Goals

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

Outline

- Homology
- Exercise 1 and 2: homology-based function annotation transfer
- Protein domains
- Protein families
- Protein family databases
- Team exercise: how to build a new (Pfam) protein family

Marco Punta

Homology

Definition:

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

Marco Punta

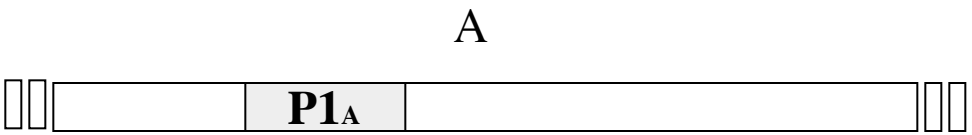
Origins of homology in proteins

Origin of homology in proteins

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

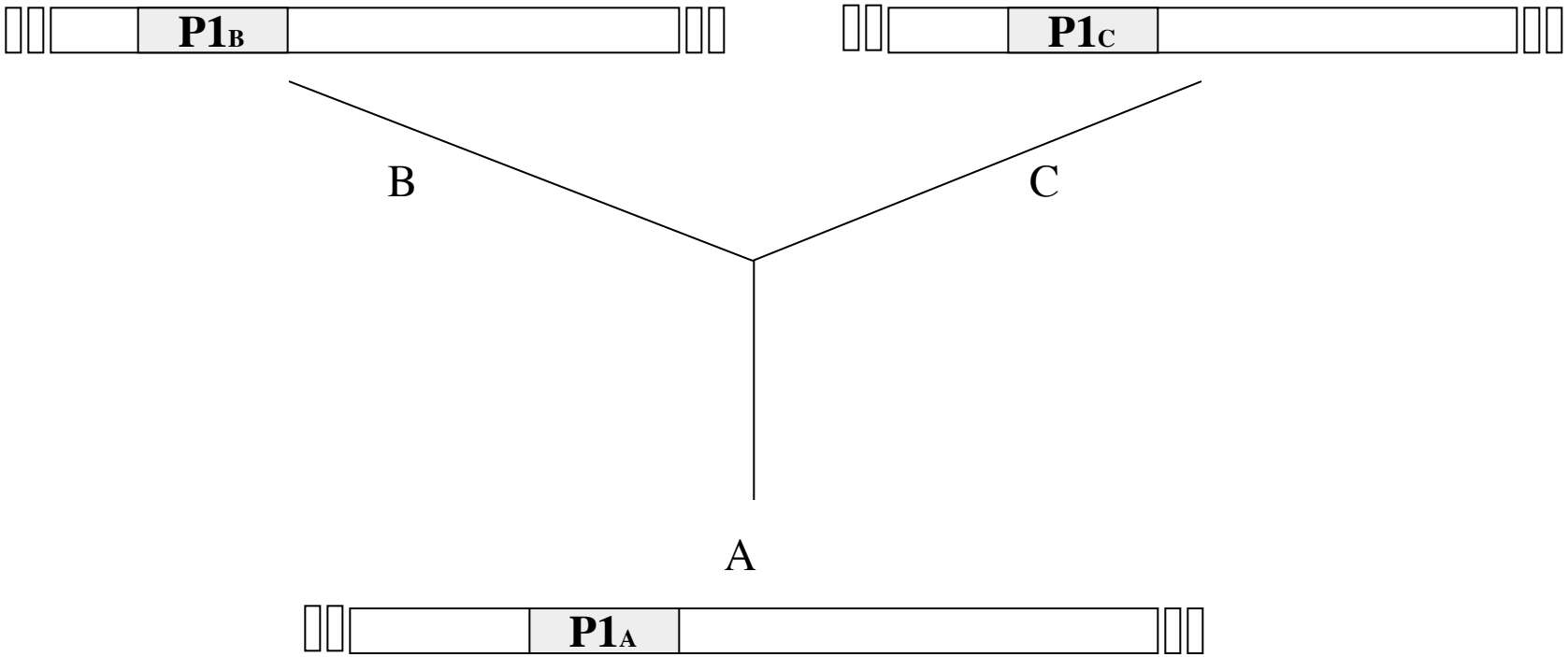
Origin of homology in proteins

Marco Punta

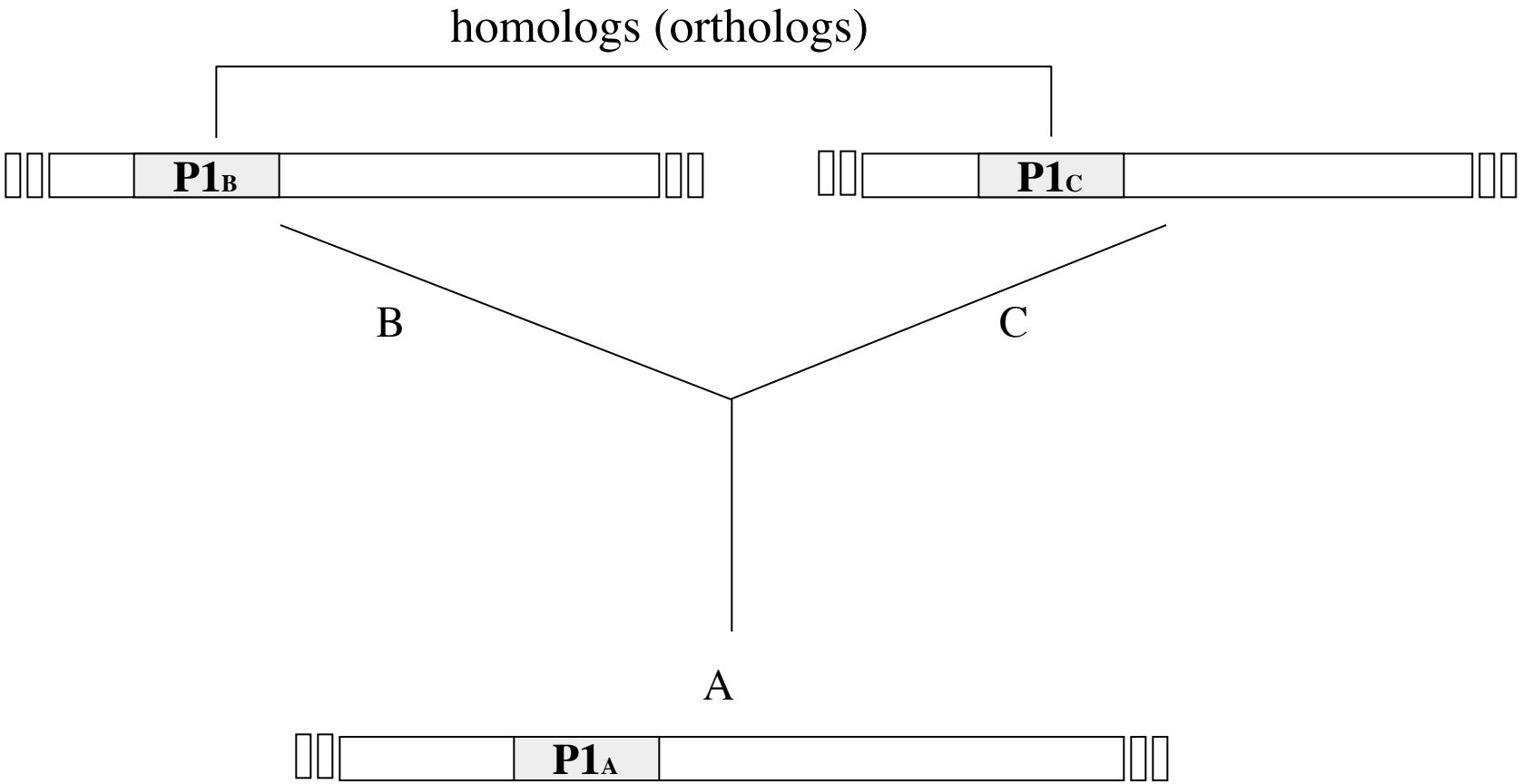


Origin of homology in proteins

