



Protein homology and resources for protein families

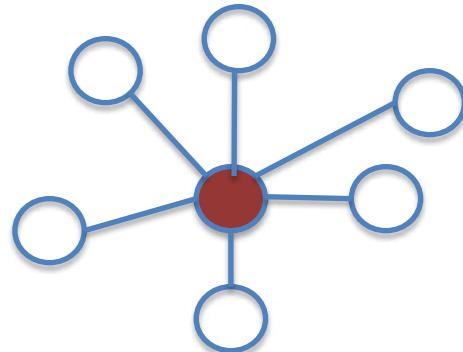
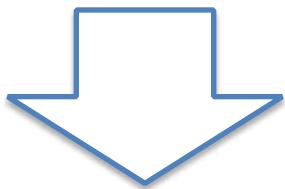
Marco Punta

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

Protein of interest



Experiments

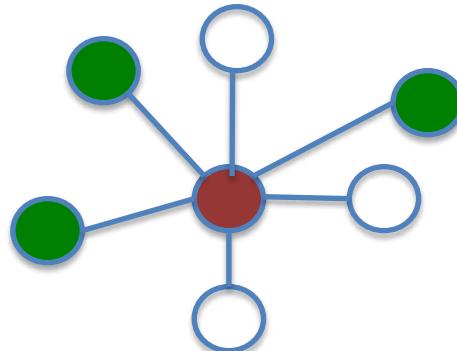
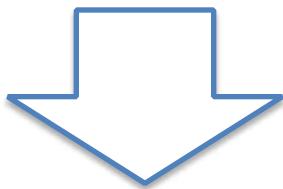


Interaction network

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

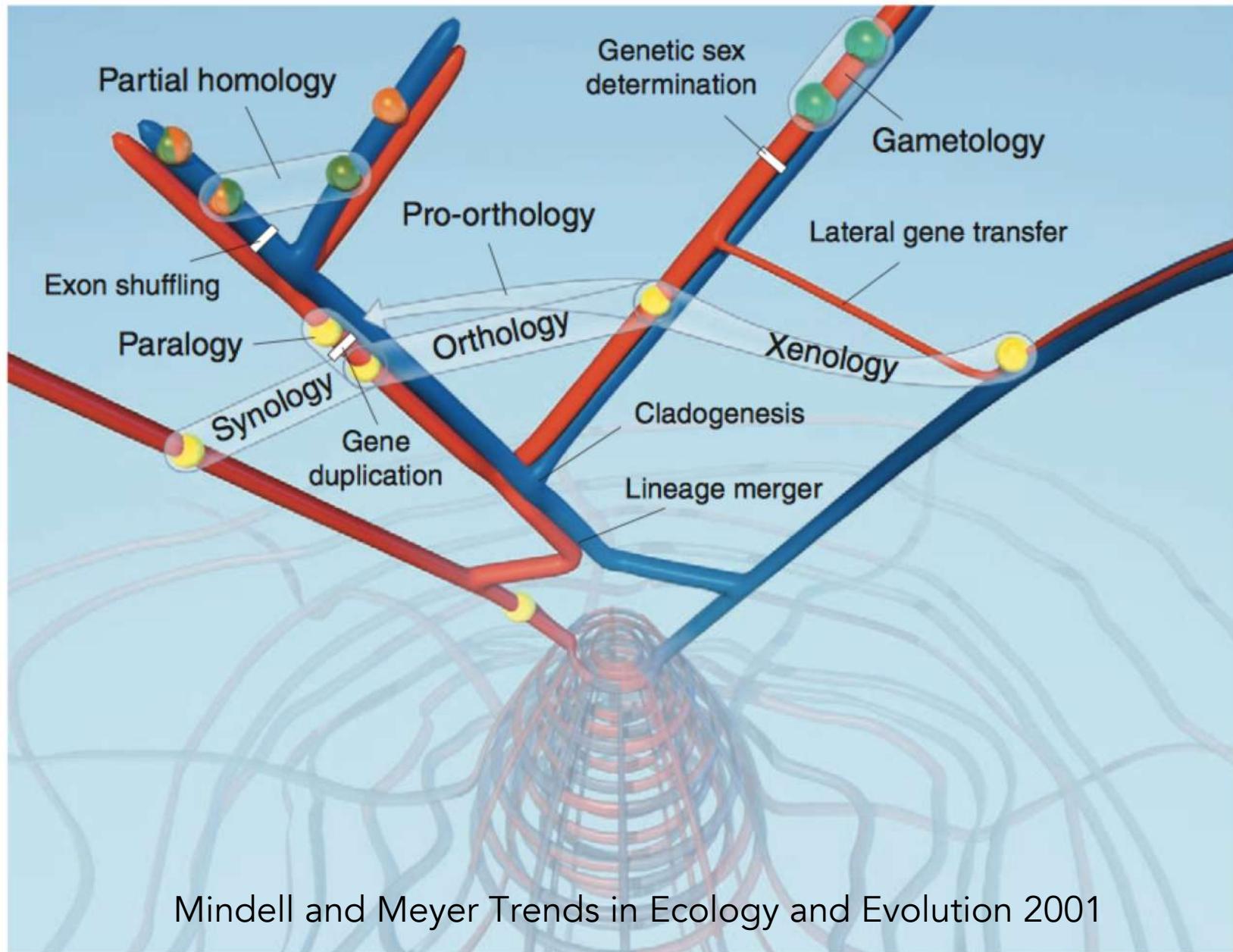
Protein Homology

Definition:

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

Origin of homology in proteins

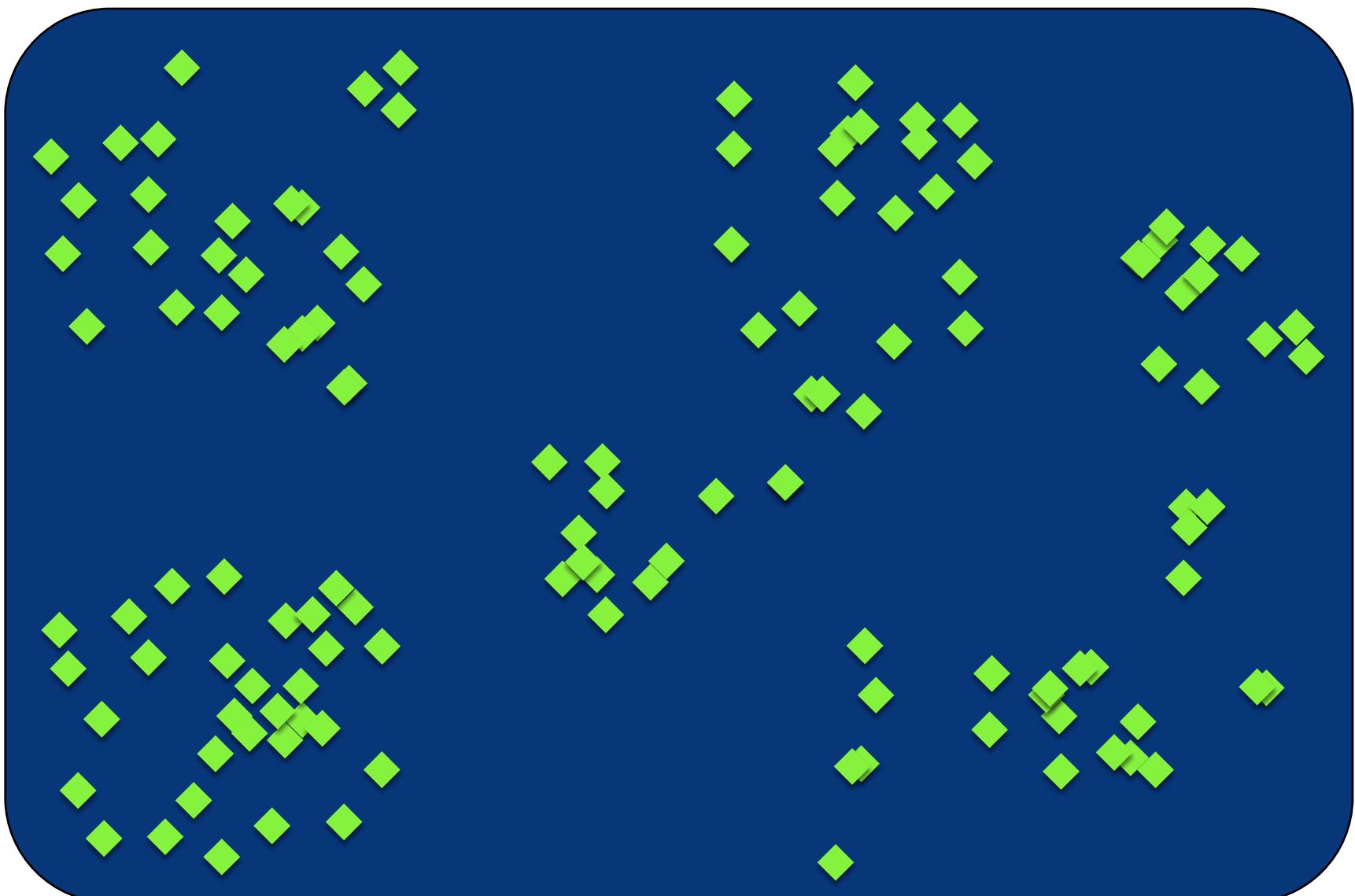
- Speciation (orthology)
- Gene duplication (paralogy)
- Horizontal gene transfer (xenology)
- Whole genome duplication (ohnology)
- etc. etc.



Definition:

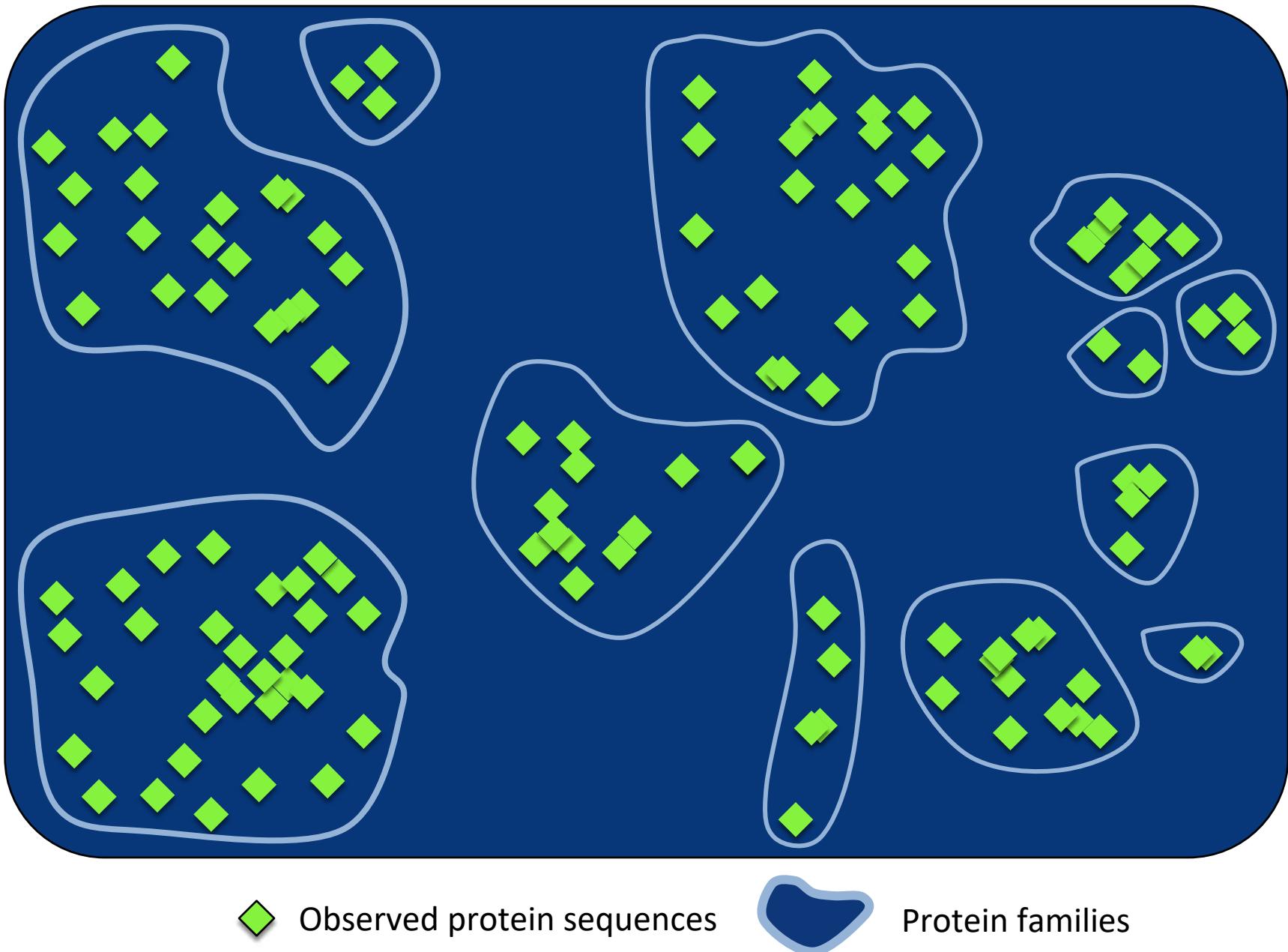
We call 'family' a group of evolutionary related proteins and/or protein regions

The sequence space



◆ Observed protein sequences

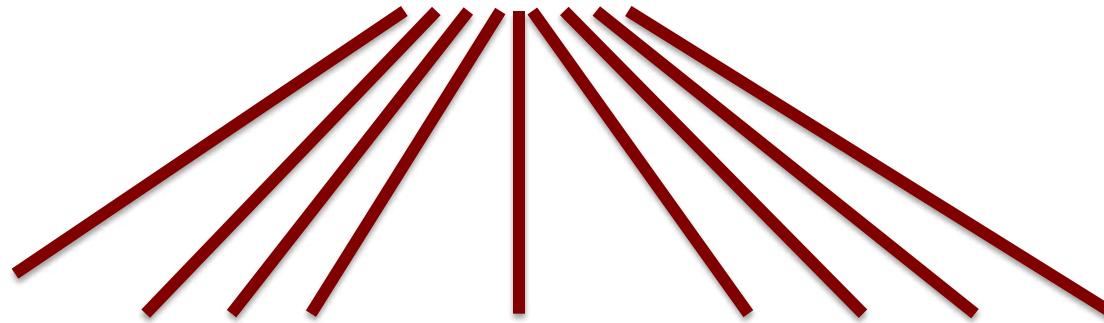
The sequence space



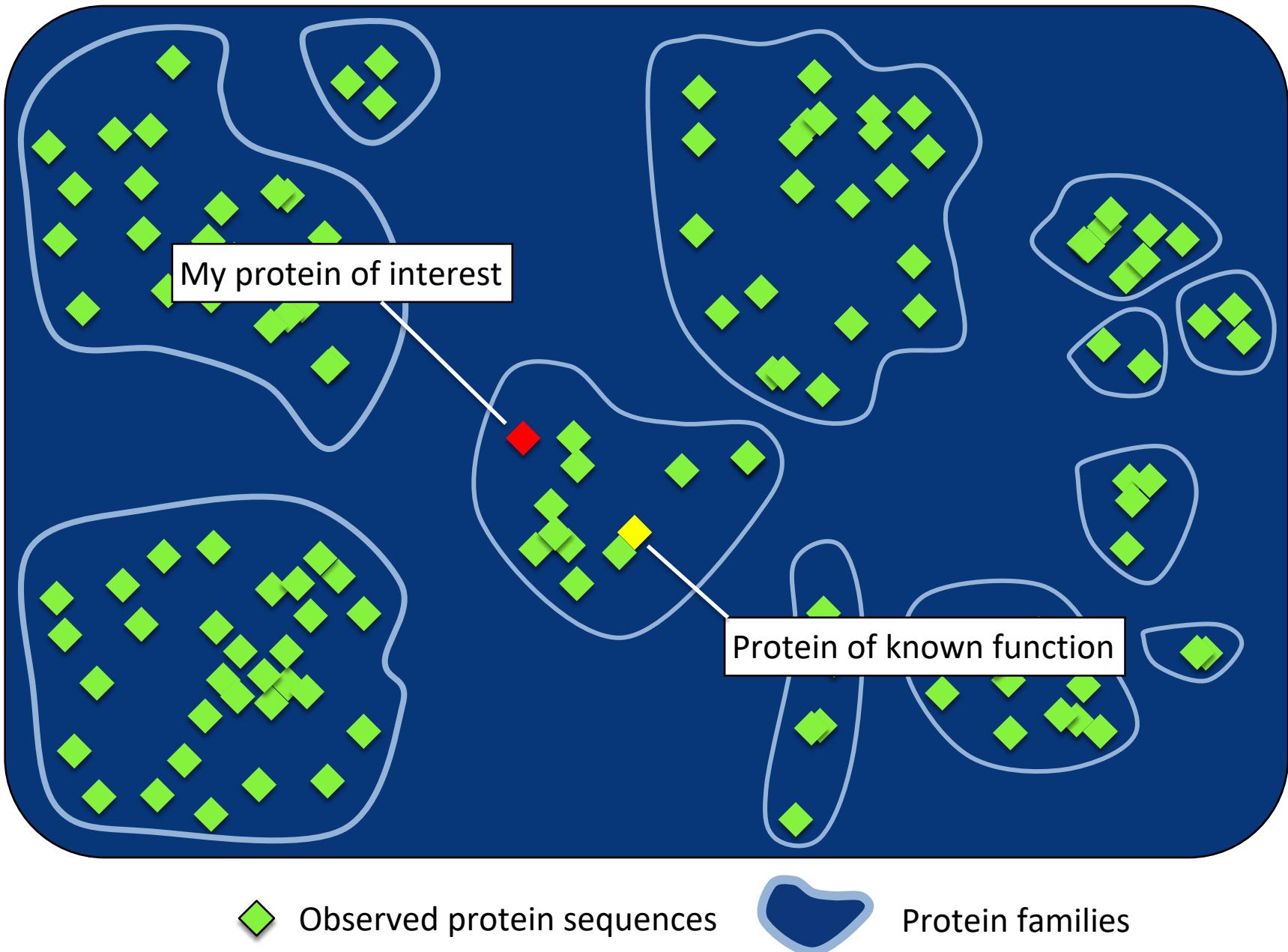
Homology: why interesting?

