Marco Punta

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

Molecular dynamics

↓

Prediction methods for protein structural and functional features

↓

Structural Genomics

↓

Pfam

Goals

- "Understand" annotation transfer by homology
- Know what protein family databases are and why they are useful

Outline

- Homology (definition, implications, how to detect)
- Exercise 1 and 2: homology-based function annotation transfer
- Protein domains
- Protein families
- Exercise 3: how to build a new (Pfam) protein family

Marco Punta

Homology

EMBO Workshop, Budapest, 2016

Definition:

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

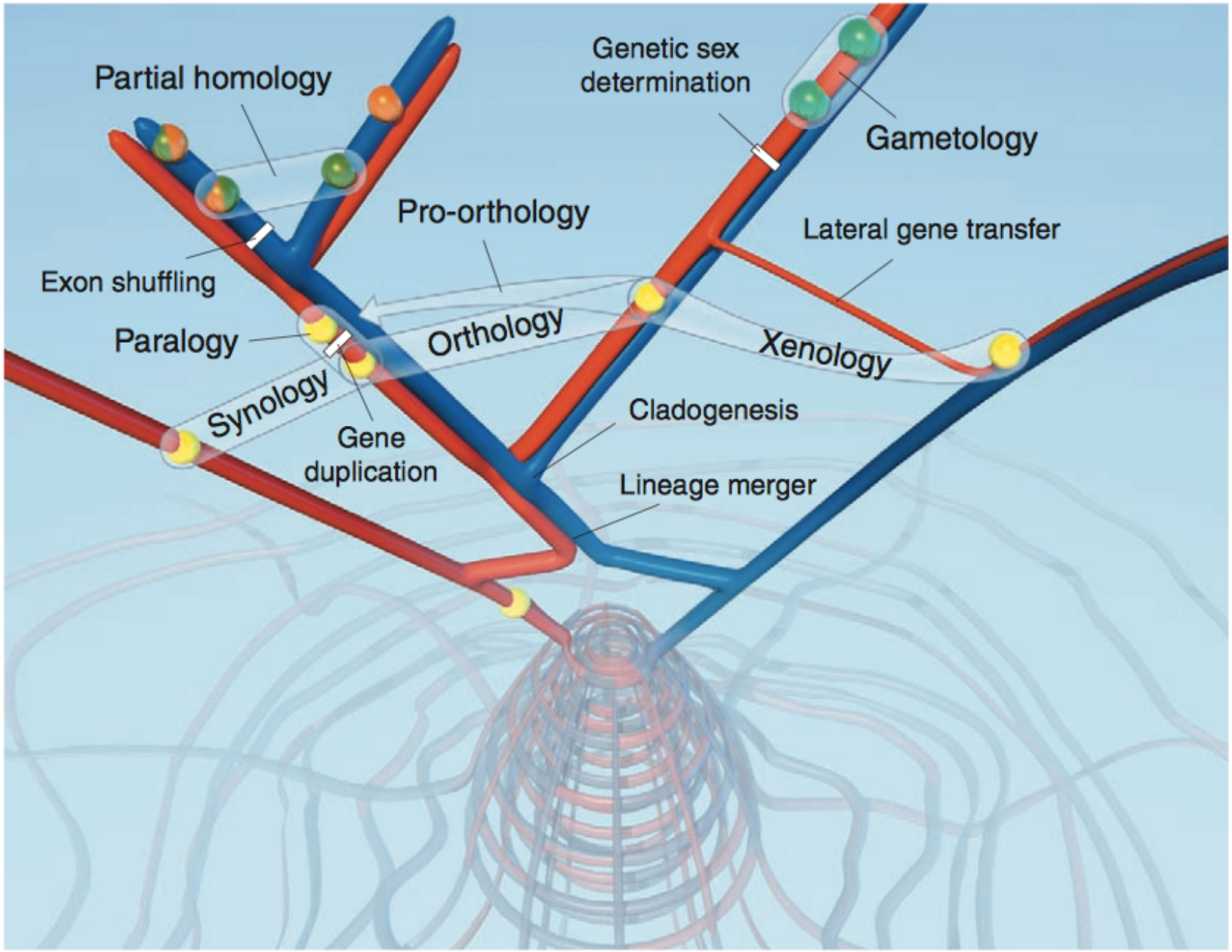
Marco Punta

Origins of homology in proteins

EMBO Workshop, Budapest, 2016

Origin of homology in proteins

- Speciation (orthology)
- Gene duplication (paralogy)
- Horizontal gene transfer (xenology)
- Whole genome duplication (ohnology)
- Gametology, Synology



Mindell and Meyer Trends in Ecology and Evolution 2001

Marco Punta

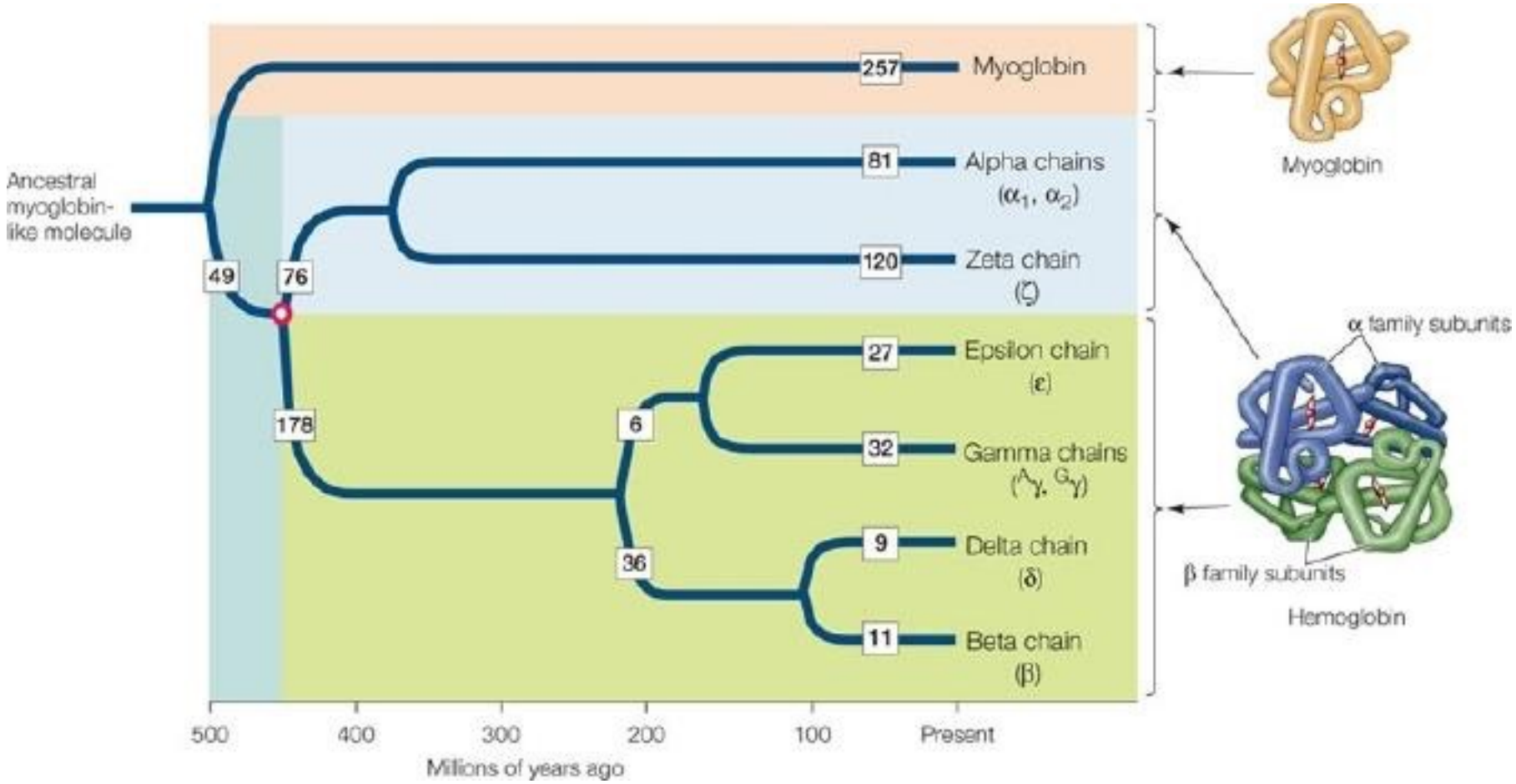
Protein Families

EMBO Workshop, Budapest, 2016

Definition:

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

Globins in Human



http://www.studyblue.com/notes/note/n/exam-3/deck/8955883

Marco Punta

Homology: why bother?

Mind the gap!

Marco Punta

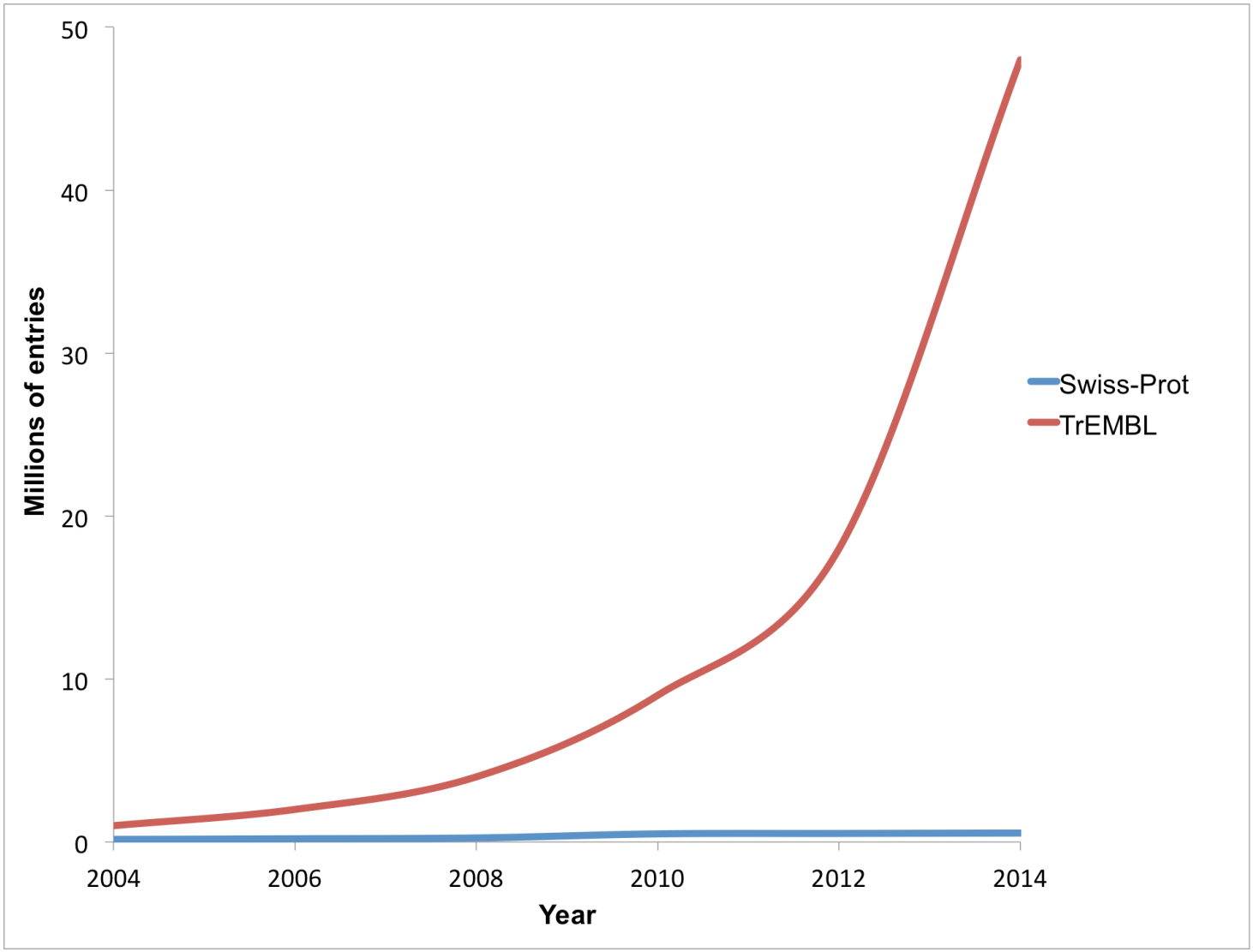


Figure courtesy of Alex Mitchell (EMBL-EBI)

Mind the gap!

Marco Punta

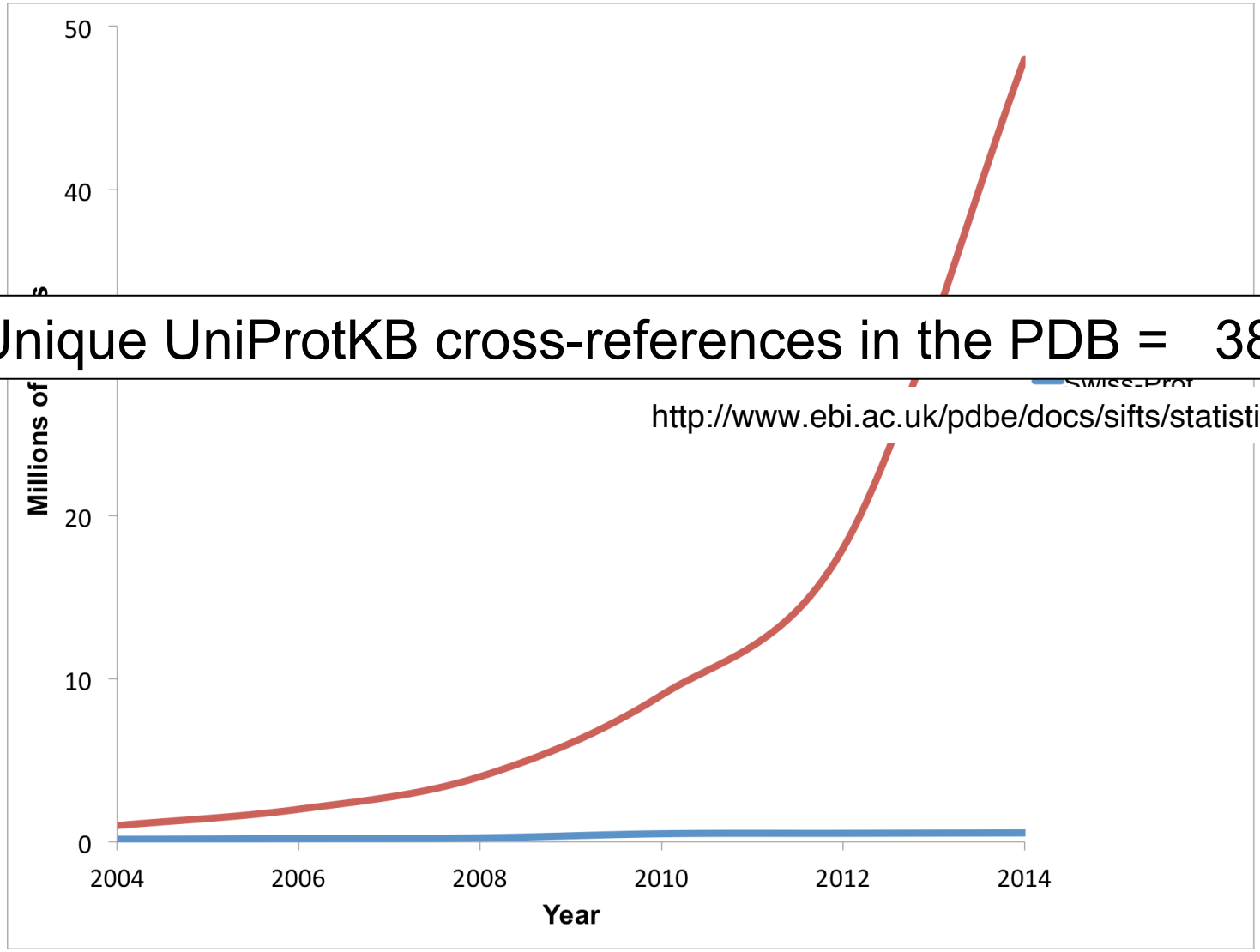


Figure courtesy of Alex Mitchell (EMBL-EBI)

Mind the gap!

Marco Punta

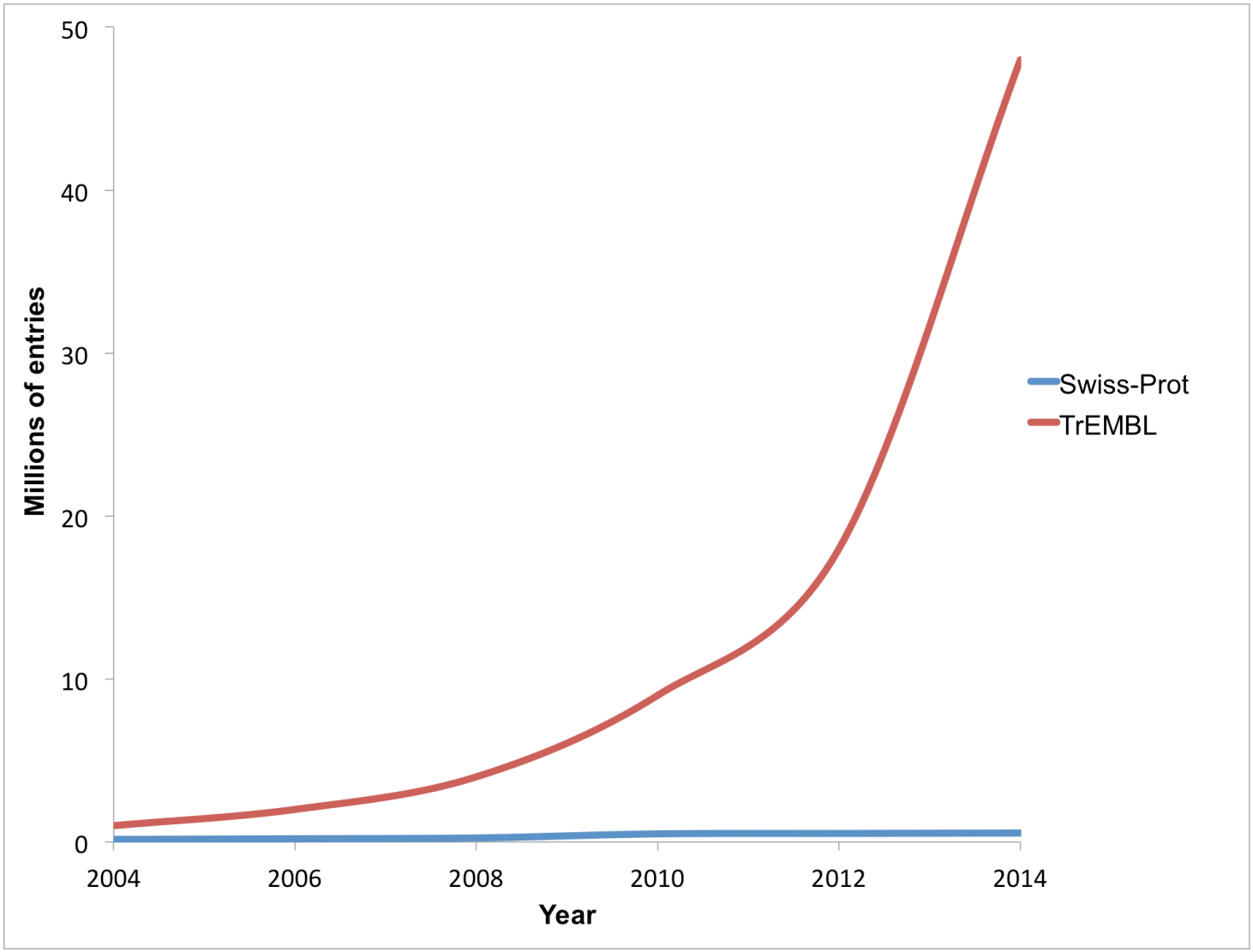


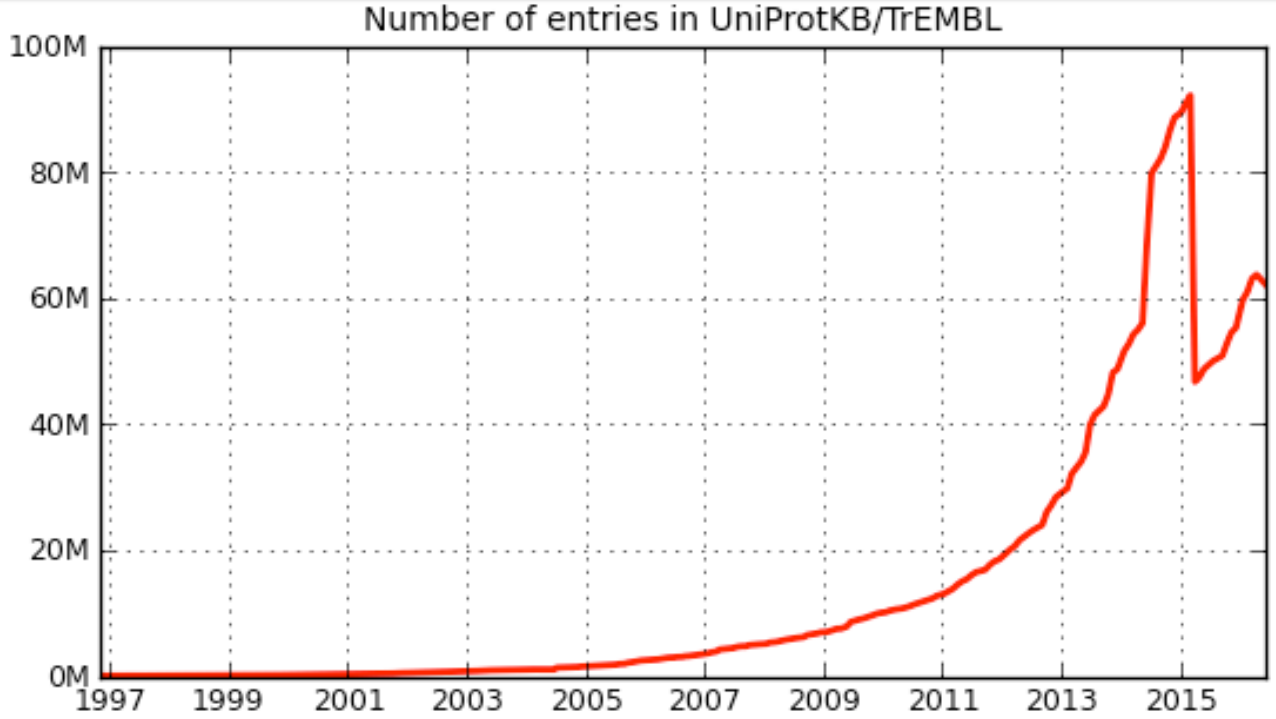
Figure courtesy of Alex Mitchell (EMBL-EBI)

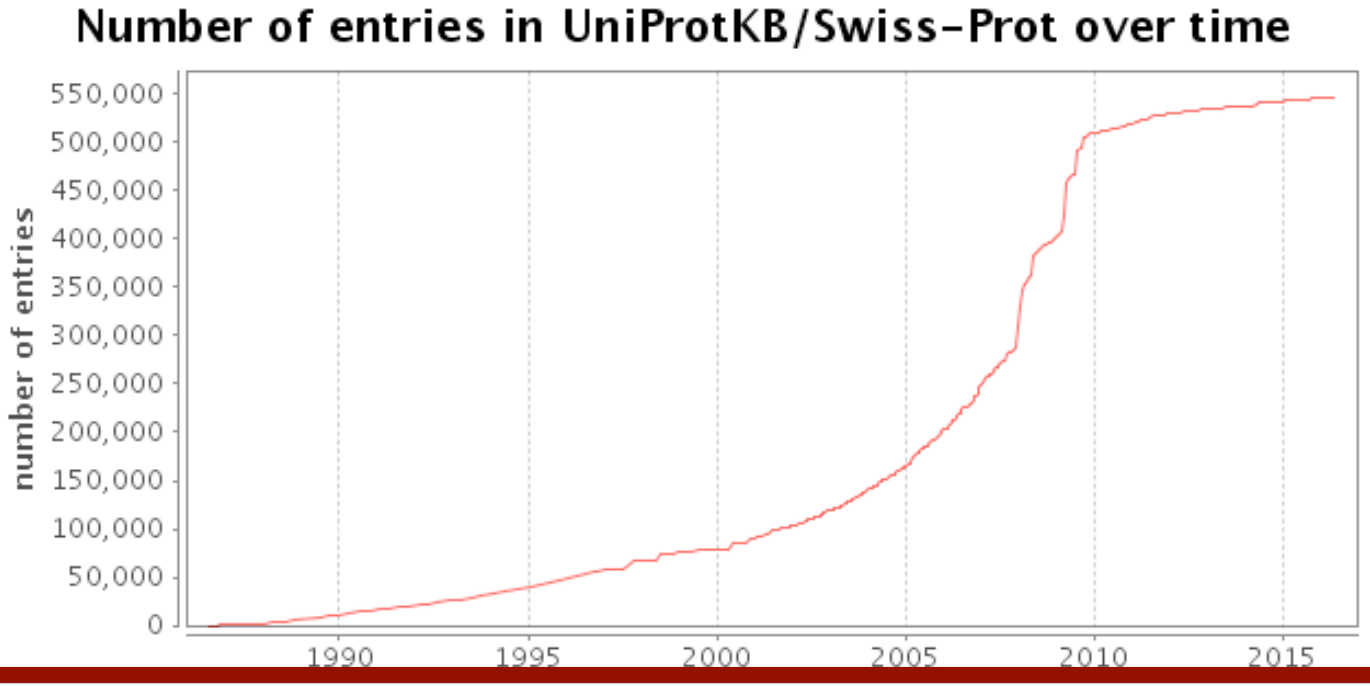
Mind the gap!

Marco Punta

	Protein Existence (PE)	Number of entries
1	Evidence at protein level	92,536
2	Evidence at transcript level	57,757
3	Inferred from homology	387,589
4	Predicted	11,358
5	Uncertain	1,953

Marco Punta


Number of entries in UniProtKB/TrEMBL
100M
80M
60M
40M
20M
0M
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015


Number of entries in UniProtKB/Swiss-Prot over time
number of entries
550,000
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
1990
1995
2000
2005
2010
2015

Marco Punta

Homology: why bother?

Homologous proteins come from a common ancestor, that is, a single ancestral protein with given sequence, structure and functions

Homology $\overset{\Rightarrow}{\Leftarrow}$ similar sequence?

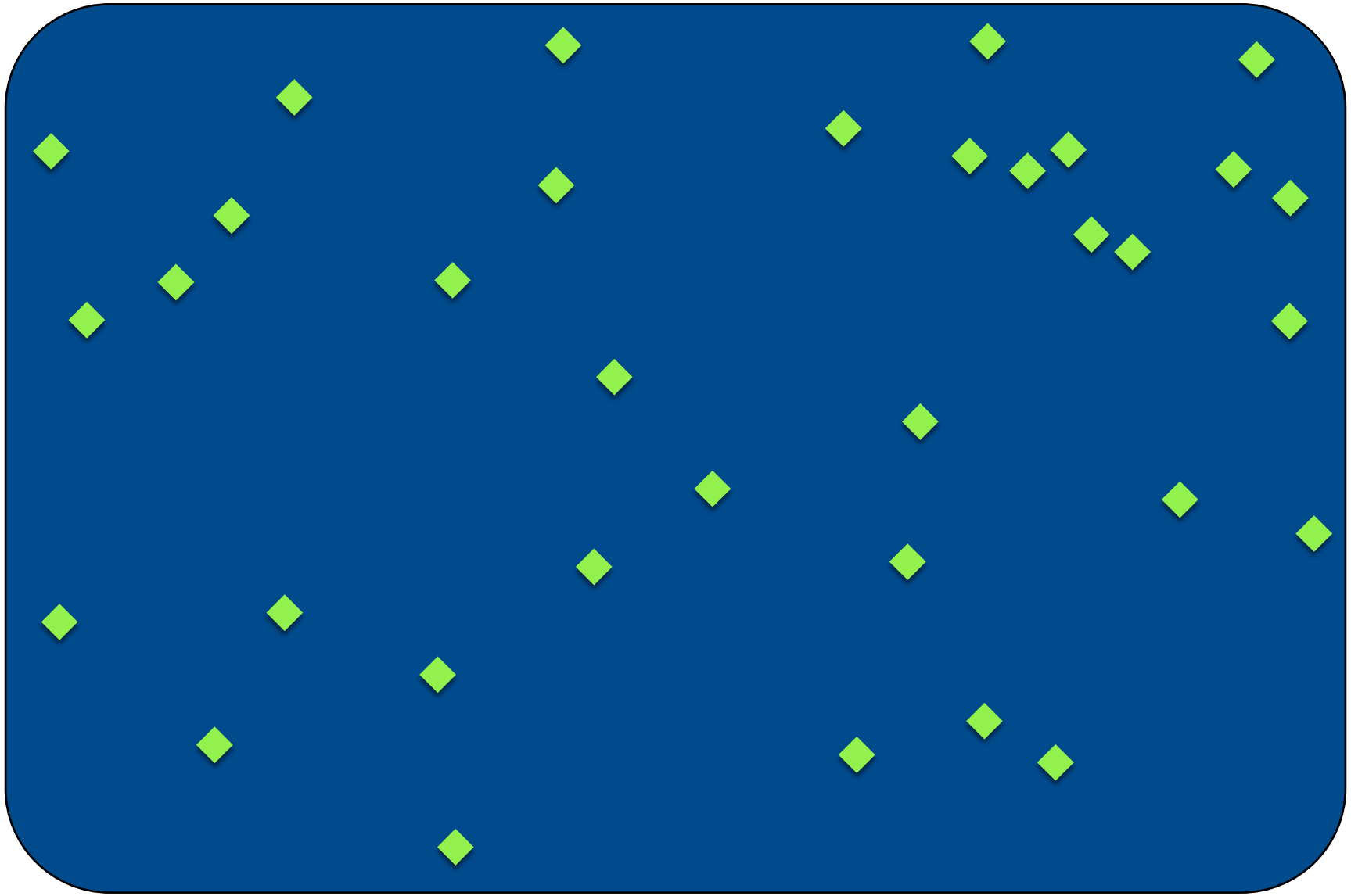
Homology $\overset{\Rightarrow}{\Leftarrow}$ similar structure?

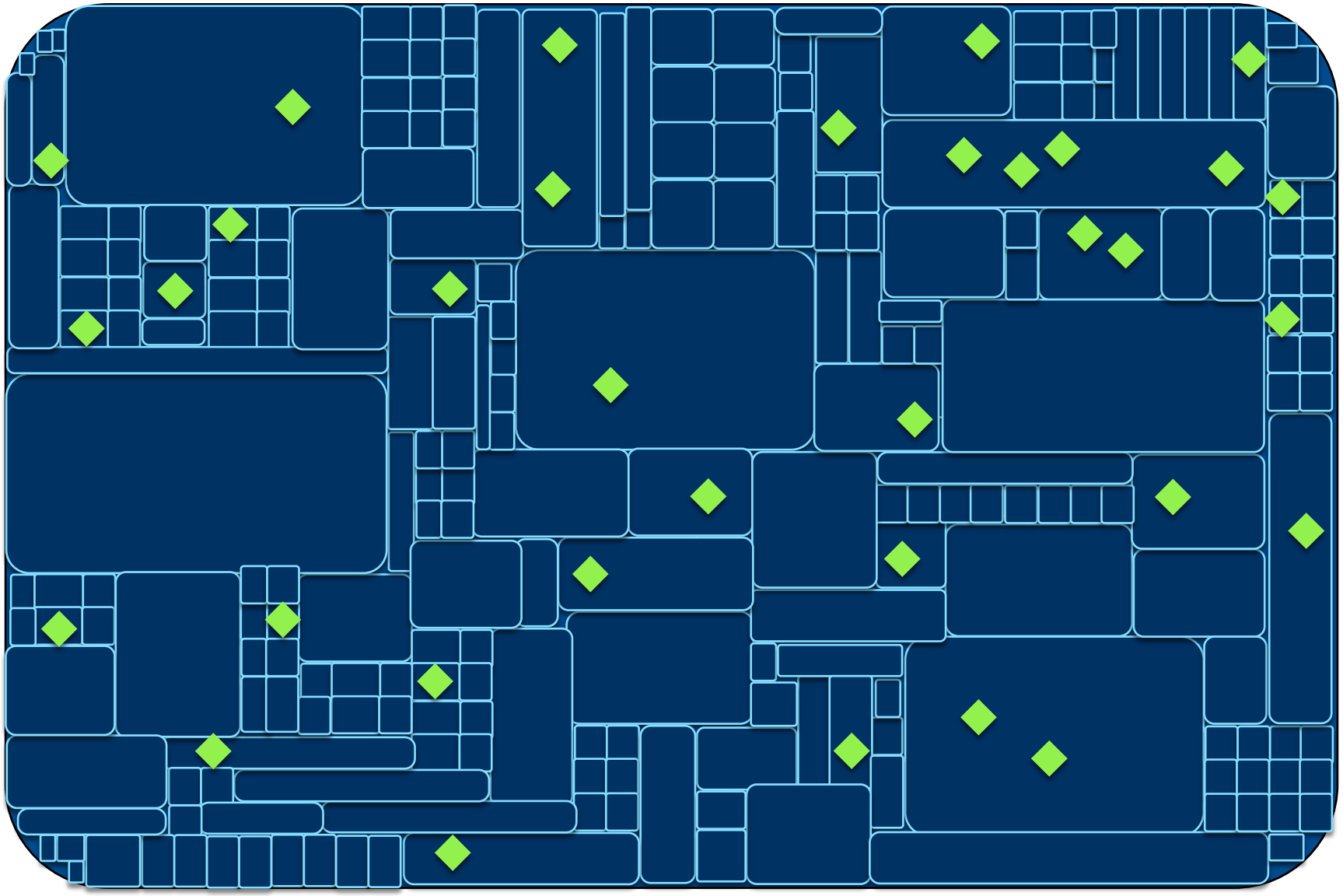
Homology $\overset{\Rightarrow}{\Leftarrow}$ similar function?

Marco Punta

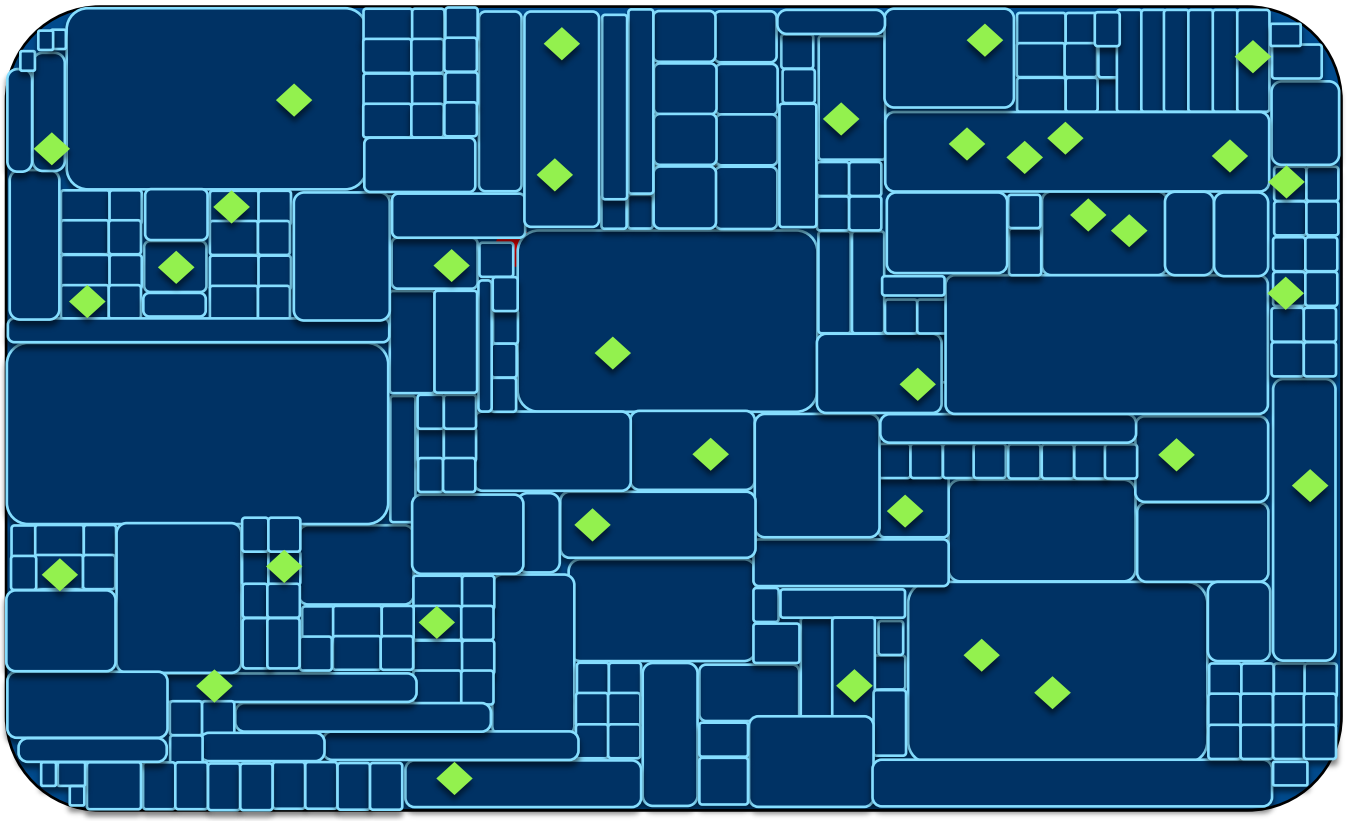
Homologous protein regions have a similar (core) structure!

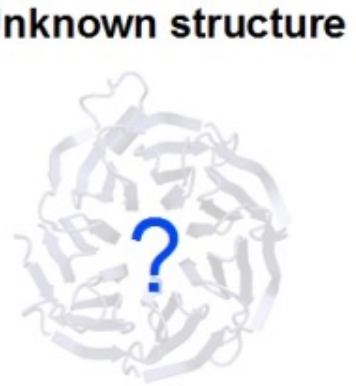
Chothia and Lesk *EMBO J* (1986)





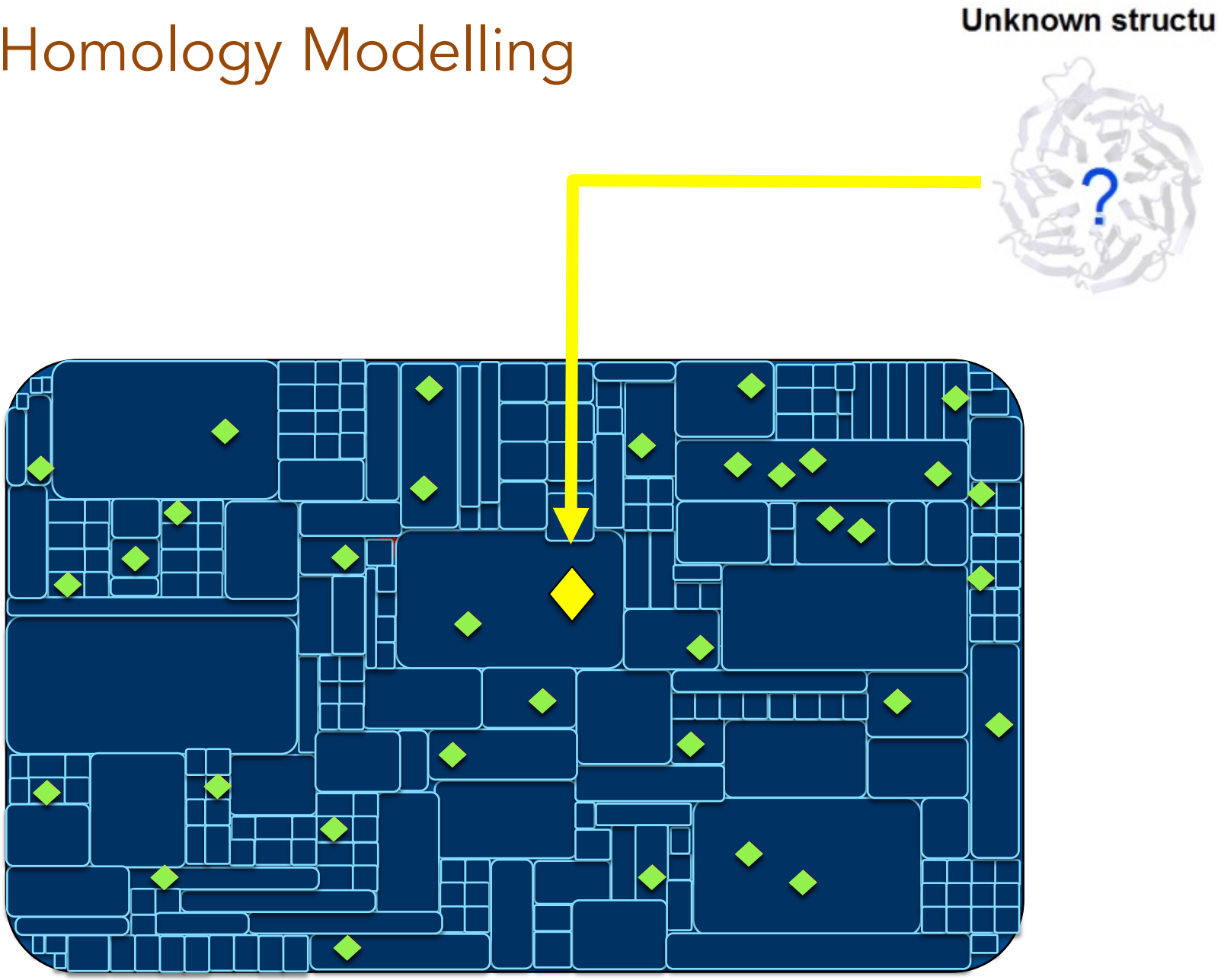
Homology Modelling





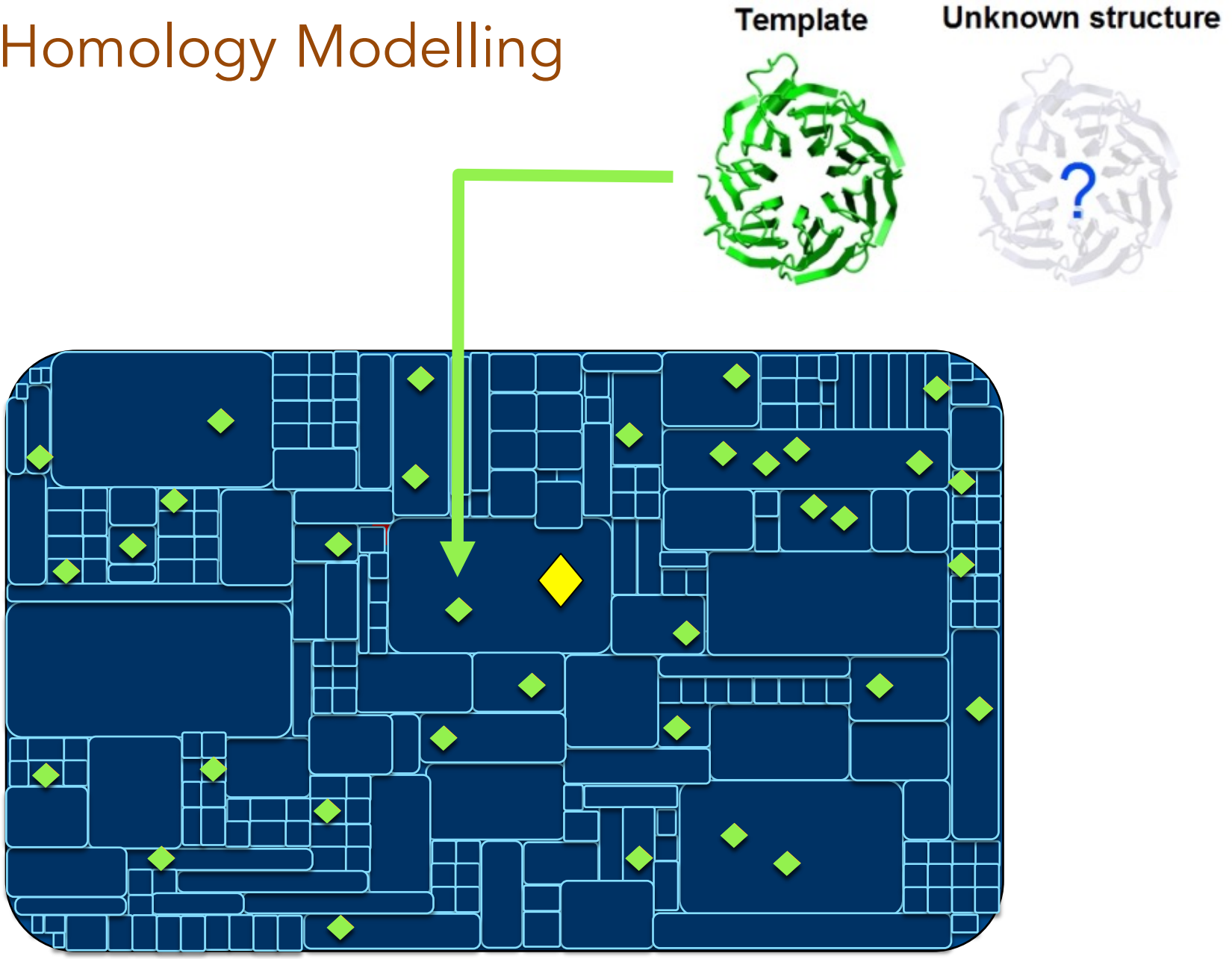
http://www.unil.ch/files/live//sites/pmf/files/shared/Technologies/Homology-modeling-450.jpg

Homology Modelling



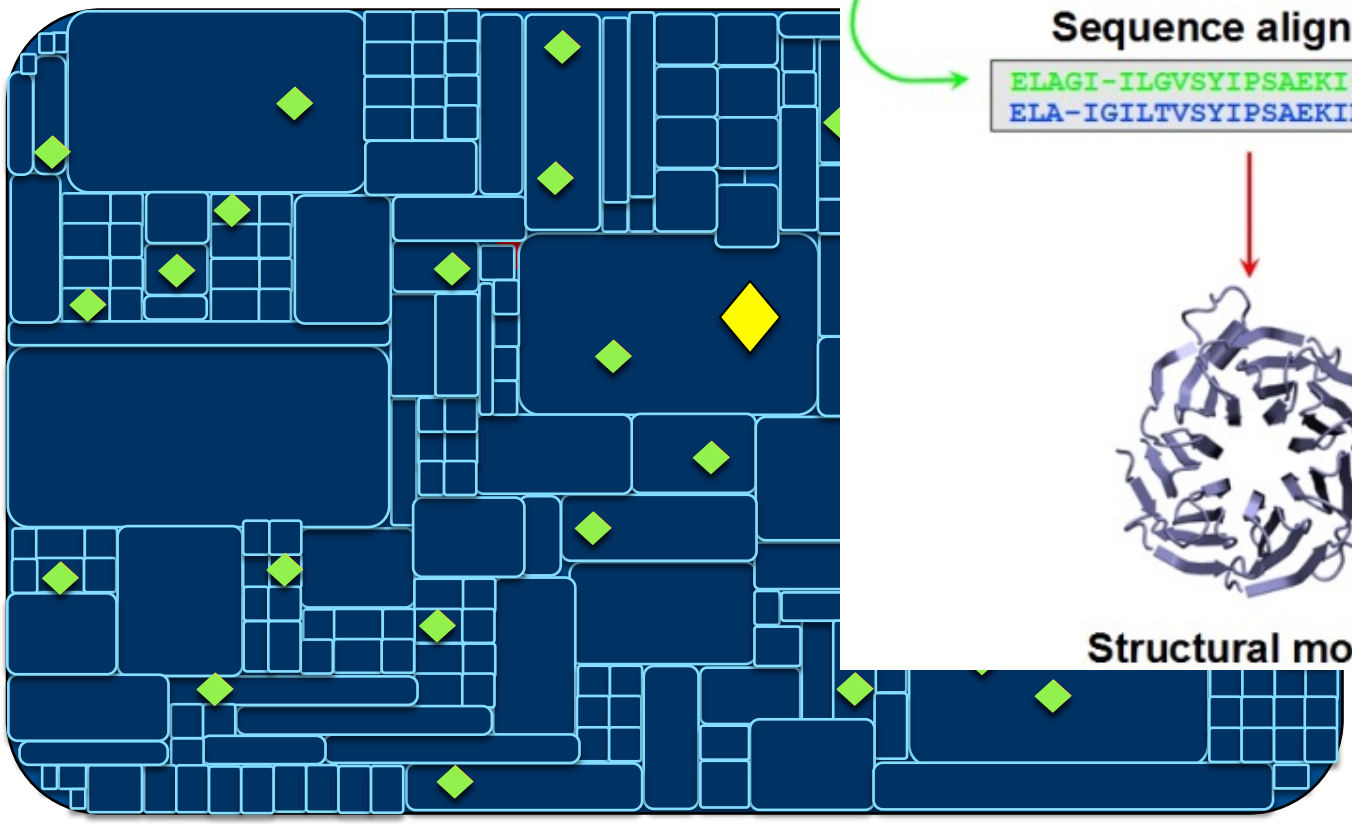
http://www.unil.ch/files/live//sites/pmf/files/shared/Technologies/Homology-modeling-450.jpg