Marco Punta

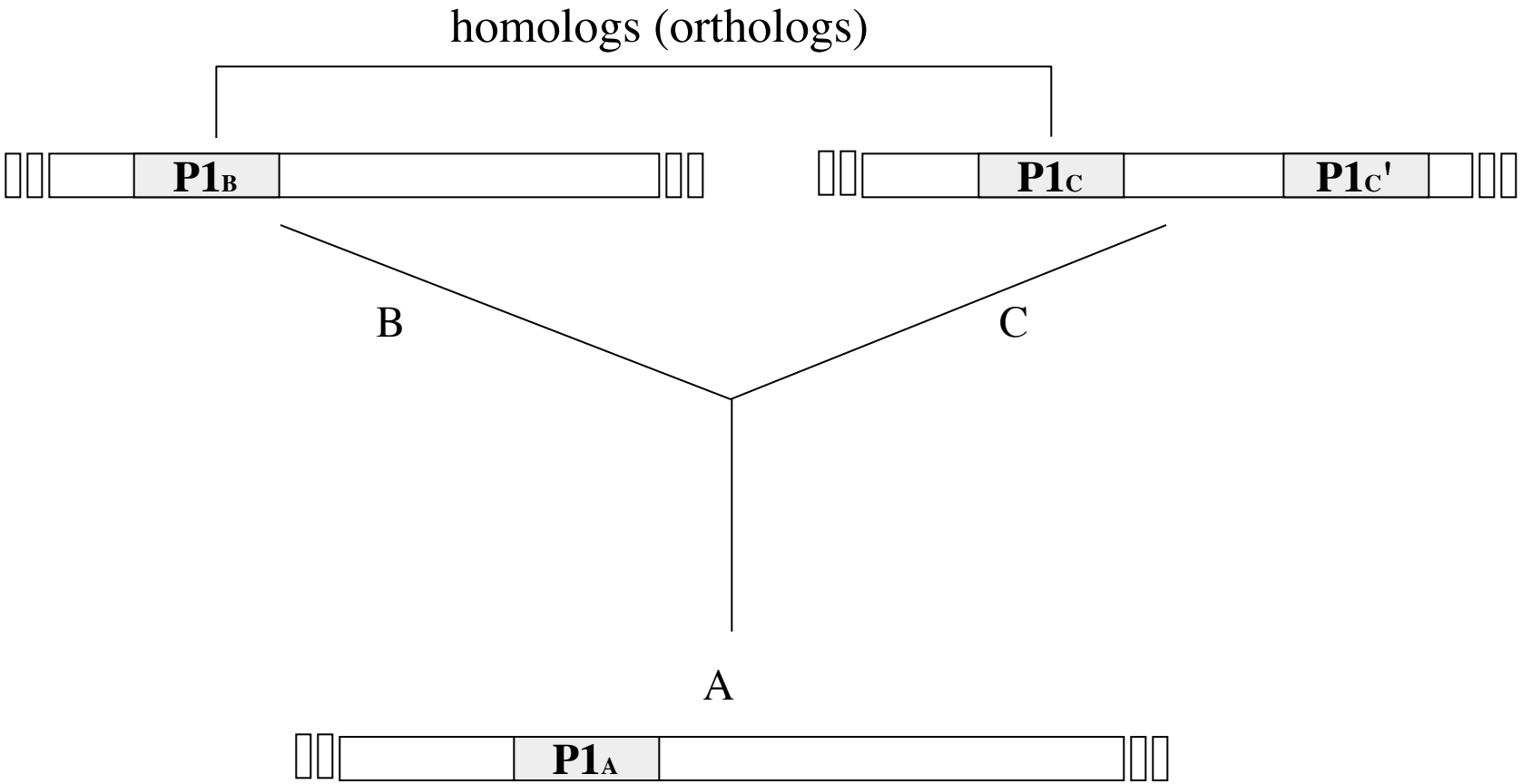


Origin of homology in proteins



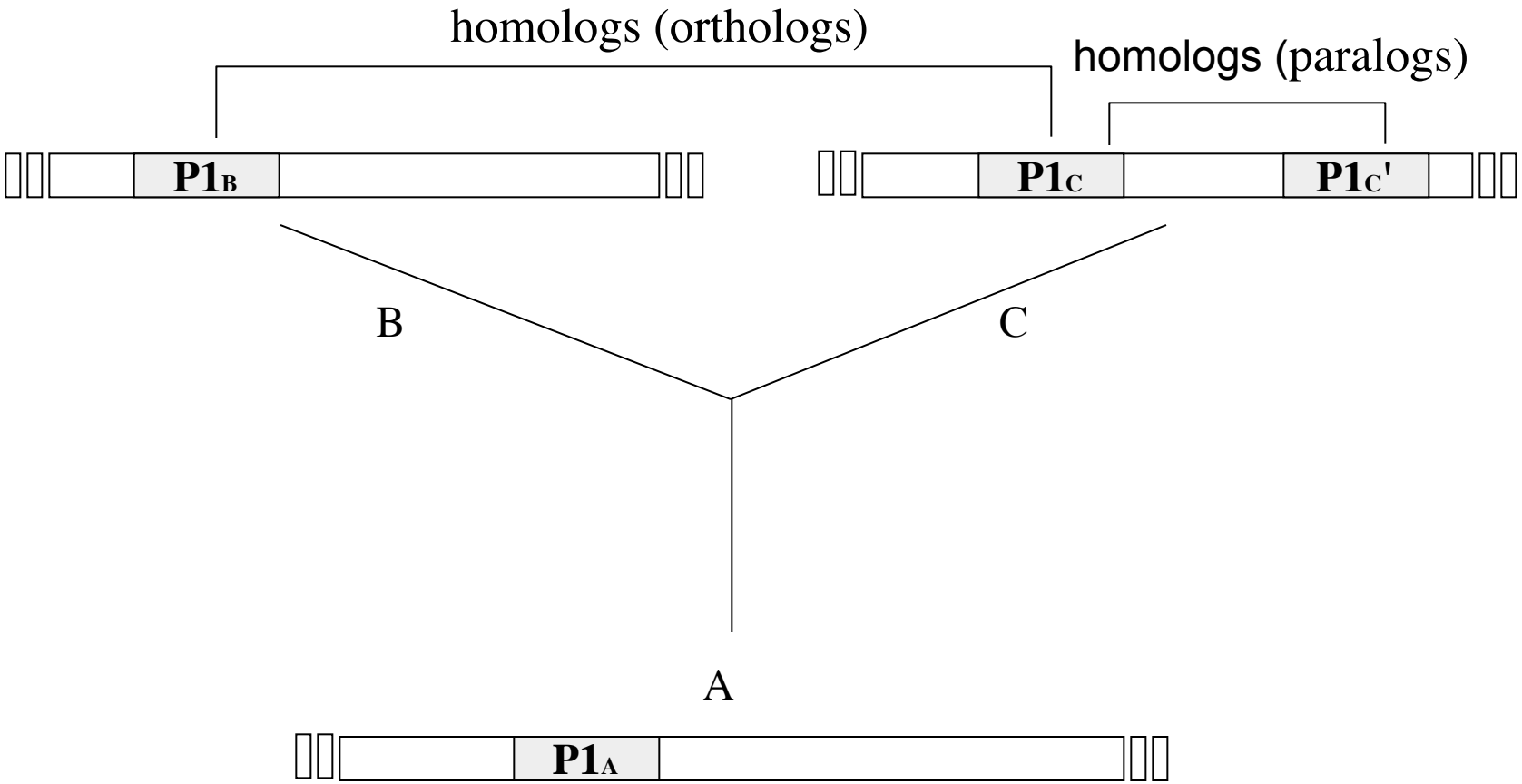
Origin of homology in proteins

Marco Punta



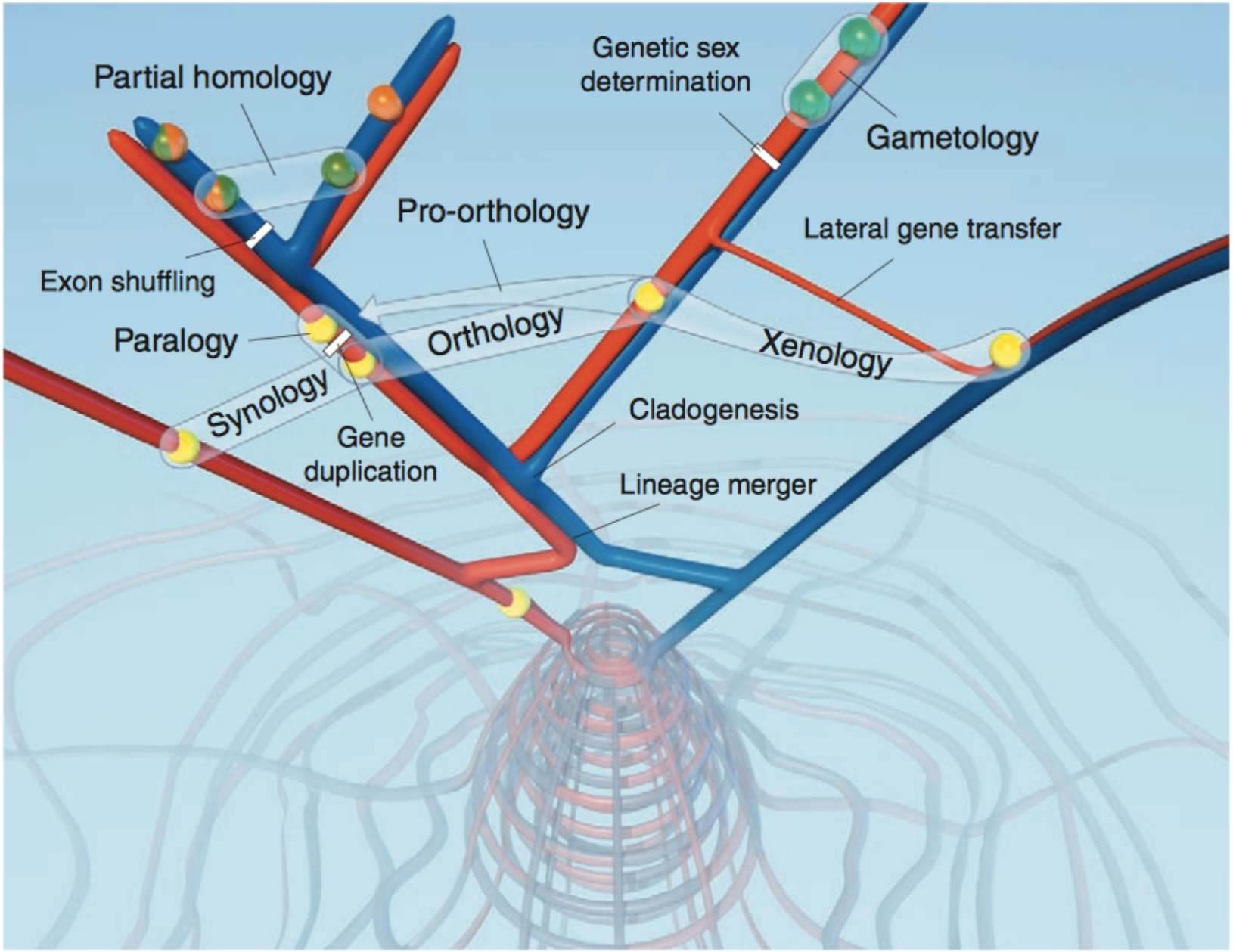
Origin of homology in proteins

Marco Punta



Origin of homology in proteins

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



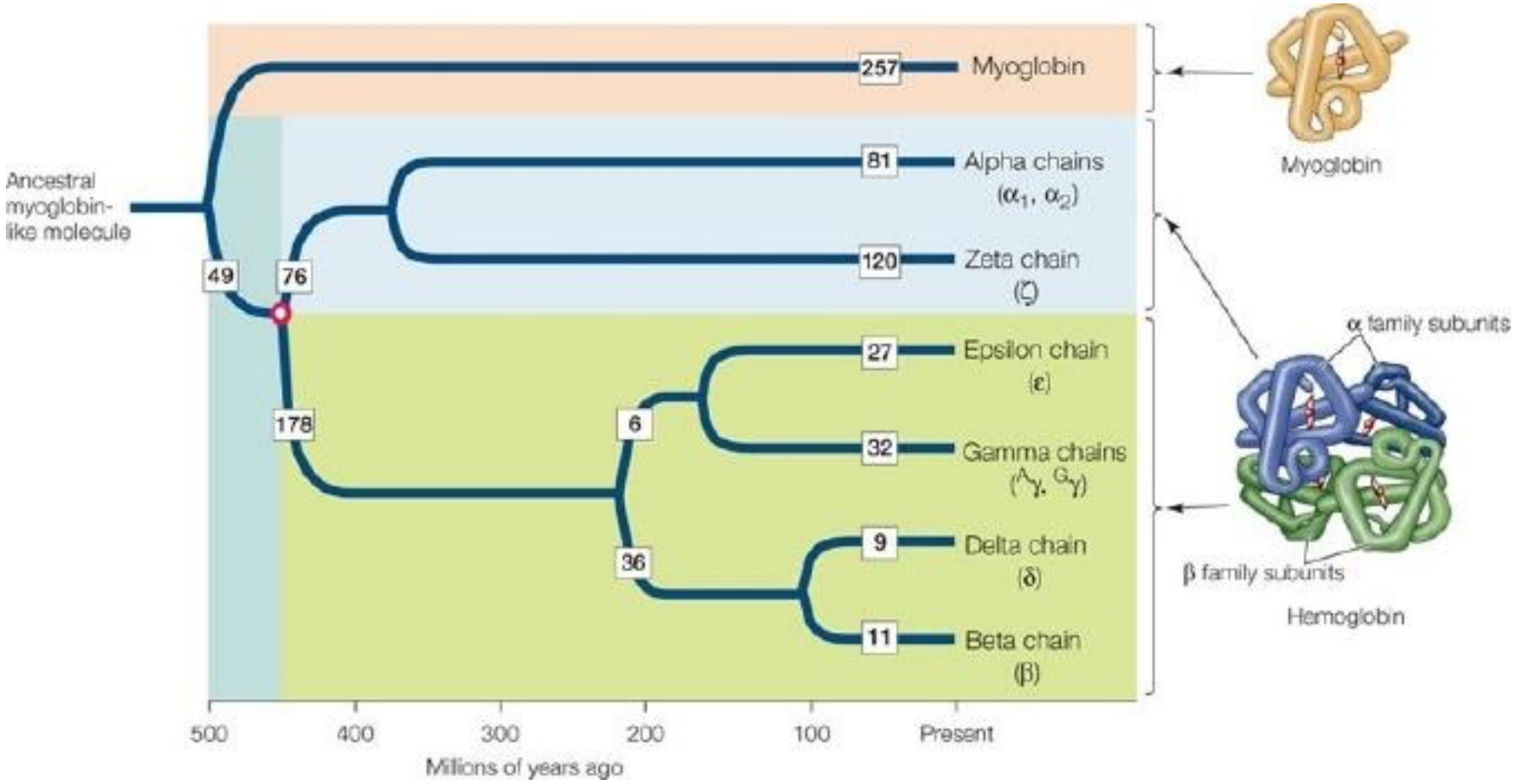
Mindell and Meyer Trends in Ecology and Evolution 2001

Protein Families

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

Homology: why bother?

Mind the gap!

Marco Punta

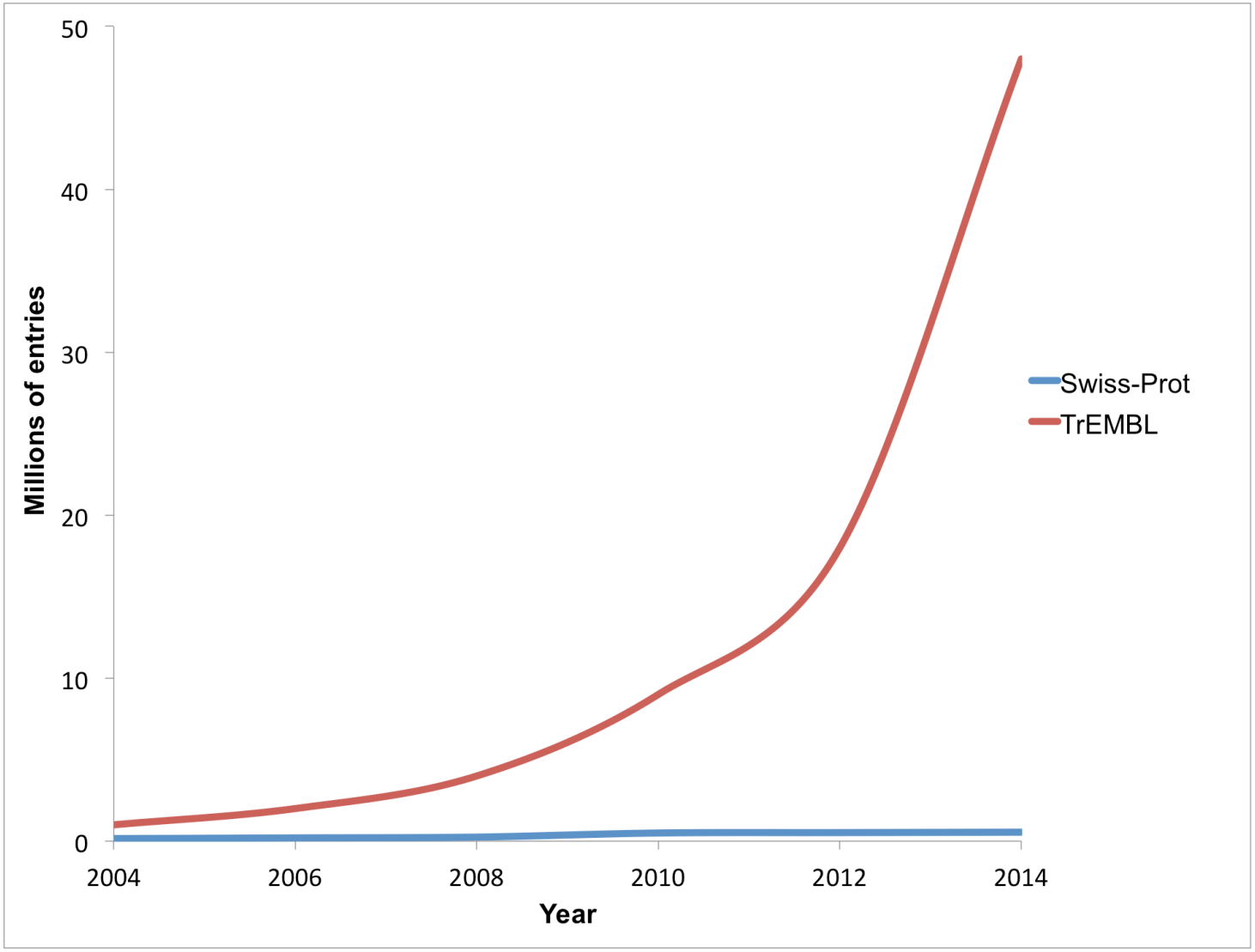


Figure courtesy of Alex Mitchell (EMBL-EBI)

Mind the gap!

