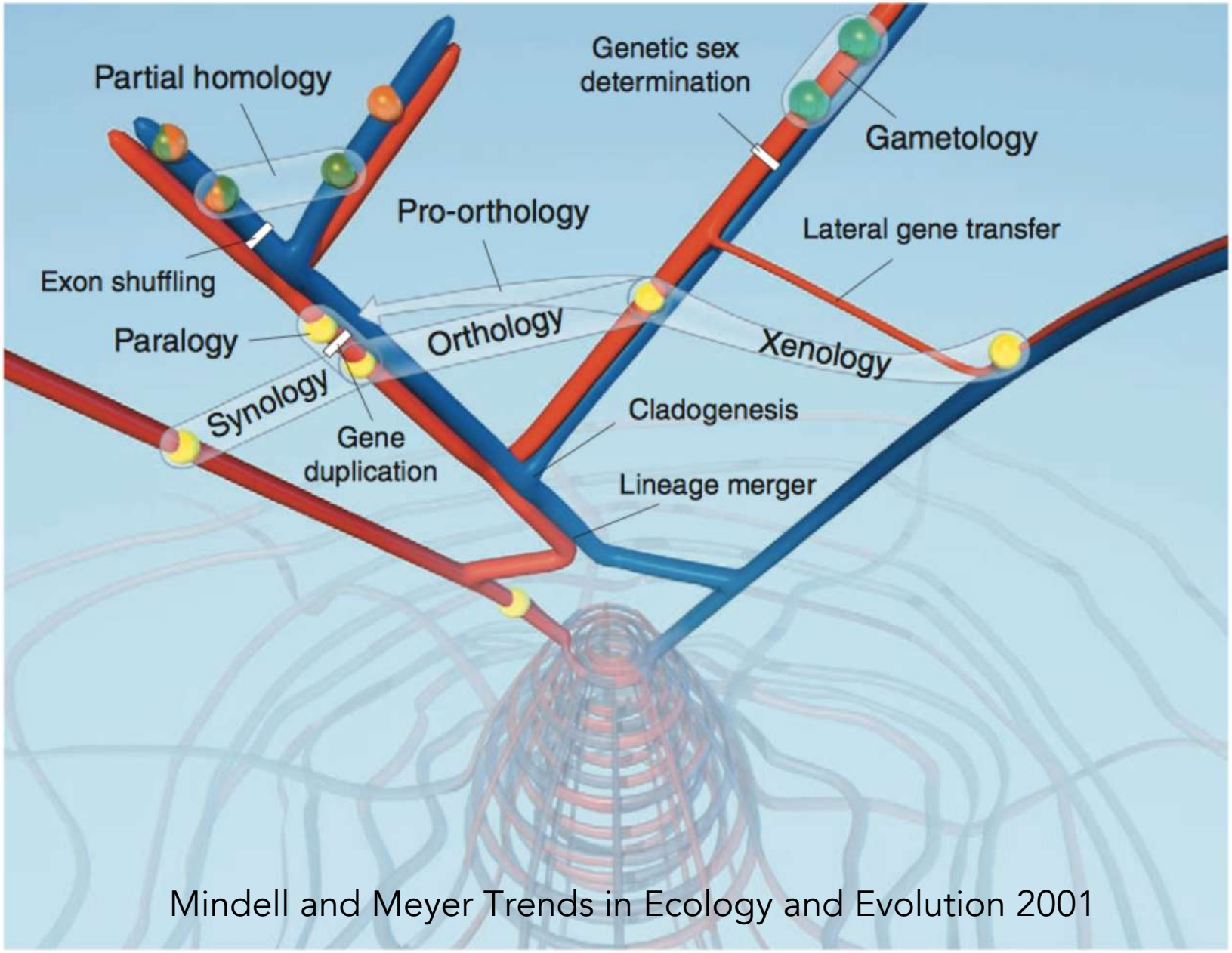
Protein Homology

Definition:

Two proteins are **homologous** if they share a common ancestor, i.e. they are evolutionary related

Origin of homology in proteins

- Speciation (orthology)
- Gene duplication (paralogy)
- Horizontal gene transfer (xenology)
- Whole genome duplication (ohnology)
- etc. etc.



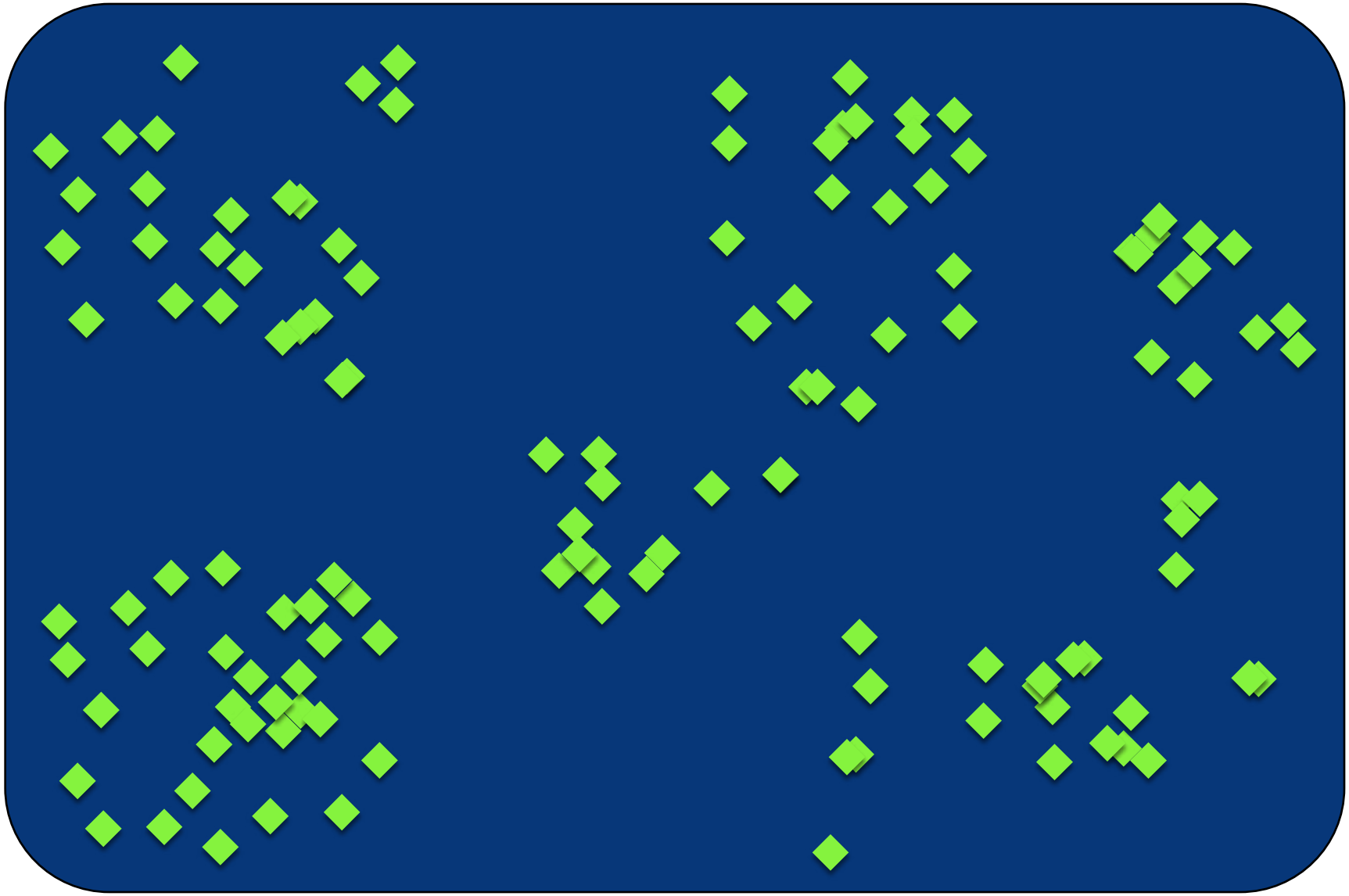
Mindell and Meyer Trends in Ecology and Evolution 2001

Definition:

We call ‘family’ a group of evolutionary related proteins and/or protein regions

The sequence space

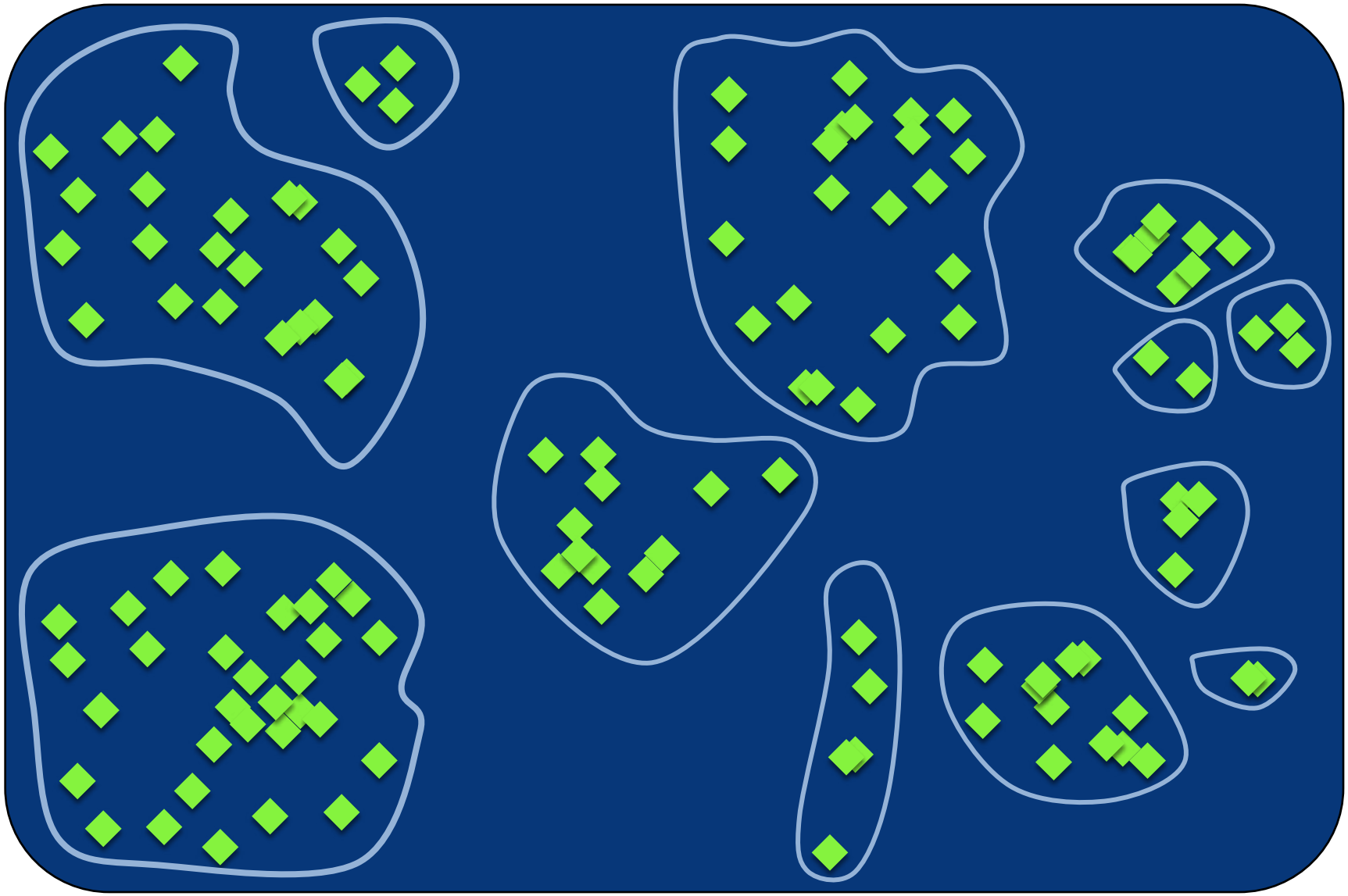
Marco Punta



Observed protein sequences

The sequence space

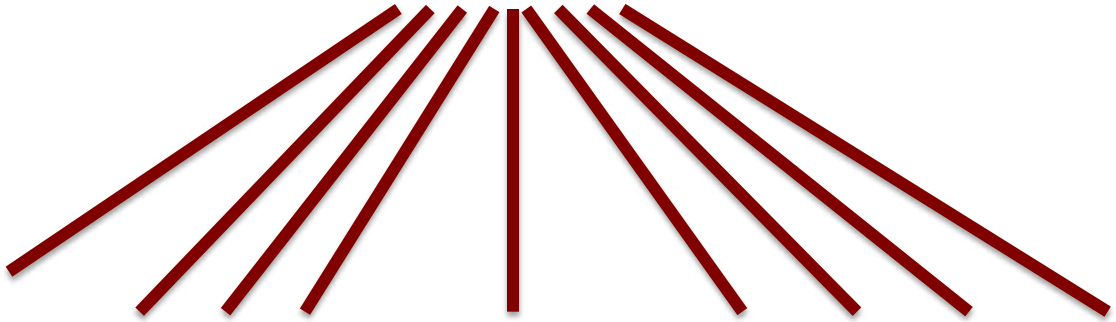
Marco Punta



Homology: why interesting?

Homology: why interesting?

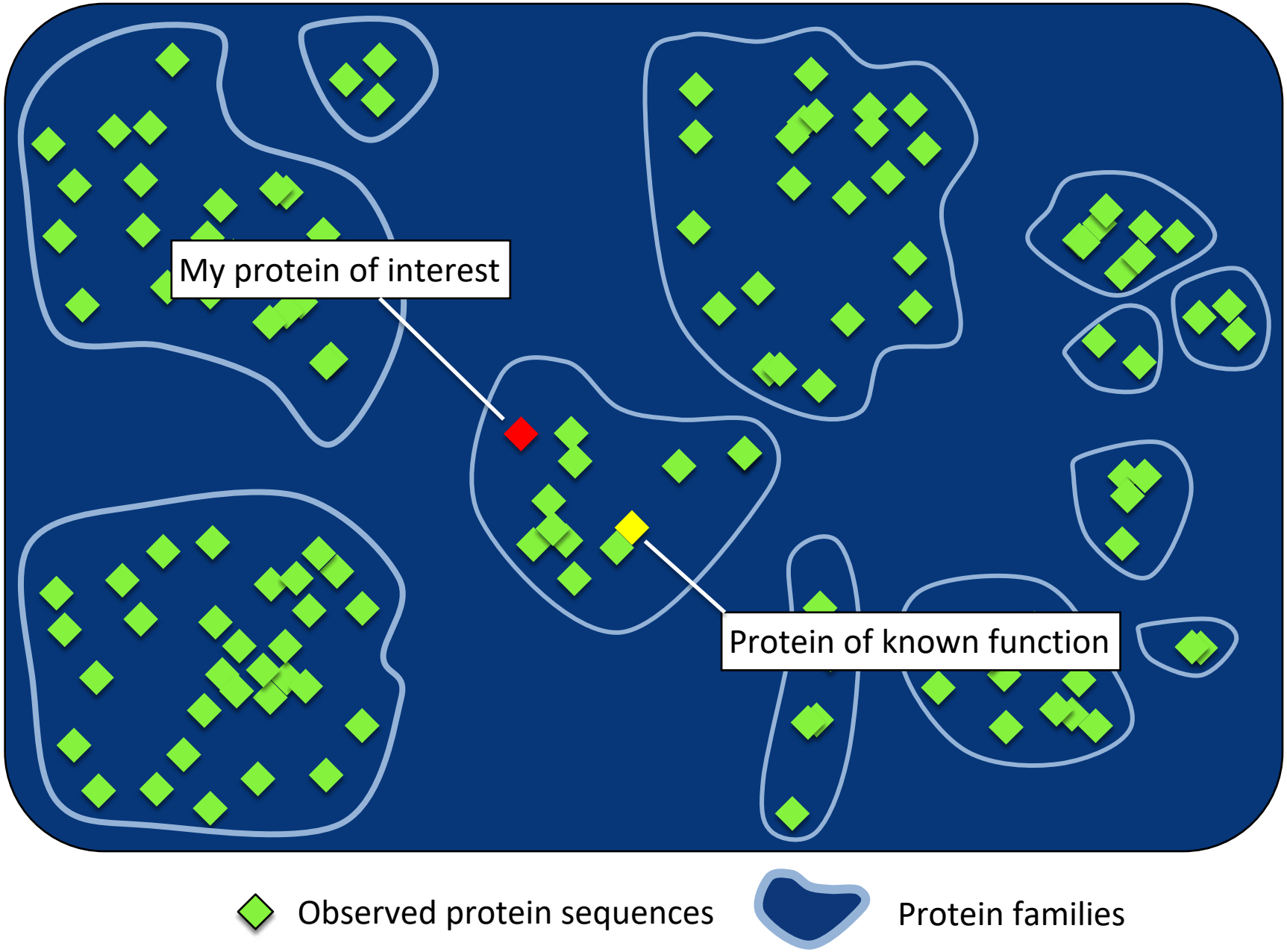
Common ancestor: one or more functions



Present day homologous proteins (family):
Share similar function(s)???

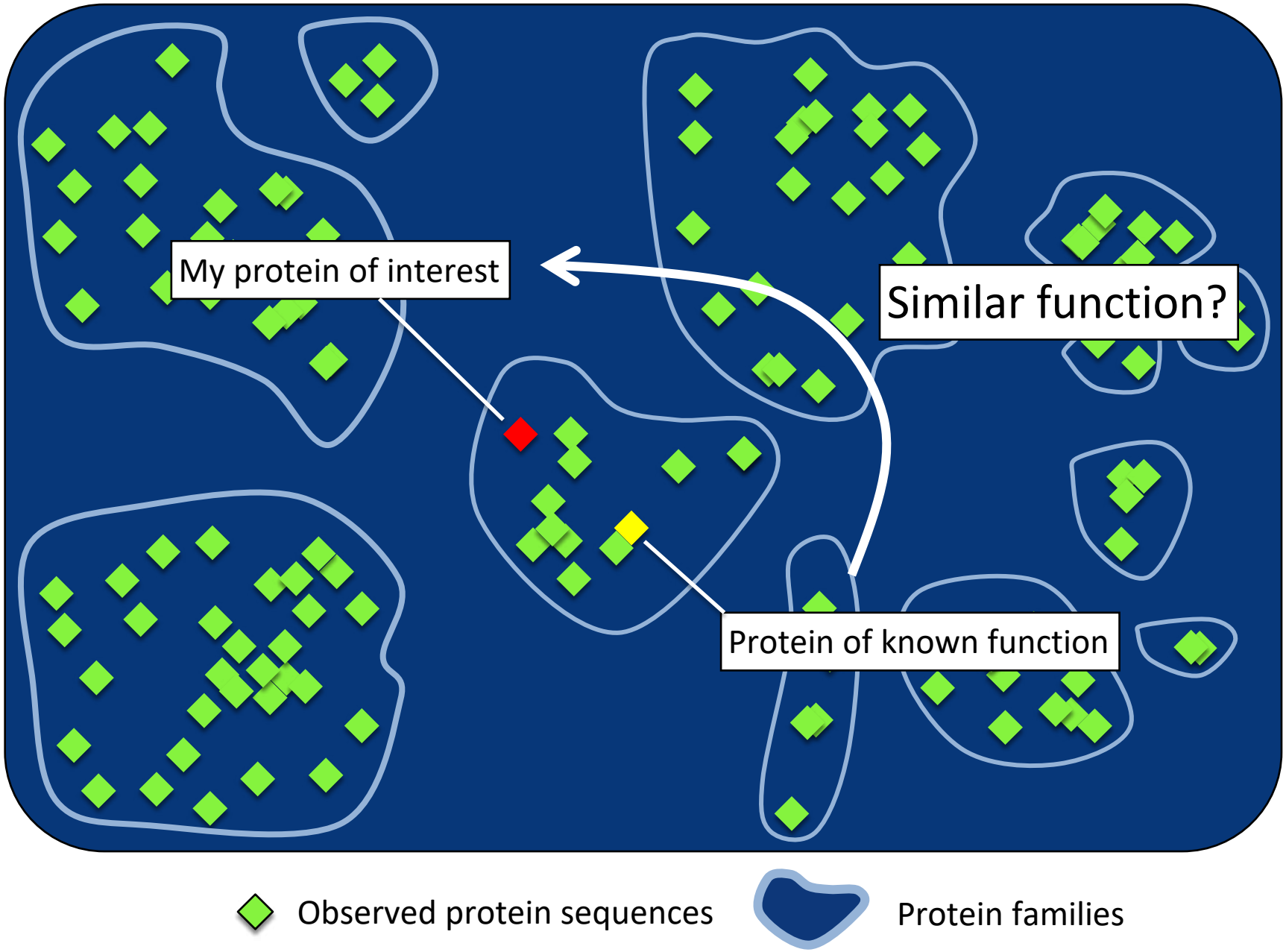
The sequence space

Marco Punta



The sequence space

Marco Punta

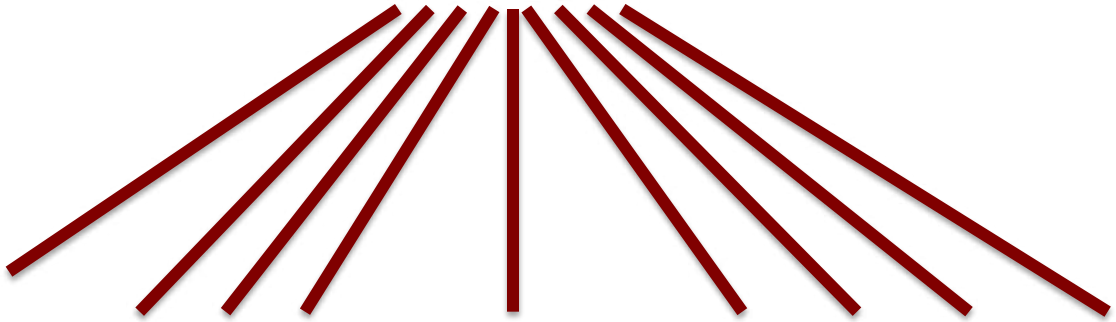


Detecting homology

How do we know two or more proteins are homologous?

Detecting homology

Common ancestor: one sequence, one structure



Present day homologous proteins (family):
Similar sequence, similar structure???