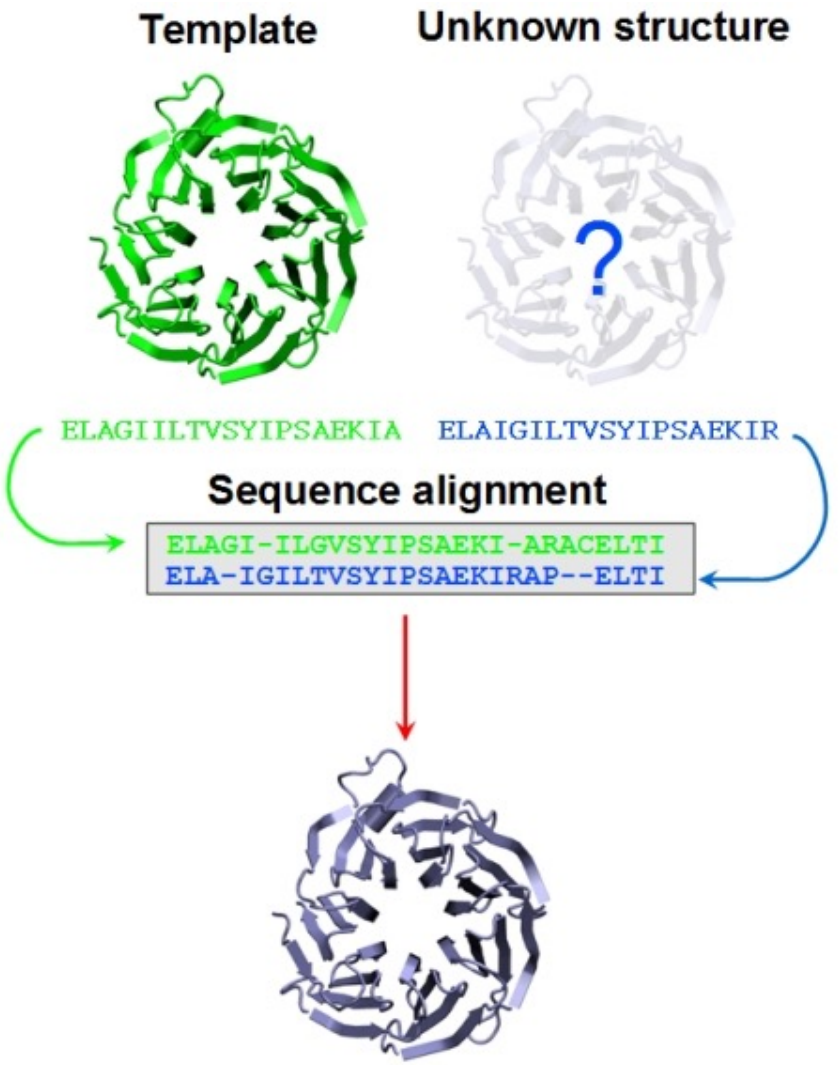
Homology Modelling



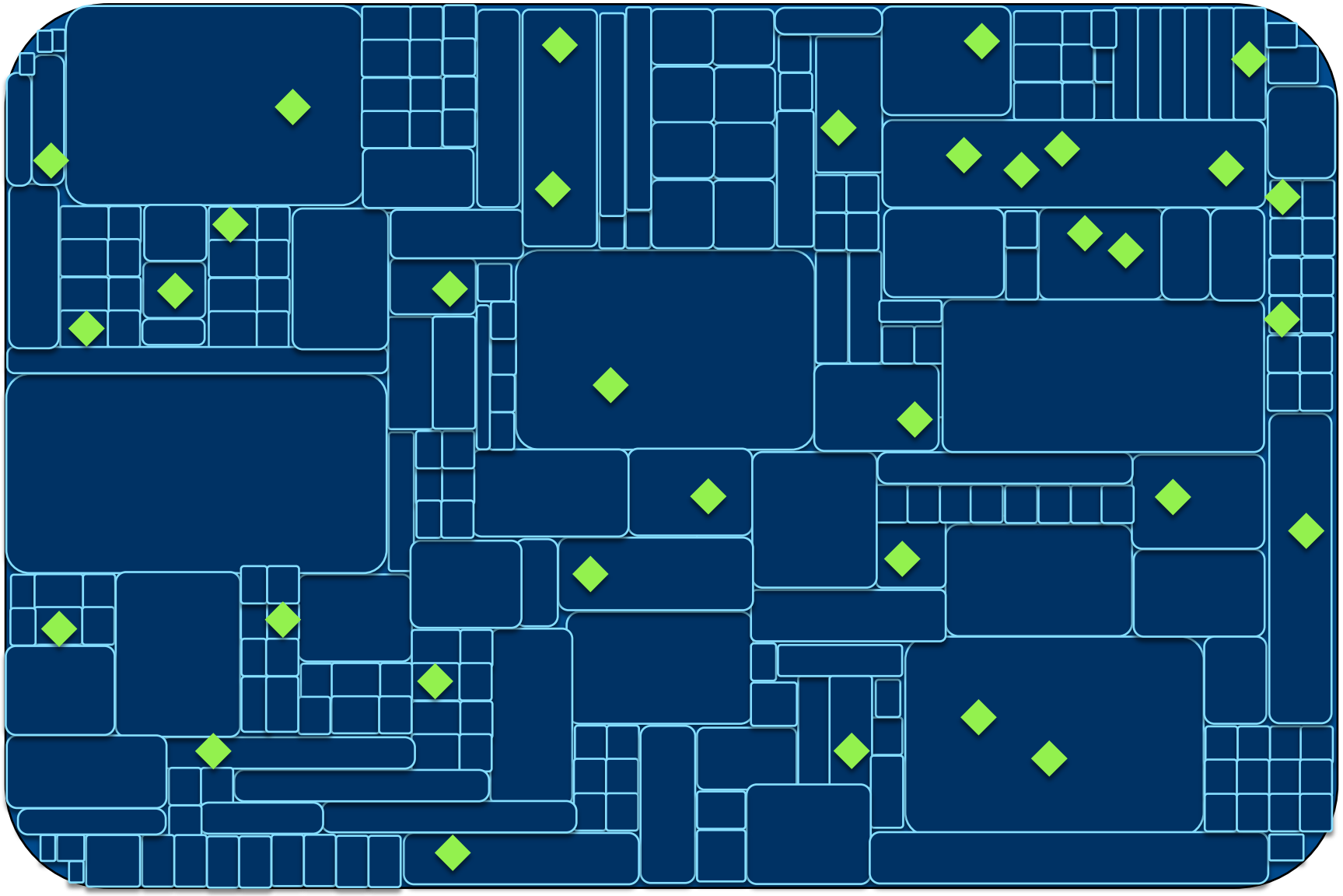
http://www.unil.ch/files/live//sites/pmf/files/shared/Technologies/Homology-modeling-450.jpg

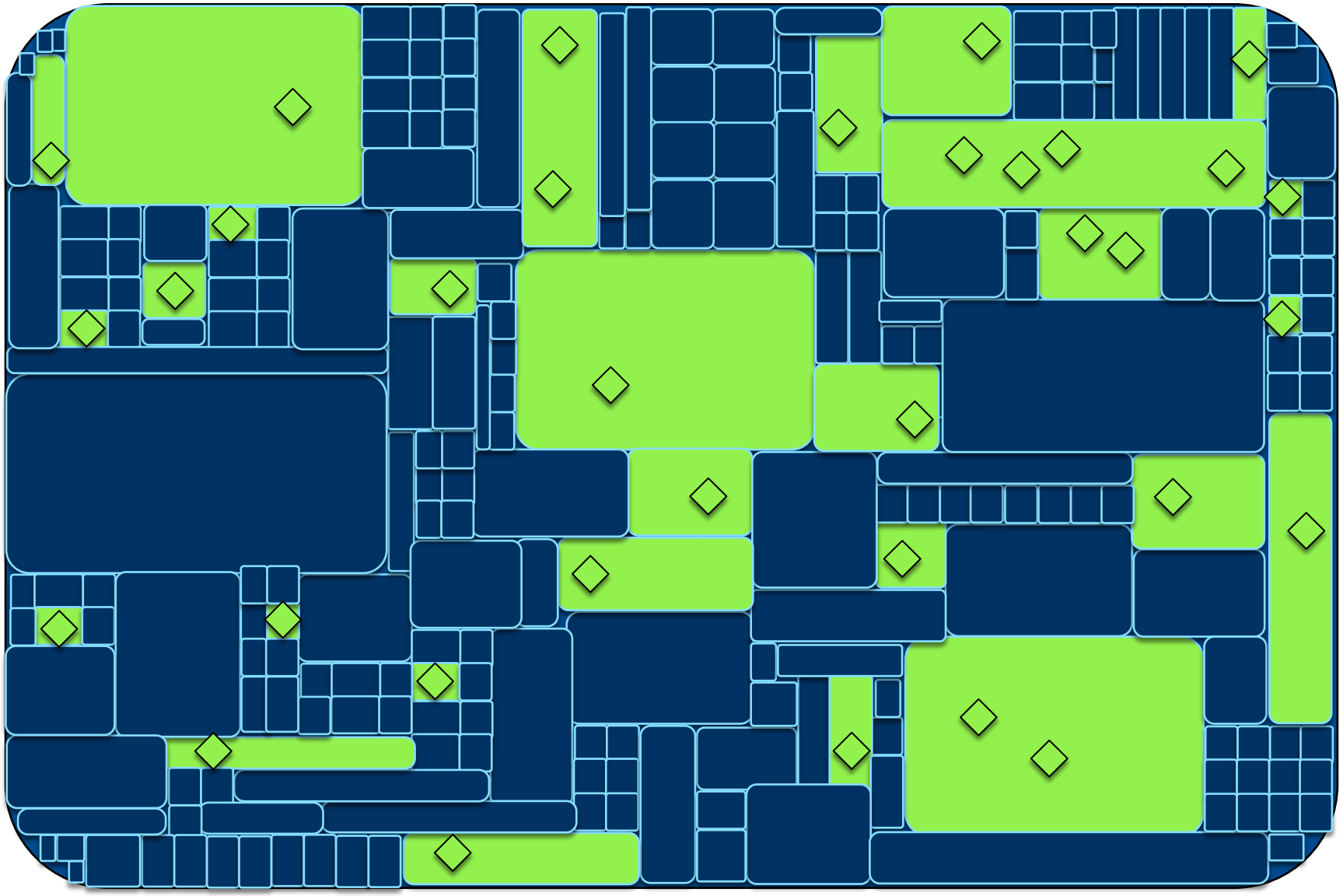
Homology Modelling





http://www.unil.ch/files/live//sites/pmf/files/shared/Technologies/Homology-modeling-450.jpg

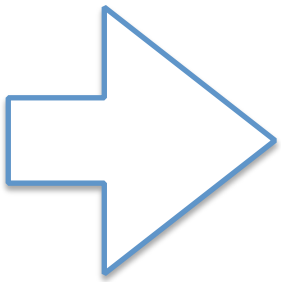




Marco Punta

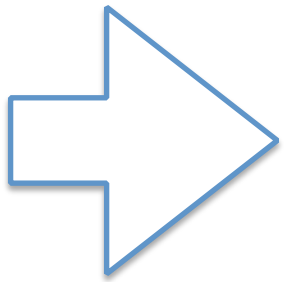
EMBO Workshop, Budapest, 2016

Homology



Structural similarity
(same fold)

Different fold



Not Homologous

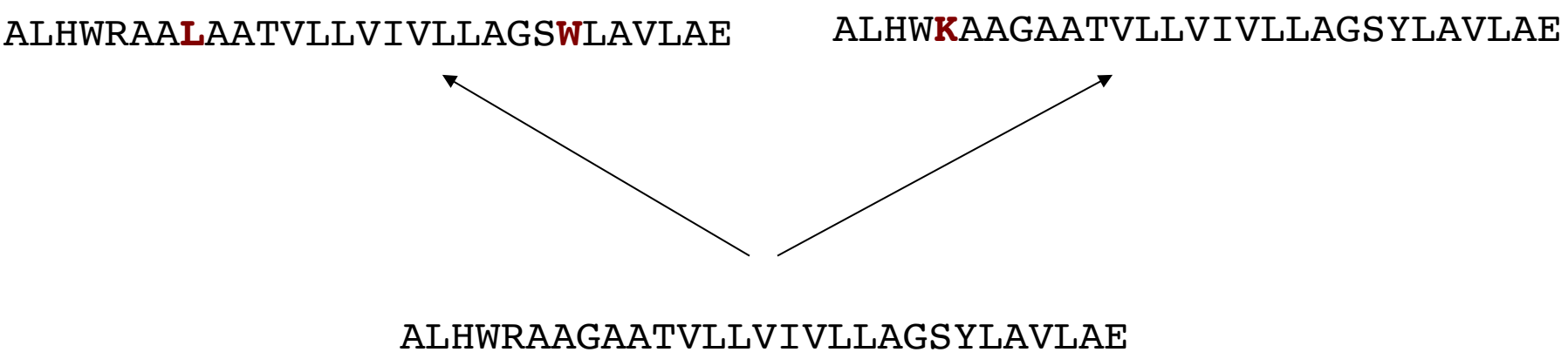
Marco Punta

Detecting homology

From sequence

Sequences of homologous proteins are related by an evolutionary process, they diverged from a common ancestor.

Modern day homologous proteins have evolved from the same ancestral sequence via a number of events (mutations, insertions, deletions, truncations,…)



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

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

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

84% identities!

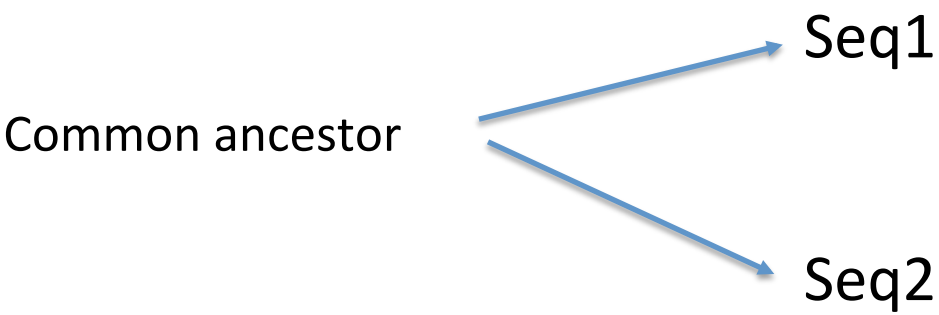
```
Human: 1   MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
           MGLSDGEWQLVLNVWGKVEAD  GHGQEVLI LFK HPETL KFDKFK LKSE  MK SE
Mouse: 1   MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSE 60

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

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

84% identities!

How can we explain “excess sequence similarity”* with respect to what we expect at random?

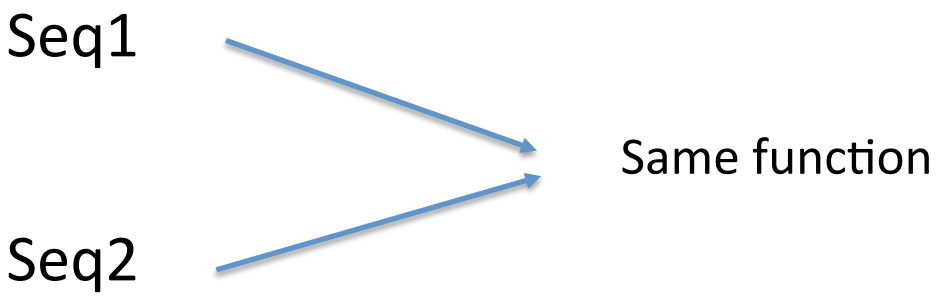
*Pearson Curr Protoc Bioinformatics 2013

Homology: divergent evolution

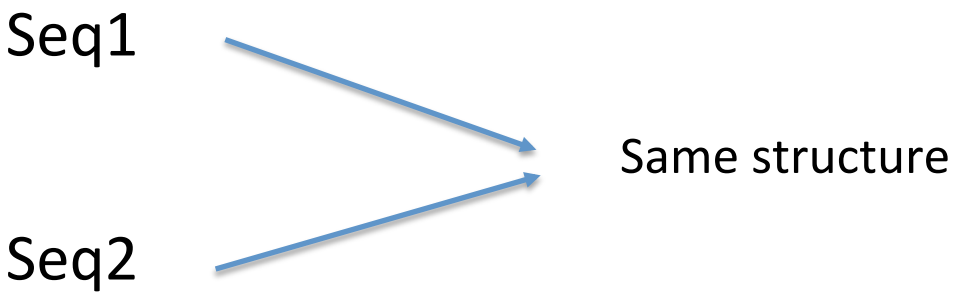


Stevens J Mol Recogn (2007)
Elias and Tawfik *J Biol Chem* (2012)

Analogy I: convergent evolution



Analogy II: concurrence

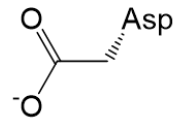


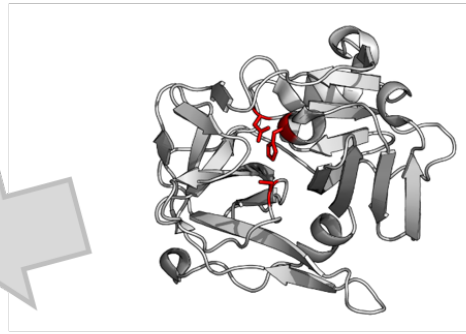
Stevens J Mol Recogn (2007)
Elias and Tawfik *J Biol Chem* (2012)

Marco Punta

Convergent evolution

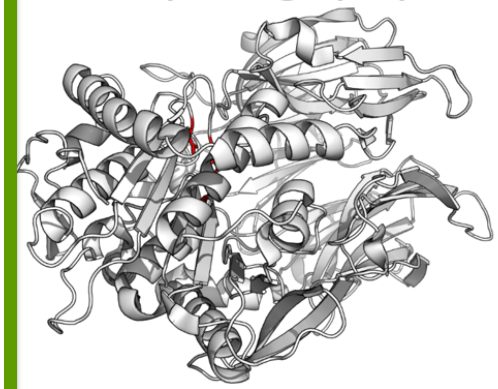
Subtilisin

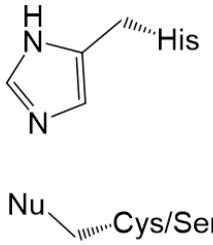


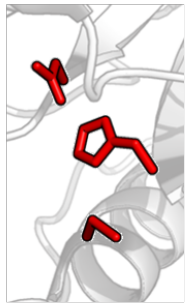


TEV protease

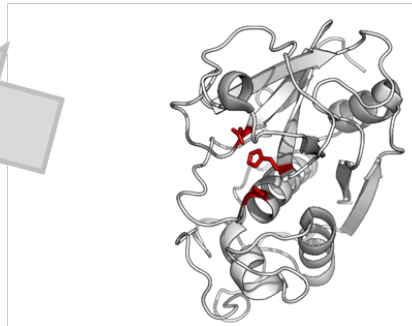
Prolyl oligopeptidase







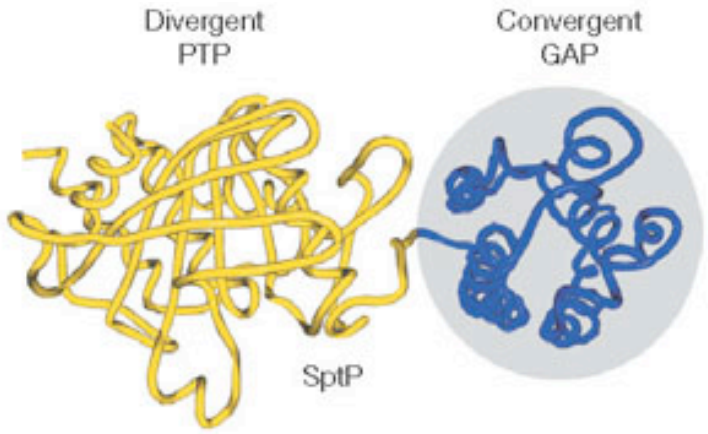
Papain



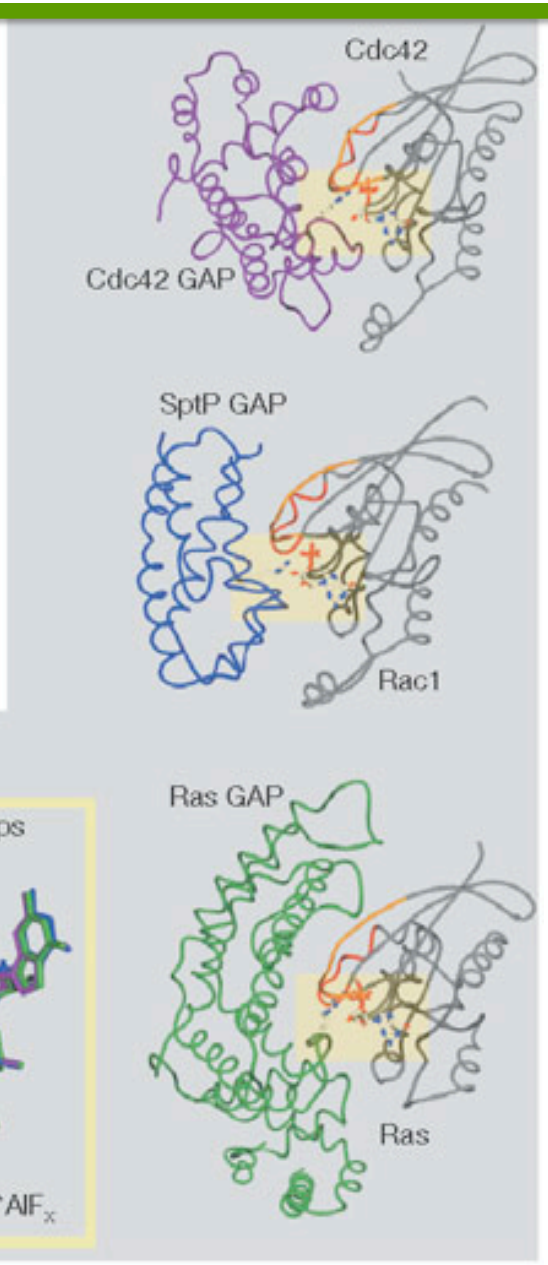
"Triad Convergence" by Thomas Shafee - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Triad_Convergence.png#/media/File:Triad_Convergence.png

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

Tyrosine phosphatase Rho GTPase activating domain



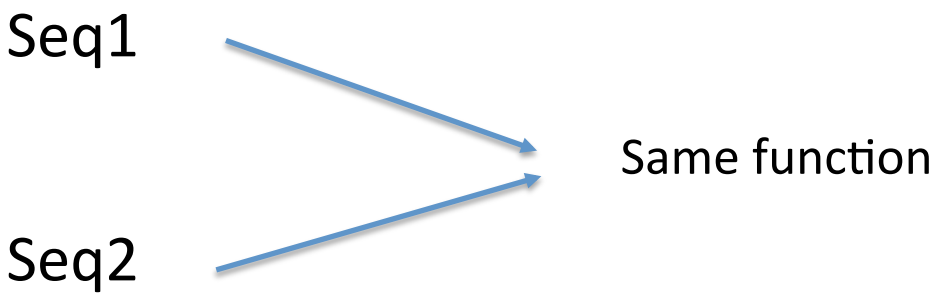
Salmonella virulence factor SptP



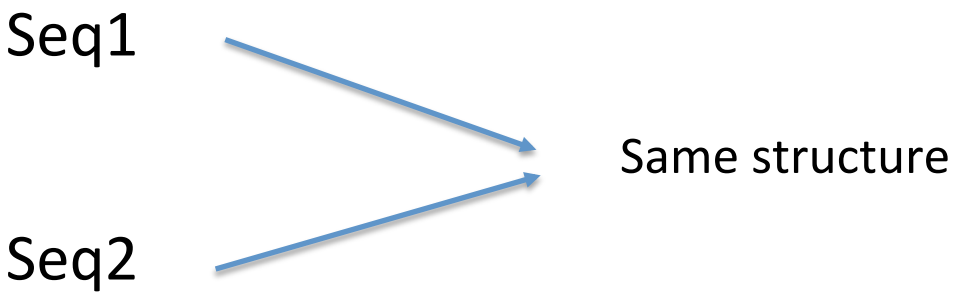
Activation of GTPases by Salmonella GAP leads to profuse rearrangements of the actin cytoskeleton and subsequent bacterial internalization into intestinal epithelial cells.

Stebbins and Galán Nature 412 (2001)

Analogy I: convergent evolution

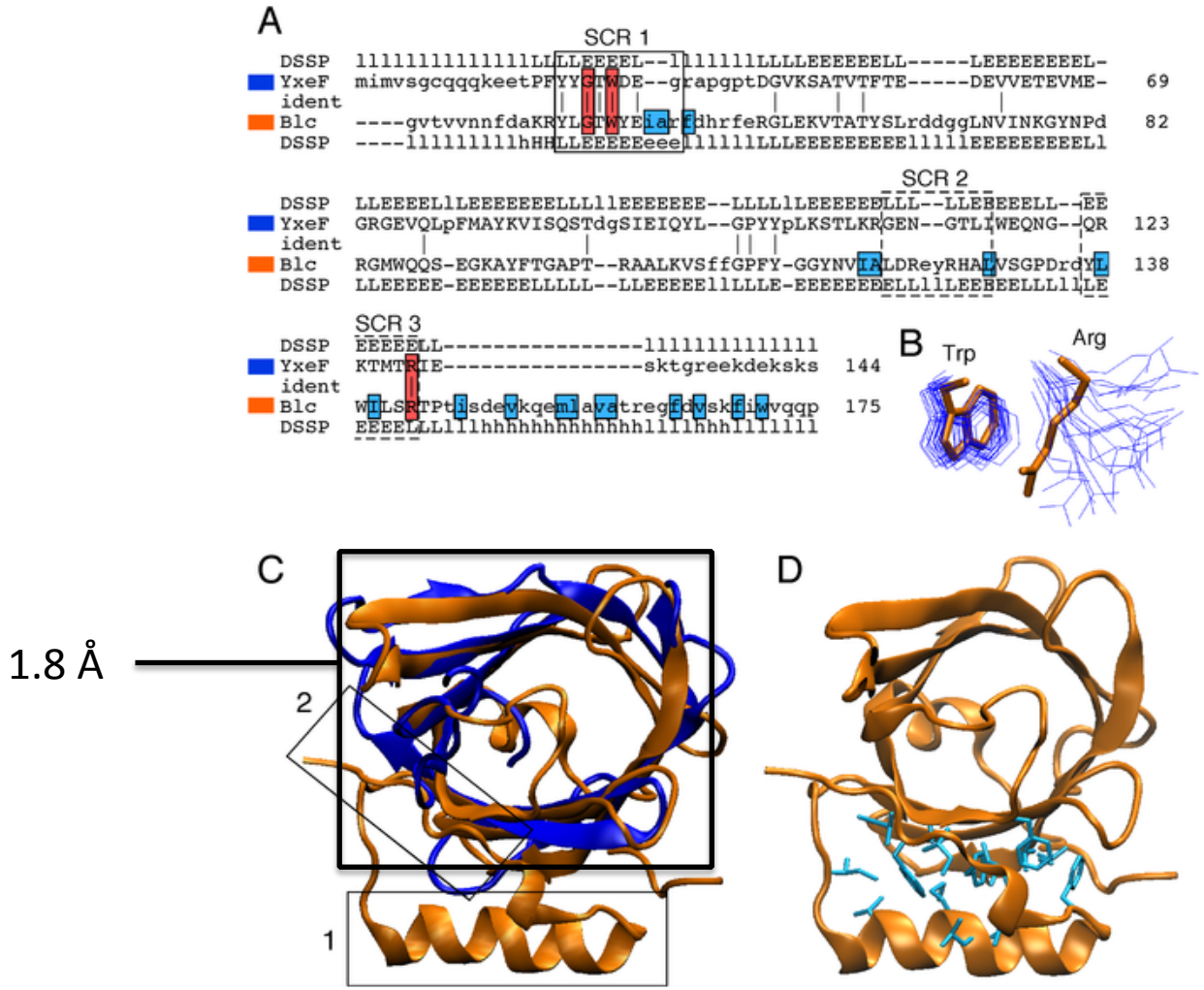


Analogy II: concurrence



Stevens J Mol Recogn (2007)
Elias and Tawfik *J Biol Chem* (2012)

Figure 5. Comparison of YxeF NMR structure (PDB ID 2JOZ, coded in blue) and Blc X-ray crystal structure (PDB ID 3MBT, orange).



Wu Punta et al. PLoS ONE (2012)

17% seq. id., Z-score of structural superposition (DALI) 6.9

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

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

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

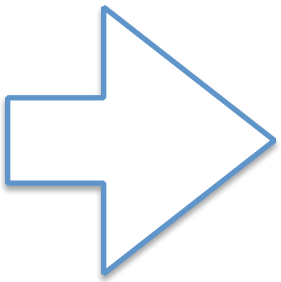
84% identities!

How can we explain “excess sequence similarity”* with respect to what we expect at random?

*Pearson Curr Protoc Bioinformatics 2013

"[…]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

Sequence alignment, 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, …

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

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

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

84% identities!

Sequence alignment, 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, …

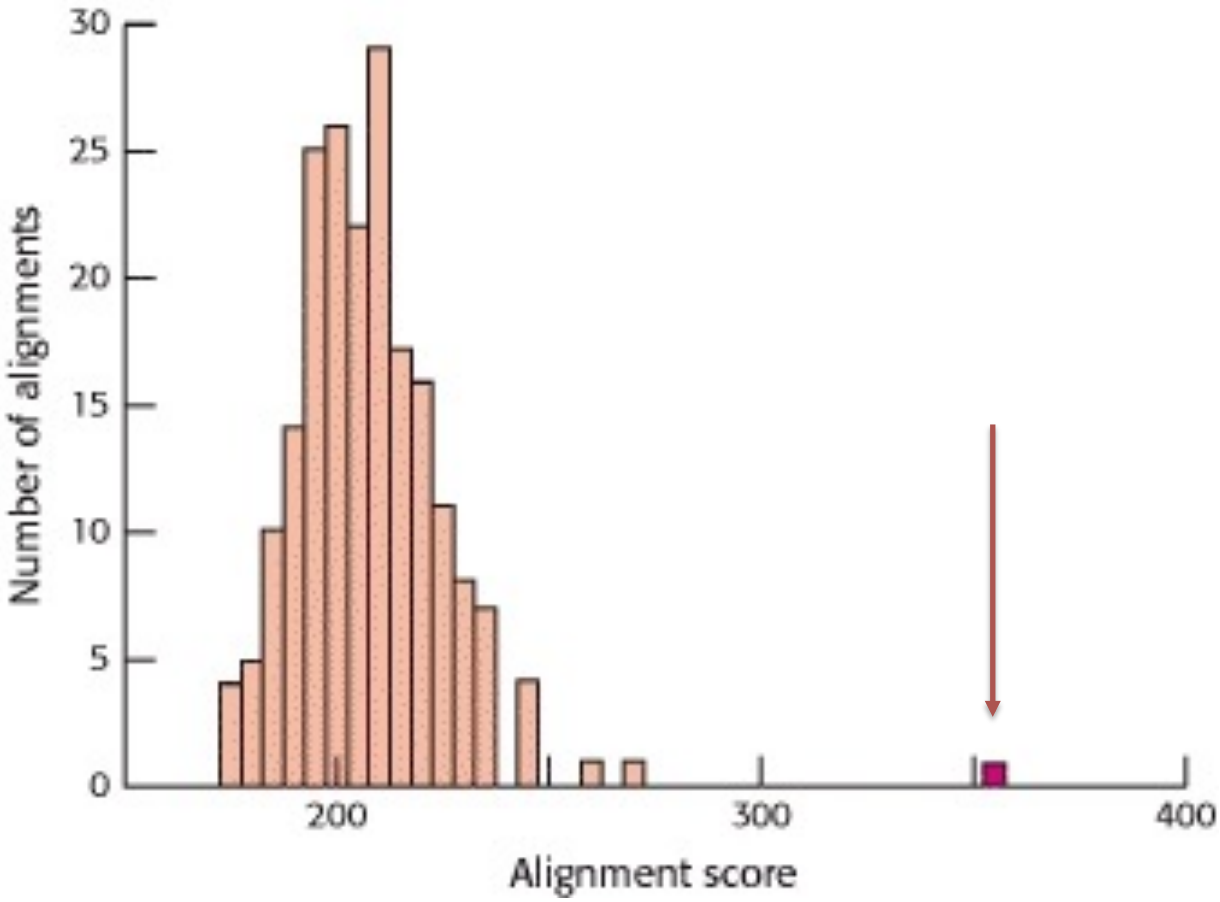


Image credits: http://www.ncbi.nlm.nih.gov/books/NBK22456/figure/A937/?report=objectonly

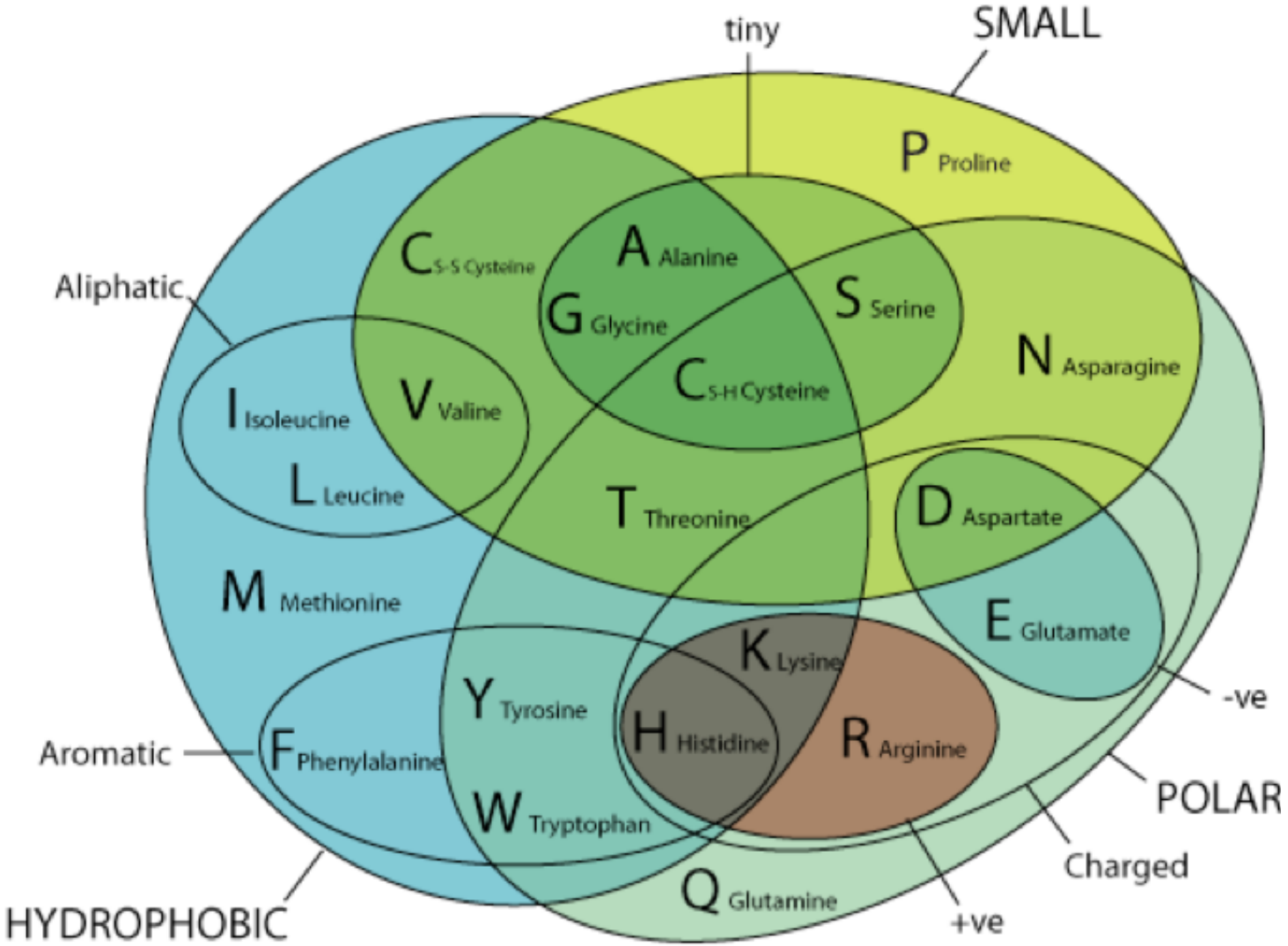
```
MYG_HUMAN      1 MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
                 M LS      V   WGKV A     G E L R F   P T   F  F      D    S
E9M4D4_HUMAN   1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSA 54