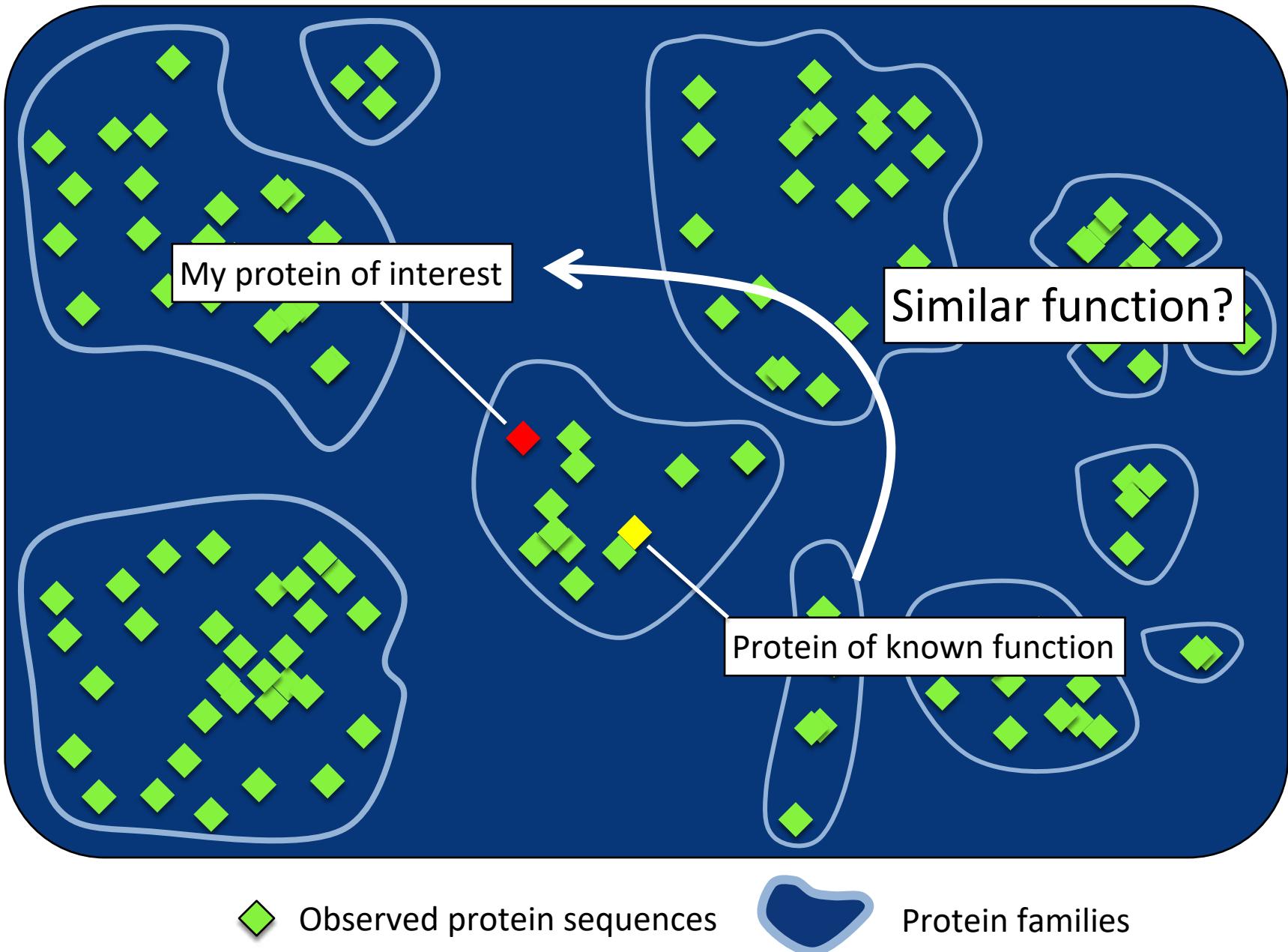
Homology: why interesting?

Common ancestor: one or more functions



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



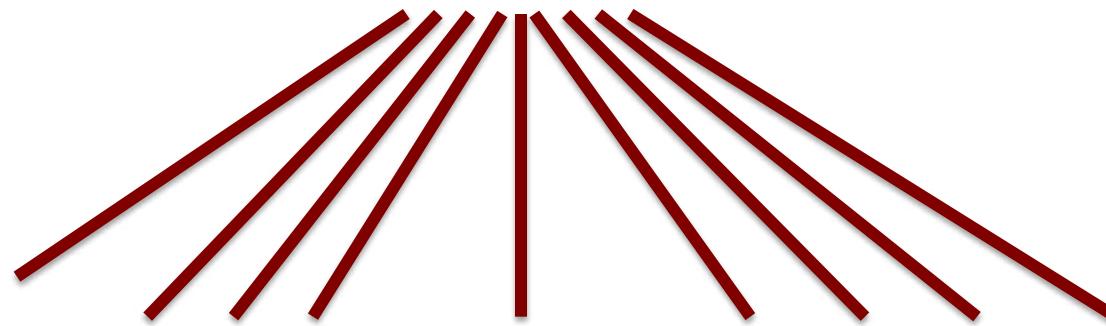


Detecting homology

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 MGLSDGEWQLVLNWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKA 60
MGLSDGEWQLVLNWGKVEAD GHGQEVL I LFK HPETL KFDKFK LKSE MK SE

Protein_2: 1 MGLSDGEWQLVLNWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKG 60

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

Protein_2: 61 DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIIEVLKKRH 120

Protein_1: 121 PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
GDFGADAQGAM KALELFR D A YKELGFQG

Protein_2: 121 SGDFGADAQGAMS KALELFRNDIAAKYKELGFQG 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 MGLSDGEWQLVLNWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKA 60
M LS V WGKV A G E L R F P T F F D S

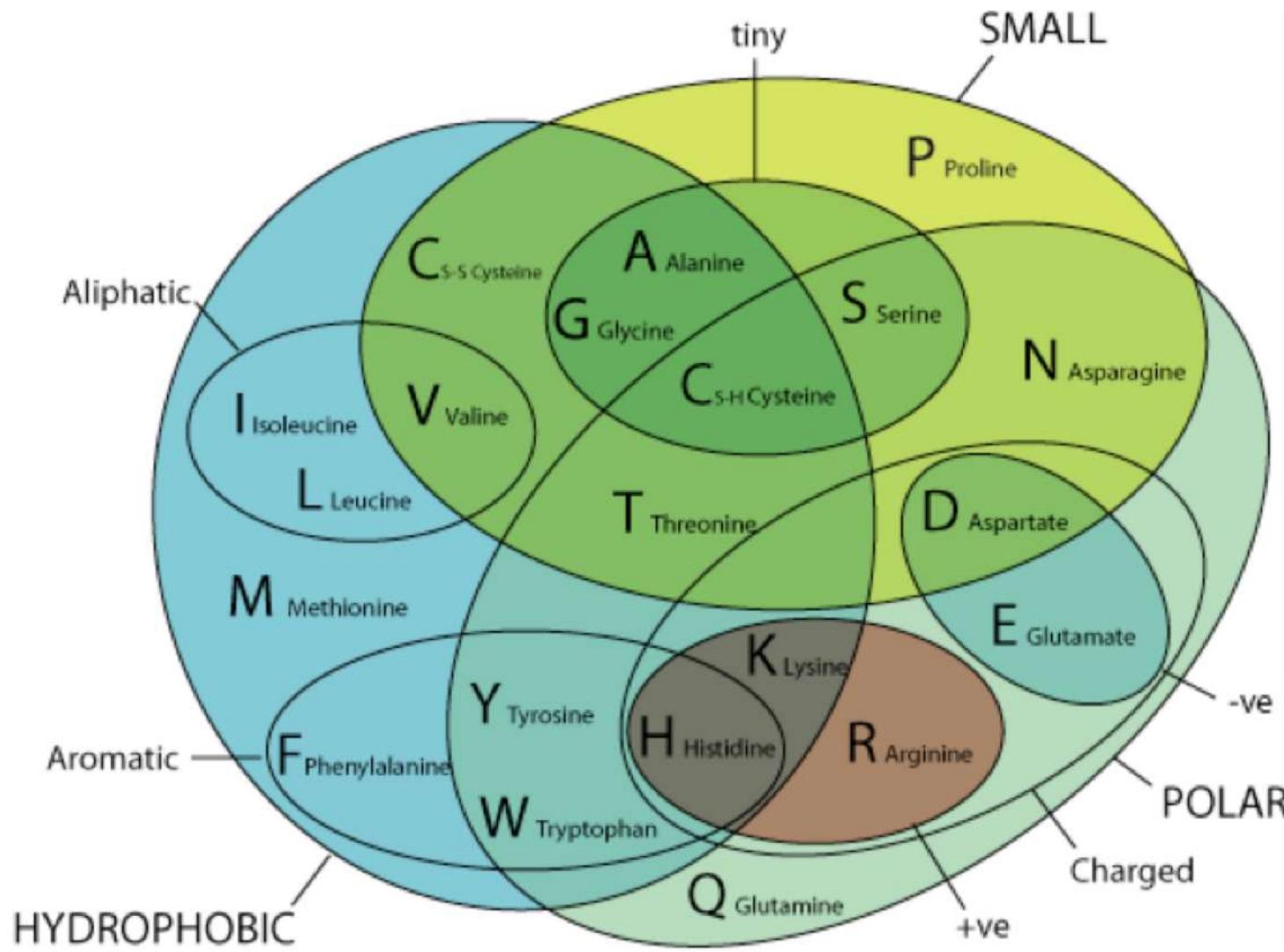
Protein 2 1 MVLSPADKTNVKAAWGKVGAGAEALERMFLSFPTTKTYFPHF-----DLSHGSA 54

Protein 1 61 DLKKHGATVLTALGGILKKKGHEAEIKPLAQSHATKHKI-PVKY 104
K H V AL L H A K P V

Protein 2 55 QVKGHSKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVPVN 99

What we need:

- Scoring system



What we need:

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

BLOSUM62 matrix

| | | | | | | | | | | | | | | | | | | | | |
|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| 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 |
| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |

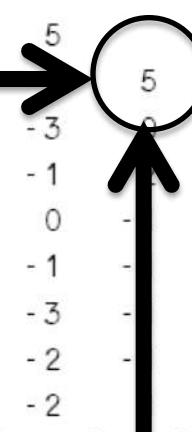
Given two protein sequences how do we know if they share excess sequence similarity? (i.e. that the observed similarity is not random)

Protein 1 MGLSDGEW...

Protein 2 MVLSPADK...



| | 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 | 0 | 2 | 1 | 2 | 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 | -2 | -3 | -2 | 2 | 7 | | |
| Val | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | -1 | -2 | -1 | -2 | 0 | -3 | 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...

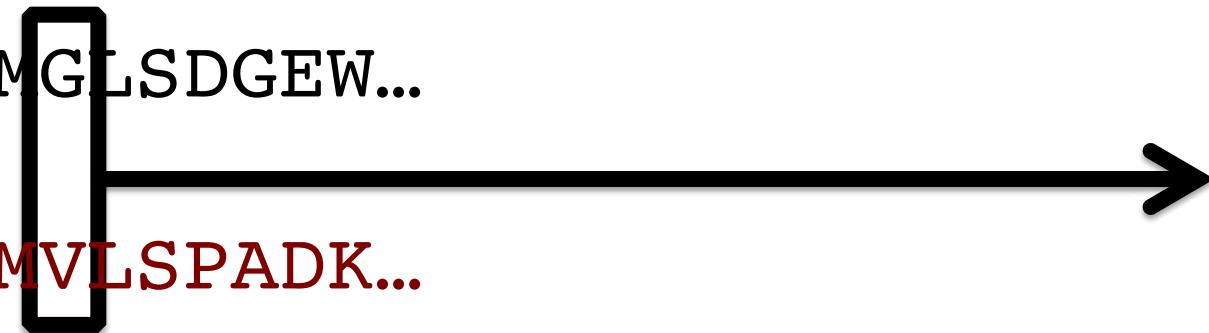


5

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 MGTSDGEW...

Protein 2 MVLSPADK...



| | | | | | | | | | | | | | | | | | | | | | | | | |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|--|--|--|--|
| 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 | 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 | 0 | 0 | 0 | 1 | 2 | -3 | -3 | 3 | 1 | -2 | 1 | -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 M**G**LSDGEW...

Protein 2 M**V**LSPADK...

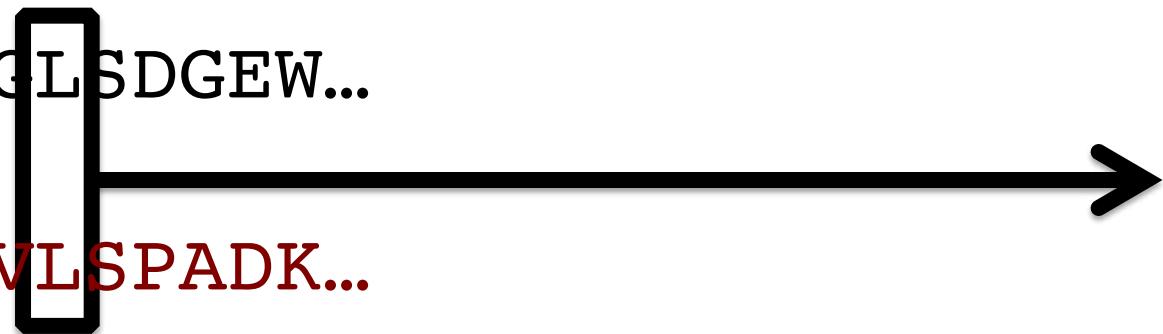


5-3

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

Protein 1 MGLSDGEW...

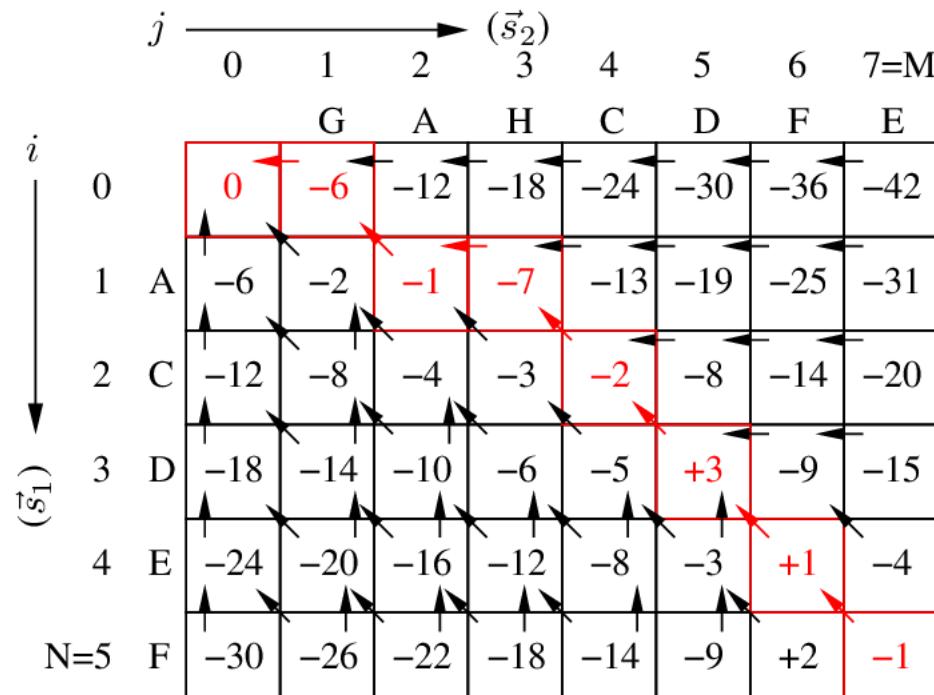
Protein 2 MVLSPADK...



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

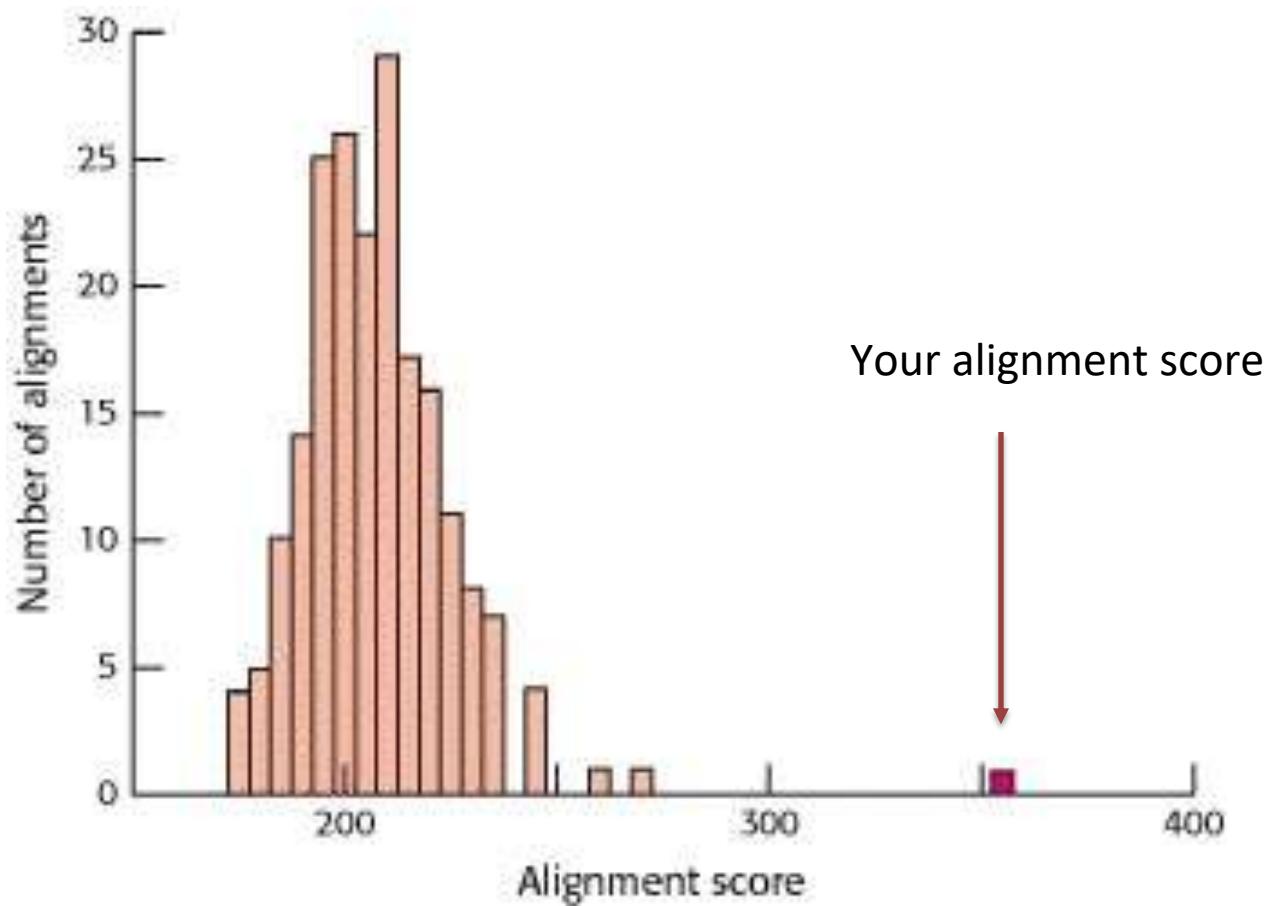
| | | | | | | |
|----|----|----|----|----|----|----|
| G | A | H | C | D | F | E |
| - | A | - | C | D | E | F |
| -6 | +5 | -6 | +5 | +5 | -2 | -2 |

Note: not same scoring matrix as before

What we need:

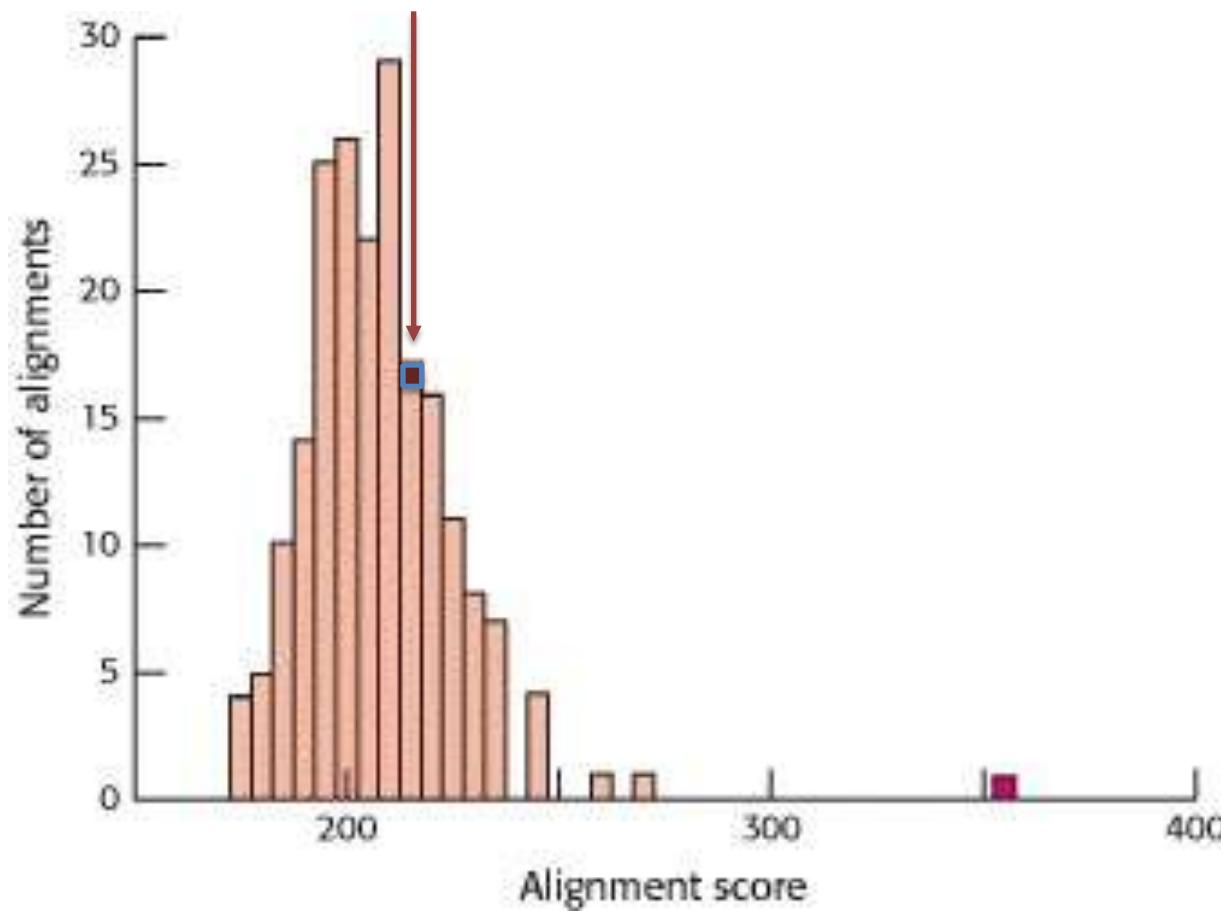
- Scoring system => empirically derived substitution matrices (PAMs, BLOSUMs,...)
- Efficient way to find highest scoring alignments => dynamic programming (Needleman-Wunsch, Smith-Waterman,...)
- Way to decide whether top score is high enough to infer homology (significance) => E-value, ...

Significance



Significance

Your alignment score



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

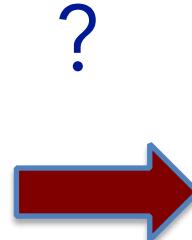
Excess sequence
similarity

?



Homology

Excess sequence
similarity

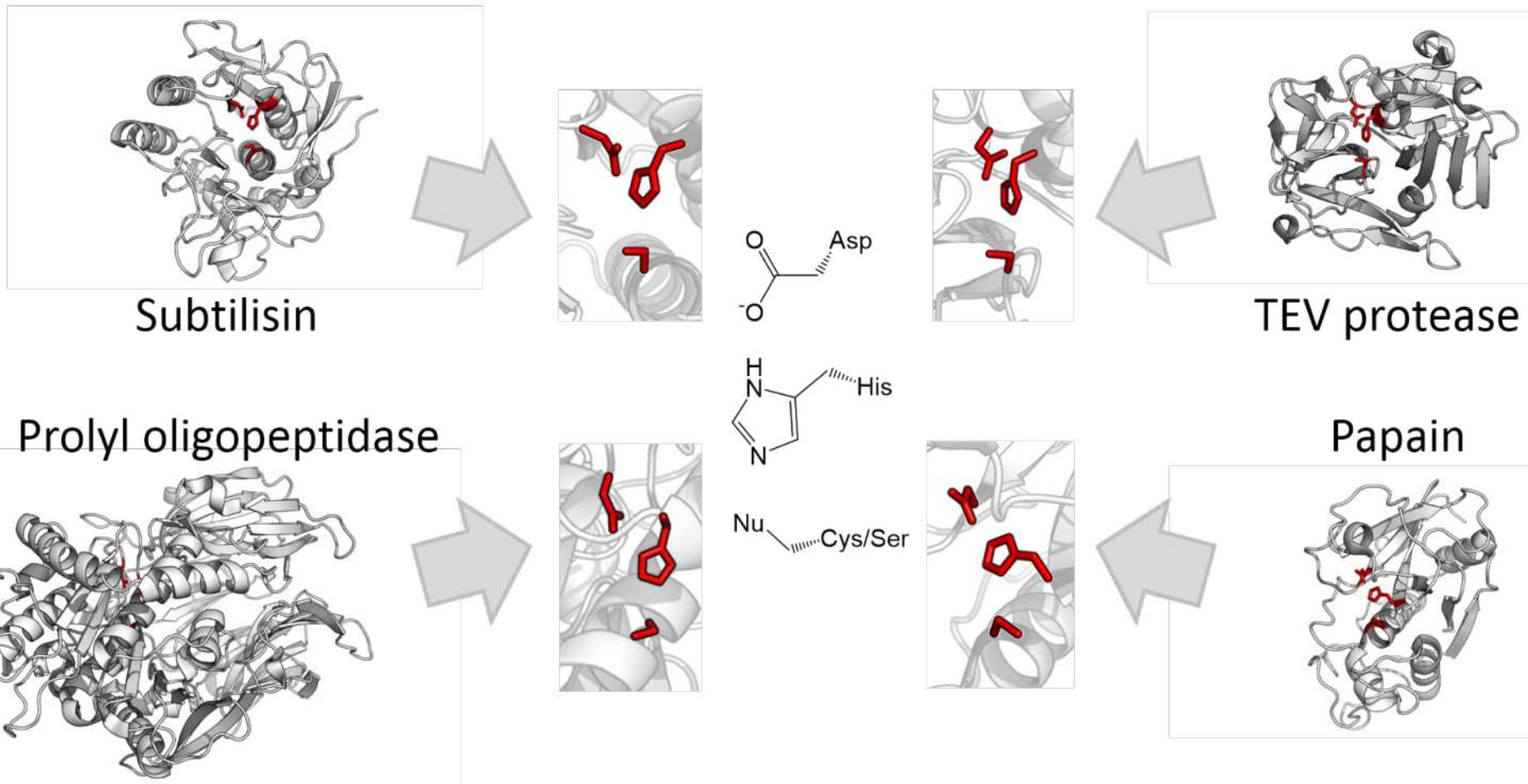


Homology

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)



"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



“[...]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.”

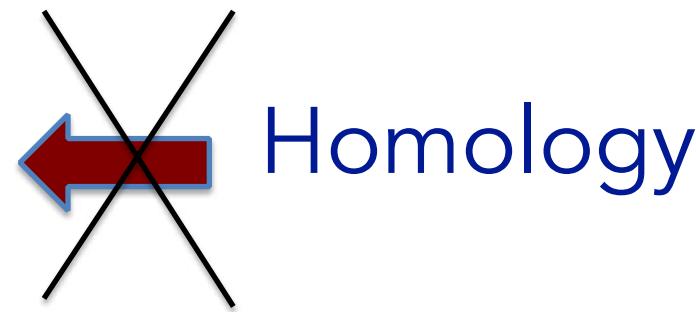
Koonin and Galperin (2003)

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



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](https://doi.org/10.1002/0471250953.bi0301s42)

PMCID: PMC3820096

NIHMSID: NIHMS519883

An Introduction to Sequence Similarity (“Homology”) Searching

[William R. Pearson¹](#)

[Author information ▶](#) [Copyright and License information ▶](#)

The publisher's final edited version of this article is available at [Curr Protoc Bioinformatics](#)

See other articles in PMC that [cite](#) the published article.

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.

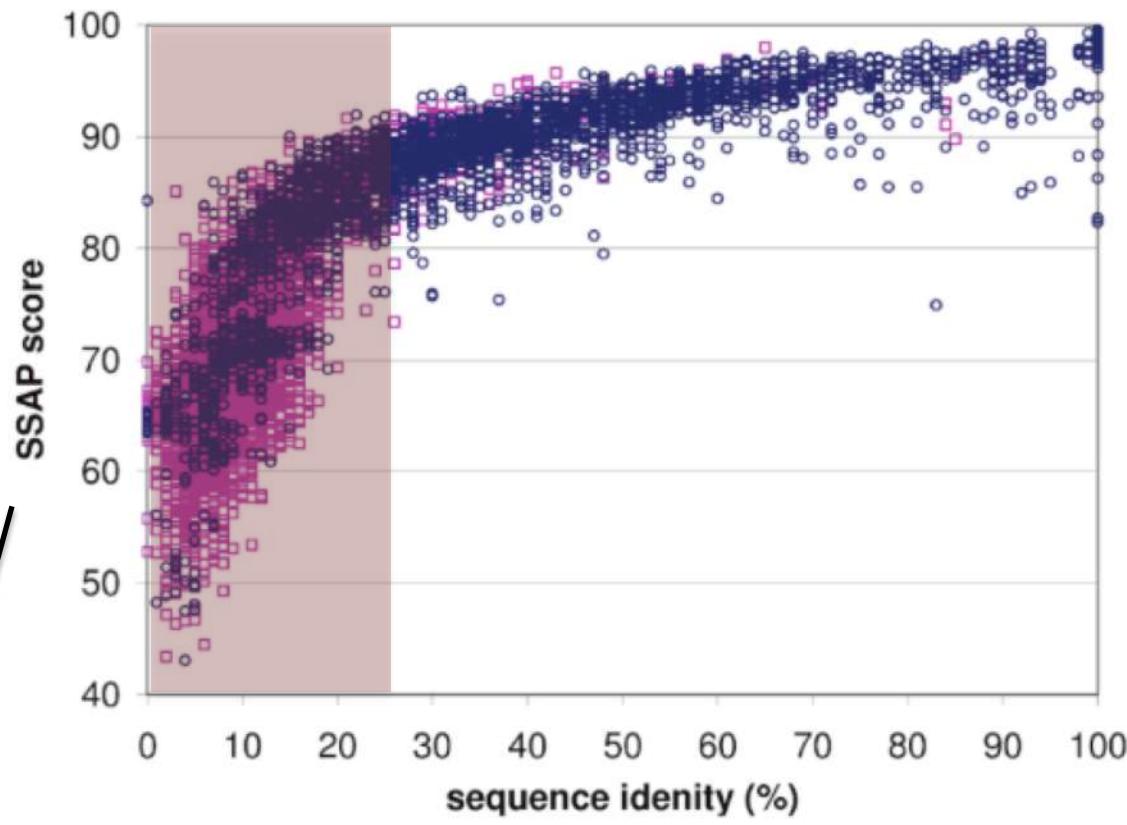
Illergård K¹, Ardell DH, Elofsson A.

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

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...

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

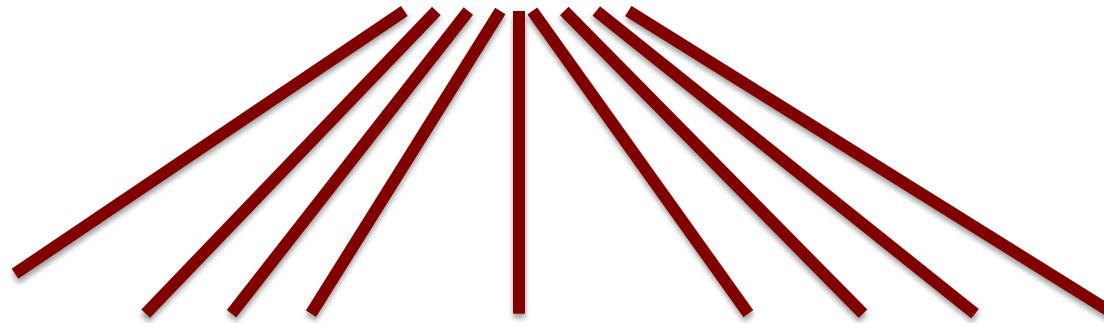
Here, we will focus on homology detection via
excess sequence similarity

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?

CbiF

Catalytic activity



Methyltransferase

precorrin-4
methyltransferase activity**CbiA**

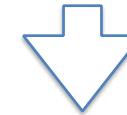
Catalytic activity



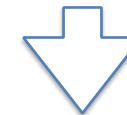
Ligase activity

cobyrinic acid a,c-diamide
synthase activity**CbiJ**

Catalytic activity



Reductase activity



precorrin-6A reductase activity

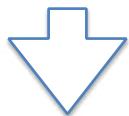
How to compare function?

CbiF

Catalytic activity



Methyltransferase

precorrin-4
methyltransferase activity**CbiA**

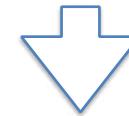
Catalytic activity



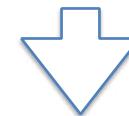
Ligase activity

cobyrinic acid a,c-diamide
synthase activity**CbiJ**

Catalytic activity



Reductase activity



precorrin-6A reductase activity

Molecular function

How to compare function?

CbiF

CbiA

CbiJ

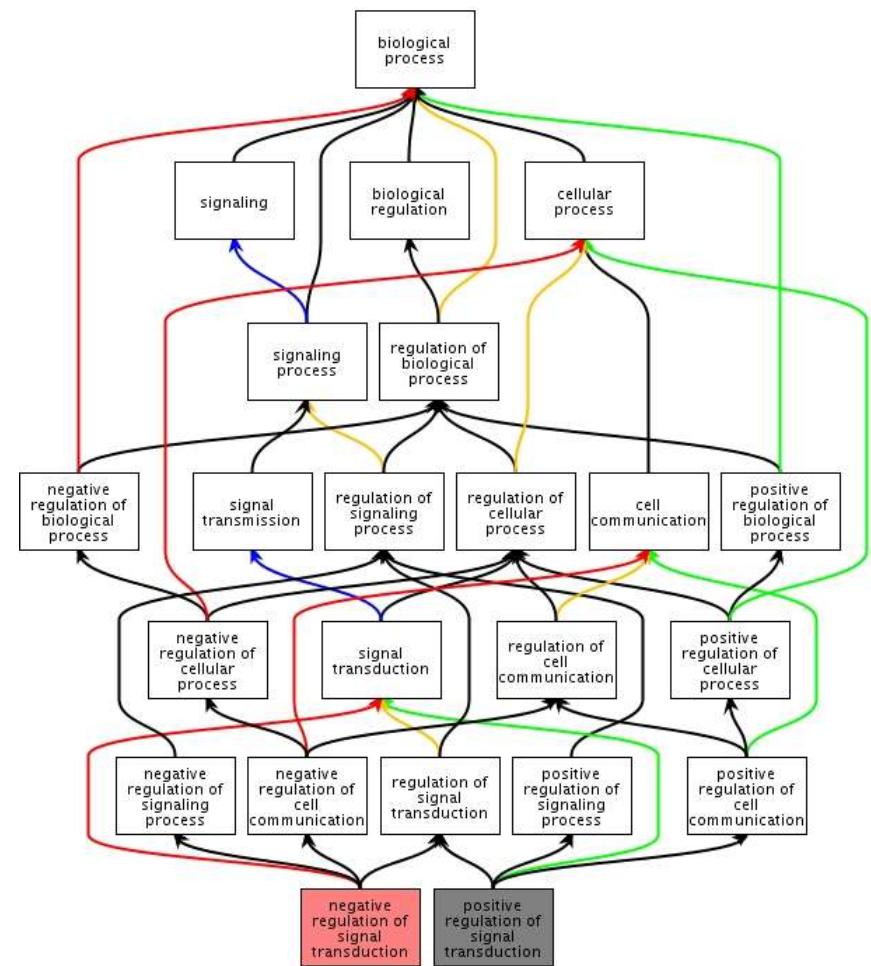


Cobalamin
biosynthetic process

Biological process

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

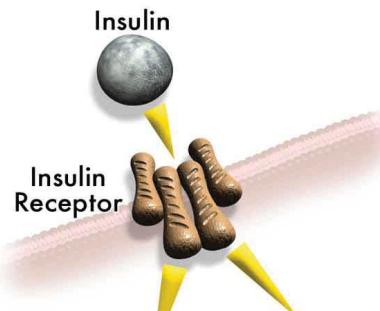
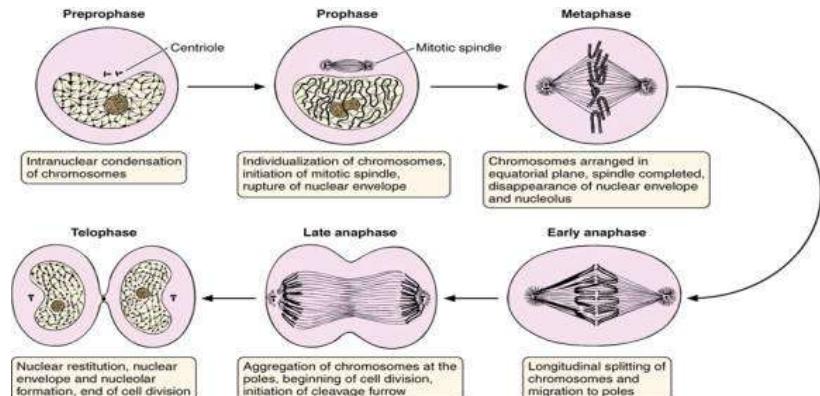


www.ebi.ac.uk/QuickGO

GO: 3 ontologies in 1

1. Molecular Function

An elemental activity or task or job

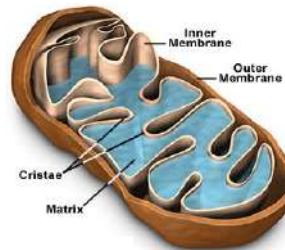


- protein kinase activity
- insulin receptor activity

2. Biological Process

A commonly recognised series of events

- cell division



- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

3. Cellular Component

Where a gene product is located

Top-level EC numbers^[5]