Given two protein sequences how do we know if they share excess sequence similarity? (i.e. that the observed similarity is unlikely to be random)

```
Protein_1:   1   MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE  60
                 MGLSDGEWQLVLNVWGKVEAD  GHGQEVLI LFK HPETL KFDKFK LKSE  MK SE
Protein_2:   1   MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSE  60

Protein_1:  61   DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH  120
                 DLKKHG TVLTALG ILKKKG H AEI PLAQSHATKHKIPVKYLEFISE II VL   H
Protein_2:  61   DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIEVLKKRH  120

Protein_1: 121   PGDFGADAQGAMNKALELFRKDMASNYKELGFQG  154
                   GDFGADAQGAM KALELFR D A  YKELGFQG
Protein_2: 121   SGDFGADAQGAMSKALELFRNDIAAKYKELGFQG  154
```

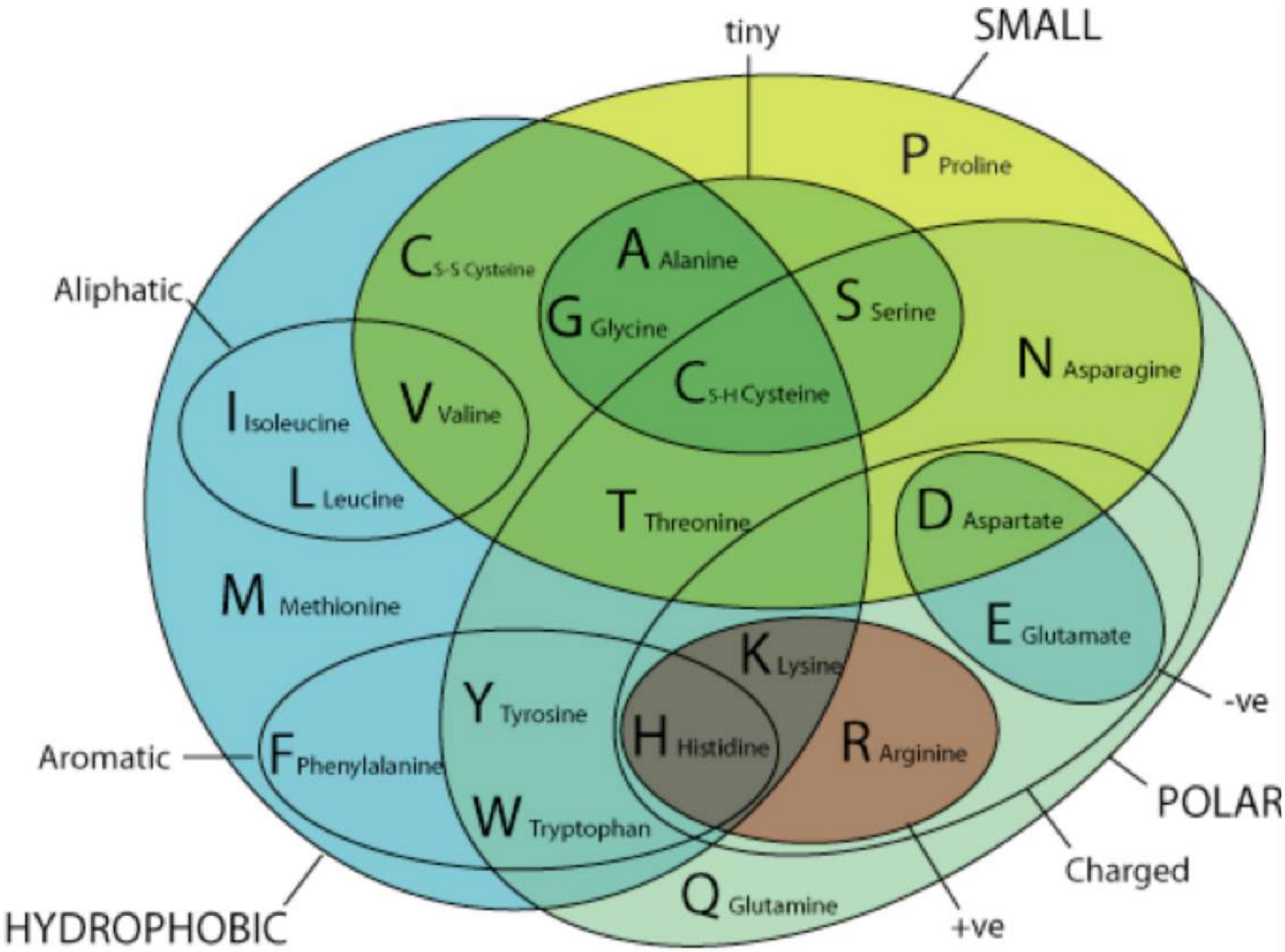
Given two protein sequences how do we know if they share excess sequence similarity? (i.e. that the observed similarity is not random)

```
Protein 1  1 MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
             M LS      V   WGKV A     G E L R F   P T   F  F       D    S
Protein 2  1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSA 54

Protein 1 61 DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKI-PVKY  104
               K H   V  AL                 L   HA K    PV
Protein 2 55 QVKGHSKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNF   99
```

What we need:

- Scoring system

aa physico-chemical properties



http://www.jalview.org/help/html/misc/properties.gif

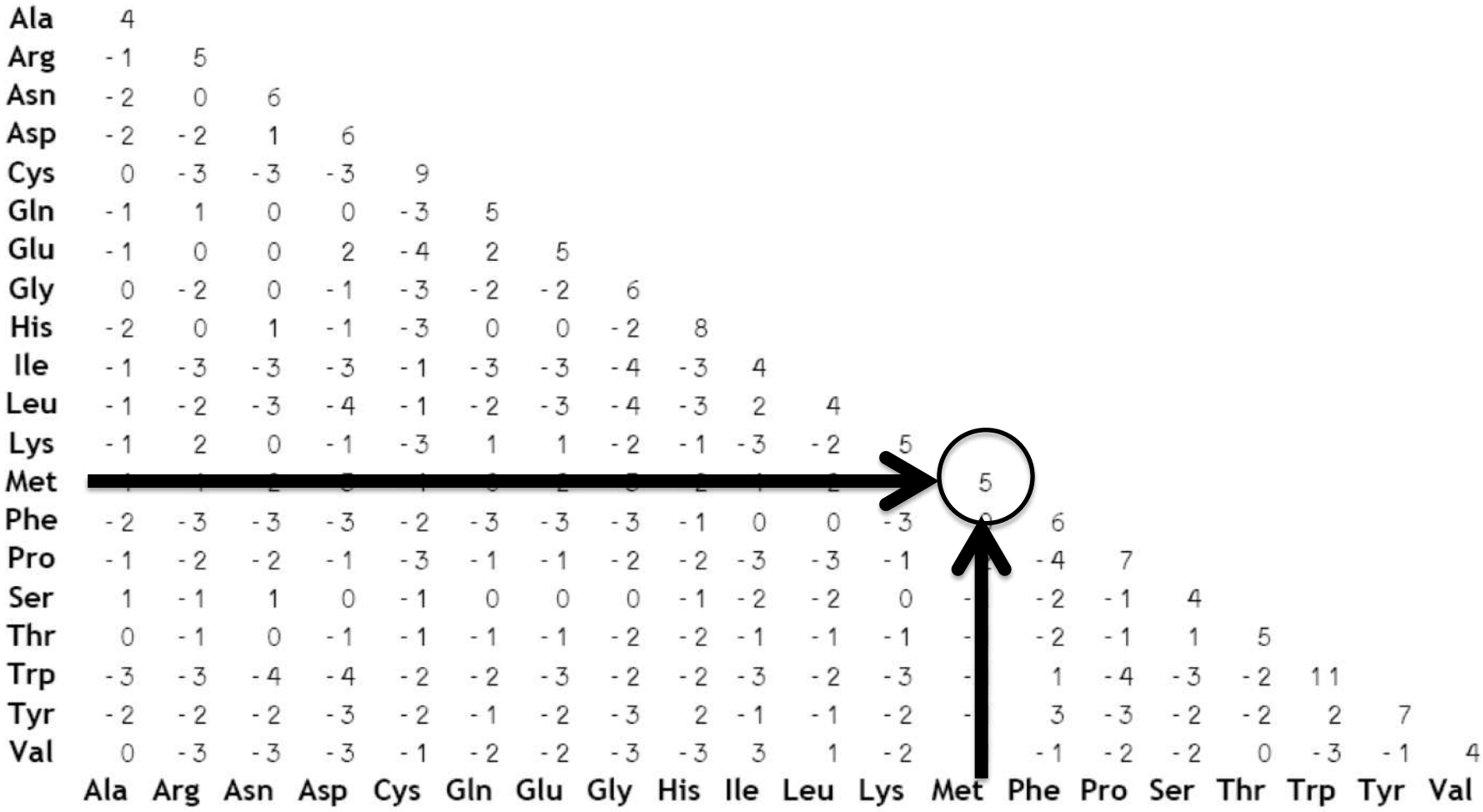
What we need:

- Scoring system => empirically derived substitution matrices (PAMs, BLOSUMs,…)

BLOSUM62 matrix

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Given two protein sequences how do we know if they share excess sequence similarity? (i.e. that the observed similarity is not random)


Ala 4
Arg -1 5
Asn -2 0 6
Asp -2 -2 1 6
Cys 0 -3 -3 -3 9
Gln -1 1 0 0 -3 5
Glu -1 0 0 2 -4 2 5
Gly 0 -2 0 -1 -3 -2 -2 6
His -2 0 1 -1 -3 0 0 -2 8
Ile -1 -3 -3 -3 -1 -3 -3 -4 -3 4
Leu -1 -2 -3 -4 -1 -2 -3 -4 -3 2 4
Lys -1 2 0 -1 -3 1 1 -2 -1 -3 -2 5
Met 5
Phe -2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 6
Pro -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -4 7
Ser 1 -1 1 0 -1 0 0 0 -1 -2 -2 0 -2 -1 4
Thr 0 -1 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -2 -1 1 5
Trp -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 1 -4 -3 -2 11
Tyr -2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 3 -3 -2 -2 2 7
Val 0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 -1 -2 -2 0 -3 -1 4
Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val

Given two protein sequences how do we know if they share excess sequence similarity? (i.e. that the observed similarity is not random)

Protein 1 MGLSDGEW…

Protein 2 MVLSPADK…

5

Given two protein sequences how do we know if they share excess sequence similarity? (i.e. that the observed similarity is not random)

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Given two protein sequences how do we know if they share excess sequence similarity? (i.e. that the observed similarity is not random)

Protein 1 MGLSDGEW…

Protein 2 MVLSPADK…

5-3

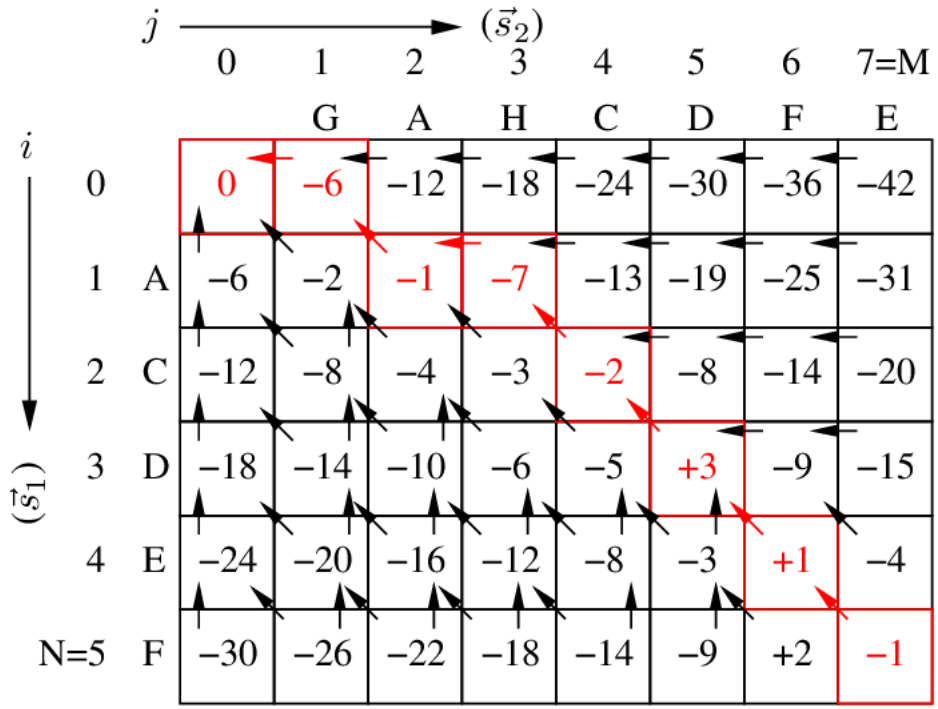
Given two protein sequences how do we know if they share excess sequence similarity? (i.e. that the observed similarity is not random)

What we need:

- Scoring system => empirically derived substitution matrices (PAMs, BLOSUMs,…)

- Efficient way to find highest scoring alignments => dynamic programming (Needleman-Wunsch, Smith-Waterman,…)

Dynamic programming



Optimum alignment score: −1

G	A	H	C	D	F	E
−	A	−	C	D	E	F
−6	+5	−6	+5	+5	−2	−2

Note: not same scoring matrix as before

What we need:

- Scoring system => empirically derived substitution matrices (PAMs, BLOSUMs,…)
- Efficient way to find highest scoring alignments => dynamic programming (Needleman-Wunsch, Smith-Waterman,…)
- Way to decide whether top score is high enough to infer homology (significance) => E-value, …

Significance

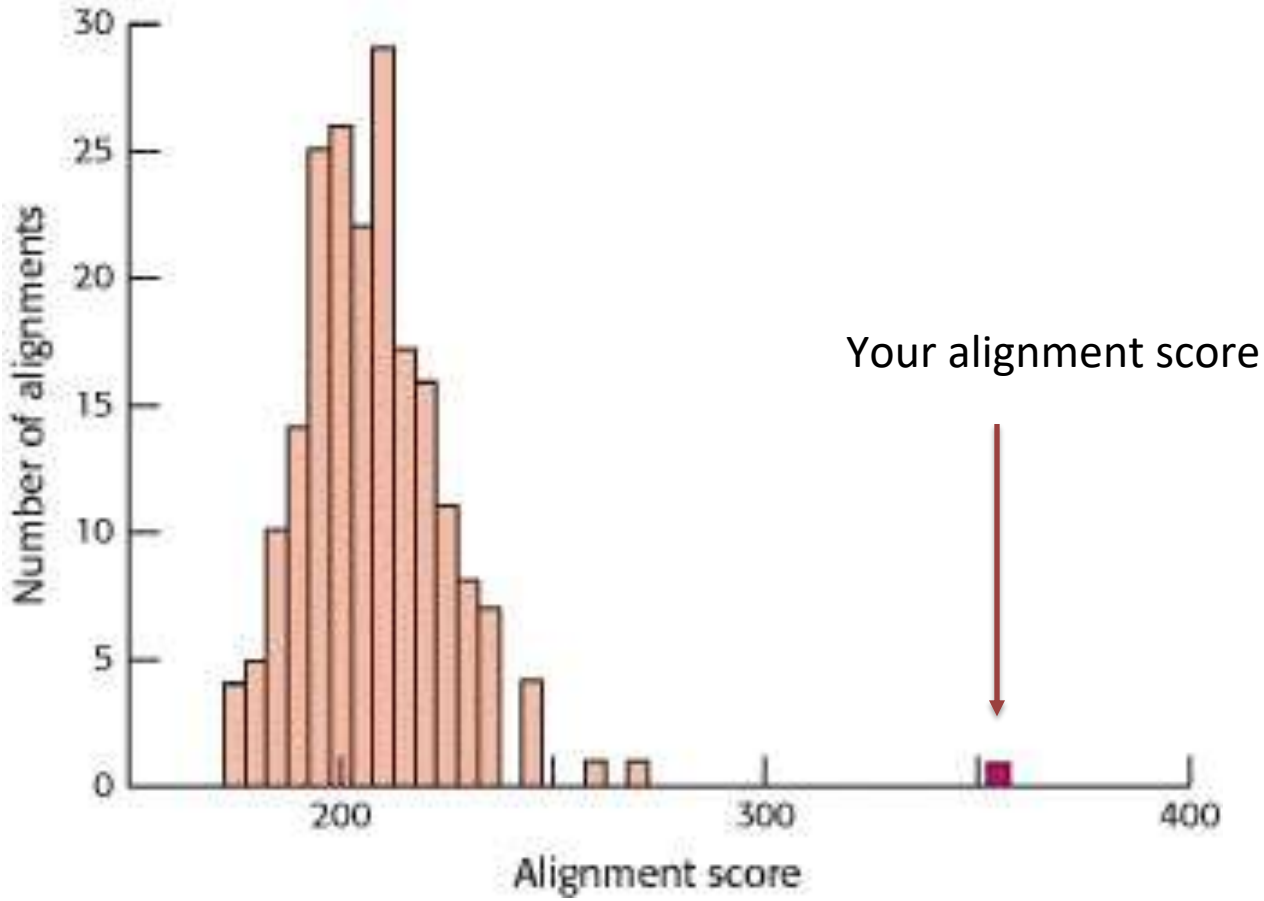


Image credits: http://www.ncbi.nlm.nih.gov/books/NBK22456/figure/A937/?report=objectonly

Significance

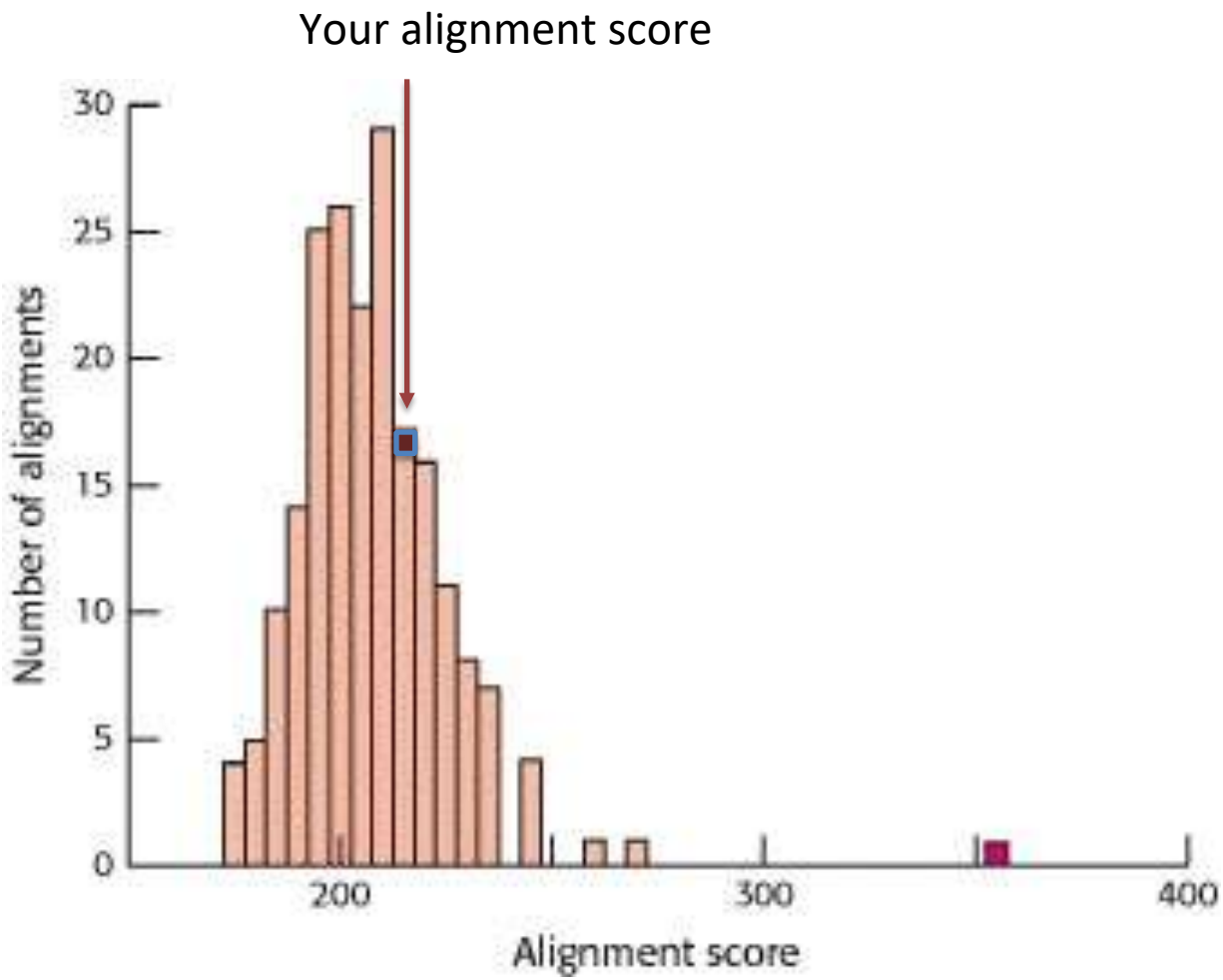


Image credits: http://www.ncbi.nlm.nih.gov/books/NBK22456/figure/A937/?report=objectonly

Statistical significance: E-values

My alignment score = S_0

The E-value of the alignment tells me how many alignments between unrelated sequences I can expect to find that have $S \geq S_0$ when searching the same database with my query sequence

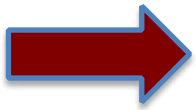
If E-value = 10, I expect to find 10 such alignments

If E-value 10^{-5}, it is estimated that I would have to run 100,000 such searches before I find one such score between unrelated sequences

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ

Excess sequence similarity

?