MYG_HUMAN     61 DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKI-PVKY  104
                   K H   V  AL                L   HA K    PV
E9M4D4_HUMAN  55 QVKGHSKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNF   99
```

BLOSUM62 matrix

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

aa physico-chemical properties

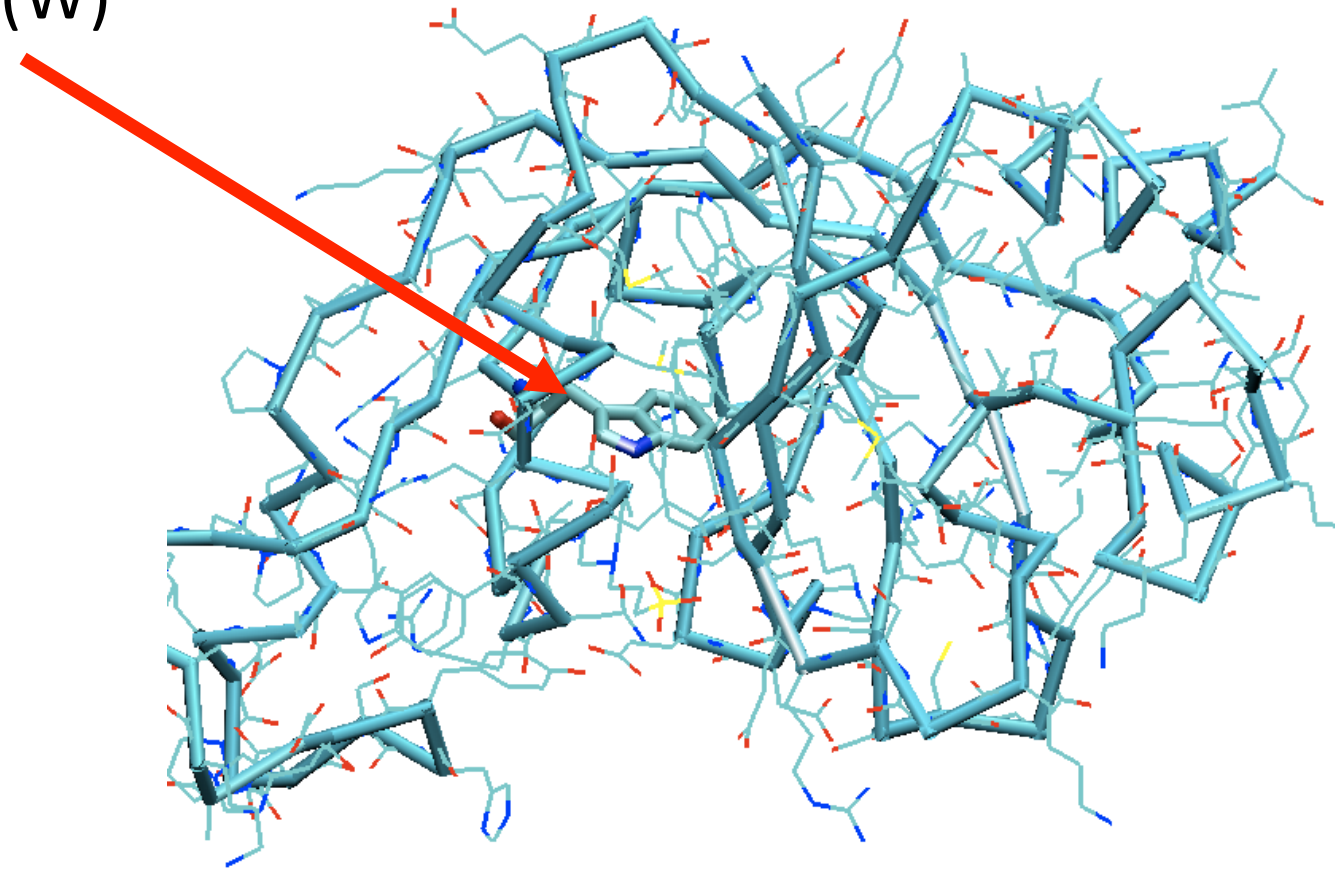
Marco Punta



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

Protein structural and functional constraints

Trp (W)



Curr Protoc Bioinformatics. Author manuscript; available in PMC 2014 Jun 1.
PMCID: PMC3820096
Published in final edited form as:
NIHMSID: NIHMS519883
Curr Protoc Bioinformatics. 2013 Jun; 0 3: 10.1002/0471250953.bi0301s42.
doi: 10.1002/0471250953.bi0301s42

An Introduction to Sequence Similarity ("Homology") Searching

William R. Pearson[1]

Author information ▶ Copyright and License information ▶

Abstract

Go to:

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

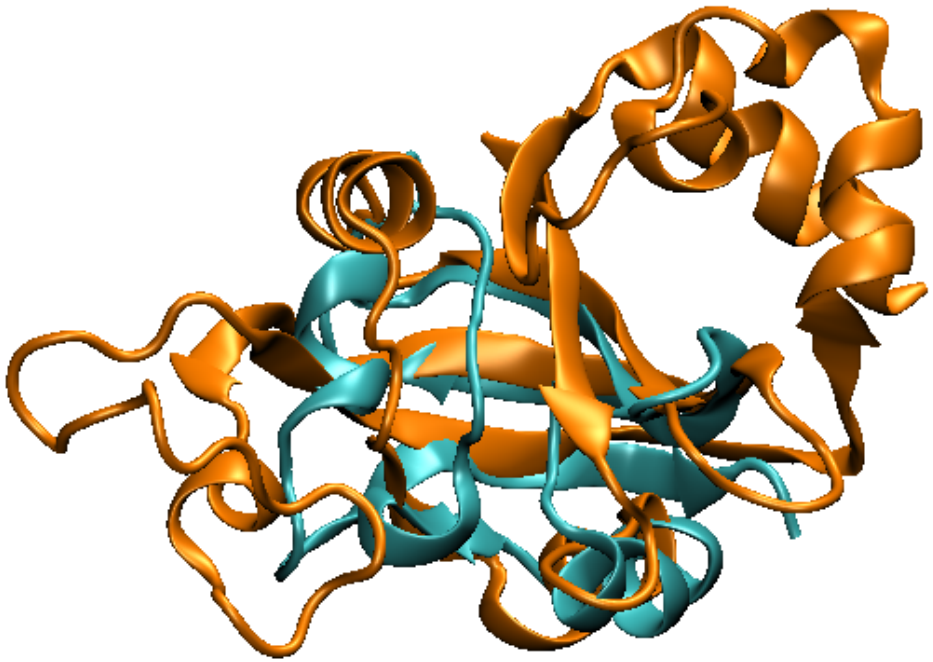
Keywords: sequence similarity, homology, orthlogy, paralogy, sequence alignment, multiple alignment, sequence evolution

Subject: Bioinformatics, Bioinformatics Fundamentals, Finding Similarities and Inferring Homologies

From structure

Marco Punta

2EVE



1J2B

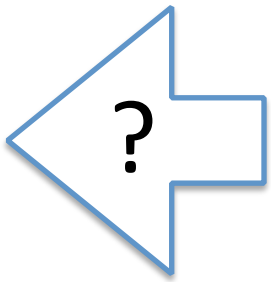
Z-score = 2.0
RMSD = 3.2
Lali = 54
%id = 6

DALI: http://ekhidna.biocenter.helsinki.fi/dali_lite/start

CE, VAST, FATCAT, PDBeFold, ...

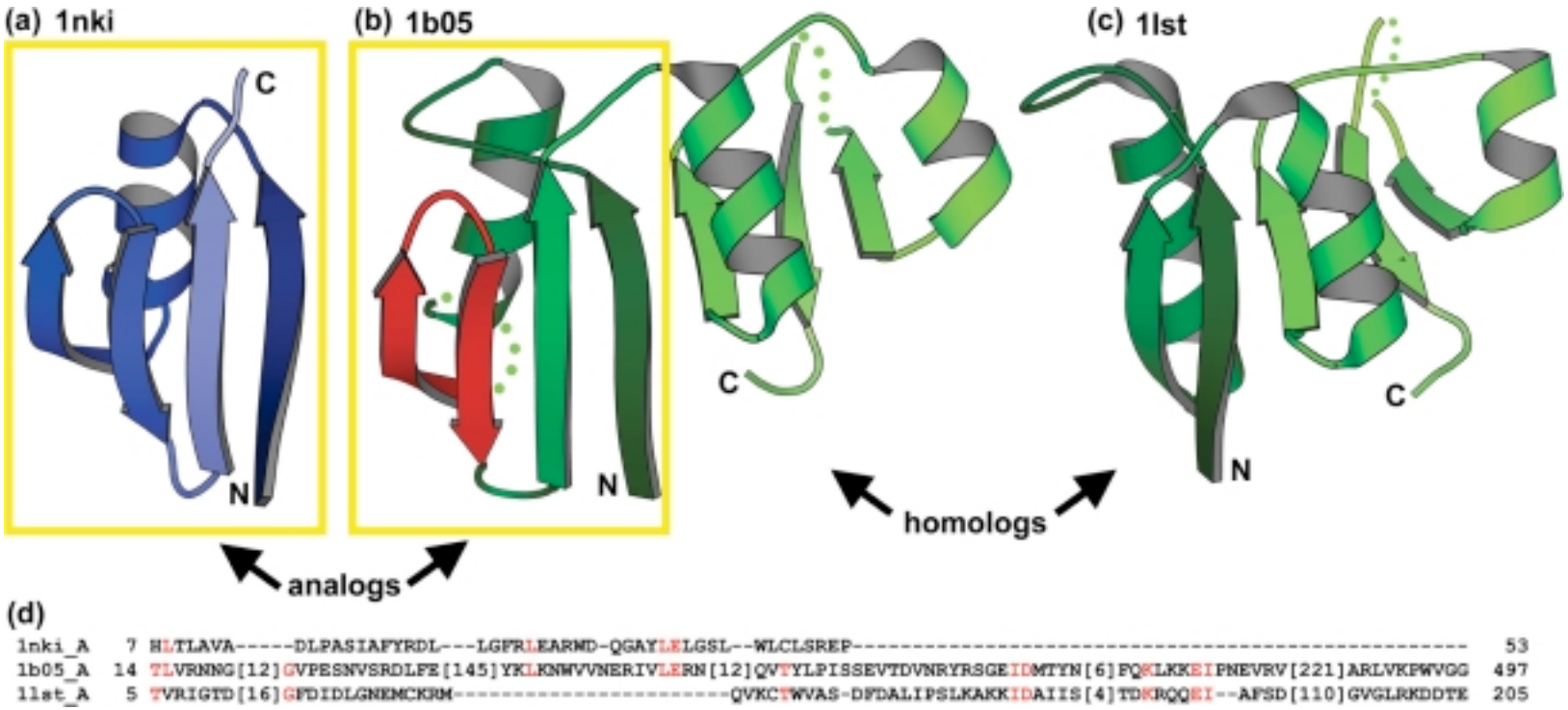
Bertonati, Punta et al. Proteins 2009

Homology

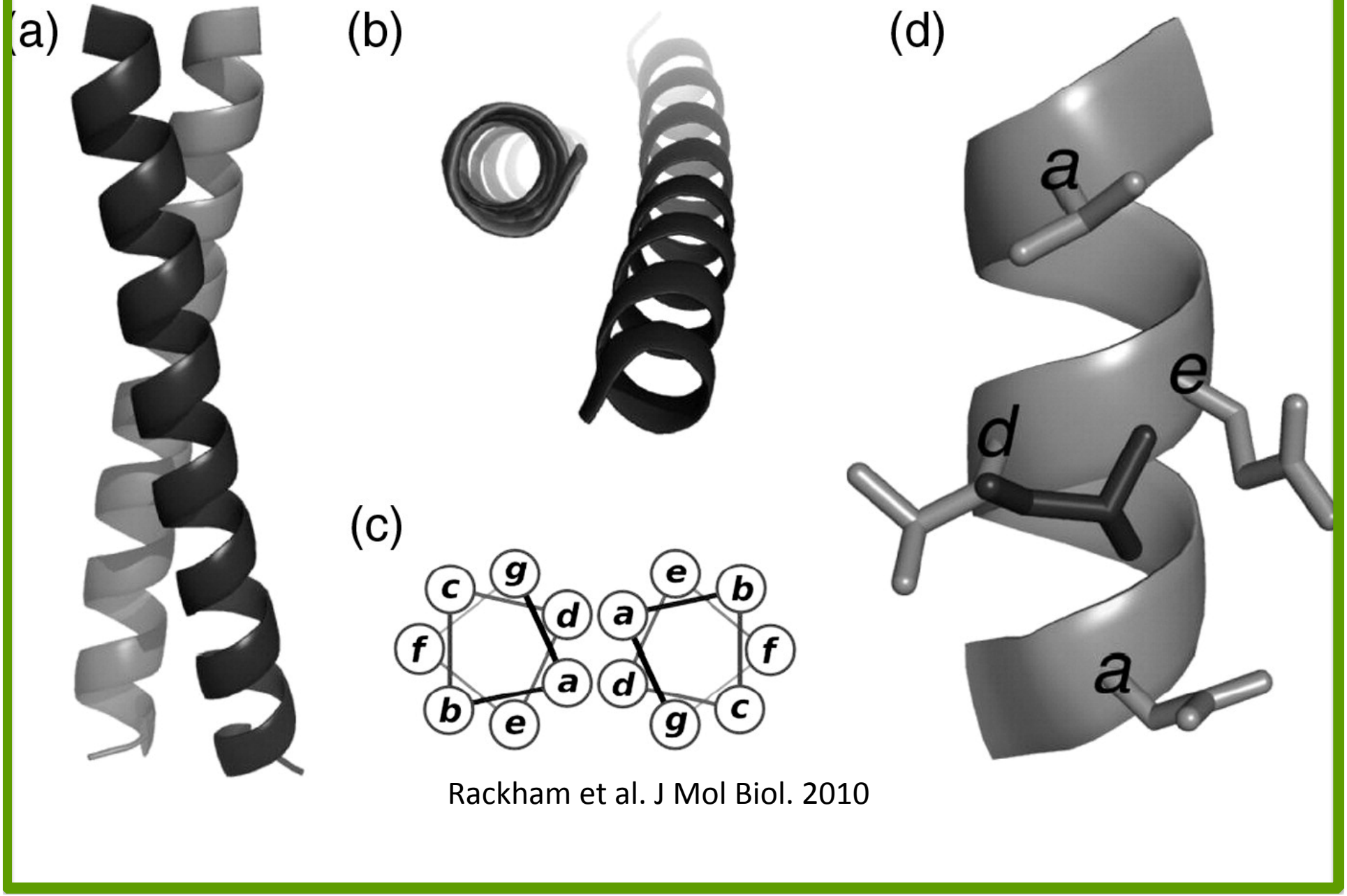


Structural similarity
(same fold)

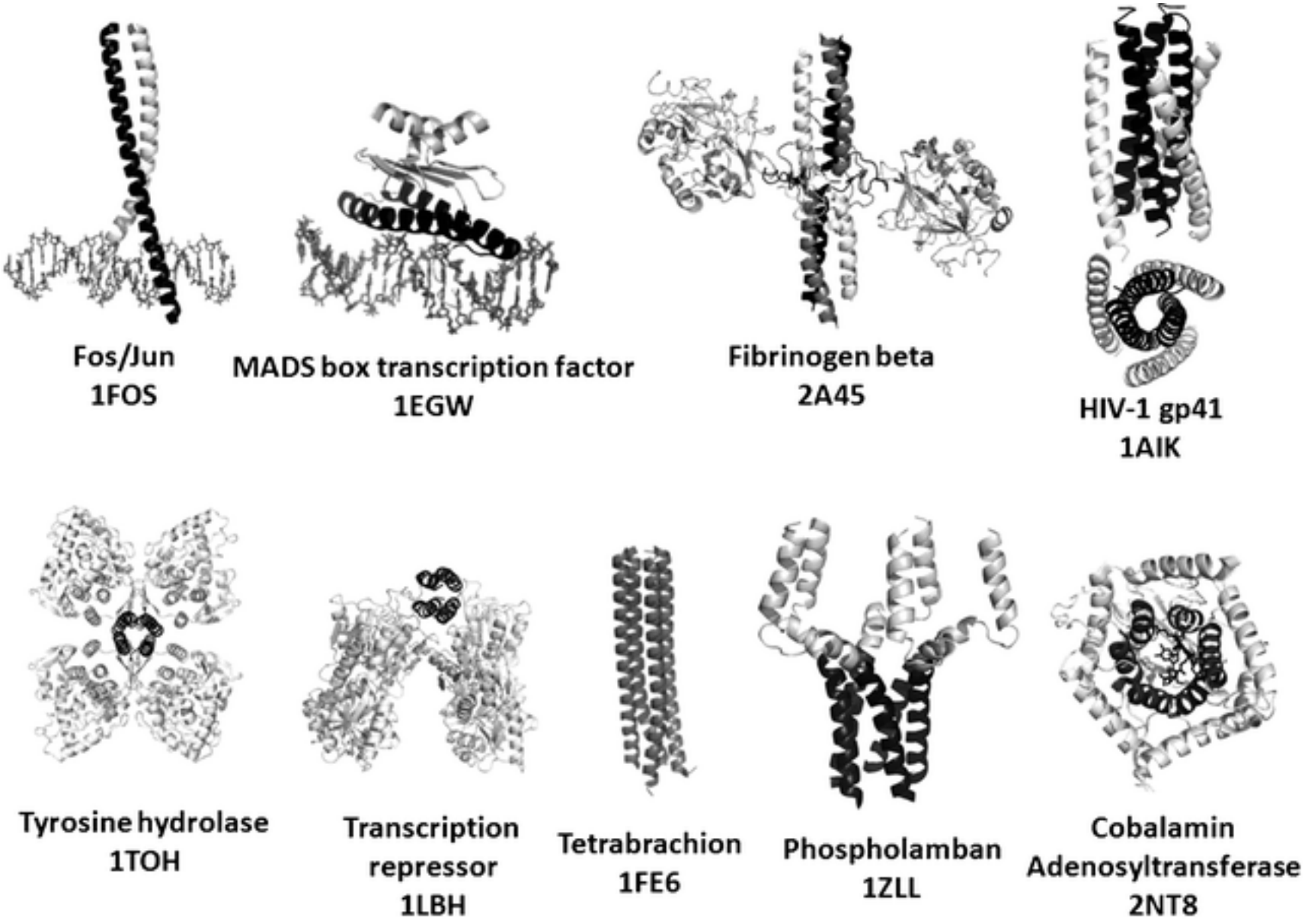
http://prodata.swmed.edu/malisam/



Cheng et al. Nucleic Acids Res. 2008

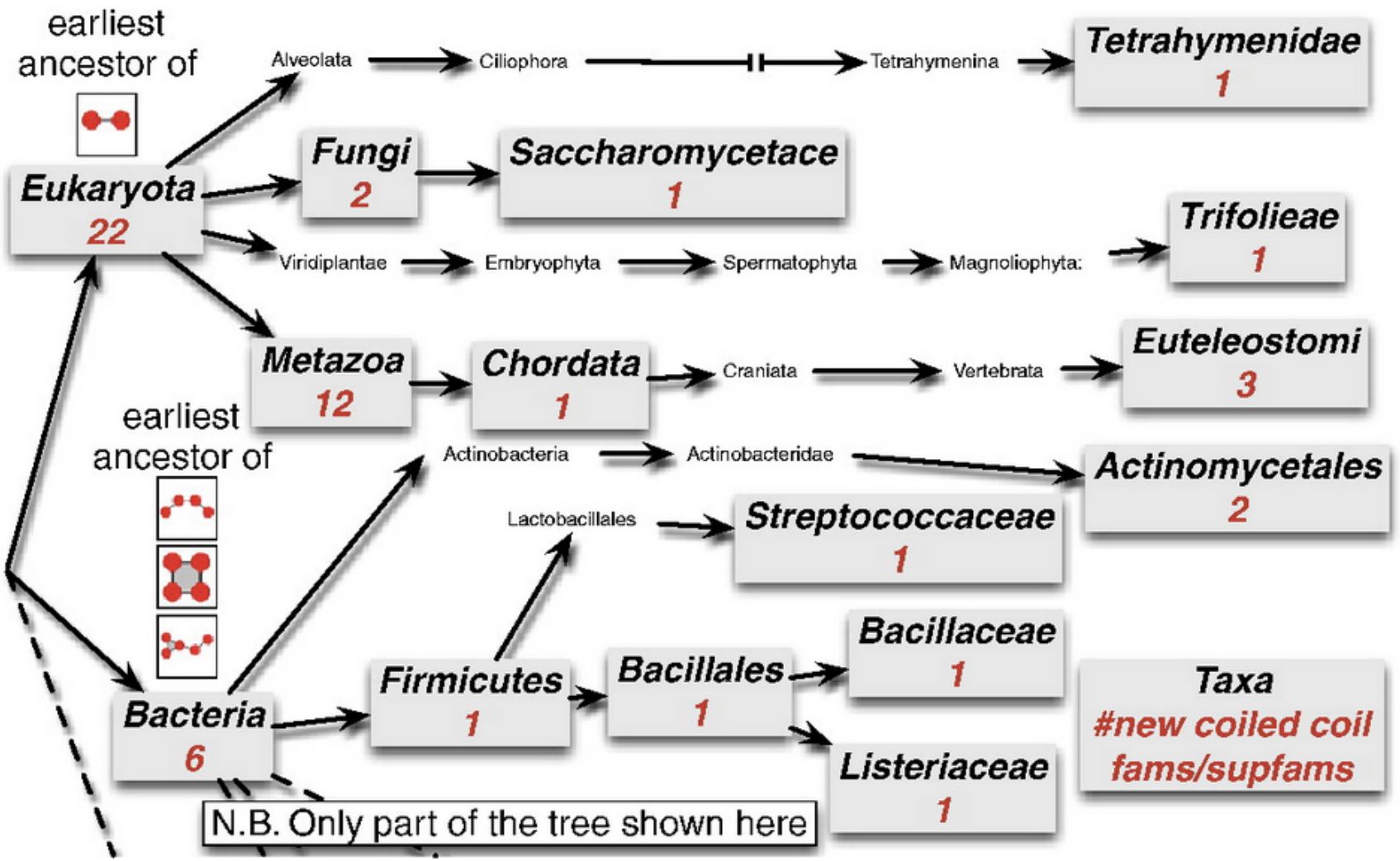


Rackham et al. J Mol Biol. 2010



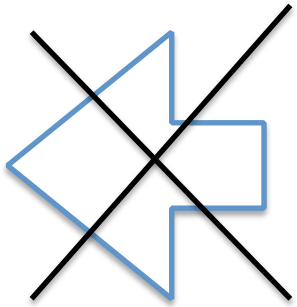
Apostolovic et al. Chem. Soc. Rev., 2010

Coiled-coil evolution



Rackham et al. J Mol Biol. 2010

Homology



Structural similarity

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

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

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

Author information

Abstract

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

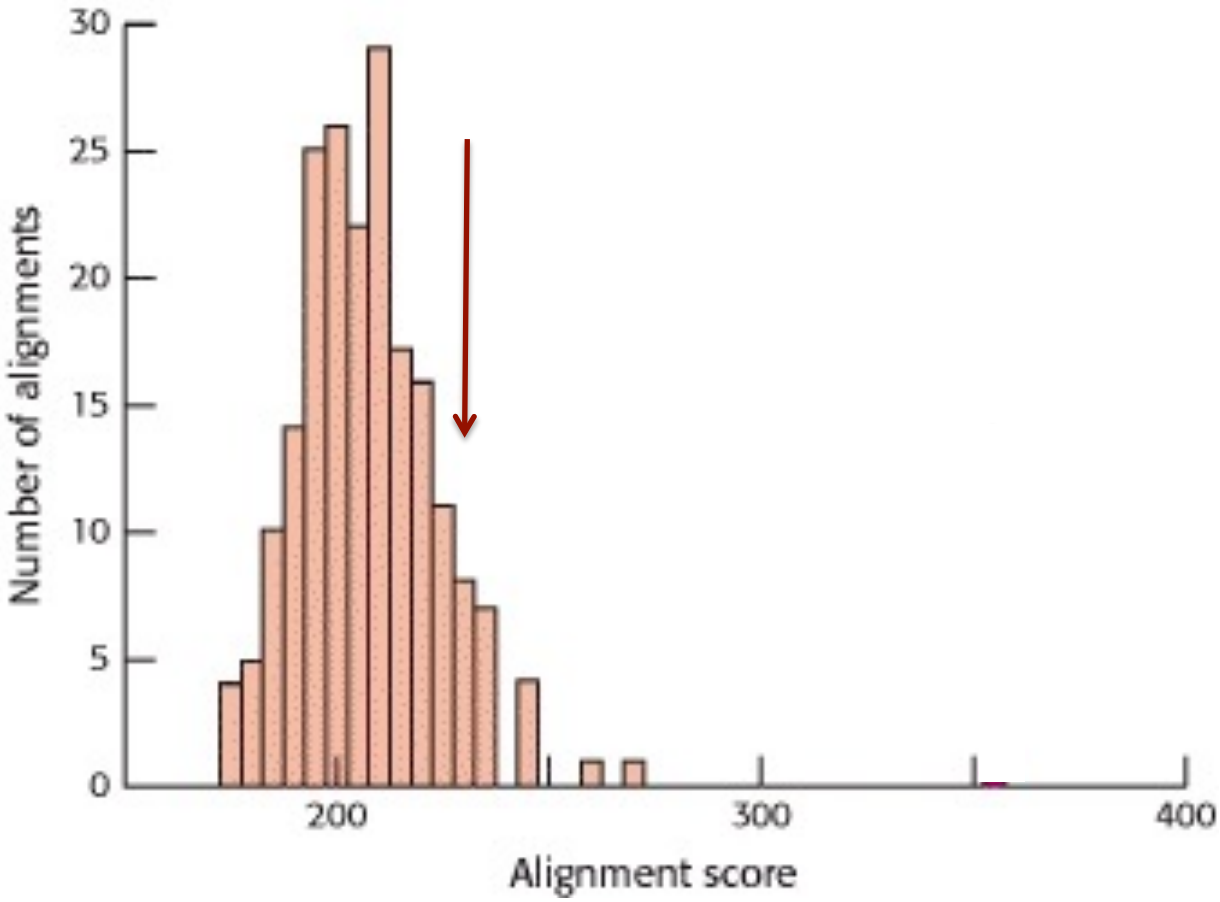
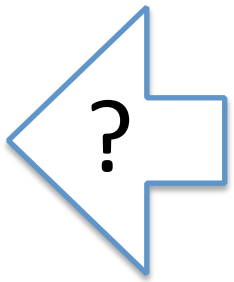
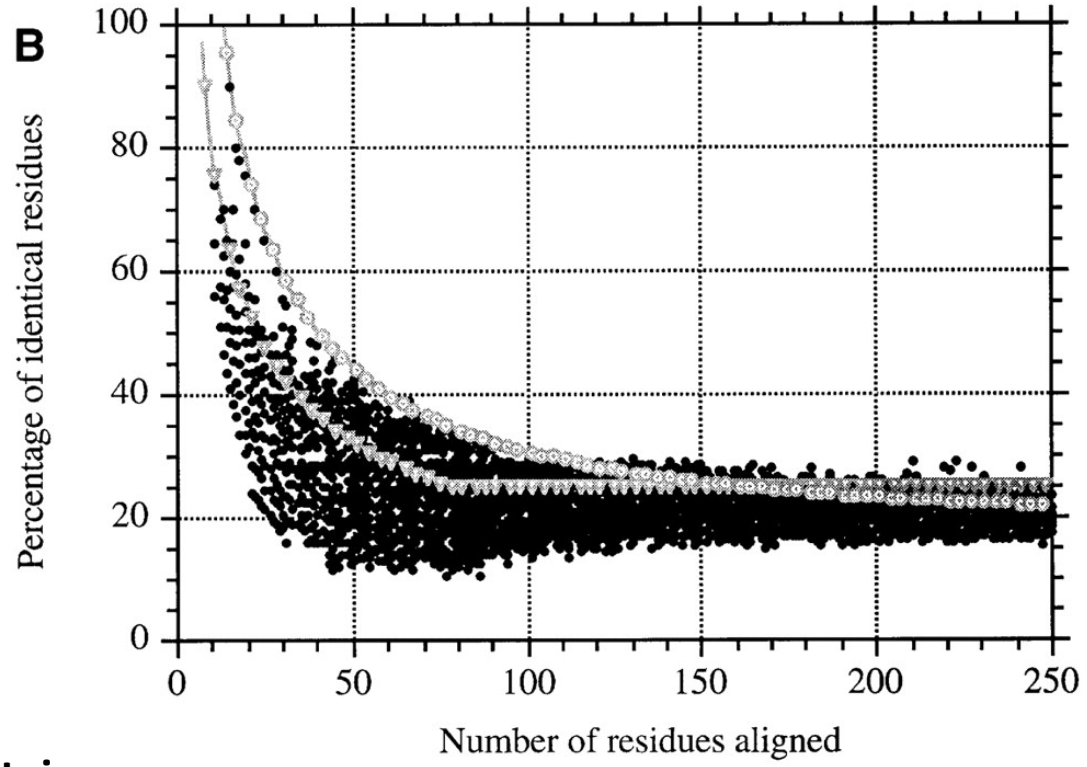


Image credits: http://www.ncbi.nlm.nih.gov/books/NBK22456/figure/A937/?report=objectonly

Excess sequence similarity  Homology

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

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

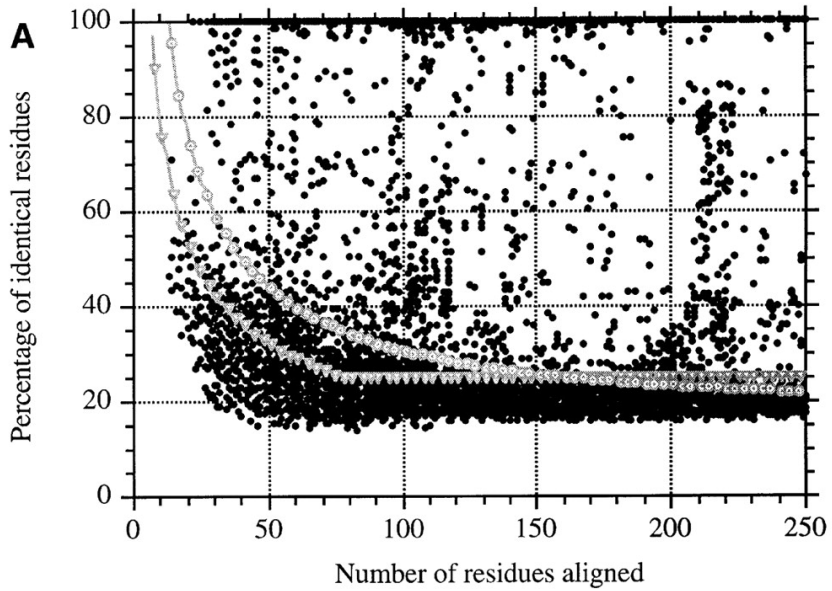


Unrelated proteins

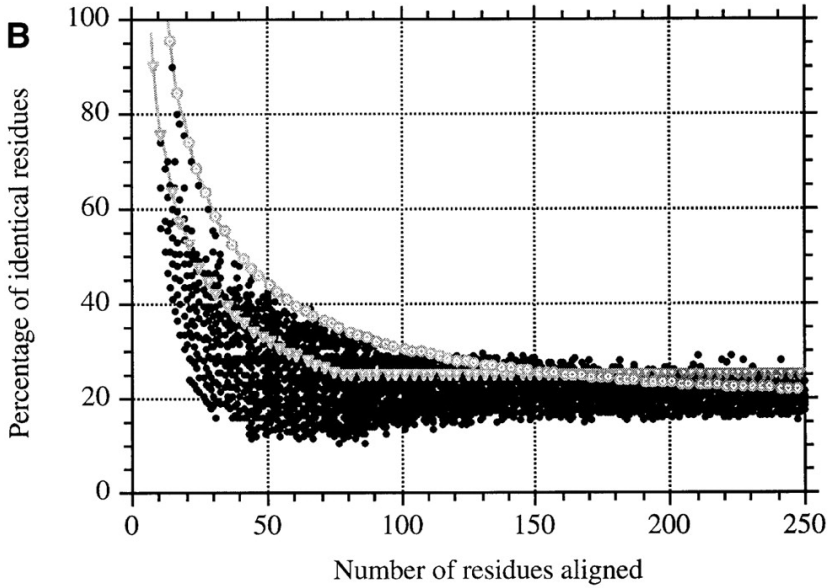
Burkhard Rost Protein Eng. 1999;12:85-94

Marco Punta

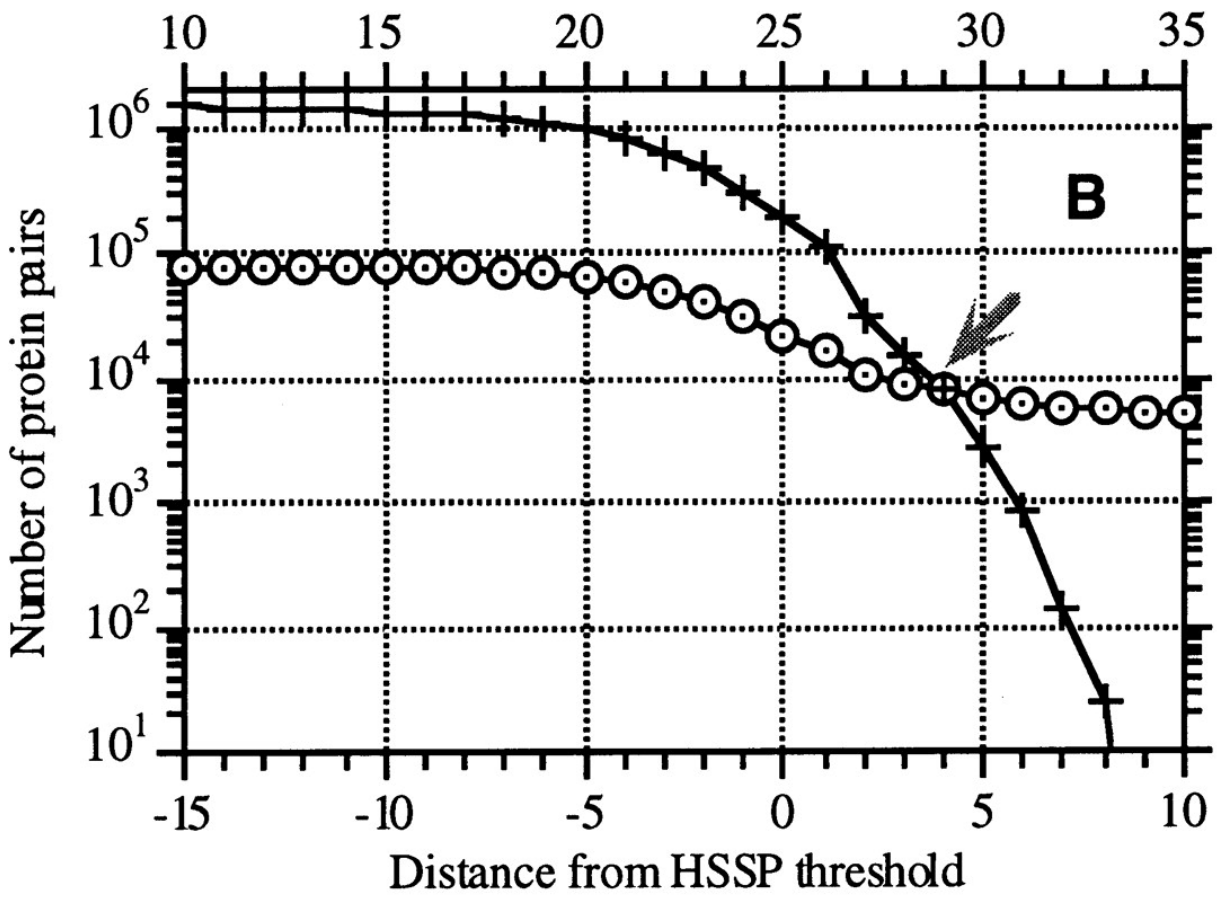
Related proteins



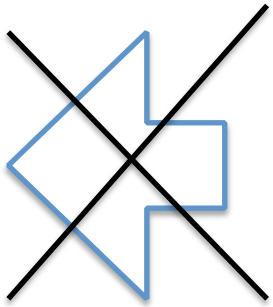
Unrelated proteins



Burkhard Rost Protein Eng. 1999;12:85-94



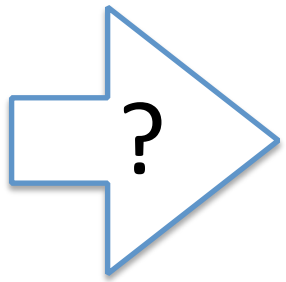
Burkhard Rost Protein Eng. 1999;12:85-94

Excess sequence similarity  Homology

Homology $\overset{\Rightarrow}{\Leftarrow}$ similar sequence?

Homology $\overset{\Rightarrow}{\Leftarrow}$ similar structure?

Homology $\overset{\Rightarrow}{\Leftarrow}$ similar function?

Homology  Similar function

Protein function(s)

Marco Punta

articles

The X-ray structure of a cobalamin biosynthetic enzyme, cobalt-precorrin-4 methyltransferase

Heidi L. Schubert[1], Keith S. Wilson[1], Evelyne Raux[2], Sarah C. Woodcock[2] and Martin J. Warren[2]

Biosynthesis of the corrin ring of vitamin B_{12} requires the action of six S-adenosyl-L-methionine (AdoMet) dependent transmethylases, closely related in sequence. The first X-ray structure of one of these, cobalt-precorrin-4 transmethylase, CbiF, from *Bacillus megaterium* has been determined to a resolution of 2.4 Å. CbiF contains two α/β domains forming a trough in which S-adenosyl-L-homocysteine (AdoHcy) binds. The location of AdoHcy and a number of conserved residues, helps define the precorrin binding site. A second crystal form determined at 3.1 Å resolution highlights the flexibility of two loops around this site. CbiF employs a unique mode of AdoHcy binding and represents a new class of transmethylase.

Schubert et al. Nat. Struct. Biol. 5 (1998)

Protein function(s)

Marco Punta

articles

The X-ray structure of a cobalamin biosynthetic enzyme, cobalt-precorrin-4 methyltransferase

Heidi L. Schubert[1], Keith S. Wilson[1], Evelyne Raux[2], Sarah C. Woodcock[2] and Martin J. Warren[2]

Biosynthesis of the corrin ring of vitamin B_{12} requires the action of six S-adenosyl-L-methionine (AdoMet) dependent transmethylases, closely related in sequence. The first X-ray structure of one of these, cobalt-precorrin-4 transmethylase, CbiF, from *Bacillus megaterium* has been determined to a resolution of 2.4 Å. CbiF contains two α/β domains forming a trough in which S-adenosyl-L-homocysteine (AdoHcy) binds. The location of AdoHcy and a number of conserved residues, helps define the precorrin binding site. A second crystal form determined at 3.1 Å resolution highlights the flexibility of two loops around this site. CbiF employs a unique mode of AdoHcy binding and represents a new class of transmethylase.

Schubert et al. Nat. Struct. Biol. 5 (1998)

Protein function(s)

Marco Punta

articles

The X-ray structure of a cobalamin biosynthetic enzyme, cobalt-precorrin-4 methyltransferase

Heidi L. Schubert[1], Keith S. Wilson[1], Evelyne Raux[2], Sarah C. Woodcock[2] and Martin J. Warren[2]

Biosynthesis of the corrin ring of vitamin B_{12} requires the action of six S-adenosyl-L-methionine (AdoMet) dependent transmethylases, closely related in sequence. The first X-ray structure of one of these, cobalt-precorrin-4 transmethylase, CbiF, from *Bacillus megaterium* has been determined to a resolution of 2.4 Å. CbiF contains two α/β domains forming a trough in which S-adenosyl-L-homocysteine (AdoHcy) binds. The location of AdoHcy and a number of conserved residues, helps define the precorrin binding site. A second crystal form determined at 3.1 Å resolution highlights the flexibility of two loops around this site. CbiF employs a unique mode of AdoHcy binding and represents a new class of transmethylase.

Schubert et al. Nat. Struct. Biol. 5 (1998)

Protein function(s)

Marco Punta

articles

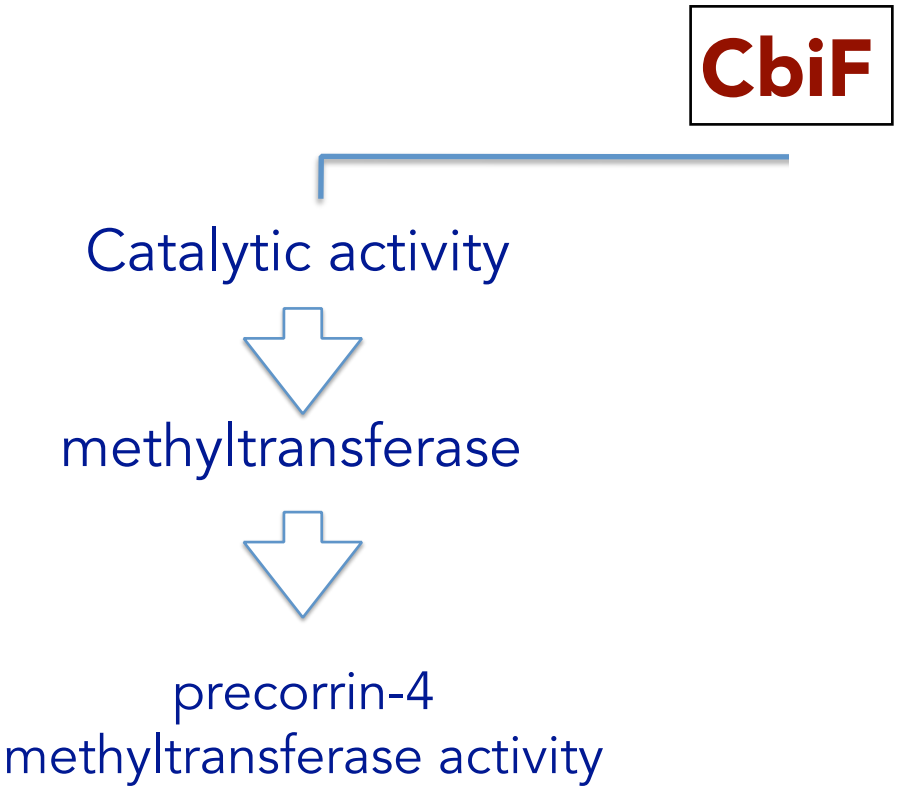
The X-ray structure of a cobalamin biosynthetic enzyme, cobalt-precorrin-4 methyltransferase

Heidi L. Schubert[1], Keith S. Wilson[1], Evelyne Raux[2], Sarah C. Woodcock[2] and Martin J. Warren[2]

Biosynthesis of the corrin ring of vitamin B_{12} requires the action of six S-adenosyl-L-methionine (AdoMet) dependent transmethylases, closely related in sequence. The first X-ray structure of one of these, cobalt-precorrin-4 transmethylase, CbiF, from *Bacillus megaterium* has been determined to a resolution of 2.4 Å. CbiF contains two α/β domains forming a trough in which S-adenosyl-L-homocysteine (AdoHcy) binds. The location of AdoHcy and a number of conserved residues, helps define the precorrin binding site. A second crystal form determined at 3.1 Å resolution highlights the flexibility of two loops around this site. CbiF employs a unique mode of AdoHcy binding and represents a new class of transmethylase.

Schubert et al. Nat. Struct. Biol. 5 (1998)

Protein function(s)



Protein function(s)

Marco Punta

articles

The X-ray structure of a cobalamin biosynthetic enzyme, cobalt-precorrin-4 methyltransferase

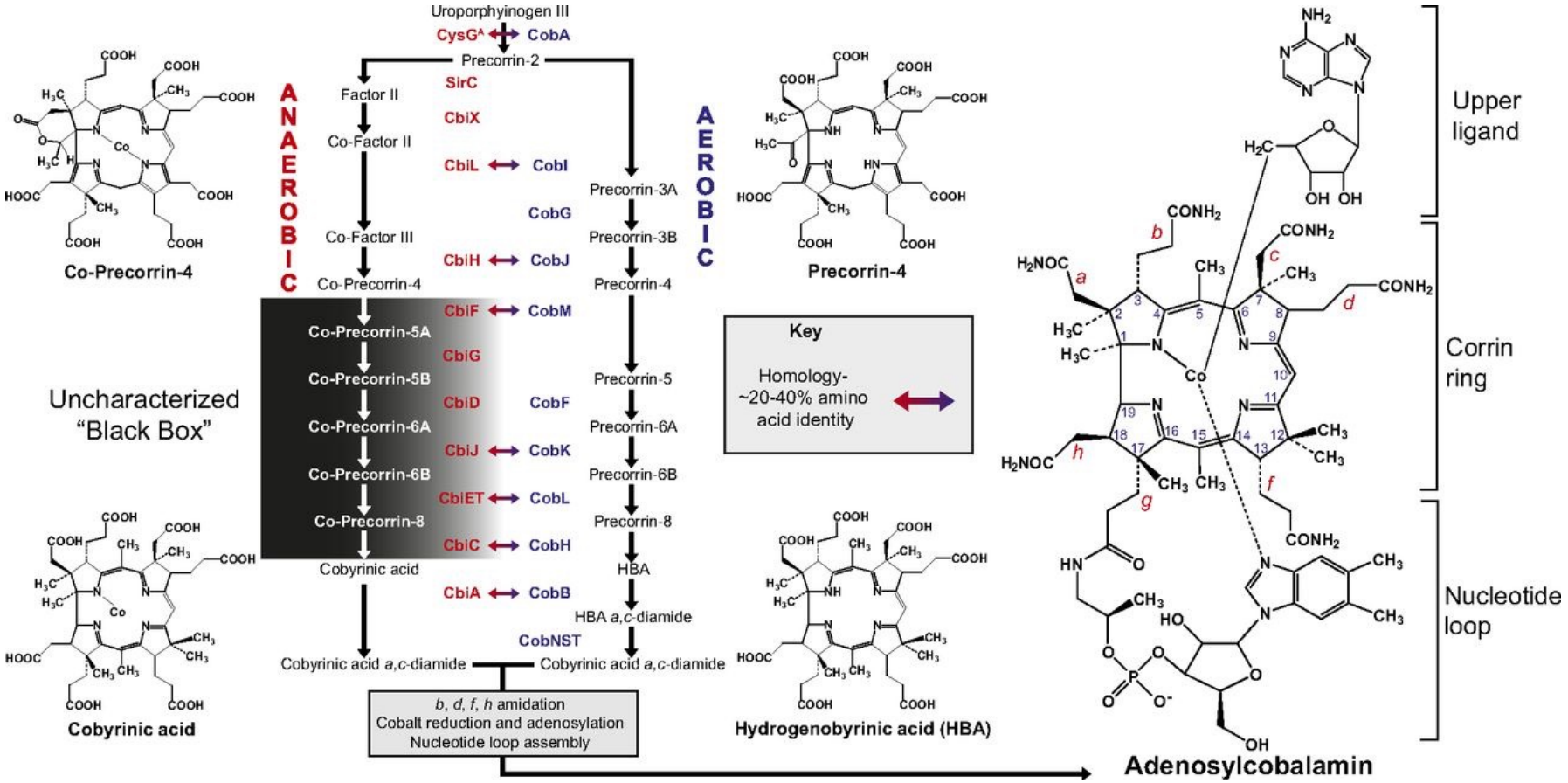
Heidi L. Schubert[1], Keith S. Wilson[1], Evelyne Raux[2], Sarah C. Woodcock[2] and Martin J. Warren[2]

Biosynthesis of the corrin ring of vitamin B_{12} requires the action of six S-adenosyl-L-methionine (AdoMet) dependent transmethylases, closely related in sequence. The first X-ray structure of one of these, cobalt-precorrin-4 transmethylase, CbiF, from *Bacillus megaterium* has been determined to a resolution of 2.4 Å. CbiF contains two α/β domains forming a trough in which S-adenosyl-L-homocysteine (AdoHcy) binds. The location of AdoHcy and a number of conserved residues, helps define the precorrin binding site. A second crystal form determined at 3.1 Å resolution highlights the flexibility of two loops around this site. CbiF employs a unique mode of AdoHcy binding and represents a new class of transmethylase.

Schubert et al. Nat. Struct. Biol. 5 (1998)

Cobalamin biosynthetic pathways

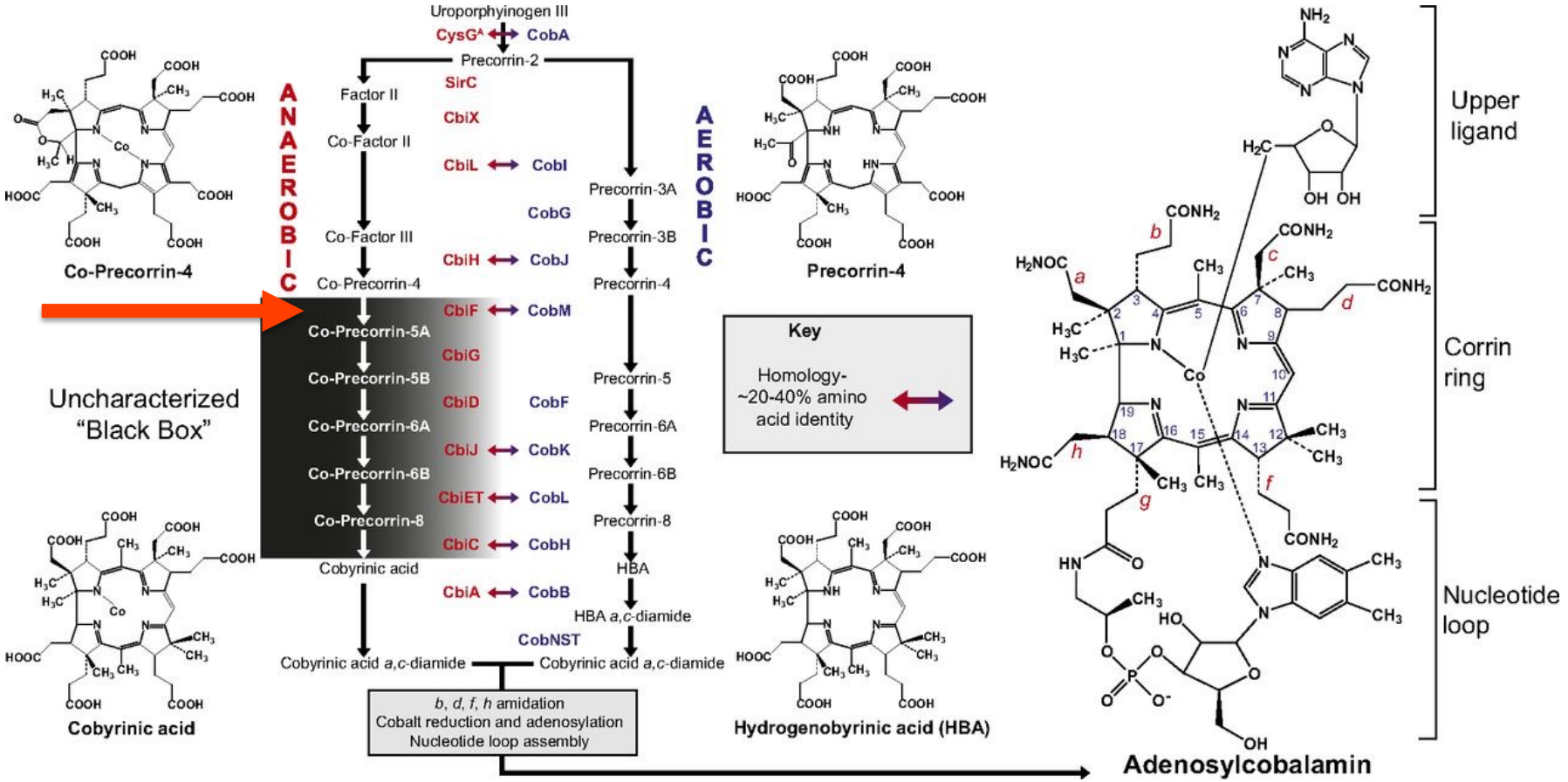
Marco Punta



Moore et al. PNAS 110 (2013)

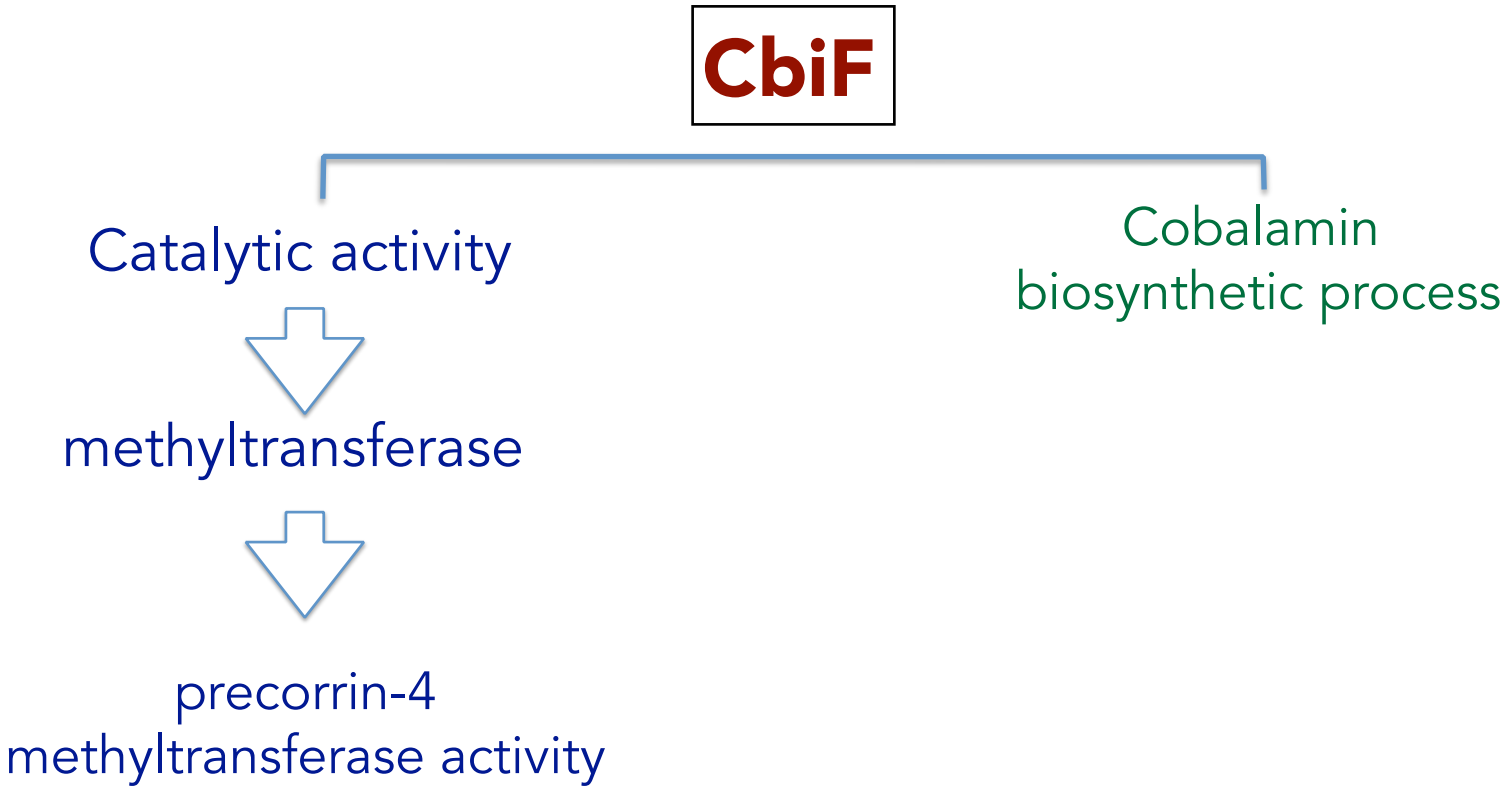
Cobalamin biosynthetic pathways

Marco Punta



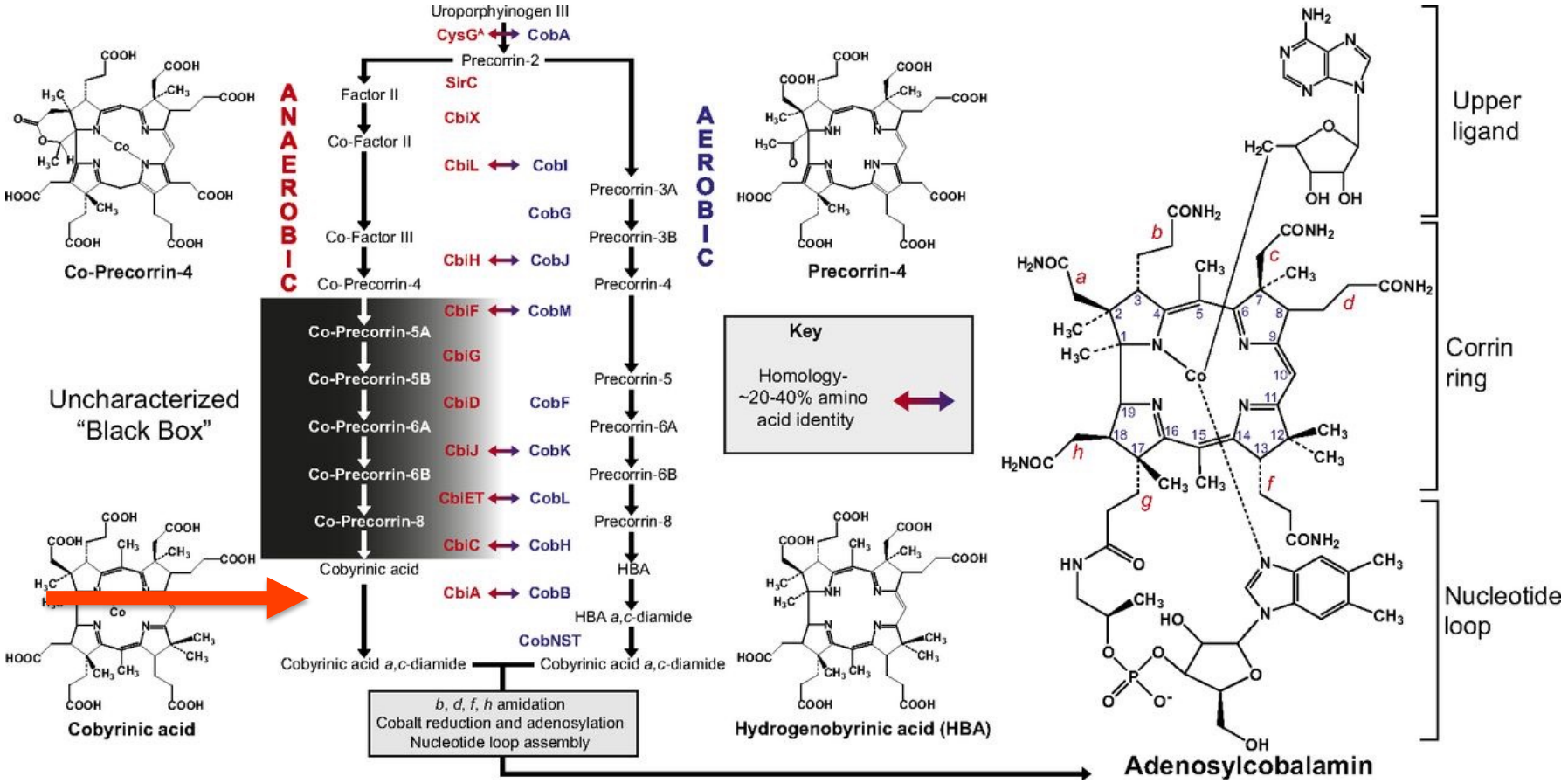
Moore et al. PNAS 110 (2013)

Protein function(s)



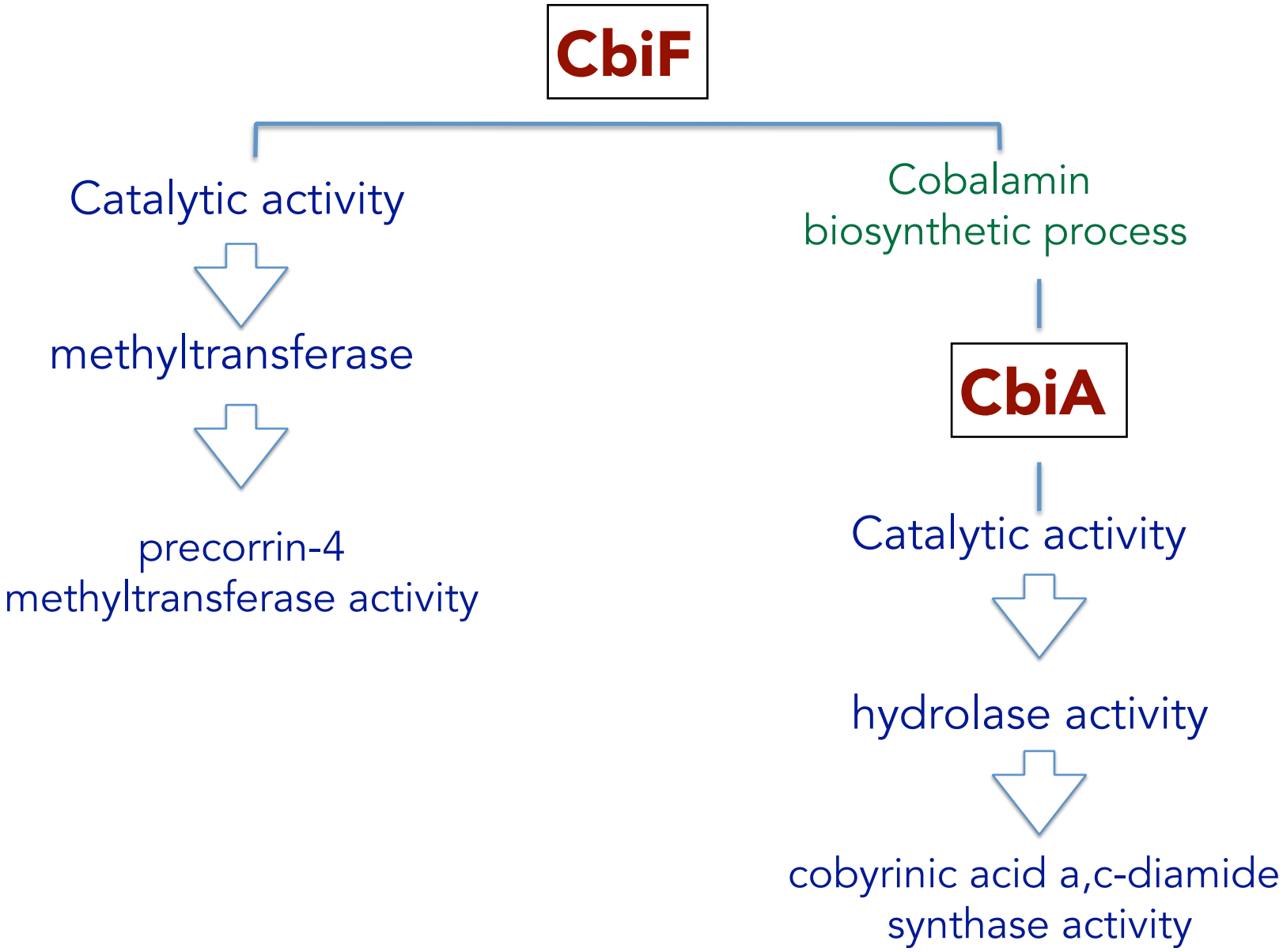
Cobalamin biosynthetic pathways

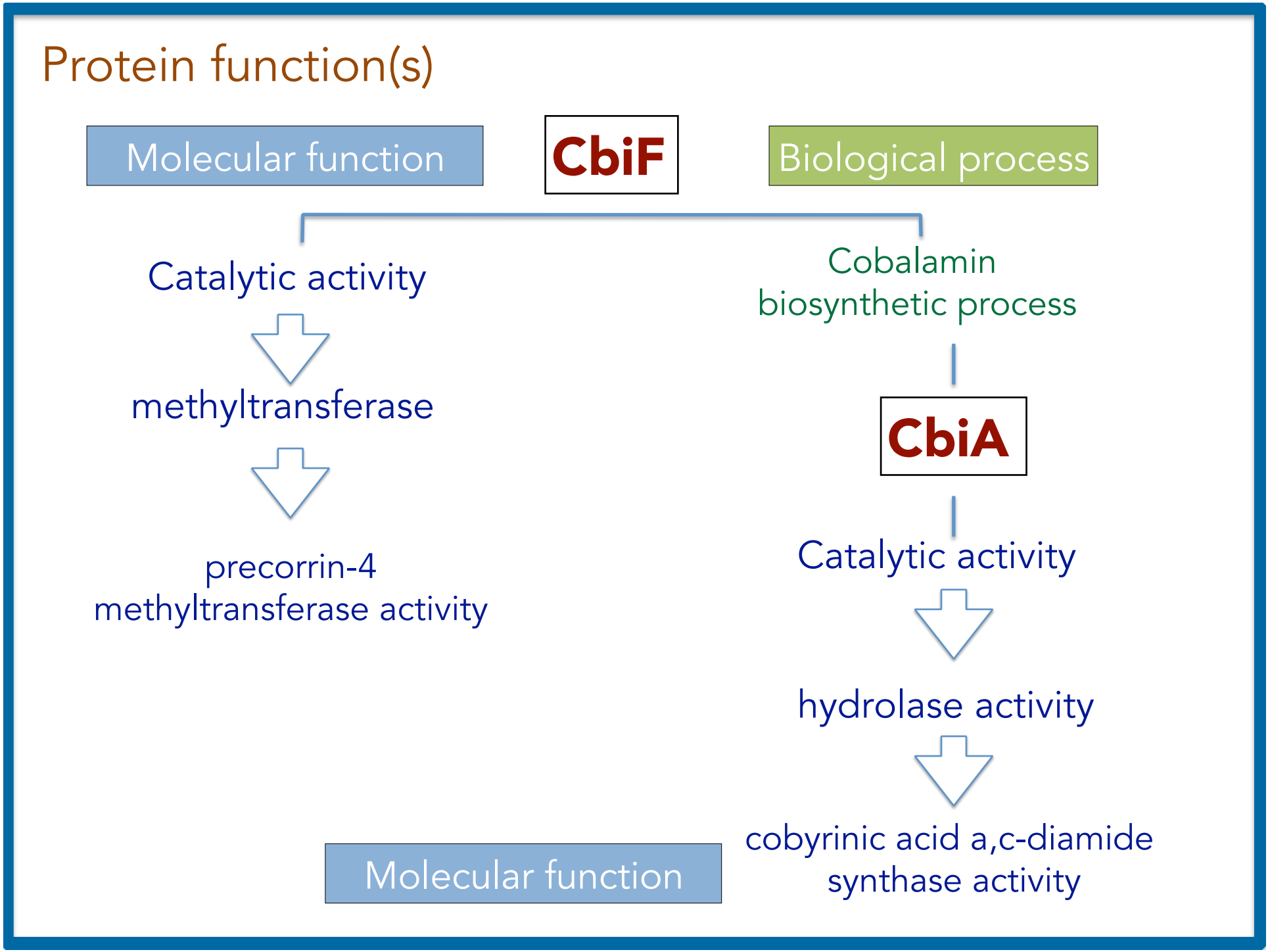
Marco Punta



Moore et al. PNAS 110 (2013)

Protein function(s)



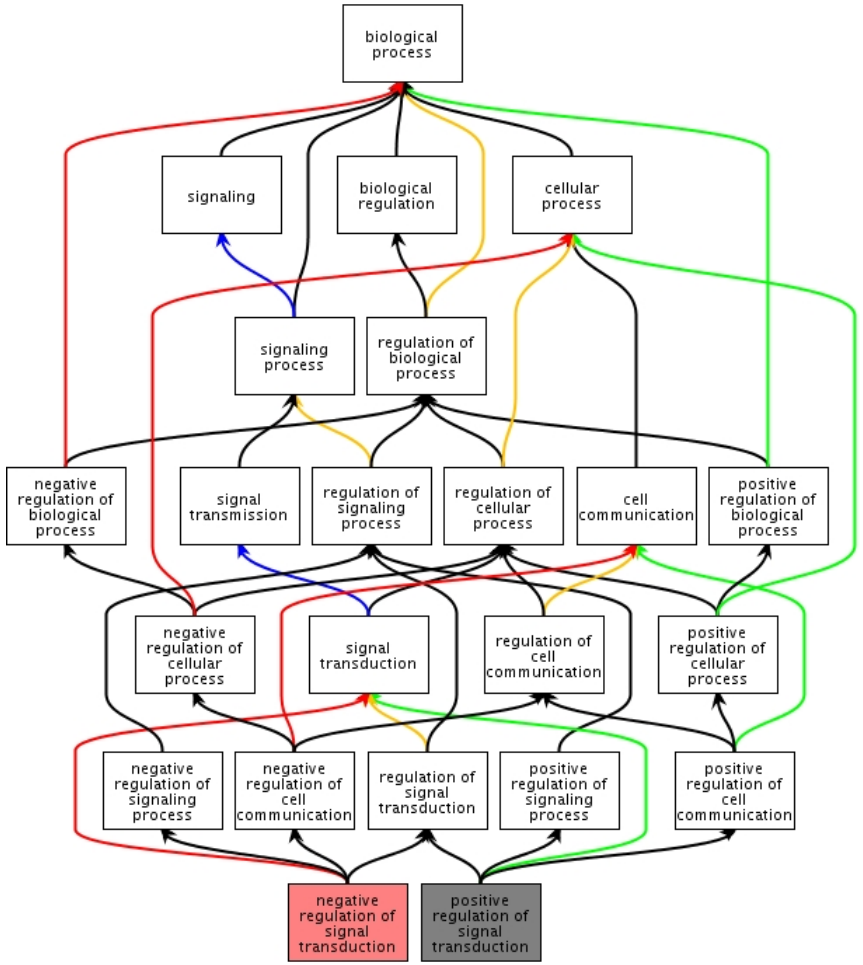

Protein function(s)
Molecular function
CbiF
Biological process
Catalytic activity
methyltransferase
precorrin-4
methyltransferase activity
Cobalamin
biosynthetic process
CbiA
Catalytic activity
hydrolase activity
cobyrinic acid a,c-diamide
synthase activity
Molecular function

The Gene Ontology (GO)

Marco Punta

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

- A set of concepts and their relationships to each other



www.ebi.ac.uk/QuickGO

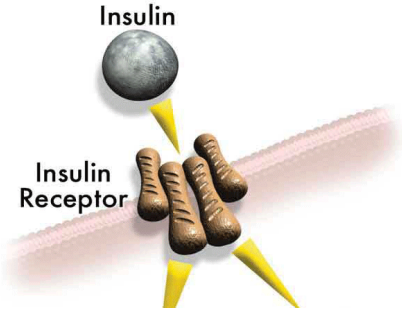
Slide courtesy of Alex Mitchell (EMBL-EBI)

GO: 3 ontologies in 1

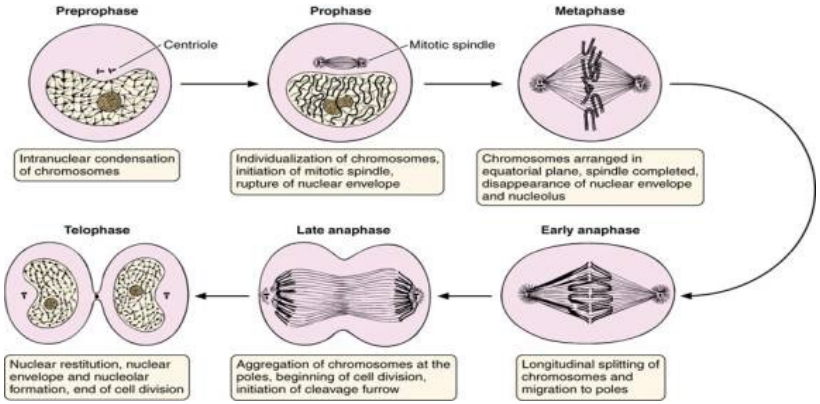
Marco Punta

1. Molecular Function

An elemental activity or task or job



- protein kinase activity
- insulin receptor activity



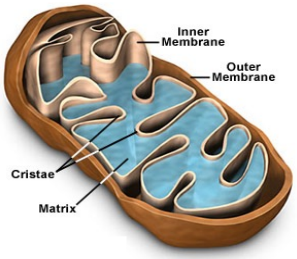
2. Biological Process

A commonly recognised series of events

- cell division

3. Cellular Component

Where a gene product is located



- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

Slide courtesy of Alex Mitchell (EMBL-EBI)

Bacillus megaterium CbiF GO annotation

Marco Punta

Database	Gene Product ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Pro Fo ID
Process													
UniProtKB	O87696	cbiF		GO:0006779	porphyrin-containing compound biosynthetic process	P	IEA	InterPro2GO	InterPro:IPR003043	1404	20151010	InterPro	
UniProtKB	O87696	cbiF		GO:0008152	metabolic process	P	IEA	InterPro2GO	InterPro:IPR000878|InterPro:IPR014776|InterPro:IPR014777	1404	20151010	InterPro	
UniProtKB	O87696	cbiF		GO:0009236	cobalamin biosynthetic process	P	IEA	InterPro2GO	InterPro:IPR006362	1404	20151010	InterPro	
UniProtKB	O87696	cbiF		GO:0009236	cobalamin biosynthetic process	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0169	1404	20151010	UniProt	
UniProtKB	O87696	cbiF		GO:0009236	cobalamin biosynthetic process	P	IEA	UniPathway2GO	UniPathway:UPA00148	1404	20151010	UniProt	
UniProtKB	O87696	cbiF		GO:0032259	methylation	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0489	1404	20151010	UniProt	
UniProtKB	O87696	cbiF		GO:0055114	oxidation-reduction process	P	IEA	InterPro2GO	InterPro:IPR003043	1404	20151010	InterPro	
Function													
UniProtKB	O87696	cbiF		GO:0008168	methyltransferase activity	F	IEA	InterPro2GO	InterPro:IPR000878|InterPro:IPR003043|InterPro:IPR014776|InterPro:IPR014777	1404	20151010	InterPro	
UniProtKB	O87696	cbiF		GO:0008168	methyltransferase activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0489	1404	20151010	UniProt	
UniProtKB	O87696	cbiF		GO:0016740	transferase activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0808	1404	20151010	UniProt	
UniProtKB	O87696	cbiF		GO:0043115	precorrin-2 dehydrogenase activity	F	IEA	InterPro2GO	InterPro:IPR003043	1404	20151010	InterPro	
UniProtKB	O87696	cbiF		GO:0046026	precorrin-4 C11-methyltransferase activity	F	IEA	InterPro2GO	InterPro:IPR006362	1404	20151010	InterPro	

Top-level EC numbers[5]

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

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

EC 1	Oxidoreductases
EC 1.3	Acting on the CH-CH Group of Donors
EC 1.3.1	With NAD^+ or $NADP^+$ as acceptor
EC 1.3.1.21	7-dehydrocholesterol reductase

IUBMB Enzyme Nomenclature

EC 1.3.1.21

Image Credits: http://legacy.earlham.edu/~merkeka/EMBL_EBI-top.gif

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

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

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

SAME

SAME

Database	Gene Product ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
Process													
UniProtKB	P02008	HBZ		GO:0000122	negative regulation of transcription from RNA polymerase II promoter	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000020531	9606	20140913	Ensembl	
UniProtKB	P02008	HBZ		GO:0006810	transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0813	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		GO:0015671	oxygen transport	P	IEA	InterPro2GO	InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0015671	oxygen transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		GO:0043249	erythrocyte maturation	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000020531	9606	20140913	Ensembl	
Function													
UniProtKB	P02008	HBZ		GO:0005344	oxygen transporter activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		GO:0005344	oxygen transporter activity	F	TAS	PMID:7555018		9606	20030904	PINC	
UniProtKB	P02008	HBZ		GO:0005506	iron ion binding	F	IEA	InterPro2GO	InterPro:IPR000971|InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0005515	protein binding	F	IPI	PMID:11159	UniProtKB:P68871	9606	20140914	IntAct	
UniProtKB	P02008	HBZ		GO:0005515	protein binding	F	IPI	PMID:1	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	

SAME

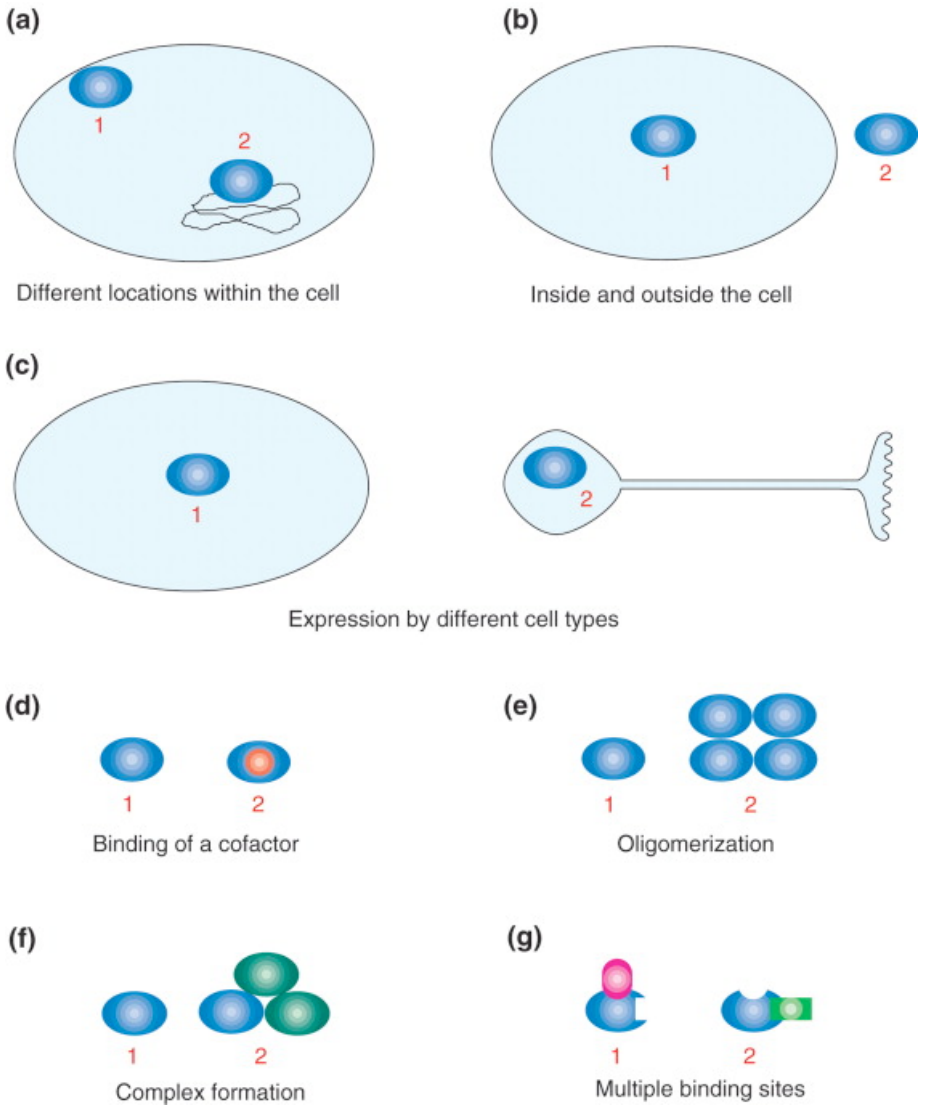
SAME

Functional property to be conserved	Sequence identity	Conservation rate	Reference
Non-enzyme	50%	98%*	[88]
All 4 EC numbers	70%**	90%	[89]
All 4 EC numbers	40%**	70%	[89]
First 3 EC numbers	50%**	90%	[89]
First 3 EC numbers	30%**	70%	[89]
All 4 EC numbers	50%	30%	[90]
First 3 EC numbers	25%	70%	[91]
SWISS-PROT keywords	40%	70%	[92]
Subcellular localization (11 classes)	70%	90%	[93]

*98% of non enzymes that have at least one enzyme homolog.
**Global identity, defined in [89].
Note: different estimates for the same functional aspects reflect the different methods, procedures, and datasets used to assess sequence similarity by the various groups.
doi:10.1371/journal.pcbi.1000160.t001

Moonlighting proteins

Marco Punta



Jeffery Trends Bioch. Sc. (1999)

Do homologous protein regions perform a similar function?

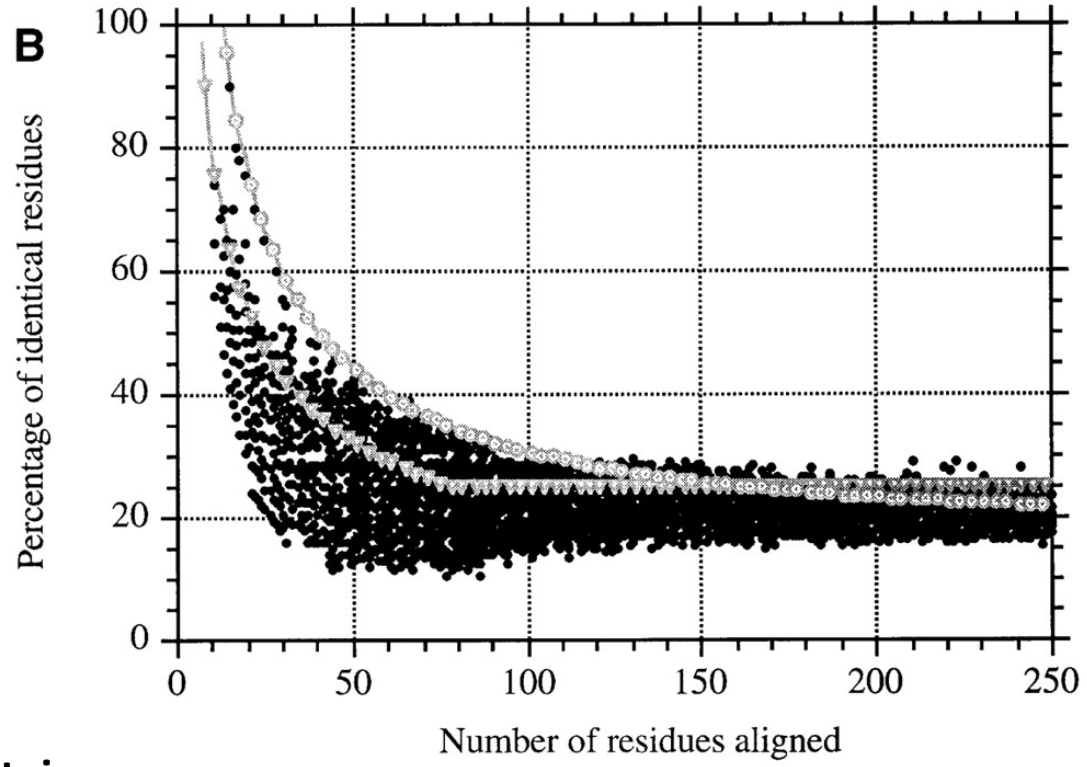
Marco Punta

Homologous proteins may share a number of functional features, however:

- functional drift can lead to different functions or aspects of function

- while functional similarity generally correlates with evolutionary distance, no distance is safe for inferring function (very closely related proteins can have slightly to radically different functions)

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

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



Unrelated proteins

Burkhard Rost Protein Eng. 1999;12:85-94

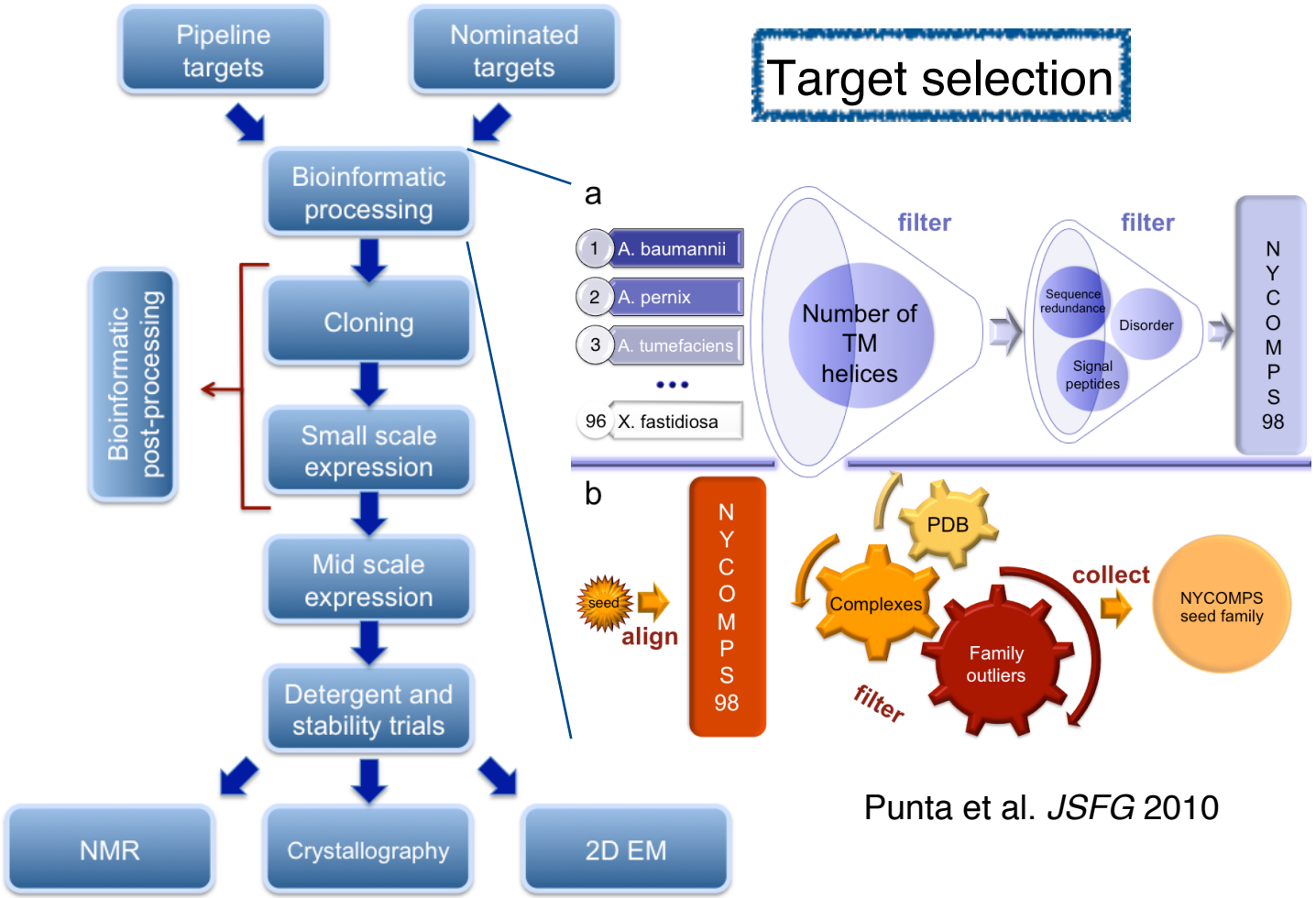
Marco Punta

Exercise

Homology-based function annotation transfer #1

NYCOMPS pipeline

Marco Punta



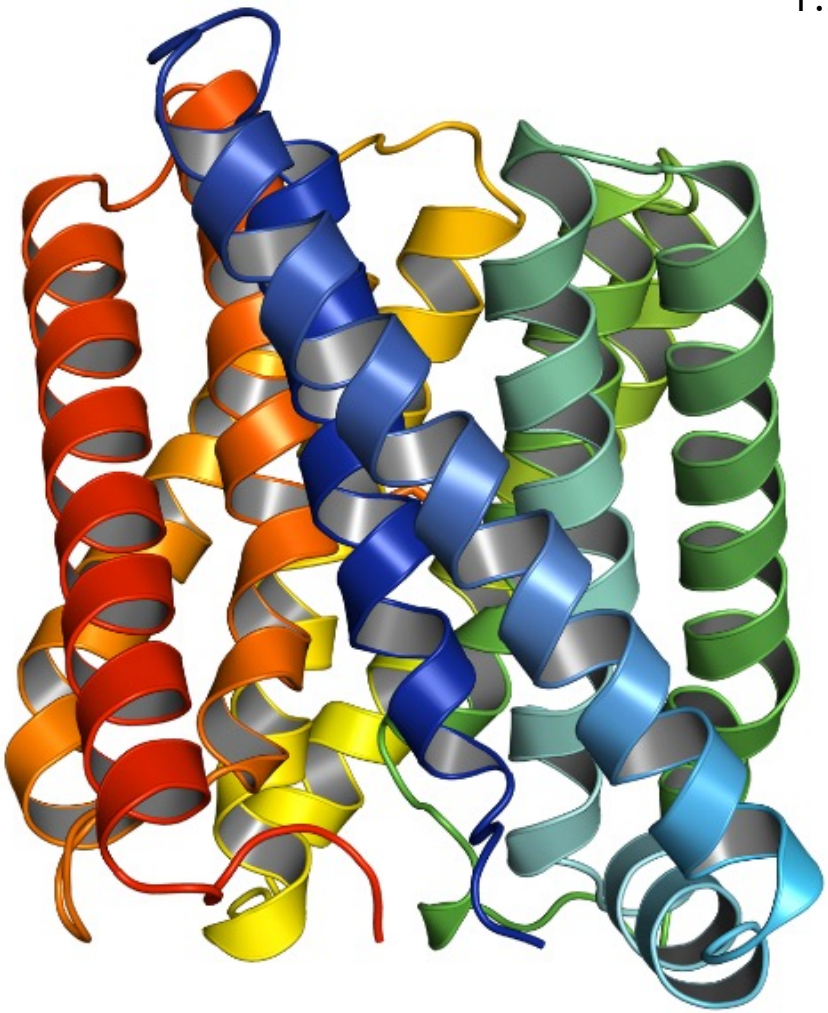
Love et al. *JSFG* 2010

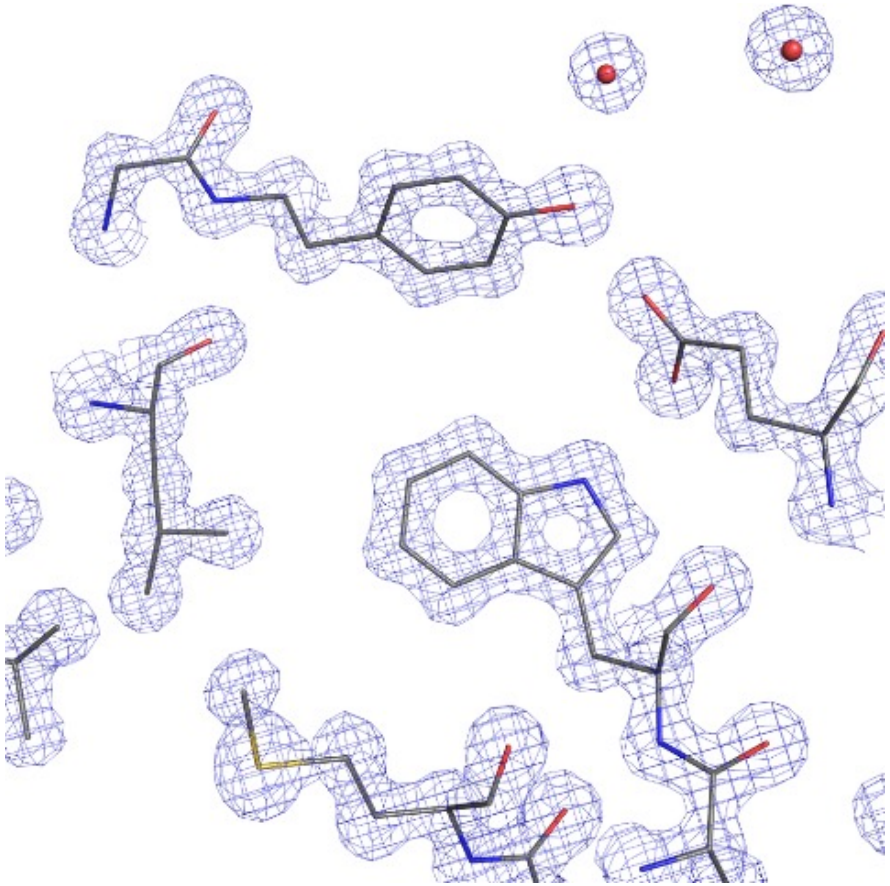
Punta et al. *JSFG* 2010

H. influenzae protein [3M71] ← PDB id

1.20 Å

Marco Punta





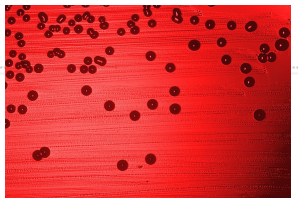
Chen et al. *Nature* 467 (2010)

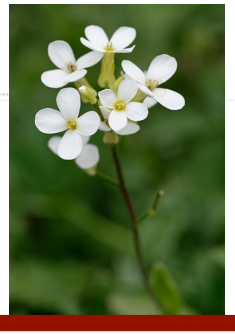
Alignment

Q9LD83 SLAC1_ARATH - **Guard cell S-type anion channel SLA...** - Arabidopsis thal...

E-value: 3e-10 **Positives :** 41.0%
Score: 160 **Query Length:** 328
Ident.: 22.0% **Match Length:** 556