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

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 | MGLSDGEWQLVNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFHKHLKSEDEMKA Sbjct 1 | E +++++W K+ G E L RLF HP+T F F D S | | | 60 |
| | MSLTKTERTIIIVSMWAKISTQADTIGTETLERLFLSHPQTKTYFPHF-----DLHPGSA | | | | 54 |
| Query 61 | DLKKHGATVL TALGGILKKKGHEAEIKPLAQSHATKHKI-PVKYLEFISECIIQVLQSK Sbjct 55 | L+ HG+ V+ A+G +K + L++ HA ++ PV + + +S C++ L ++ | | | 119 |
| | QLRAHGSKVVAAVGDAVKSIDDIGGALSKLSELHAYILRVDPVNF-KLLSHCILLVT LAAR | | | | 113 |
| Query 120 | HPGDFGADAQGAMNKALELFRKDMASNYK Sbjct 114 | P DF A+A A +K L + + Y+ FPADFTAEAAWDKFLSVVSSVLTEKYR | 148 | 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 | |
| Function | | | | | | | | | | | | | |
| 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:ENSRNOP0000006184 | 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:ENSRNOP0000006184 | 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:ENSRNOP0000006184 | 9606 | 20140913 | Ensembl | |
| UniProtKB | P02144 | MB | | GO:0042542 | response to hydrogen peroxide | P | IEA | Ensembl Compara | Ensembl:ENSRNOP0000006184 | 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 | |
| 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:ENSMUSP0000020531 | 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:ENSMUSP0000020531 | 9606 | 20140913 | Ensembl | |
| 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 Product By | 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:ENSRNOP0000006184 | 9606 | 20140913 | Ensembl | |
| UniProtKB | P02144 | MB | GO:0015671 | oxygen transport | P | IEA | InterProGO | InterPro:IPR002335 InterPro:IPR012292 | 9606 | 20140913 | InterPro | |
| UniProtKB | P02144 | MB | GO:0015671 | oxygen transport | P | IEA | Ensembl Compara | Ensembl:ENSRNOP0000006184 | 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:ENSRNOP0000006184 | 9606 | 20140913 | Ensembl | |
| UniProtKB | P02144 | MB | GO:0042542 | response to hydrogen peroxide | P | IEA | Ensembl Compara | Ensembl:ENSRNOP0000006184 | 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 | InterProGO | InterPro:IPR000971 InterPro:IPR012292 | 9606 | 20140913 | InterPro | |
| UniProtKB | P02144 | MB | GO:0019825 | oxygen binding | F | IEA | InterProGO | InterPro:IPR002335 InterPro:IPR012292 | 9606 | 20140913 | InterPro | |
| UniProtKB | P02144 | MB | GO:0019825 | oxygen binding | F | IEA | Ensembl Compara | Ensembl:ENSRNOP0000006184 | 9606 | 20140913 | Ensembl | |
| UniProtKB | P02144 | MB | GO:0020037 | heme binding | F | IEA | InterProGO | 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 Product By | 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 | InterProGO | 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 | |
| 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 | InterProGO | InterPro:IPR000971 InterPro:IPR002338 InterPro:IPR002340 InterPro:IPR012292 | 9606 | 20140913 | InterPro | |
| UniProtKB | P02008 | HBZ | GO:0005515 | protein binding | F | IPI | PMID:115871 | InterPro:IPR000971 IPR002338 IPR002340 IPR012292 | 9606 | 20140914 | IntAct | |
| UniProtKB | P02008 | HBZ | GO:0005515 | protein binding | F | IPI | PMID:6 | InterPro:IPR000971 IPR002338 IPR002340 IPR012292 | 9606 | 20140914 | IntAct | |
| UniProtKB | P02008 | HBZ | GO:0019825 | oxygen binding | F | IEA | InterProGO | InterPro:IPR002338 InterPro:IPR002340 InterPro:IPR012292 | 9606 | 20140913 | InterPro | |
| UniProtKB | P02008 | HBZ | GO:0020037 | heme binding | F | IEA | InterProGO | 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 | InterProGO | 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

SAME

SAME

Conservation of function (EC numbers) in homologous proteins

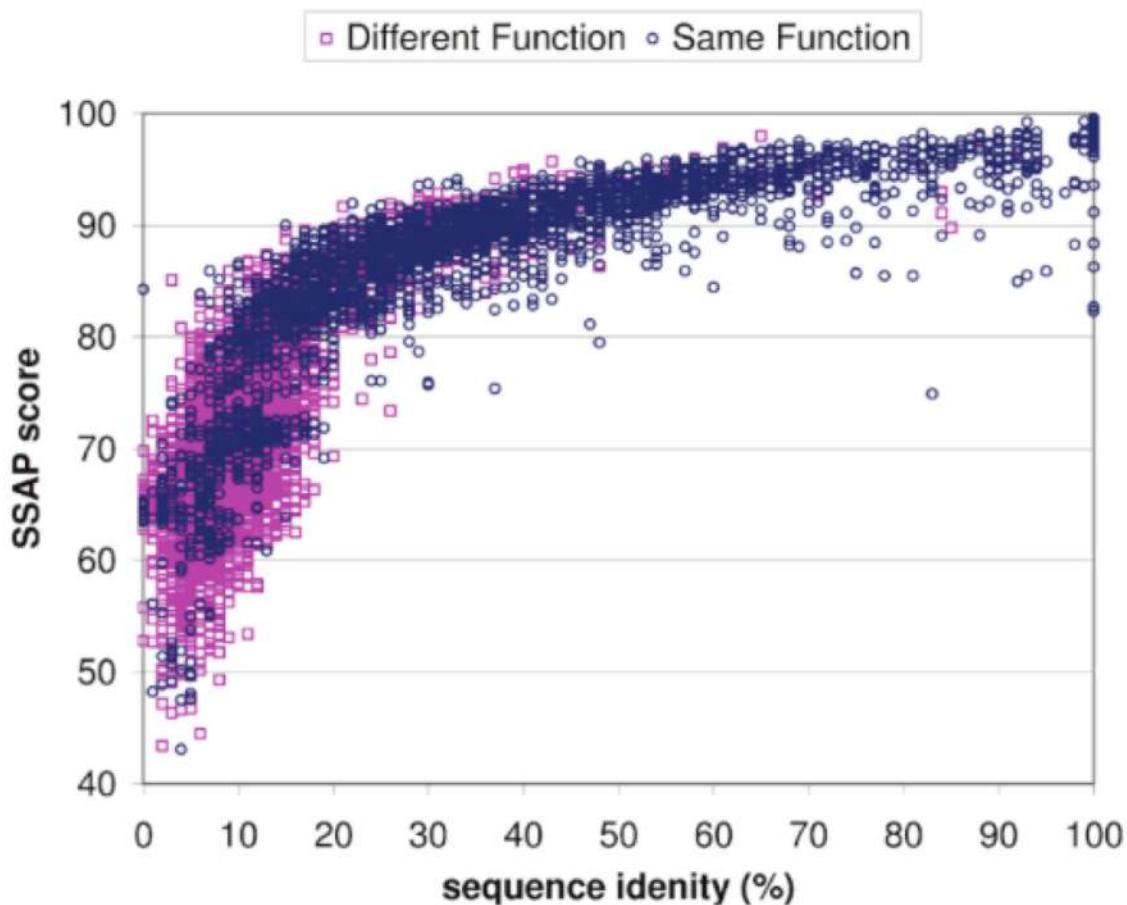
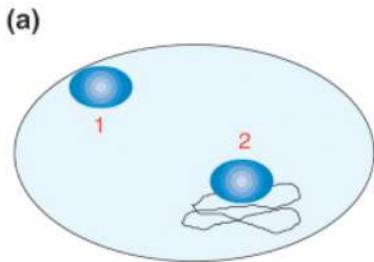


Fig. 7.8 Scatter plot showing the relationship between sequence, structure, and function of all homologues in enzyme superfamilies. Relatives having the same EC classification number are shown in blue. Those with different EC numbers are shown in pink.

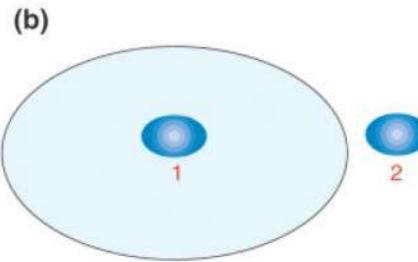
Conservation of function in homologs

- 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.

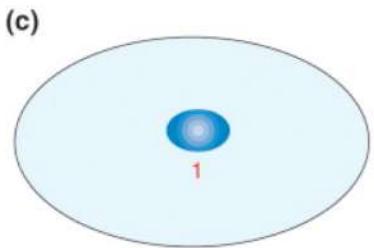
Moonlighting proteins



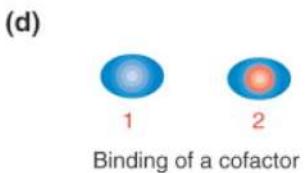
Different locations within the cell



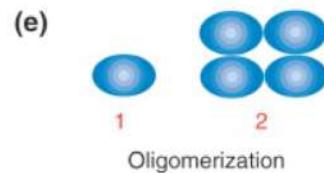
Inside and outside the cell



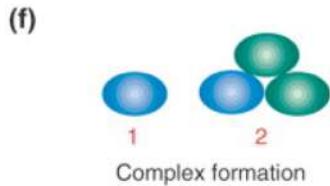
Expression by different cell types



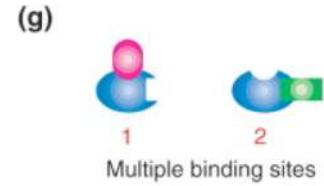
Binding of a cofactor



Oligomerization



Complex formation

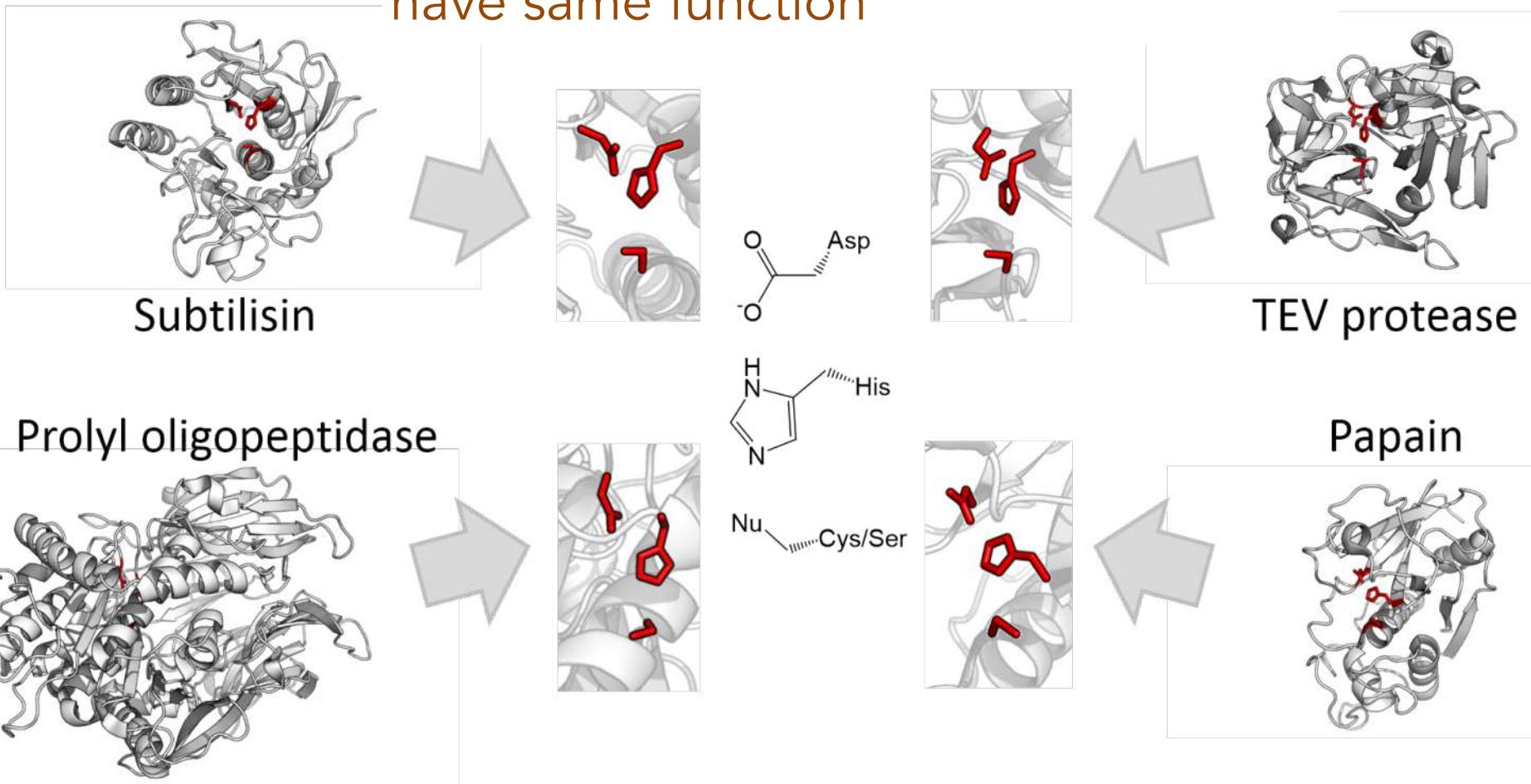


Multiple binding sites

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)

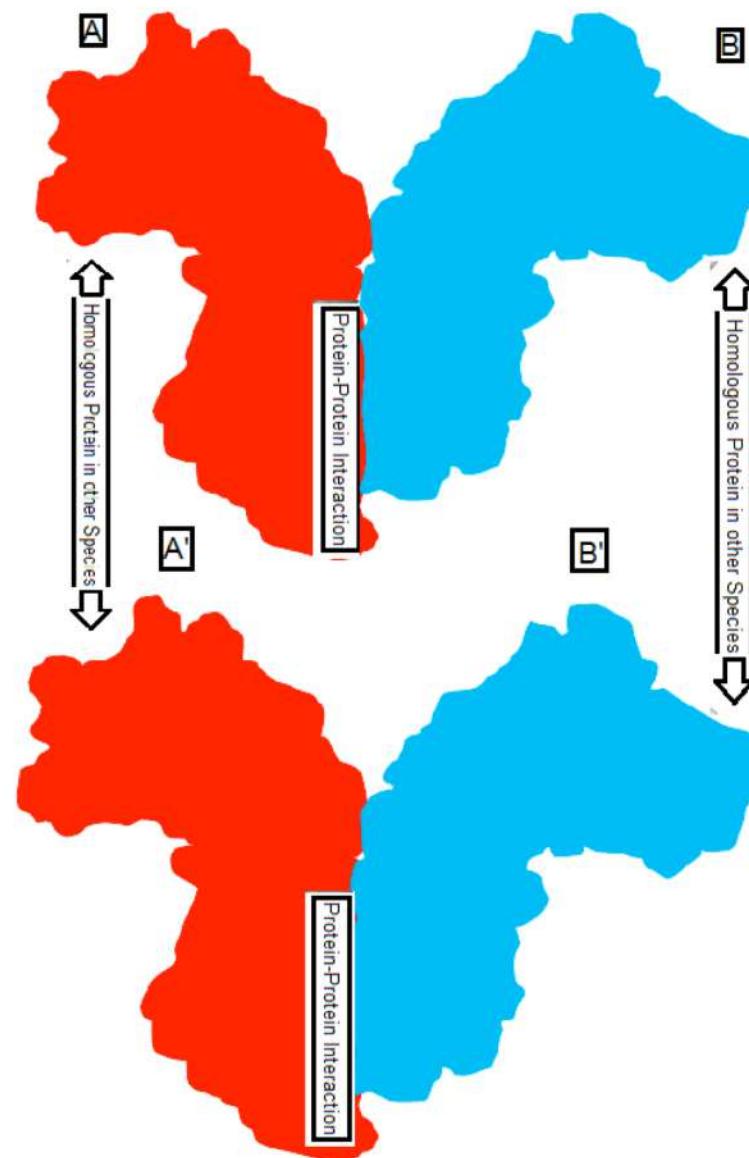
Note: homology not necessary to have same function



"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

How about homology and
protein-protein interactions?

Interologs



Interologs

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>

| Article | Authors | Metrics | Comments | Related Content |
|---------|---------|---------|----------|-----------------|
| ▼ | | | | |

Abstract

Author Summary

Introduction

Results/Discussion

Materials and Methods

Supporting Information

Acknowledgments

Author Contributions

References

Reader Comments (1)

Media Coverage (0)

Figures

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).

Control interactive selection and zoom in the structure panel

Basic swap between positions 31 and 35

Residues Focus: R31A, N35A, E41A, F31B, E35B, A178

Invariant position at E35

Multiple sequence Alignment:

| gi/Index | R31A | N35A | E41A | F31B | E35B | A178 |
|-------------------------------|------|------|------|------|------|------|
| Anopheles_gambiae | H | - | - | - | - | - |
| Aedes_aegypti | H | - | - | - | - | - |
| Drosophila_melanogaster | H | - | - | - | - | - |
| Ixodes_scapularis | H | - | - | - | - | - |
| Acyrtosiphon_pistaci | H | - | - | - | - | - |
| Schistosoma_mansoni | H | - | - | - | - | - |
| Trichoplax_adhaerens | S | - | - | - | - | - |
| Loa_loa | S | - | - | - | - | - |
| Dictyostelium_discoidium | S | - | - | - | - | - |
| Malassezia_globosa | S | - | - | - | - | - |
| Ustilago_maydis | S | - | - | - | - | - |
| Laccaria_bicolor | S | - | - | - | - | - |
| Schizophyllum_comune | S | - | - | - | - | - |
| Schizosaccharomyces_pombe | S | - | - | - | - | - |
| Schizosaccharomyces_japonicus | S | - | - | - | - | - |
| Tuber_melanoporum | R | - | - | - | - | - |
| Magnaporthe_grisea | R | - | - | - | - | - |
| Uncinocarpus_reesii | R | - | - | - | - | - |
| Coccidioides_posadasii | R | - | - | - | - | - |
| Trichophyton_verrucosum | R | - | - | - | - | - |
| Aspergillus_nidulans | R | - | - | - | - | - |
| Aspergillus_fumigatus | R | - | - | - | - | - |
| Talaromyces_stipitatus | R | - | - | - | - | - |
| Verticillium_albo-atrum | R | - | - | - | - | - |
| Gibberella_zaeae | R | - | - | - | - | - |
| Hectria_haematoceca | R | - | - | - | - | - |
| Botryotinia fuckeliana | R | - | - | - | - | - |
| Podospora_anserina | R | - | - | - | - | - |
| Neurospora_crassa | S | S | S | S | S | T |

Chain A (2c35)

Chain B (2c35)

Chain A (2c35)
Chain A (1y14)

Chain B (2c35)
Chain B (1y14)

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:

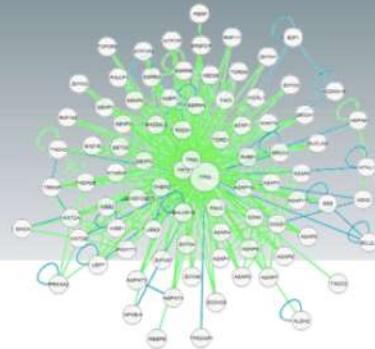
NCBI IDs are preferred, but gene names, Ensembl IDs,
and several other identifiers are translated
example: 672, TP53, ENSG00000107331

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.



Home

Home

Sample 1

Sample 2

Help

Contact us

Protein-Protein Interaction Search

Input an interacting protein pair as a query to search its homologous interactions across multiple species

Press the **?** to obtain more information on that specific field.