Homology

Excess sequence similarity ? Homology

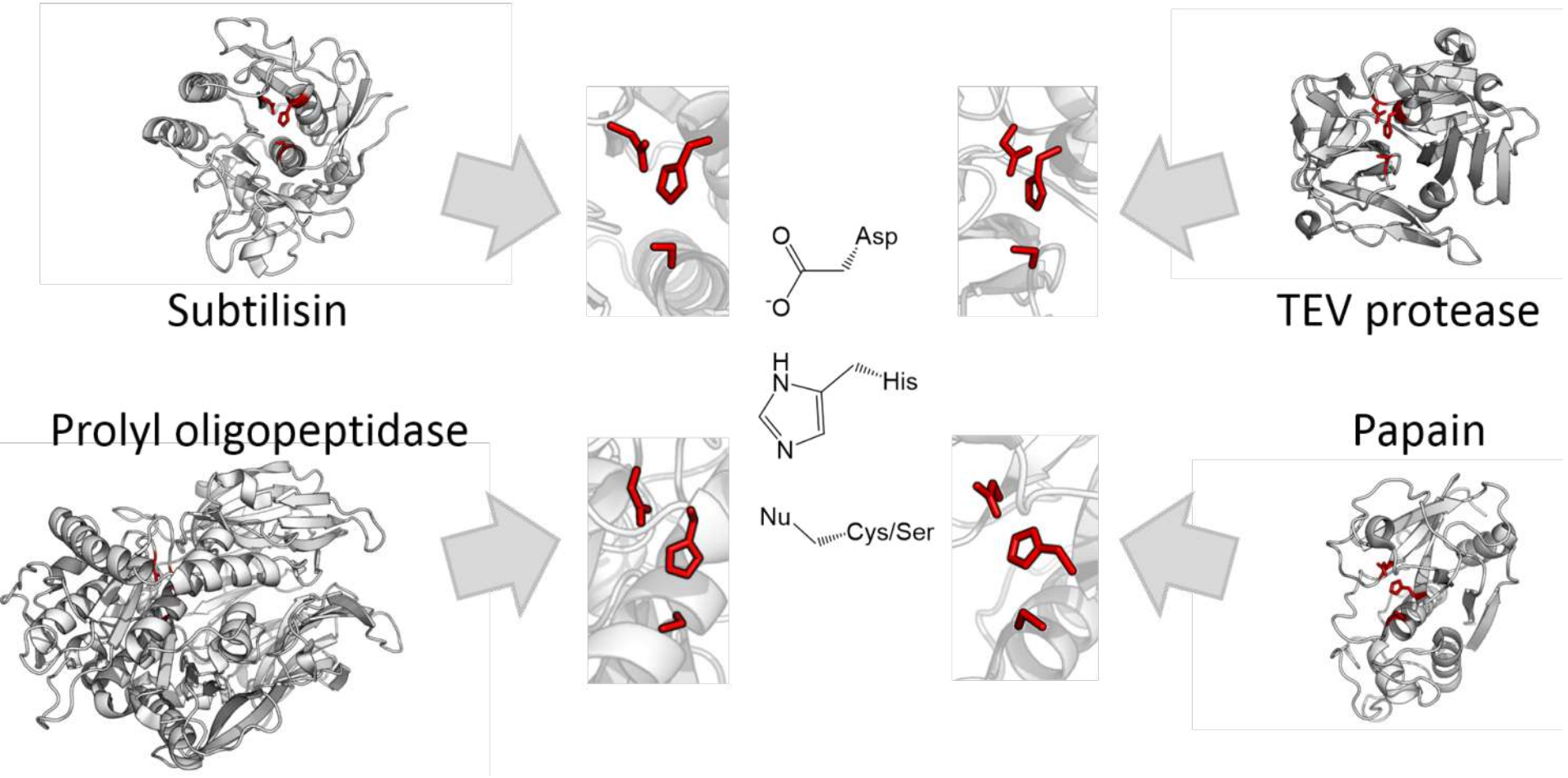
There are alternative possible explanations for excess sequence similarity (i.e. analogy due to functional or structural convergence)

In general, we are guided by observation (what is known today) and "the principles of parsimony (Occams' razor) and likelihood"*

*Computational Structural Biology: Methods and Applications Torsten Schwede (2008)

Convergent evolution of function

Marco Punta

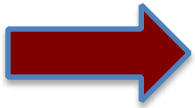


"Triad Convergence" by Thomas Shafee - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Triad_Convergence.png#/media/File:Triad_Convergence.png

See the Book: Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Koonin and Galperin (2003)

"[…]we are justified to conclude that whenever statistically significant sequence or structural similarity between proteins or protein domains is observed, this is an indication of their divergent evolution from a common ancestor or, in other words, evidence of homology."

Koonin and Galperin (2003)

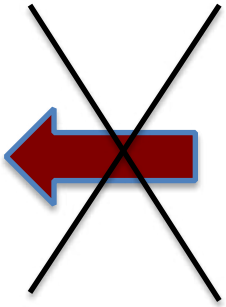
Excess sequence similarity  Homology

Cases where significance may misguide: low complexity regions (homopolymers) (filter out, penalise), coiled-coils

(see e.g. Mistry et al NAR 2013)

The reverse statement is not true (plenty of examples)

Excess sequence similarity



Homology

Suggested reading

Curr Protoc Bioinformatics. Author manuscript; available in PMC 2014 Jun 1.
Published in final edited form as:
Curr Protoc Bioinformatics. 2013 Jun; 0 3: 10.1002/0471250953.bi0301s42.
doi: 10.1002/0471250953.bi0301s42

PMCID: PMC3820096
NIHMSID: NIHMS519883

An Introduction to Sequence Similarity ("Homology") Searching

William R. Pearson[1]

Author information ▶ Copyright and License information ▶

Abstract Go to:

Sequence similarity searching, typically with BLAST (units 3.3, 3.4), is the most widely used, and most reliable, strategy for characterizing newly determined sequences. Sequence similarity searches can identify "homologous" proteins or genes by detecting excess similarity – statistically significant similarity that reflects common ancestry. This unit provides an overview of the inference of homology from significant similarity, and introduces other units in this chapter that provide more details on effective strategies for identifying homologs.

Keywords: sequence similarity, homology, orthlogy, paralogy, sequence alignment, multiple alignment, sequence evolution
Subject: Bioinformatics, Bioinformatics Fundamentals, Finding Similarities and Inferring Homologies

"[…]we are justified to conclude that whenever statistically significant sequence or structural similarity between proteins or protein domains is observed, this is an indication of their divergent evolution from a common ancestor or, in other words, evidence of homology."

Koonin and Galperin (2003)

Structure more conserved than sequence

Proteins. 2009 Nov 15;77(3):499-508. doi: 10.1002/prot.22458.

Structure is three to ten times more conserved than sequence--a study of structural response in protein cores.

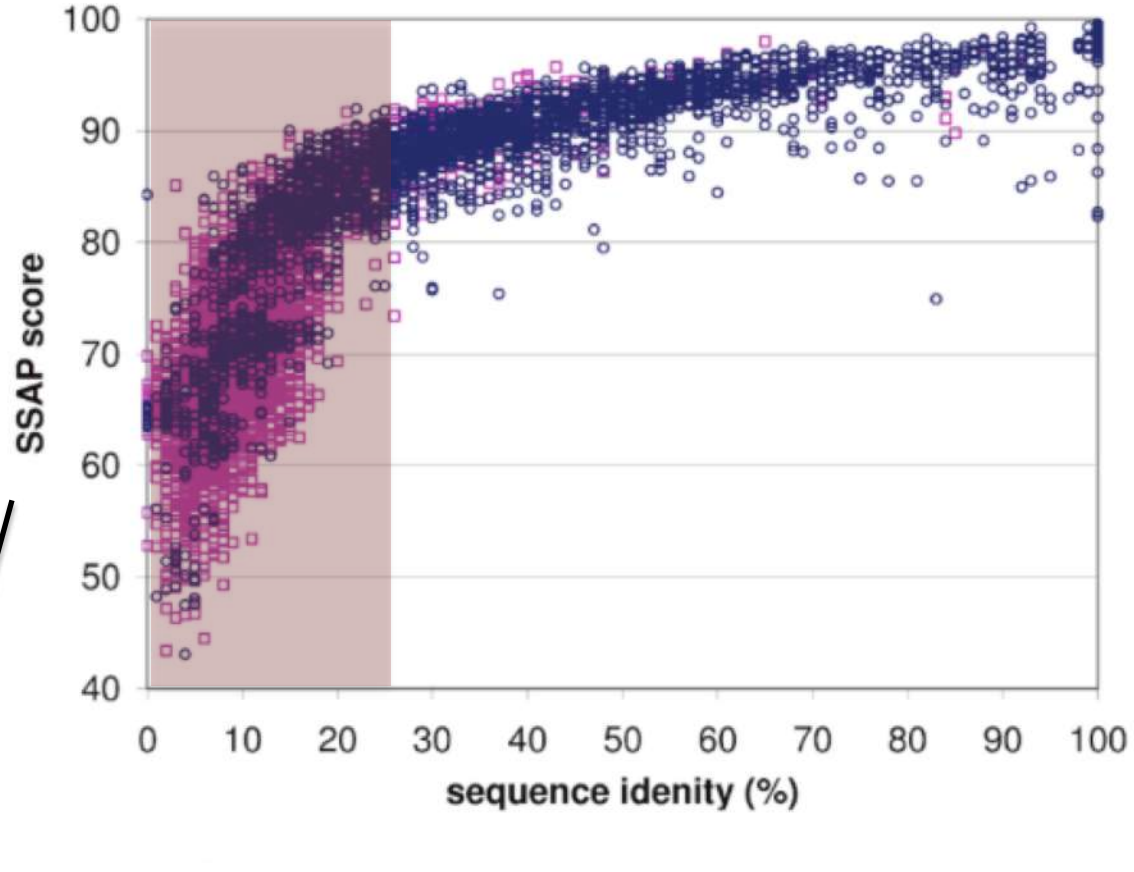
Illergård K[1], Ardell DH, Elofsson A.

Author information

Abstract

Protein structures change during evolution in response to mutations. Here, we analyze the mapping between sequence and structure in a set of structurally aligned protein domains. To avoid artifacts, we restricted our attention only to the core components of these structures. We found that on average, using different measures of structural change, protein cores evolve linearly with evolutionary distance (amino acid substitutions per site). This is true irrespective of which measure of structural change we used, whether RMSD or discrete structural descriptors for secondary structure, accessibility, or contacts. This linear response allows us to quantify the claim that structure is more conserved than sequence. Using structural alphabets of similar cardinality to the sequence alphabet, structural cores evolve three to ten times slower than sequences. Although we observed an average linear response, we found a wide variance. Different domain families varied fivefold in structural response to evolution. An attempt to categorically analyze this variance among subgroups by structural and functional category revealed only one statistically significant trend. This trend can be explained by the fact that beta-sheets change faster than alpha-helices, most likely due to that they are shorter and that change occurs at the ends of the secondary structure elements.

Structure vs Sequence similarity of homologous proteins



Structural similarity

Modified from Computational Structural Biology: Methods and Applications Torsten Schwede (2008)

Issues in structural comparison

- Statistical framework for structural alignments less solid than the one for sequence alignments:
- How should we define the alignment score?
- How do we find the optimal structural alignment?
- How do we define significance?
- Different methods may give different answers…

see discussion in Colliers et al. Bioinformatics (2017)

In practice

- Structural similarity important for suggesting homology in protein regions sharing insignificant sequence similarity, however:

- we generally have to look for additional signs of homology for validation (e.g. conservation of structural and/or functional residues, conserved domain architectures)

- Many proteins with no known function have no known structure

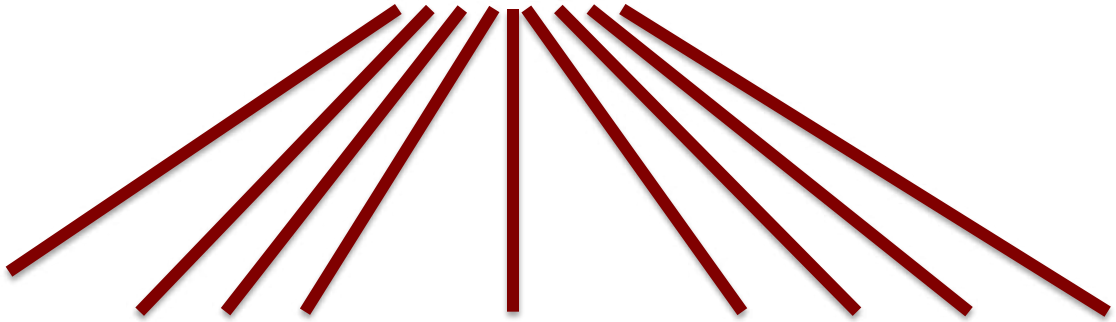
Here, we will focus on homology detection via excess sequence similarity

Homology modeling

Same structural core ⟸ Homology

Homology: why interesting?

Common ancestor: one or more functions



Present day homologous proteins (family):
Share similar function(s)???

How to compare function?

How to compare function?

Marco Punta

CbiF

Catalytic activity

Methyltransferase

precorrin-4 methyltransferase activity

CbiA

Catalytic activity

Ligase activity

cobyrinic acid a,c-diamide synthase activity

CbiJ

Catalytic activity

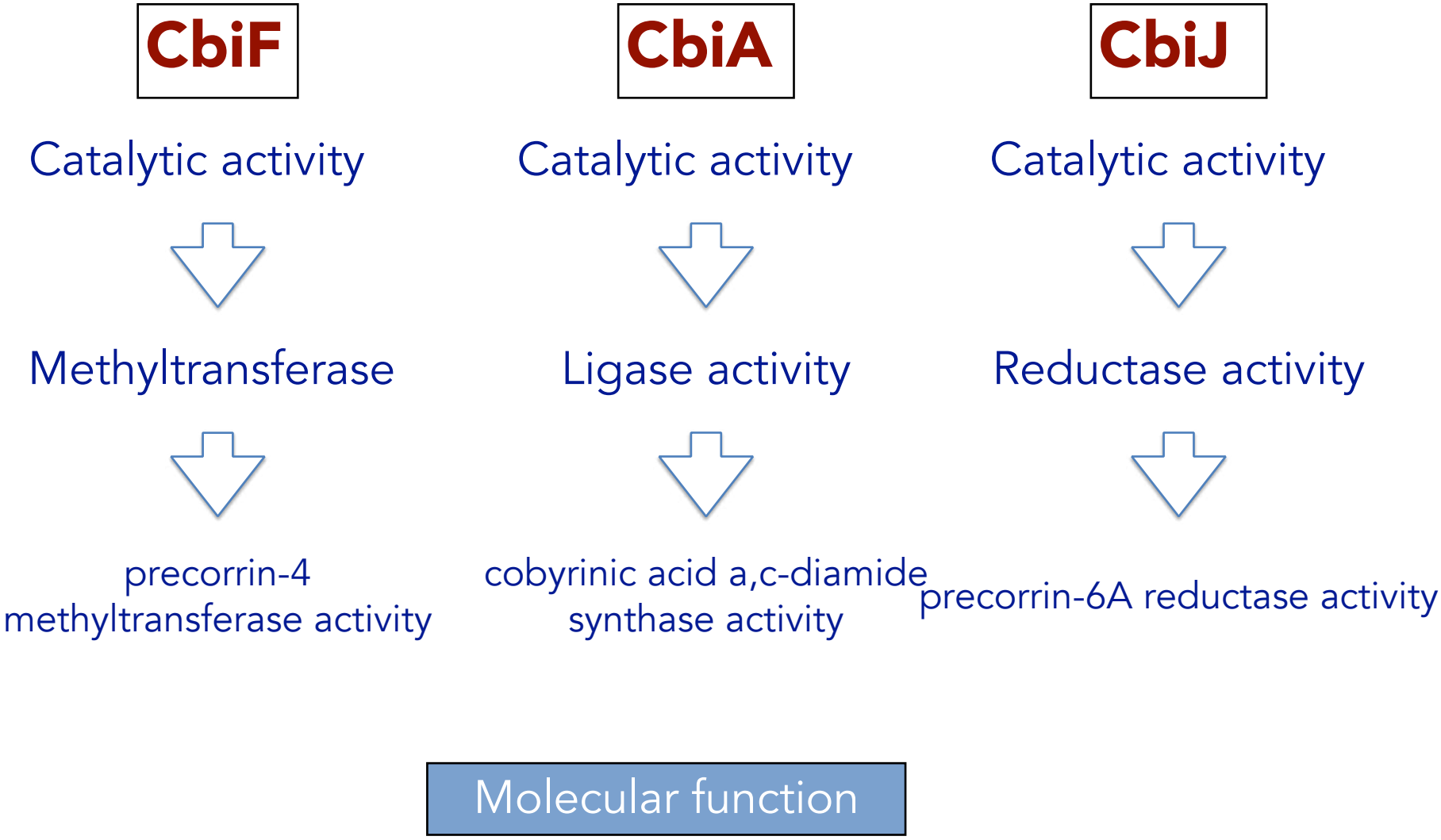
Reductase activity

precorrin-6A reductase activity

Moore et al. PNAS 110 (2013)

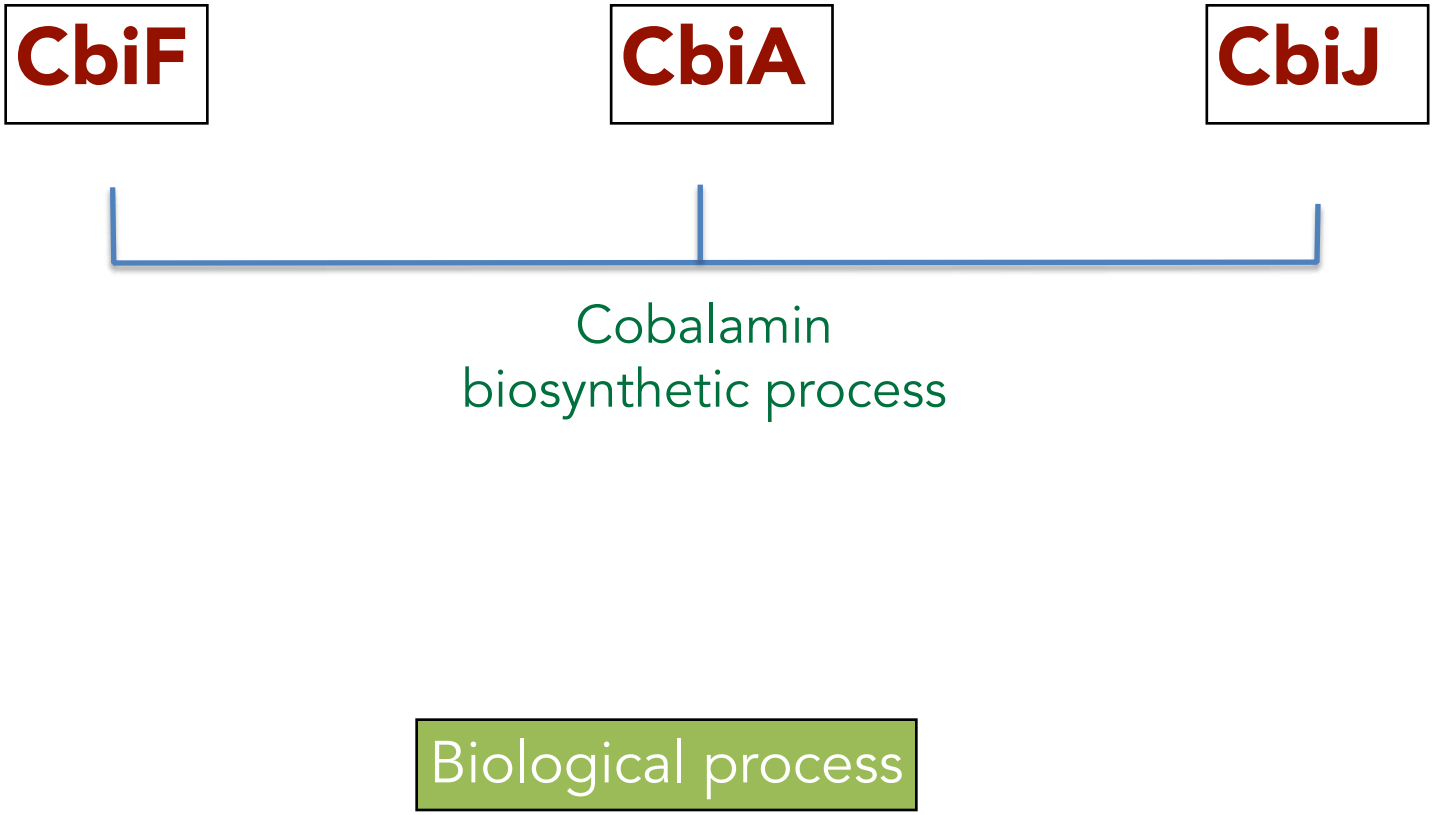
How to compare function?

Marco Punta



Moore et al. PNAS 110 (2013)

How to compare function?

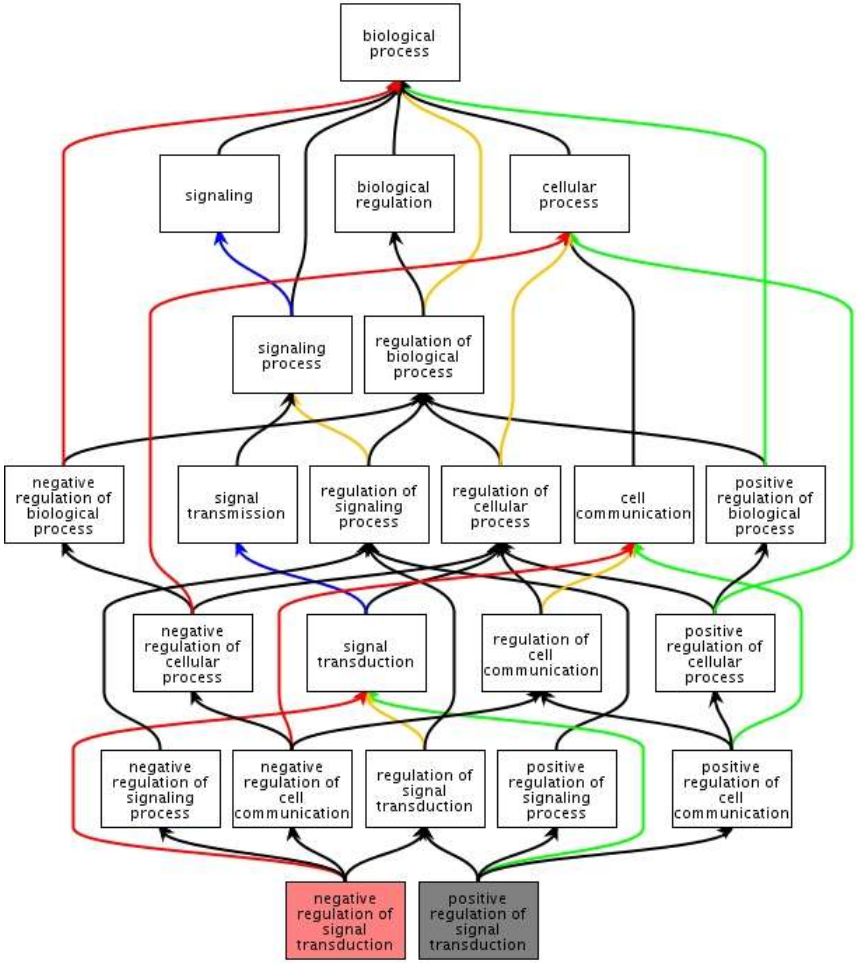


Moore et al. PNAS 110 (2013)

The Gene Ontology (GO)

- A way to capture biological knowledge in a written and computable form

- A set of concepts and their relationships to each other



www.ebi.ac.uk/QuickGO

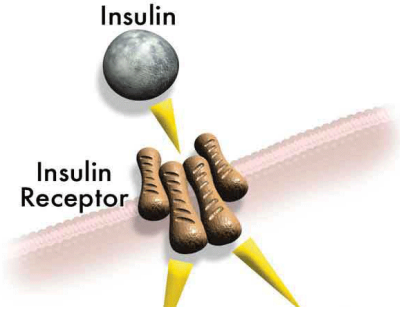
Slide courtesy of Alex Mitchell (EMBL-EBI)