```
P44741      20   PFPL--PTGYFGIPLGLAALSLAWFHLE-----NLFPAARMVSDVLGIVASAVWILFILM   72
                 PF L  P G FGI LGL++ ++ W  L      N      +++ V+ + +  V +
Q9LD83     183   PFLLRFPIGCFGICLGLSSQAVLWLALAKSPATNFLHITPLINLVVWLFSLVVLVSVSFT  242

P44741      73   YAYKLRYYFEEVRAEYHSPVRFSFIALIPITTMLVG---DILYRWNPLIAEVLIWIGTIG  129
                 Y  K  +YFE V+ EY  PVR +F     +  M +      ++  N       IW   +G
Q9LD83     243   YILKCIFYFEAVKREYFHPVRVNFFFAPWVVCMFLAISVPPMFSPNRKYLHPAIWCVFMG  302

P44741     130   QLLFSTLRVSELWQGGVFEQ--KSTHPSFYLPAVAANFTSASSLALLGYHDLGYLFFGAG  187
                    F  L++   W  G   +  K  +PS +L +V  NF  A   + +G+ ++    +  G
Q9LD83     303   PYFFLELKIYGQWLSGGKRRLCKVANPSSHL-SVVGNFVGAILASKVGWDEVAKFLWAVG  361

P44741     188   MIAWIIFEPVLLQHLRISSLEPQFRATMGIVLAPAFVCVSAYLSINHGEVDTLAKILWGY  247
                    +++    L Q L  S   P+    +  +   A    S   +  +G+ D  ++  +
Q9LD83     362   FAHYLVVFVTLYQRLPTSEALPKELHPVYSMFIAAPSAASIAWNTIYGQFDGCSRTCFFI  421

P44741     248   GFLQLFFLLRLFPWIVEKGLNIGLWAFSFGLASMANSATAFY----HGNVLQGVSIFAFV  303
                       L+    +      ++  W++++F  + + A+ AT  Y     G   + +++
Q9LD83     422   ALFLYISLVARINFFTGFKFSVAWWSYTFFMTT-ASVATIKYAEAVPGYPSRALALTLSF  480