Marco Punta

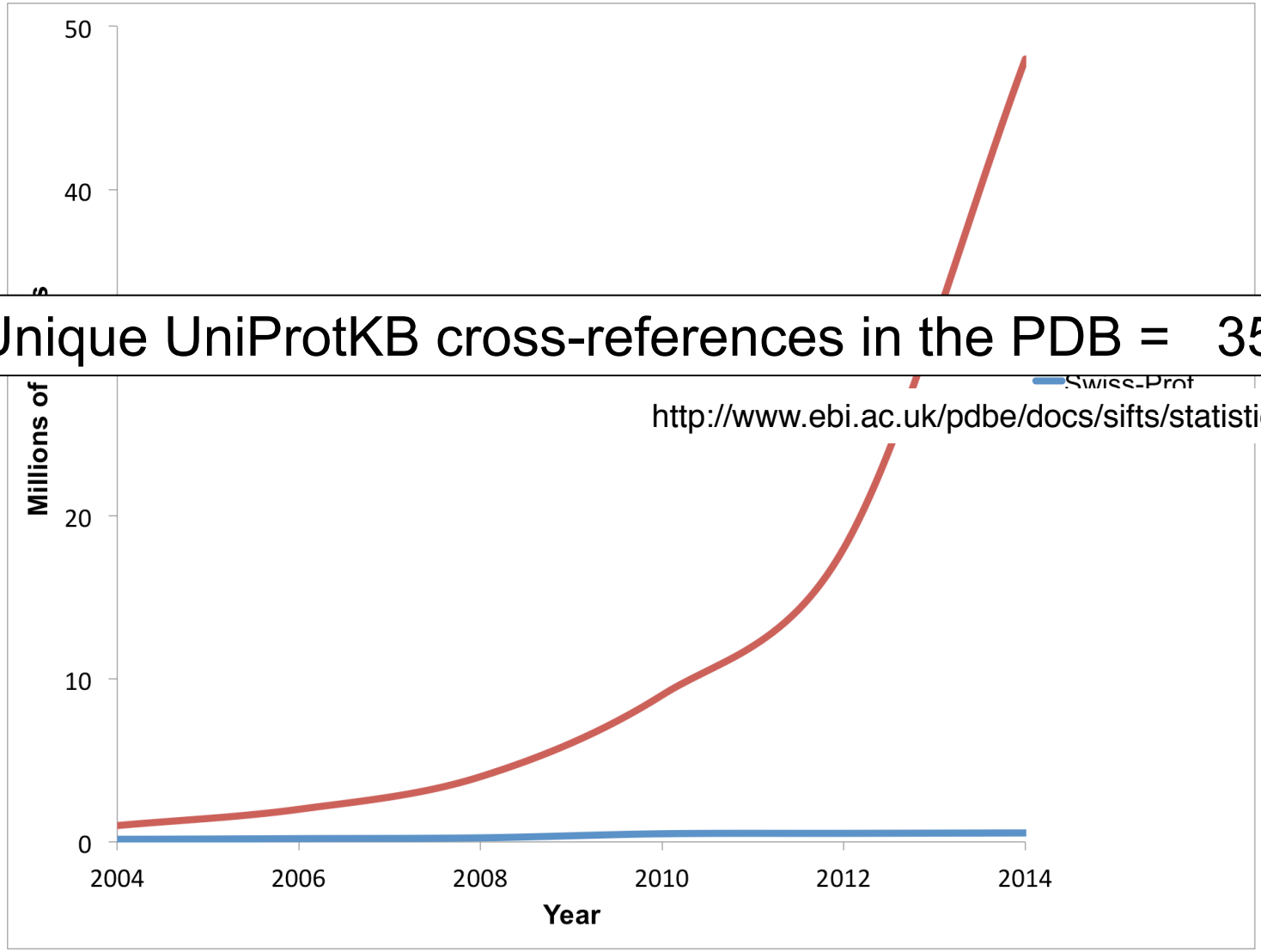


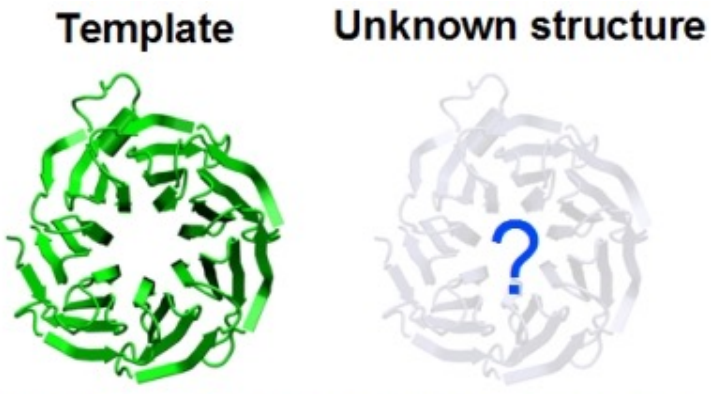
Figure courtesy of Alex Mitchell (EMBL-EBI)

Homologous protein regions have a similar (core) structure!

Chotia and Lesk *EMBO J* (1986)

Homology Modelling

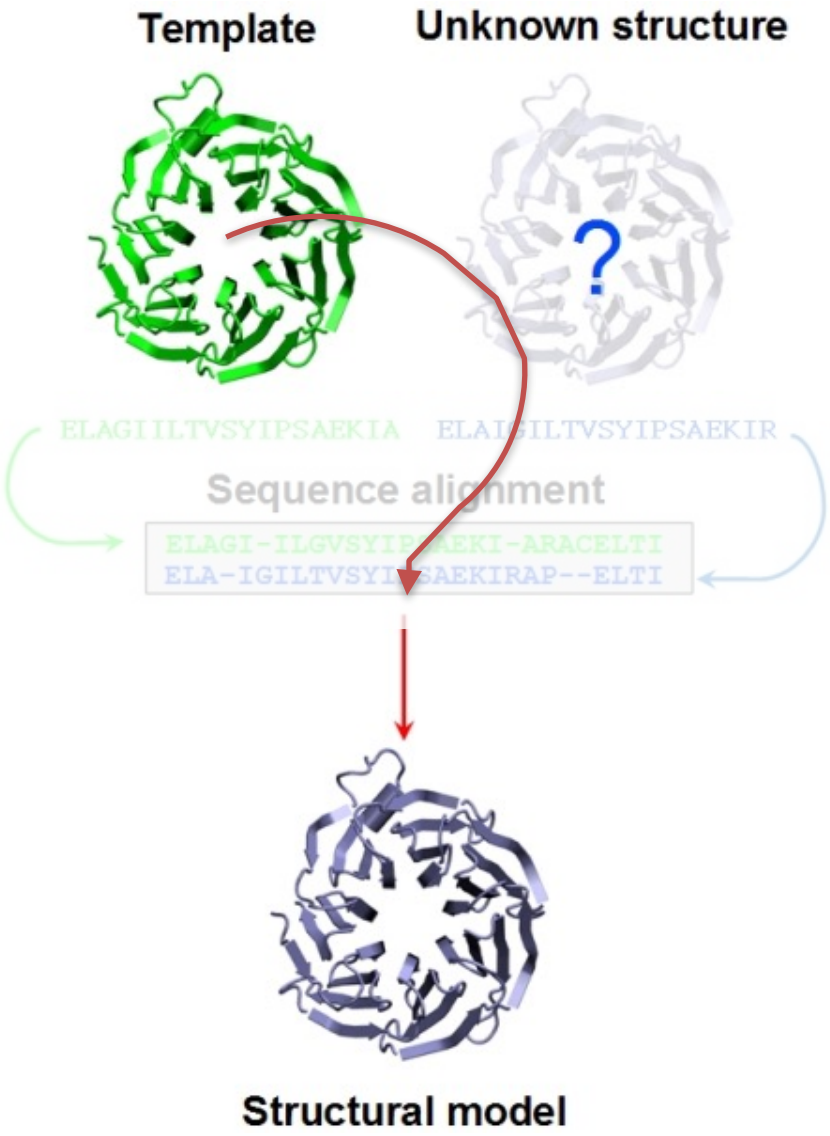
Marco Punta



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

Homology Modelling

Marco Punta



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

Homology: why bother?

Do homologous protein regions perform a similar function?

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)

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)

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)

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.

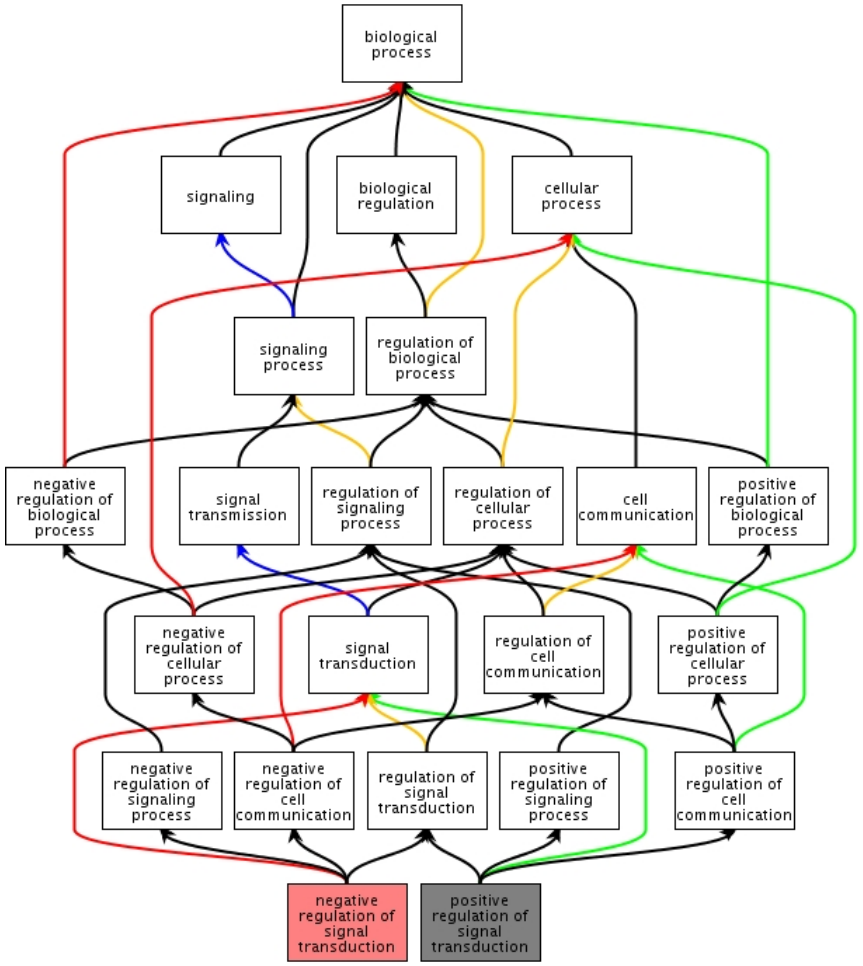
Schubert et al. Nat. Struct. Biol. 5 (1998)

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

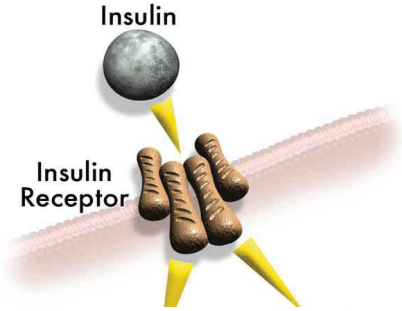
Slide courtesy of Alex Mitchell (EMBL-EBI)

GO: 3 ontologies in 1

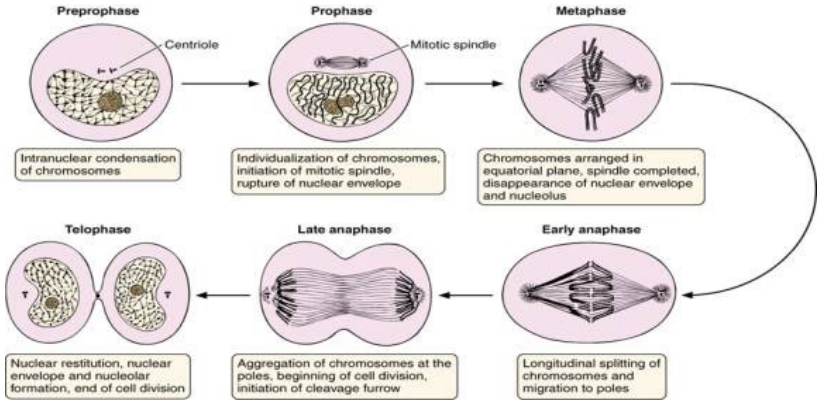
Marco Punta

1. Molecular Function

An elemental activity or task or job



- protein kinase activity
- insulin receptor activity



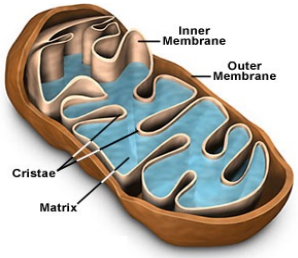
2. Biological Process

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)

CbiF GO annotation

Marco Punta

Database	Gene Product ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
Process													
UniProtKB	O87696	cbiF		GO:0006779	porphyrin-containing compound biosynthetic process	P	IEA	InterPro2GO	InterPro:IPR003043	1404	20150919	InterPro	
UniProtKB	O87696	cbiF		GO:0008152	metabolic process	P	IEA	InterPro2GO	InterPro:IPR000878|InterPro:IPR014776|InterPro:IPR014777	1404	20150919	InterPro	
UniProtKB	O87696	cbiF		GO:0009236	cobalamin biosynthetic process	P	IEA	InterPro2GO	InterPro:IPR006362	1404	20150919	InterPro	
UniProtKB	O87696	cbiF		GO:0009236	cobalamin biosynthetic process	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0169	1404	20150919	UniProt	
UniProtKB	O87696	cbiF		GO:0009236	cobalamin biosynthetic process	P	IEA	UniPathway2GO	UniPathway:UPA00148	1404	20150912	UniProt	
UniProtKB	O87696	cbiF		GO:0032259	methylation	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0489	1404	20150919	UniProt	
UniProtKB	O87696	cbiF		GO:0055114	oxidation-reduction process	P	IEA	InterPro2GO	InterPro:IPR003043	1404	20150919	InterPro	
Function													
UniProtKB	O87696	cbiF		GO:0008168	methyltransferase activity	F	IEA	InterPro2GO	InterPro:IPR000878|InterPro:IPR003043|InterPro:IPR014776|InterPro:IPR014777	1404	20150919	InterPro	
UniProtKB	O87696	cbiF		GO:0008168	methyltransferase activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0489	1404	20150919	UniProt	
UniProtKB	O87696	cbiF		GO:0016740	transferase activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0808	1404	20150919	UniProt	
UniProtKB	O87696	cbiF		GO:0043115	precorrin-2 dehydrogenase activity	F	IEA	InterPro2GO	InterPro:IPR003043	1404	20150919	InterPro	
UniProtKB	O87696	cbiF		GO:0046026	precorrin-4 C11-methyltransferase activity	F	IEA	InterPro2GO	InterPro:IPR006362	1404	20150919	InterPro	

Database	Gene Product ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Referen
UniProtKB	P02144	MB		GO:0001666	response to hypoxia	P	IEA	Ensemb
UniProtKB	P02144	MB		GO:0006810	transport	P	IEA	UniProt Prot ent
UniProtKB	P02144	MB		GO:0007507	heart development	P	IEA	Ensemb
UniProtKB	P02144	MB		GO:0009725	response to hormone	P	IEA	Ensemb
UniProtKB	P02144	MB		GO:0015671	oxygen transport	P	IEA	InterPro
UniProtKB	P02144	MB		GO:0015671	oxygen transport	P	IEA	Ensemb
UniProtKB	P02144	MB		GO:0015671	oxygen transport	P	IEA	UniProt Prot ent
UniProtKB	P02144	MB		GO:0031444	slow-twitch skeletal muscle fiber contraction	P	IEA	Ensemb
UniProtKB	P02144	MB		GO:0042542	response to hydrogen peroxide	P	IEA	Ensemb
UniProtKB	P02144	MB		GO:0043353	enucleate erythrocyte differentiation	P	IEA	Ensemb
UniProtKB	P02144	MB		GO:0050873	brown fat cell differentiation	P	IEA	Ensemb
UniProtKB	P02144	MB		GO:0005344	oxygen transporter activity	F	IEA	UniProt Prot ent
UniProtKB	P02144	MB		GO:0005506	iron ion binding	F	IEA	InterPro
UniProtKB	P02144	MB		GO:0019825	oxygen binding	F	IEA	InterPro
UniProtKB	P02144	MB		GO:0019825	oxygen binding	F	IEA	Ensemb
UniProtKB	P02144	MB		GO:0020037	heme binding	F	IEA	InterPro
UniProtKB	P02144	MB		GO:0046872	metal ion binding	F	IEA	UniProt Prot ent
UniProtKB	P02144	MB		GO:0070062	extracellular vesicular exosome	C	IDA	PMID:23

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:11597	UniProtKB:P68871	9606	20140914	IntAct	
UniProtKB	P02008	HBZ		GO:0005515	protein binding	F	IPI	PMID:6	UniProtKB:P68871	9606	20140914	IntAct	
UniProtKB	P02008	HBZ		GO:0019825	oxygen binding	F	IEA	InterPro2GO	InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0020037	heme binding	F	IEA	InterPro2GO	InterPro:IPR000971|InterPro:IPR002338|InterPro:IPR002340|InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0046872	metal ion binding	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0479	9606	20140913	UniProt	
Component													
UniProtKB	P02008	HBZ		GO:0005833	hemoglobin complex	C	IEA	InterPro2GO	InterPro:IPR002338|InterPro:IPR002340	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		GO:0005833	hemoglobin complex	C	TAS	PMID:7555018		9606	20030904	PINC	
UniProtKB	P02008	HBZ		GO:0070062	extracellular vesicular exosome	C	IDA	PMID:23533145		9606	20140714	UniProt	

SAME

SAME

Do homologous protein regions perform a similar function?