GO: 3 ontologies in 1

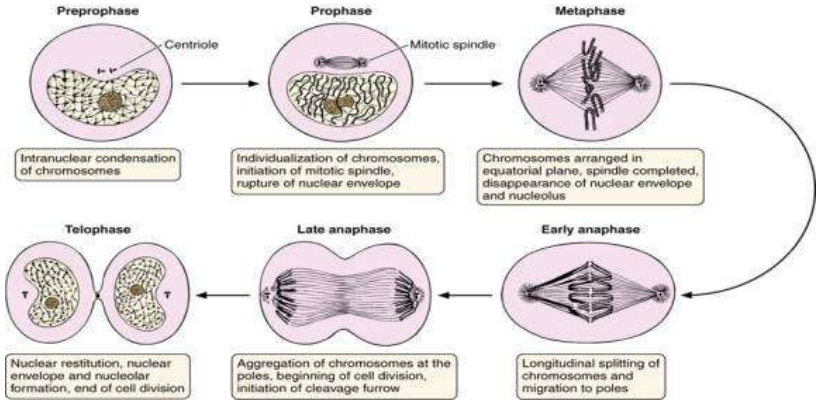
Marco Punta

1. Molecular Function

An elemental activity or task or job



- protein kinase activity
- insulin receptor activity



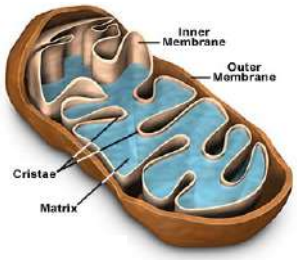
2. Biological Process

A commonly recognised series of events

- cell division

3. Cellular Component

Where a gene product is located



- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

Slide courtesy of Alex Mitchell (EMBL-EBI)

Top-level EC numbers[5]

Group	Reaction catalyzed	Typical reaction	Enzyme example(s) with trivial name
EC 1 ***Oxidoreductases***	To catalyze oxidation/reduction reactions; transfer of H and O atoms or electrons from one substance to another	AH + B → A + BH (**reduced**) A + O → AO (**oxidized**)	Dehydrogenase, oxidase
EC 2 ***Transferases***	Transfer of a functional group from one substance to another. The group may be methyl-, acyl-, amino- or phosphate group	AB + C → A + BC	Transaminase, kinase
EC 3 ***Hydrolases***	Formation of two products from a substrate by hydrolysis	AB + H_2O → AOH + BH	Lipase, amylase, peptidase
EC 4 ***Lyases***	Non-hydrolytic addition or removal of groups from substrates. C-C, C-N, C-O or C-S bonds may be cleaved	RCOCOOH → RCOH + CO_2 or [X-A-B-Y] → [A=B + X-Y]	Decarboxylase
EC 5 ***Isomerases***	Intramolecule rearrangement, i.e. isomerization changes within a single molecule	ABC → BCA	Isomerase, mutase
EC 6 ***Ligases***	Join together two molecules by synthesis of new C-O, C-S, C-N or C-C bonds with simultaneous breakdown of ATP	X + Y+ ATP → XY + ADP + Pi	Synthetase

https://en.wikipedia.org/wiki/Enzyme_Commission_number

Human myoglobin vs human hemoglobin

sp|P02008|HBAZ_HUMAN Hemoglobin subunit zeta OS=Homo sapiens OX=9606 GN=HBZ PE=1 SV=2

Sequence ID: Query_24201 Length: 142 Number of Matches: 1

Range 1: 1 to 142 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
71.6 bits(174)	1e-21	Compositional matrix adjust.	42/149(28%)	72/149(48%)	8/149(5%)

```
Query  1    MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE  60
            M L+  E  +++++W K+       G E L RLF  HP+T   F  F      D    S
Sbjct  1    MSLTKTERTIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYFPHF------DLHPGSA  54

Query  61   DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKI-PVKYLEFISECIIQVLQSK  119
             L+ HG+ V+ A+G  +K        +  L++ HA   ++ PV + + +S C++  L ++
Sbjct  55   QLRAHGSKVVAAVGDAVKSIDDIGGALSKLSELHAYILRVDPVNF-KLLSHCLLVTLAAR  113

Query  120  HPGDFGADAQGAMNKALELFRKDMASNYK  148
             P DF A+A  A +K L +    +   Y+
Sbjct  114  FPADFTAEAHAAWDKFLSVVSSVLTEKYR  142
```

Database	Gene Product ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
Process													
UniProtKB	P02144	MB		GO:0001666	response to hypoxia	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0006810	transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0813	9606	20140913	UniProt	
UniProtKB	P02144	MB		GO:0007507	heart development	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0009725	response to hormone	P	IEA	Ensembl Compara	Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0015671	oxygen transport	P	IEA	InterPro2GO	InterPro:IPR002335|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02144	MB		GO:0015671	oxygen transport	P	IEA	Ensembl Compara	Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0015671	oxygen transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02144	MB		GO:0031444	slow-twitch skeletal muscle fiber contraction	P	IEA	Ensembl Compara	Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0042542	response to hydrogen peroxide	P	IEA	Ensembl Compara	Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0043353	enucleate erythrocyte differentiation	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0050873	brown fat cell differentiation	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
Function													
UniProtKB	P02144	MB		GO:0005344	oxygen transporter activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02144	MB		GO:0005506	iron ion binding	F	IEA	InterPro2GO	InterPro:IPR000971|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02144	MB		GO:0019825	oxygen binding	F	IEA	InterPro2GO	InterPro:IPR002335|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02144	MB		GO:0019825	oxygen binding	F	IEA	Ensembl Compara	Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0020037	heme binding	F	IEA	InterPro2GO	InterPro:IPR000971|InterPro:IPR002335|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02144	MB		GO:0046872	metal ion binding	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0479	9606	20140913	UniProt	
Component													
UniProtKB	P02144	MB		GO:0070062	extracellular vesicular exosome	C	IDA	PMID:23533145		9606	20140714	UniProt	

Database	Gene Product ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
Process													
UniProtKB	P02008	HBZ		GO:0000122	negative regulation of transcription from RNA polymerase II promoter	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000020531	9606	20140913	Ensembl	
UniProtKB	P02008	HBZ		GO:0006810	transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0813	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		GO:0015671	oxygen transport	P	IEA	InterPro2GO	InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0015671	oxygen transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		GO:0043249	erythrocyte maturation	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000020531	9606	20140913	Ensembl	
Function													
UniProtKB	P02008	HBZ		GO:0005344	oxygen transporter activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		GO:0005344	oxygen transporter activity	F	TAS	PMID:7555018		9606	20030904	PINC	
UniProtKB	P02008	HBZ		GO:0005506	iron ion binding	F	IEA	InterPro2GO	InterPro:IPR000971|InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0005515	protein binding	F	IPI	PMID:11159543	UniProtKB:P68871	9606	20140914	IntAct	
UniProtKB	P02008	HBZ		GO:0005515	protein binding	F	IPI	PMID:6683087	UniProtKB:P68871	9606	20140914	IntAct	
UniProtKB	P02008	HBZ		GO:0019825	oxygen binding	F	IEA	InterPro2GO	InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0020037	heme binding	F	IEA	InterPro2GO	InterPro:IPR000971|InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0046872	metal ion binding	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0479	9606	20140913	UniProt	
Component													
UniProtKB	P02008	HBZ		GO:0005833	hemoglobin complex	C	IEA	InterPro2GO	InterPro:IPR002338|InterPro:IPR002340	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0005833	hemoglobin complex	C	TAS	PMID:7555018		9606	20030904	PINC	
UniProtKB	P02008	HBZ		GO:0070062	extracellular vesicular exosome	C	IDA	PMID:23533145		9606	20140714	UniProt	

Database	Gene Product ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
Process													
UniProtKB	P02144	MB		GO:0001666	response to hypoxia	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0006810	transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0813	9606	20140913	UniProt	
UniProtKB	P02144	MB		GO:0007507	heart development	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0009725	response to hormone	P	IEA	Ensembl Compara	Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0015671	oxygen transport	P	IEA	InterPro2GO	InterPro:IPR002335|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02144	MB		GO:0015671	oxygen transport	P	IEA	Ensembl Compara	Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0015671	oxygen transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02144	MB		GO:0031444	slow-twitch skeletal muscle fiber contraction	P	IEA	Ensembl Compara	Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0042542	response to hydrogen peroxide	P	IEA	Ensembl Compara	Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0043353	enucleate erythrocyte differentiation	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0050873	brown fat cell differentiation	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
Function													
UniProtKB	P02144	MB		GO:0005344	oxygen transporter activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02144	MB		GO:0005506	iron ion binding	F	IEA	InterPro2GO	InterPro:IPR000971|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02144	MB		GO:0019825	oxygen binding	F	IEA	InterPro2GO	InterPro:IPR002335|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02144	MB		GO:0019825	oxygen binding	F	IEA	Ensembl Compara	Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
UniProtKB	P02144	MB		GO:0020037	heme binding	F	IEA	InterPro2GO	InterPro:IPR000971|InterPro:IPR002335|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02144	MB		GO:0046872	metal ion binding	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0479	9606	20140913	UniProt	
Component													
UniProtKB	P02144	MB		GO:0070062	extracellular vesicular exosome	C	IDA	PMID:23533145		9606	20140714	UniProt	

SAME

SAME

Database	Gene Product ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
Process													
UniProtKB	P02008	HBZ		GO:0000122	negative regulation of transcription from RNA polymerase II promoter	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000020531	9606	20140913	Ensembl	
UniProtKB	P02008	HBZ		GO:0006810	transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0813	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		GO:0015671	oxygen transport	P	IEA	InterPro2GO	InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0015671	oxygen transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		GO:0043249	erythrocyte maturation	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000020531	9606	20140913	Ensembl	
Function													
UniProtKB	P02008	HBZ		GO:0005344	oxygen transporter activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		GO:0005344	oxygen transporter activity	F	TAS	PMID:7555018		9606	20030904	PINC	
UniProtKB	P02008	HBZ		GO:0005506	iron ion binding	F	IEA	InterPro2GO	InterPro:IPR000971|InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0005515	protein binding	F	IPI	PMID:15...	UniProtKB:P68871	9606	20140914	IntAct	
UniProtKB	P02008	HBZ		GO:0005515	protein binding	F	IPI	PMID:6...	UniProtKB:P68871	9606	20140914	IntAct	
UniProtKB	P02008	HBZ		GO:0019825	oxygen binding	F	IEA	InterPro2GO	InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0020037	heme binding	F	IEA	InterPro2GO	InterPro:IPR000971|InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0046872	metal ion binding	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0479	9606	20140913	UniProt	
Component													
UniProtKB	P02008	HBZ		GO:0005833	hemoglobin complex	C	IEA	InterPro2GO	InterPro:IPR002338|InterPro:IPR002340	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0005833	hemoglobin complex	C	TAS	PMID:7555018		9606	20030904	PINC	
UniProtKB	P02008	HBZ		GO:0070062	extracellular vesicular exosome	C	IDA	PMID:23533145		9606	20140714	UniProt	

SAME

SAME

Conservation of function (EC numbers) in homologous proteins

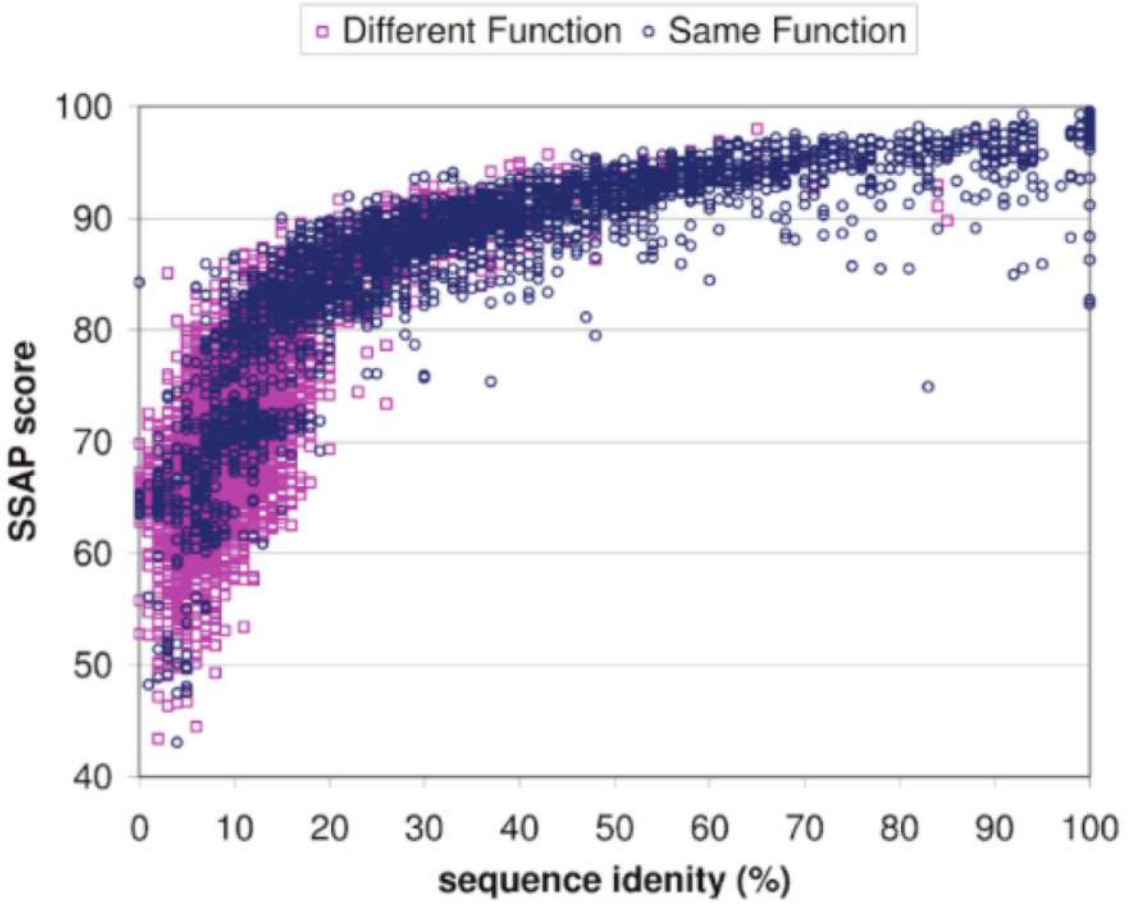


Fig. 7.8 Scatter plot showing the relationship between sequence, structure, and function of all homologues in enzyme superfamilies. Relatives having the same EC classification number are shown in blue. Those with different EC numbers are shown in pink.

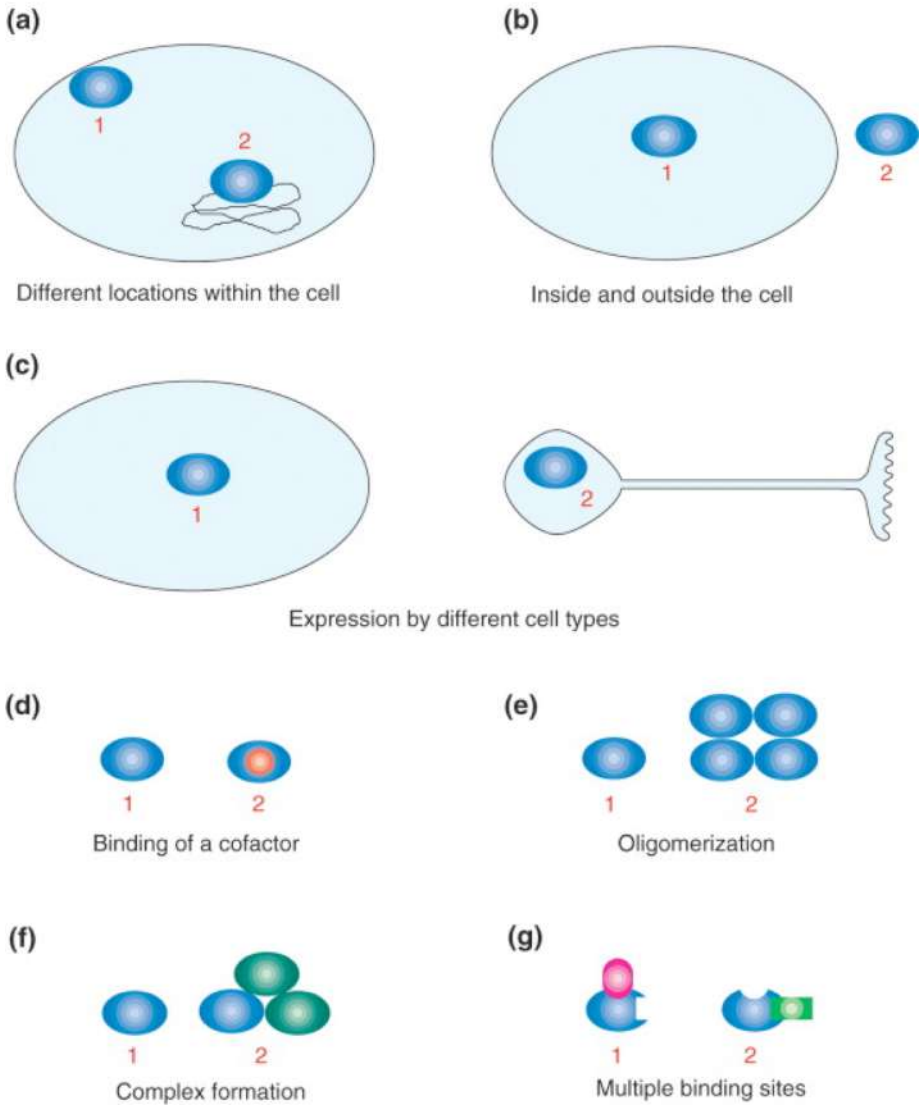
Computational Structural Biology: Methods and Applications Torsten Schwede (2008)

Conservation of function in homologs

- Function not always conserved in homologous proteins
- Correlates with sequence/structural similarity
- No safe threshold
- Even single mutations can induce important changes in function (e.g. oncogenic and resistance point mutations in cancer)
- Protein may share only some of their functions
- Always look for additional evidence related to function (functional residues, functional motifs, cellular localisation, patterns of expression etc.) and eventually perform experiments.

Moonlighting proteins

Marco Punta



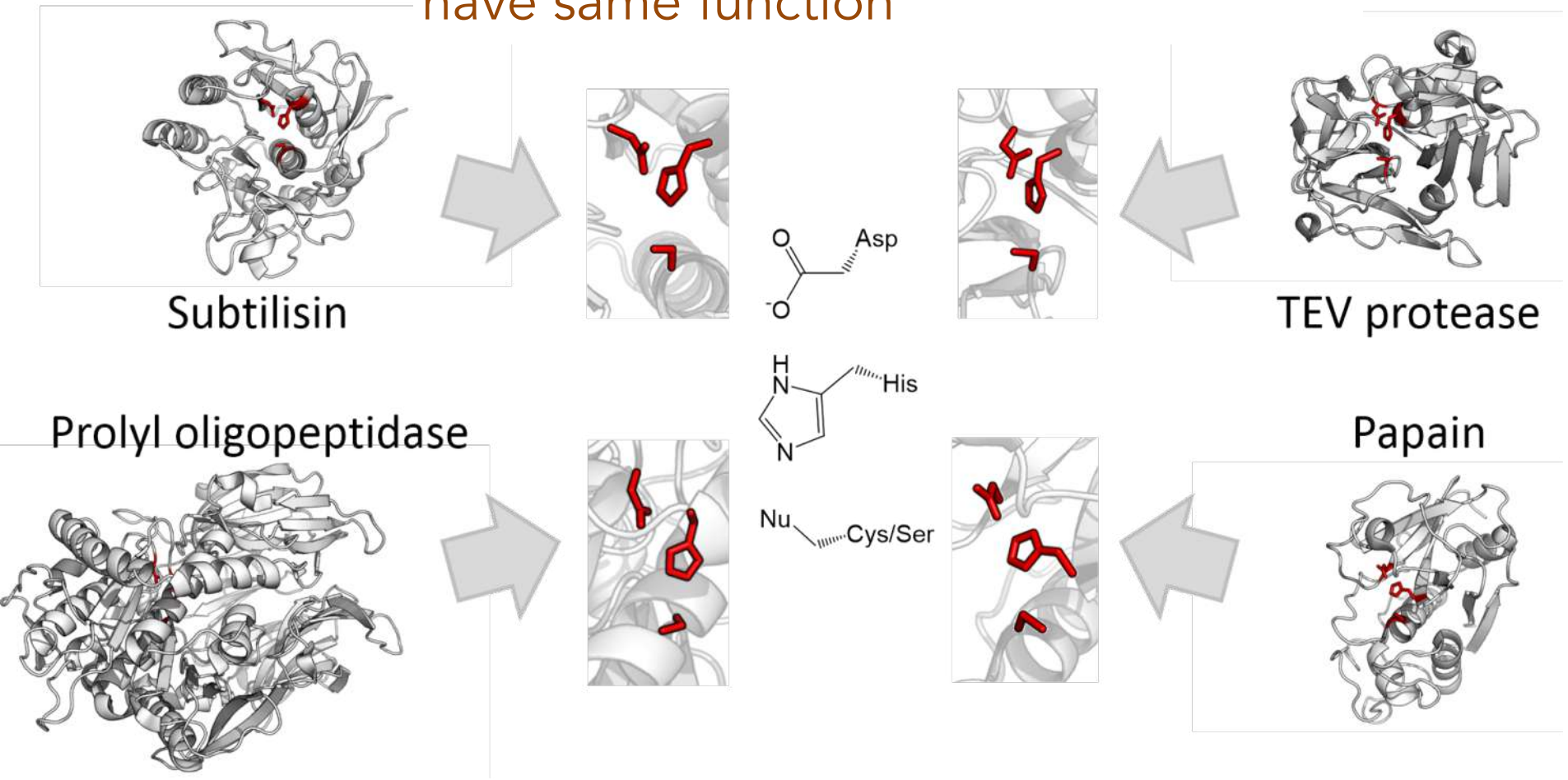
Jeffery Trends Bioch. Sc. (1999)

Still…

"Detecting sequence similarity in order to uncover homologous relationships between proteins remains the single most powerful tool for function prediction" Ochoa….Singh PLOS CB (2015)

Note: homology not necessary to have same function

Marco Punta



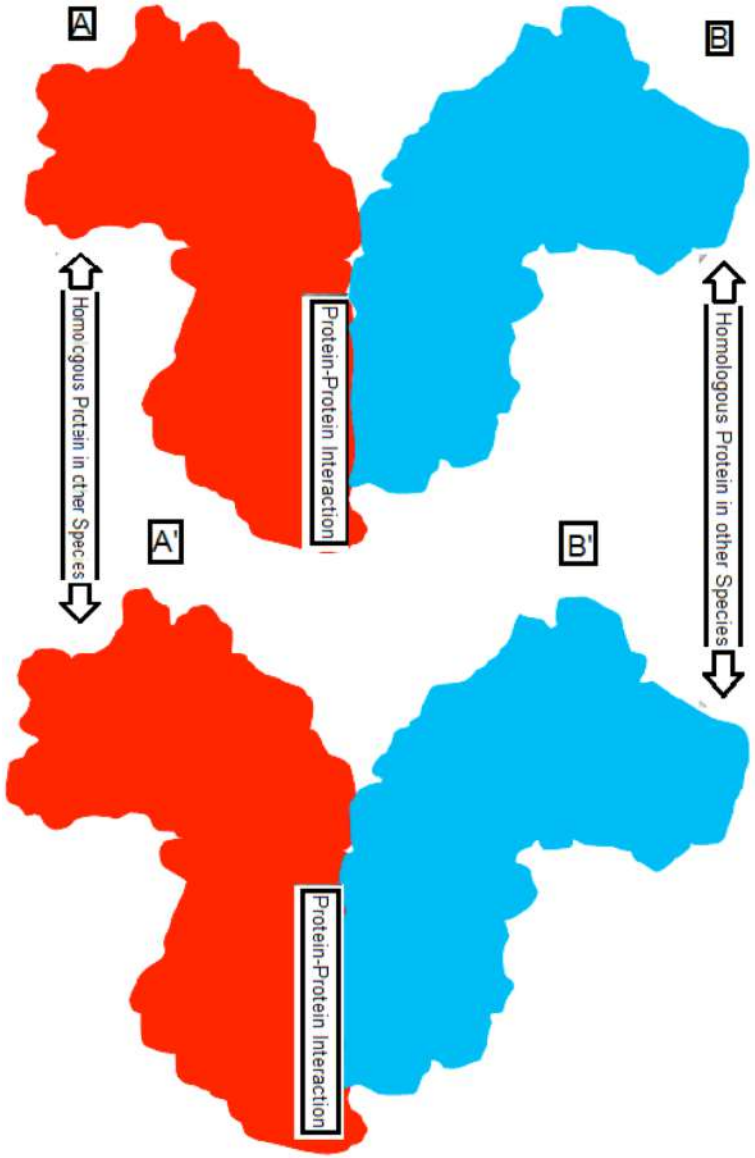
"Triad Convergence" by Thomas Shafee - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Triad_Convergence.png#/media/File:Triad_Convergence.png

See the Book: Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Koonin and Galperin (2003)

How about homology and protein-protein interactions?