Query protein pair (sequences in FASTA format or [UniProt ID](#)) :

Input sequences in FASTA format

Interacting partner 1:

```
>sp|P61967|AP1S1_MOUSE  
MMRFMLLFSRQGKRLQKWYLATSDKERKKMVRELMQVVLARKPKMCSFLEWRDLKVYYK  
RYASLYFCCAIEGQDNELTITLELIHYVELLDKYFGSVCELDIIFNFKEAYFILDEFMLMG  
GDVQDTSKKSVLKAIEQADLLQEEDESPRSVLEEMGLA
```

Interacting partner 2:

```
>sp|P22892|AP1G1_MOUSE  
MPAPIRLRELIRLIRTARTQAEEREMIQKECAAIRSSFREEDNTYRCRNVAKLLYMHMLG  
YPAHFGQLECLKLIAQSQKFTDKRIGYLGAMLLLDERQDVHLLMTNCIKNDLNHSTQFVQG  
LALCTLGCMGSSEMCRDLAGEVEKLLKTSNSYLRKKAALCAHVIRKPELMEMFLPATK  
NLLNEKNHGVLHTSVVLLTEMCERSPDMLAHFRKLVPQLVRILKNLIMSGYSPEHDVSGI
```

Input UniProt ID (Ex: AP1S1_MOUSE)

Interacting partner 1:

Interacting partner 2:

Options:

E-value cut-off threshold for homolog searching **?**

10^{-1}

10^{-10} (Default)

Other:

(Ex: -50 = 10^{-50})

Joint E-value **?**

10^{-100}

10^{-40} (Default)

10^{-10}

Other:

(Ex: -50 = 10^{-50})

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 mystery 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® » blastp suite

Align Sequences Protein BLAST

[blastn](#) [blastp](#) [blastx](#) [tblastn](#) [tblastx](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)[Clear](#)Query subrange [?](#)

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 [?](#)

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® » 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 [?](#)

Upload file **mystery_protein.fasta**

Click



Choose File mystery_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 [?](#)

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® » blastp suite

Align Sequences Protein BLAST

[blastn](#) [blastp](#) [blastx](#) [tblastn](#) [tblastx](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)[Clear](#)Query subrange [?](#)

```
>mystery_protein
MGNGSVKPKHSKHPDGHSGNLTTDALRNKVTELERELRRKDAEIQEREYHLKELREQLSK
QTVAIAEELTEELQNKCQLNKLQDVHMQGGSPHQASPDVKPLEVHRKTSGLVLHSRRG
AKAGVSAEPTTRTYDLNKPPEFSFEKARVRKDSSEKKLTDALNKQFLKRLDPQQIKDM
VECMYGRNYQQGSYIJKQGEPGNHIFVLAEGRLEVWKLSSIPMWTTFGELAILYNC
```

Or, upload file [Choose File](#) no file selected

Job Title

 mystery_proteinEnter 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 [?](#)

```
>sp|P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens
OX=9606 GN=CNGA1 PE=1 SV=3
MKLSMKNNIINTQQSFVTMPNVIVPDIEKEIRRMEENGACSSFSEDDDSASTSEESENNP
HARGSF SYKSLRKGGPSQREQYLPGAIALFVNNSNNKDQEPEEKKKKKKEKKSKSDDKN
ENKNDPEKKKKKKDKKEKKKEEKS KDKEEEKKEVVVIDPSGN TYYNWLF CILPV MYNW
```

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](#)

UniProtKB - P29973 (CNGA1_HUMAN)

Basket

Display

[BLAST](#) [Align](#) [Format](#) [Add to basket](#) [History](#)[Help video](#) [Add a publication](#) [Feedback](#)Entry Protein **cGMP-gated cation channel alpha-1**Gene **CNGA1**Organism **Homo sapiens (Human)**

Status Reviewed - Annotation score: - Experimental evidence at protein level

None

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.

Caution

It is uncertain whether Met-1 or Met-5 is the initiator.

Sites

| Feature key | Position(s) | Description | Actions | Graphical view | Length |
|---------------------------|-------------|-------------|---------|----------------|--------|
| Binding site ⁱ | 546 | cGMP | | | |
| Binding site ⁱ | 561 | cGMP | | | |

Regions

| Feature key | Position(s) | Description | Actions | Graphical view | Length |
|---------------------------------|-------------|-------------|---------|----------------|--------|
| Nucleotide binding ⁱ | 487 – 609 | cGMP | | | 122 |

Entry

Publications

Feature viewer

Feature table

 Function Names & Taxonomy Subcellular location Pathology & Biotech PTM / Processing Expression Interaction Structure Family & Domains Sequences (2+) Similar proteins Cross-references



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](#)

```
>mystery_protein
MGNGSVKPKHSKHPDGHSGNLTTDALRNKVTELERELRRKDAEIQEREYHLKELREQLSK
QTVAIAEELTEELQNKCQLNKLQDVHMQGGSPHQASPDVKPLEVHRKTSGLVLHSRRG
AKAGVSAEPTTRTYDLNKPPEFSFEKARVRKDSSEKKLTDALKNQFLKRLDPQQIKDM
VECMYGRNYQQGSYIJKQGEPGNHIFVLAEGRLEVFGKEKLSSIPMWTFGELAILYNC
```

Or, upload file

[Choose File](#) no file selected [?](#)

Job Title

mystery_protein

Enter a descriptive title for your BLAST search [?](#) Align two or more sequences [?](#)Query subrange [?](#)

From

To

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)[Clear](#)

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

Or, upload file

[Choose File](#) no file selected [?](#)Subject subrange [?](#)

From

To

Program Selection

Algorithm

 blastp (protein-protein BLAST)[Choose a BLAST algorithm](#) [?](#)[BLAST](#)Search protein sequence
 Show results in a new window**Then BLAST!**[Algorithm parameters](#)

BLAST Results

Edit and Resubmit Save Search Strategies > Formatting options > Download

YouTube How to read this page Blast report des...

Job title: mystery_protein (762 letters)

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

Query ID Icl|Query_208167

Description mystery_protein

Molecule type amino acid

Query Length 762

Subject ID Icl|Query_208169

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

Molecule type amino acid

Subject Length 690

Program BLASTP 2.8.1+ > Citation

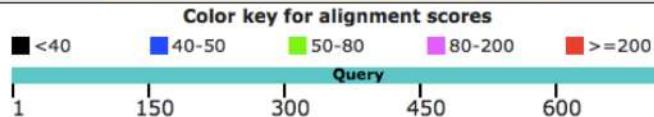
Other reports: > Search Summary [Multiple alignment] [MSA viewer]

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

All Alignments Download Graphics Multiple alignment



| | Description | Max score | Total score | Query cover | E value | Ident | Accession |
|--------------------------|--|-----------|-------------|-------------|---------|-------|--------------|
| <input type="checkbox"/> | 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...

Job title: mystery_protein (762 letters)

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

Query ID Icl|Query_208167

Description mystery_protein

Molecule type amino acid

Query Length 762

Subject ID Icl|Query_208169

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

Molecule type amino acid

Subject Length 690

Program BLASTP 2.8.1+ > Citation

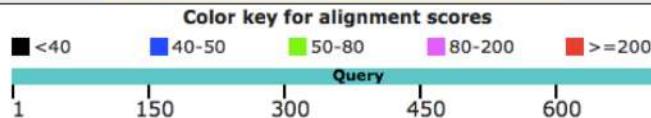
Other reports: > Search Summary [Multiple alignment] [MSA viewer]

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

All Alignments Download Graphics Multiple alignment

| | Description | Max score | Total score | Query cover | E value | Ident | Accession |
|--------------------------|--|-----------|-------------|-------------|---------|-------|--------------|
| <input type="checkbox"/> | sp P29973 CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens OX=9606 GN=CNGA1 PE=1 SV=3 | 41.6 | 138 | 65% | 1e-07 | 27% | Query_208169 |

E-value=10⁻⁷

Is our mystery 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 | LRSVSSLKKNLPEDKLTKIIDCLEVEYYDKGDYIIREGEEGSTFFILAKGVKVQSTEGLH | 340 | | | |
| | L+ V + + 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 | DALNKNQFLKRLLDPQQIKDMVECMYGRNYQQGSYIJKQGEPEGNHIFVLAEGRLEVFGQEK | 220 | | | |
| | D L K + + + +V + + Y G YI K+G+ G +++++ EG+L V + | | | | |
| Sbjct 472 | DTLKKVRIFADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVVADDG | 531 | | | |
| Query 221 | LLSSIPM--WTFGELAIL-----YNCTRTASVKAITNVKTWALDR | 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 | VDFGFAKKIGSGQKTWTCGTPEYVAPEV-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 | LRSVSLKLKNLPEDKLTKIIDCLEVEYYDKGDYIIREGEESTFFILAKGKVKTQSTEGH | 340 |
| | | L+ V + + 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 | DALKNKNQFLKRLDPQQIKDMVEC MYGRNYQQGSYIIKQGE PGNHIFVLAEGRLEV FQGEK | 220 |
| | | D L K + + + ++V + + Y G YI K+G+ G +++++ EG+L V + | |
| Sbjct | 472 | DTLKKVRIFADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVVADDG | 531 |
| Query | 221 | LLSSIPM--WTTFGELAIL-----YNCTRTASVKAITNVKTWALDR | 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 | LRSVSLKLKNLPEDKLTKIIDCLEVEYYDKGDYIIREGEESTFFILAKGKVKTQSTEGH | 340 |
| | | L+ V + + 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 | DALKNQFLKRLDPQQIKDMVEC MYGRNYQQGSYI IKQGE PGNHIFVLAEGRLEV FQGEK | 220 |
| | | D L K + + + +V + + Y G YI K+G+ G + + + EG+L V + | |
| Sbjct | 472 | DTLKKVRIFADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVVADDG | 531 |

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



cGMP-gated cation channel alpha-1

P29973 (CNGA1_HUMAN)

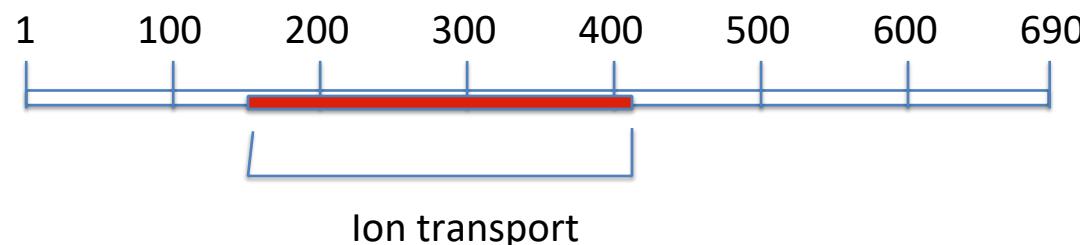
Marco Punta



cGMP-gated cation channel alpha-1

P29973 (CNGA1_HUMAN)

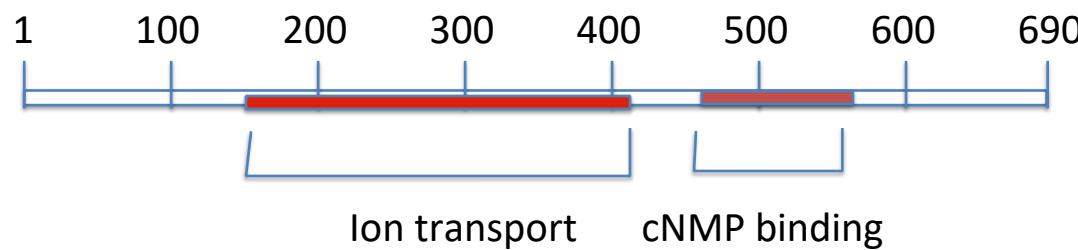
Marco Punta

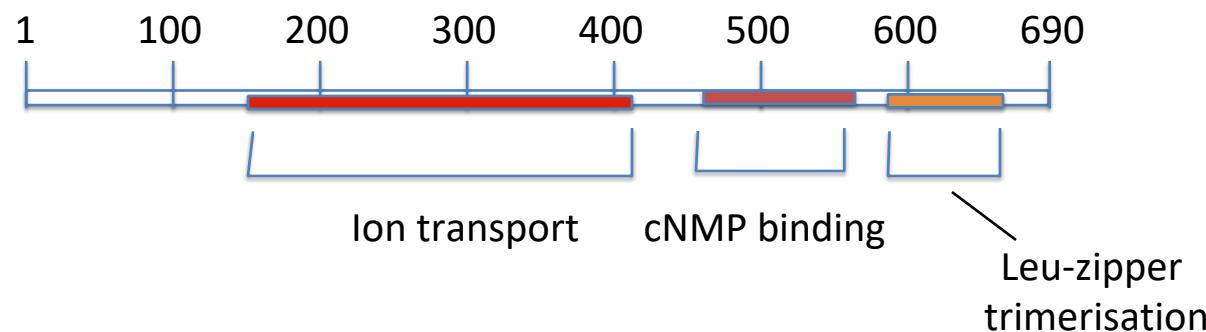


cGMP-gated cation channel alpha-1

P29973 (CNGA1_HUMAN)

Marco Punta

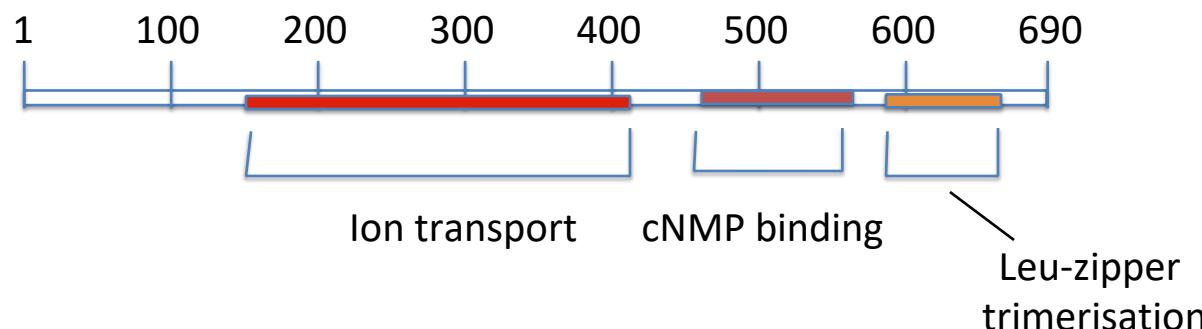




cGMP-gated cation channel alpha-1

P29973 (CNGA1_HUMAN)

Marco Punta



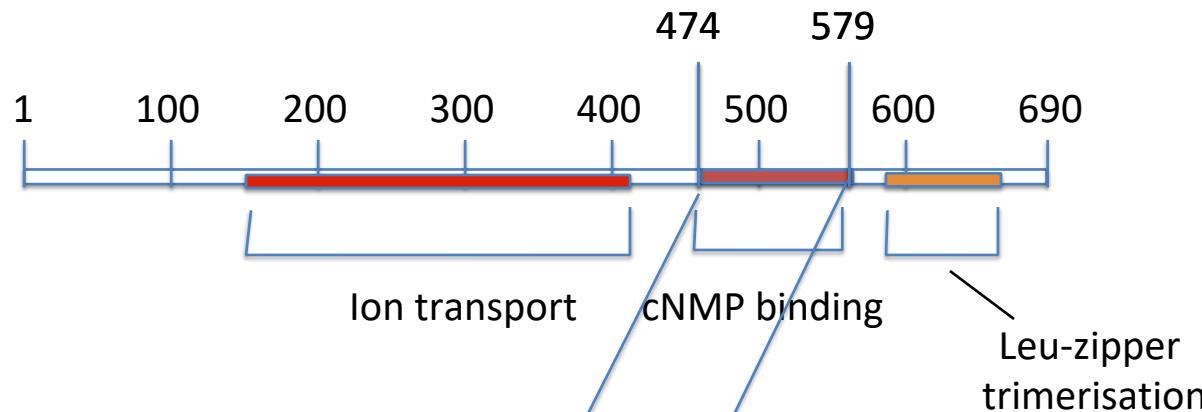
Mistery protein



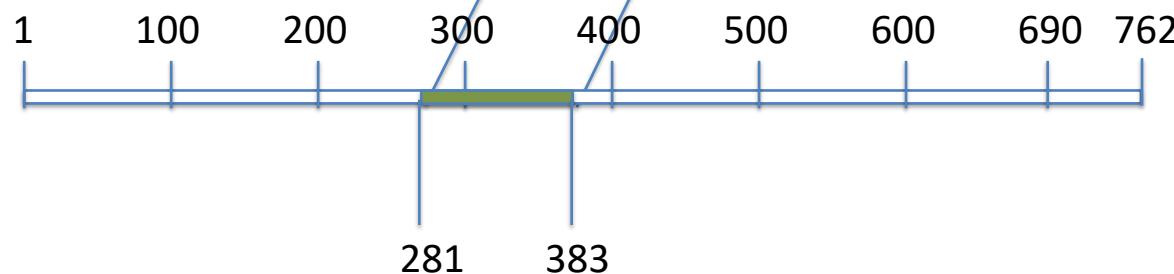
cGMP-gated cation channel alpha-1

P29973 (CNGA1_HUMAN)

Marco Punta



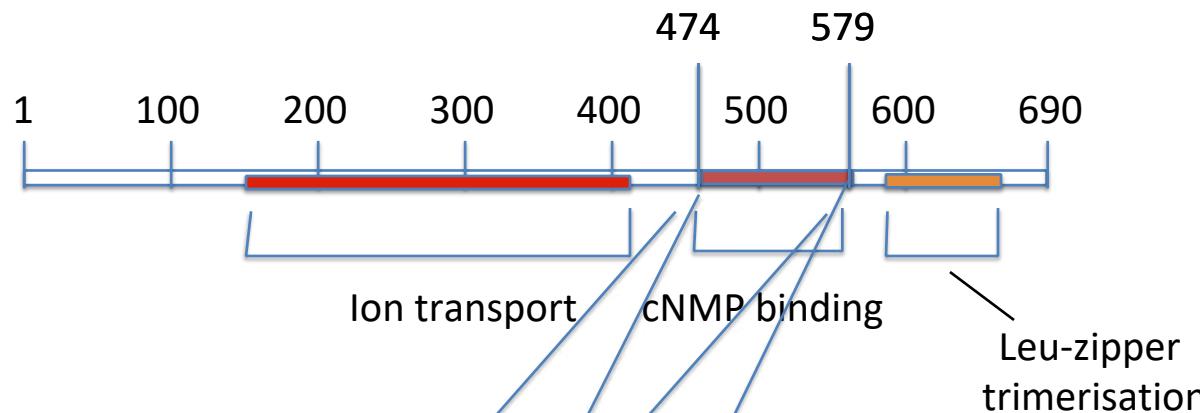
Mistery protein



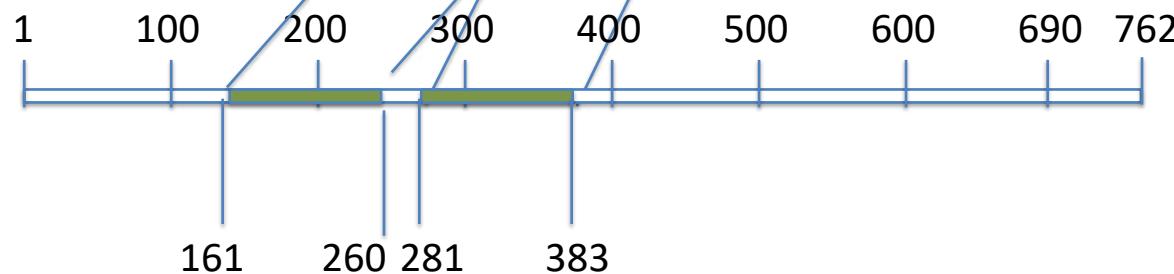
cGMP-gated cation channel alpha-1

P29973 (CNGA1_HUMAN)