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)

We can integrate homology with other information, for example:

- Functional motifs

- Conservation of functional residues

- Genomic context (mostly in bacteria)

If structure available:

- structural motifs

- Electrostatic, cavities, etc.

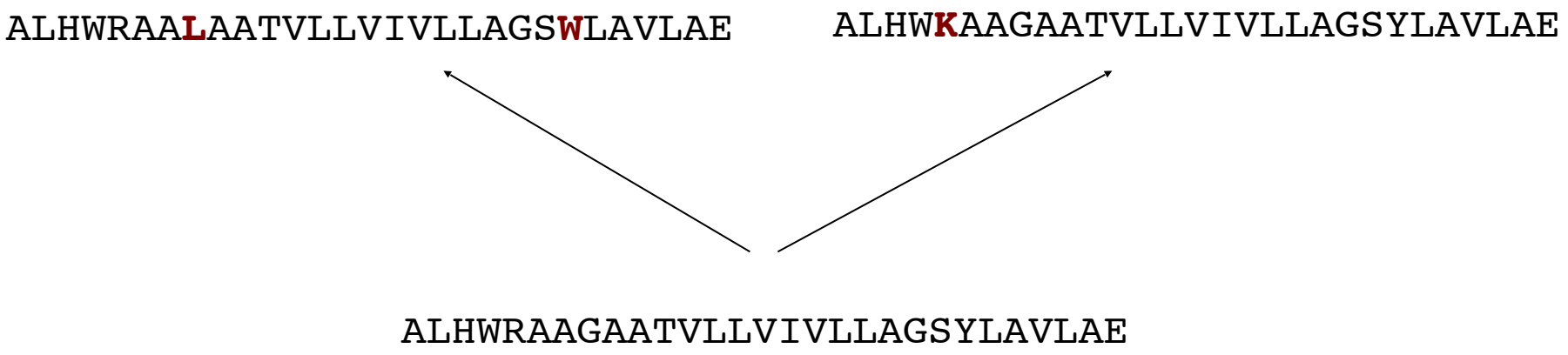
Detecting homology

Marco Punta

From sequence

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

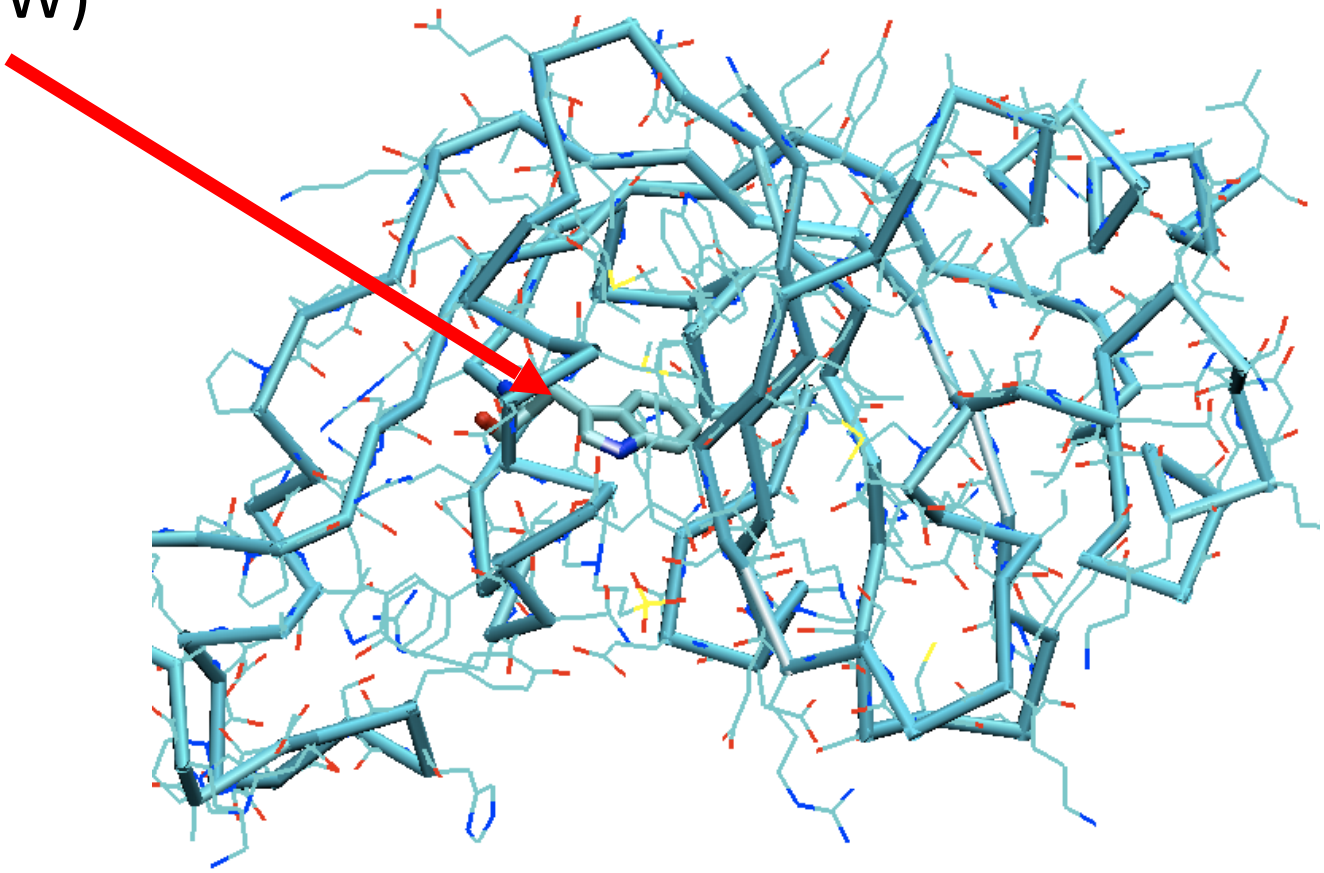
Modern day homologous proteins have evolved from the same sequence via a number of events (mutations, insertions, deletions, duplications,…)



```
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
```

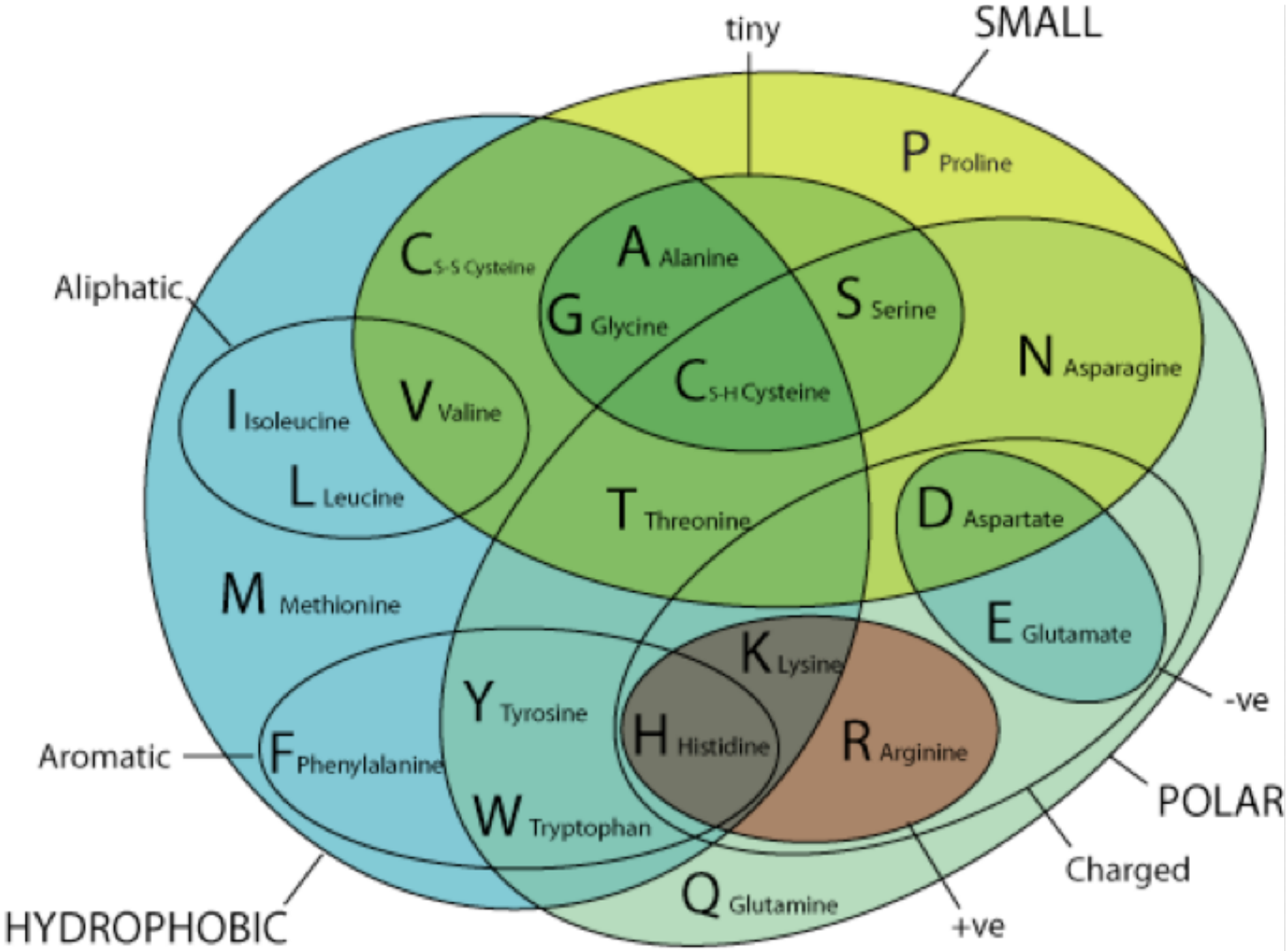
Protein structural and functional constraints

Trp (W)

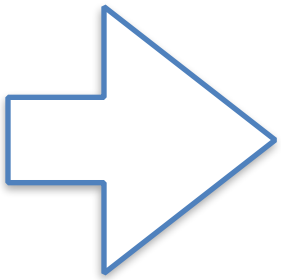


aa physico-chemical properties

Marco Punta



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



If divergence not too large we can hope to use sequence similarity to detect homology (excess sequence similarity* -> homology)

*Pearson Curr Protoc Bioinformatics. 2013

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

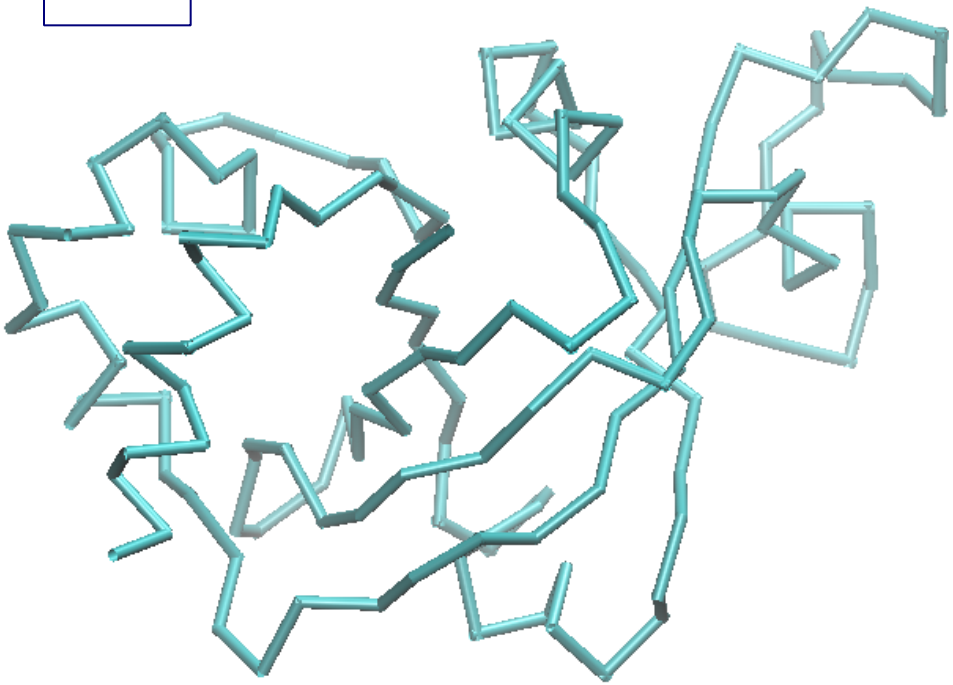
Sequence alignment, what we need:

- 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, …

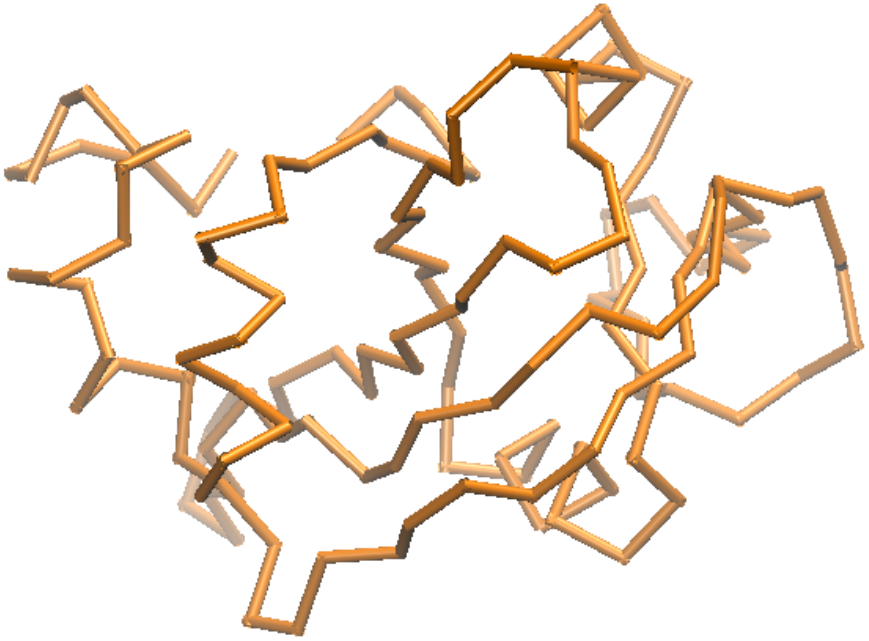
From structure

Marco Punta

2G2X



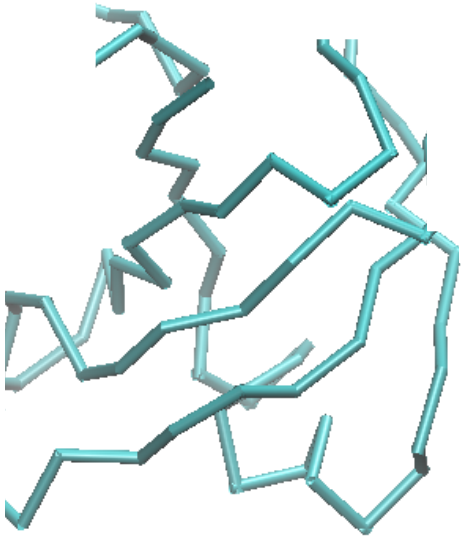
2P5D



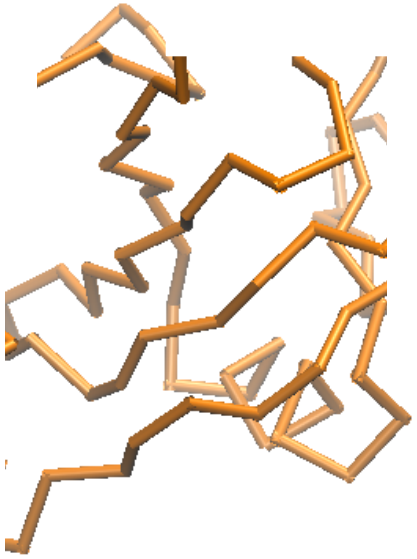
Structural similarity

Marco Punta

2G2X



2P5D

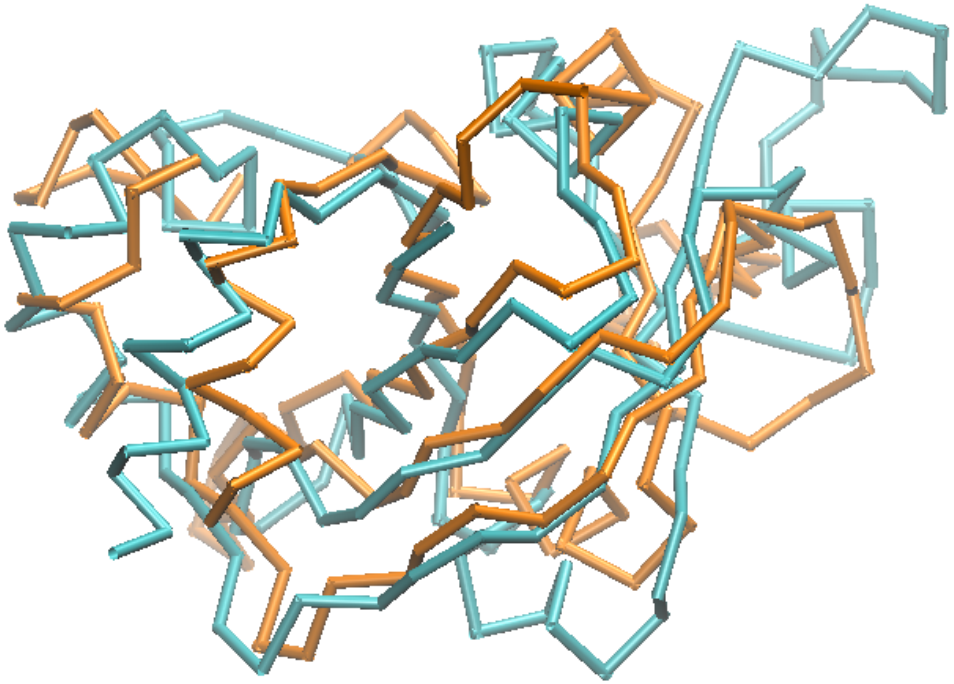


Structural similarity

Marco Punta

2G2X

2P5D



Z-score = 12.2
RMSD = 2.9
Lali = 122
%id =20

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

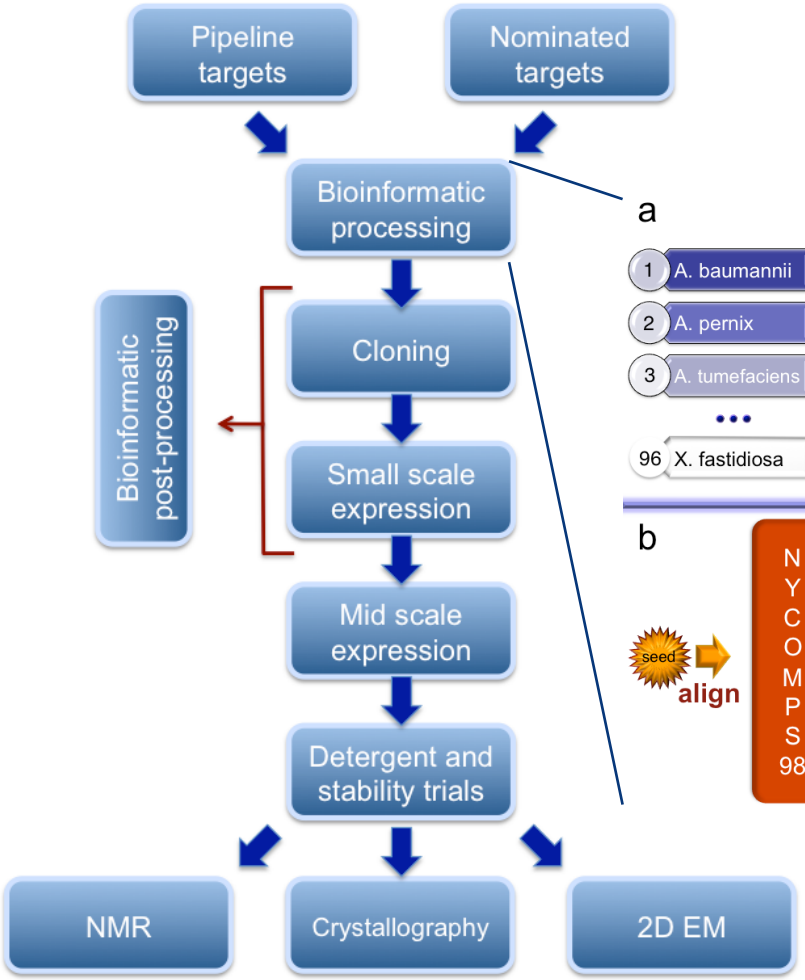
Marco Punta

Exercise

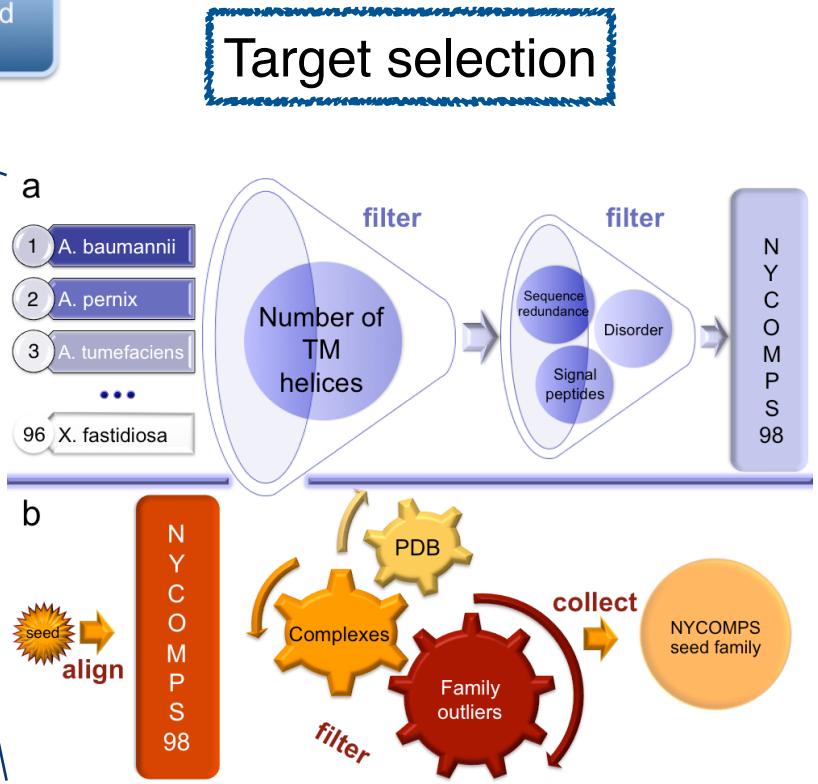
Homology-based function annotation transfer #1

Marco Punta

NYCOMPS pipeline



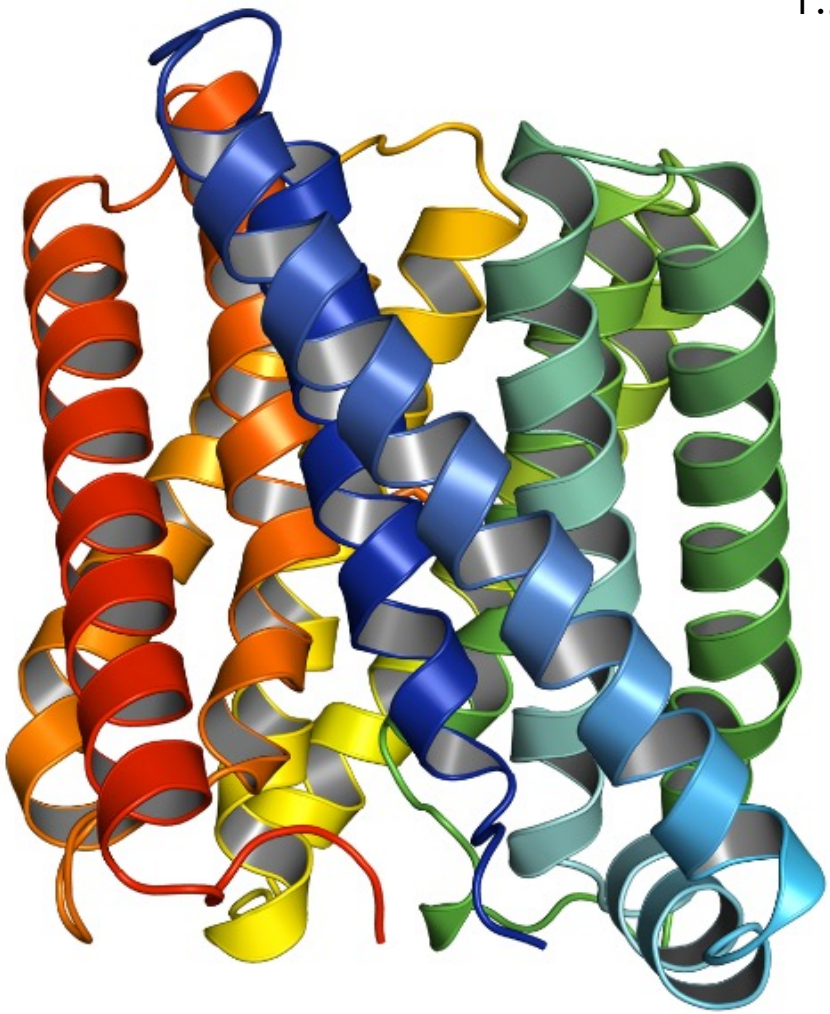
Love et al. *JSFG* 2010

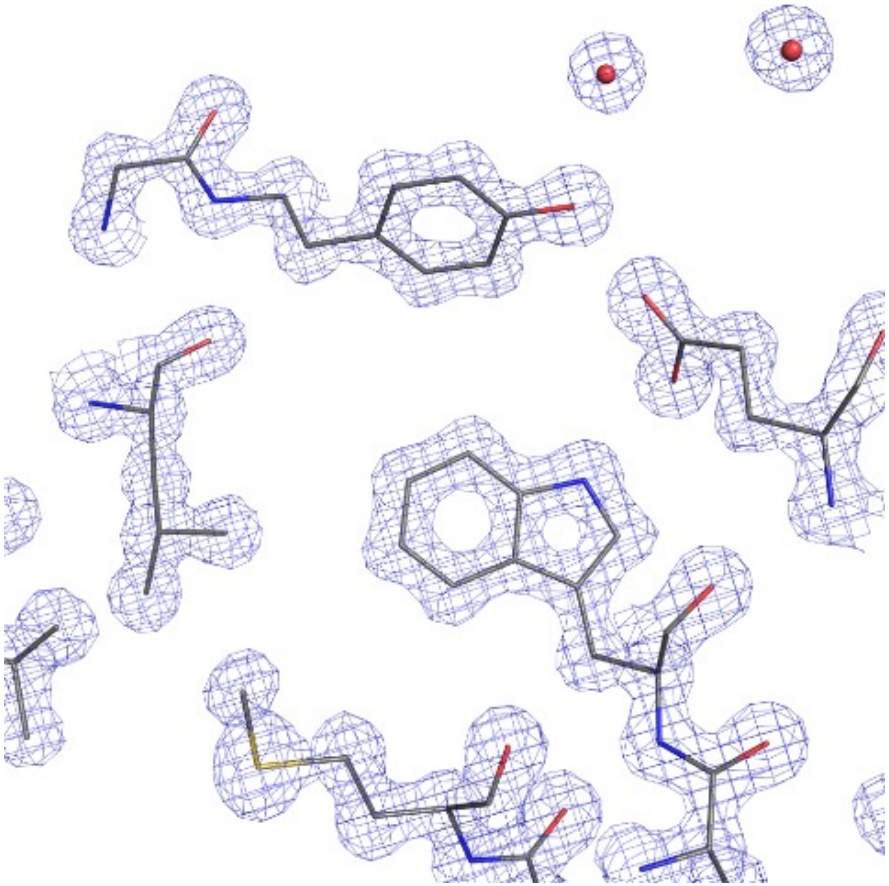


Punta et al. *JSFG* 2010

H. influenzae protein [3M71] ← PDB id

1.20 Å





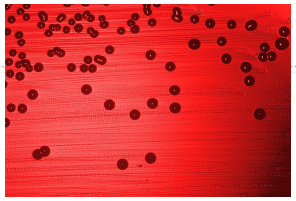
Chen et al. *Nature* 467 (2010)

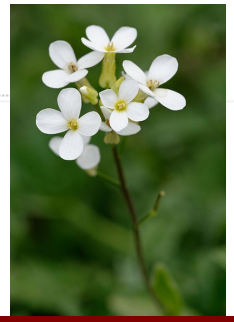
Alignment

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

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

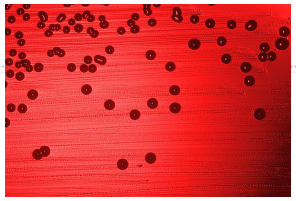
For an alignment with score S and E-value=1, we expect to have by chance 1 match with the same or higher score. If E-value is 0.001 then we expect by chance 0.001 matches with the same or higher score.