Interologs

Marco Punta



https://en.wikipedia.org/wiki/Interolog#/media/File:HypotheticalInterologExample.png

Interologs

Marco Punta

What Evidence Is There for the Homology of Protein-Protein Interactions?

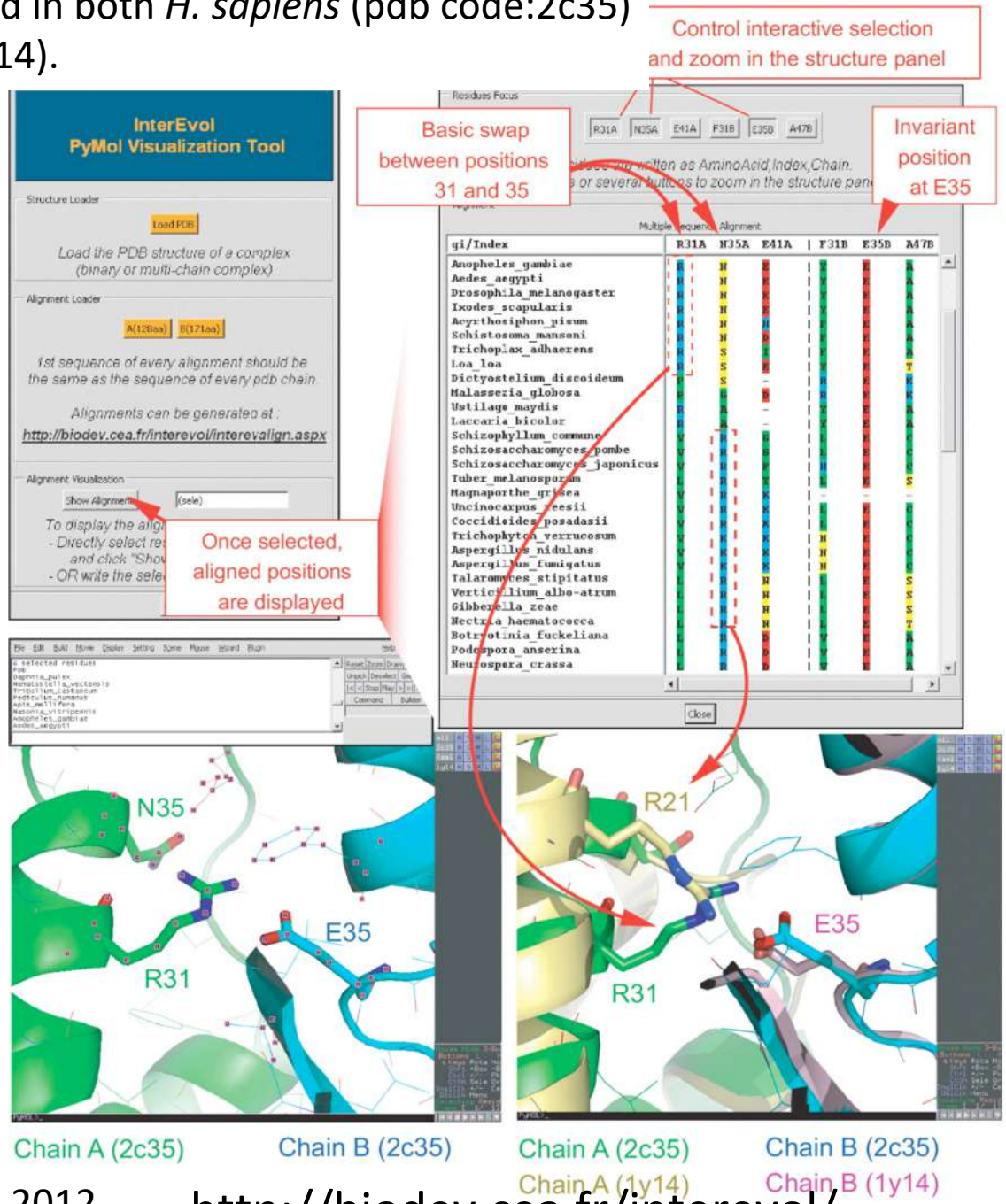
Anna C. F. Lewis, Nick S. Jones, Mason A. Porter, Charlotte M. Deane

Published: September 20, 2012 • http://dx.doi.org/10.1371/journal.pcbi.1002645

Abstract

The notion that sequence homology implies functional similarity underlies much of computational biology. In the case of protein-protein interactions, an interaction can be inferred between two proteins on the basis that sequence-similar proteins have been observed to interact. The use of transferred interactions is common, but the legitimacy of such inferred interactions is not clear. Here we investigate transferred interactions and whether data incompleteness explains the lack of evidence found for them. Using definitions of homology associated with functional annotation transfer, we estimate that conservation rates of interactions are low even after taking interactome incompleteness into account. For example, at a blastp E-value threshold of 10^{-70}, we estimate the conservation rate to be about 11% between *S. cerevisiae* and *H. sapiens*. Our method also produces estimates of interactome sizes (which are similar to those previously proposed). Using our estimates of interaction conservation we estimate the rate at which protein-protein interactions are lost across species. To our knowledge, this is the first such study based on large-scale data. Previous work has suggested that interactions transferred within species are more reliable than interactions transferred across species. By controlling for factors that are specific to within-species interaction prediction, we propose that the transfer of interactions within species might be less reliable than transfers between species. Protein-protein interactions appear to be very rarely conserved unless very high sequence similarity is observed. Consequently, inferred interactions should be used with care.

Rpb4-Rpb7 complex crystallized in both *H. sapiens* (pdb code:2c35) and *S. cerevisiae* (pdb code:1y14).

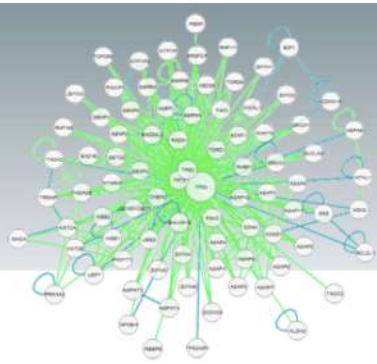


Faure et al. Nucleic Acids Res. 2012

http://biodev.cea.fr/interevol/

INTEROLOG FINDER

Start New Analysis Page



navigation

Pages

Start New Analysis
Downloads
About Us
Help and FAQ

Protein or Proteins of interest:

Paste identifiers, comma separated or list format:
NCBI IDs are preferred, but gene names, Ensembl IDs, and several other identifiers are translated
example: 672, TP53, ENSG00000107331

Species:

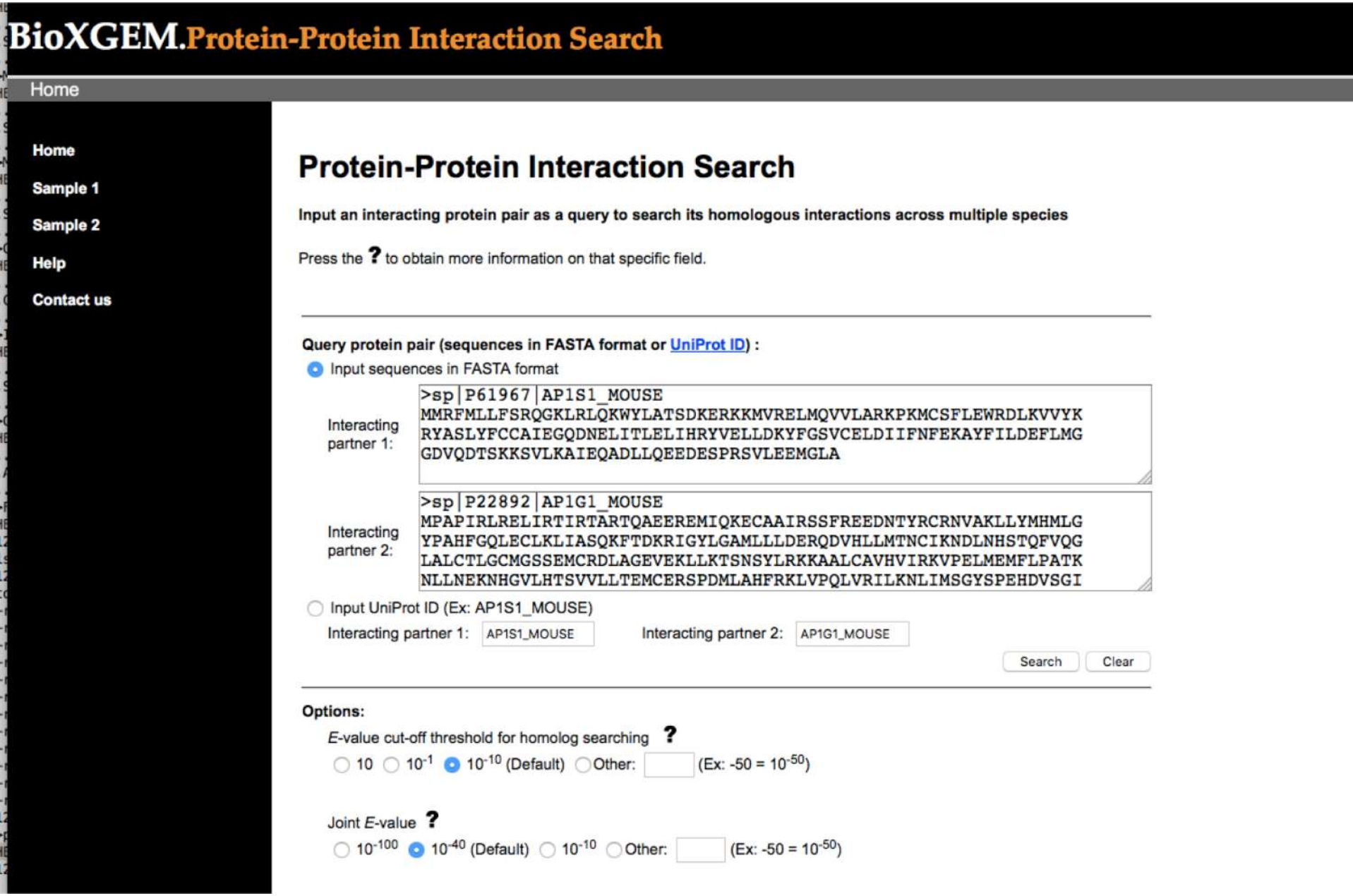
- *Homo sapiens*
- *Mus musculus*
- *Drosophila melanogaster*
- *Caenorhabditis elegans*
- *Saccharomyces cerevisiae*

Get Interologs

On the next page results and possible synonyms to your input proteins will be displayed; chose the correct ID from the following synonym list and click extend.
If you need further help with identifiers, please visit Ensembl or NCBI. Click on the button to add selected genes.

http://interologfinder.org/simpleIfWS?operation=doStart

Wiles et al. BMC Syst Biol. 2010



http://gemdock.life.nctu.edu.tw/ppisearch/

Chen et al. NAR 2009

Exercise

We are going to align two protein sequences.

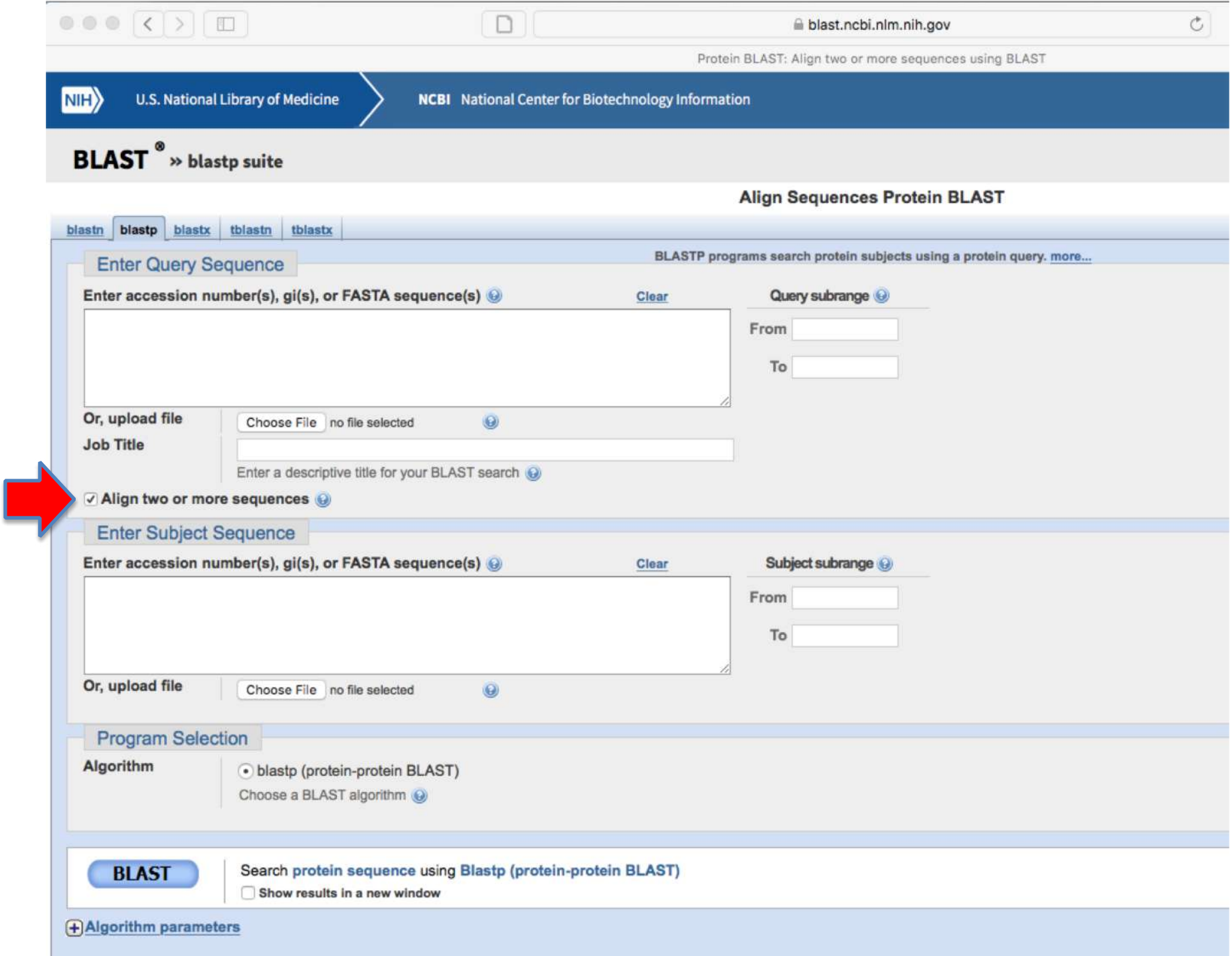
The first one is:
P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1

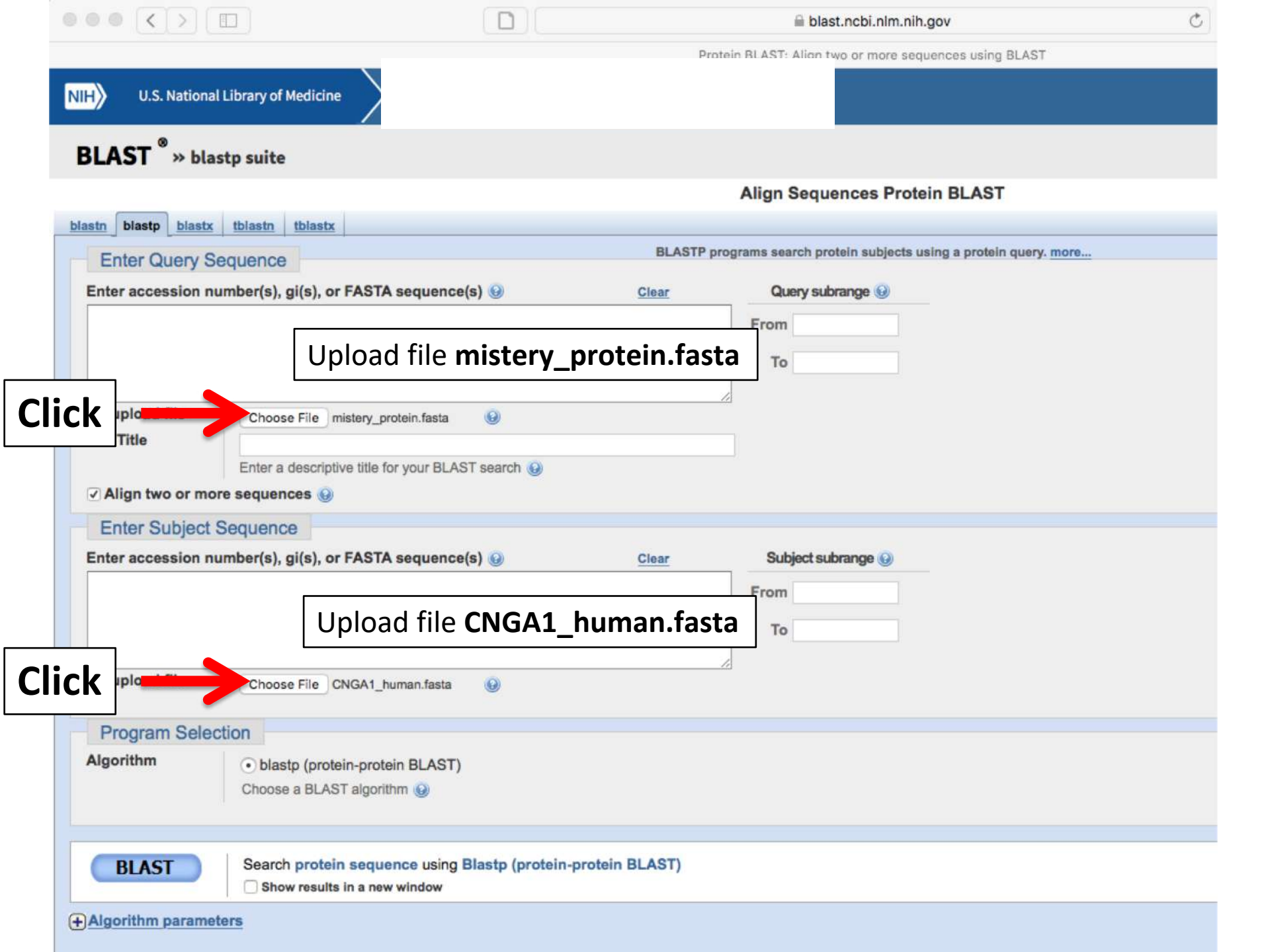
the second one is a mistery protein…

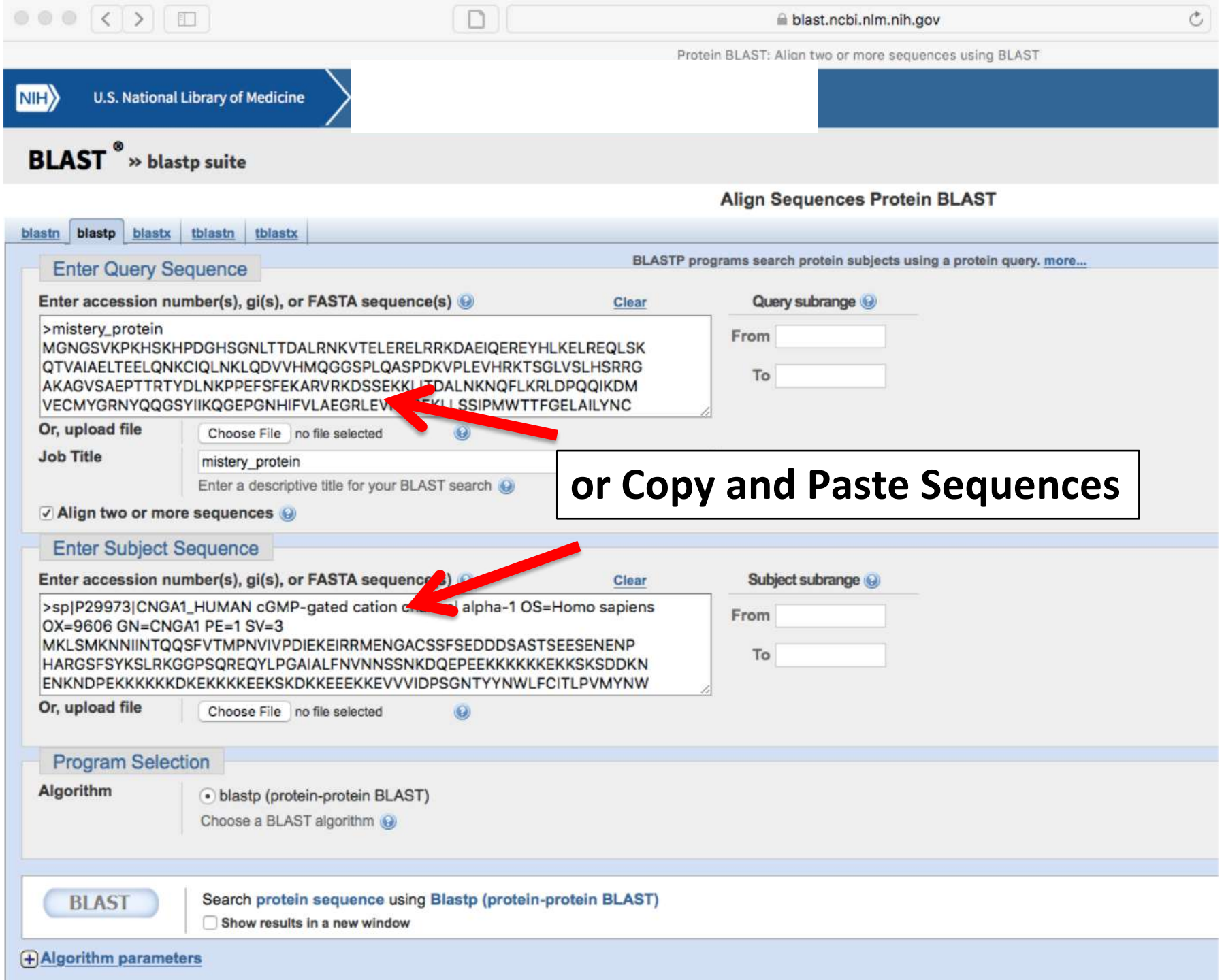
Let’s go to the BLAST website:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