P44741     304   FSNVMIGLLVLMTI   317
                  S  M+ +L + T+
Q9LD83     481   ISTAMVCVLFVSTL   494
```



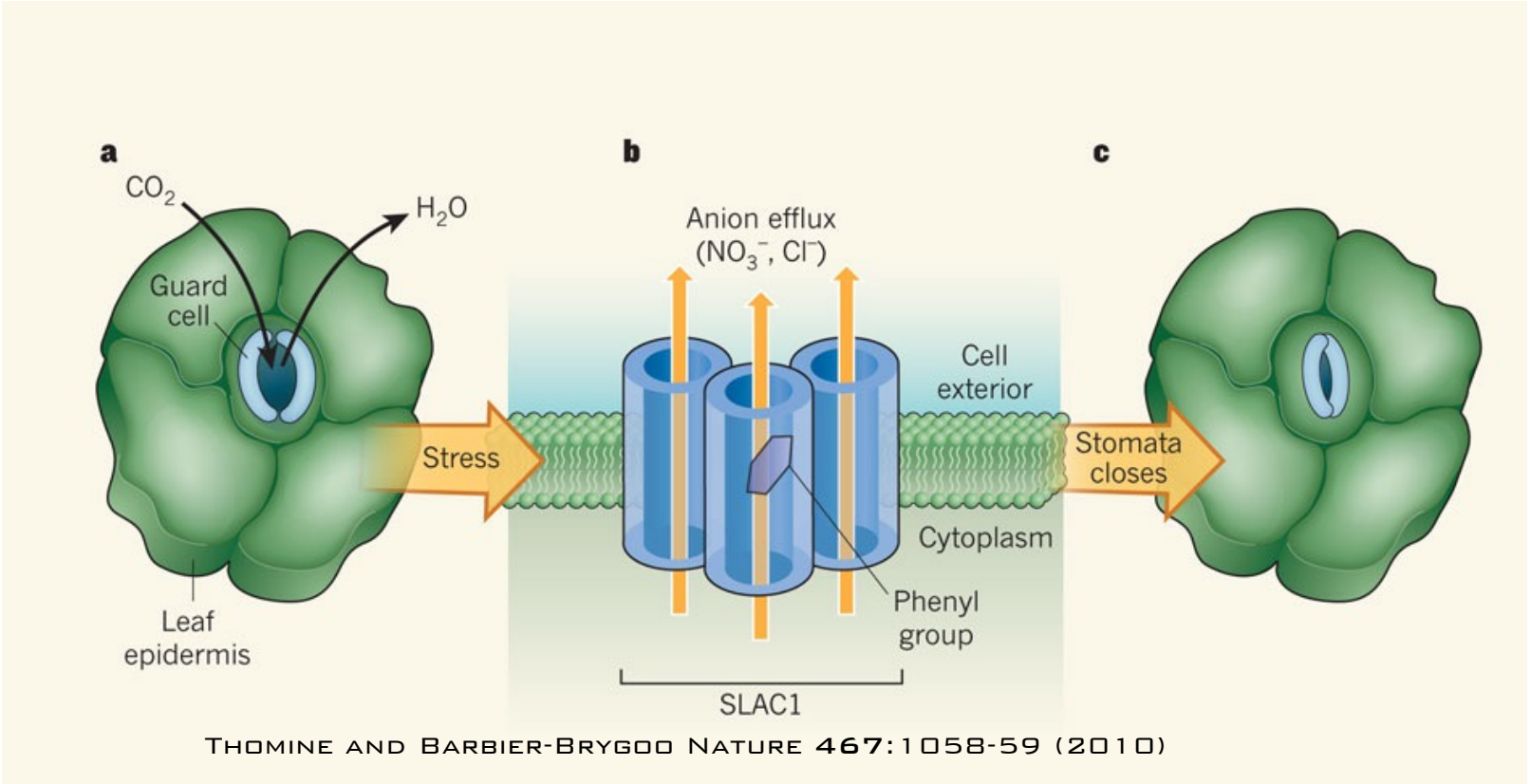


E-value is the number of matches with a given score (or higher) that we expect to occur by chance.

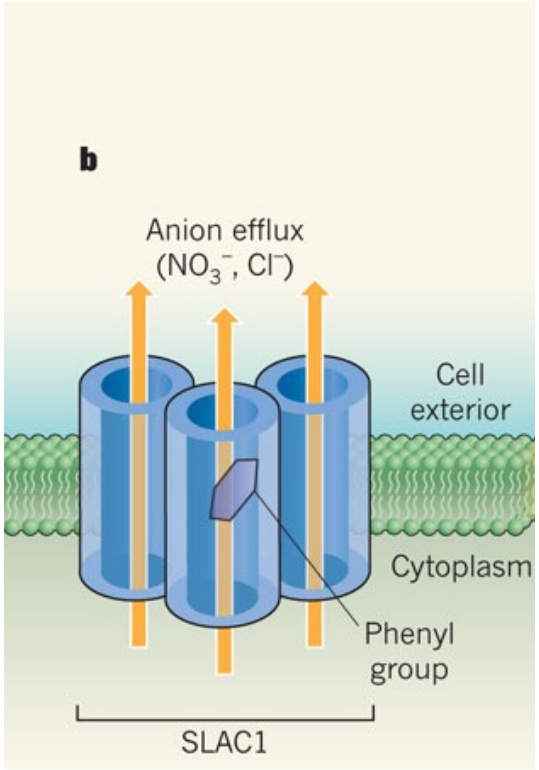
This depends on database size!

For an alignment with score S and E-value=1, we expect to have by chance 1 match with the same or higher score.

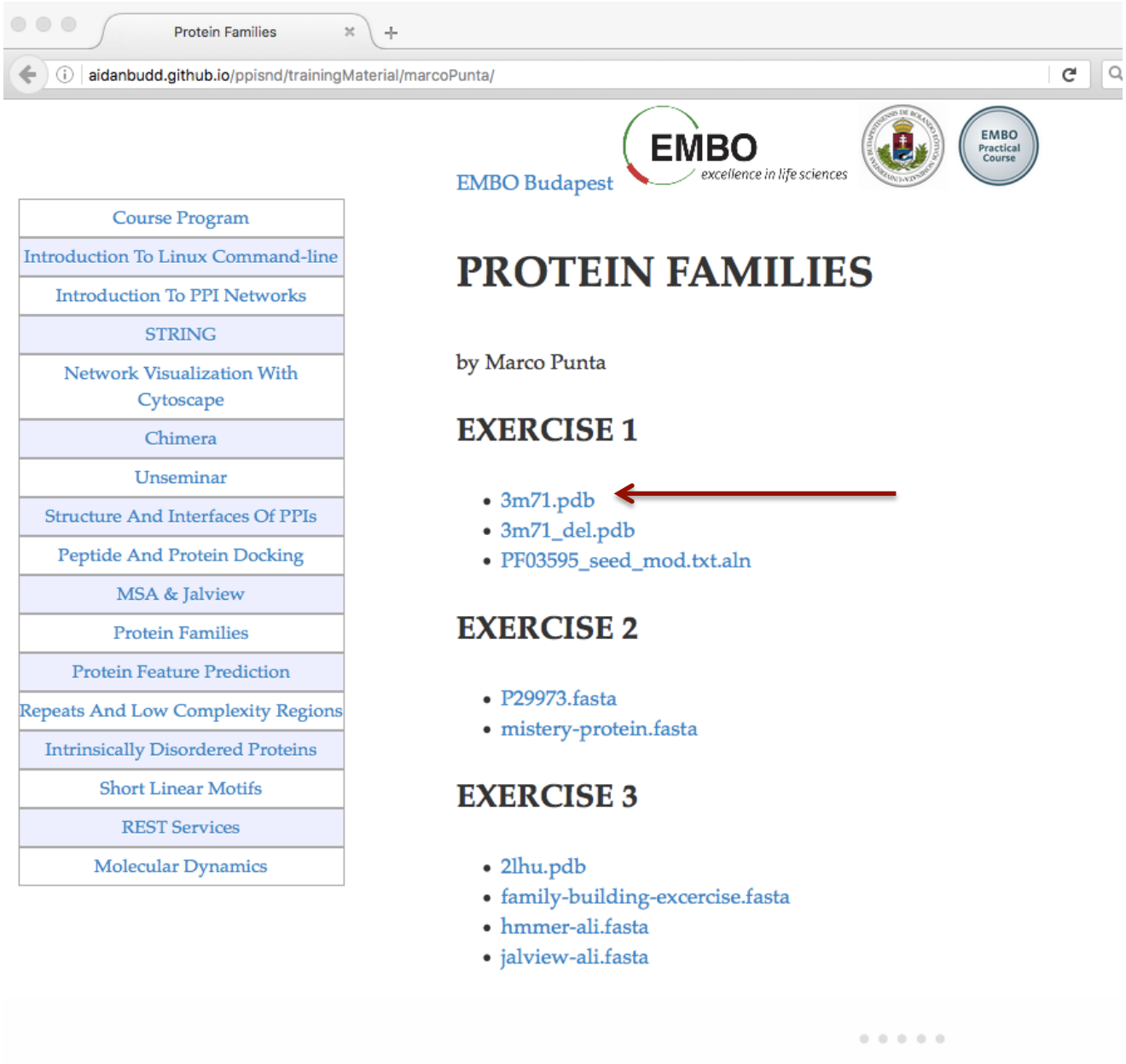
For an alignment with score S and E-value=1, we expect to have by chance 1 match with the same or higher score. If E-value is 0.001 then we expect by chance 0.001 matches with the same or higher score.



Thomine and Barbier-Brygoo Nature **467**:1058-59 (2010)


b
Anion efflux
(NO_3^-, Cl^-)
Cell
exterior
Cytoplasm
Phenyl
group
SLAC1

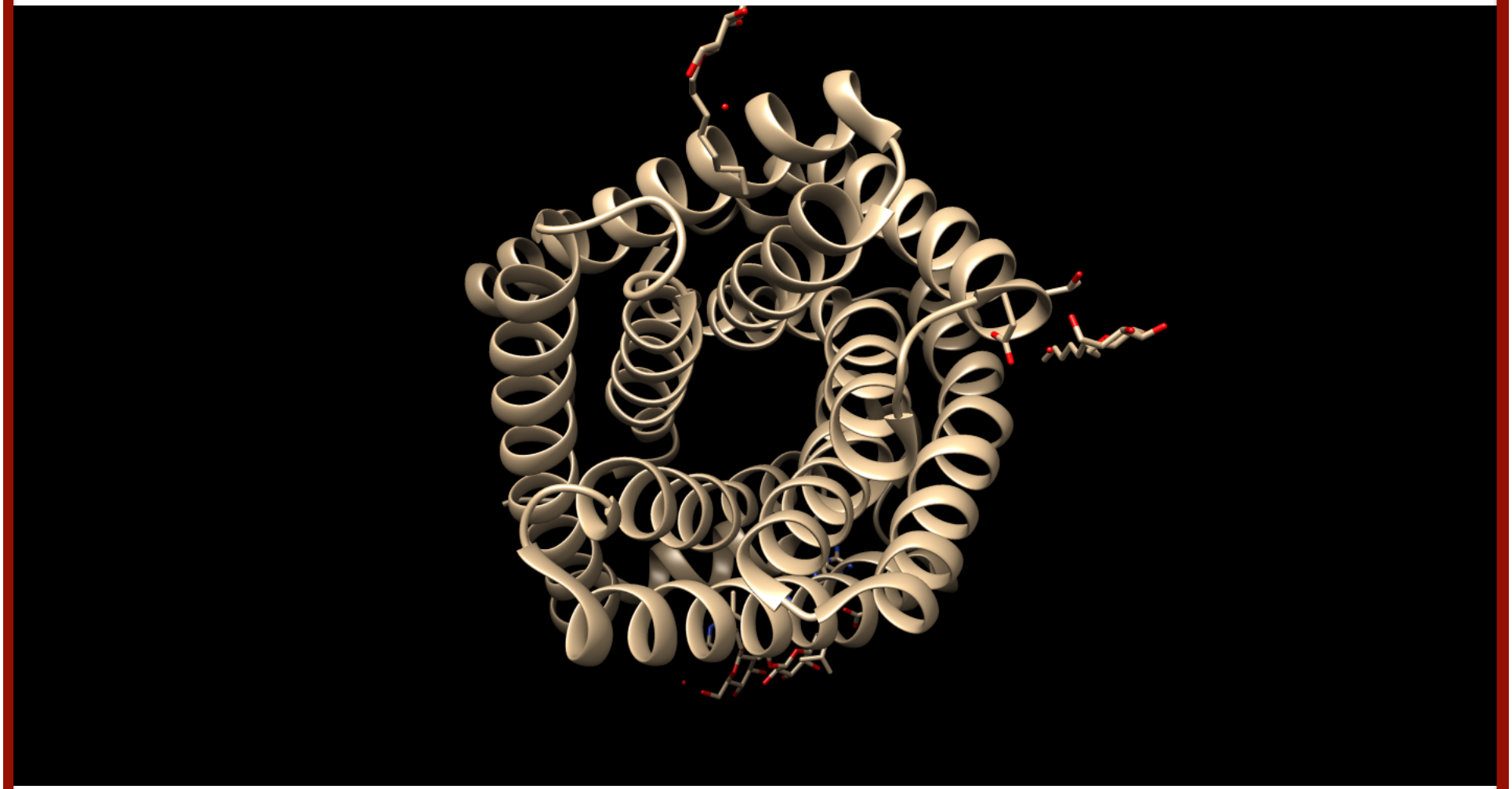
Marco Punta



http://aidanbudd.github.io/ppisnd/trainingMaterial/marcoPunta/

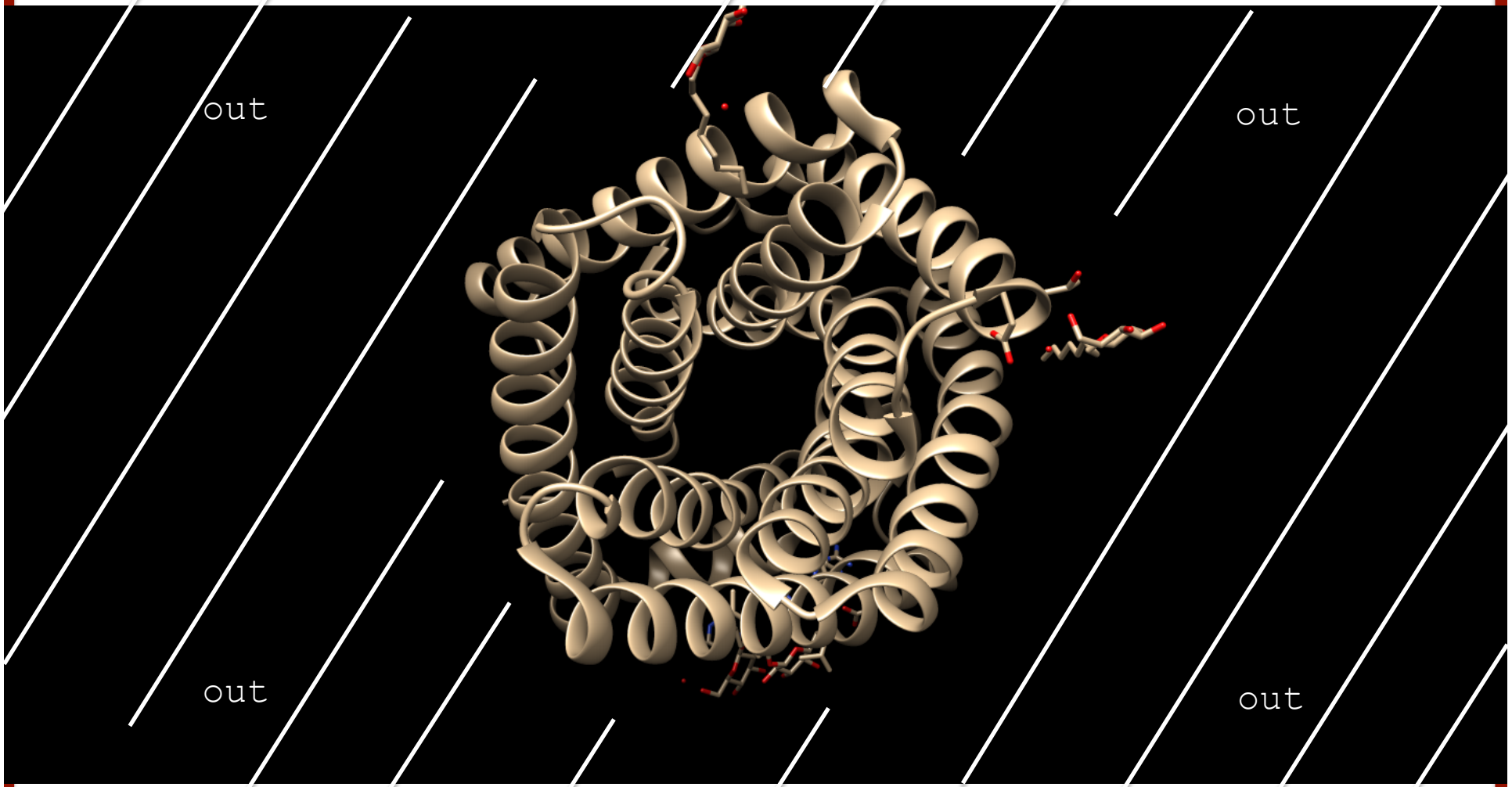
1. OPEN Chimera
2. File -> Open “3M71.pdb”

Marco Punta



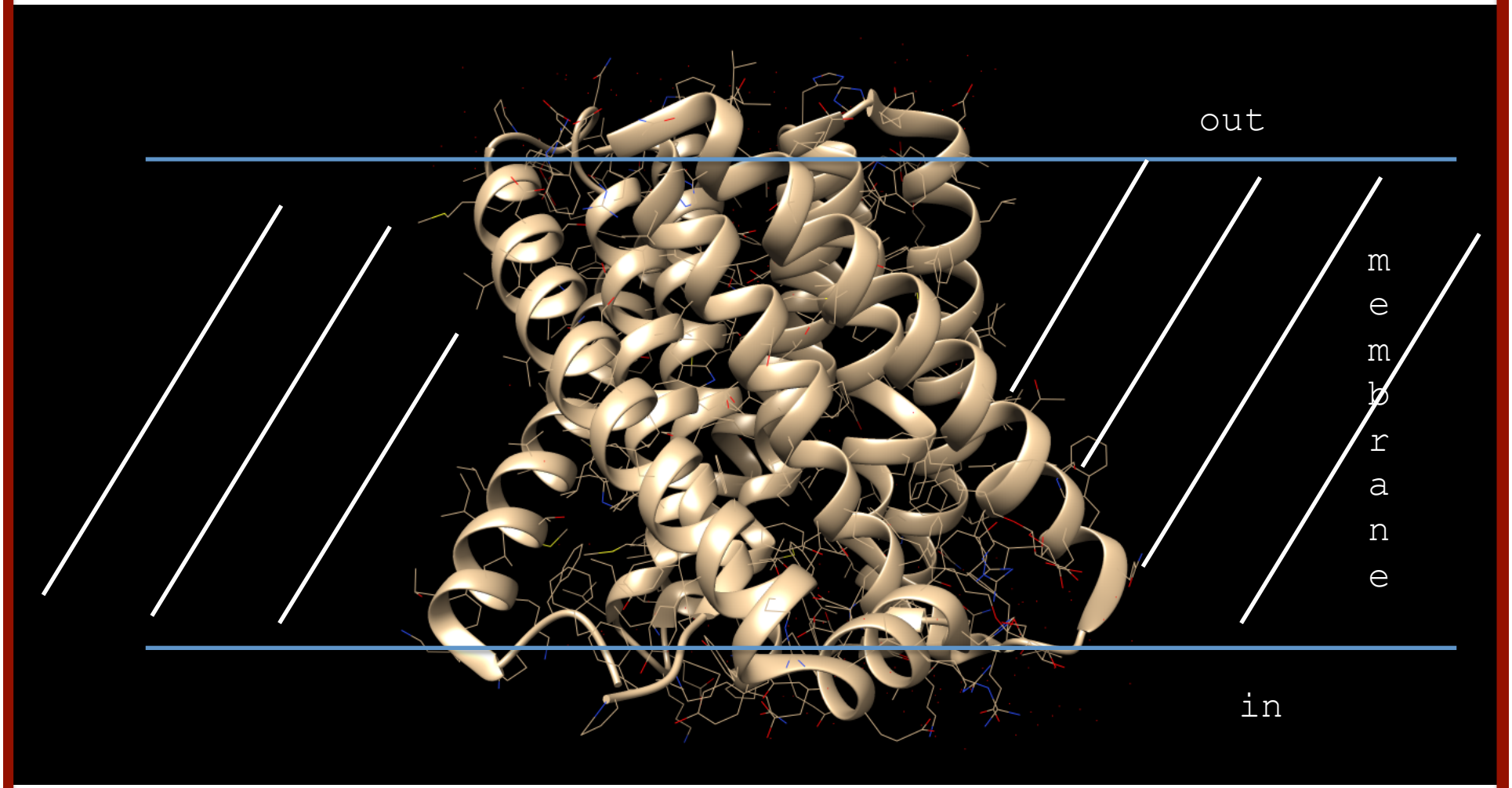
EMBO Workshop, Budapest, 2016

Marco Punta

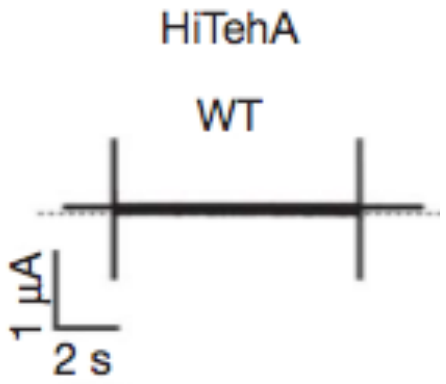


EMBO Workshop, Budapest, 2016

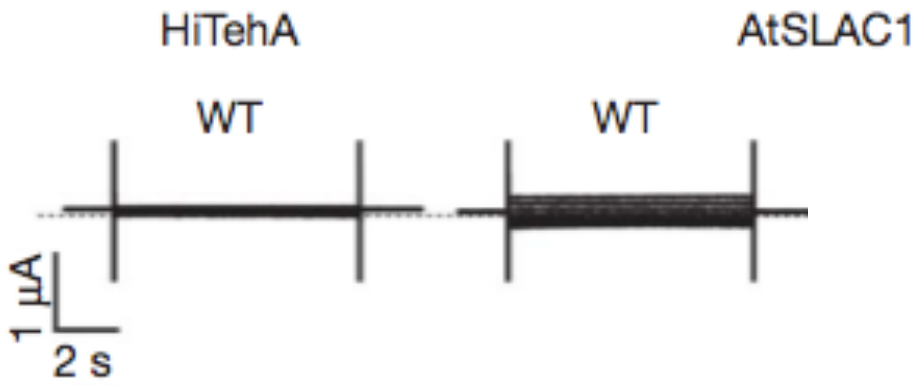
Marco Punta



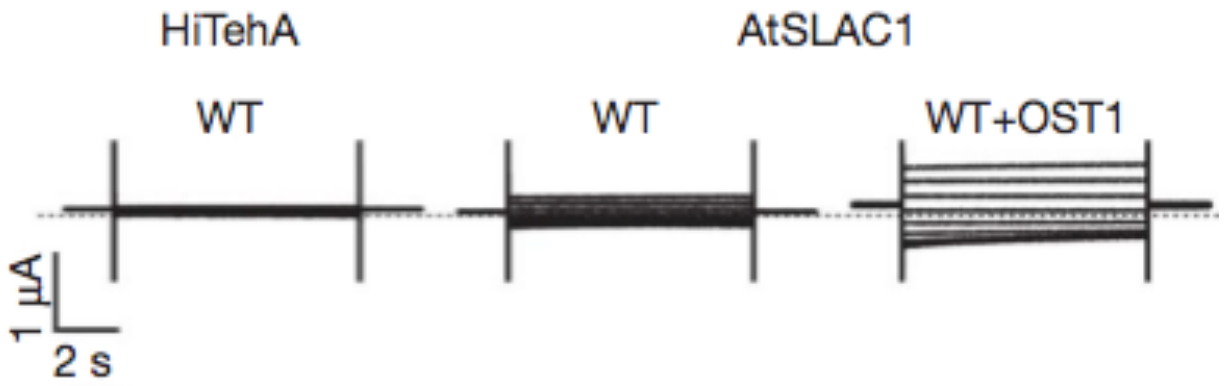
EMBO Workshop, Budapest, 2016



Chen et al. *Nature* 467 (2010)



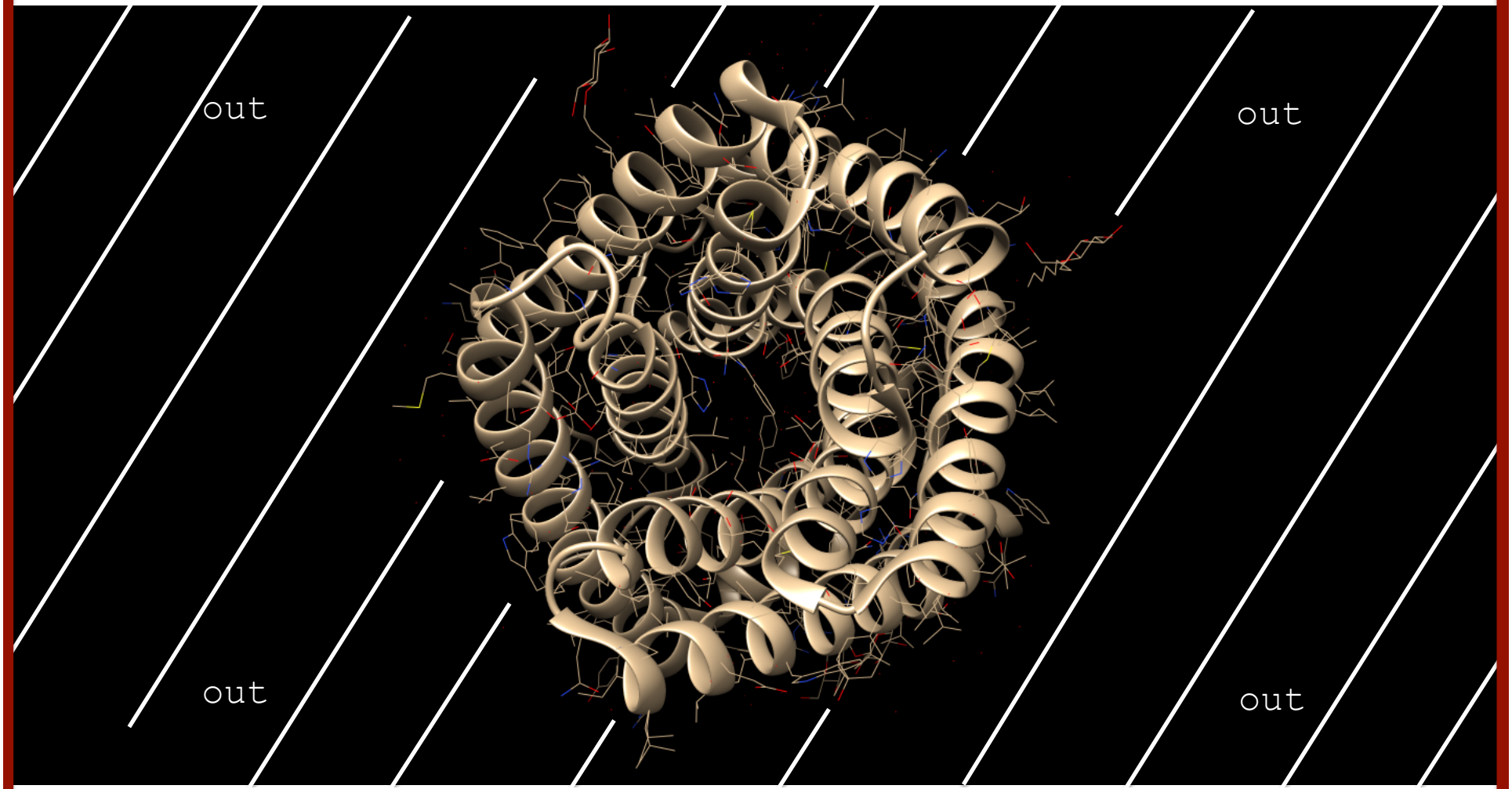
Chen et al. *Nature* 467 (2010)



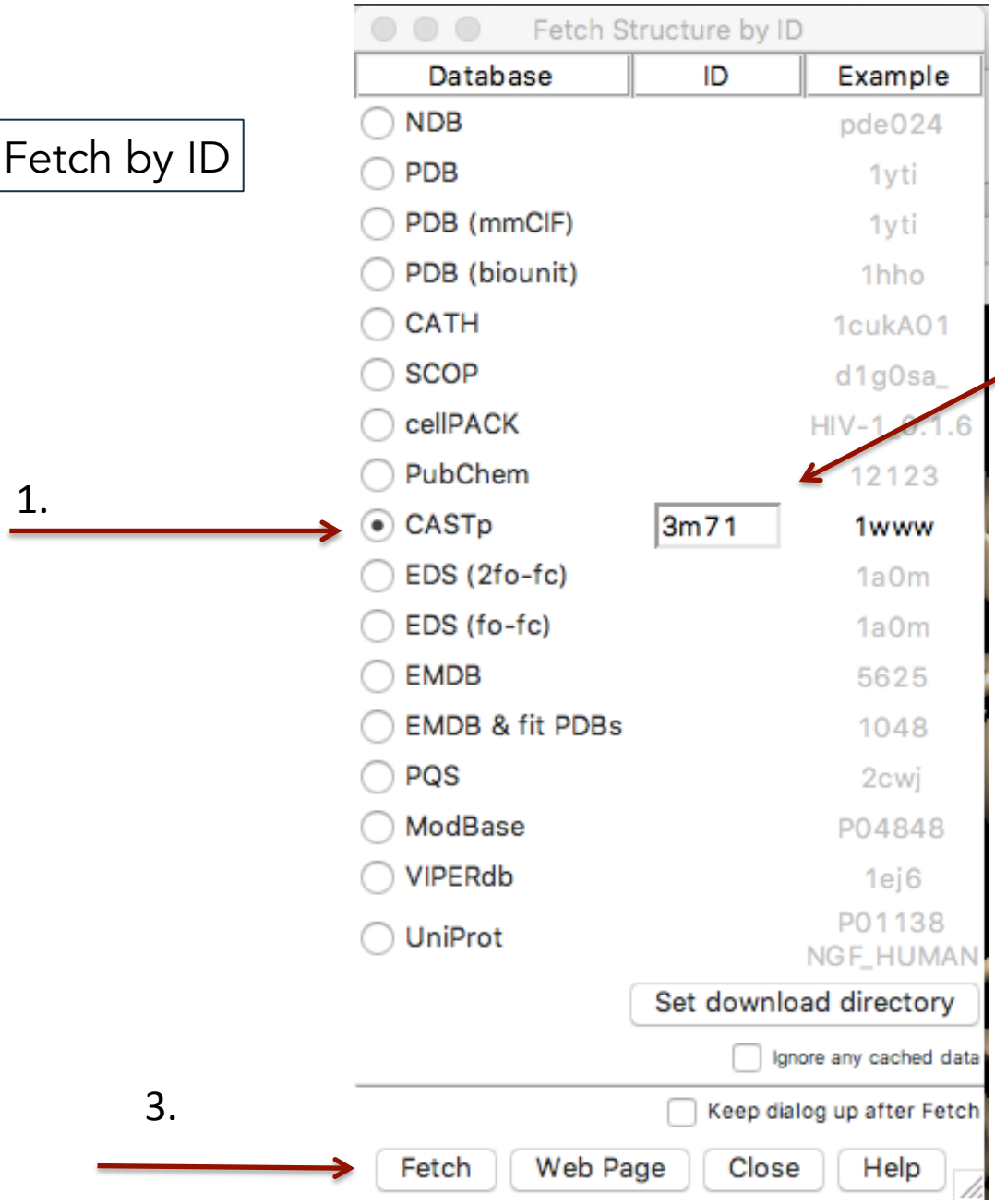
Chen et al. *Nature* 467 (2010)

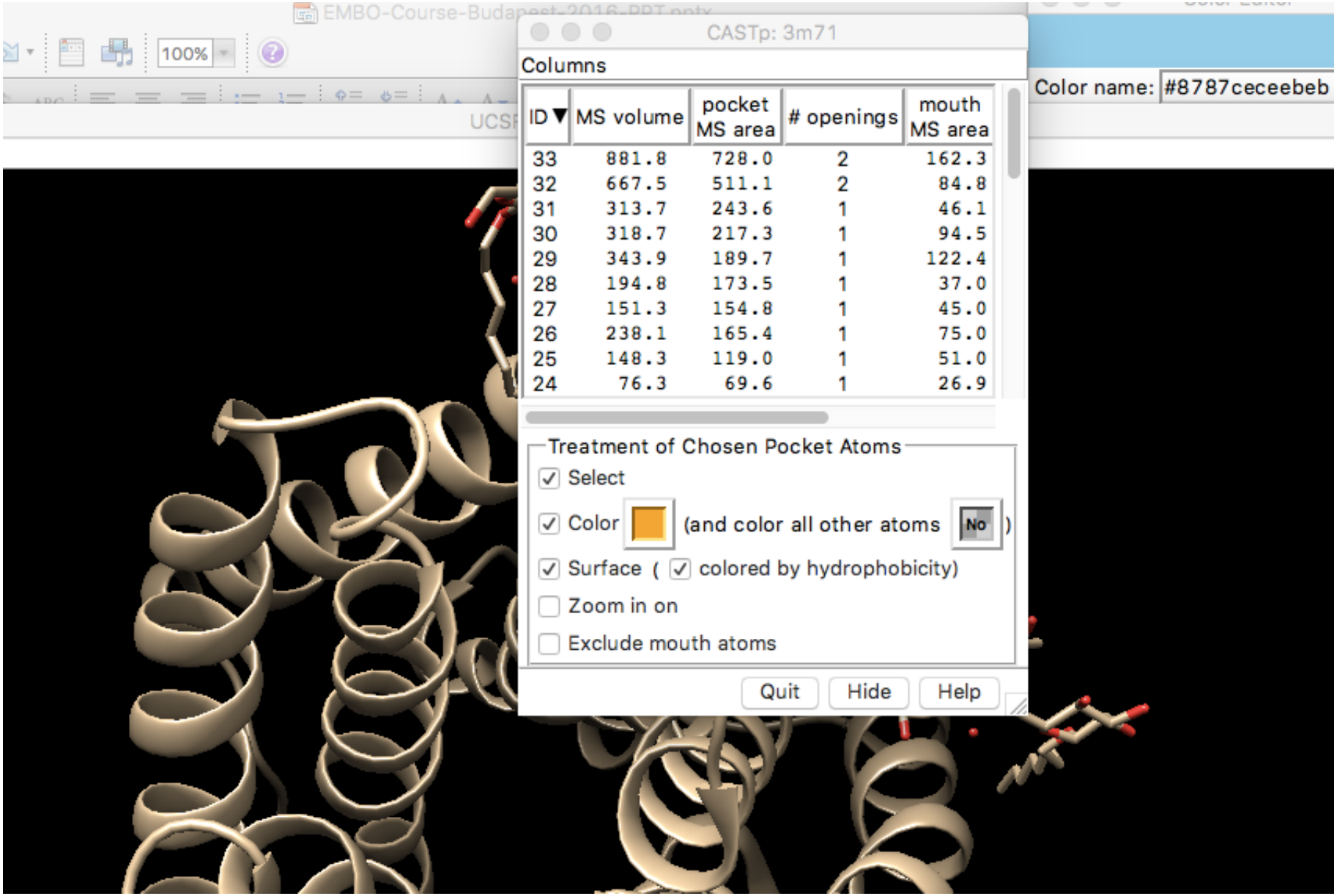
1. Actions -> Atoms/Bonds -> wire
2. Actions -> Atoms/Bonds -> show

1. Actions -> Atoms/Bonds -> wire
2. Actions -> Atoms/Bonds -> show

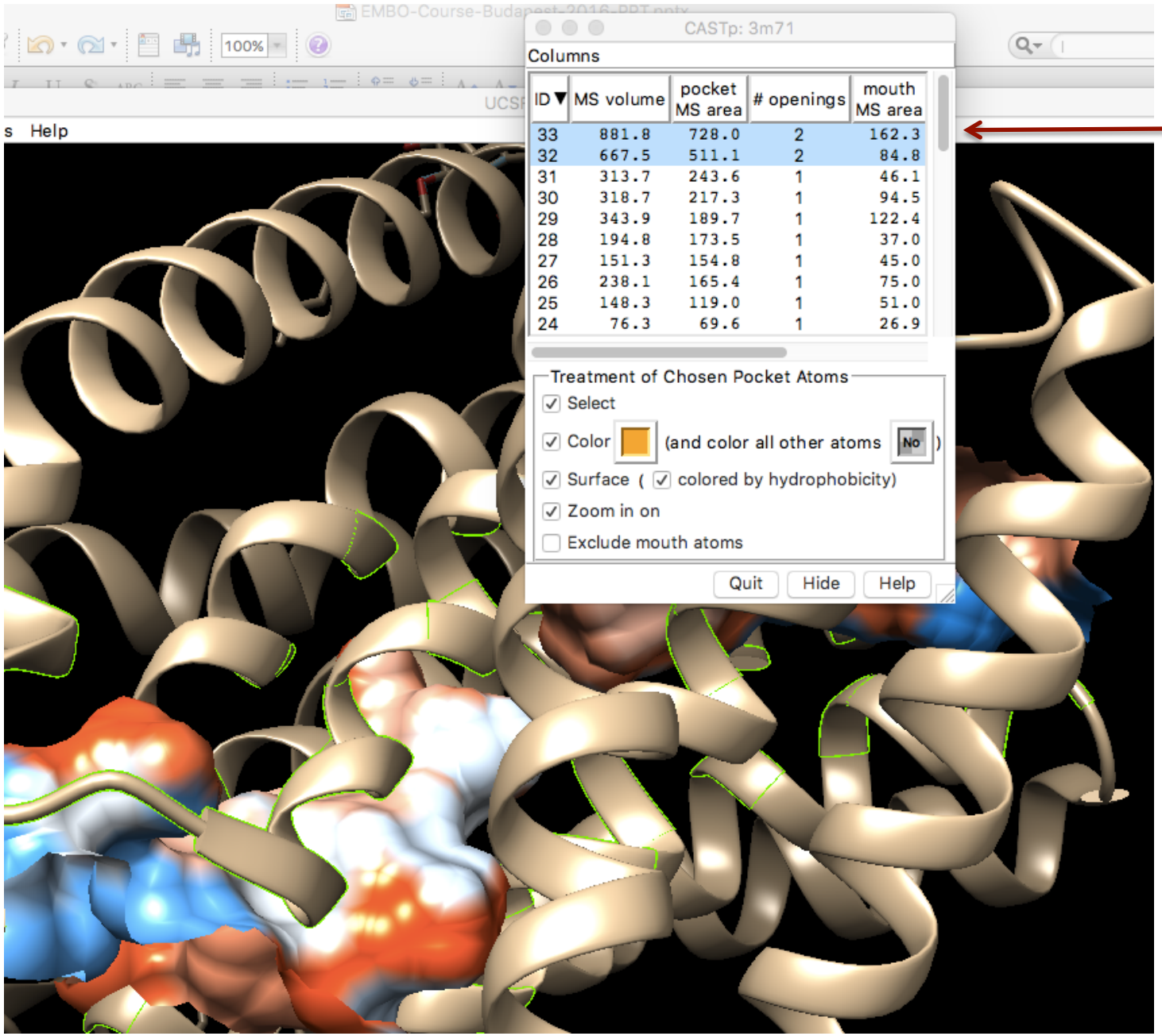


1. File -> Fetch by ID

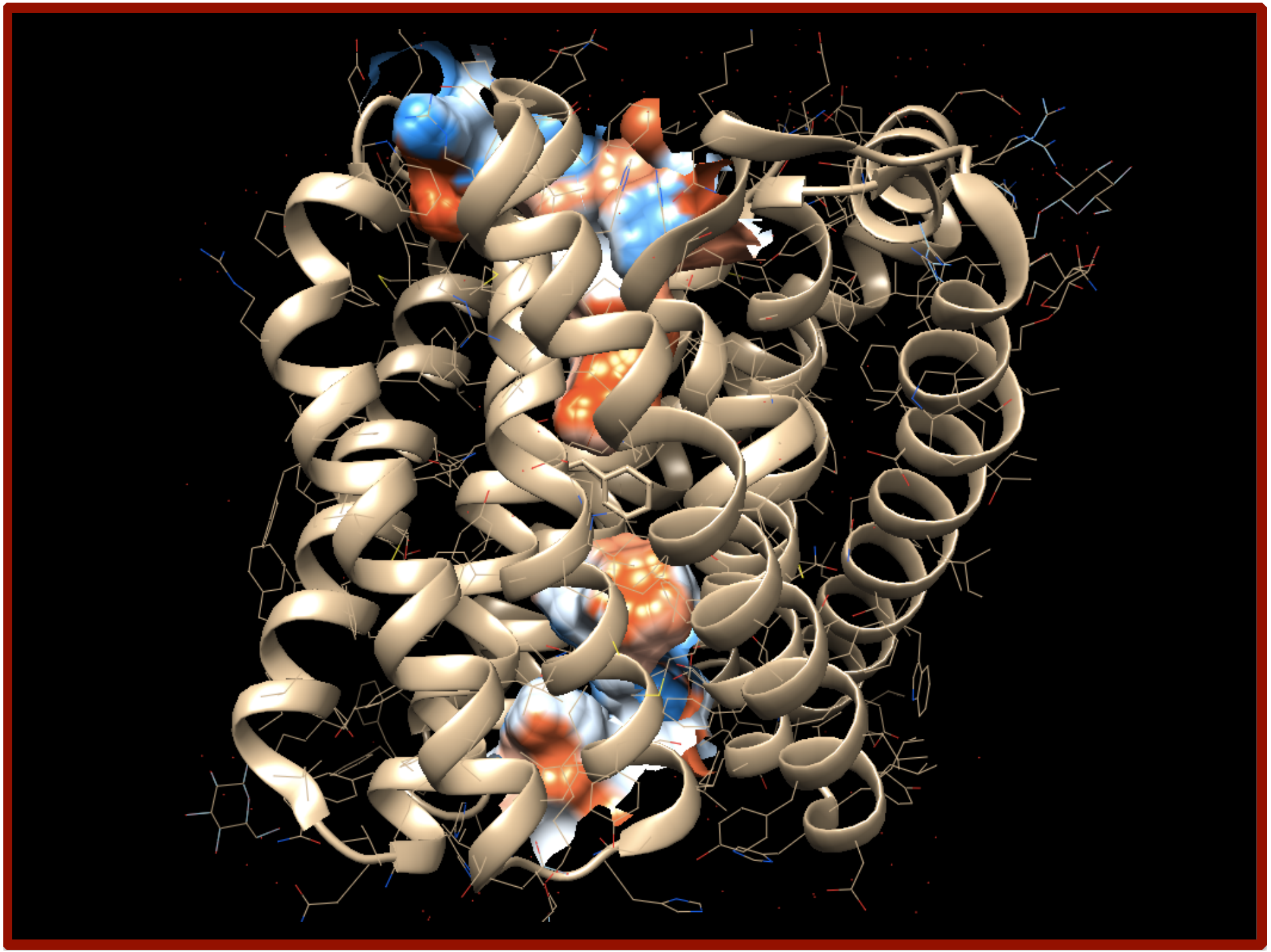


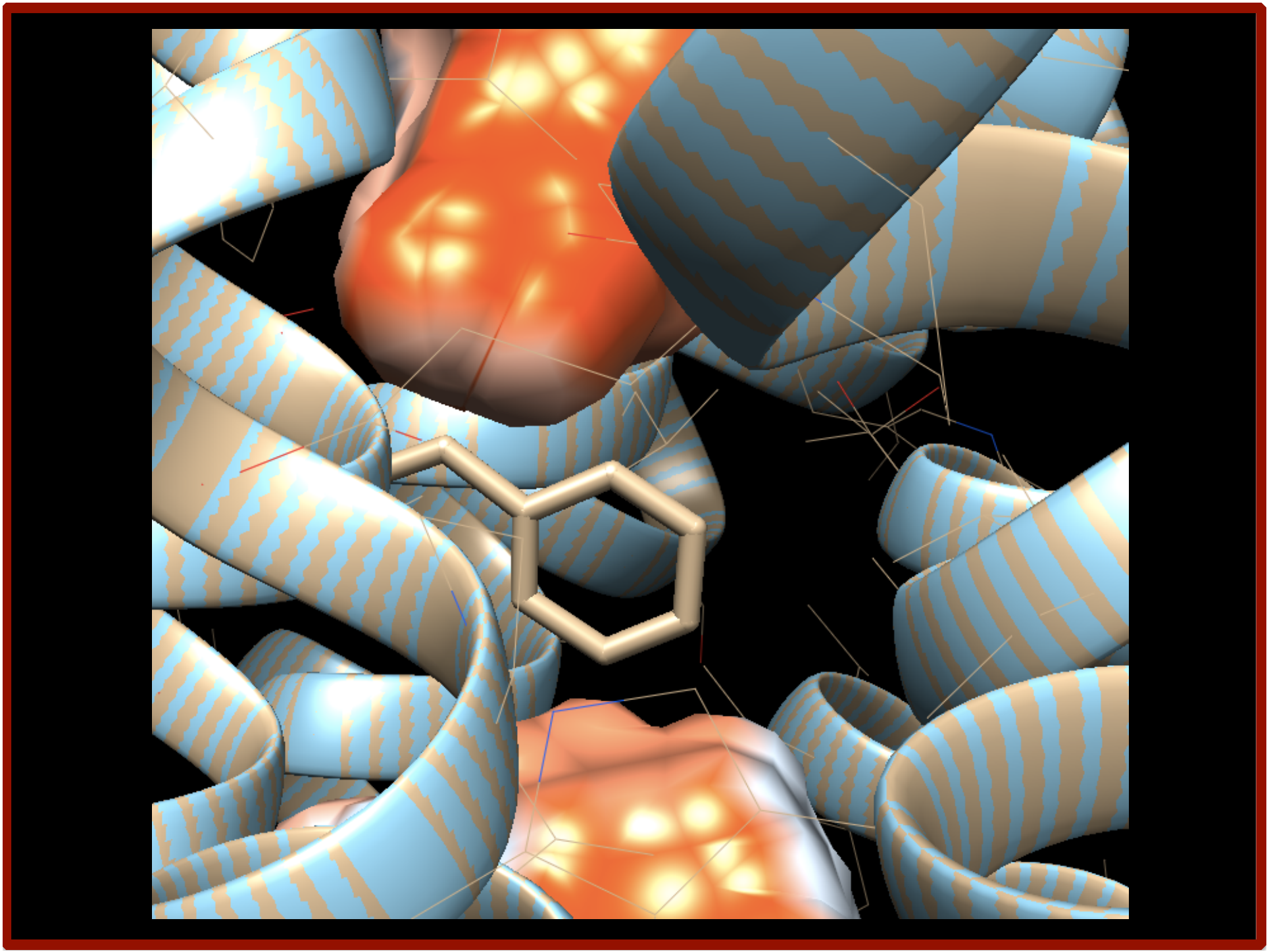

EMBO-Course-Budapest-2016-PPT.pptx
100%
CASTp: 3m71
Columns
ID ▼ MS volume pocket MS area # openings mouth MS area
33 881.8 728.0 2 162.3
32 667.5 511.1 2 84.8
31 313.7 243.6 1 46.1
30 318.7 217.3 1 94.5
29 343.9 189.7 1 122.4
28 194.8 173.5 1 37.0
27 151.3 154.8 1 45.0
26 238.1 165.4 1 75.0
25 148.3 119.0 1 51.0
24 76.3 69.6 1 26.9
Color name: #8787ceceebeb
Treatment of Chosen Pocket Atoms
Select
Color
(and color all other atoms
No
)
Surface (colored by hydrophobicity)
Zoom in on
Exclude mouth atoms
Quit
Hide
Help

Marco Punta



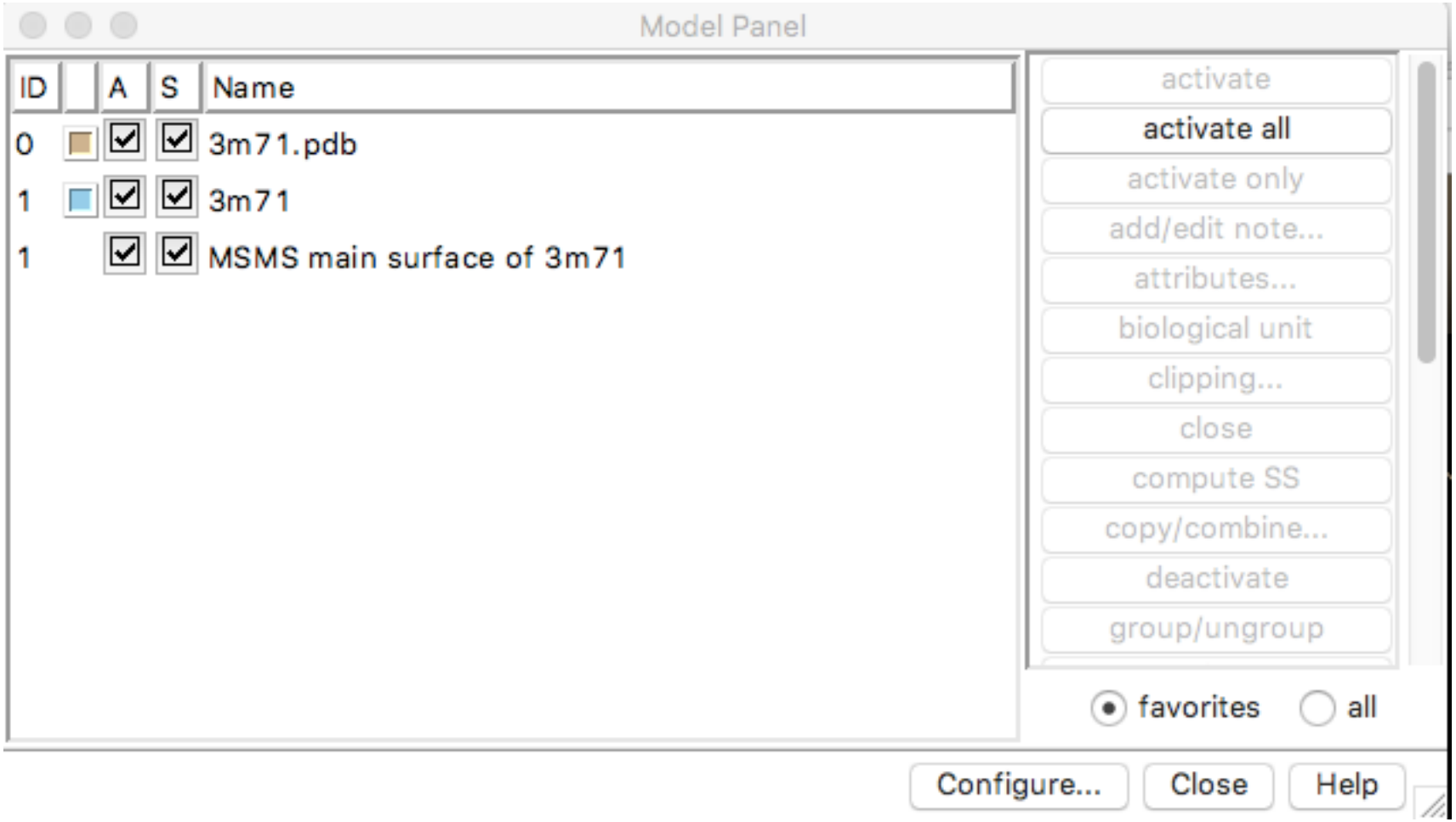
Hold down ‘Shift’ key to select multiple cavities



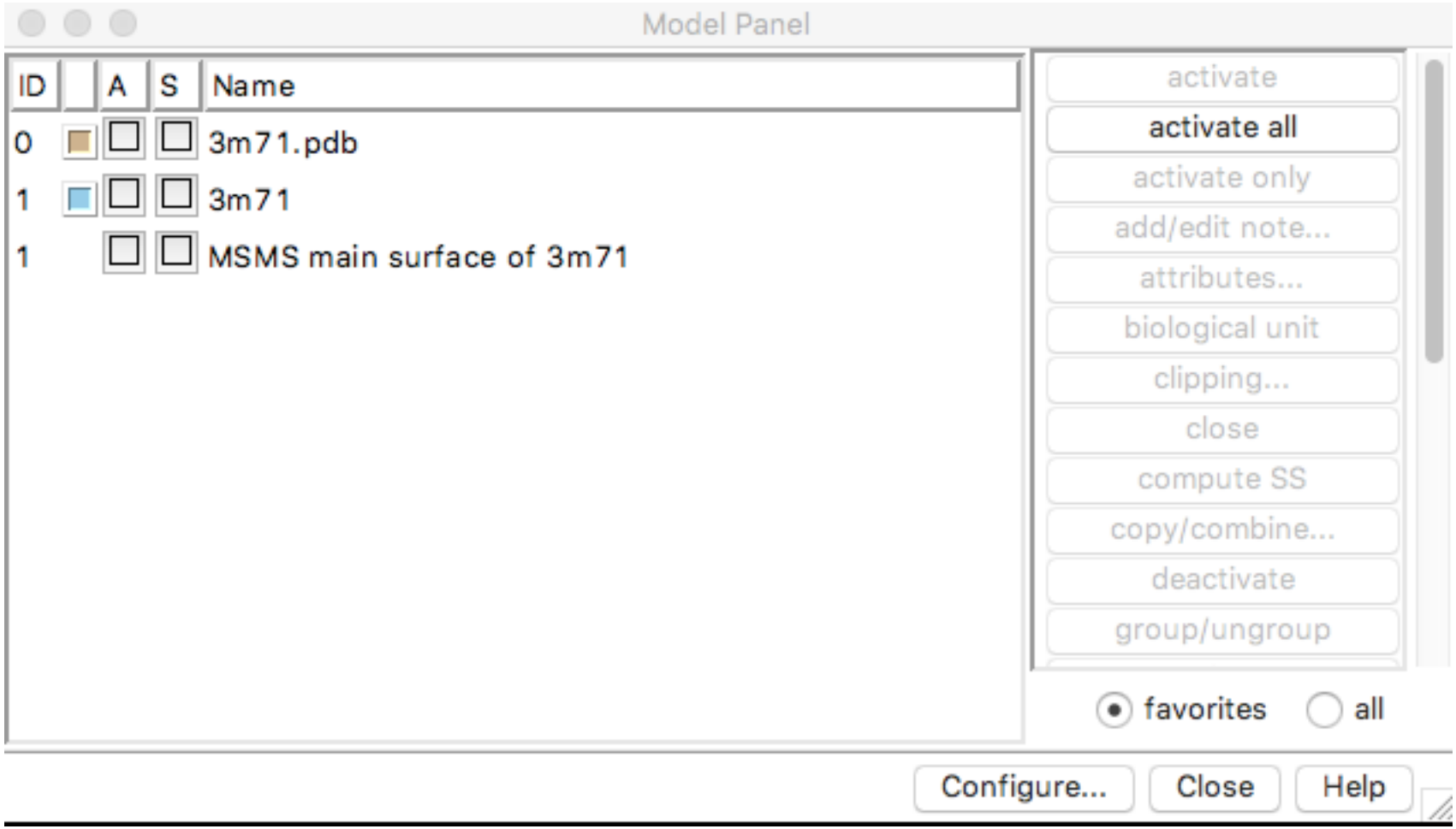


1. Favorites -> Model panel

1. Favorites -> Model panel

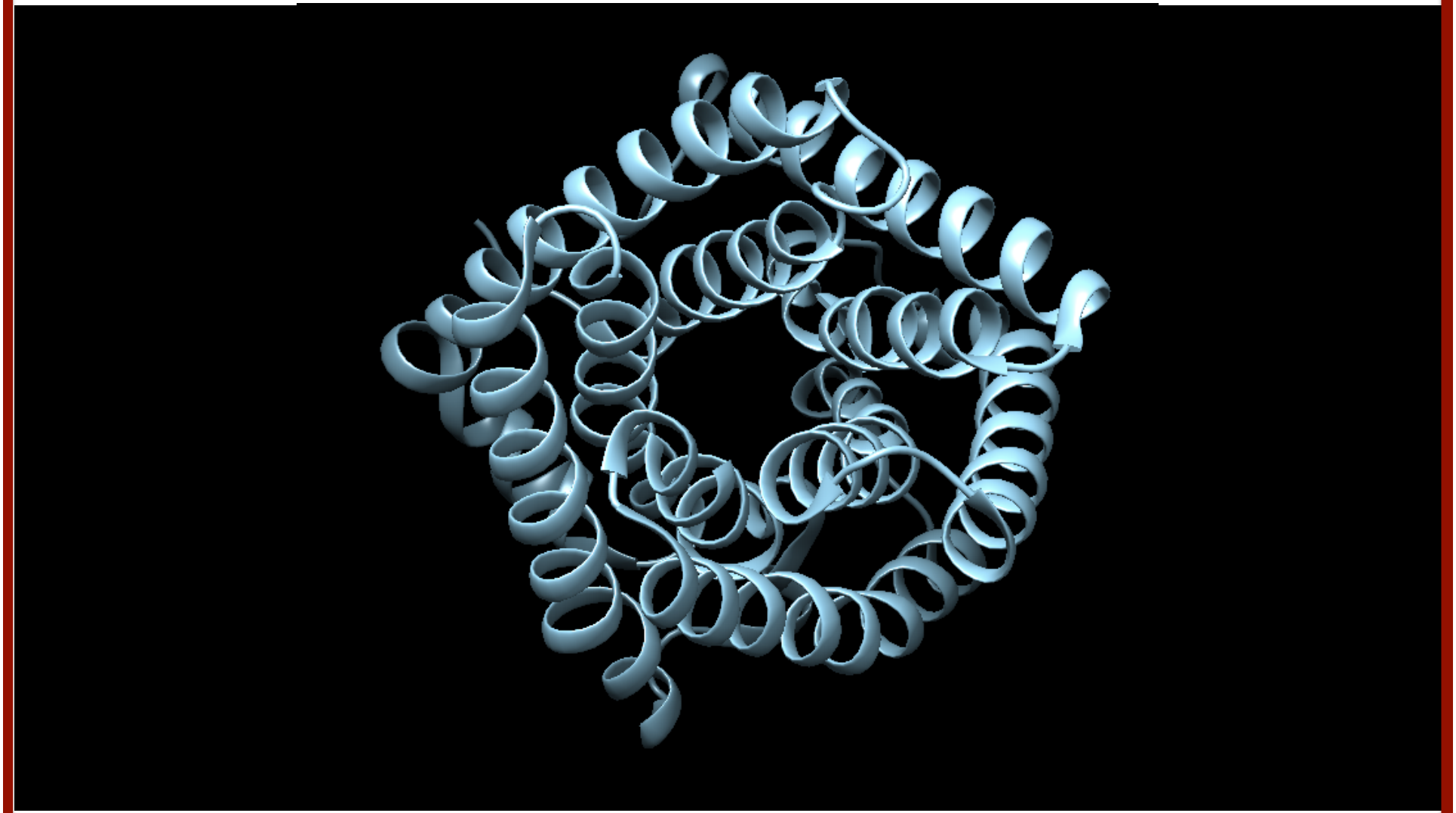


1. Favorites -> Model panel



1. File -> Open “3M71_del.pdb”

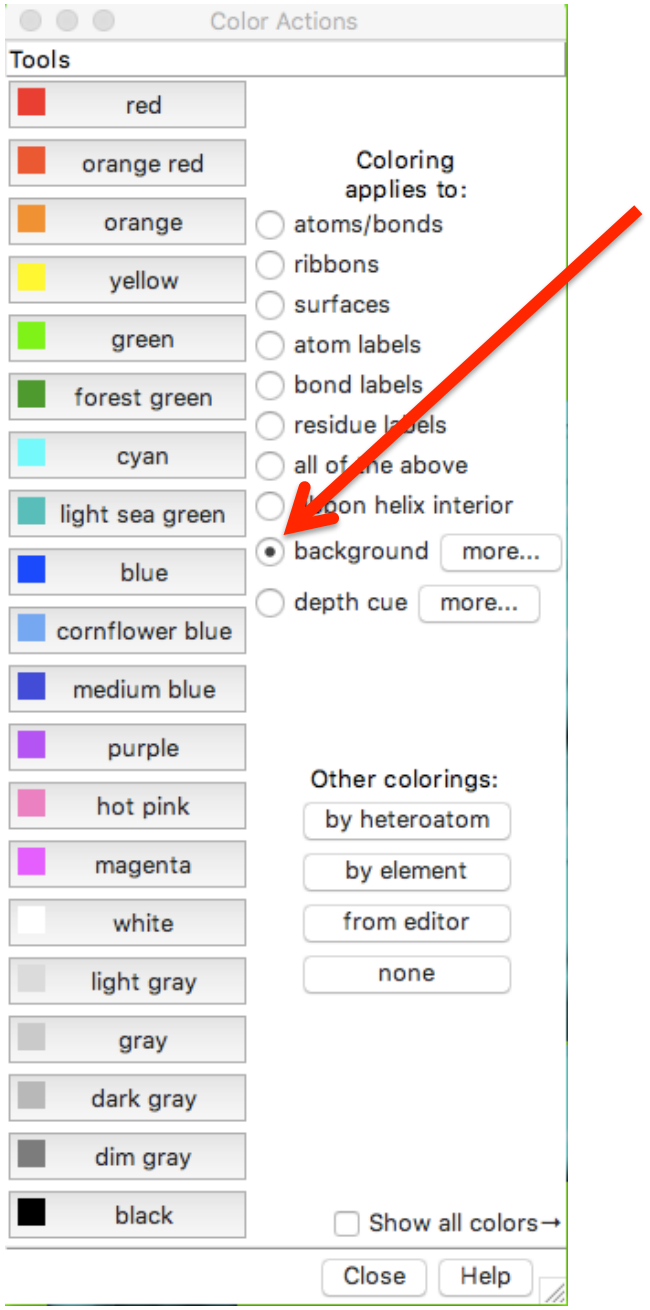
1. File -> Open “3M71_del.pdb”



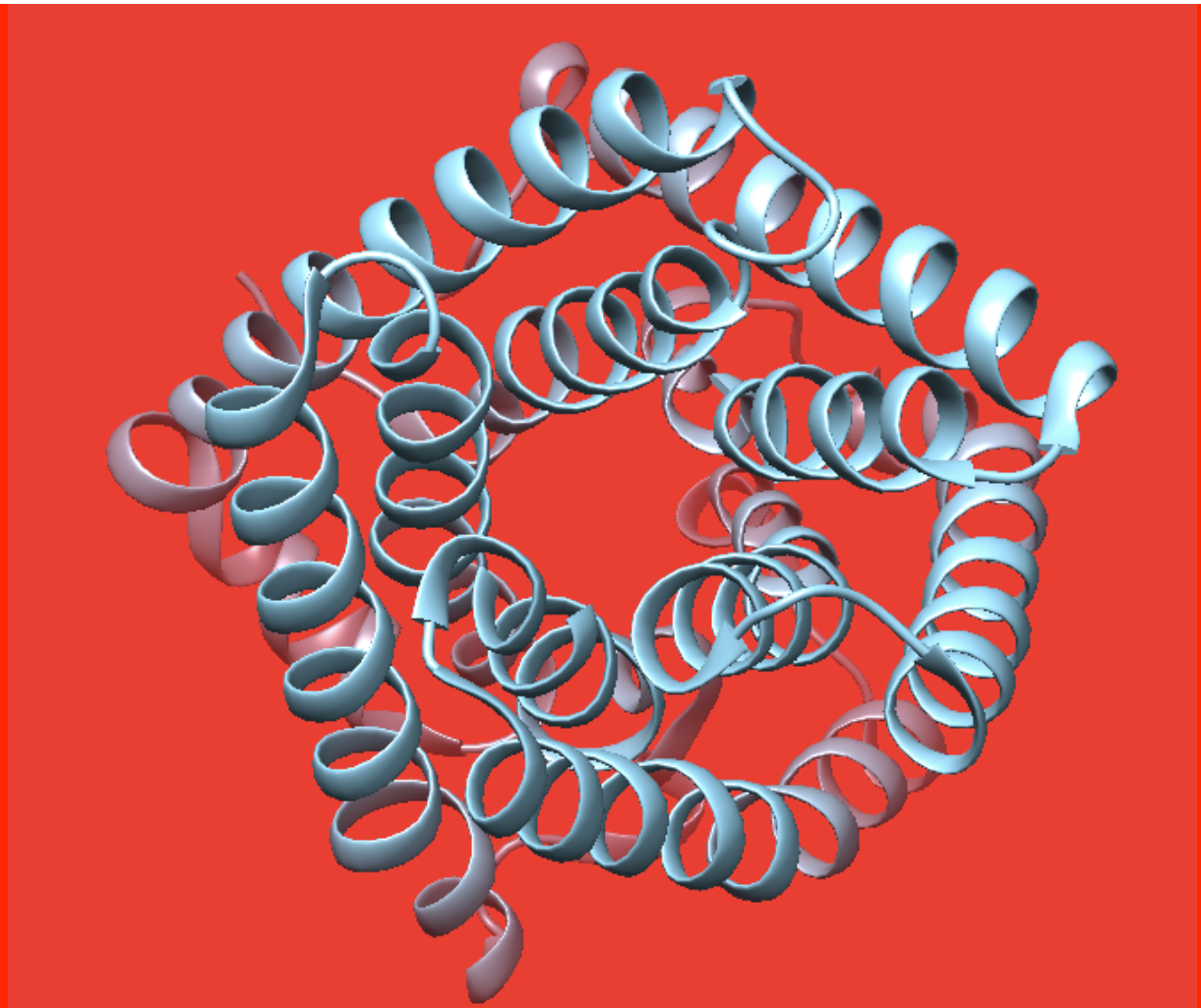
1. File -> Open “3M71_del.pdb”
2. Actions -> Color -> all options

Marco Punta

1. File -> Open “3M71_del.pdb”
2. Actions -> Color -> all options
3. Click on ‘background’ and select red



Marco Punta

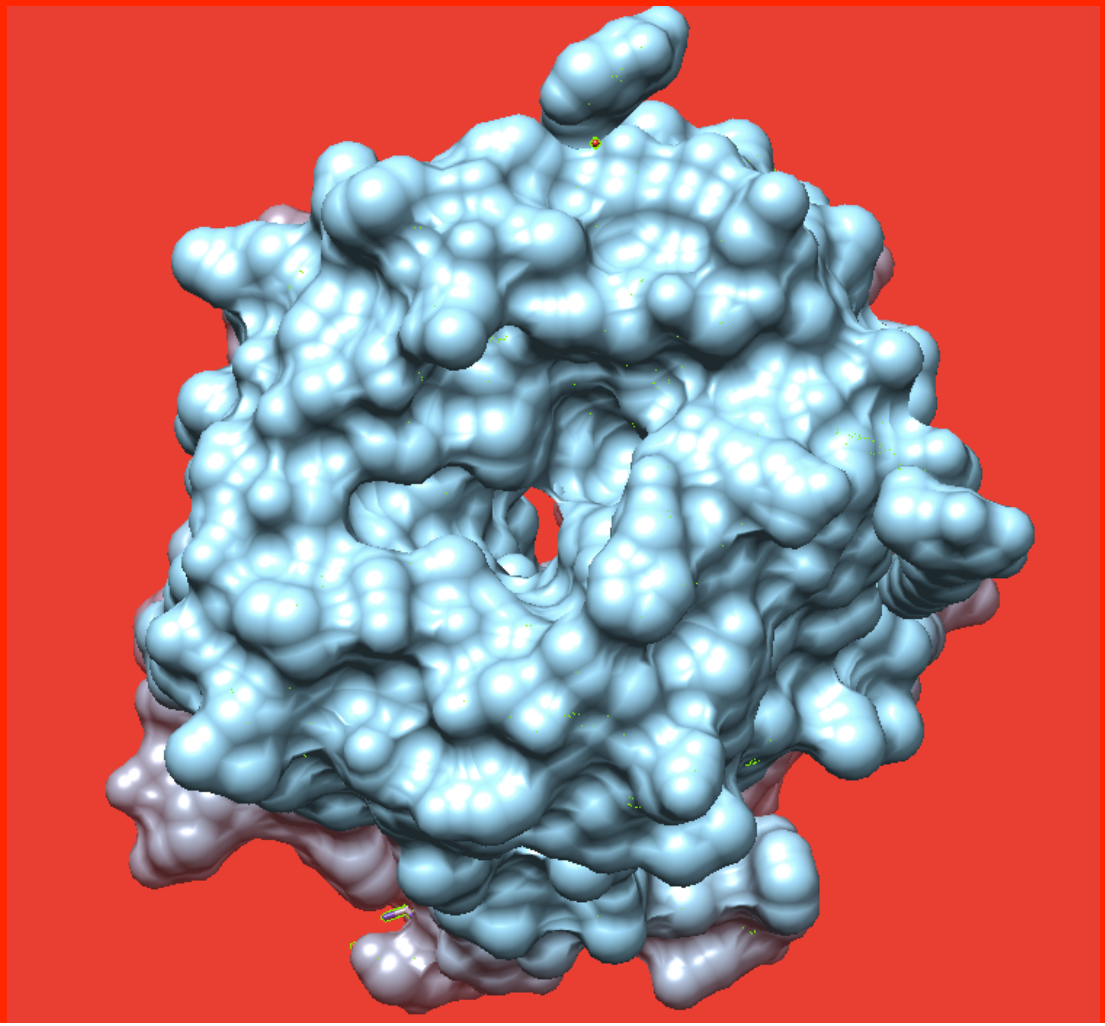


EMBO Workshop, Budapest, 2016

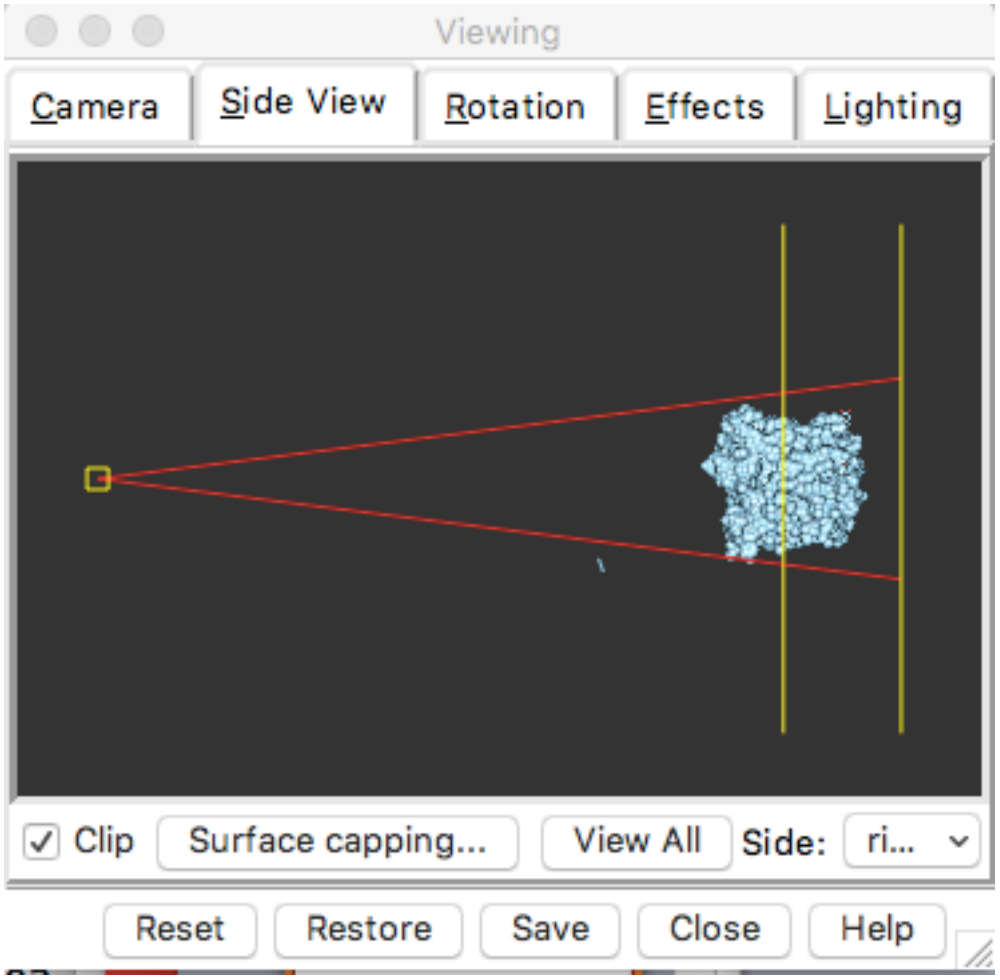
1. Select -> Chain -> A -> 3m71_del.pdb (#1)
2. Actions-> Surface -> show

Marco Punta

1. Select -> Chain -> A -> 3m71_del.pdb (#1)
2. Actions-> Surface -> show

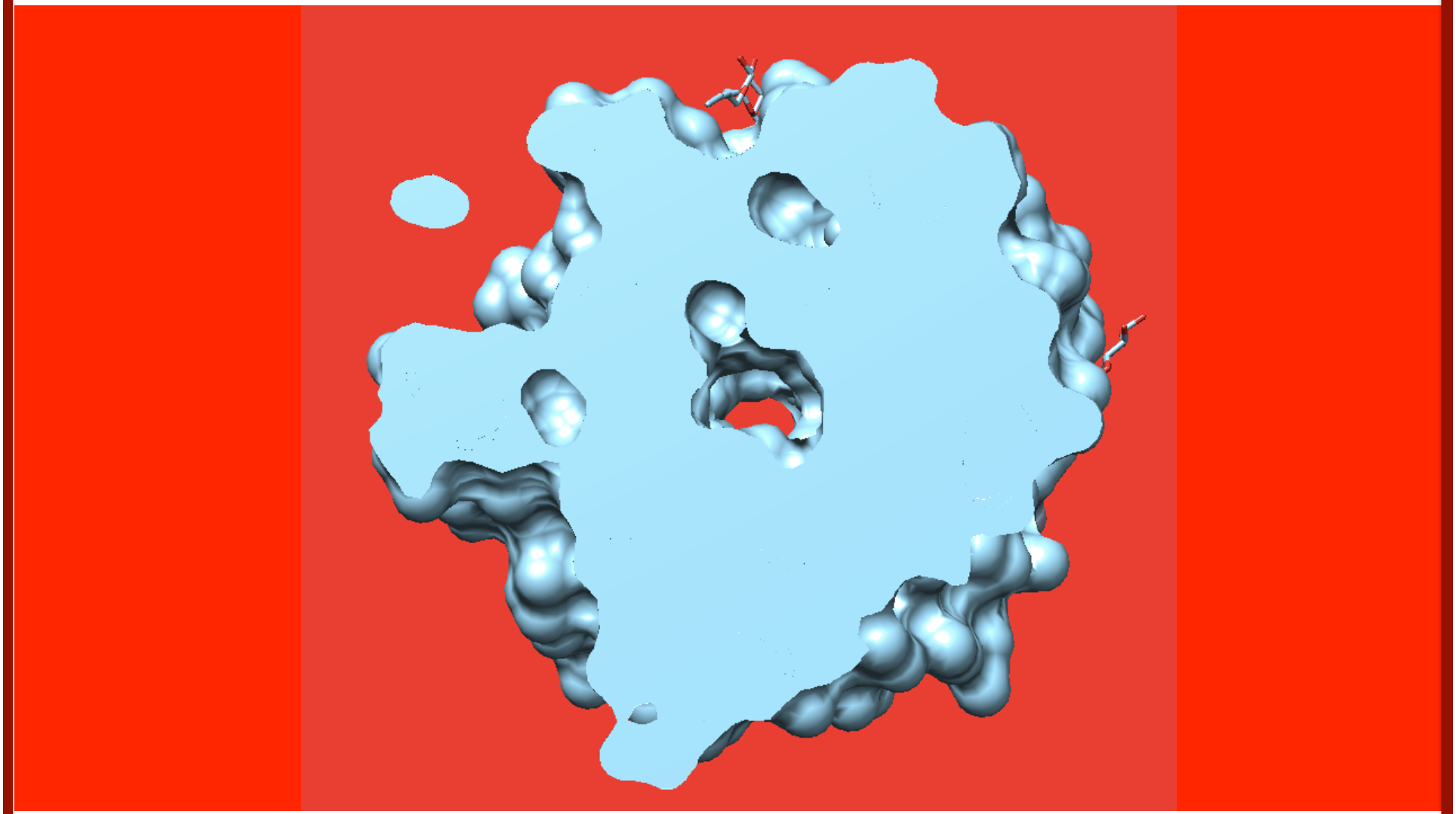


1. Favorites -> Side View

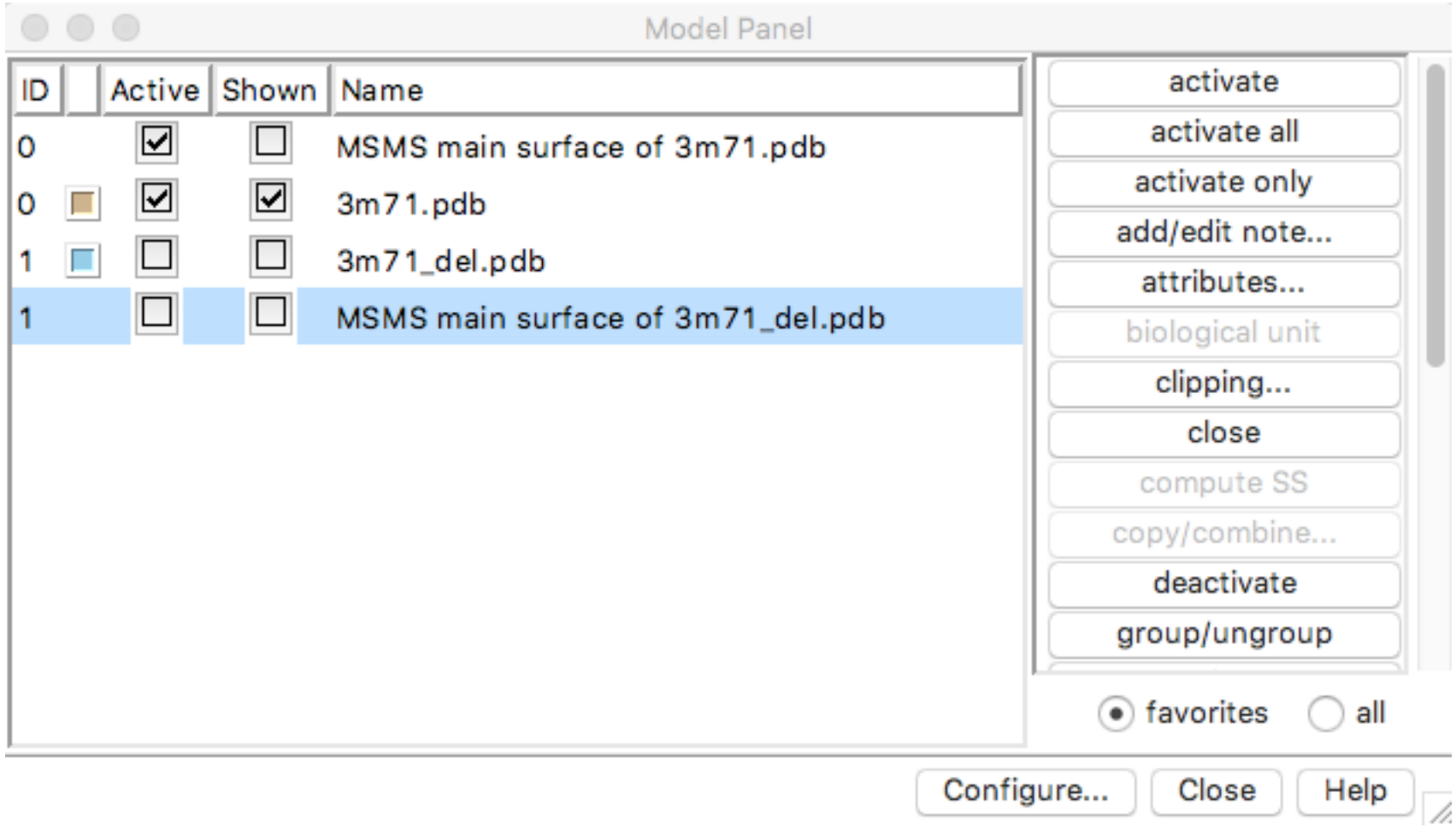


Marco Punta

1. Favorites -> Side View

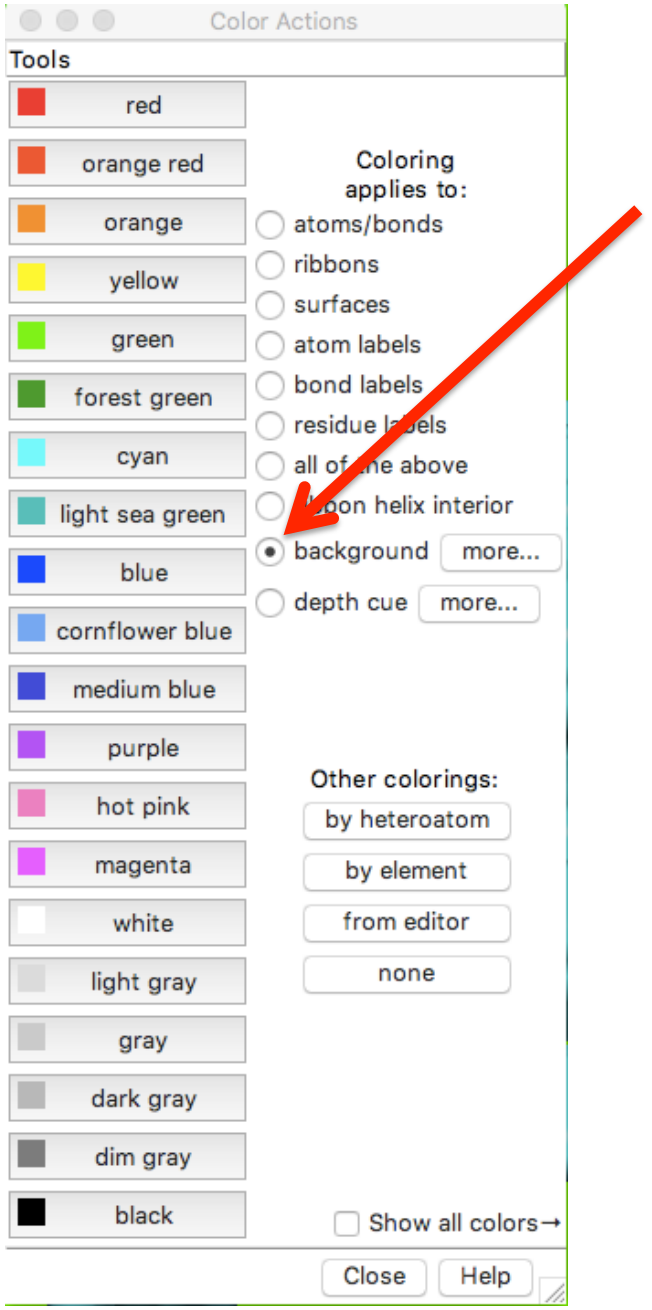


1. Favorites -> Model panel

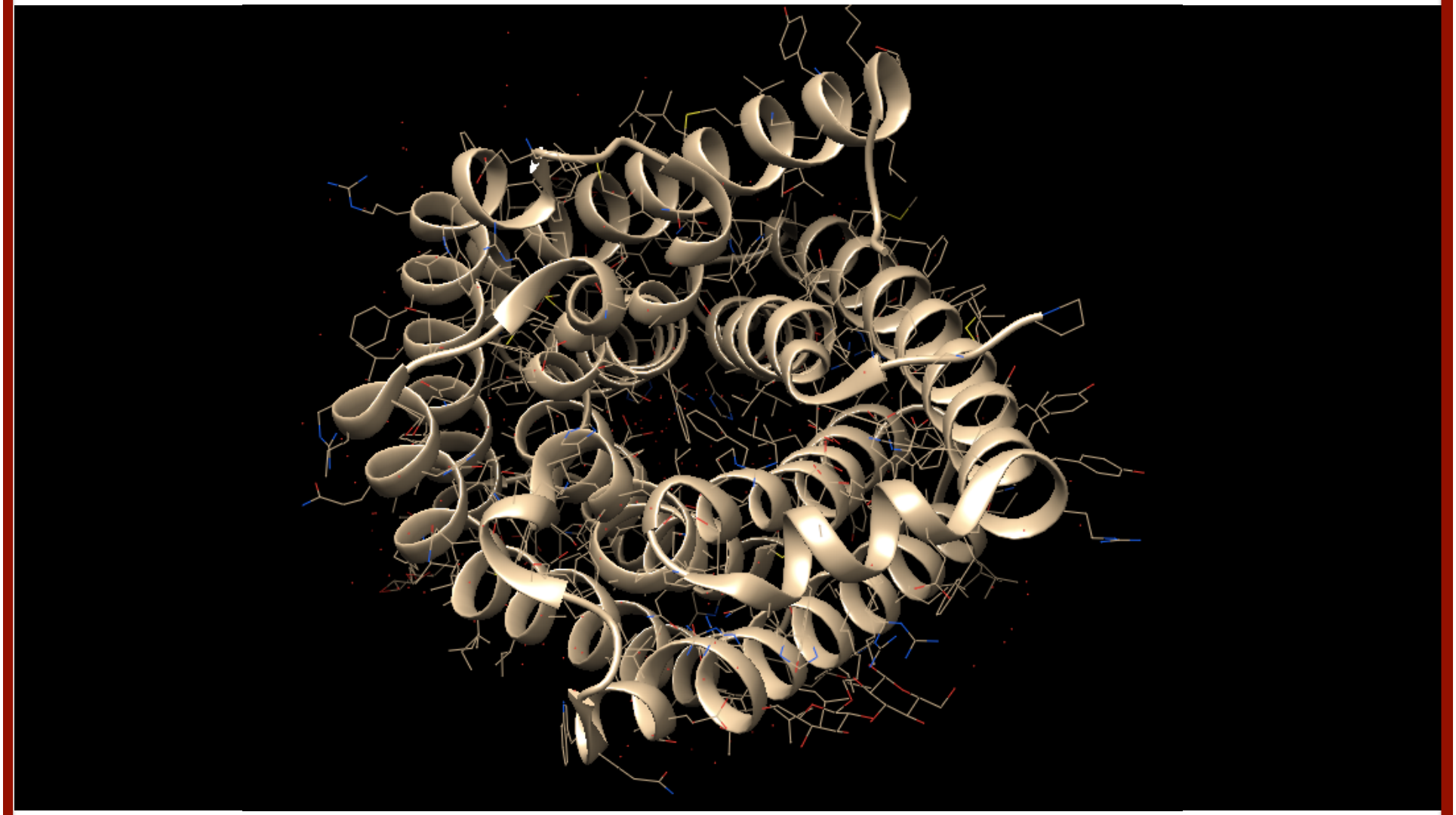


Marco Punta

1. File -> Open “3M71_del.pdb”
2. Actions -> Color -> all options
3. Click on ‘background’ and select black



Marco Punta



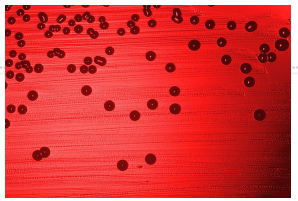
EMBO Workshop, Budapest, 2016

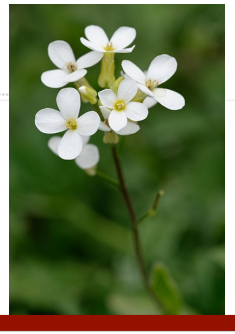
Alignment

Q9LD83 SLAC1_ARATH - **Guard cell S-type anion channel SLA...** - Arabidopsis thal...

E-value: 3e-10 **Positives :** 41.0%
Score: 160 **Query Length:** 328
Ident.: 22.0% **Match Length:** 556

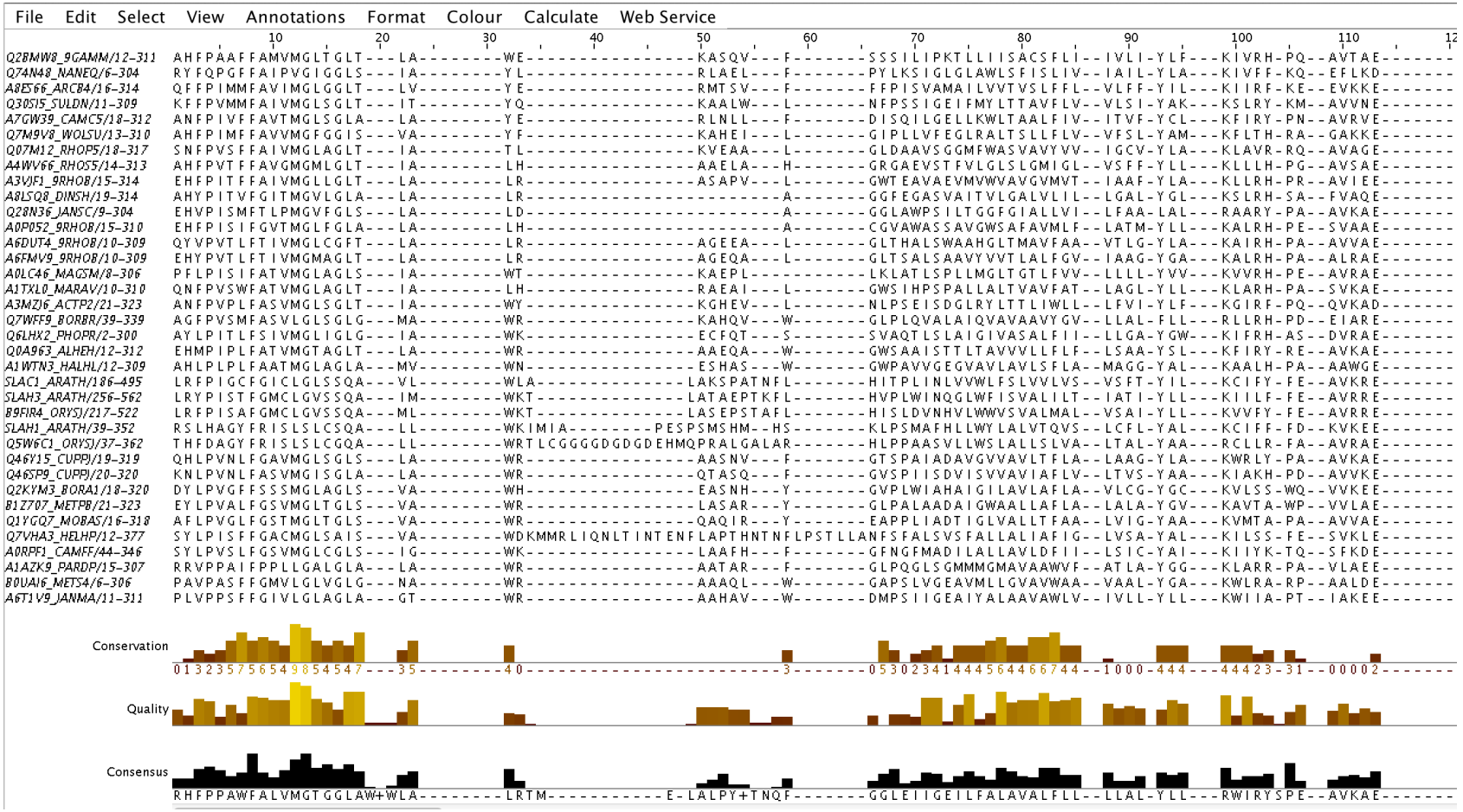
```
P44741      20   PFPL--PTGYFGIPLGLAALSLAWFHLE-----NLFPAARMVSDVLGIVASAVWILFILM   72
                 PF L  P G FGI LGL++ ++ W  L      N      +++ V+ + +  V +
Q9LD83     183   PFLLRFPIGCFGICLGLSSQAVLWLALAKSPATNFLHITPLINLVVWLFSLVVLVSVSFT  242