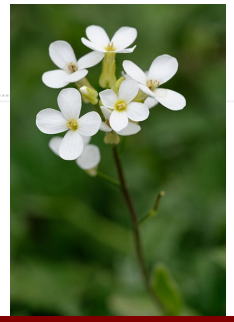
Alignment

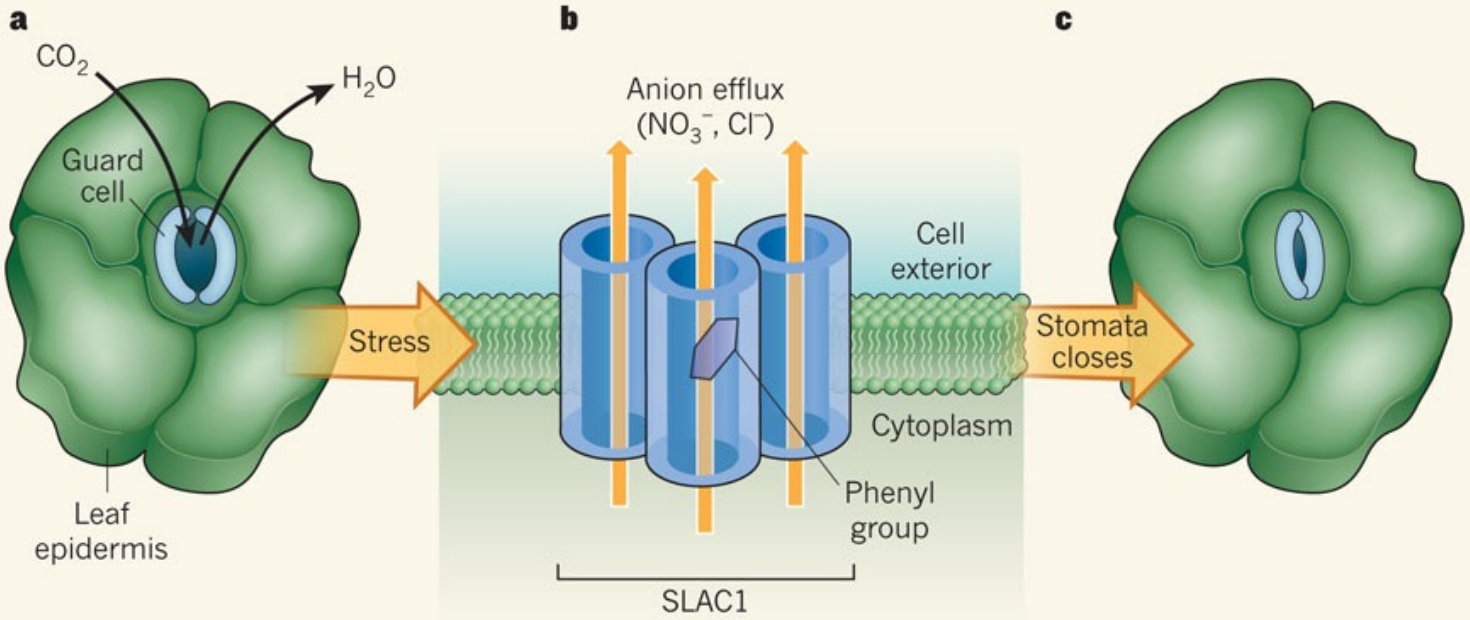
Q9LD83 SLAC1_ARATH - **Guard cell S-type anion channel SLA...** - Arabidopsis thal...