Marco Punta



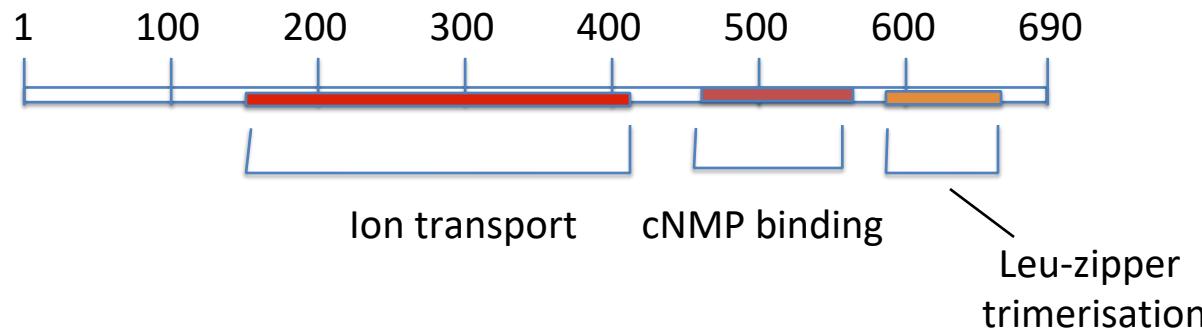
Mistery protein



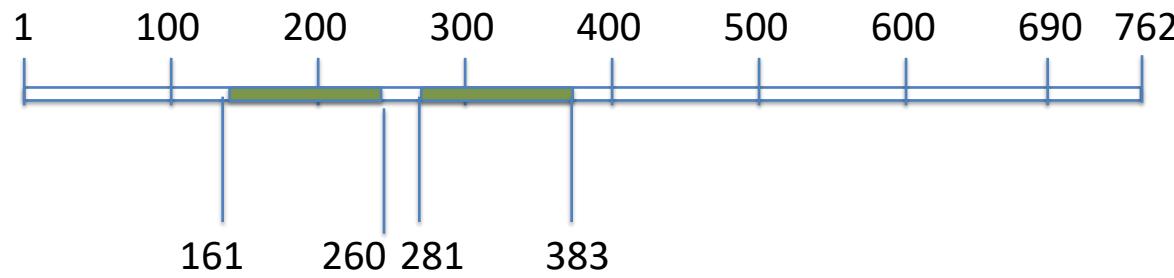
cGMP-gated cation channel alpha-1

P29973 (CNGA1_HUMAN)

Marco Punta
[color scale]



Mistery protein

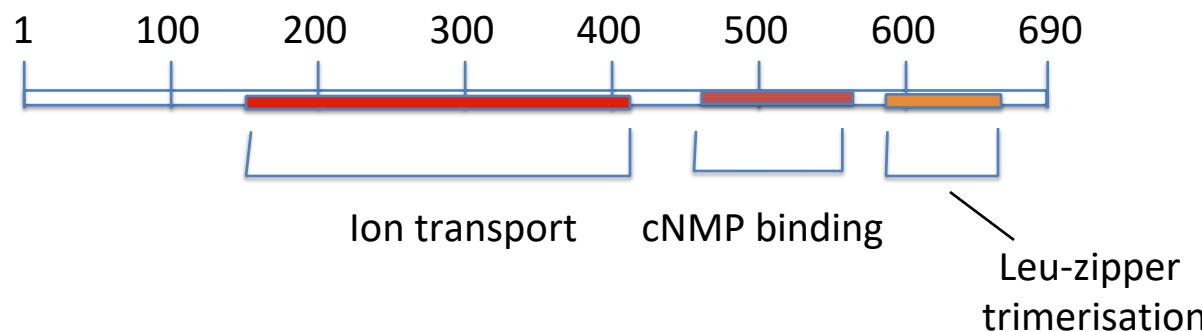


cNMP binding?

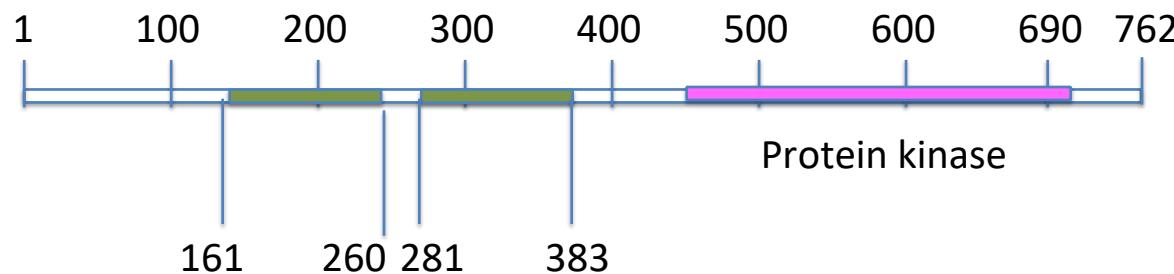
cGMP-gated cation channel alpha-1

P29973 (CNGA1_HUMAN)

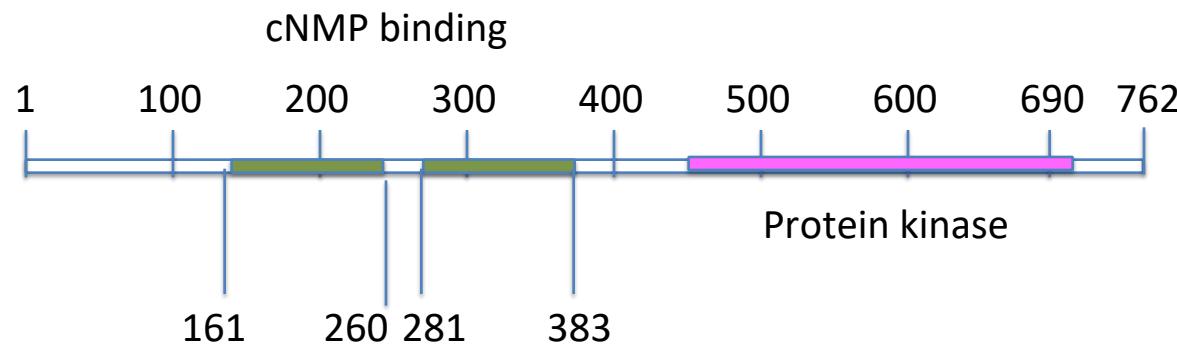
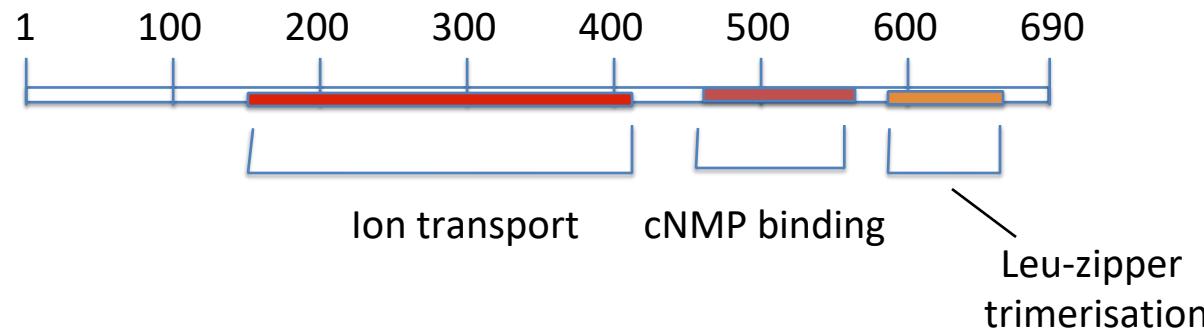
Marco Punta



Mistery protein



cNMP binding?



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

Definition (Wikipedia):

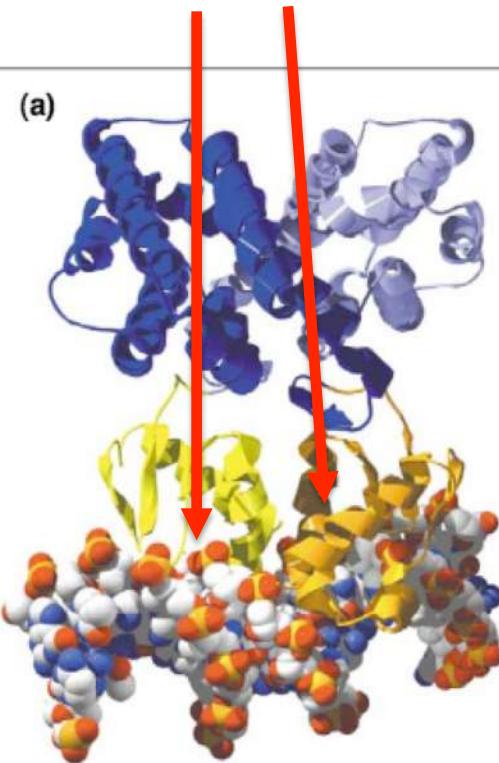
A protein domain is a conserved part of a given protein sequence and (tertiary) structure that can evolve, function, and exist independently of the rest of the protein chain. Each domain forms a compact three-dimensional structure and often can be independently stable and folded. Many proteins consist 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

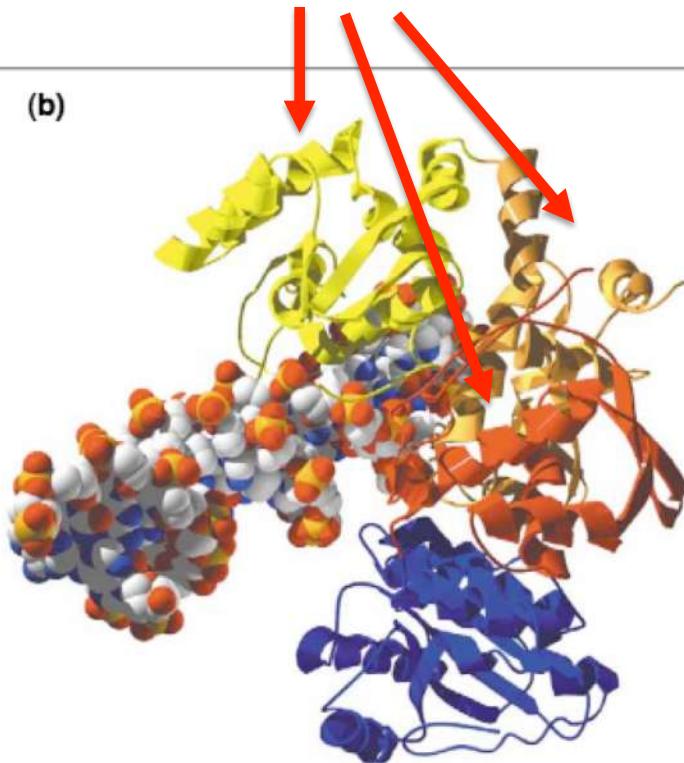
- Proteins may be homologous only in some regions (domains), this is especially true at longer evolutionary distance
- If so, function annotation transfer possible (still not safe) only between these regions

Winged helix domain

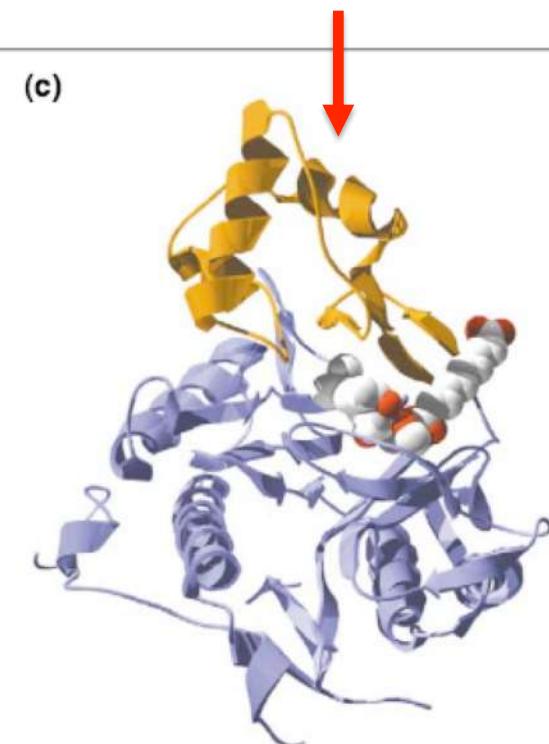
DNA binding



DNA binding



substrate specificity pocket



Current Opinion in Structural Biology

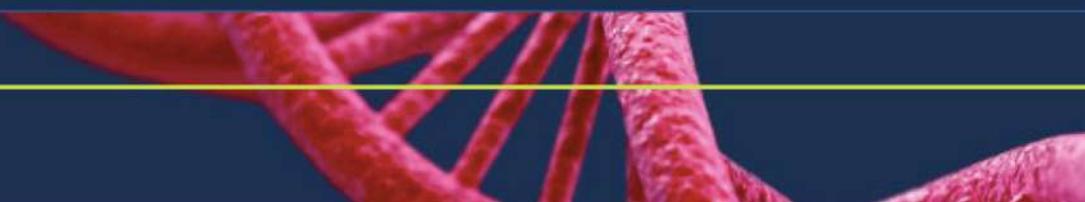
Transcription
factor

Restriction
endonuclease

Human methionine
aminopeptidase 2

Beyond just homology

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

Prokaryotes both domains and full-length equivalents



Signalling, extracellular and chromatin-associated proteins



Superfamily 1.75
HMM library and genome assignments server



Structural domains from SCOP

Gene3D
Structural domains from CATH

No limits, domains

Marco Punta



Pfam

ProDom



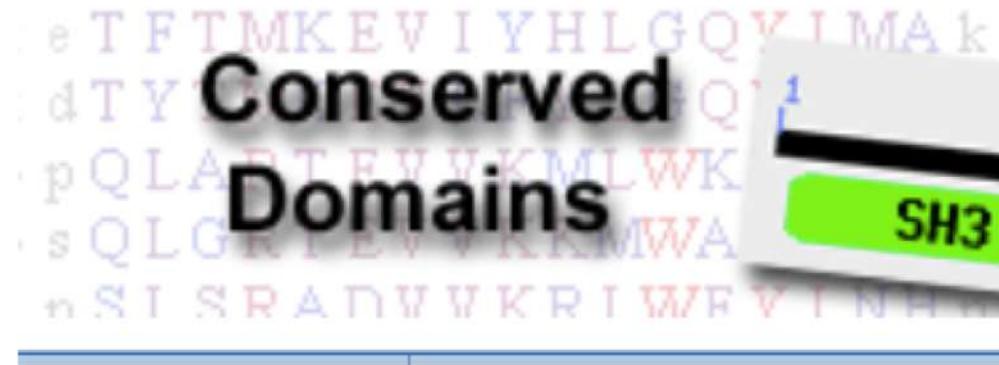
No limits, full-length proteins



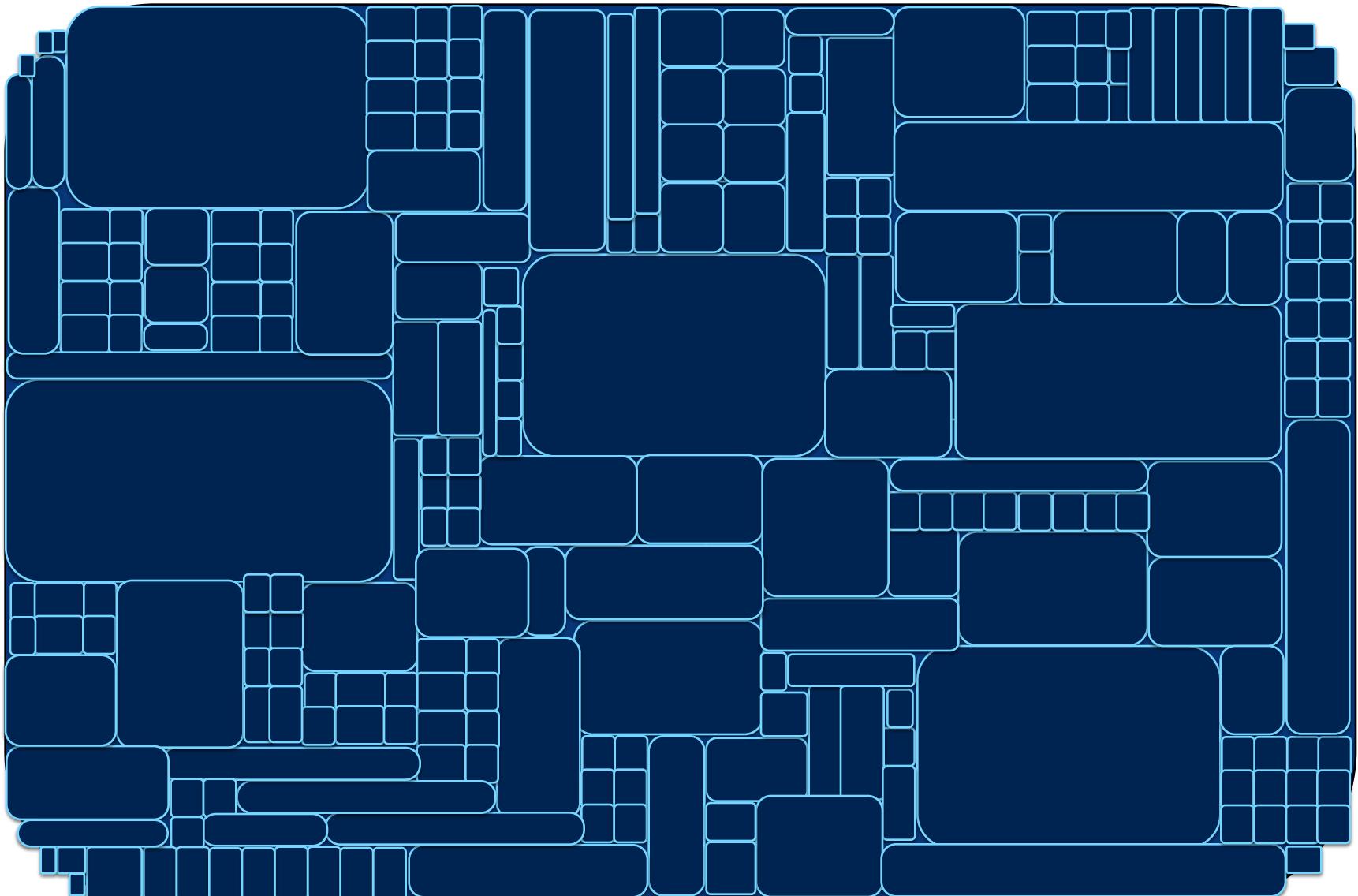


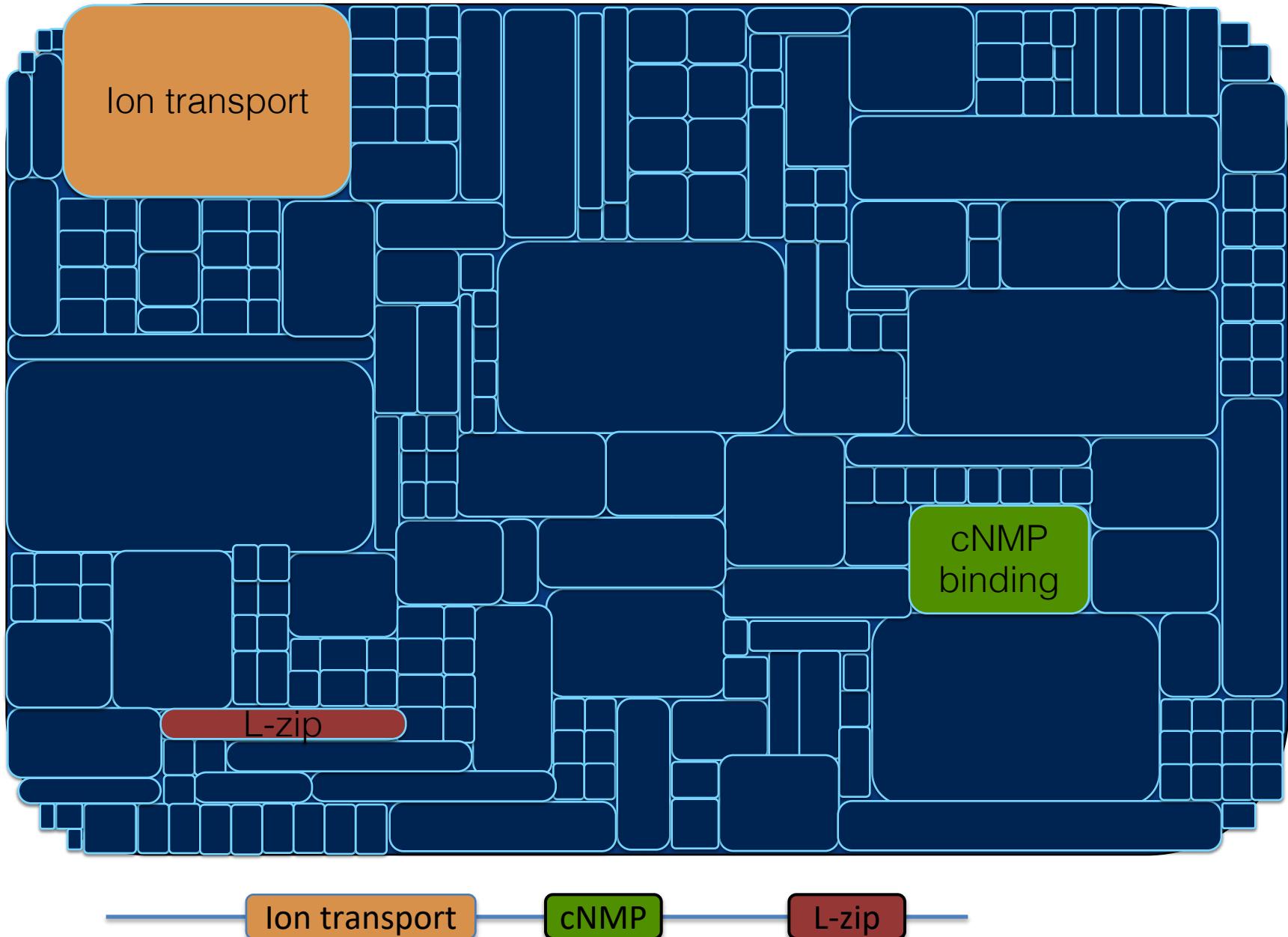
Integration

CDD









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.”