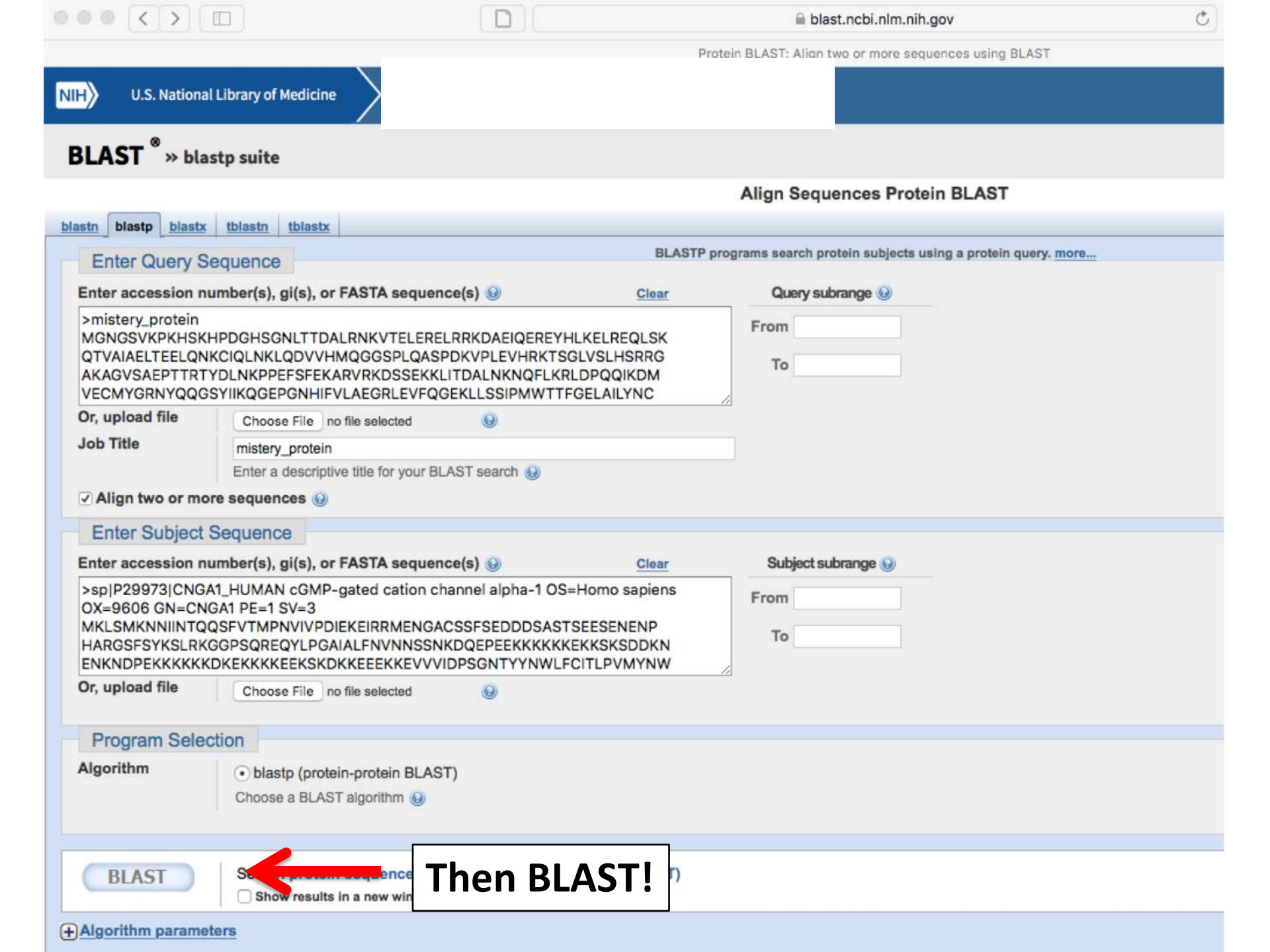
See file: blast_pairwise_link.txt


blast.ncbi.nlm.nih.gov
Protein BLAST: Align two or more sequences using BLAST
NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
BLAST ® » blastp suite
Align Sequences Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
BLASTP programs search protein subjects using a protein query. more...
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Choose File
no file selected
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Enter Subject Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Subject subrange
From
To
Or, upload file
Choose File
no file selected
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters


blast.ncbi.nlm.nih.gov
Protein BLAST: Align two or more sequences using BLAST
U.S. National Library of Medicine
BLAST® » blastp suite
Align Sequences Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
BLASTP programs search protein subjects using a protein query. more...
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Upload file mistery_protein.fasta
Click
Choose File
mistery_protein.fasta
Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Enter Subject Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Subject subrange
From
To
Upload file CNGA1_human.fasta
Click
Choose File
CNGA1_human.fasta
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters


blast.ncbi.nlm.nih.gov
Protein BLAST: Align two or more sequences using BLAST
NIH
U.S. National Library of Medicine
BLAST® » blastp suite
Align Sequences Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
BLASTP programs search protein subjects using a protein query. more...
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
>mistery_protein
MGNGSVKPKHSKHPDGHSGNLTTDALRNKVTELERELRRKDAEIQEREYHLKELREQLSK
QTVAIAELTEELQNKCIQLNKLQDVVHMQGGSPLQASPDKVPLEVHRKTSGLVSLHSRRG
AKAGVSAEPTTRTYDLNKPPEFSFEKARVRKDSSEKKL
DALNKNQFLKRLDPQQIKDM
VECMYGRNYQQGSYIIKQGEPGNHIFVLAEGRLEV
SSIPMWTTFGELAILYNC
Or, upload file
Choose File
no file selected
Job Title
mistery_protein
Enter a descriptive title for your BLAST search
Align two or more sequences
or Copy and Paste Sequences
Enter Subject Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Subject subrange
From
To
>sp|P29973|CNGA1_HUMAN cGMP-gated cation
alpha-1 OS=Homo sapiens
OX=9606 GN=CNGA1 PE=1 SV=3
MKLSMKNNIINTQQSFVTMPNVIVPDIEKEIRRMENGACSSFSEDDDSASTSEESENENP
HARGSFSYKSLRKGGPSQREQYLPGAIALFNVNNSSNKDQEPEEKKKKKKEKKSKSDDKN
ENKNDPEKKKKKDKEKKKKEEKSKDKKEEEKKEVVVIDPSGNTYYNWLFCITLPVMYNW
Or, upload file
Choose File
no file selected
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Search protein sequence using Blastp (protein-protein BLAST)
Show results in a new window
Algorithm parameters


UniProt
UniProtKB
Advanced
Search
BLAST Align Retrieve/ID mapping Peptide search
Help Contac
UniProtKB - P29973 (CNGA1_HUMAN)
Basket
Display
Entry
Publications
Feature viewer
Feature table
None
Function
Names & Taxonomy
Subcellular location
Pathology & Biotech
PTM / Processing
Expression
Interaction
Structure
Family & Domains
Sequences (2+)
Similar proteins
Cross-references
BLAST Align Format Add to basket History
Help video Add a publication Feedba
Protein cGMP-gated cation channel alpha-1
Gene CNGA1
Organism Homo sapiens (Human)
Status Reviewed - Annotation score: - Experimental evidence at protein level
Function
Subunit of the rod cyclic GMP-gated cation channel, which is involved in the final stage of the phototransduction pathway. When light hits rod photoreceptors, cGMP concentrations decrease causing rapid closure of CNGA1/CNGB1 channels and, therefore, hyperpolarization of the membrane potential. By similarity
Caution
It is uncertain whether Met-1 or Met-5 is the initiator. Curated
Sites
Feature key Position(s) Description Actions Graphical view Lengt
Binding site 546 cGMP Sequence analysis
Binding site 561 cGMP Sequence analysis
Regions
Feature key Position(s) Description Actions Graphical view Lengt
Nucleotide binding 487 – 609 cGMP Sequence analysis Add BLAST 12


blast.ncbi.nlm.nih.gov
Protein BLAST: Align two or more sequences using BLAST
NIH
U.S. National Library of Medicine
BLAST® » blastp suite
Align Sequences Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
Enter Query Sequence
BLASTP programs search protein subjects using a protein query. more...
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
>mistery_protein
MGNGSVKPKHSKHPDGHSGNLTTDALRNKVTELERELRRKDAEIQEREYHLKELREQLSK
QTVAIAELTEELQNKCIQLNKLQDVVHMQGGSPLQASPDKVPLEVHRKTSGLVSLHSRRG
AKAGVSAEPTTRTYDLNKPPEFSFEKARVRKDSSEKKLITDALNKNQFLKRLDPQQIKDM
VECMYGRNYQQGSYIIKQGEPGNHIFVLAEGRLEVFQGEKLLSSIPMWTTFGELAILYNC
Or, upload file
Choose File
no file selected
Job Title
mistery_protein
Enter a descriptive title for your BLAST search
Align two or more sequences
Enter Subject Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Subject subrange
From
To
>sp|P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens
OX=9606 GN=CNGA1 PE=1 SV=3
MKLSMKNNIINTQQSFVTMPNVIVPDIEKEIRRMENGACSSFSEDDDSASTSEESENENP
HARGSFSYKSLRKGGPSQREQYLPGAIALFNVNNSSNKDQEPEEKKKKKKEKKSKSDDKN
ENKNDPEKKKKKDKEKKKKEEKSKDKKEEEKKEVVVIDPSGNTYYNWLFCITLPVMYNW
Or, upload file
Choose File
no file selected
Program Selection
Algorithm
blastp (protein-protein BLAST)
Choose a BLAST algorithm
BLAST
Show results in a new win
Algorithm parameters
Then BLAST!

BLAST Results

Edit and Resubmit | Save Search Strategies | ▷Formatting options | ▷Download | How to read this page | Blast report des

Blast 2 sequences

Job title: mistery_protein (762 letters)

RID XXGSCDW8114 (Expires on 11-05 16:59 pm)

Query ID lcl|Query_208167
Description mistery_protein
Molecule type amino acid
Query Length 762

Subject ID lcl|Query_208169
Description sp|P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens OX=
GN=CNGA1 PE=1 SV=3
Molecule type amino acid
Subject Length 690
Program BLASTP 2.8.1+ ▷Citation

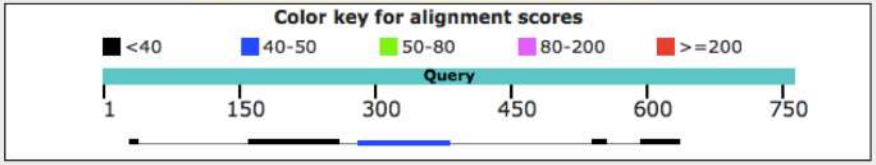
Other reports: ▷Search Summary [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

Graphic Summary

Distribution of the top 5 Blast Hits on 1 subject sequences

Mouse over to see the title, click to show alignments



Dot Matrix View

Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | Graphics | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
sp|P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens OX=9606 GN=CNGA1 PE=1 SV=3	41.6	136	35%	1e-07	27%	Query_208169

BLAST Results

Edit and Resubmit | Save Search Strategies | ▷Formatting options | ▷Download | YouTube How to read this page | Blast report des

Blast 2 sequences

Job title: mistery_protein (762 letters)

RID XXGSCDW8114 (Expires on 11-05 16:59 pm)

Query ID lcl|Query_208167
Description mistery_protein
Molecule type amino acid
Query Length 762

Subject ID lcl|Query_208169
Description sp|P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens OX=
GN=CNGA1 PE=1 SV=3
Molecule type amino acid
Subject Length 690
Program BLASTP 2.8.1+ ▷Citation

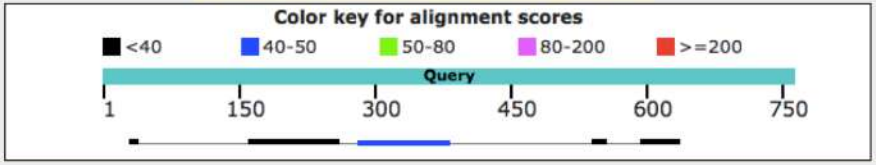
Other reports: ▷Search Summary [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

Graphic Summary

Distribution of the top 5 Blast Hits on 1 subject sequences

Mouse over to see the title, click to show alignments



Dot Matrix View

Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download Graphics Multiple alignment

Description	Max score	Total sc	Query cover	E value	Ident	Accession
sp|P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens OX=9606 GN=CNGA1 PE=1 SV=3	41.6	136	35%	1e-07	27%	Query_208169

E-value=10^{-7}

Is our mistery protein a cGMP-gated cation channel?

Alignments

Download Graphics Sort

sp|P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens OX=9606 GN=CNGA1 PE=1 SV=3

Sequence ID: Query_89161 Length: 690 Number of Matches: 5

Range 1: 474 to 579 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
41.6 bits(96)	1e-07	Compositional matrix adjust.	30/110(27%)	54/110(49%)	11/110(10%)

```
Query  281  LRSVSLLKNLPEDKLTKIIDCLEVEYYDKGDYIIREGEEGSTFFILAKGKVKVTQSTEGH  340
            L+ V +  +     L +++  L+ + Y  GDYI ++G+ G   +I+ +GK+ V
Sbjct  474  LKKVRIFADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVV----AD  529

Query  341  DQPQLIKTLQKGEYFGEKALI------SDDVRSANIIA-EENDVACLVID  383
            D       L  G YFGE +++       + + R+ANI +   +D+ CL  D
Sbjct  530  DGVTQFVVLSDGSYFGEISILNIKGSKAGNRRTANIKSIGYSDLFCLSKD  579
```

E-value=10^{-7}

Range 2: 472 to 579 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
37.7 bits(86)	1e-06	Compositional matrix adjust.	26/108(24%)	53/108(49%)	8/108(7%)

```
Query  161  DALNKNQFLKRLDPQQIKDMVECMYGRNYQQGSYIIKQGEPGNHIFVLAEGRLEVFQGEK  220
            D L K +     +    + ++V  +   + Y  G YI K+G+ G   ++++ EG+L V   +
Sbjct  472  DTLKKVRIFADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVVADDG  531

Query  221  LLSSIPM--WTTFGELAIL------YNCTRTASVKAITNVKTWALDRE  260
            +   + +   + FGE++IL           RTA++K+I     + L ++
Sbjct  532  VTQFVVLSDGSYFGEISILNIKGSKAGNRRTANIKSIGYSDLFCLSKD  579
```

E-value=10^{-6}

Range 3: 317 to 356 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
21.9 bits(45)	0.11	Compositional matrix adjust.	14/44(32%)	23/44(52%)	5/44(11%)

```
Query  593  VDFGFAKKIGSGQKTWTFCGTPEYVAPEV-ILNKGHDFSVDFWS  635
            V +  +K IG G  TW +   P+   PE   L + + +S+ +WS
Sbjct  317  VFYSISKAIGFGNDTWVY---PDINDPEFGRLARKYVYSL-YWS  356
```

Range 4: 180 to 196 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
18.5 bits(36)	1.2	Compositional matrix adjust.	6/17(35%)	10/17(58%)	0/17(0%)

```
Query  539  WSILRDRGSFDEPTSKF  555
```

Alignments

Download Graphics Sort by: E value

sp|P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens OX=9606 GN=CNGA1 PE=1 SV=3

Sequence ID: Query_89161 Length: 690 Number of Matches: 5

Range 1: 474 to 579 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
41.6 bits(96)	1e-07	Compositional matrix adjust.	30/110(27%)	54/110(49%)	11/110(10%)

```
Query  281  LRSVSLLKNLPEDKLTKIIDCLEVEYYDKGDYIIREGEEGSTFFILAKGKVKVTQSTEGH  340
            L+ V +  +     L +++  L+ + Y  GDYI ++G+ G   +I+ +GK+ V
Sbjct  474  LKKVRIFADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVV----AD  529

Query  341  DQPQLIKTLQKGEYFGEKALI------SDDVRSANIIA-EENDVACLVID  383
            D       L  G YFGE +++      + + R+ANI +   +D+ CL  D
Sbjct  530  DGVTQFVVLSDGSYFGEISILNIKGSKAGNRRTANIKSIGYSDLFCLSKD  579
```

Range 2: 472 to 579 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
37.7 bits(86)	1e-06	Compositional matrix adjust.	26/108(24%)	53/108(49%)	8/108(7%)

```
Query  161  DALNKNQFLKRLDPQQIKDMVECMYGRNYQQGSYIIKQGEPGNHIFVLAEGRLEVFQGEK  220
            D L K +     +   + ++V  +  + Y  G YI K+G+ G  ++++ EG+L V   +
Sbjct  472  DTLKKVRIFADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVVADDG  531

Query  221  LLSSIPM--WTTFGELAIL------YNCTRTASVKAITNVKTWALDRE  260
            +   + +   + FGE++IL           RTA++K+I     + L ++
Sbjct  532  VTQFVVLSDGSYFGEISILNIKGSKAGNRRTANIKSIGYSDLFCLSKD  579
```

Alignments

Download Graphics Sort by: E value

sp|P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens OX=9606 GN=CNGA1 PE=1 SV=3

Sequence ID: Query_89161 Length: 690 Number of Matches: 5

Range 1: 474 to 579 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
41.6 bits(96)	1e-07	Compositional matrix adjust.	30/110(27%)	54/110(49%)	11/110(10%)

```
Query  281  LRSVSLLKNLPEDKLTKIIDCLEVEYYDKGDYIIREGEEGSTFFILAKGKVKVTQSTEGH  340
            L+ V +  +     L +++  L+ + Y  GDYI ++G+ G   +I+ +GK+ V
Sbjct  474  LKKVRIFADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVV----AD  529

Query  341  DQPQLIKTLQKGEYFGEKALI------SDDVRSANIIA-EENDVACLVID  383
            D       L  G YFGE +++      + + R+ANI +   +D+ CL  D
Sbjct  530  DGVTQFVVLSDGSYFGEISILNIKGSKAGNRRTANIKSIGYSDLFCLSKD  579
```

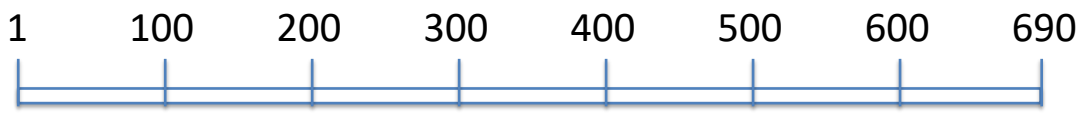
Range 2: 472 to 579 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
37.7 bits(86)	1e-06	Compositional matrix adjust.	26/108(24%)	53/108(49%)	8/108(7%)

```
Query  161  DALNKNQFLKRLDPQQIKDMVECMYGRNYQQGSYIIKQGEPGNHIFVLAEGRLEVFQGEK  220
            D L K +     +   + ++V  +  + Y  G YI K+G+ G  ++++ EG+L V   +
Sbjct  472  DTLKKVRIFADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVVADDG  531

Query  221  LLSSIPM--WTTFGELAIL------YNCTRTASVKAITNVKTWALDRE  260
            +   + +   + FGE++IL           RTA++K+I     + L ++
Sbjct  532  VTQFVVLSDGSYFGEISILNIKGSKAGNRRTANIKSIGYSDLFCLSKD  579
```

cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)

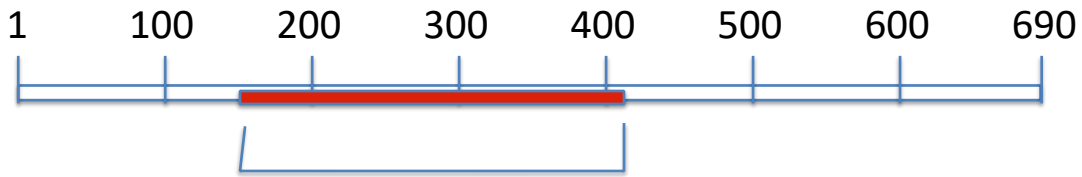
Marco Punta



ICR The Institute of Cancer Research

cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)

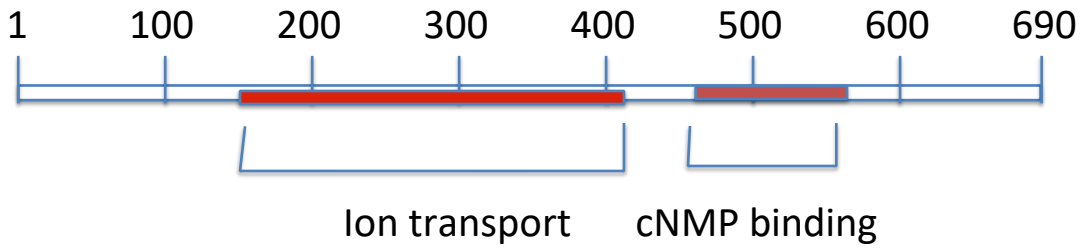
Marco Punta



ICR The Institute of Cancer Research

cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)