P44741      73   YAYKLRYYFEEVRAEYHSPVRFSFIALIPITTMLVG---DILYRWNPLIAEVLIWIGTIG  129
                 Y  K  +YFE V+ EY  PVR +F     +  M +      ++  N       IW   +G
Q9LD83     243   YILKCIFYFEAVKREYFHPVRVNFFFAPWVVCMFLAISVPPMFSPNRKYLHPAIWCVFMG  302

P44741     130   QLLFSTLRVSELWQGGVFEQ--KSTHPSFYLPAVAANFTSASSLALLGYHDLGYLFFGAG  187
                    F  L++   W  G   +  K  +PS +L +V  NF  A   + +G+ ++    +  G
Q9LD83     303   PYFFLELKIYGQWLSGGKRRLCKVANPSSHL-SVVGNFVGAILASKVGWDEVAKFLWAVG  361

P44741     188   MIAWIIFEPVLLQHLRISSLEPQFRATMGIVLAPAFVCVSAYLSINHGEVDTLAKILWGY  247
                    +++    L Q L  S   P+    +  +   A    S   +  +G+ D  ++  +
Q9LD83     362   FAHYLVVFVTLYQRLPTSEALPKELHPVYSMFIAAPSAASIAWNTIYGQFDGCSRTCFFI  421

P44741     248   GFLQLFFLLRLFPWIVEKGLNIGLWAFSFGLASMANSATAFY----HGNVLQGVSIFAFV  303
                       L+    +      ++  W++++F  + + A+ AT  Y     G   + +++
Q9LD83     422   ALFLYISLVARINFFTGFKFSVAWWSYTFFMTT-ASVATIKYAEAVPGYPSRALALTLSF  480

P44741     304   FSNVMIGLLVLMTI   317
                  S  M+ +L + T+
Q9LD83     481   ISTAMVCVLFVSTL   494
```





1. OPEN Jalview

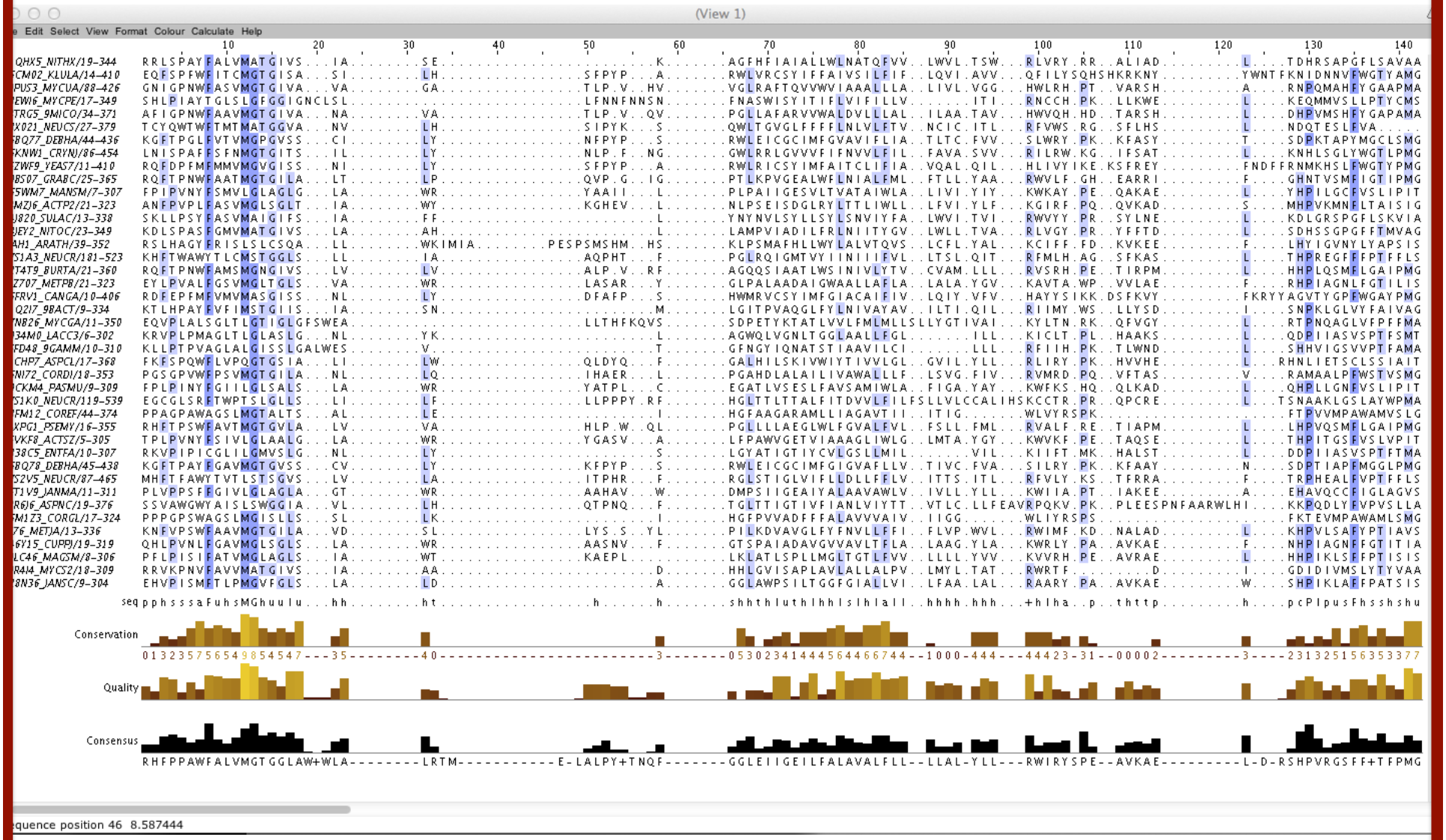
2. File -> Input Alignment -> From File “PF03595_seed.txt”



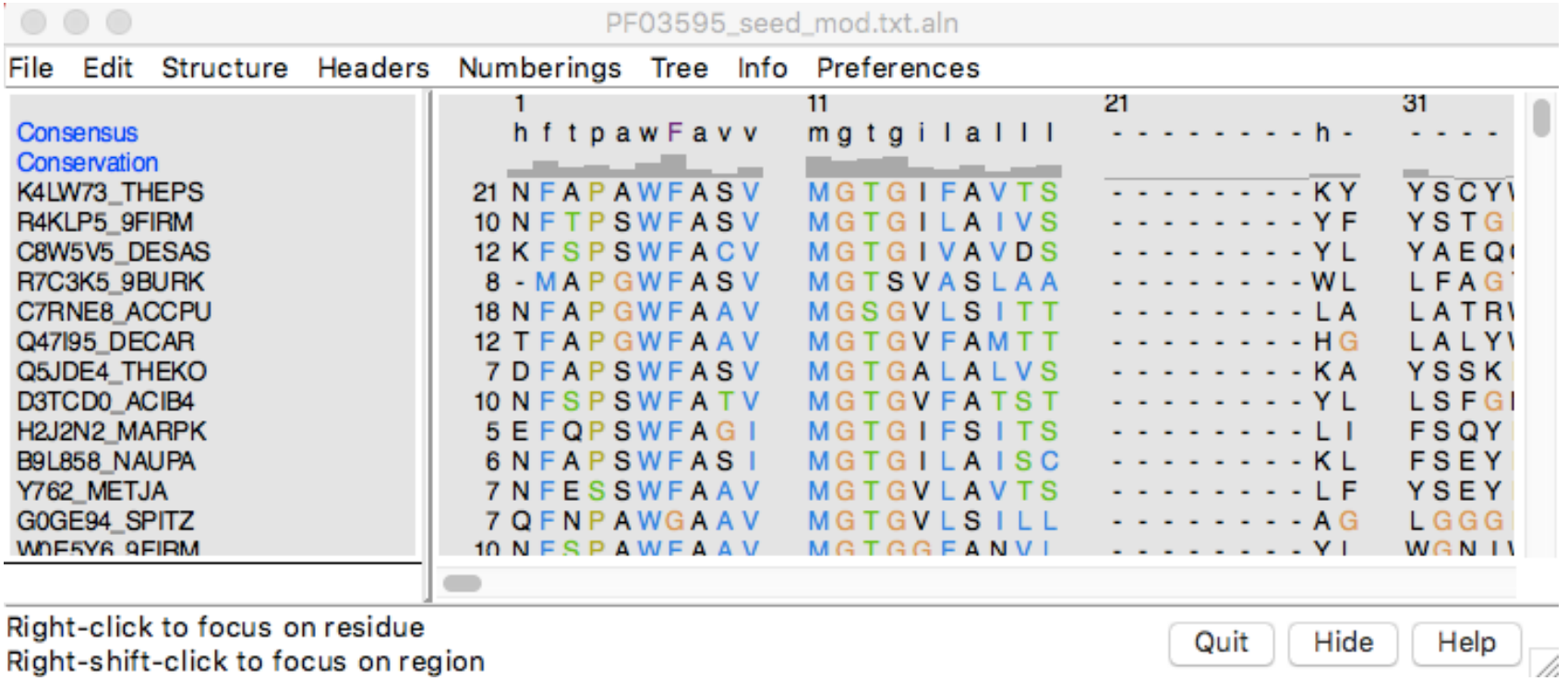
Marco Punta

Marco Punta

1. Colour -> BLOSUM62

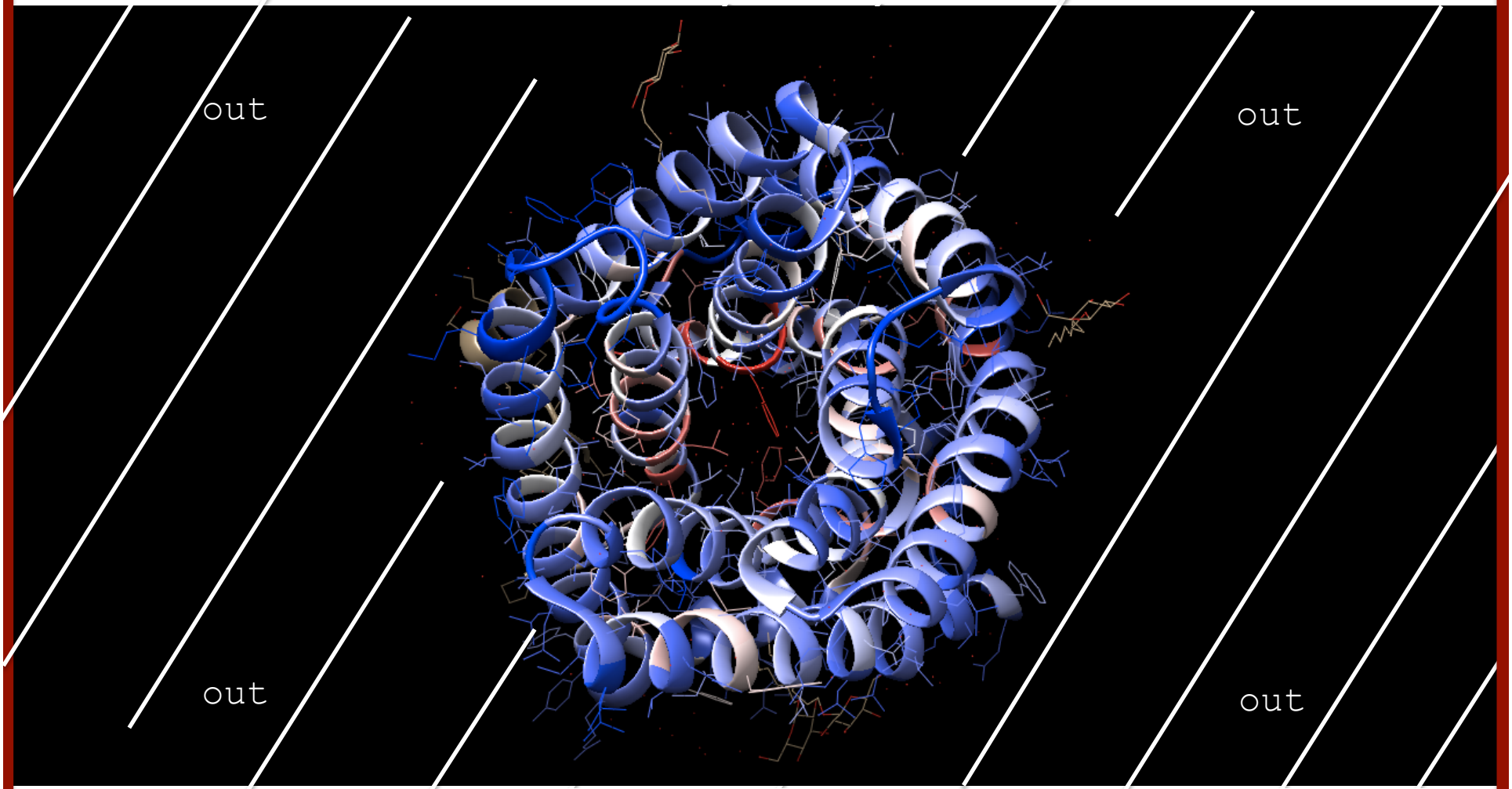


1. Tools-> Sequence -> Multialign Viewer
2. Choose "PF03595_seed_mod.txt"
3. Select Aligned FASTA



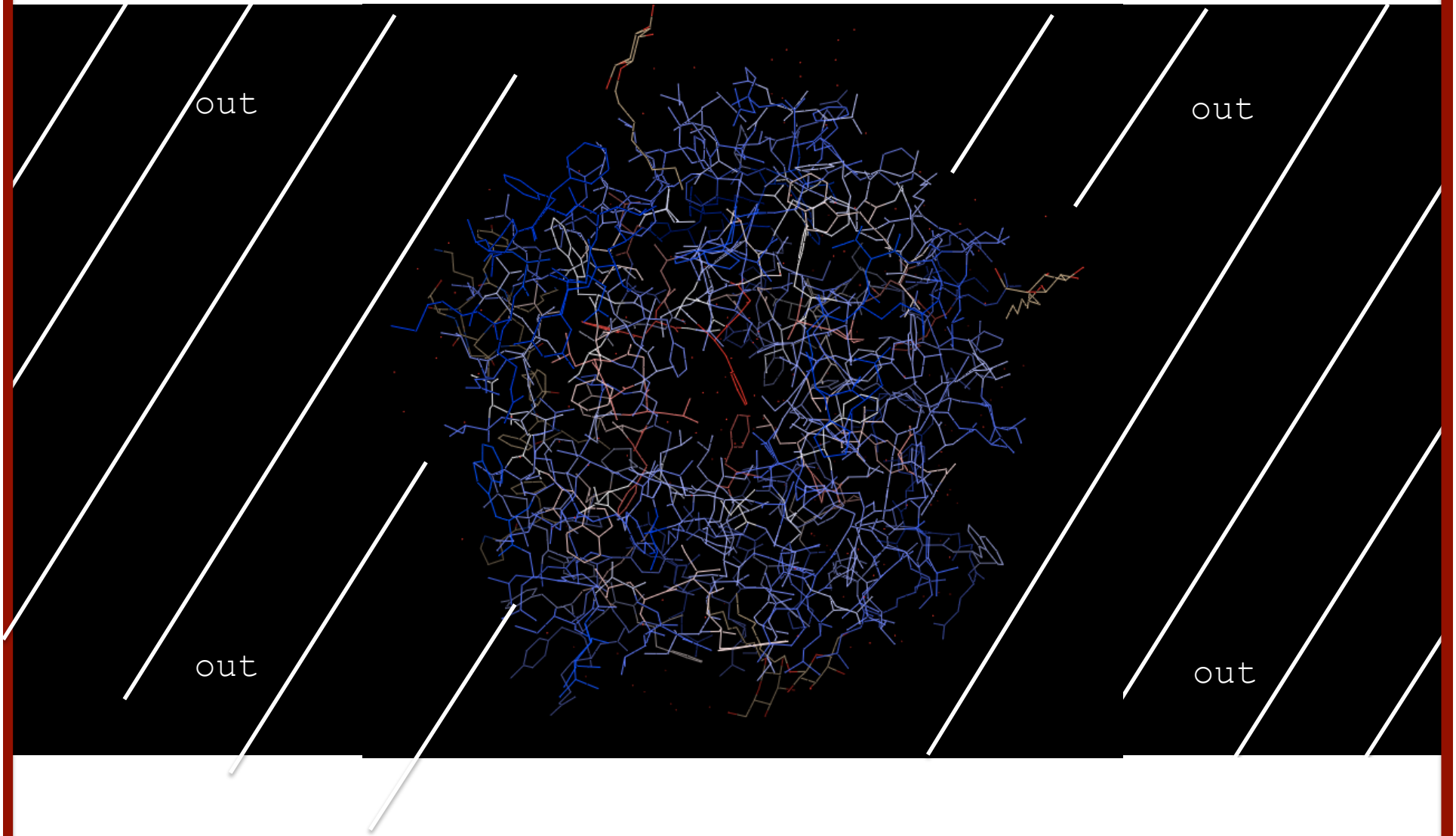
1. Tools-> Sequence -> Multialign Viewer
2. Choose "PF03595_seed_mod.txt"
3. Select Aligned FASTA

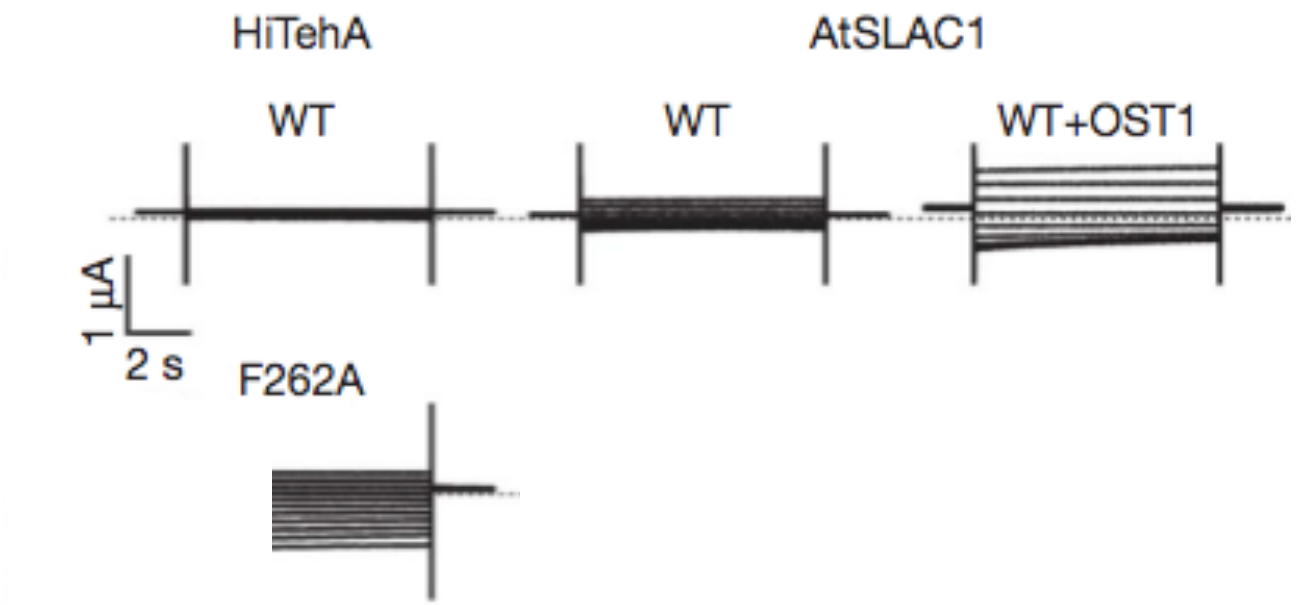
4. Structure -> Render by Conservation


out
out
out
out

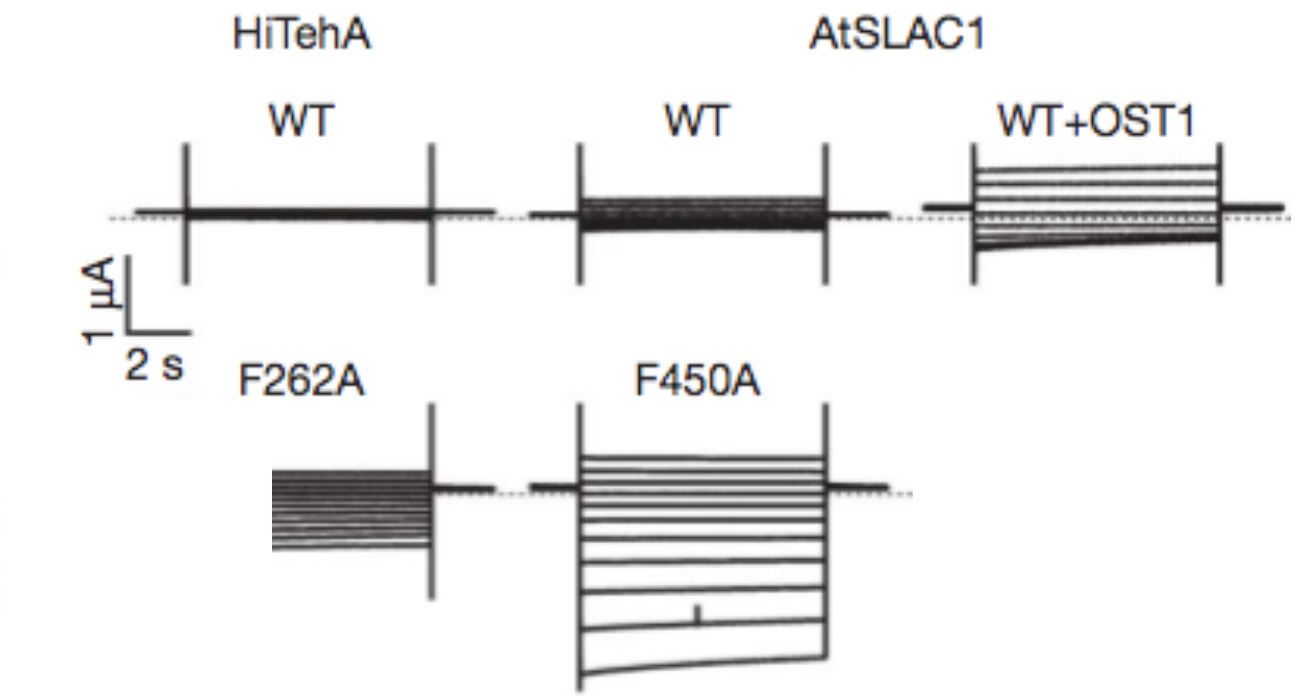
Marco Punta

1. Actions-> Ribbon -> hide

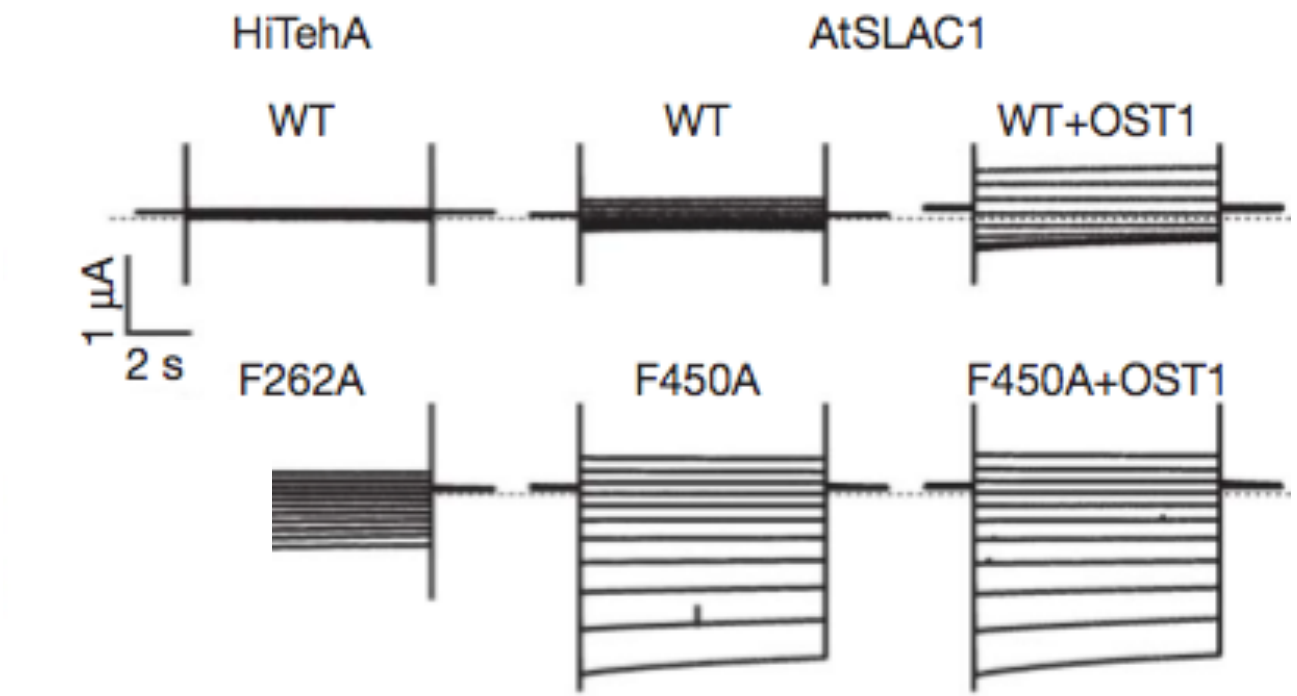




Chen et al. *Nature* 467 (2010)

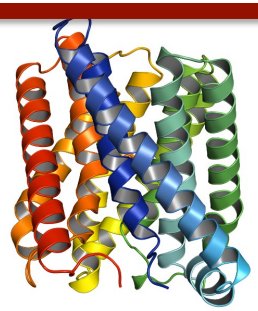


Chen et al. *Nature* 467 (2010)

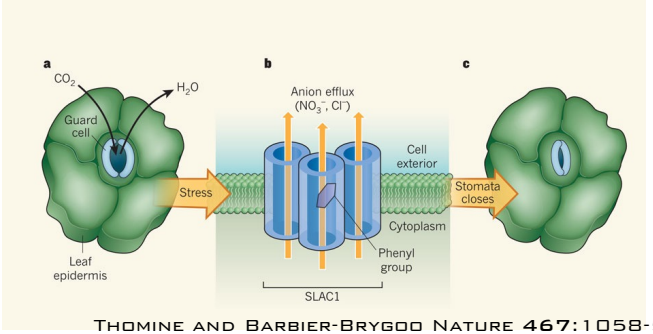


Chen et al. *Nature* 467 (2010)

H. influenzae protein structure

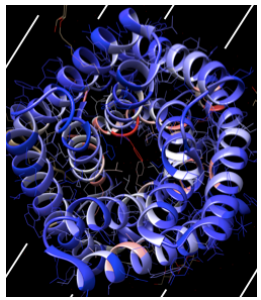


Functional hypothesis via homology to SLAC1

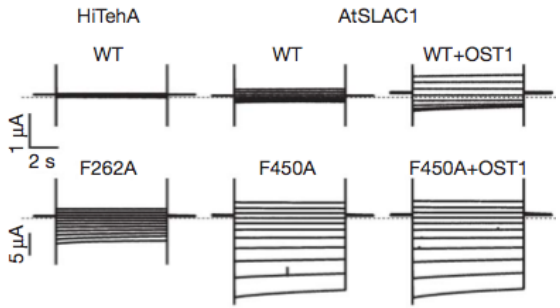


Thomine and Barbier-Brygoo Nature 467:1058-59 (201

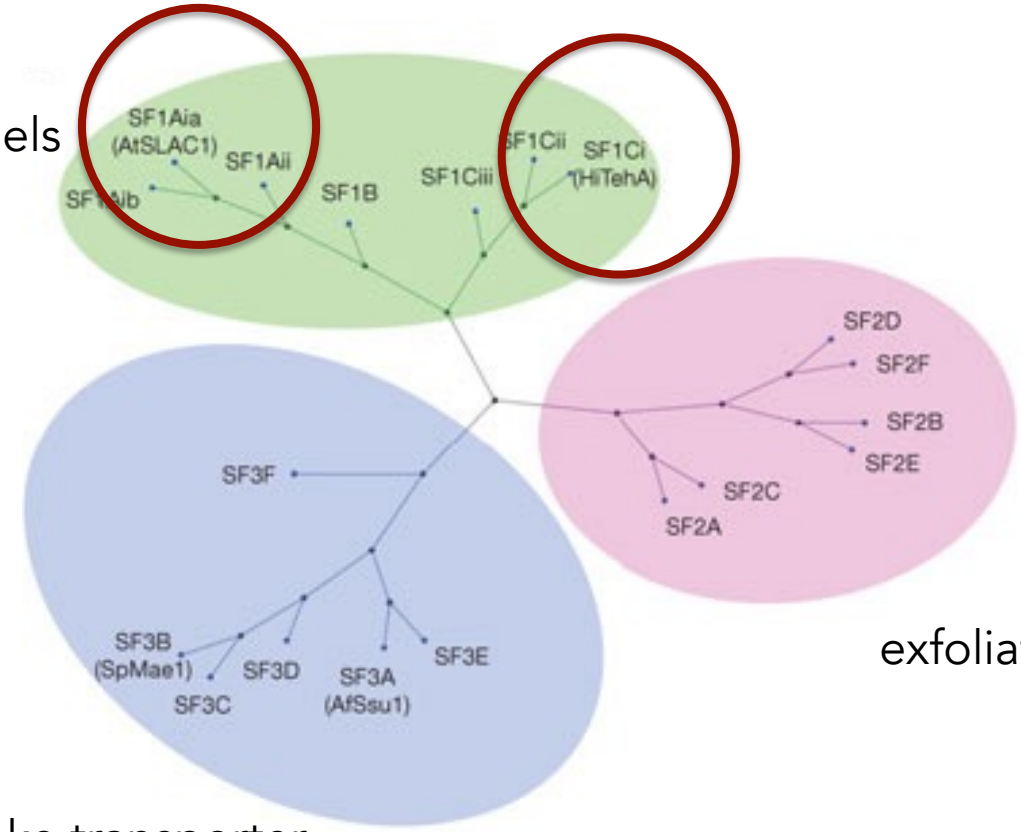
Identification potential functional residues using sequence conservation across the family and structural knowledge



Suggested experiments to test functional hypothesis



Anion channels



exfoliative toxins

malate uptake transporter
sulphite efflux pump

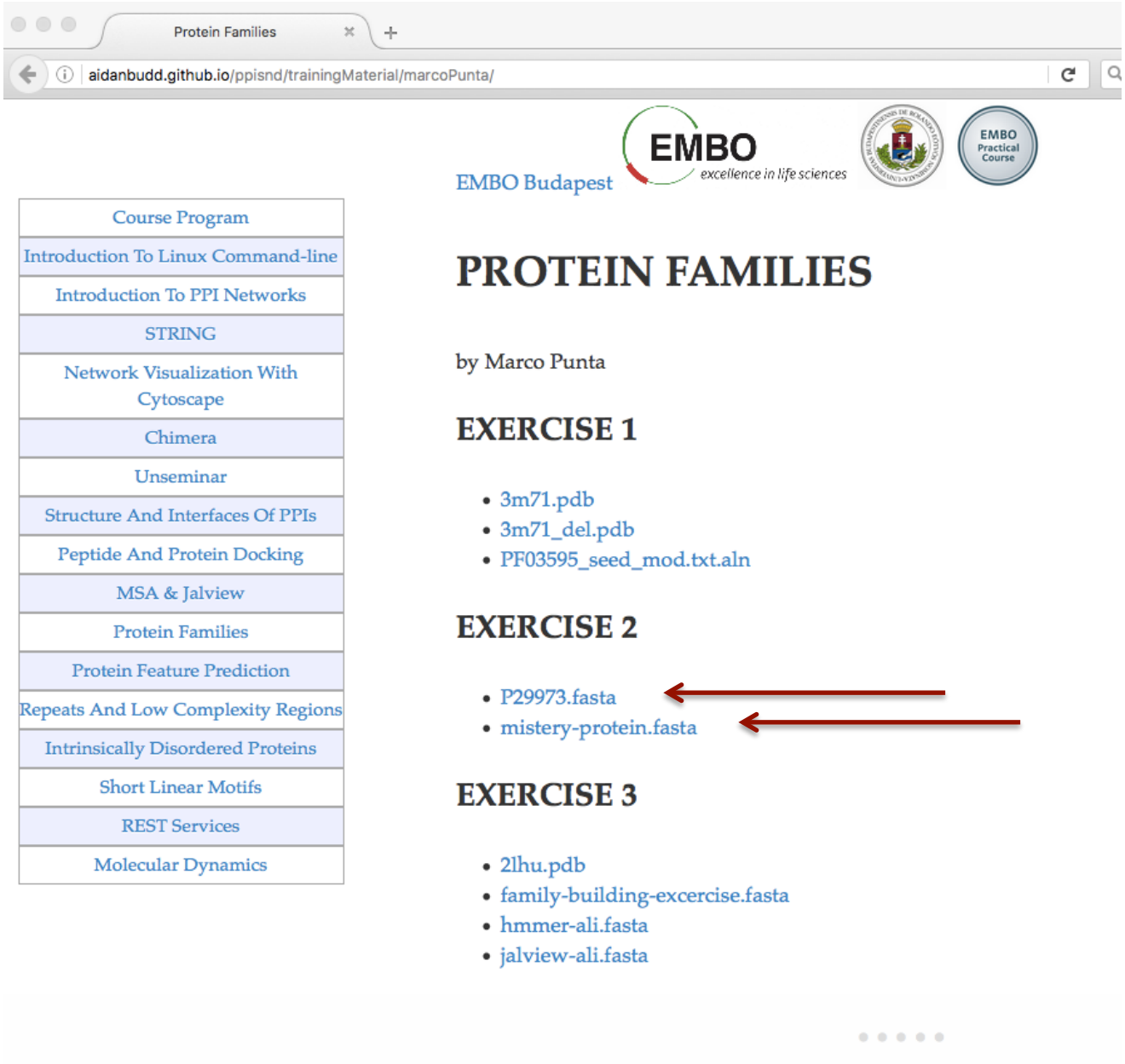
Chen et al. *Nature* 467 (2010)

Marco Punta

Exercise

Homology-based function annotation transfer #2

Marco Punta



http://aidanbudd.github.io/ppisnd/trainingMaterial/marcoPunta/

>sp|P29973|CNGA1_HUMAN cGMP-gated cation channel alpha-1 OS=Homo sapiens

MKLSMKNNIINTQQSFVTMPNVIVPDIEKEIRRMENGACSSFSEDDDSASTSEESENENP
HARGSFSYKSLRKGGPSQREQYLPGAIALFNVNNSSNKDQEPEEKKKKKKEKKSKSDDKN
ENKNDPEKKKKKKDKEKKKKEEKSKDKKEEEKKEVVVIDPSGNTYYNWLFCITLPVMYNW
TMVIARACFDELQSDYLEYWLILDYVSDIVYLIDMFVRTRTGYLEQGLLVKEELKLINKY
KSNLQFKLDVLSLIPTDLLYFKLGWNYPEIRLNRLLRFSRMFEFFQRTETRTNYPNIFRI
SNLVMYIVIIIHWNACVFYSISKAIGFGNDTWVYPDINDPEFGRLARKYVYSLYWSTLTL
TTIGETPPPVRDSEYVFVVVDFLIGVLIFATIVGNIGSMISNMNAARAEFQARIDAIKQY
MHFRNVSKDMEKRVIKWFDYLWTNKKTVDEKEVLKYLPDKLRAEIAINVHLDTLKKVRIF
ADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVVADDGVTQFVVLSD
GSYFGEISILNIKGSKAGNRRTANIKSIGYSDLFCLSKDDLMEALTEYPDAKTMLEEKGK
QILMKDGLLDLNIANAGSDPKDLEEKVTRMEGSVDLLQTRFARILAEYESMQQKLKQRLT
KVEKFLKPLIDTEFSSIEGPGAESGPIDST

```
>mistery protein
MGNGSVKPKHSKHPDGHSGNLTTDALRNKVTELERELRRKDAEIQEREYHLKELREQLSK
QTVAIAELTEELQNKCIQLNKLQDVVHMQGGSPLQASPDKVPLEVHRKTSGLVSLHSRRG
AKAGVSAEPTTRTYDLNKPPEFSFEKARVRKDSSEKKLITDALNKNQFLKRLDPQQIKDM
VECMYGRNYQQGSYIIKQGEPGNHIFVLAEGRLEVFQGEKLLSSIPMWTTFGELAILYNC
TRTASVKAITNVKTWALDREVFQNIMRRTAQARDEQYRNFLRSVSLLKNLPEDKLTKIID
CLEVEYYDKGDYIIREGEEGSTFFILAKGKVKVTQSTEGHDQPQLIKTLQKGEYFGEKAL
ISDDVRSANIIAEENDVACLVIDRETFNQTVGTFEELQKYLEGYVANLNRDDEKRHAKRS
MSNWKLSKALSLEMIQLKEKVARFSSSSPFQNLEIIATLGVGGFGRVELVKVKNENVAFA
MKCIRKKHIVDTKQQEHVYSEKRILEELCSPFIVKLYRTFKDNKYVYMLLEACLGGELWS
ILRDRGSFDEPTSKFCVACVTEAFDYLHRLGIIYRDLKPENLILDAEGYLKLVDFGFAKK
IGSGQKTWTFCGTPEYVAPEVILNKGHDFSVDFWSLGILVYELLTGNPPFSGVDQMMTYN
LILKGIEKMDFPRKITRRPEDLIRRLCRQNPTERLGNLKNGINDIKKHRWLNGFNWEGLK
ARSLPSPLQRELKGPIDHSYFDKYPPEKGMPPDELSGWDKDF
```

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins

BLAST®

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastp suite

Align Sequences Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein subjects using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