Sensitivity

Family databases build multiple sequence alignments
between member sequences

MSAs

| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 | 120 |
|----------------------|---------------------------|---|---|---|---|---|---|---------------------------------|----|-----|-----|-----|
| HBAZ_CAPI/7-107 | ERT I ILS LWS K - | I ST ----- | Q ADV I G T E T LER L F SCY P Q A K T Y F P H F - | D L H ----- | S G S A Q L R A H G S K V V A A V G D A V K S I - | - - - - - | D - N V T S A L S K L S E L H A Y V L - - - | R V D P V N F K I L S H C L | | | | |
| HBA3_PLEWA/7-107 | E K A L V V G L C G K - | I S G ----- | H C D A L G G E A L D R L F A S F G Q R T Y F S H F - | - D L S ----- | P G S A D V K R H G G K V L S A I G E A A K H I - | - - - - - | D - S M D Q A S K L S D L H A Y N L - - - | R V D P G N F Q L L S H C I | | | | |
| HBA_CATCL/6-107 | D K A D V K I A W A K - | I S P ----- | R A D E I G A E A L G R M L T V Y P Q T K T Y F A H W - | A D L S ----- | P G S G P V K H G K K V I M G A I G D A V T K F - | - - - - - | D - D L L G G L A S L S E L H A S K L - - - | R V D P S N F K I L A N C I | | | | |
| HBB_HETPO/7-106 | E L H E I T T T W K S - | I ----- | D K H S L G A K A L A R M F I V Y P W T T R Y F G N L - | K E F T - - - - | A C S Y G V K E H A K K V T G A L G V A V T H L - - - | G - D V K S Q F T D L S K K H A E E L - - - | H V D V E S F K L L A K C F | | | | | |
| HBB_SQUAC/7-107 | E K A L V N A V W T K - | T ----- | D H Q A V V A K A L E R L F V V V Y P W T K T Y F V K F N G K F H - | - - - - - | A S D S T V Q T H A G K V V S A L T V A Y N H I - | - - - - - | D - D V K P H F V E L S K K H Y E E L - - - | H V D P E N F K L L A N C I | | | | |
| HBB1_CYGMA/8-112 | E L T I I N D I F S H - | L ----- | D Y D D I G P K A L S R C L I V Y P W T Q R Y F S G F - | G N L Y N A E A I I G N A N V A A H G I K V L H G L D R G L K N M - | - - - - - | D - N I V D A Y A E L S T L H S E K L - - - | H V D P D N F K L L S D C I | | | | | |
| HBB1_XENBO/7-111 | D R Q L I N S T W G K - | V ----- | C A K T I G K E A L G R L L W T Y P W T Q R Y F S S F - | G N L N S A D A V F H E A V A A H G E K V V T S I G E A I K H M - | - - - - - | D - D I K G Y Y A Q L S K Y H S E T L - - - | H V D P C N F K R F G G C L | | | | | |
| HBB_LITCT/1-105 | G G S D M V S A F L A K - | V ----- | D K R A V G G E A L A R L L I V Y P W T Q R Y F S T F - | G N L G S A D A I S H N S K V L A H G Q R V L D S I E E G L K H P - | - - - - - | Z - B L K A Y Y A K L S E R H S G E L - - - | H V D P A N F Y R L G N V L | | | | | |
| HBB1_LEPPA/7-111 | E K Q Y I V S V F S K - | I ----- | D V D H V G A N T L E R V L I V F P W T K R Y F N S F - | G D L S S P G A I K H N N K V S A H G R K V L A A I I E C T R H F - | - - - - - | G - N I K G H A L N L S H L S E K L - - - | H V D P H N F R V L G Q C L | | | | | |
| HBB2_XENBL/8-112 | E K A A I T S V W Q K - | V ----- | N V E H D G H D A L G R L L I V Y P W T Q R Y F S N F - | G N L S N S A A V A G N A Q V O A H G K V V L S A V G N A I S H I - | - - - - - | D - S V K S S L Q Q L S K I H A T E L - - - | F V D P E N F K R F G G V L | | | | | |
| HBB_ALLMI/7-111 | E R K F I V D L W A K - | V ----- | D V A Q C G A D A L S R M L I V Y P W K R R Y F E H F - | G K M C N A H D I L H N S K V Q E H G K K V L A S F G E A V K H L - | - - - - - | D - N I K G H F A N L S K L H C E K F - - - | H V D P E N F K L L G D I I | | | | | |
| HBB0_MOUSE/8-112 | E K A A I T S I W D K - | V ----- | D L E K V G U G E T L G R L L I V Y P W T Q R Y F F D K F - | G N L L S S A Q A I M G N P R I K A H G K V V L T S I G L A V K N M - | - - - - - | D - N L K E T F A H L S E L H C D K L - - - | H A D P E N F K L L G N M L | | | | | |
| HBBN_AMMLE/2-106 | B K A L I T G F W S K - | V ----- | K V B Z V G A Z A L G R L L V V V Y P W T Z R F F Z H F - | G B L S S A B A V M B B A V K V A H G K V V L B S F S B G L K H L - | - - - - - | B - B L K G A F A S L Z L H C B K L - - - | H V B P Z B F R L L G B V L | | | | | |
| HBA1_LEPPA/7-108 | D E V L I K E A W G L - | L - H - | Q I P N A G G E A L A R M F S C Y P G T K S Y F P H F G H D F S - - - | - A N N E K V K H H G K V V D A I G Q G V Q H L - - - | - - - - - | H - D L S S C L H T L S E K H A R E L - - - | M V D P C N F Q Y L I E A I | | | | | |
| HBA1_TORMA/6-107 | N K K A I K N L L Q K - | I H S - - - - | Q T E V L G A E A L A R F C H P Q T S Y F P K F - - - | S G F S - - - - - | A N D K R V K H G A L V L K A L V D T N K H L - - - | - - - - - | D - D L P H H L N K L A E K H G K G L - - - | L V D P H N K L F S D C I | | | | |
| HBA3_SQUAC/6-107 | D K T A I K H L T G S - | L R T - - - - | N A E A F G A E A L A R M F A T T P S T K T Y F S K F - | T D F S - - - - - | A N G K R V K A H G G K V L V A N A V A D A T D H L - - - | - - - - - | D - N V A G H L D P L A V L H G T T L - - - | C V D P H N F P L L T Q C I | | | | |
| HBA_HETPO/13-114 | D R A E L A A L S K V - | LA Q - - - - | N A E A F G A E A L A R M F T V Y A A T K S Y F K D Y - | K D F T - - - - - | A A A P S I K A H G A K V V T A L A K A C D H L - - - | - - - - - | D - D L K T H L H K L A T F H G S E L - - - | K V D P A N F Q Y L S Y C L | | | | |
| GLB1_TYLHET/7-110 | Q R I K V K Q Q W A Q - | V Y S V - - - | G E S R T D F A I D V F N N F R T N P D T R S - | L F N R V N G D N V - - - | Y S P E F K A H M V R V F A G F D I L S V L - - - | - - - - - | D D K P V L D Q A L A H Y A F H K Q F G - - - | T I P - - - - - F K A F Q G T M | | | | |
| GLB4_LUMTE/11-120 | D R R E I R H I W D D - | V W S S - - - | F T D R R V A I V R A V F D D L F K H Y P T S K A L E R V K I D E P - - - | - - - - - | E S G E F K S H L V R V A N G L D L L I N L L - - - | - - - - - | D D T L V L Q S H L G H L A D O H I Q R K - - - | G V T K E Y F R G I G E A F | | | | |
| GLB3_TYLHE/8-117 | D R H E M L D N W K G - | I W S A E - | F T G R R V A I G Q A I F Q E L F A L D P N A K G V F G R V N V D - | K - - - - - | P S E A D W K A H V I R V I N G L D L A V N L L - - - | - - - - - | E D P K A L Q E E L K H L A R Q H R E R S - - - | G V K A V Y F D E M E K A L | | | | |
| GLB4_TYLHE/8-117 | D R R E M Q A L W R S - | I W S A E - | D T G R R T L I G R L L F E E L F E I D G A T K G L K F K R V N V D D T - - - | - - - - - | H S P E F F A V H L V R V N G L D T L I G V L - - - | - - - - - | G D S D - T L N S L I N D L H A E Q H K A R A - - - | G F K T V Y F K E F G K A L | | | | |
| GLB2_TYLHE/9-115 | Q R L K V K Q Q W A K - | A Y G V - - - | G H E R V E L G I A L W K S M F A Q D N D A R D L F K R V H G E D V - - - | - - - - - | H S P A F E A H M A R V F N G L D R V I S S L - - - | - - - - - | T D E P V L N A Q L E H L R Q O H I K L G - - - | I T G H M F N L M R T G L | | | | |
| GLB2_LUMTE/8-114 | E G L K V K S E W G R - | A Y G S - - - | G H D R E A F S Q A I W R A T F A Q V P E S R S L F K R V H G D D T - - - | - - - - - | S H P A F I A H A E R V L G G L D I A I S T L - - - | - - - - - | D Q P A T L K E E L D H L Q V Q H E G R K - - - | I P D N Y F D A F K T A I | | | | |
| GLB2_TUBT6/6-112 | Q R F K V K H Q W A E - | A F G T - - - | S H H R L D F G L K L L W N S I F R D A P E I R G L F K R V D G D - - - | - - - - - | A Y S A E F E A H E A R V L G G L D M T I S L L - - - | - - - - - | D D Q A A F D Q A L A H L K S Q H A E R N - - - | I K A D Y Y G V F V N E L | | | | |
| GLB2_LAMSP/7-113 | Q R L K V K R Q W A E - | A Y G S - - - | G N D R E E F G H F I W T H V F K D A P S A R D L F K R V R G D N I - - - | - - - - - | H T P A F R A H A T R V L G G L D M C I A L L - - - | - - - - - | D D E G V L N T Q L A H L A S O H S S R G - - - | V S A A Q Y D V V E H S V | | | | |
| GLB2_PAREP/8-117 | Q D I L L K E L G P H - | V - D T - - - | P A H I V E T G L G A Y H A L F T A H P Q Y I I H F S R L - - - | E G - H T I E N V M Q S E G I K H Y A R T L T E A I V H M L K E I - - - | S N D A E V K K I A A Q Y G K D H T S R K - - - | - - - - - | V T K D E F M S G E P I F | | | | | |
| Q21978_CAEEL/165-283 | S C E V V A D S W R L - | W E S R S S A A E T S A C F G L F V F Q R V F S K I P M P L R P F G - | L - S E S D D V F D L P D N H P V R R H A L F T S I L H I S V K N V - | - - - - - | D E L E A Q V A P T V F K Y G E R H Y R P D I T P H M T E E N V R V F C A Q I | | | | | | | |
| GLB_PSEDC/21-134 | T R E L C M K S L E H - | A K V G T - - - | S K E A K Q D G I D L Y K H M F E H Y P A M K K Y F K H R - | - - - - - | E N Y T P A D V V Q K D P F I I K Q G Q N I I L L A C H V L C A T Y - | - - - - - | D D R E T F D A Y V G E L M A R H E R D H V - - - | K I P N D V W N H F W E H F | | | | |
| GLB_ACSU/21-134 | T R E L C M K S L E H - | A K V D T - - - | S N E A R Q D G I D L Y K H M F E N Y P P L R K Y F K N R - | - - - - - | E E Y T A E D V Q N D P F F A K Q Q G Q K I L L A C H V L C A T Y - | - - - - - | D D R E T F N A Y T R E L L D R H A R D H V - - - | H M P P E V W T D F W K L F | | | | |
| GLB_NIPPR/21-135 | D V K - - K H T V E S - | M K A V P - | V G R D K A Q N G I D F Y K F F T H H K D L R K F K G A - | - - - - - | E N F G A D D V Q K S K R F E K Q G T A L L L A V H V L A N V Y - | - - - - - | D N Q A V F H G F V R E L M N R H E K R G V D P K L W K I F F D D V W V P F | | | | | |
| GLB_CAEEL/10-119 | D L C - V K S L E G R - | M V G T E - - - | A Q N I - E N G N A F Y R Y F F T N F P D L R V Y F K G A - | - - - - - | E K Y T A D D V Q K S S E R F D K Q Q G R I I L L A C H L L A N V Y - | - - - - - | T N E E V F K G Y V V R E T I N R H R I Y K - - - | M D P A L W M A F T V F | | | | |
| GLB2_NIPPR/16-114 | P I S K A Q Q - - - - - | A Q - - - - - | V G K D F Y K F F F T N H P D L R K Y F K G A - | - - - - - | E N F T A D D V Q K S D R F E K L G S G L L L S H I L A N T F - | - - - - - | D N E D V F R A F C R E T I D R H V G R G - - - | L D P P A L W K A F W S V W | | | | |
| GLB_TRICO/30-132 | D V V P L G S T P E K - | L - - - - - | — E N G R E F K Y F F T N H Q D L R K Y F K G A - | - - - - - | E T F T A D D I A K S D R F K K L G N Q L L L S V H L A A D T Y - | - - - - - | D N E M I F R A F V R D T I D R H V D R G - - - | L D P K L W K E F W S I Y | | | | |
| Q20638_CAEEL/74-184 | E K E L L R T W S D - | E F D - - - | — N L Y E L G S A I Y I F D H N P N C Q K L Q P F - | - - - - - | I S K Y Q G D E W K E S K E F R S Q A L K F V Q T L A Q V V K N I Y H M E R T E S F L Y M V G Q K H V K F A D R G - - - | - - - - - | F K H E Y W D I F Q D A M | | | | | |
| Q19601_CAEEL/105-25 | E R I L L E Q S W R K - | T R K - - - | T G A D H I G S K I F F M V L T A Q P D I K A F G - | L - - - - | E K I P T G R L K Y D P R F Q H O A L V Y T K T L D F V I R N L - - - | - - - - - | D Y P G K L E V Y F E N L G K R H V A M Q G - R G F E P G Y W E T F A E C M | | | | | |
| Q18311_CAEEL/32-140 | T K K L V I Q E W P R - | V L A - - - | — Q C P E L F T E I W H K S A T R S T S I K L A F G - | I - A E - N - - | E S P M Q N A A F L G L S S T I Q A F F Y K L I T Y E - L - N D D Q V R E A C E Q L G A R H V D F I S - R G F N S H F W D I F L V C M | | | | | | | |



Functionally and/or structurally relevant?