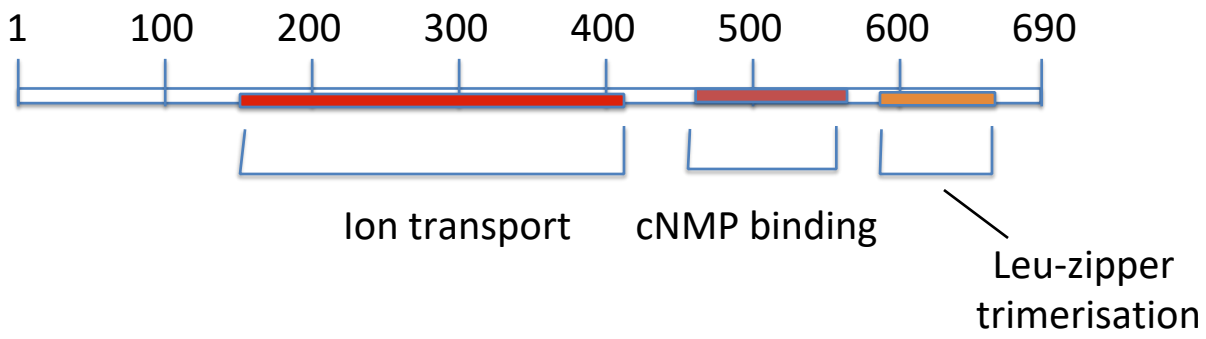
Marco Punta



ICR The Institute of Cancer Research

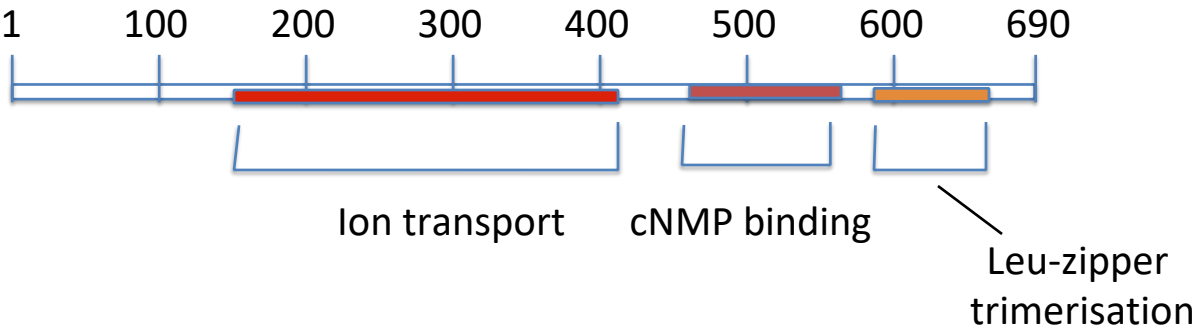
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)

Marco Punta

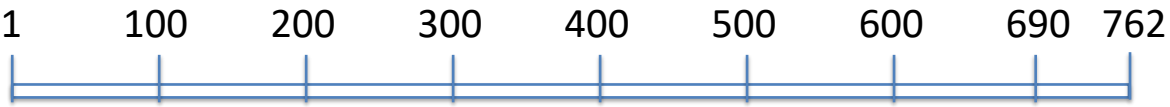


ICR The Institute of Cancer Research

cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)

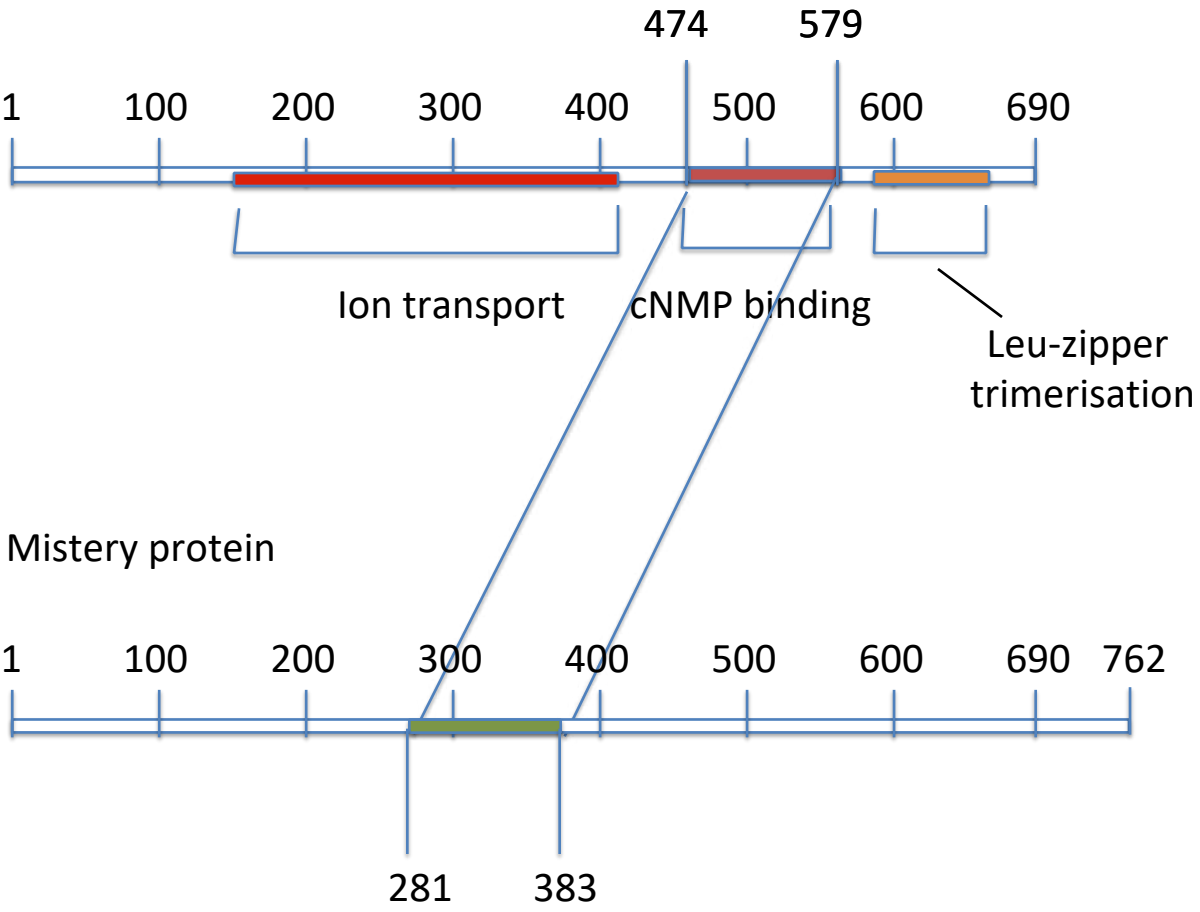


Mistery protein



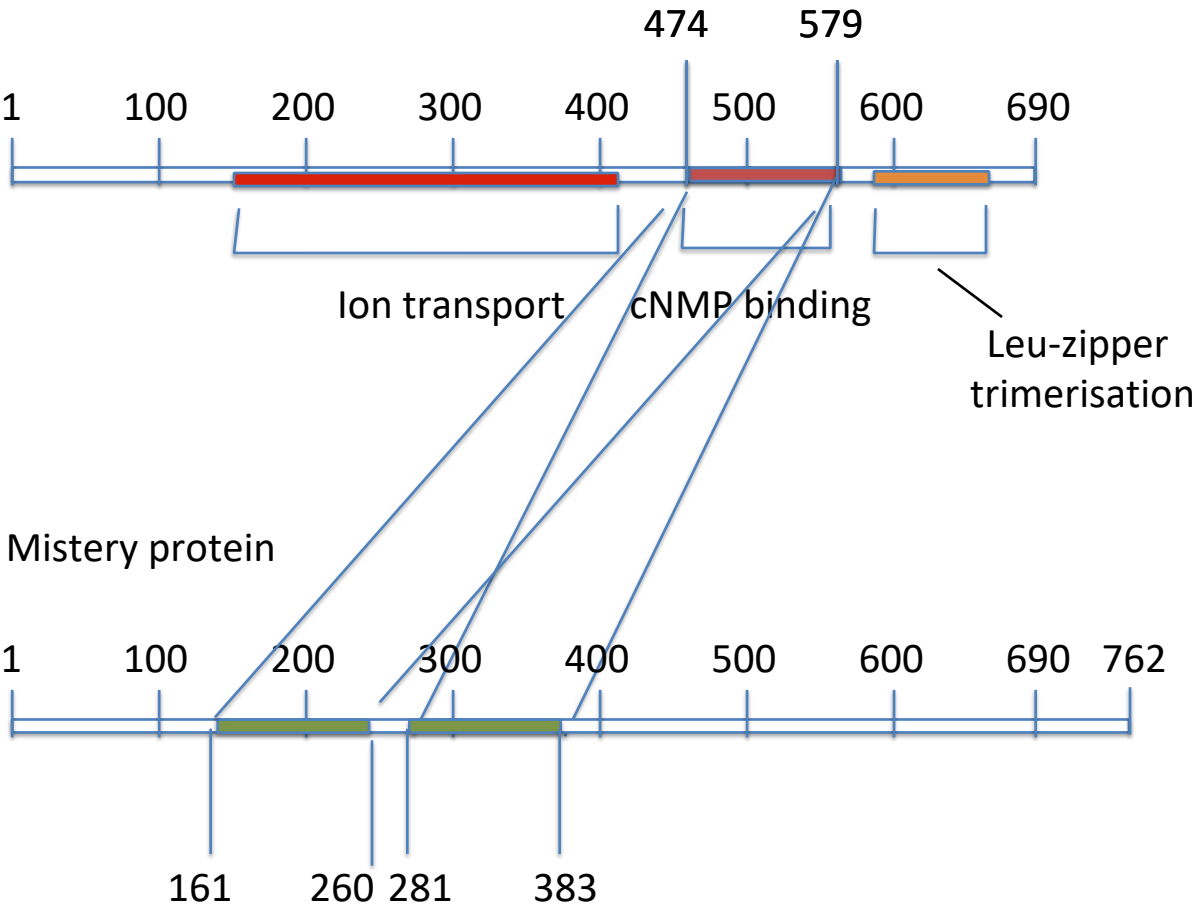
Marco Punta

cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)



Marco Punta

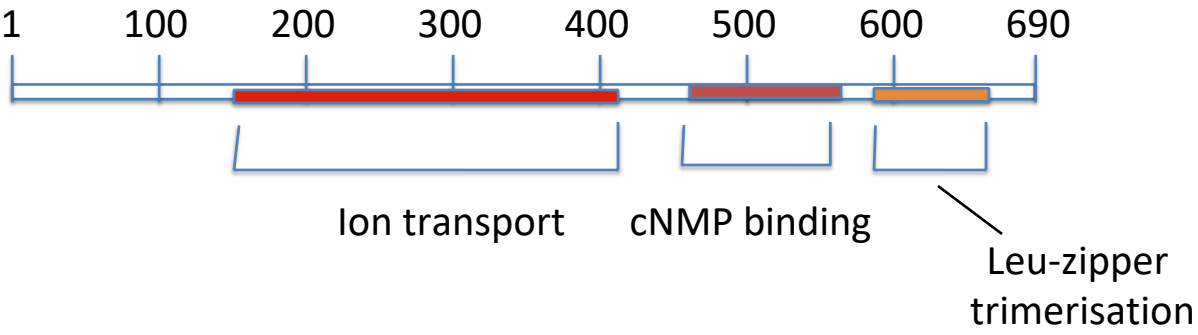
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)



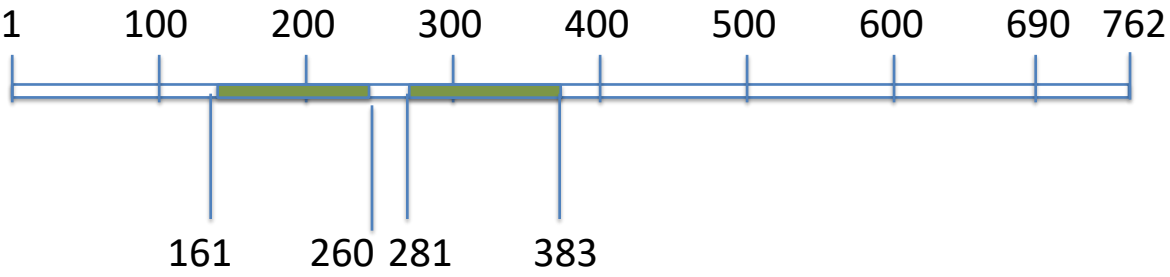
ICR The Institute of Cancer Research

cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)

Marco Punta



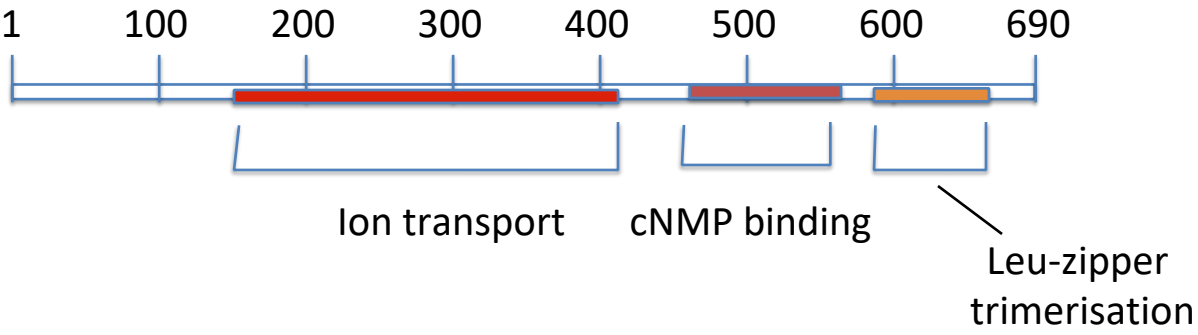
Mistery protein



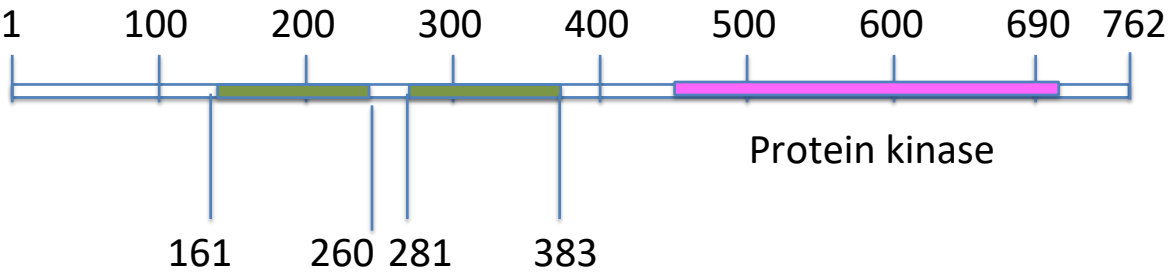
cNMP binding?

ICR The Institute of Cancer Research

cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)

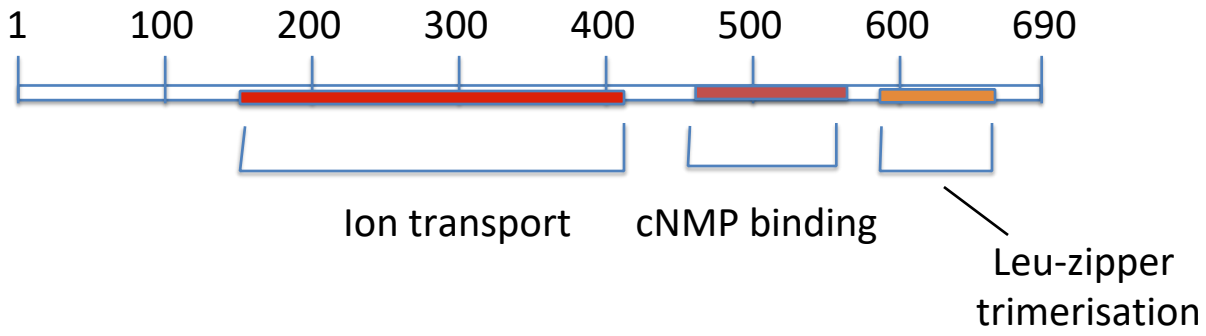


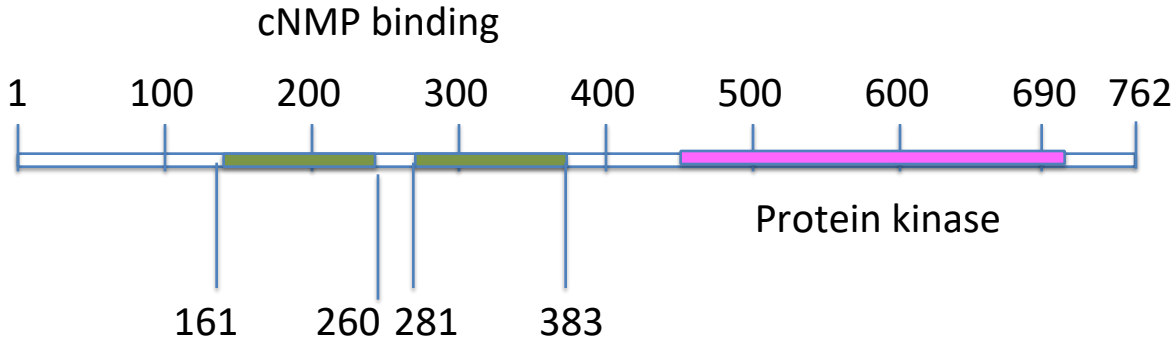
Mistery protein



cNMP binding?

cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)





Mystery protein is a cGMP-dependent protein kinase 2
Q13237 (KGP2_HUMAN)

Definition (Wikipedia):

A protein domain is a conserved part of a given protein sequence and (tertiary) structure that can evolve, function, and exist independently of the rest of the protein chain. Each domain forms a compact three-dimensional structure and often can be independently stable and folded. Many proteins consist of several structural domains. One domain may appear in a variety of different proteins. Molecular evolution uses domains as building blocks and these may be recombined in different arrangements to create proteins with different functions.

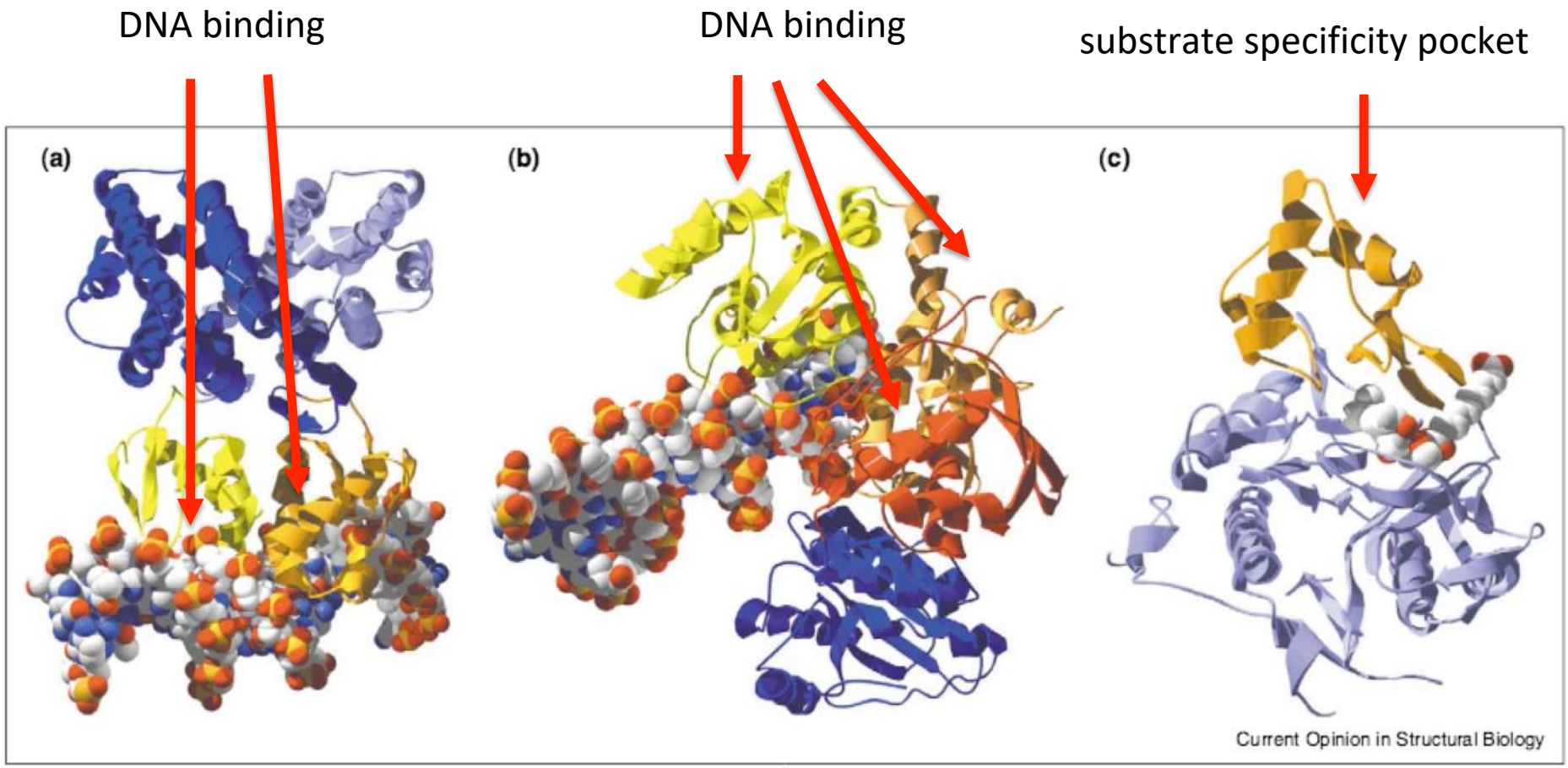
Domains and function annotation

- Proteins may be homologous only in some regions (domains), this is especially true at longer evolutionary distance

- If so, function annotation transfer possible (still not safe) only between these regions

Winged helix domain

Marco Punta



Transcription factor

Restriction endonuclease

Human methionine aminopeptidase 2

Vogel et al. Curr. Opin. Struct. Biol. 14 (2004)

Beyond just homology

BMC Part of Springer Nature

Search Explore

Genome Biology

Home About Articles Submission Guidelines

Research | Open Access | Published: 19 November 2019

The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens

Naihui Zhou, Yuxiang Jiang, [...] Iddo Friedberg

Genome Biology **20**, Article number: 244 (2019) | Cite this article

1157 Accesses | **36** Altmetric | Metrics

Abstract

Background

The Critical Assessment of Functional Annotation (CAFA) is an ongoing, global, community-driven effort to evaluate and improve the computational annotation of protein function.