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

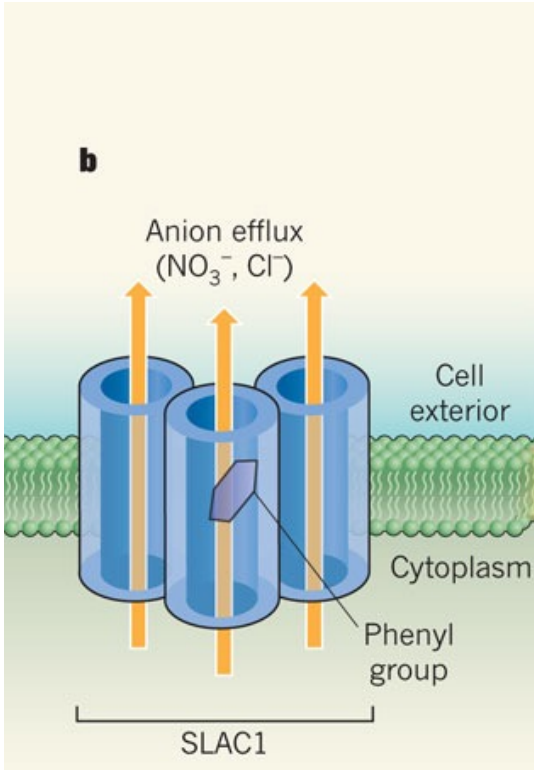
```
P44741    20   PFPL--PTGYFGIPLGLAALSLAWFHLE-----NLFPAARMVSDVLGIVASAVWILFILM   72
               PF L   P G FGI LGL++ ++ W  L       N       +++ V+ + +  V +
Q9LD83   183   PFLLRFPIGCFGICLGLSSQAVLWLALAKSPATNFLHITPLINLVVWLFSLVVLVSVSFT  242

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   ALFLYISLVARINFFTGFKFSVAWWSYTFPMTT-ASVATIKYAEAVPGYPSRALALTLSF  480

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



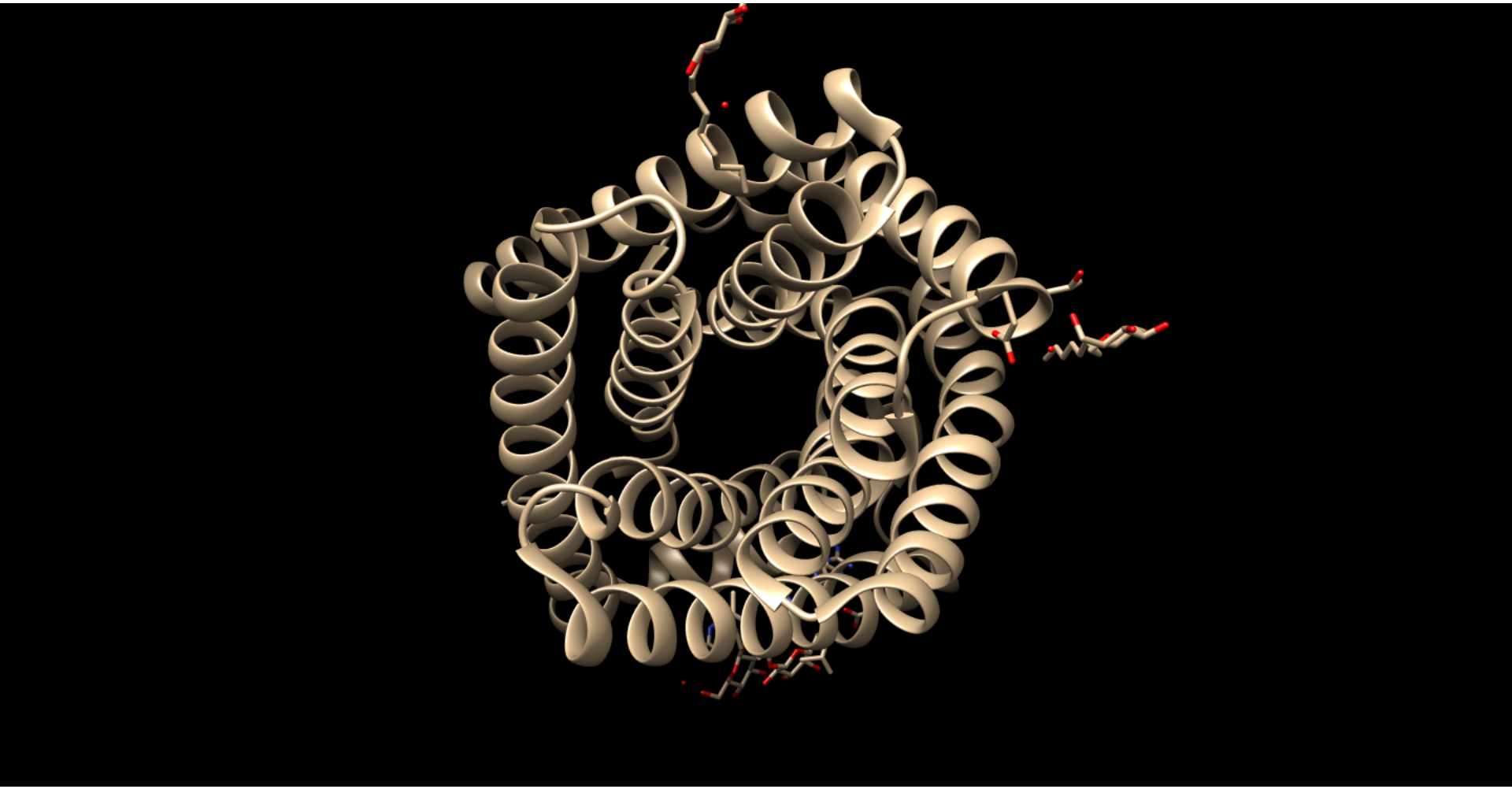


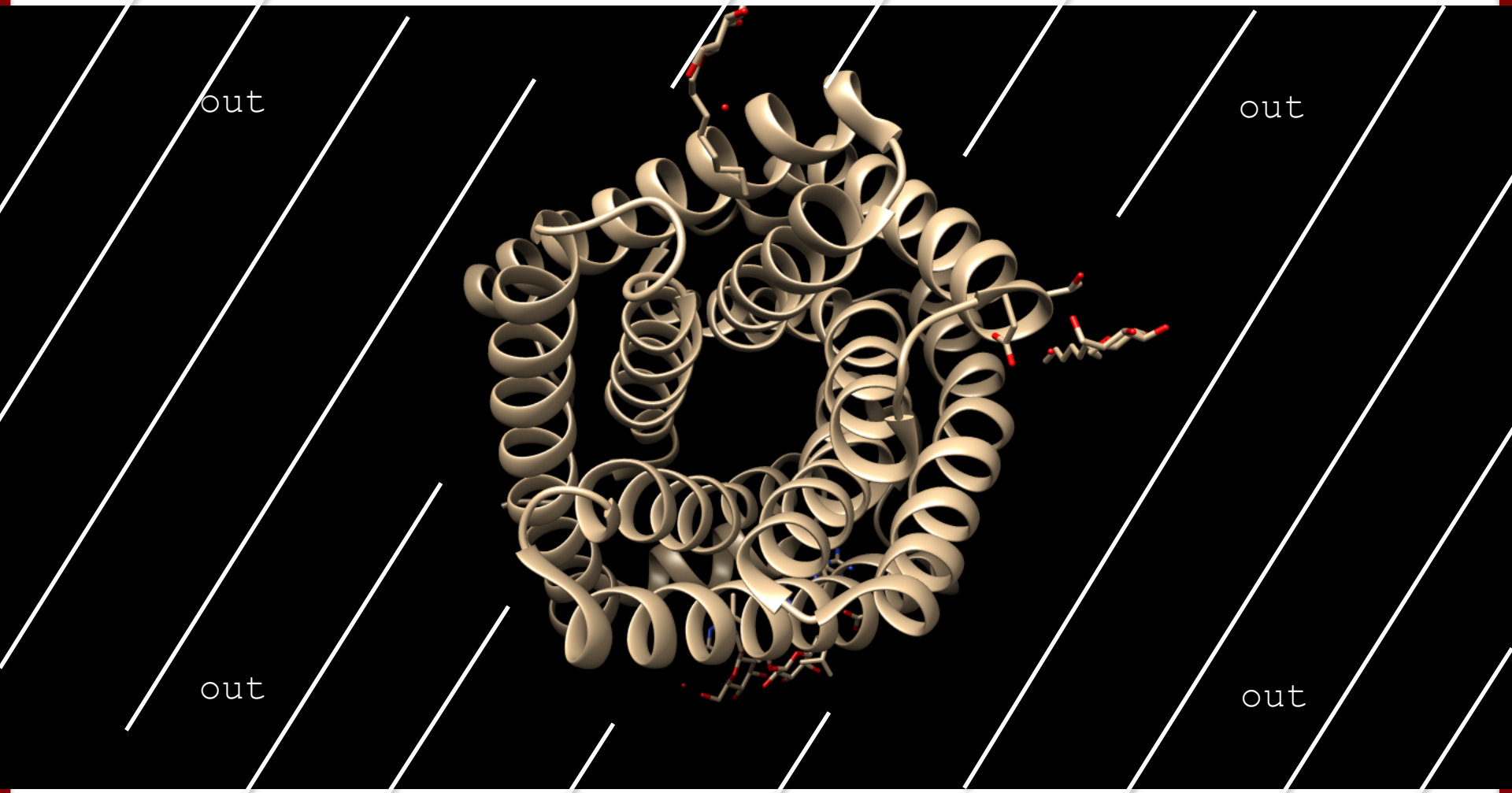


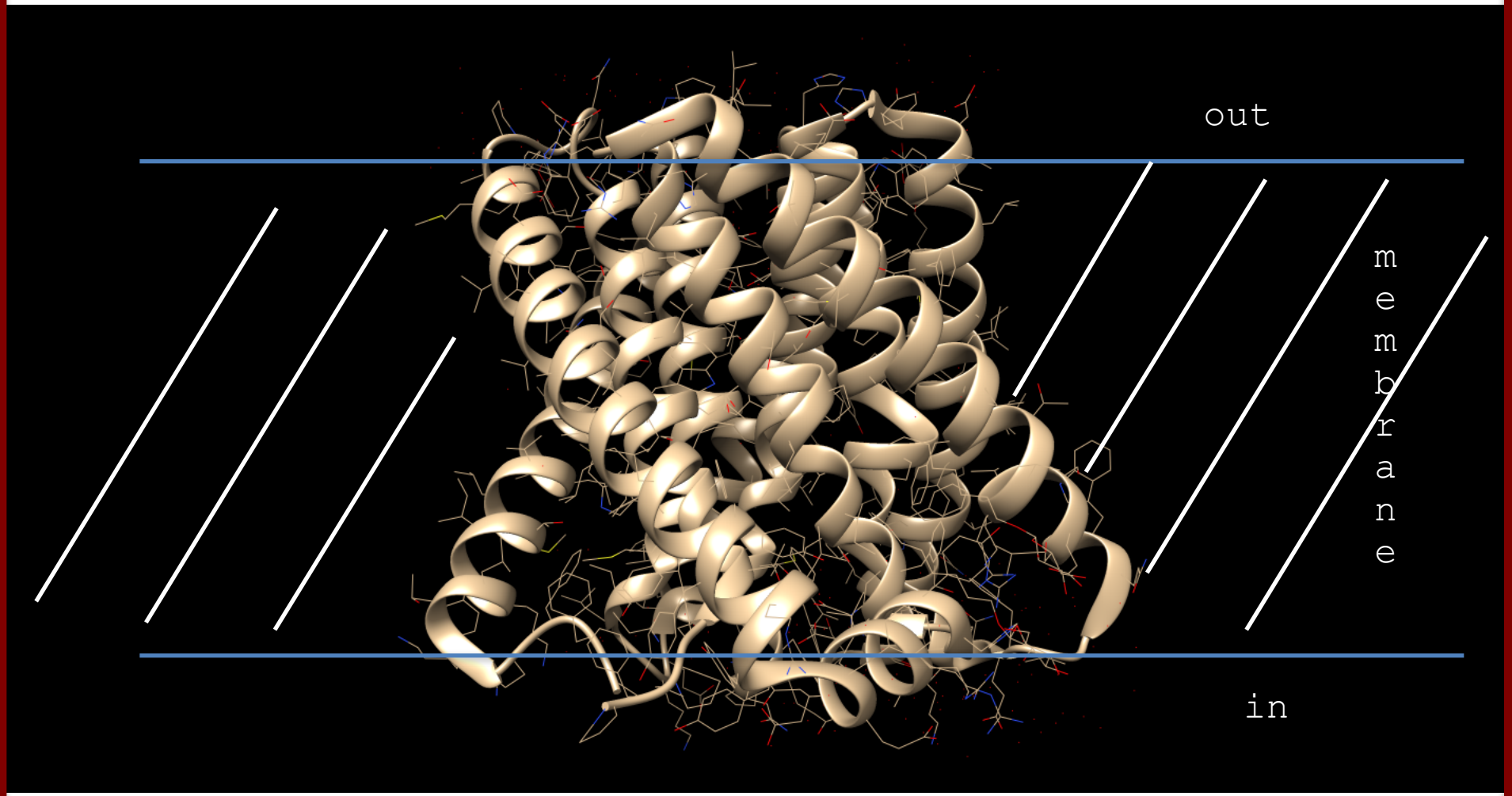
Thomine and Barbier-Brygoo Nature **467**:1058-59 (2010)

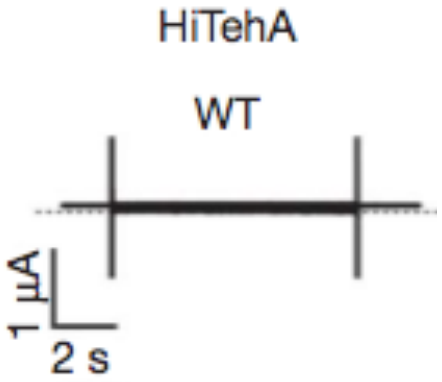

b
Anion efflux
(NO_3^-, Cl^-)
Cell
exterior
Cytoplasm
Phenyl
group
SLAC1

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

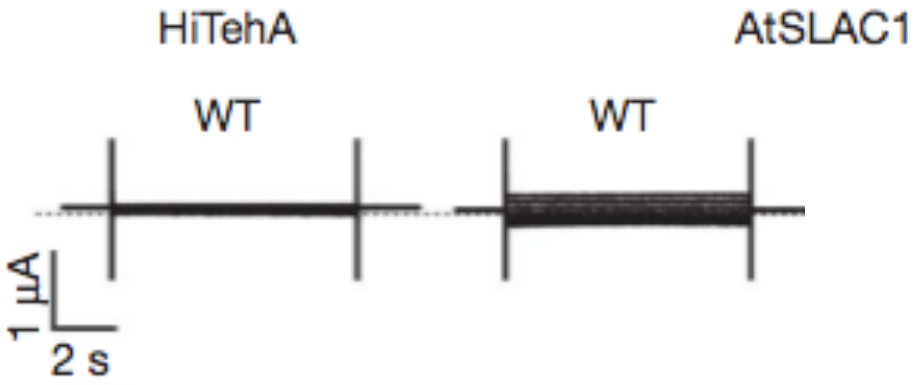



out
out
out
out

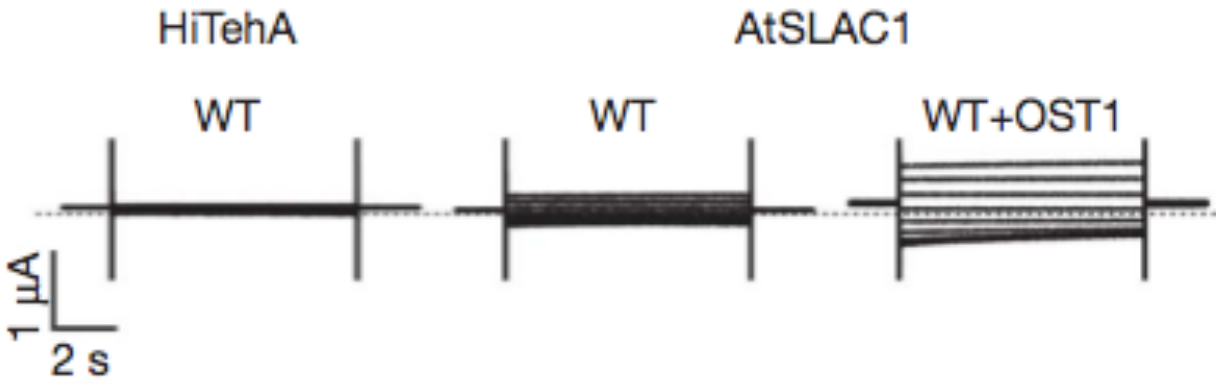

out
membrane
in



Chen et al. *Nature* 467 (2010)



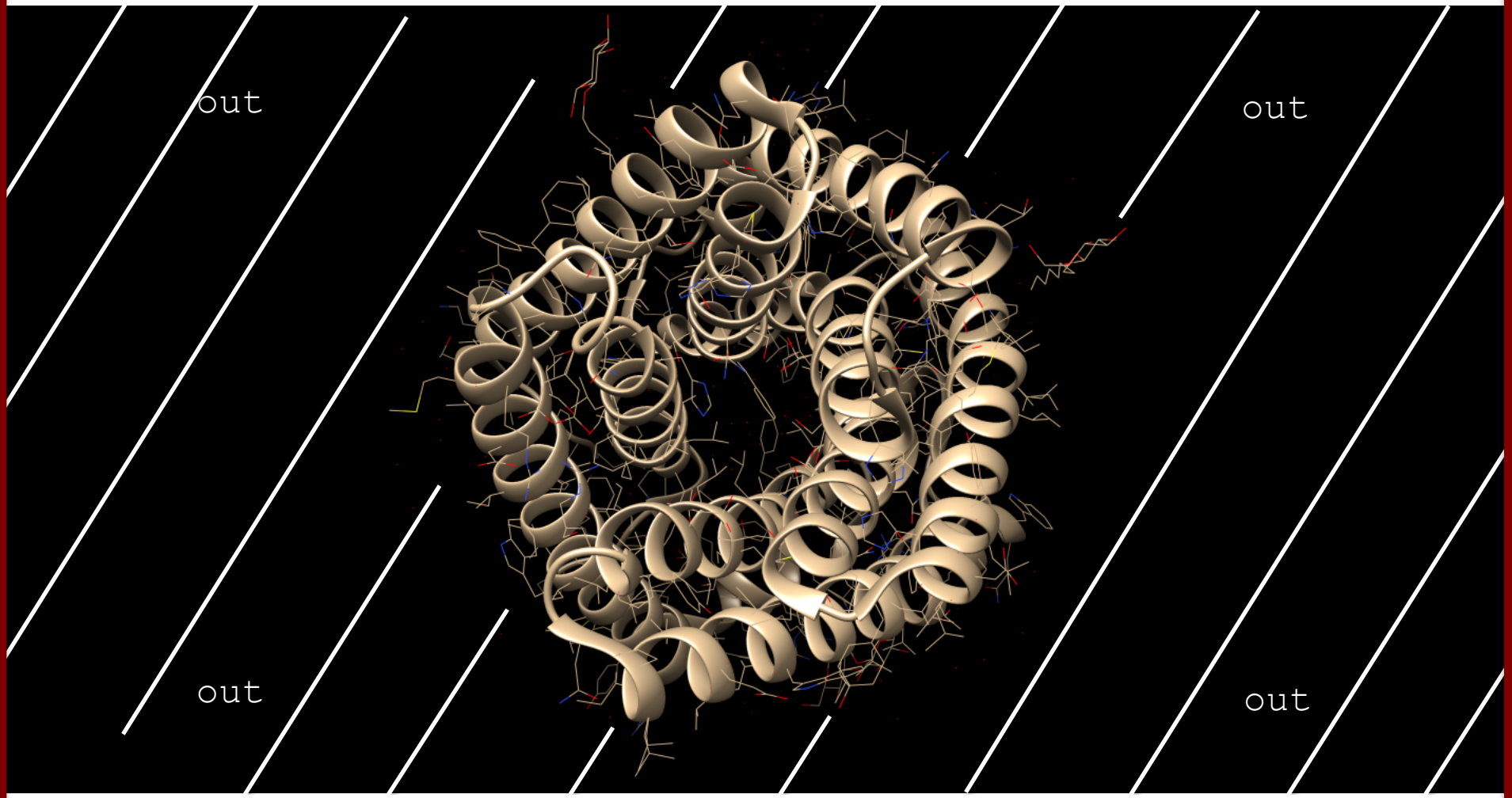
Chen et al. *Nature* 467 (2010)



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

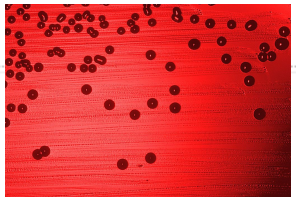


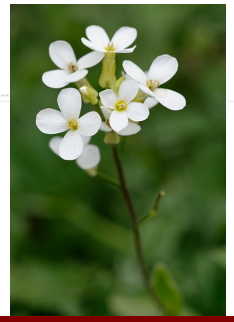
Alignment

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

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

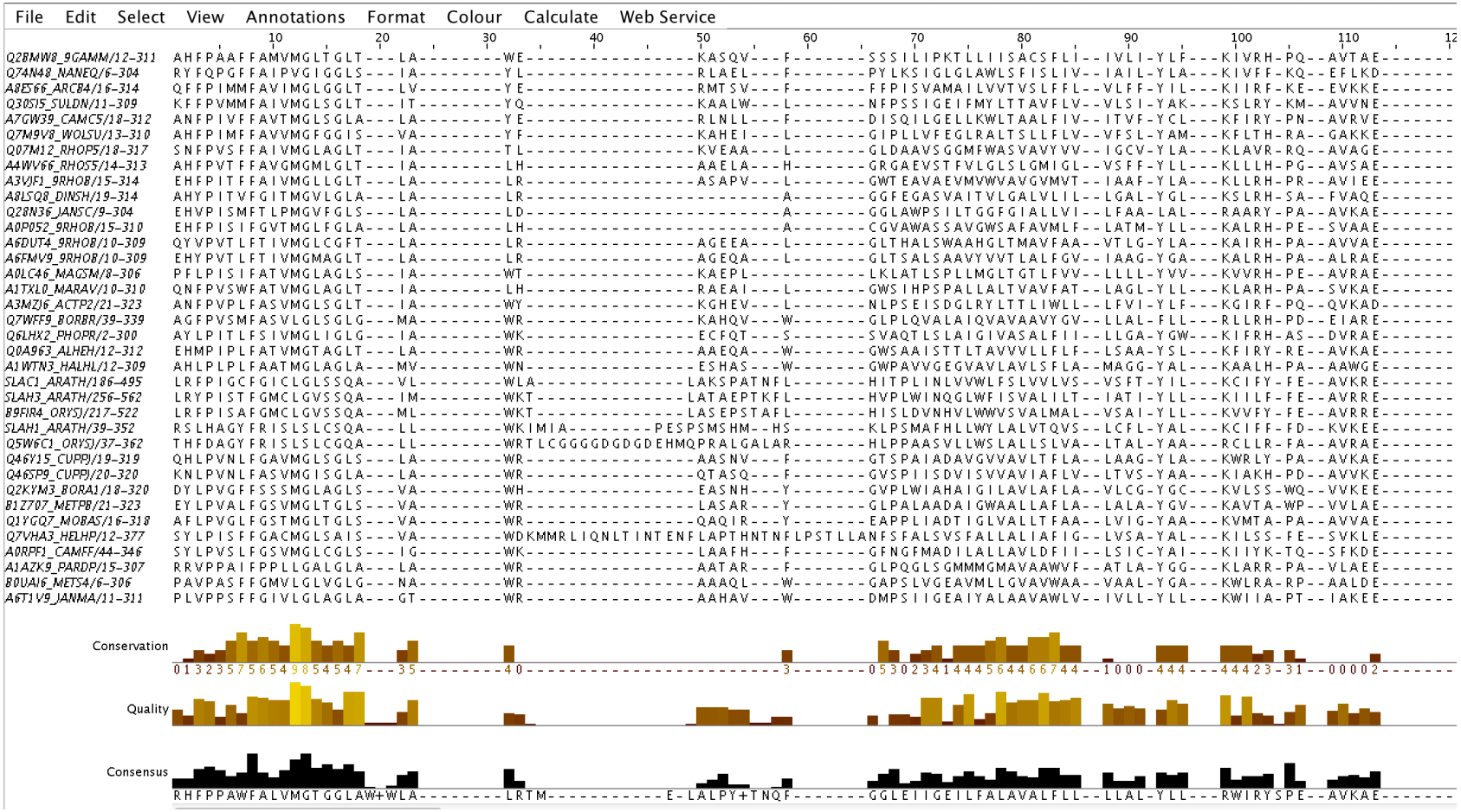
```
P44741    20   PFPL--PTGYFGIPLGLAALSLAWFHLE-----NLFPAARMVSDVLGIVASAVWILFILM   72
               PF L   P G FGI LGL++ ++ W  L       N       +++ V+ + +  V +
Q9LD83   183   PFLLRFPIGCFGICLGLSSQAVLWLALAKSPATNFLHITPLINLVVWLFSLVVLVSVSFT  242