Family power

Human: 1 MGLSDGEWQLVNVWGKVEADIPGHGQEVLIRLFKGPETLEKFDKFKHLKSEDEMKGASE 60
MGLSDGEWQLVNVWGKVEAD GHGQEVLILFK HPETL KFDKFK LKSE MK SE

Mouse: 1 MGLSDGEWQLVNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSE 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 MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDFKF KHLKSEDEMKA 60
 MGLSDGEWQLVLNVWGKVEAD GHGQEVL I LFK HPETL KFDKF K LKSE MK SE

Mouse: 1 MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDFKF KNLKSEEDMKGSE 60



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

Mouse: 61 DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIIEVLKKRH 120

Human: 121 PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
 GDFGADAQGAM KALELFR D A YKELGFQG

Mouse: 121 SGDFGADAQGAMS KALELFRNDIAAKYKELGFQG 154

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 -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

Phe -2 -3 -3 -3 -2 -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 -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

BLOSUM62

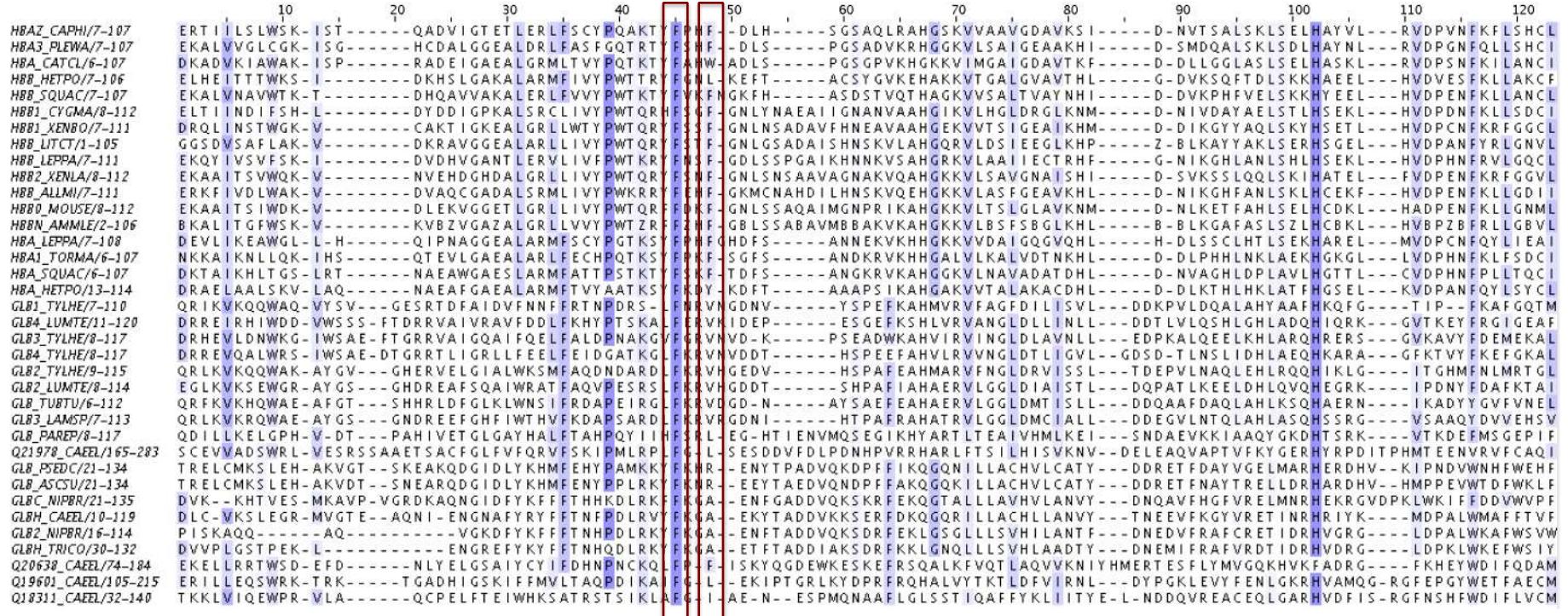
e.g. F->F Same score irrespective of position along protein sequence

Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val

Profile-sequence alignments

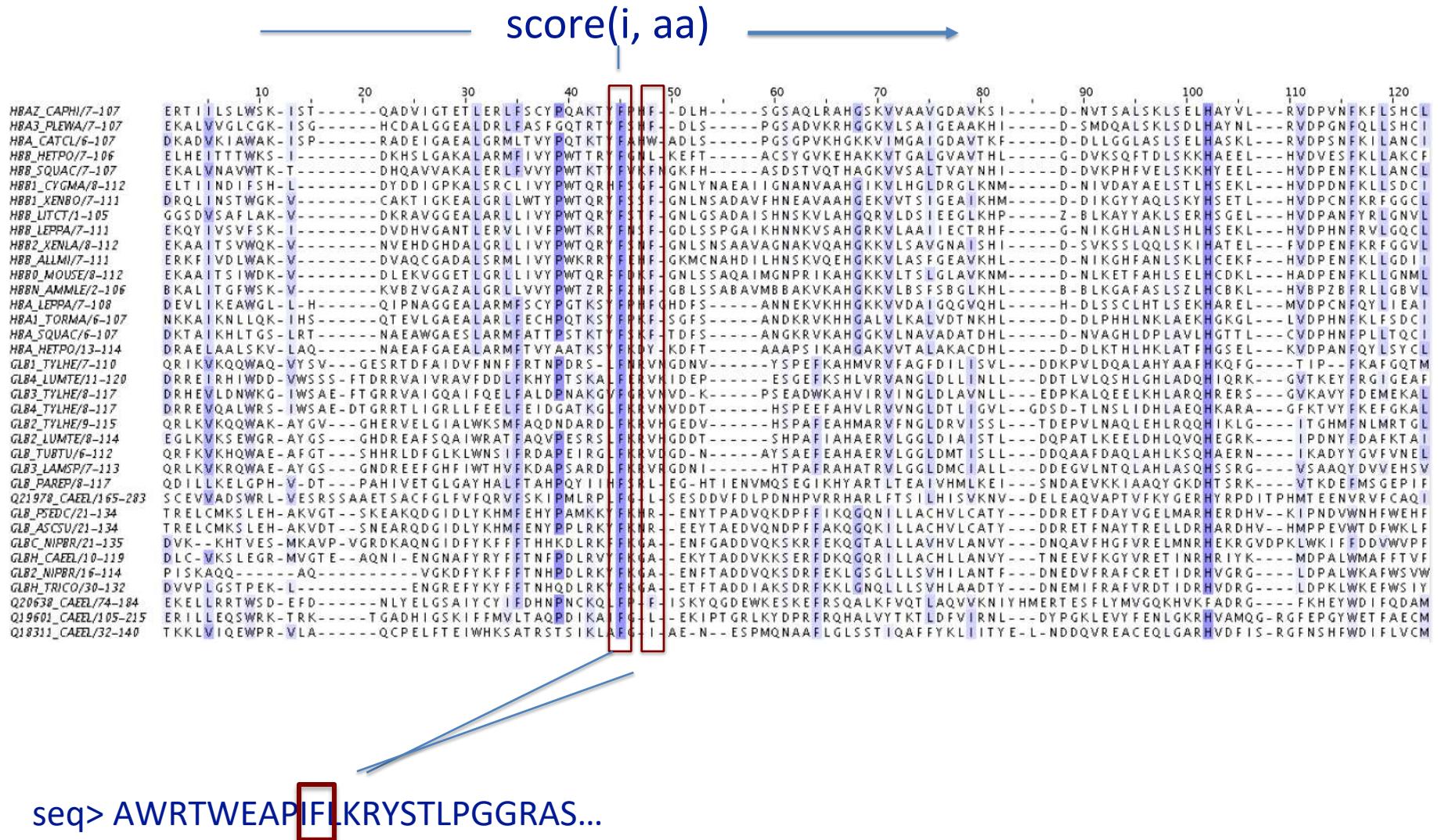
| | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 | 120 |
|----------------------|---|---|---|---|-----------------------------|----|----|----|----|-----|-----|-----|
| HBAZ_CAPII/7-107 | ERT I I LS WLS K - I S T - - - | QAD VIG T E T L E R L F S C Y P Q A K T Y F P H F - - D L H - - - | S G S A Q L R A H G S K V V A A V G D A V K S I - - - | D - N V T S A L S K L S E L H A Y V L - - - | R V D P V N F K F L S H C L | | | | | | | |
| HBA3_PLEWA/7-107 | E K A L V V G L C G K - I S G - - - | H C D A L G E A L D R L F A S G Q T R T Y F S H F - - D L S - - - | P G S A D V K R H G G K V L S A I G E A A K H I - - - | D - S M D Q A L S K L S D L H A Y N L - - - | R V D P G N F Q L L S H C I | | | | | | | |
| HBA_CATCL/6-107 | D K A D V K I A W A K - I S P - - - | R A D E I G A E A L G R M L T V Y P Q T K T Y F A H W - A D L S - - - | P G S G P V K H G G K V I M G A I G D A V T K F - - - | D - D L L G G L A S L S E L H A S K L - - - | R V D P S N F K I L A N C I | | | | | | | |
| HBB_HETPO/7-106 | E L H E I T T W K S - I - - - | D K H S L G A K A L A R M F I V Y P W T T R Y F G N L - K E F T - - - | A C S Y G V K E H A K K V T G A L G V A V T H L - - - | G - D V K S Q F T D L S K K H A E E L - - - | H V D V E S F K L L A K C F | | | | | | | |
| HBB_SQUAC/7-107 | E K A L V N A V W T K - T - - - | D H Q A V V V A K A L E R L F V V Y P W T K T Y F V K F N G K F H - - - | - A S D S T V Q T H A G K V V S A L T V A Y N H I - - - | D - D V K P H F V E L S K K H Y E E L - - - | H V D P E N F K L L A K C F | | | | | | | |
| HBB1_CYGMA/8-112 | E L T I I N D I F S H - L - - - | D Y D D I G P K A L S R C L I V Y P W T Q R H F S G - G N L Y N A E A I I G N A N V A A H G I K V L H G L D R G L K N M - - - | D - N I V D A Y A E L S T H E S K L - - - | H V D P D N F K L L A N C I | | | | | | | | |
| HBB1_XENB0/7-111 | D R Q L I N S T W G K - V - - - | C A K T I G K E A L G R L L W T Y P W T Q R Y F S S F - G N L N S A D A V F H N E A V A A H G E K V V T S I G E A I K H M - - - | D - D I K G Y Y A Q L S K Y H S E T L - - - | H V D P C N F K R F G G C L | | | | | | | | |
| HBB_LITCT/1-105 | G G S D V S A F L A K - V - - - | D K R A V G G E A L L R L I V Y P W T Q R Y F S T F - G N L G S A D A I S H N S K V L A H G Q R V L D S I I E E G L K H P - - - | Z - B L K A Y Y A K L S E R H S G E L - - - | H V D P A N F Y R L G N V L | | | | | | | | |
| HBB_LEPPA/7-111 | E K Q Y I V S V F S K - I - - - | D V D H V G A N T L E R V L I V P W T Q R Y F S N F - G D L L S P G A I K H N N K V S A H G R K V L A A I I E C T R H F - - - | G - N I K G H L A N L S H L E S K L - - - | H V D P H N F R V L G Q C L | | | | | | | | |
| HBB2_XENLA/8-112 | E K A A I T S V W Q K - V - - - | N V E H D G H D A L G R L L I V Y P W T Q R Y F S N F - G N L S N S A A V A G N A K V O A H G K K V L S A V G N A I S H I - - - | D - S V K S S L Q Q L S K I H A T E L - - - | F V D P E N F K R F G G V L | | | | | | | | |
| HBB2_XENLII/7-111 | E R K F I V D L W A K - V - - - | D V A Q C G A D A L S R M L I V Y P W T Q R Y F E H F - G K M C N A H D I L H N S N S V K Q E H G K K V L S A F G E A V K H L - - - | D - N I K G H F A N L S K L H C E K F - - - | H V D P E N F K R F G G V L | | | | | | | | |
| HBB2_MOUSE/8-112 | E K A A I T S I W D K - V - - - | D L E K V G G E T L G R L L I V Y P W T Q R F F D K F - G N L S S A Q A I M G N P R I K A H G K K V L T S L G L A V K N M - - - | D - N L K E T F A H F E L S L H C D K L - - - | H A D P E N F K L L G N M L | | | | | | | | |
| HBBN_AMMLE/2-106 | B K A L I T G F W S K - V - - - | K V B Z V G A Z A L G R L L V V Y P W T Z R F F Z H F - G B L S S A B A V M B B A K V K A H G K K V L B S F S B G L K H L - - - | B - B L K G A F A S L S Z L H C B K L - - - | H V B P Z B F R L L G B V L | | | | | | | | |
| HBA_LEPPA/7-108 | D E V L I K E A W G L - L - H - - - | Q I P N A G G E A L A R M F S C Y P G T K S Y F P H F G H D F S - - - | A N N E K V K H H G K V V D A I Q G Q V Q H L - - - | H - D L S S C L H T L S E K H A R E L - - - | M V D P C N F Q Y L I E A I | | | | | | | |
| HBA1_TORMA/6-107 | N K K A I K N L L Q L K - I H S - - - | Q T E V L G A E A L A R L F E C H P Q T K S Y F P K F - S G F S - - - | A N D K R V K H H G K V L A L V D T N K H L - - - | D - D L P H H L N K L A E K H G K G L - - - | L V D P H N F K L F S D C I | | | | | | | |
| HBA_HETPO/13-114 | D K T A I K H L T G S - L R T - - - | N A E A W G A E S L A R M F A T T P S T K T Y F S K F - T D F S - - - | A N G K R V K H A G G K V L N A V A D A T D H L - - - | D - N V A G H L D P L A V L H G T T L - - - | C V D P H N F P L L T Q C I | | | | | | | |
| GLB1_TYLHE/7-110 | D R A E L A A L S K V - L A Q - - - | N A E A F G A E A L A R M F T V Y A A T K S Y F K D Y - K D F T - - - | A A P S I S K A H G A K V V T A L A K A C D H L - - - | D - D L K T H L H K L A T F H G S E L - - - | K V D P A N F Q Y L S Y C L | | | | | | | |
| GLB4_LUMTE/11-120 | Q R I K V K Q Q W A Q - V Y S V - - - | G E S R T D F A I D V F N N F F R T N P D R S - L F N R V N G D N V - - - | Y S P E F K A H M V R V F A G F D I L S V L - - - | D D Q P V L D L Q A L A H Y A A F H K Q F G - - - | T I P - - F K A F G Q T M | | | | | | | |
| GLB3_LUMTE/8-112 | D R R E I R H I W D D - - V W S S S - | F T D R R V A I V R A V F D D L F K H Y P T S K A L F E R V K I D E P - - - | E S G E F K S H L V R V A N G L D L L I N L L - - - | D D T L V L Q S H L G H L A D Q H I Q R K - - - | G V T K E Y F R G I G E A F | | | | | | | |
| GLB3_TYLHE/8-117 | D R H E V L D N W K G - I W S A E - | F T G R R V A I Q G Q A I Q E L F A L D P N A K G V F G R V N V D - K - - - | P S E A D W K A H V I R V I N G L D L A V N L L - - - | E D P K A L Q E E L K H L A R Q H R E S - - - | G V K A V Y F D E M E K A L | | | | | | | |
| GLB2_TYLHE/8-117 | D R R E V Q A L W R S - I W S A E - | D T G R R T L I G R L L F E E L F E I D G A T K G L F K R V N V D D T - - - | H S P E E F A H V L R V V N G L D T L I G V L - - - | G D S D - T L N S L I D H L A E Q H K A R A - - - | G F K T V Y F K E F G K A L | | | | | | | |
| GLB2_TYLHE/9-115 | Q R L K V K Q Q W A K - A Y G V - - - | G H E R V E L G I A L W K S M F A Q D N D A R D L F K R V H G E D V - - - | H S P A F E A H M A R V F N G L D R V I S S L - - - | T D E P V L N A Q L E H L R Q R Q H I K L G - - - | I T G H M F N L M R T G L | | | | | | | |
| GLB2_LUMTE/8-114 | E G L K V K S E W G R - A Y G S - - - | G H D R E A F S Q A I W R A T F A Q V P E S R S L F K R V H G D D T - - - | S H P A F I A H A E R V F N G L D I A I S T L - - - | D Q P A T L K E E D L H Q V Q H E G R K - - - | I P D N Y F D A F K T A I | | | | | | | |
| GLB2_TUBTU/6-112 | Q R F K V K H Q W A E - A Y G S - - - | S H H R L D F G L K L W N S I F R D A P E I R G L F K R V D G D - N - - - | A Y S A E F A H A E R V L G G L D M T I S L L - - - | D D Q A A F D A Q L A H L K S Q H A E R N - - - | I K A D Y Y G V F V N E L | | | | | | | |
| GLB3_LAMSP/7-113 | Q R L K V K R Q W A E - A Y G S - - - | G N D R E E F G H F I W T H V F K D A P S A R D L F K R V R G D N I - - - | H T P A F R A H A T R V L G G L D M C I A L L - - - | D D E G V L N T Q L A H L A S Q H S R G - - - | V S A A Q Y D V V E H S V | | | | | | | |
| GLB_PAREP/8-117 | Q D I L L K E L G P H - V - D T - - | P A H I V E T G L G A Y H A L F T A H P Q Y I I H F S R L - E G - H T I E N V M Q S E G I K H Y A R T L T E A I V H M L K E I - - - | S N D A E V K K I A A Q Y G K D H T S R K - - - | V T K D E F M S G E P I F | | | | | | | | |
| Q21978_CAEEL/165-283 | S C E V V A D S W P L - V E R S S A A E T S A C F G L F V F Q R V F S K I P M L R P L F G - L - | S E S D D V F D L P D N H P V R R H A R L F T S I I H I S V K N V - - - | D E L E A Q V P A T V F K Y G E R H Y R P D I T P H M T E E N V R V F C A Q I | | | | | | | | | |
| GLB_PSED/C/21-134 | T R E L C M K S L E H - A K V G T - - - | S K E A K Q D G I D L Y K H M F E H Y P A M K K Y F K H R - - | E N Y T P A D V Q K D P F F I K Q G Q N I I L L A C H V L C A T Y - - - | D D R E T F D A Y V G E L M A R H E R D H V - - | K I P N D V W N H F W E H F | | | | | | | |
| GLB_ACSVU/21-134 | T R E L C M K S L E H - A K V D T - - - | S N E A R Q D G I D L Y K H M F E H Y P P L R K Y F K N R - - | E E Y T P A D V Q N D P F F A K Q G Q K I I L L A C H V L C A T Y - - - | D D R E T F N A Y T R E L L D R H A R D H V - - | H M P P E V W T D F W K L F | | | | | | | |
| GLB_C_NIPPR/21-135 | D V K - - K H T V E S - M K A V P - | V G R D K A Q N G I D F Y K F F T H H K D L R K F F K G A - - | E N F G A D D V Q K S K R F E K Q G T A L L L A V H V L A N V Y - - - | D N Q A V F H G F V R E L M N R H E K R G V D P K L W K I F F D D V W V P F | | | | | | | | |
| GLBH_CAEEL/10-119 | D L C - V K S L E G R - M V G T E - - | A Q N I - E N G N A F Y R Y F F T N F P D L R V Y F K G A - - | E K Y T A D D V K K S E R F D K Q G Q R I I L L A C H V L L A N V Y - - - | T N E E V F K G Y V R E T I N R H R I Y K - - - | M D P A L W M A F F T V F | | | | | | | |
| GLB2_NIPPR/16-114 | P I S K A Q Q - - - A Q - - - | V G K D F Y K F F F T N H P D L R K Y F K G A - - | E N F T A D D V Q K S D R F E K L G S G L L L S V H I L A N T F - - - | D N E D V F R A F C R E T I D R H V G R G - - - | L D P A L W K A F W S V W | | | | | | | |
| GLB2_TRICO/30-132 | D V V P L G S T P E K - L - - - | E N G R E F Y K F F T N H Q D L R K Y F K G A - - | E T F T A D D I A K S D R F K K L G N Q L L L S V H L A A D T Y - - - | D N E M I F R A F V R D T I D R H V D R G - - - | L D P K L W K E F W S I Y | | | | | | | |
| Q20638_CAEEL/74-184 | E K E L L R R T W S D - E F D - - - | N L Y E L G S A I Y C Y I F D H N P N C K Q L F P - F - | I S K Y Q G D E W K E S K E F R S Q A L K F V Q T L A Q V V K N I Y H M E R T E S F L Y M V G Q K H V K F A D R G - - - | F K H E Y W D I F Q D A M | | | | | | | | |
| Q19601_CAEEL/105-215 | E R I L L E Q S W R K - T R K - - - | T G A D H I G S K I F F M V L T A Q P D I K A I F G - L - | E K I P T G R L K Y D P R F R Q H A L V Y T K T L D F V I R N L - - - | D Y P G K L E V Y F E L G R K R H V A M Q G - R G F E P G Y W E T F A E C M | | | | | | | | |
| Q18311_CAEEL/32-140 | T K K L V I Q E W P R - V L A - - - | Q C P E L F T E I W H K S A T R S T S I K L A F G - I - A E - N - - | E S P M Q N A A F L G L S S T I Q A F F Y K L I I T Y E - L - N D D Q V R E A C E Q L G A R H V D F I S - R G F N S H F W D I F L V C M | | | | | | | | | |

AA probabilities within the family are position specific



$$F_i \rightarrow F_i \neq F_j \rightarrow F_j$$

Position-specific scores



Sequence-profile alignments

- Position specific substitution matrices
- profile-hidden Markov models

EDITORIAL

New computational approaches to understanding molecular protein function

Jacquelyn S. Fetrow¹*, Patricia C. Babbitt²

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