```
TTIGETPPPVRDSEYVFVVVDFLIGVLIFATIVGNIGSMISNMNAARAEFQARIDAIKQY
MHFRNVSKDMEKRVIKWFDYLWTNKKTVDEKEVLKYLPDKLRAEIAINVHLDTLKKVRIF
ADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVVADDGVTQFVVLSD
GSYFGEISILNIKGSKAGNRRTANIKSIGYSDLFCLSKDDLMEALTEYPDAKTMLEEKGK
QILMKDGLLDLNIANAGSDPKDLEEKVTRMEGSVDLLQTRFARILAEYESMQQKLKQRLT
KVEKFLKPLIDTEFSSIEGPGAESGPIDST
```

Query subrange

From

To

Or, upload file Browse... 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

```
MSNWKLSKALSLEMIQLKEKVARFSSSSPFQNLEIIATLGVGGFGRVELVKVKNENVAFA
MKCIRKKHIVDTKQQEHVYSEKRILEELCSPFIVKLYRTFKDNKYVYMLLEACLGGELWS
ILRDRGSFDEPTSKFCVACVTEAFDYLHRLGIIYRDLKPENLILDAEGYLKLVDFGFAKK
IGSGQKTWTFCGTPEYVAPEVILNKGHDFSVDFWSLGILVYELLTGNPPFSGVDQMMTYN
LILKGIEKMDFPRKITRRPEDLIRRLCRQNPTERLGNLKNGINDIKKHRWLNGFNWEGLK
ARSLPSPLQRELKGPIDHSYFDKYPPEKGMPPDELSGWDKDF
```

Subject subrange

From

To

Or, upload file Browse... No file selected.

Alignments

Download Graphics Sort by: E value

unnamed protein product

Sequence ID: lcl|Query_22995 **Length:** 762 **Number of Matches:** 3

Range 1: 281 to 383 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
41.2 bits(95)	1e-07	Compositional matrix adjust.	30/110(27%)	54/110(49%)	11/110(10%)

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

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

Range 2: 161 to 260 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
38.1 bits(87)	1e-06	Compositional matrix adjust.	26/108(24%)	53/108(49%)	8/108(7%)

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

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

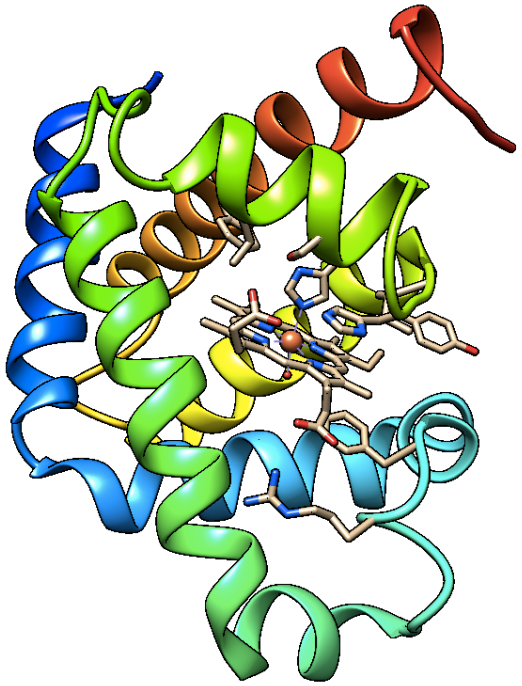
Range 3: 593 to 649 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
22.3 bits(46)	0.081	Compositional matrix adjust.	17/58(29%)	27/58(46%)	7/58(12%)

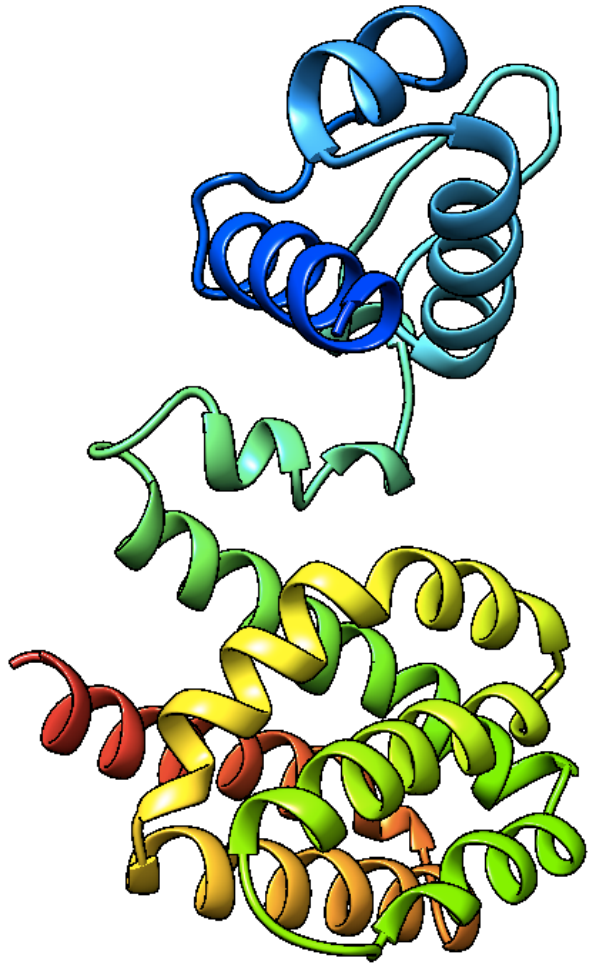
```
Query  317  VFYSISKAIGFGNDTWVY---PDINDPEFGRLARKYVYSL-YWS--TLTLTTIGETPP  368
            V +  +K IG G  TW +   P+   PE   L + + +S+ +WS   L    +   PP
Sbjct  593  VDFGFAKKIGSGQKTWTFCGTPEYVAPEV-ILNKGHDFSVDFWSLGILVYELLTGNPP  649
```

Mystery protein is a cGMP-gated cation channel?

Marco Punta



1MBN

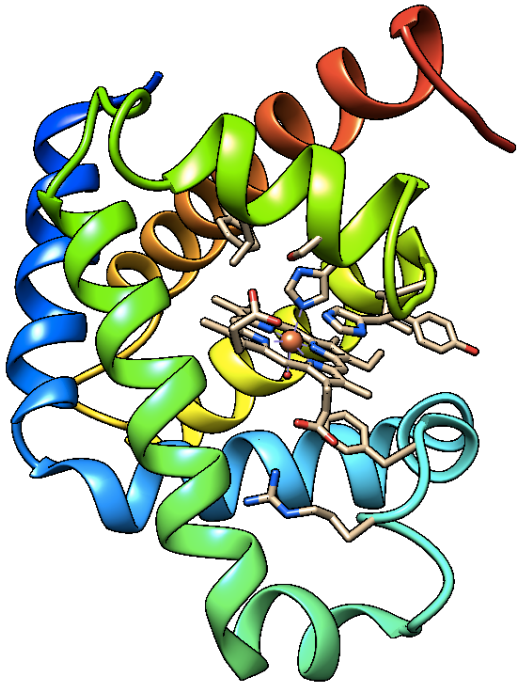


1HW2

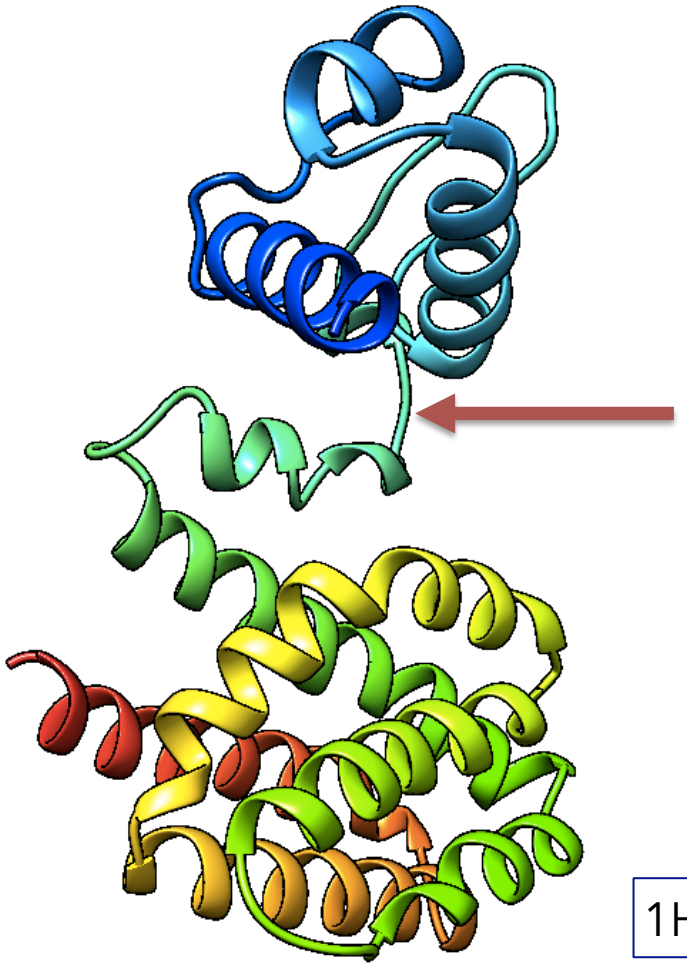
Colour Scheme:

N’ **C’**

Marco Punta



1MBN

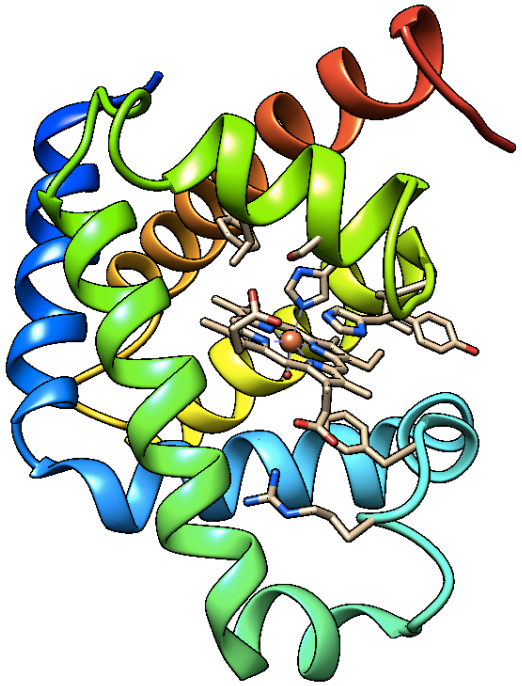


1HW2

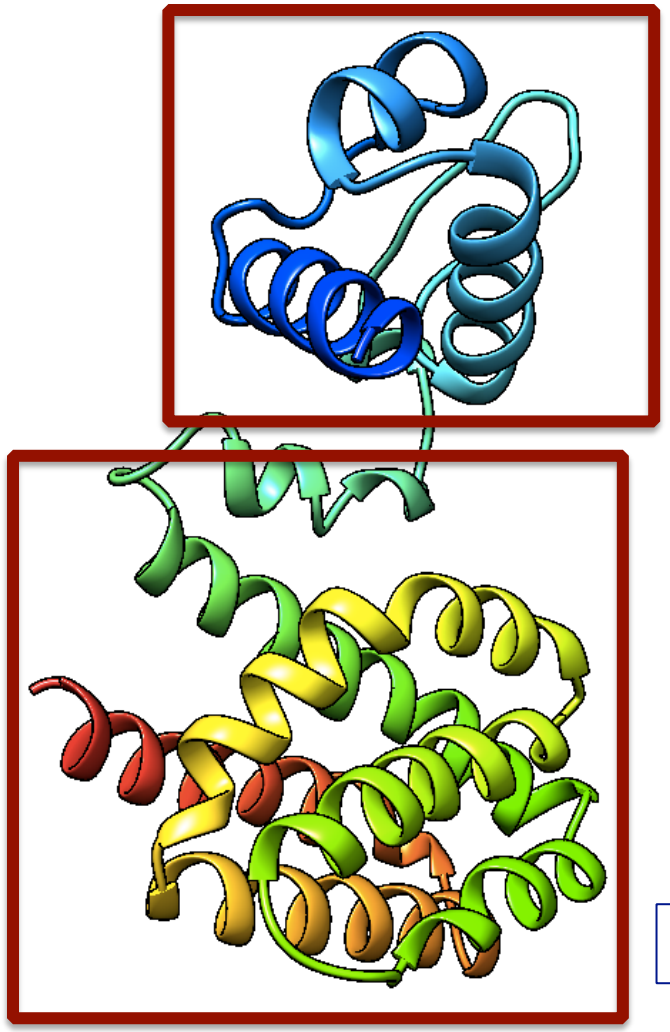
Colour Scheme:

N’ **C’**

Marco Punta



1MBN



1HW2

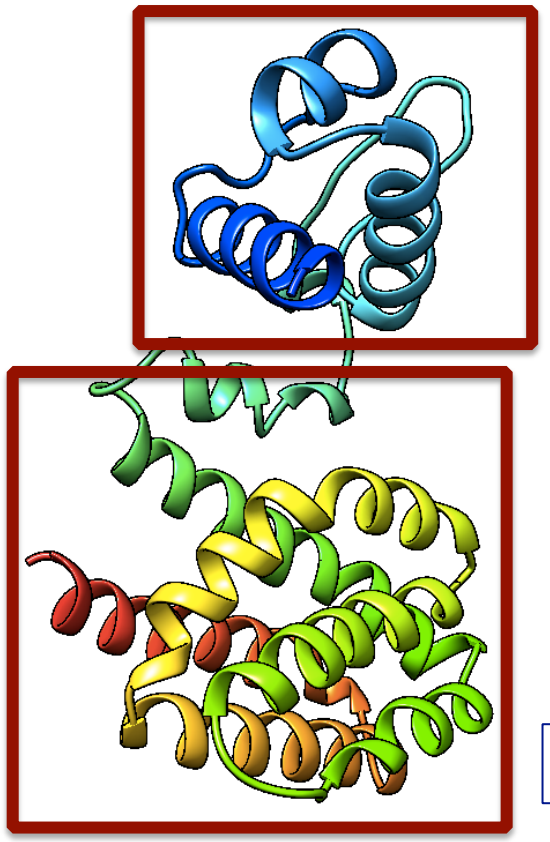
Colour Scheme:

N’ **C’**

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.

Marco Punta



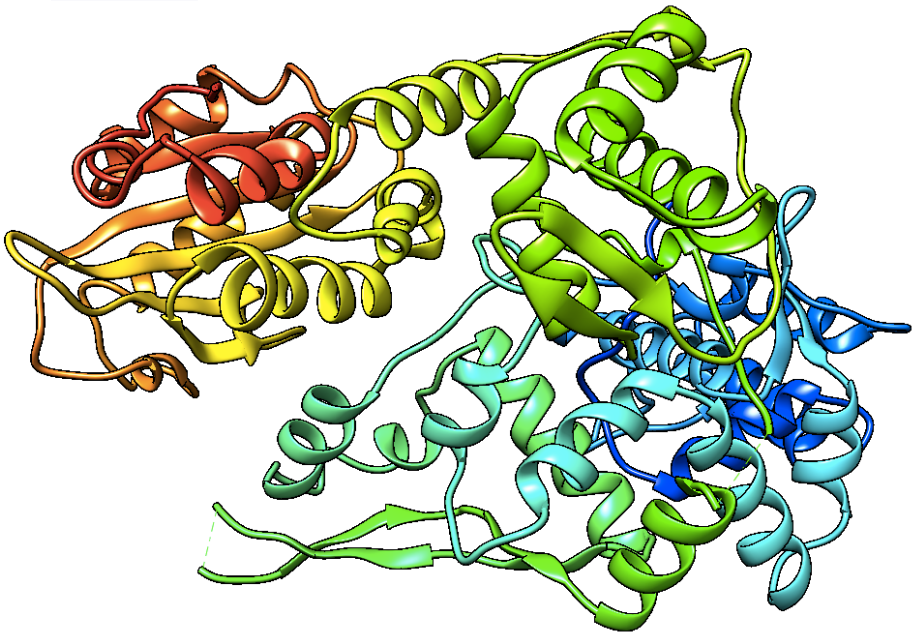
1HW2

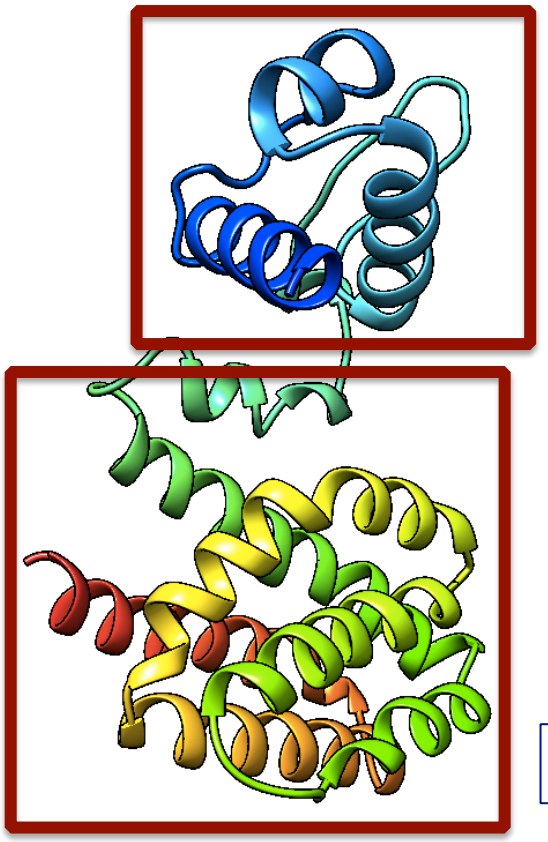
Colour Scheme:

N’ **C’**

Marco Punta

1FOK





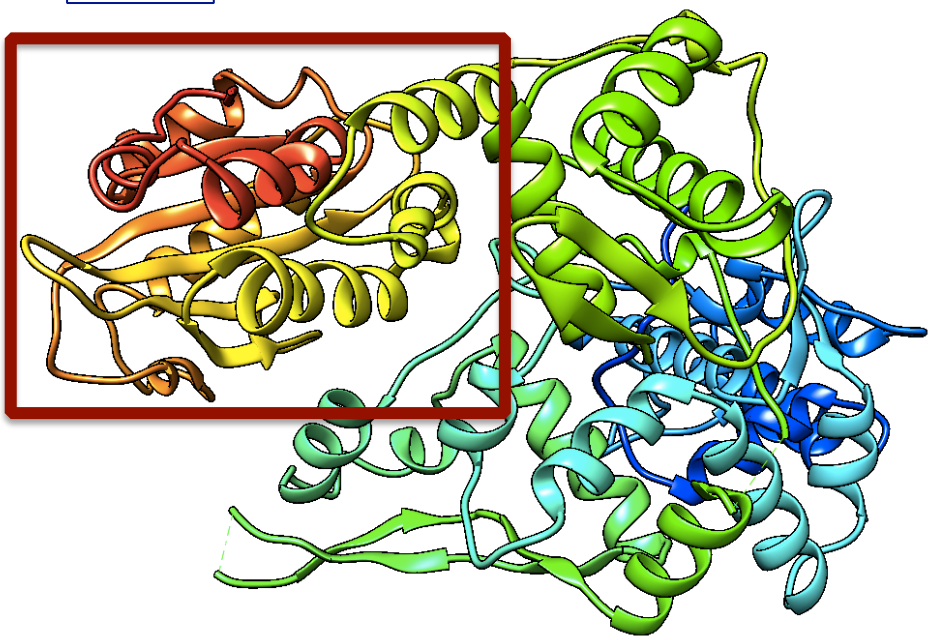
1HW2

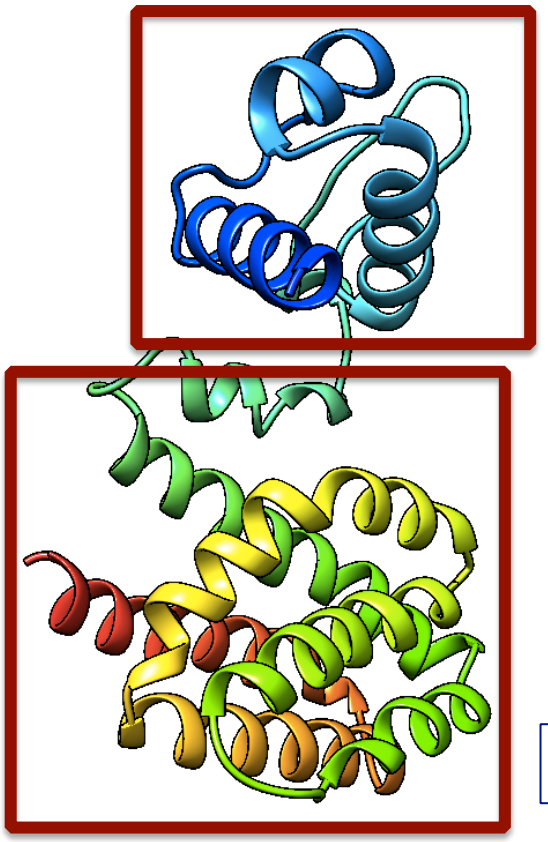
Colour Scheme:

N’ **C’**

Marco Punta

1FOK





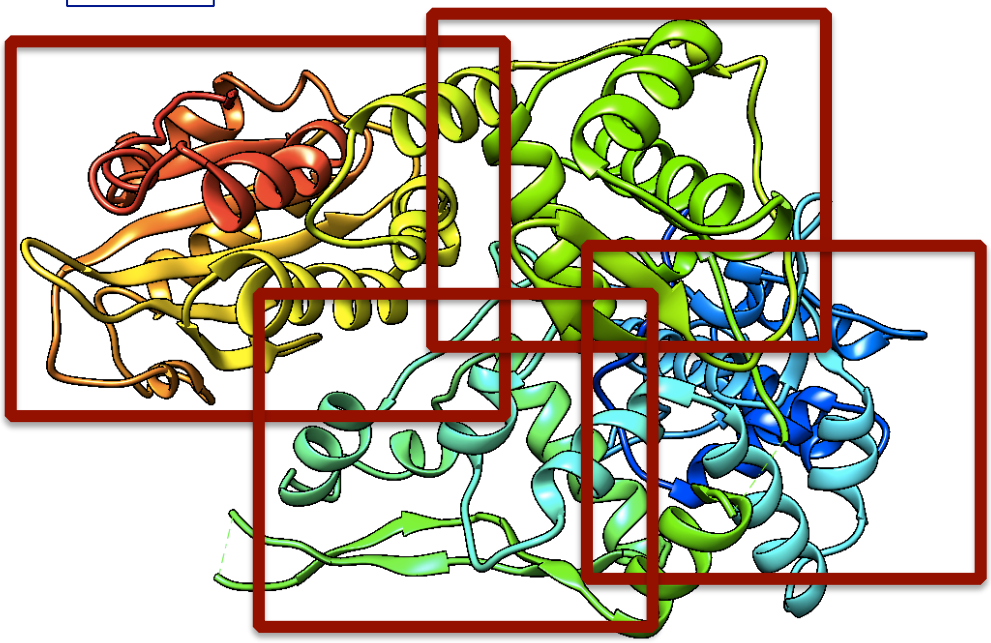
1HW2

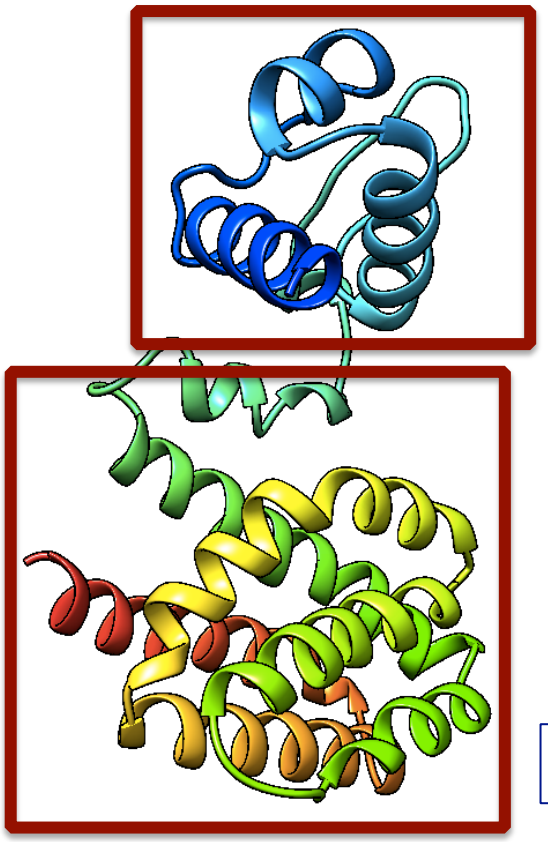
Colour Scheme:

N’ **C’**

Marco Punta

1FOK





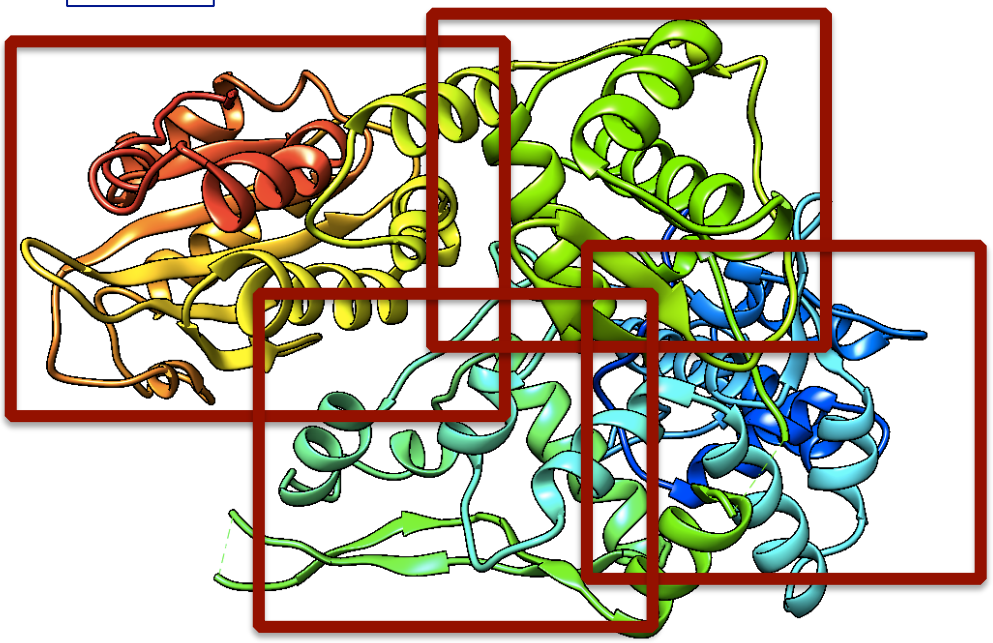
1HW2

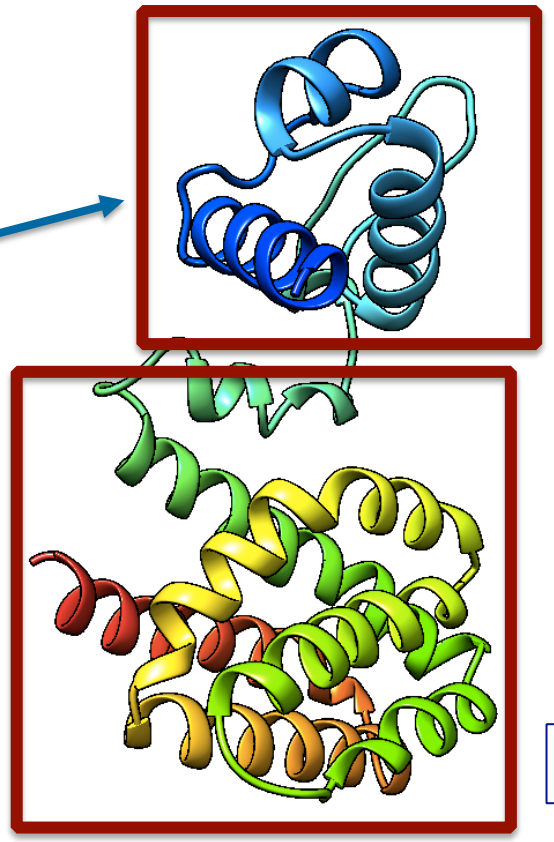
Colour Scheme:

N’ **C’**

Marco Punta

1FOK



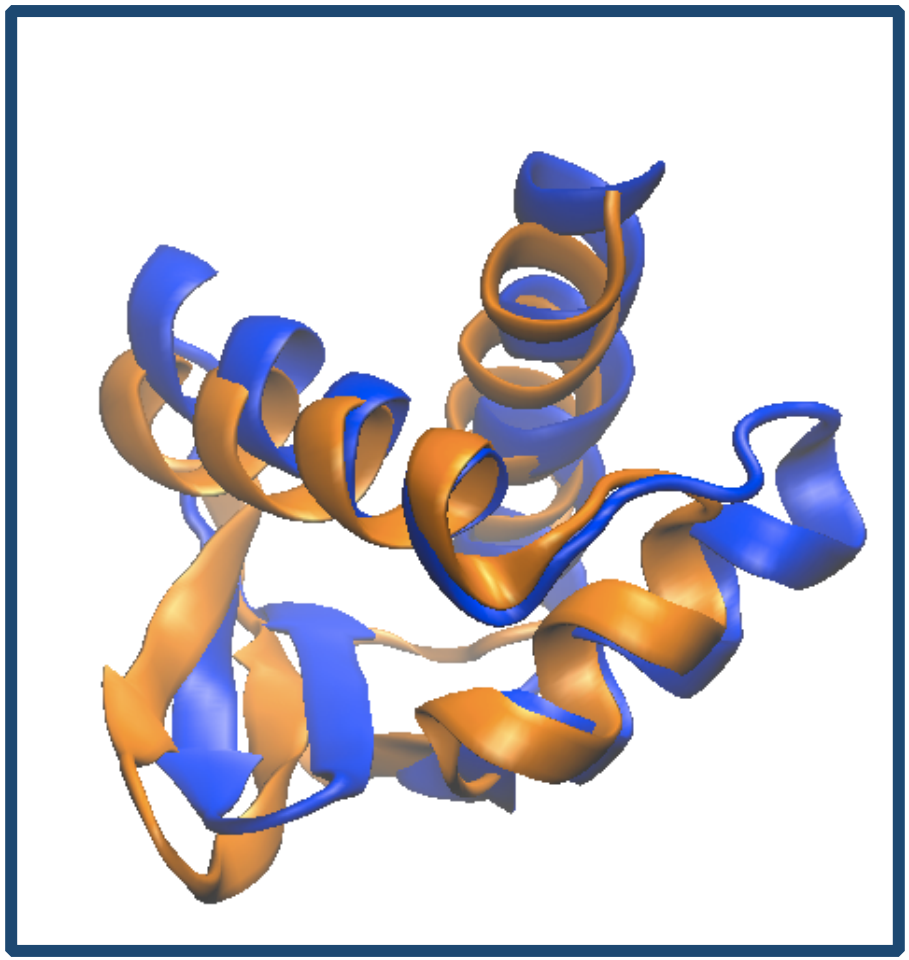


1HW2

Colour Scheme:

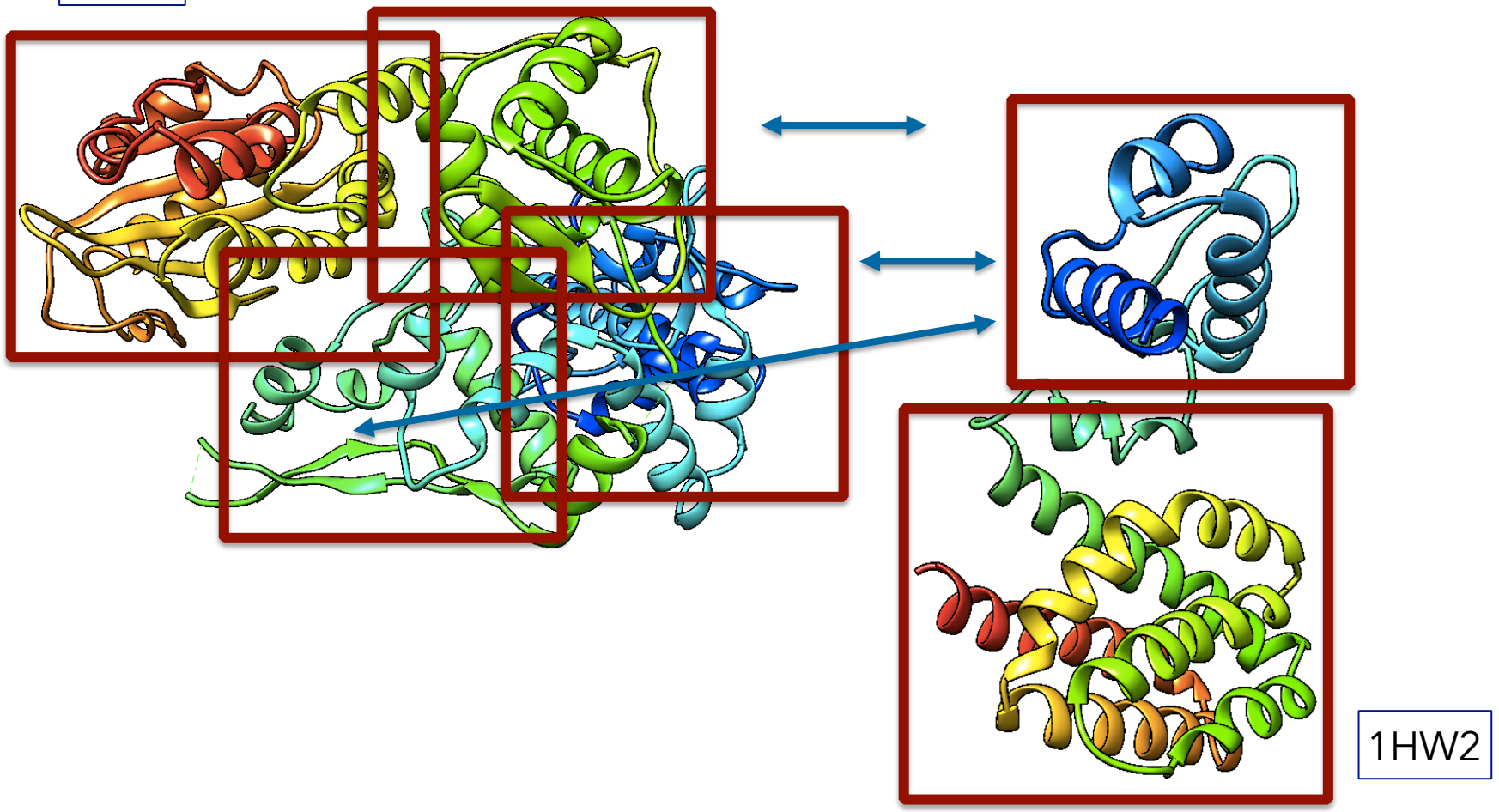
N’ **C’**

Winged helix domain (WHD)

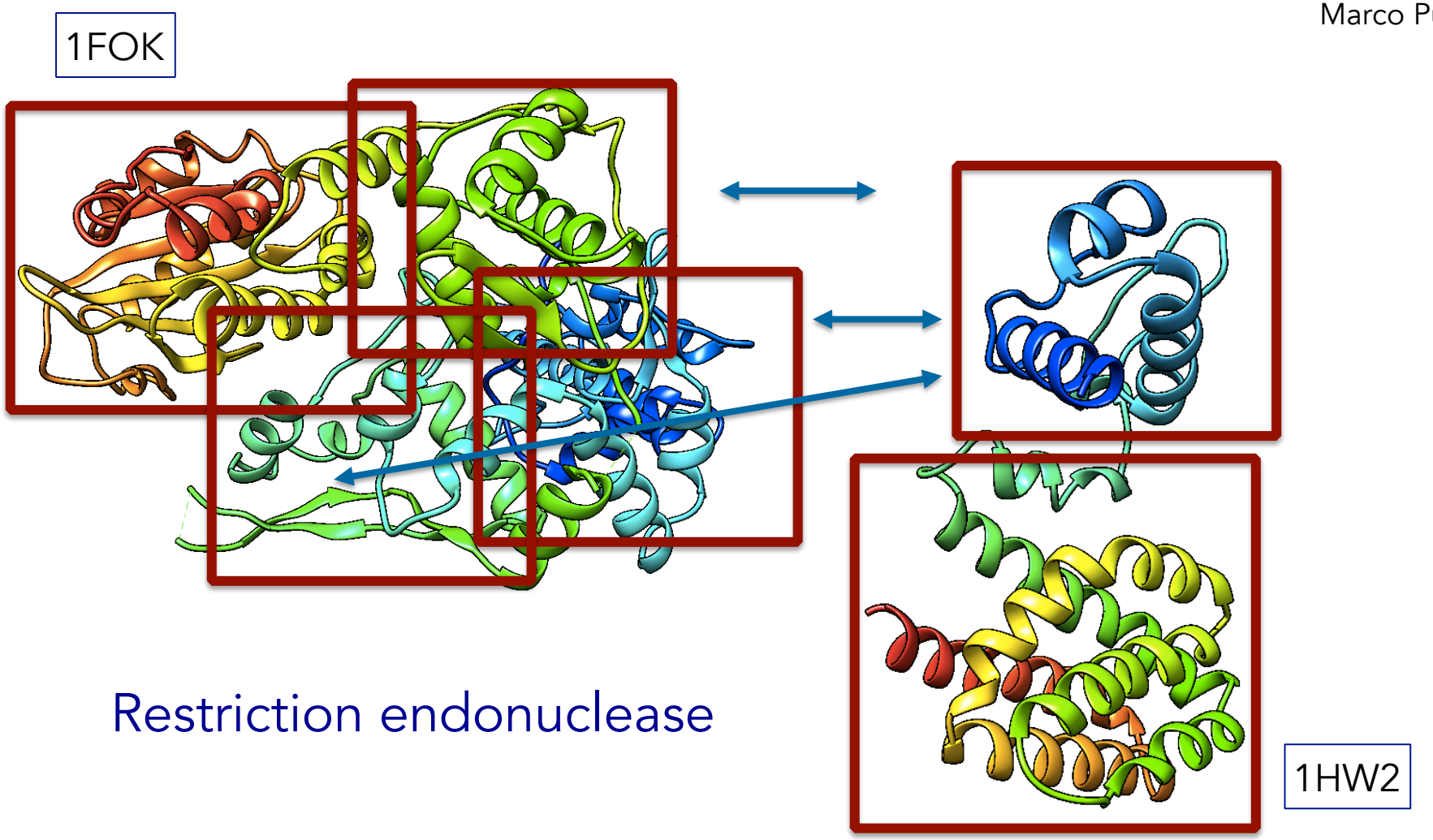


Z-score = 4.0
%id = 8%
RMSD = 2.7Å

Marco Punta

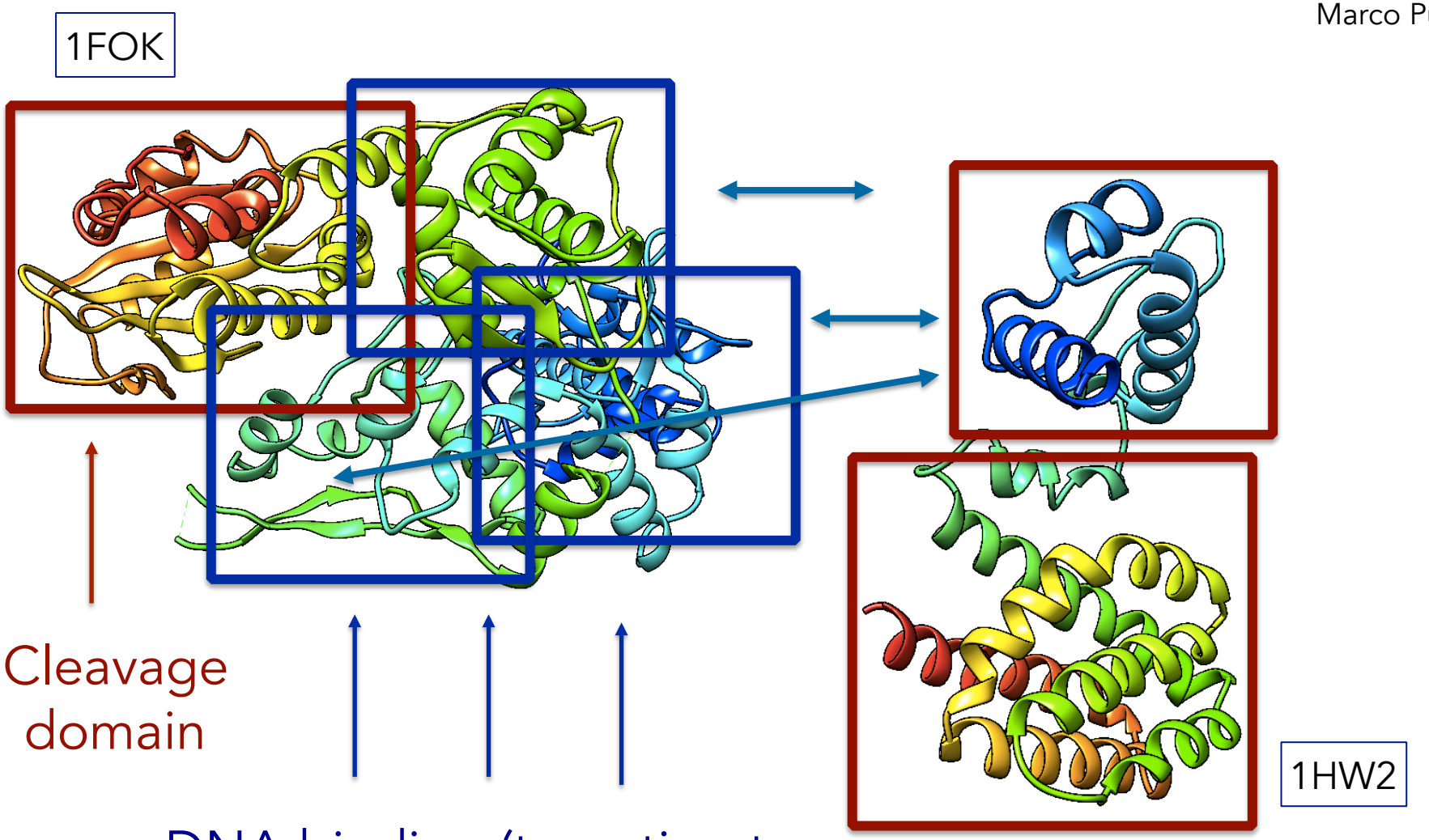


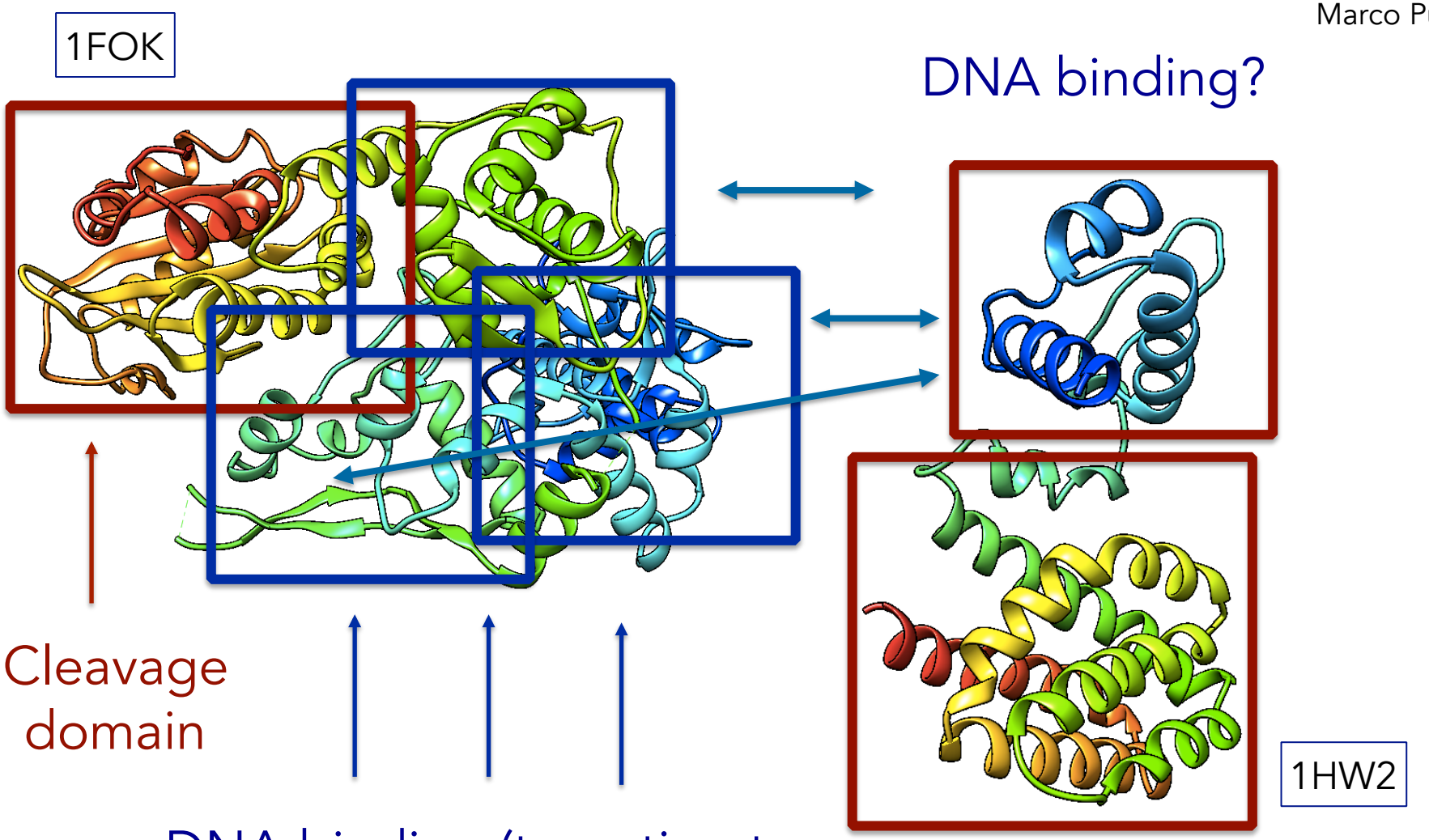
Colour Scheme:



Restriction endonuclease

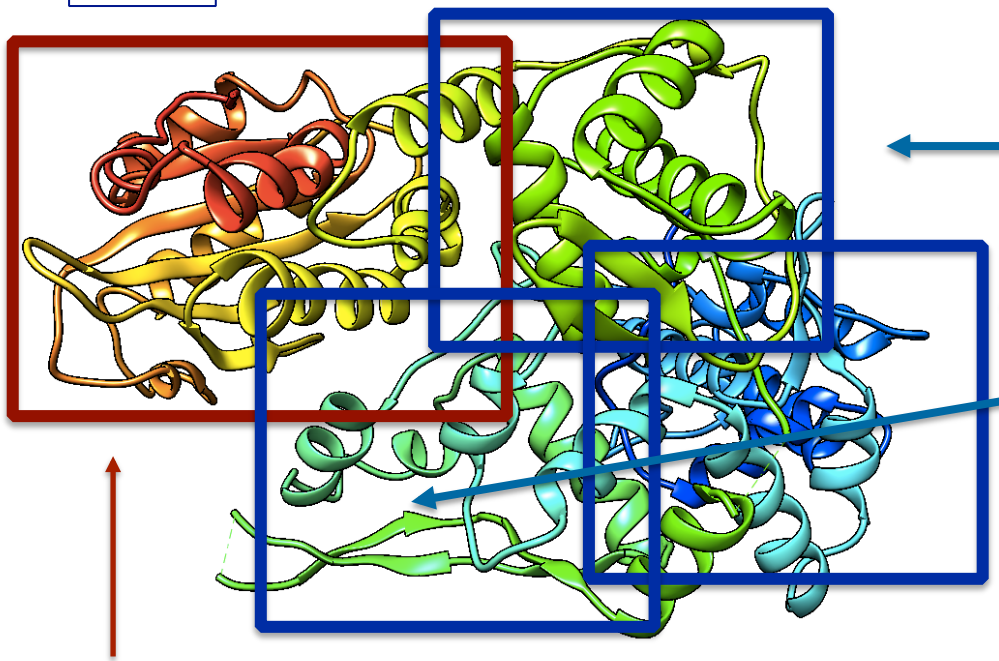
Marco Punta


1FOK
1HW2
Cleavage domain
DNA binding (targeting to a specific DNA sequence)

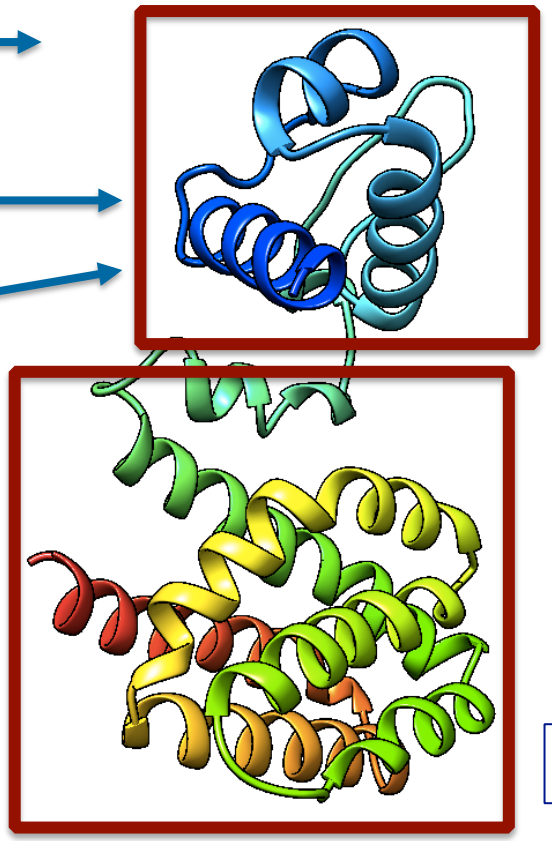

1FOK
DNA binding?
Cleavage domain
DNA binding (targeting to a specific DNA sequence)
1HW2

Marco Punta

1FOK



DNA binding?



1HW2

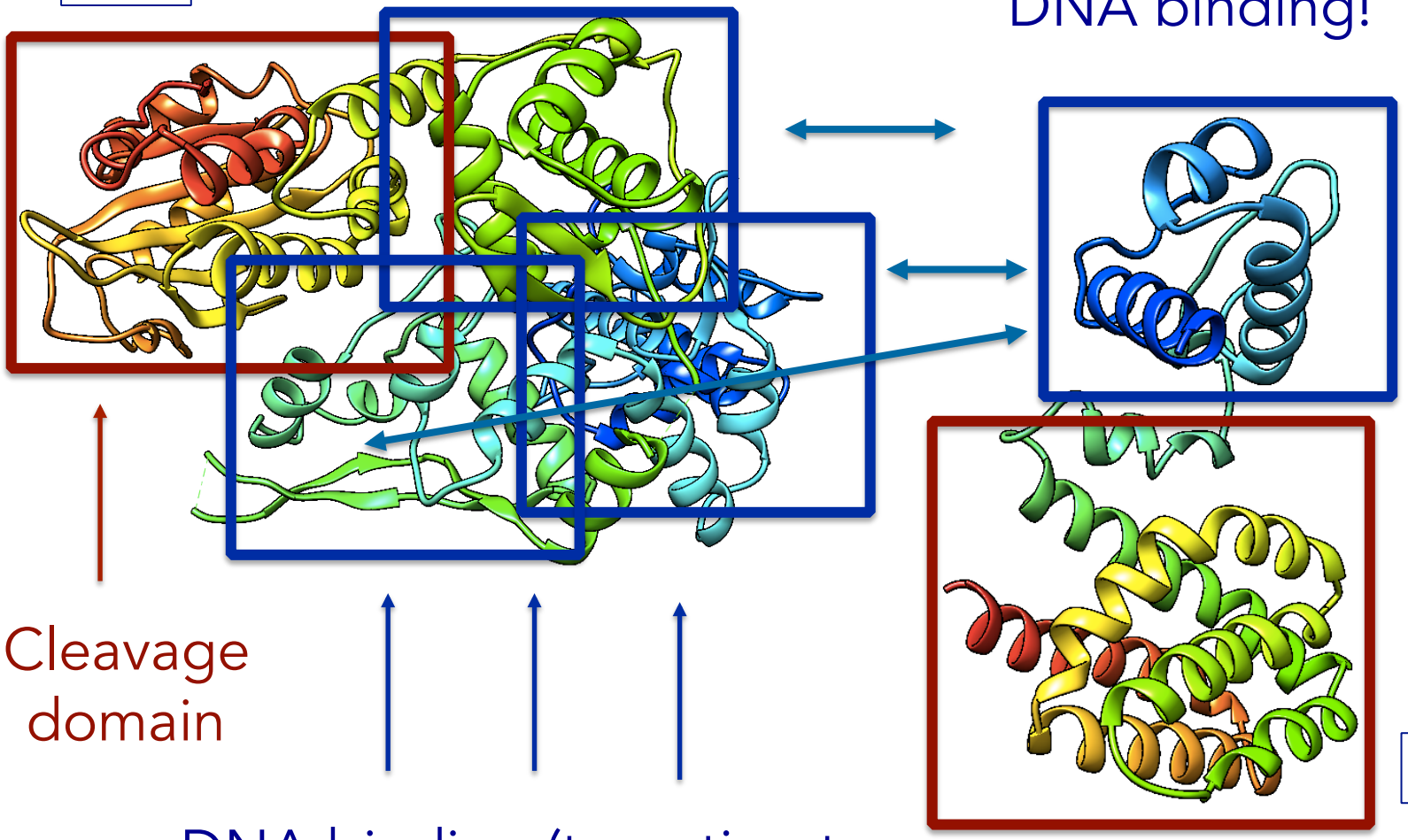
Cleavage domain

DNA binding (targeting to a specific DNA sequence)

?

1FOK

DNA binding!



Cleavage domain

DNA binding (targeting to a specific DNA sequence)

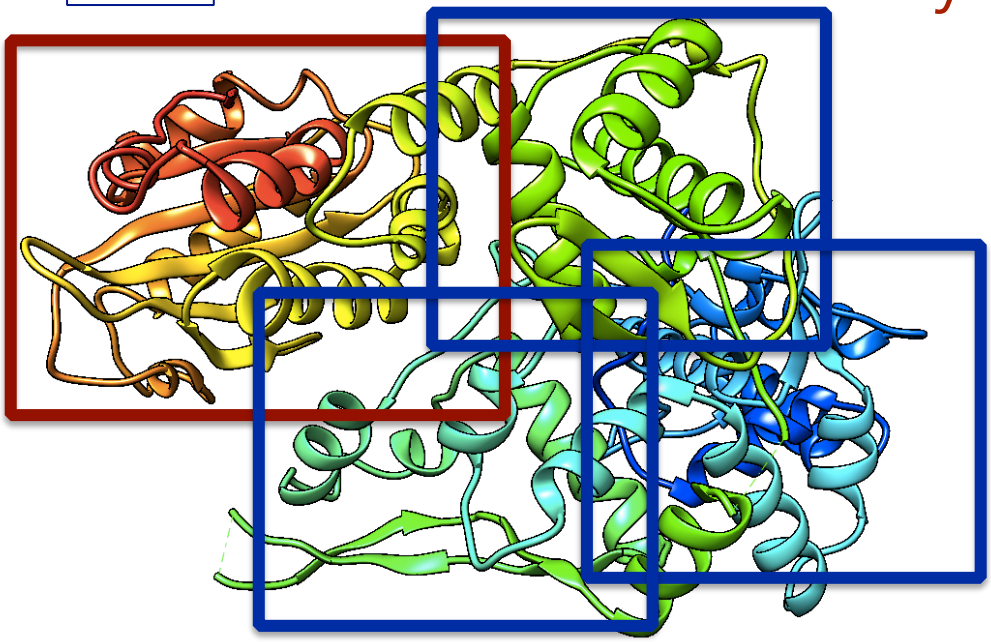
1HW2

acyl-CoA binding domain controls affinity

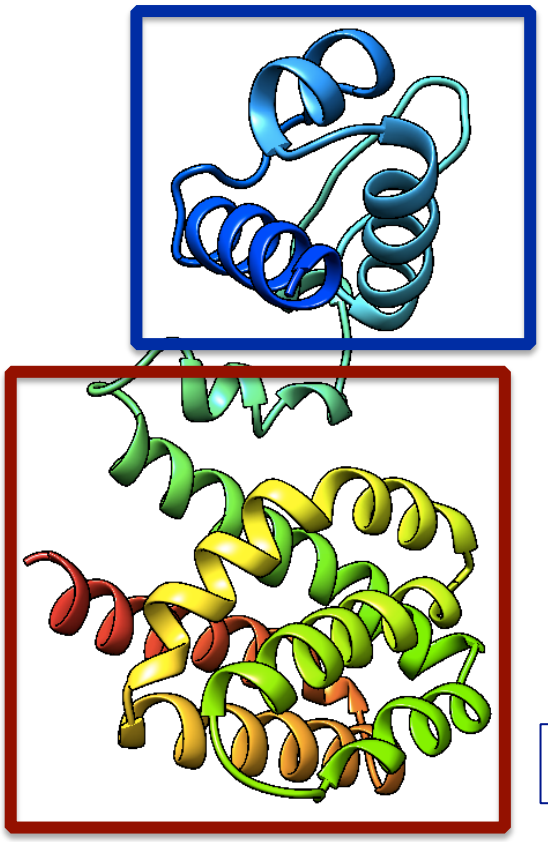
Marco Punta

1FOK

"syntactical change"



Restriction endonuclease

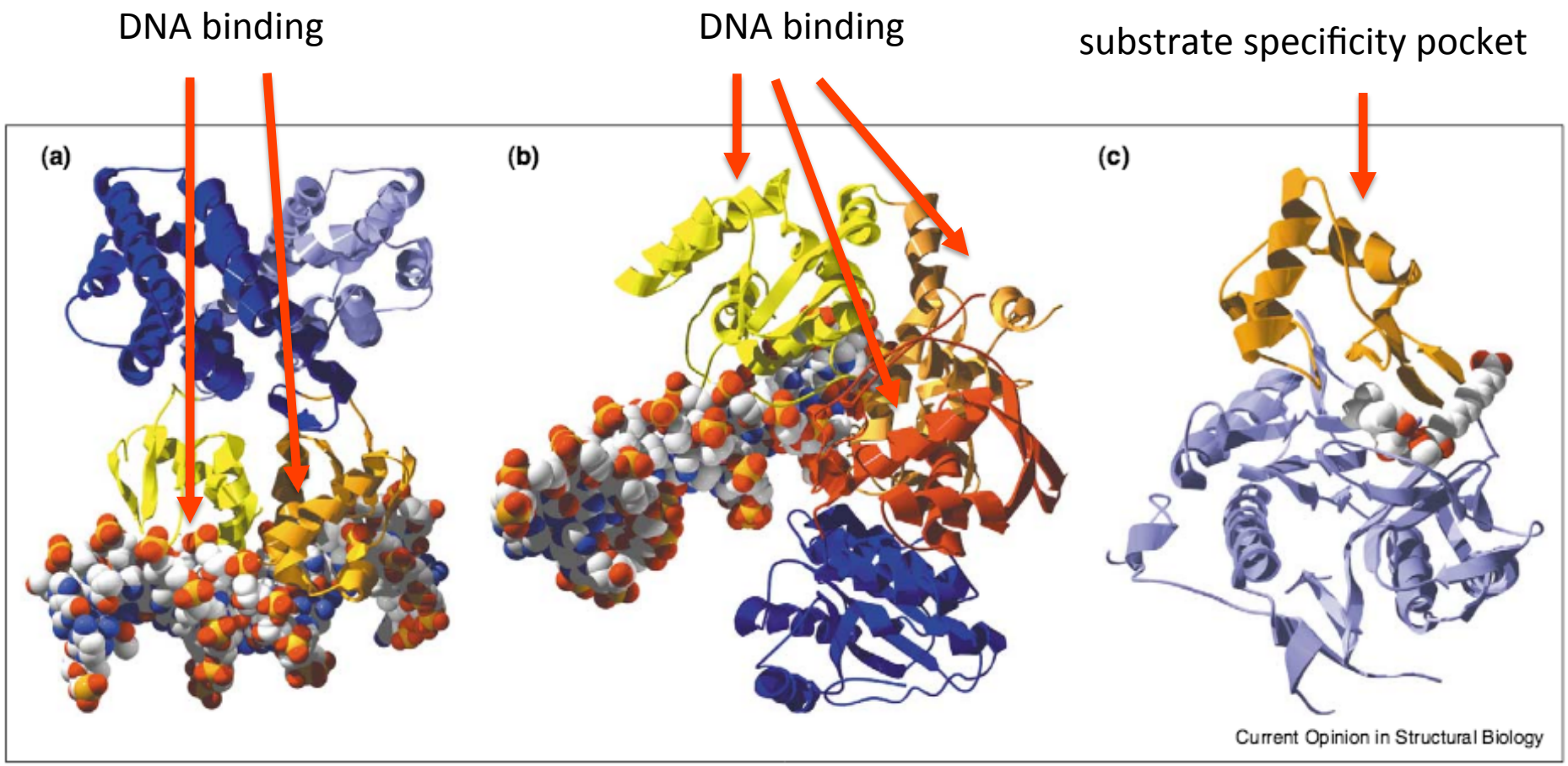


1HW2

Transcription factor

Semantic change

Marco Punta



Transcription factor

Restriction endonuclease

Human methionine aminopeptidase 2

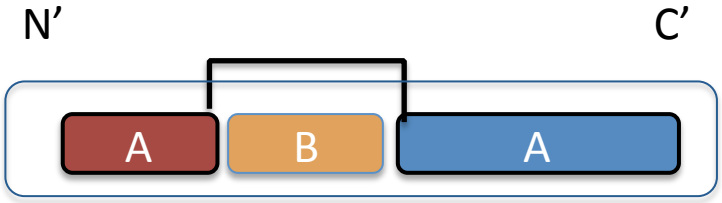
Vogel et al. Curr. Opin. Struct. Biol. 14 (2004)

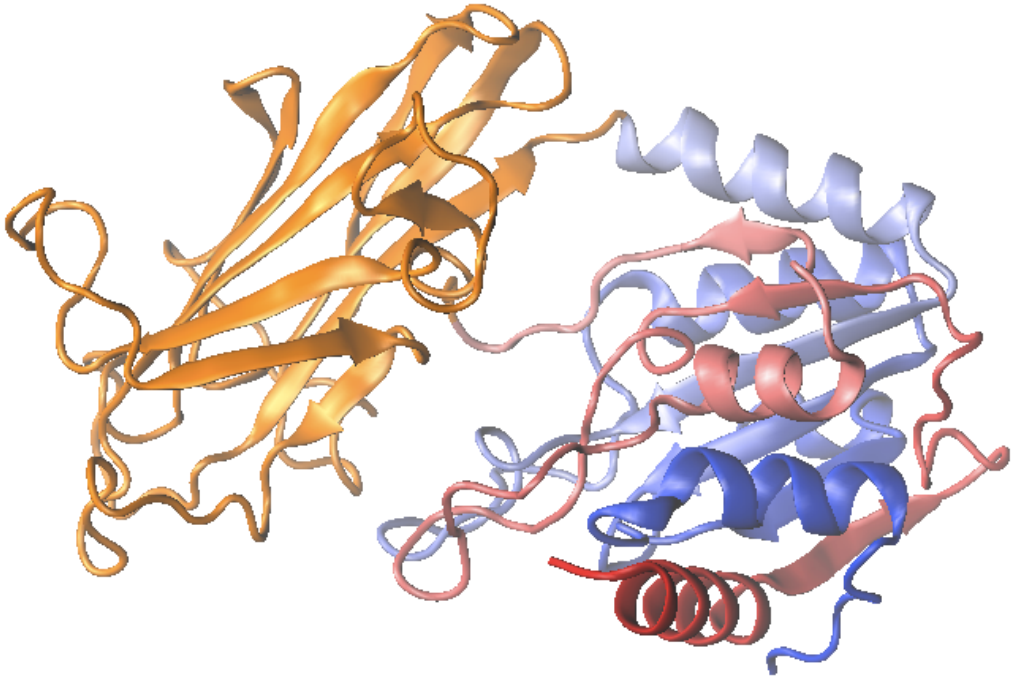
"syntactical change"

DNA sequence recognised

- Restriction endonuclease
  5'-GGATG-3'
- Transcription factor
  5'-TGGNNNNNCCA-3'

"Nested" domains





3ABZ-truncated

Alignments

Download Graphics Sort by: E value

unnamed protein product

Sequence ID: lcl|Query_22995 **Length:** 762 **Number of Matches:** 3

Range 1: 281 to 383 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
41.2 bits(95)	1e-07	Compositional matrix adjust.	30/110(27%)	54/110(49%)	11/110(10%)

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

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

Range 2: 161 to 260 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
38.1 bits(87)	1e-06	Compositional matrix adjust.	26/108(24%)	53/108(49%)	8/108(7%)

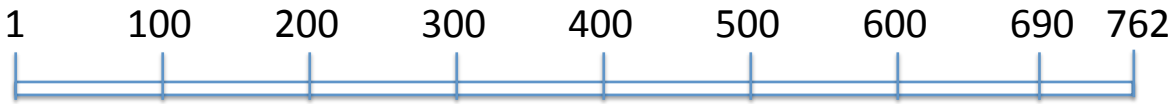
```
Query  472  DTLKKVRIFADCEAGLLVELVLKLQPQVYSPGDYICKKGDIGREMYIIKEGKLAVVADDG  531
            D L K +     +   + ++V  +  + Y  G YI K+G+ G  ++++ EG+L V   +
Sbjct  161  DALNKNQFLKRLDPQQIKDMVECMYGRNYQQGSYIIKQGEPGNHIFVLAEGRLEVFQGEK  220