Protein families

- Members will share a structural core (homology modeling)
- Members may share aspects of function (function prediction)
- The whole set of members may reveal elements of protein and organism evolution (phylogenies)

Protein family databases

(most but not all try to build families for domains)

Functions, organisms, structures

Marco Punta

Prokaryotes both domains and full-length equivalogs

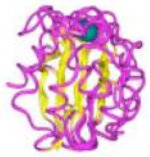
TIGRFAMs

Signalling, extracellular and chromatin-associated proteins

SMART

Superfamily 1.75

HMM library and genome assignments server



Structural domains from SCOP

Gene3D

Structural domains from CATH

ICR The Institute of Cancer Research

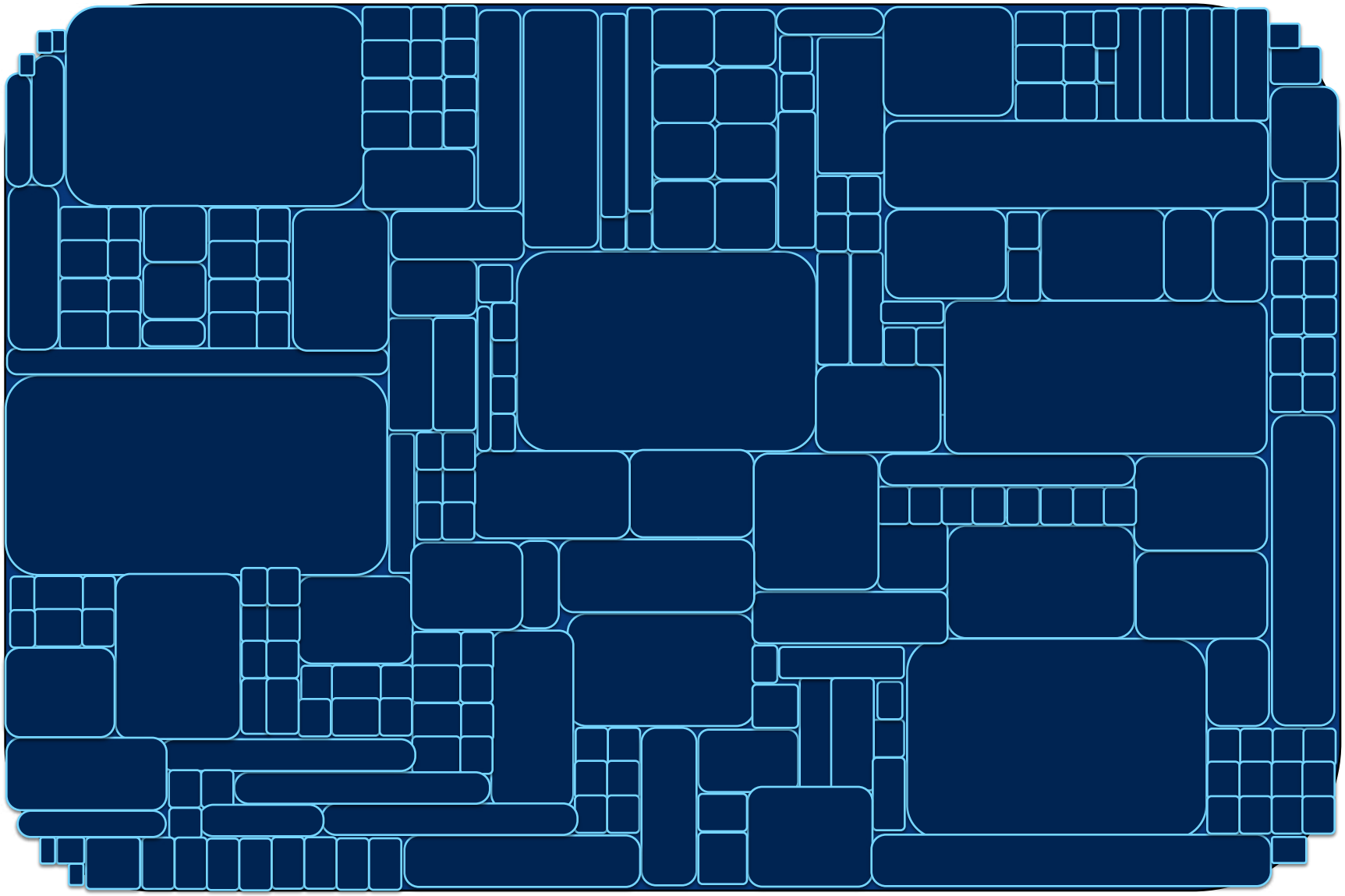
No limits, domains

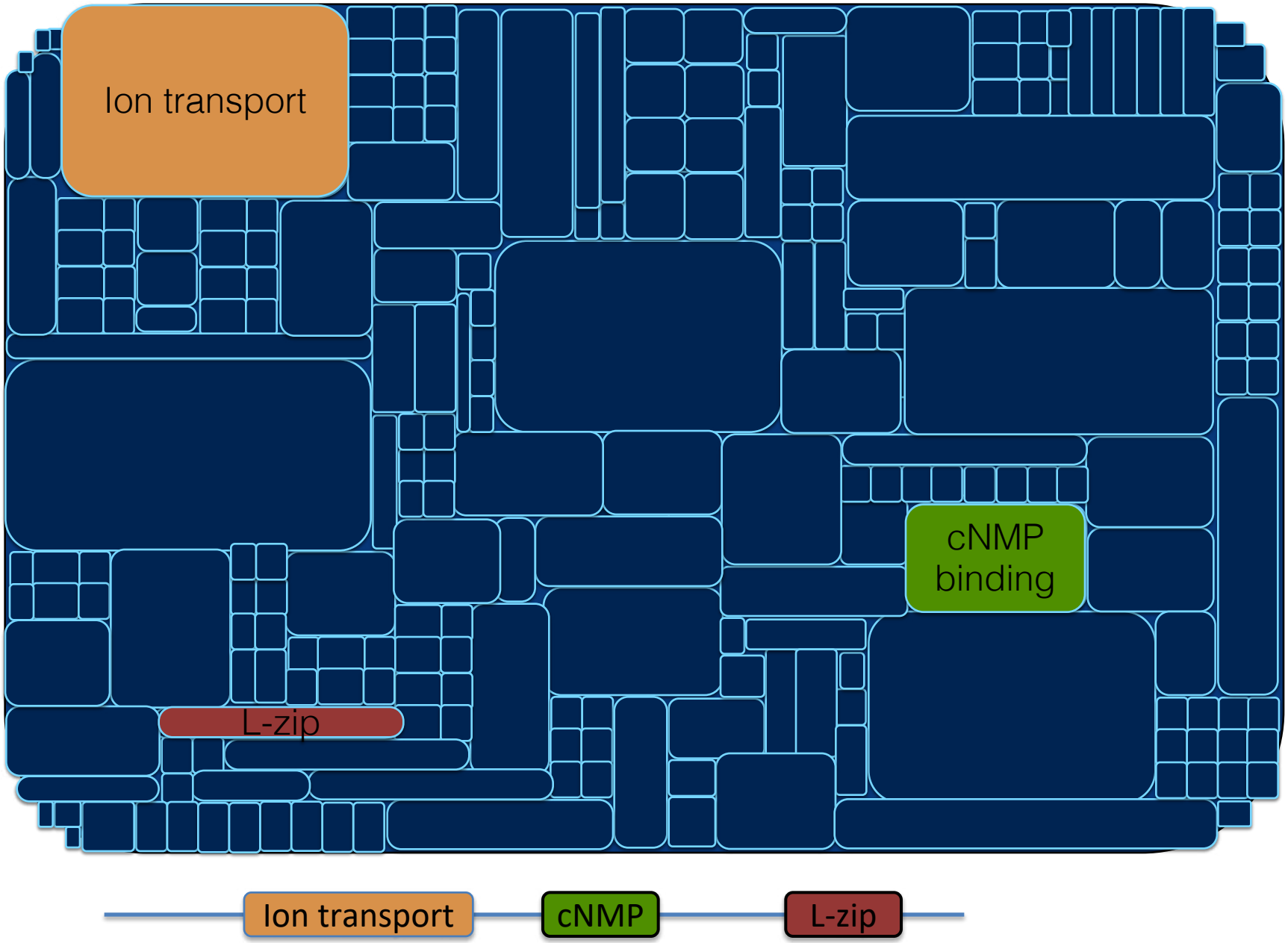
No limits, full-length proteins

Integration

CDD


InterPro
Protein sequence analysis & classification




Ion transport
cNMP binding
L-zip
Ion transport
cNMP
L-zip

How much of the sequence space is currently covered?

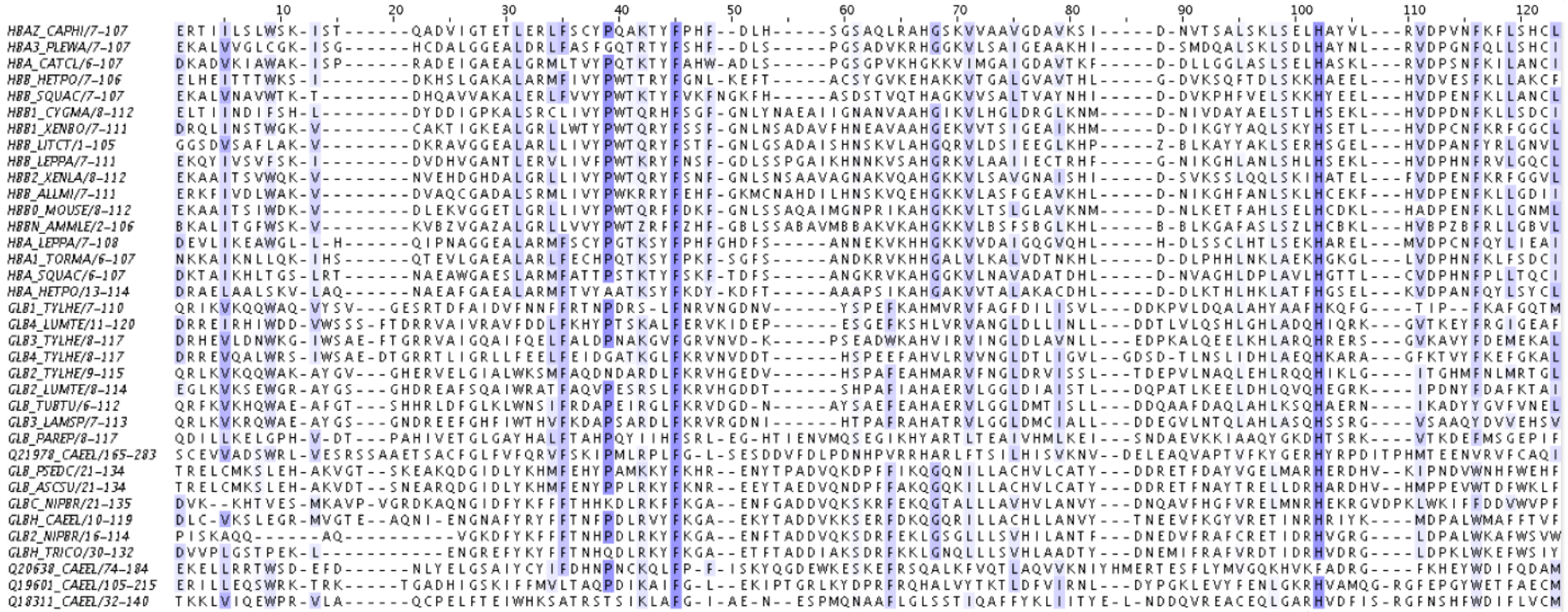
"The Uniprot reference proteomes set that we based Pfam 31.0 on contains 26.7 million sequences, which is an increase in size of 51% compared to when we made Pfam 30.0. Of the proteins in the Uniprot reference proteomes, 73% have a match to at least one Pfam entry, and 48% of all residues fall within a Pfam family."

https://xfam.wordpress.com/2017/03/08/pfam-31-0-is-released/

Sensitivity

Family databases build multiple sequence alignments between member sequences

MSAs



Functionally and/or structurally relevant?

Family power

```
Human: 1   MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
           MGLSDGEWQLVLNVWGKVEAD  GHGQEVLI LFK HPETL KFDKFK LKSE  MK SE
Mouse: 1   MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSE 60

Human: 61  DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120
           DLKKHG TVLTALG ILKKKG H AEI PLAQSHATKHKIPVKYLEFISE II VL   H
Mouse: 61  DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIEVLKKRH 120

Human: 121 PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
            GDFGADAQGAM KALELFR D A  YKELGFQG
Mouse: 121 SGDFGADAQGAMSKALELFRNDIAAKYKELGFQG 154
```

Sequence-sequence alignments

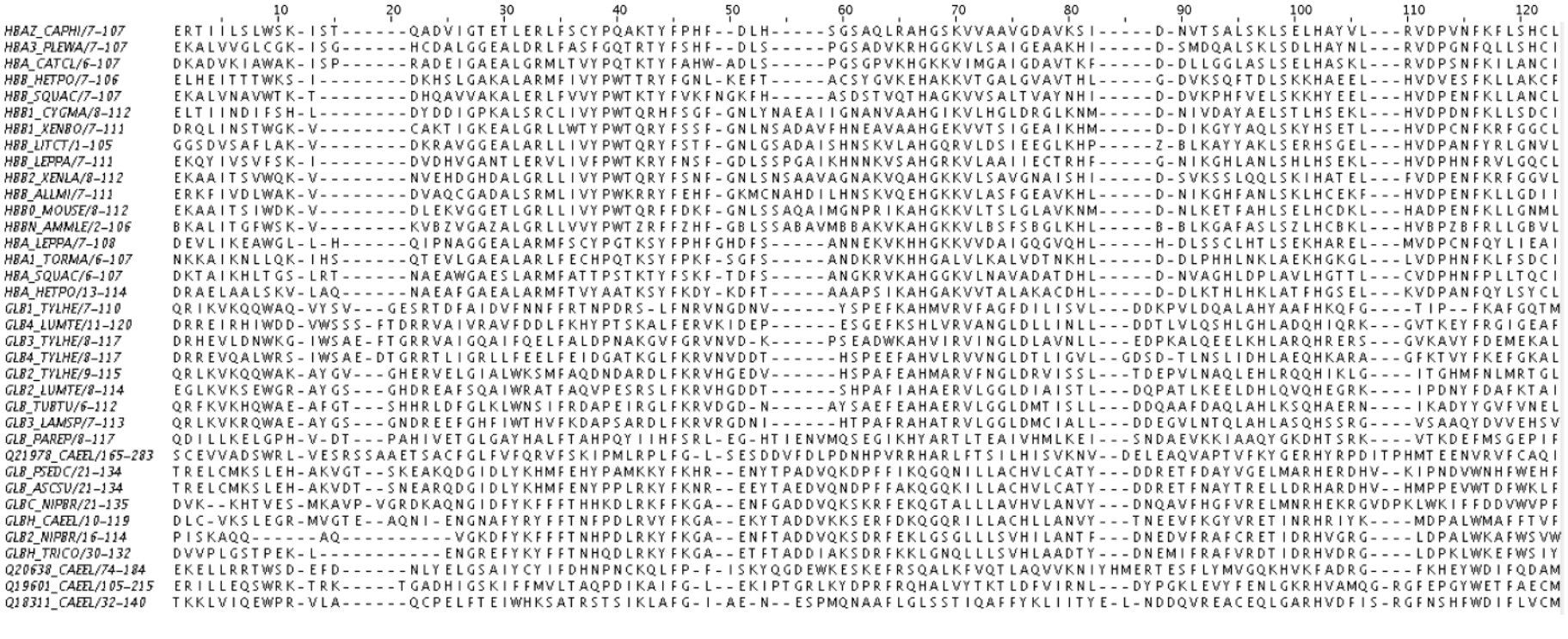
```
Human: 1    MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
           MGLSDGEWQLVLNVWGKVEAD  GHGQEVLI LFK HPETL KFDKFK LKSE  MK SE
Mouse: 1    MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSE 60

Human: 61   DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120
           DLKKHG TVLTALG ILKKKG H AEI PLAQSHATKHKIPVKYLEFISE II VL   H
Mouse: 61   DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIEVLKKRH 120

Human: 121  PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
             GDFGADAQGAM KALELFR D A  YKELGFQG
Mouse: 121  SGDFGADAQGAMSKALELFRNDIAAKYKELGFQG 154
```

BLOSUM62

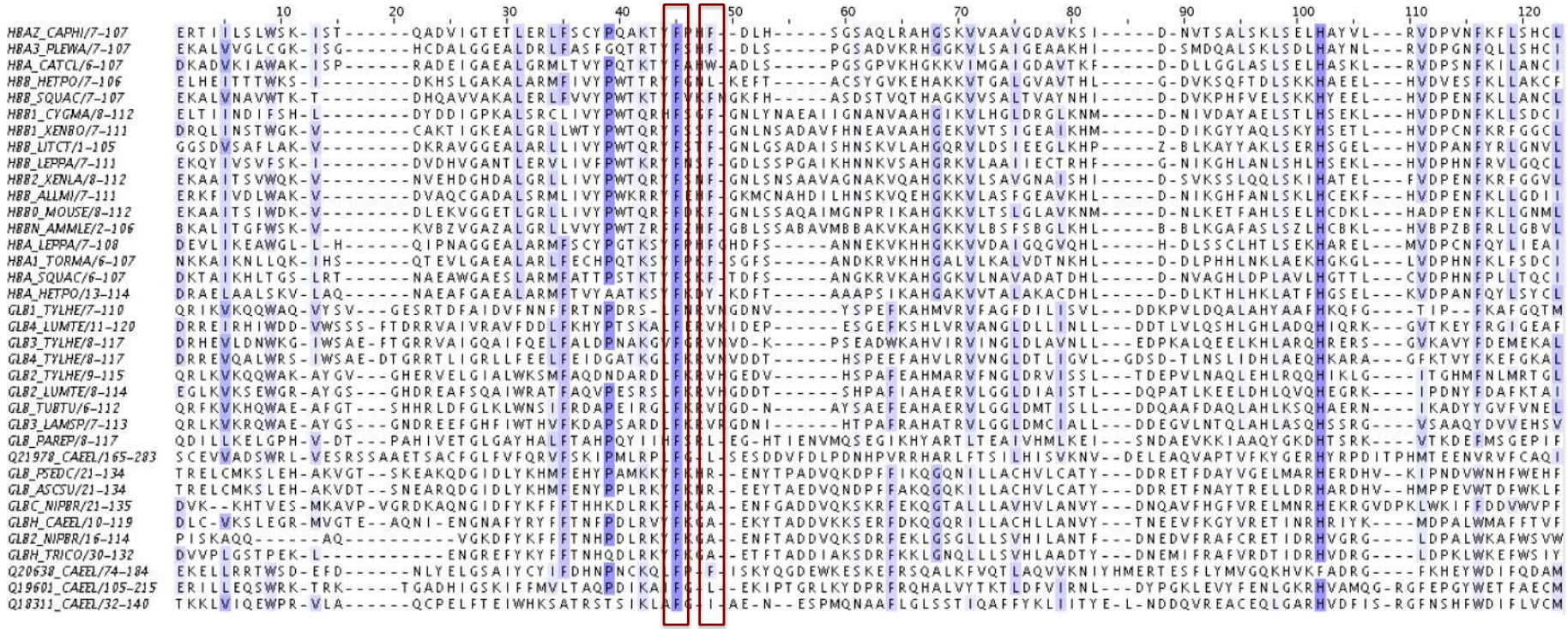
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

e.g. F->F Same score irrespective of position along protein sequence

Profile-sequence alignments



AA probabilities within the family are position specific

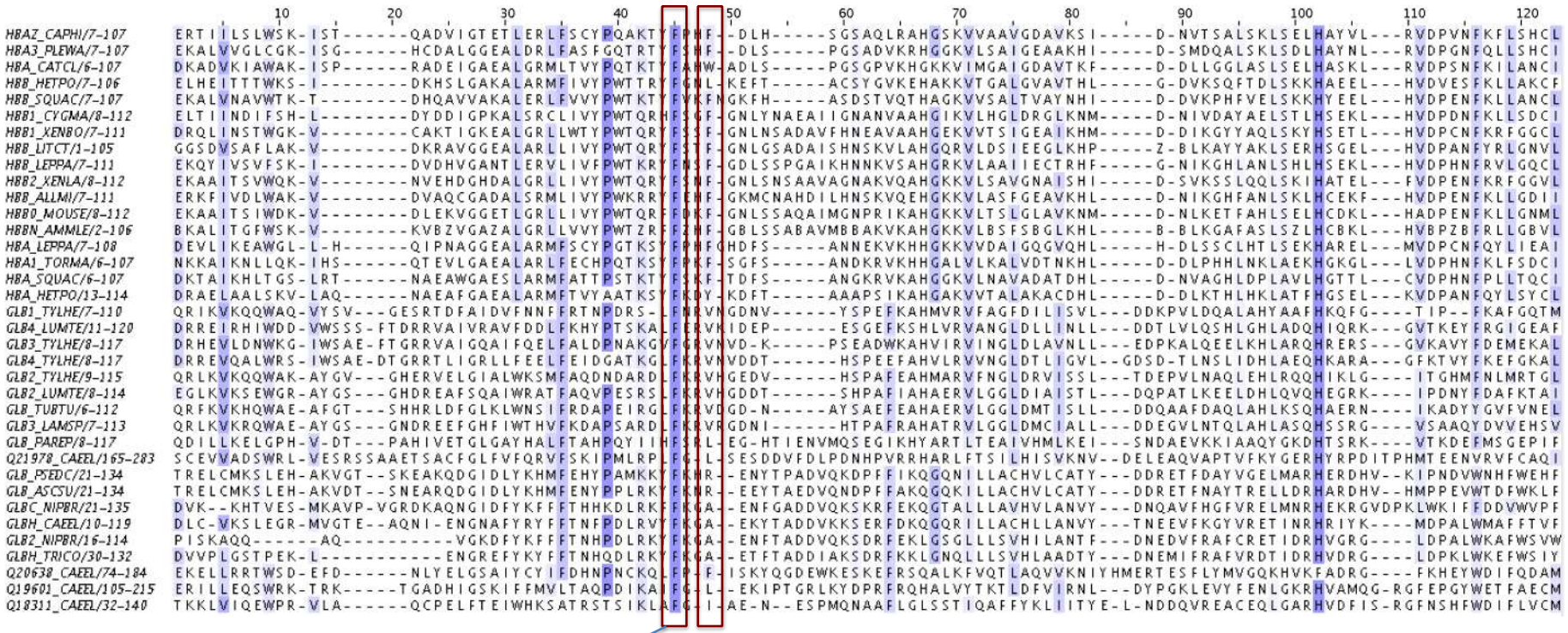


$$F_i \text{->} F_i \neq F_j \text{->} F_j$$

Position-specific scores

Marco Punta

score(i, aa)



seq> AWRTWEAPIFLKRYSTLPGGRAS…

Sequence-profile alignments

- Position specific substitution matrices
- profile-hidden Markov models

PLOS | COMPUTATIONAL BIOLOGY

EDITORIAL

New computational approaches to understanding molecular protein function

Jacquelyn S. Fetrow[1]*, Patricia C. Babbitt[2]

1 Office of the President, Albright College, Reading, Pennsylvania, United States of America, **2** Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, United States of America

* jfetrow@albright.edu

Defining function

Function is like beauty—its definition lies in the eye of the beholder or, in this case, the researcher. At the broadest level, we define organismal function—the function that the protein plays in the overall organism. This function can be observed by understanding the impact on the organism of deletion or mutation of the protein. Physiological function is the function the protein plays in pathways, such as metabolic or signaling pathways. Another level of function

Next:

- Practical: Homology based annotation of interactors' function in P-P interaction networks