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

Range 3: 593 to 649 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

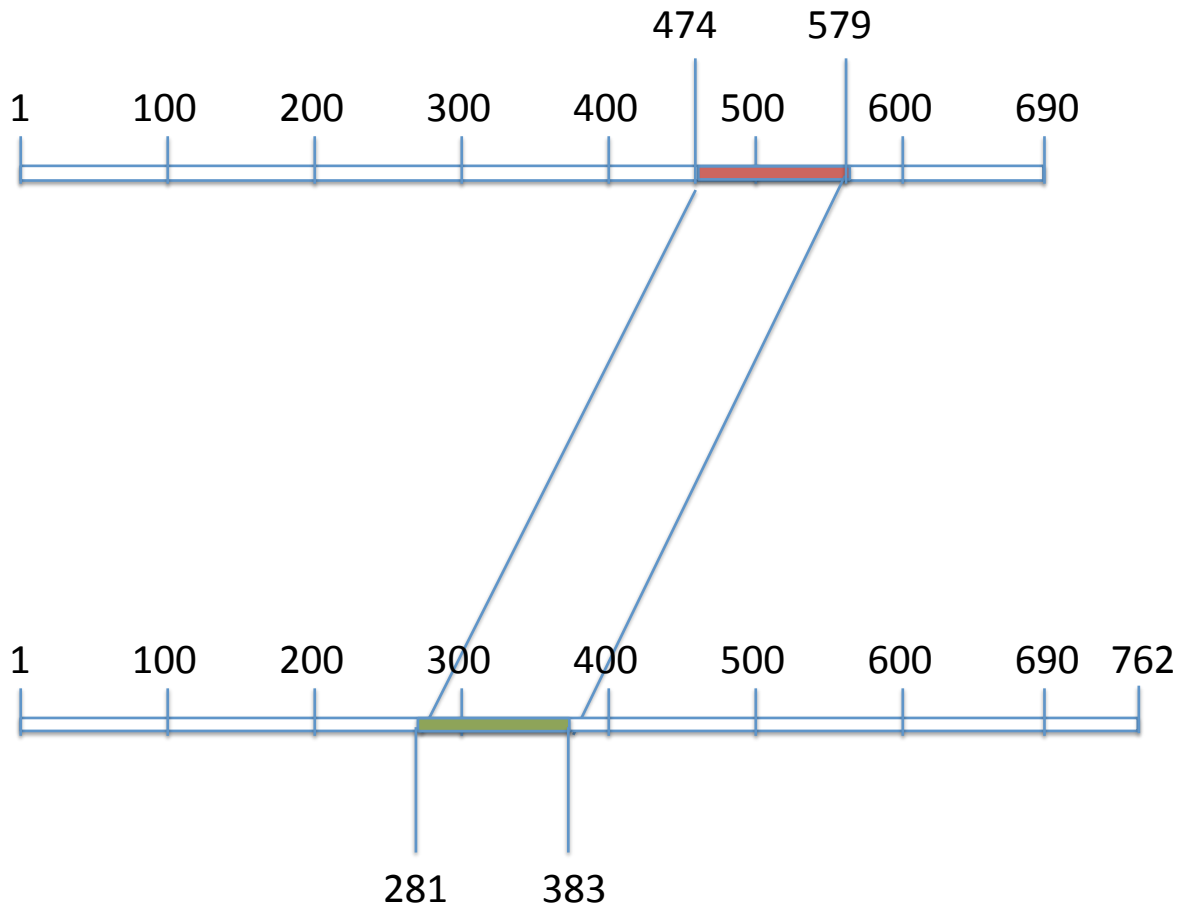
Score	Expect	Method	Identities	Positives	Gaps
22.3 bits(46)	0.081	Compositional matrix adjust.	17/58(29%)	27/58(46%)	7/58(12%)

```
Query  317  VFYSISKAIGFGNDTWVY---PDINDPEFGRLARKYVYSL-YWS--TLTLTTIGETPP  368
            V +  +K IG G  TW +   P+   PE   L + + +S+ +WS   L    +   PP
Sbjct  593  VDFGFAKKIGSGQKTWTFCGTPEYVAPEV-ILNKGHDFSVDFWSLGILVYELLTGNPP  649
```

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

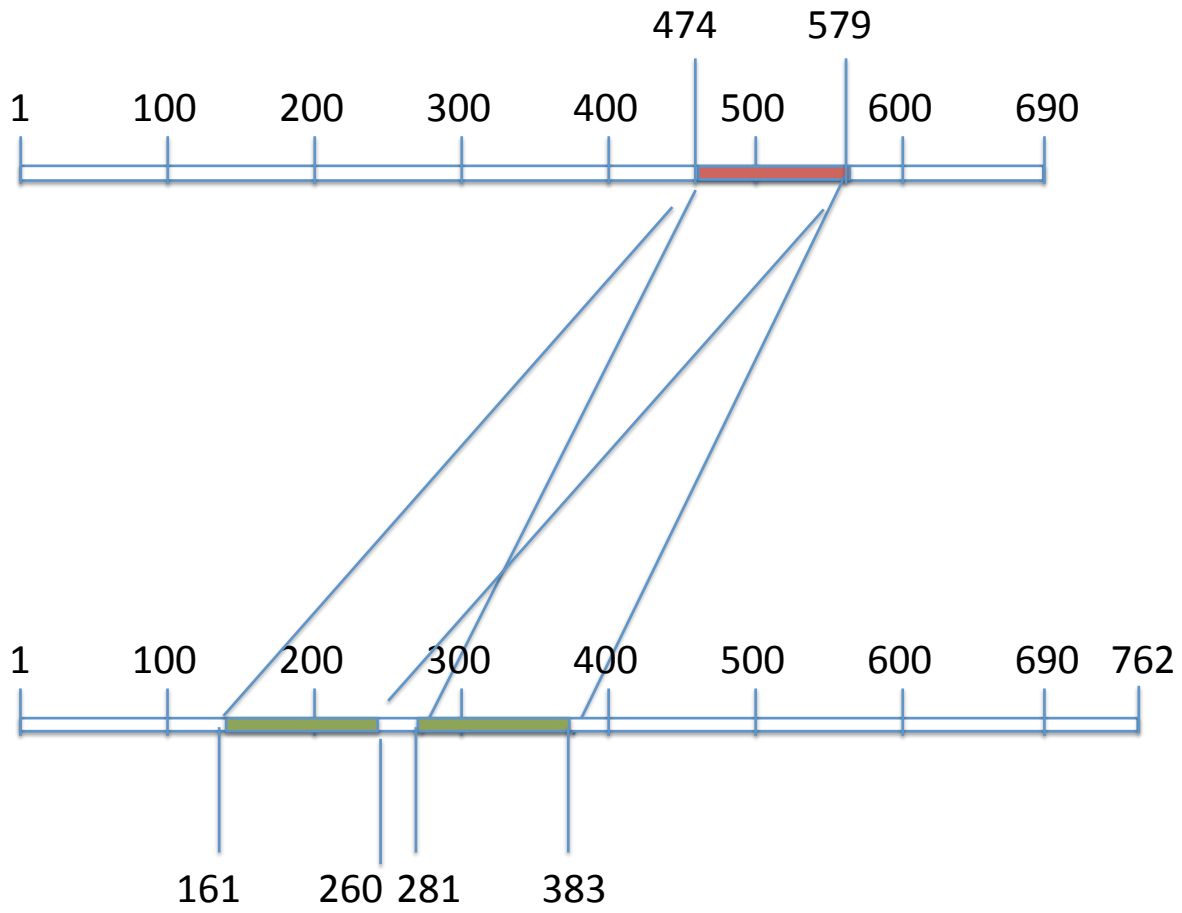


cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN) Marco Punta



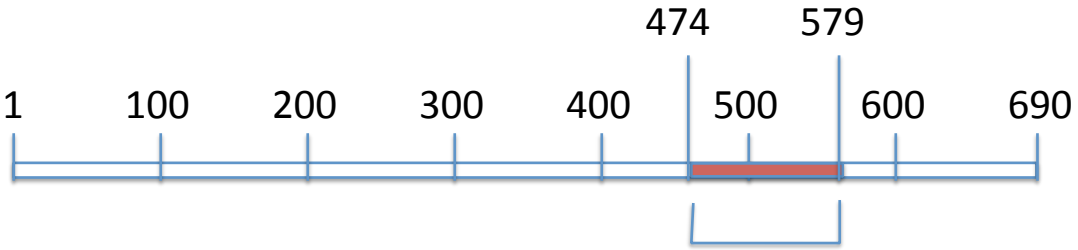
EMBO Workshop, Budapest, 2016

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

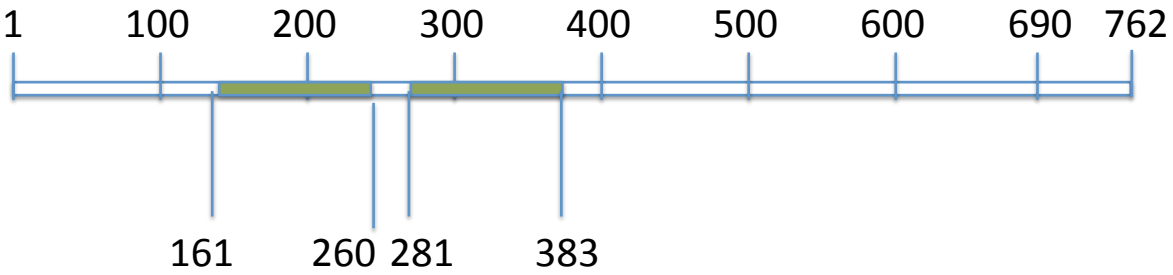


EMBO Workshop, Budapest, 2016

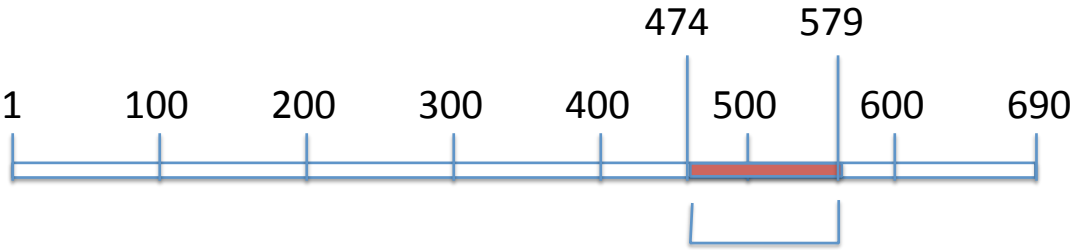
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN) Marco Punta



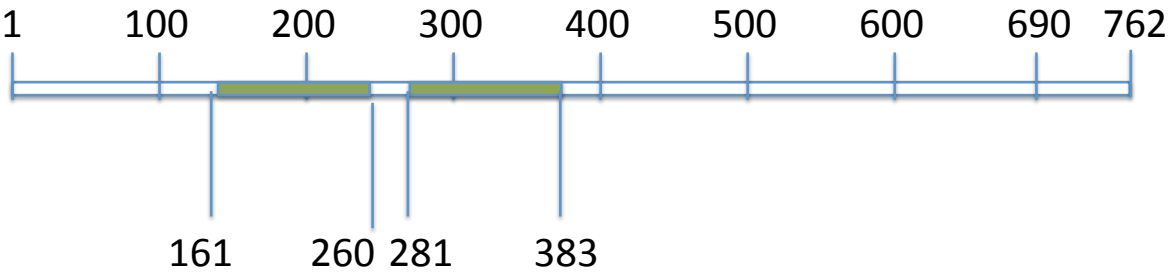
cNMP binding domain



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

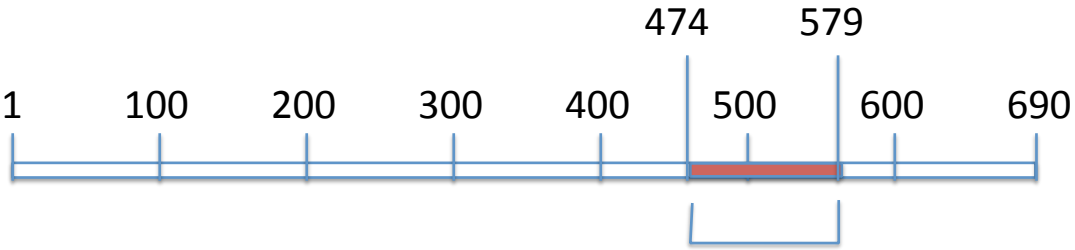


cNMP binding domain

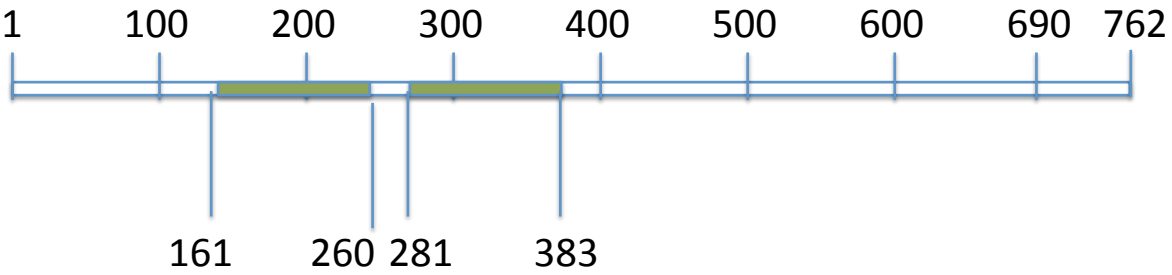


cNMP binding domains?

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

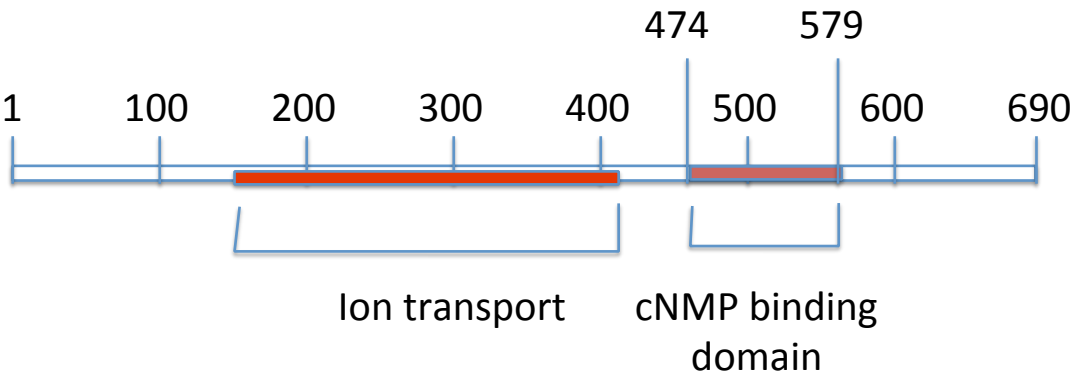


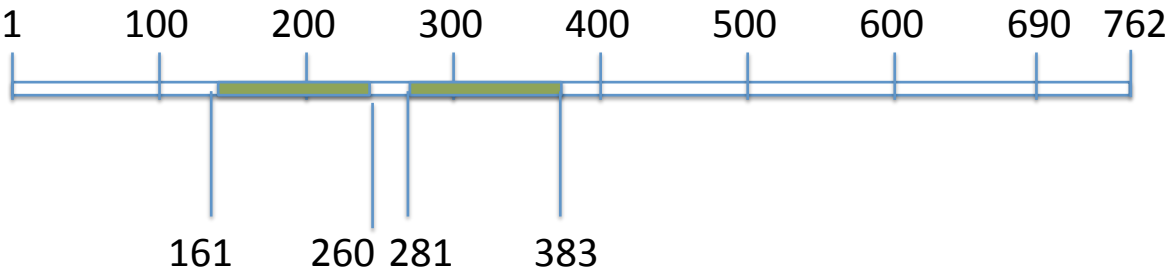
cNMP binding domain



cNMP binding domains?

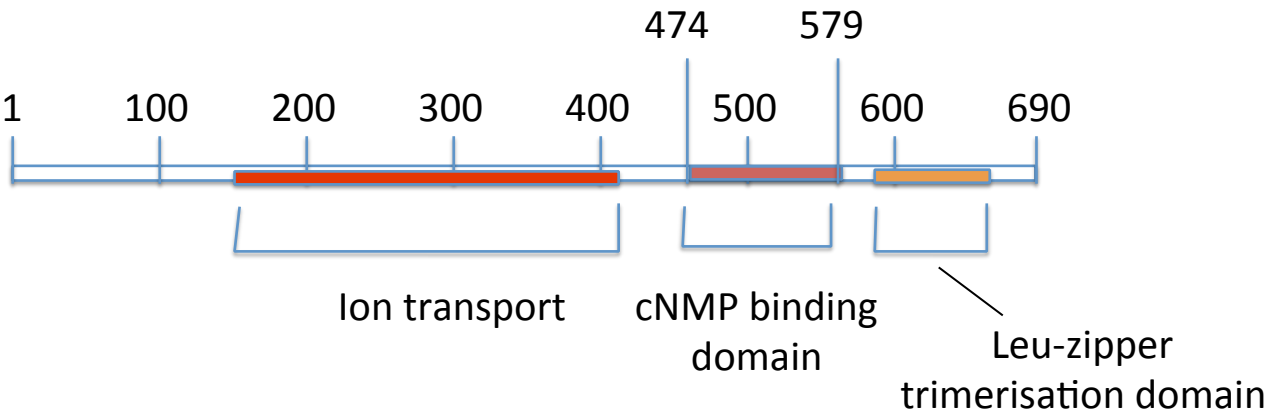
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN) Marco Punta

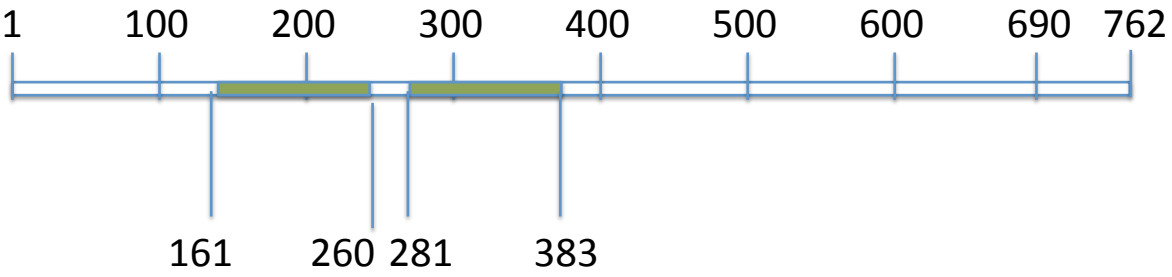




cGMP binding domains?

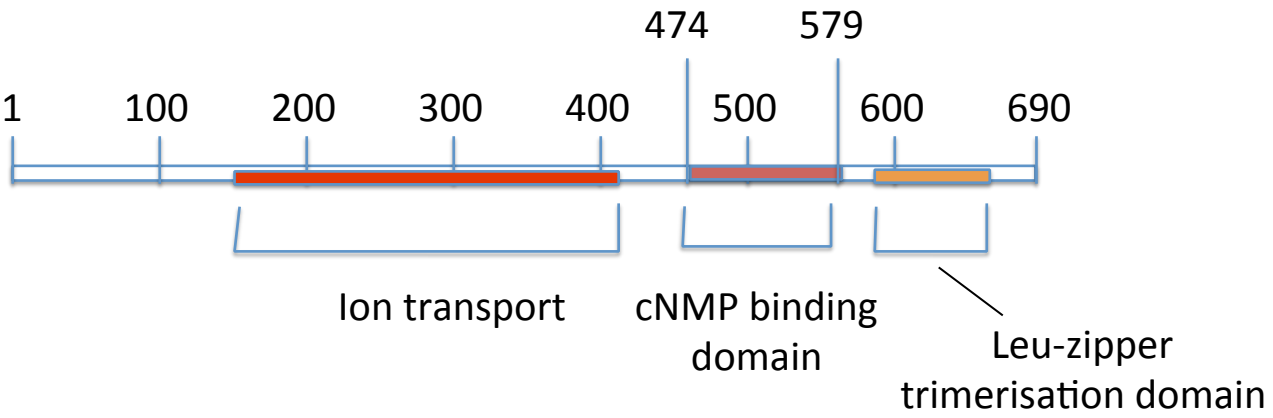
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN) Marco Punta

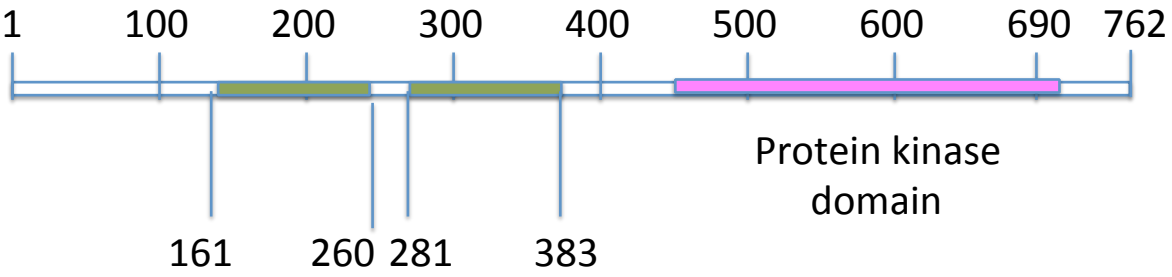




cGMP binding domains?

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





cGMP binding domains?

Marco Punta

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

EMBO Workshop, Budapest, 2016

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.

Function annotation transfer by homology

Homologous proteins can share a number of functional features, however:

- functional drift can lead to radically different functions

- while functional similarity correlates with function, no similarity threshold is safe for transfer

- Proteins may have multiple functions

- if more than one functional domain is present annotation transfer can be attempted only between domains that are homologous and NOT for the full-length protein function (but still not 'safe')

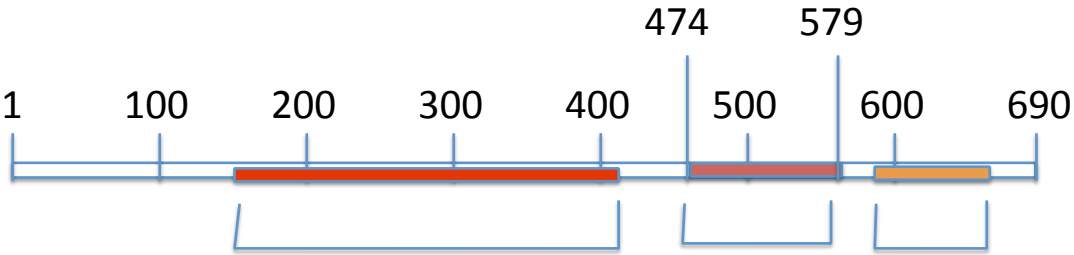
Globular proteins?

Water-soluble, globe-like shape.

Non globular proteins?

Membrane proteins, fibrous proteins, disordered proteins

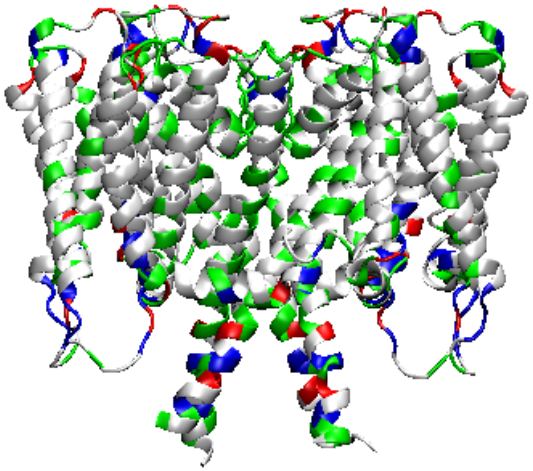
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN) Marco Punta



Ion transport

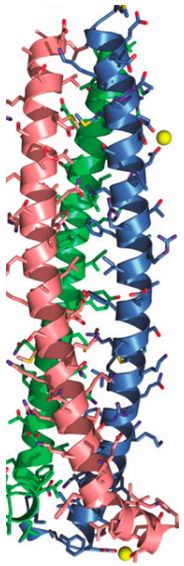
cNMP binding domain

Leu-zipper trimerisation domain



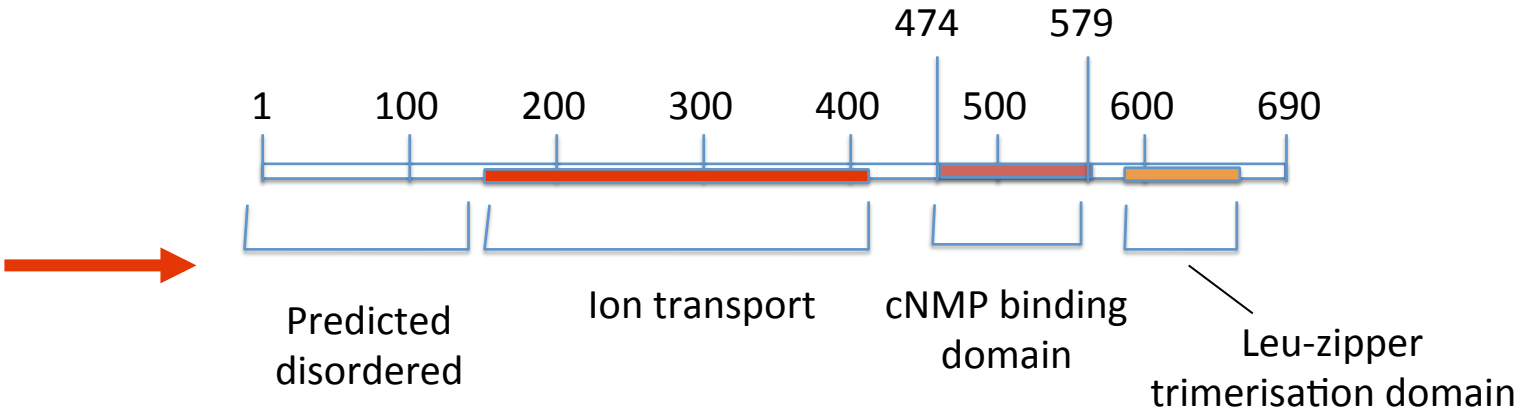
Transmembrane

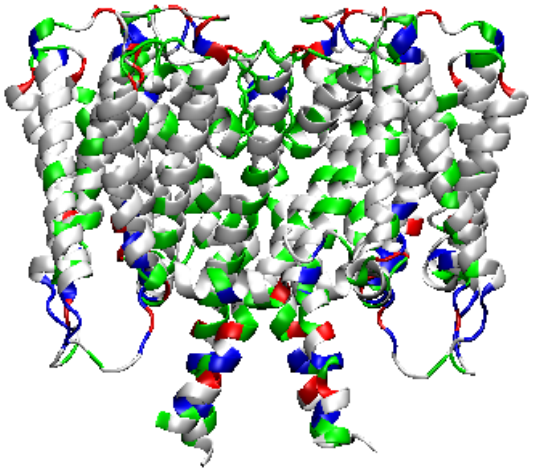
Globular



Coiled-coil

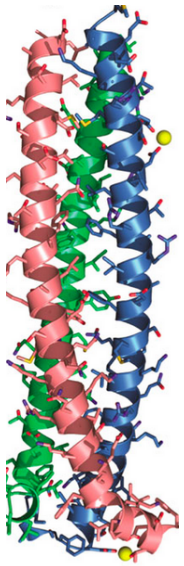
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN) Marco Punta





Transmembrane

Globular



Coiled-coil

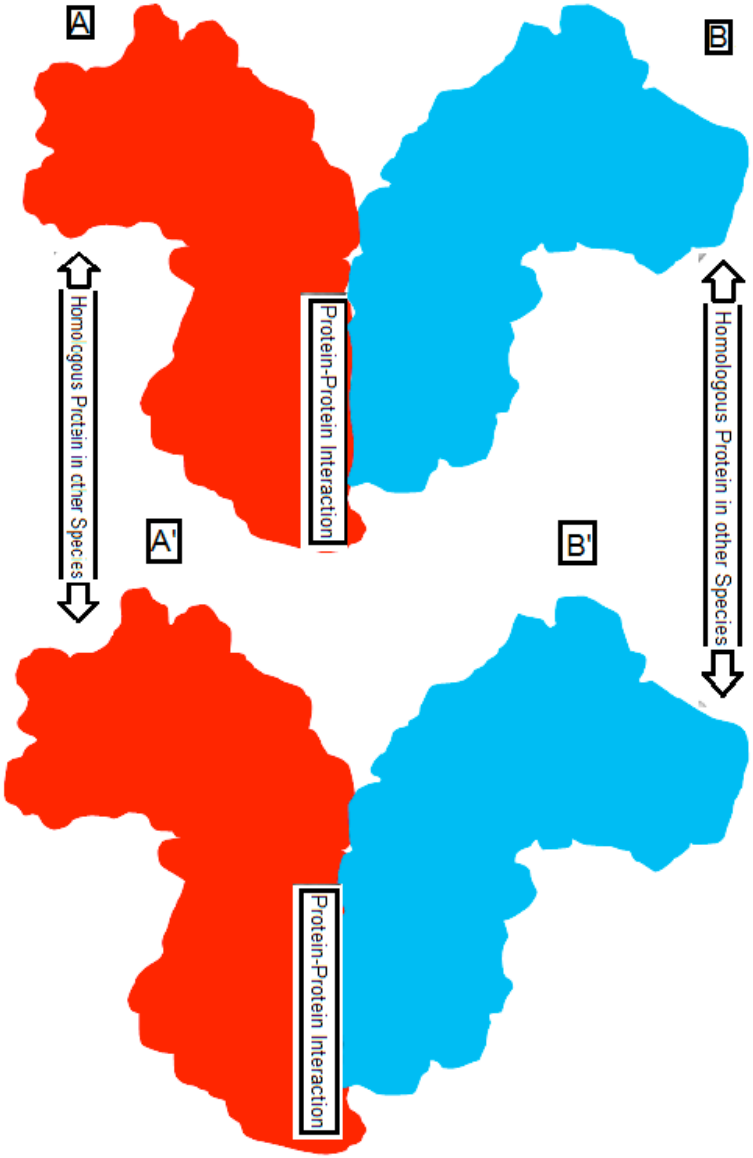
EMBO Workshop, Budapest, 2016

Homology $\overset{\Rightarrow}{\Leftarrow}$ similar sequence?

Homology $\overset{\Rightarrow}{\Leftarrow}$ similar structure?

Homology $\overset{\Rightarrow}{\Leftarrow}$ similar function?

Marco Punta



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

Interologs

Marco Punta

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

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

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

Article | Authors | Metrics | Comments | Related Content

Abstract
Author Summary
Introduction
Results/Discussion
Materials and Methods
Supporting Information
Acknowledgments
Author Contributions
References

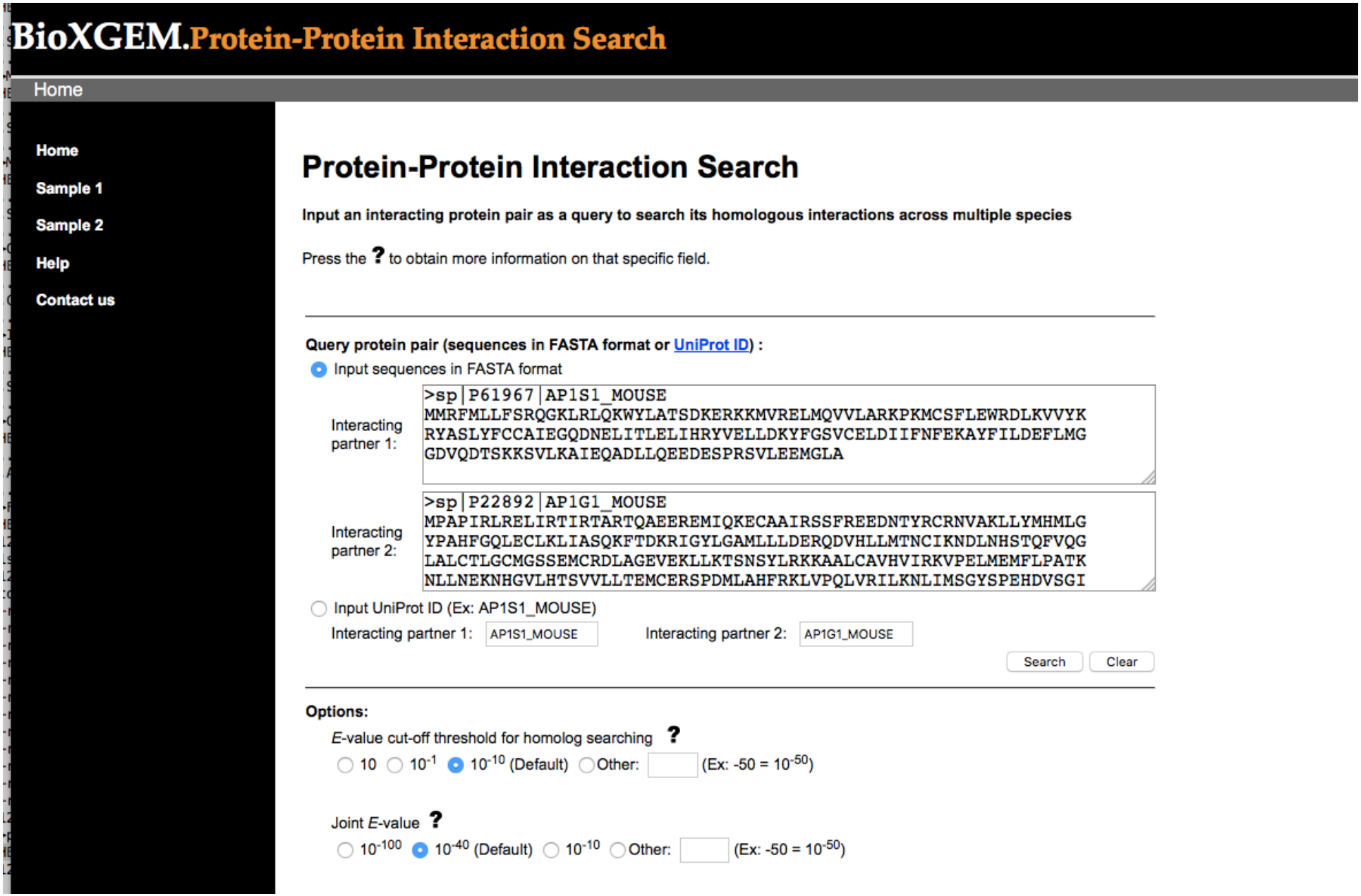
Reader Comments (1)
Media Coverage (0)
Figures

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.

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

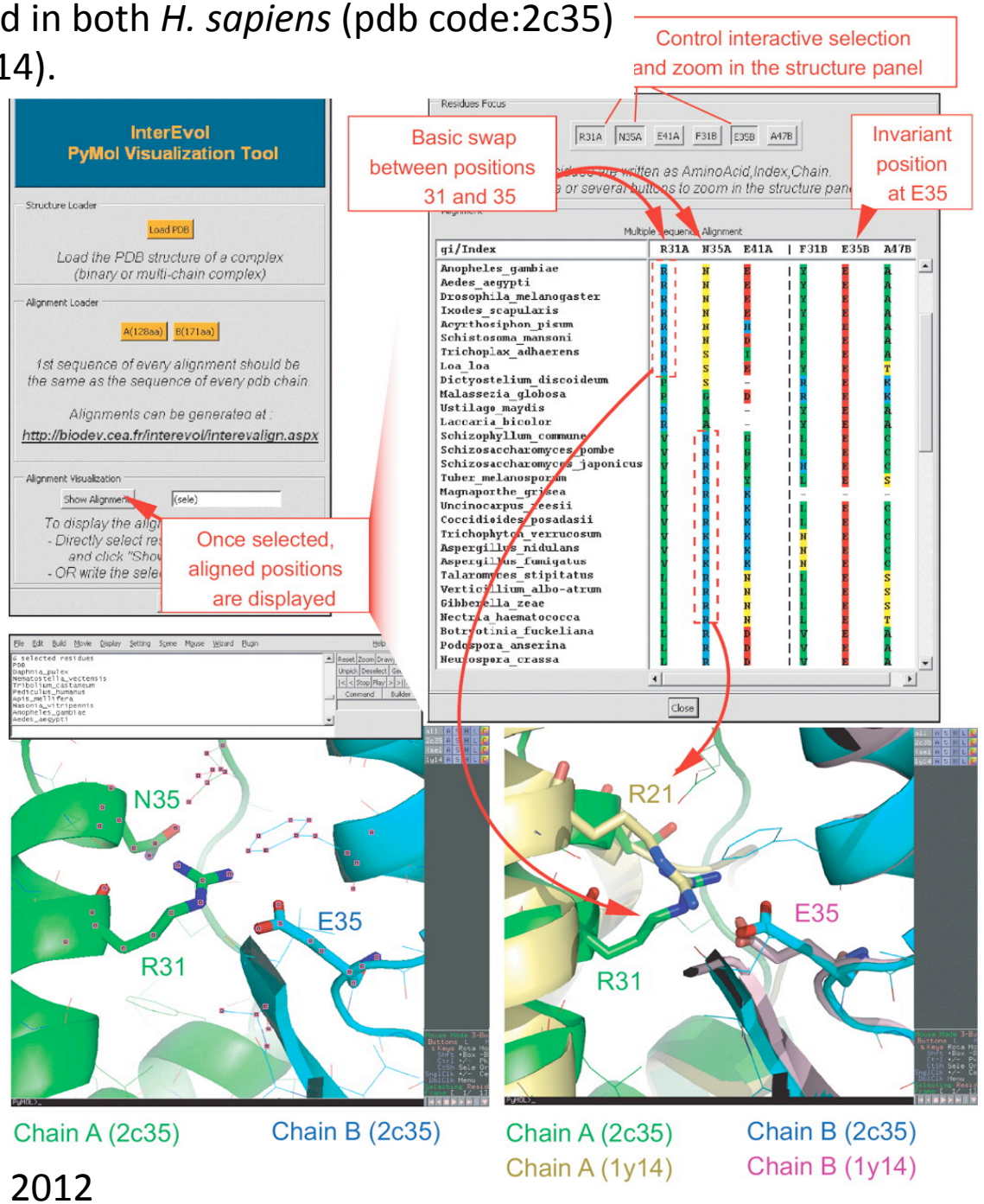
Wiles et al. BMC Syst Biol. 2010



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

Chen et al. NAR 2009

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



Faure et al. Nucleic Acids Res. 2012