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

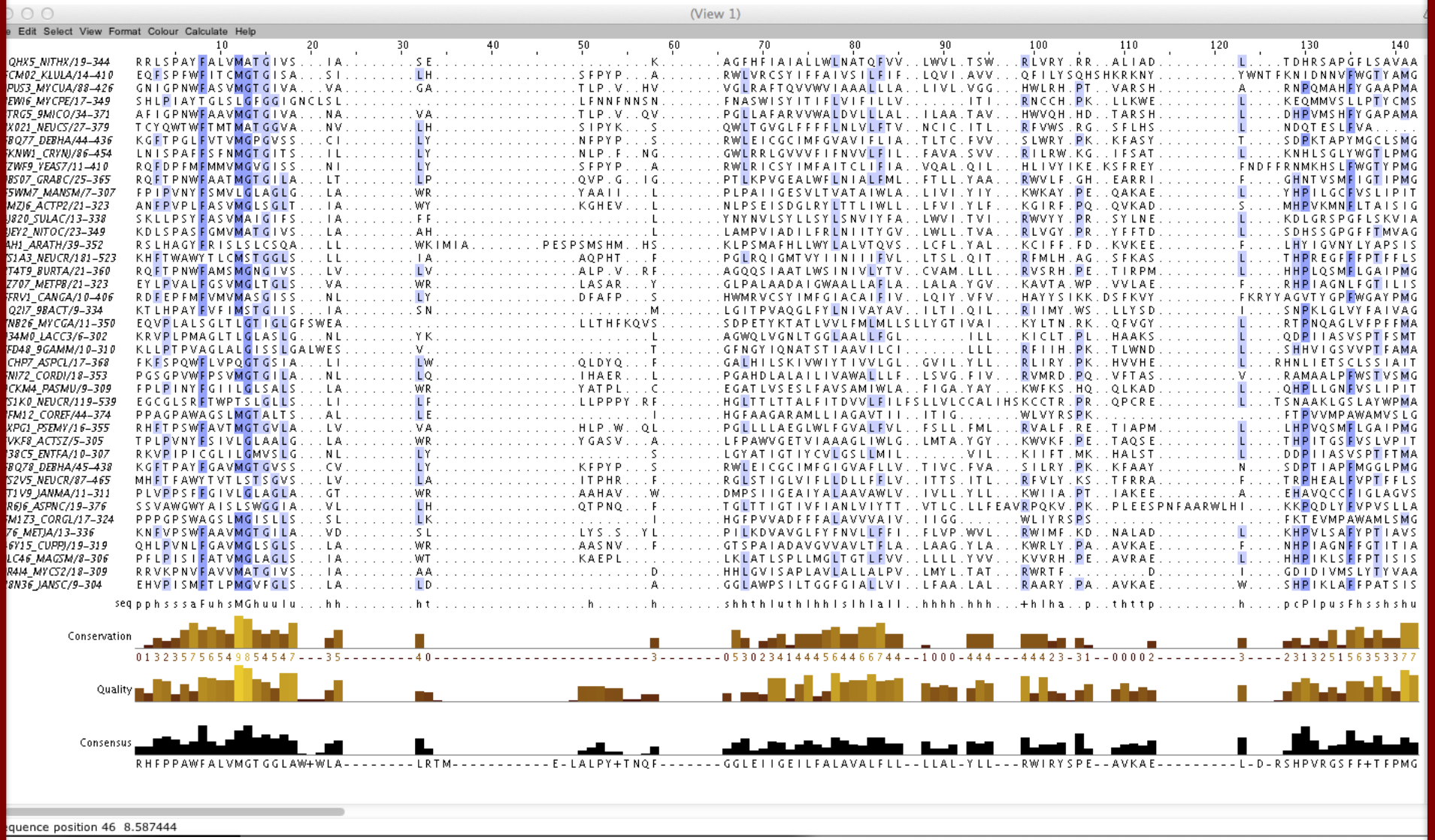
2. File -> Input Alignment -> From File “PF03595_seed.txt”



Marco Punta

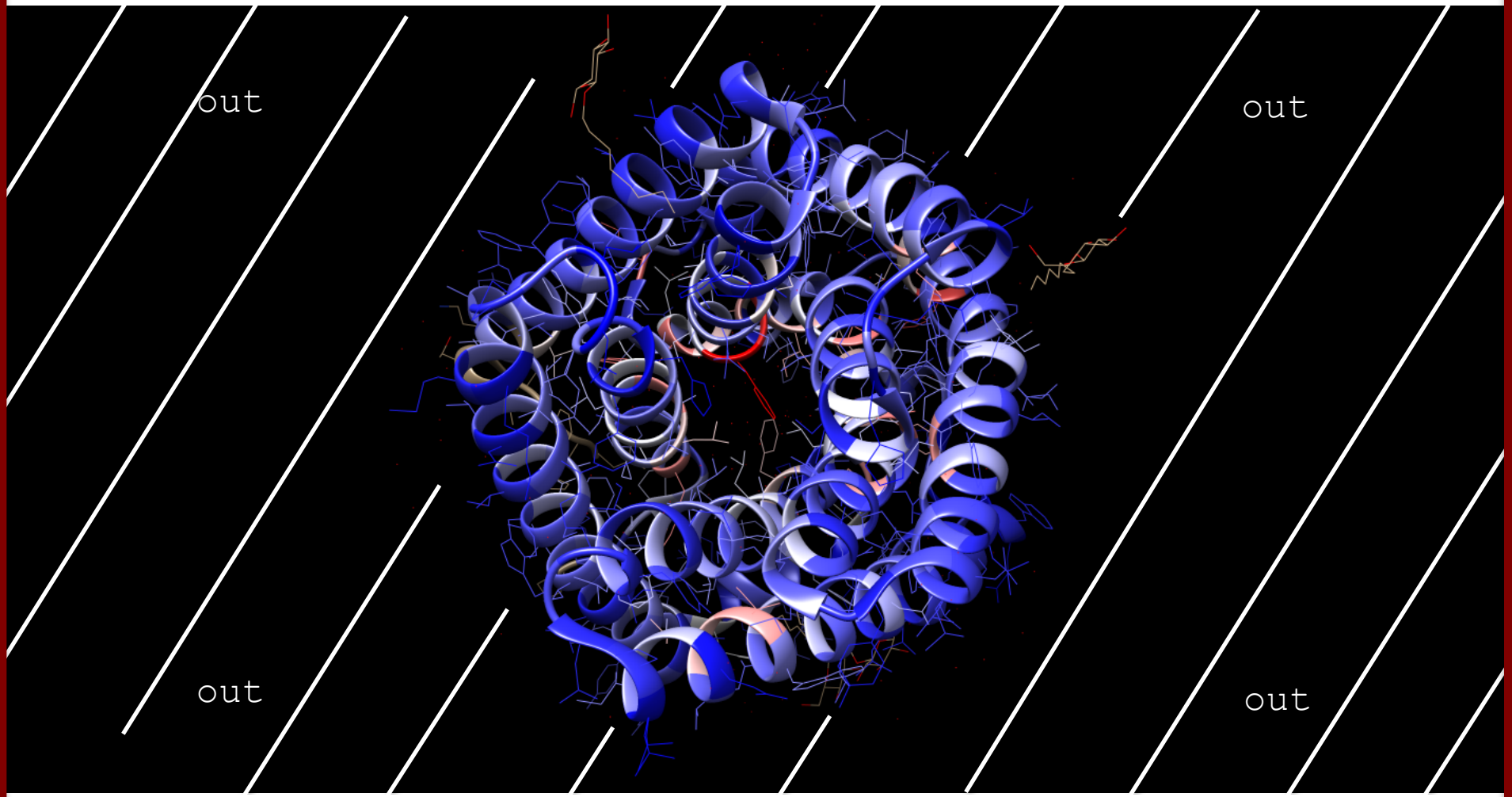
Marco Punta

1. Colour -> BLOSUM62

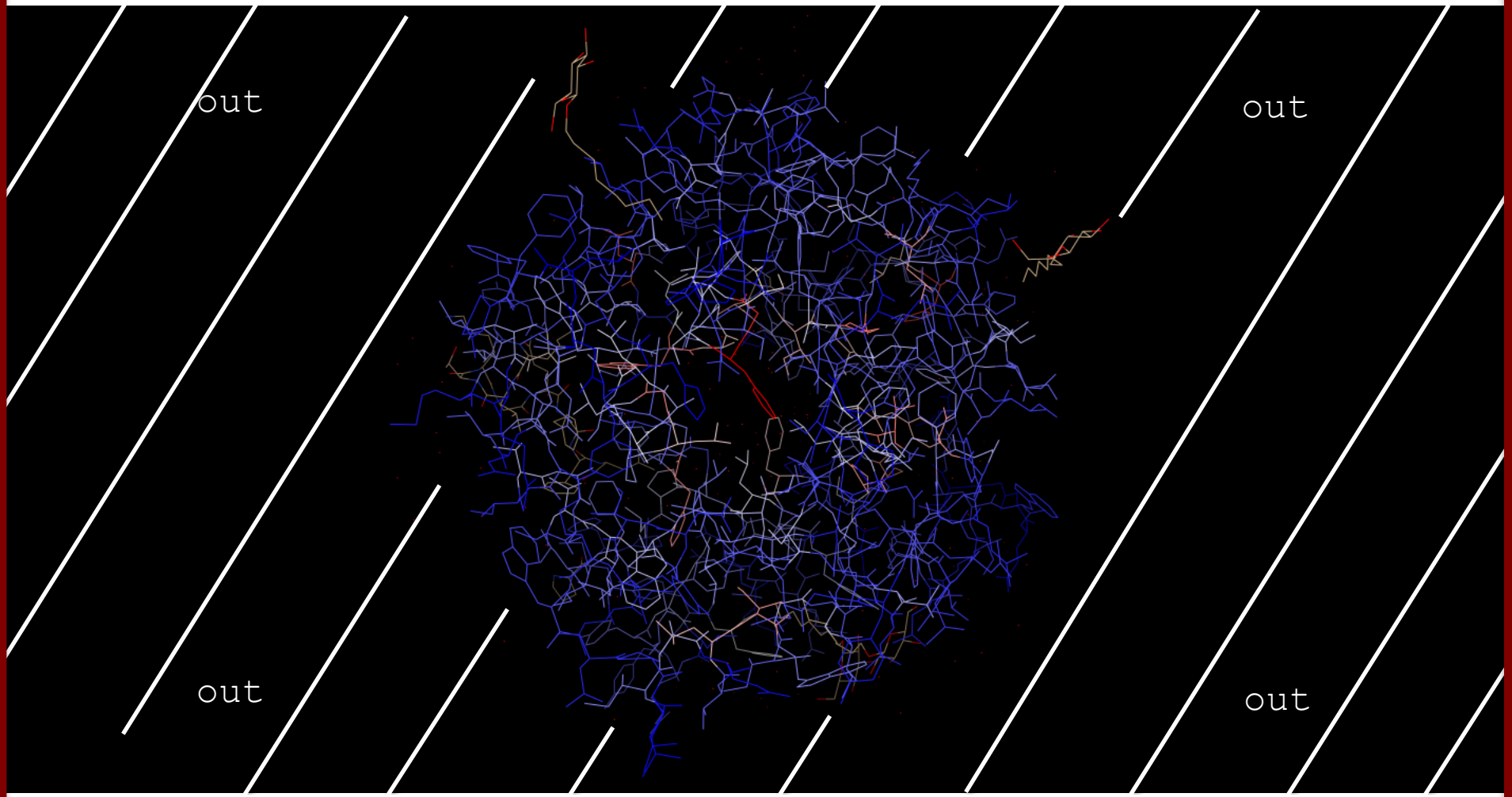


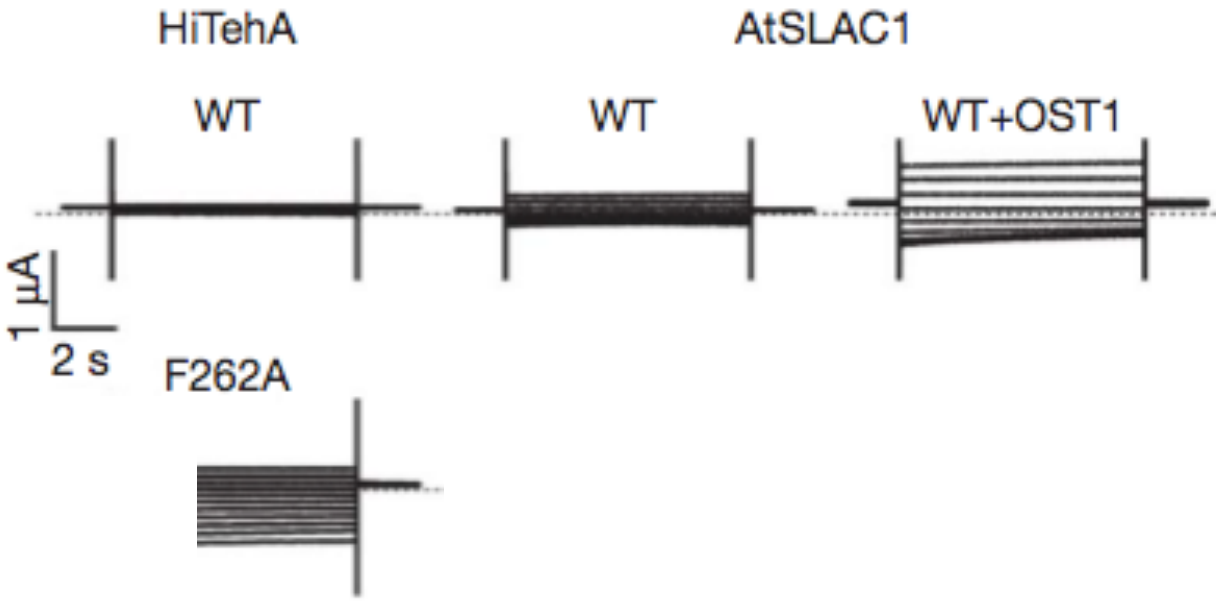
1. Tools-> Sequence -> Multialign Viewer
2. Choose “PF03595_seed.txt”
3. Select Aligned FASTA

4. Structure -> Render by Conservation

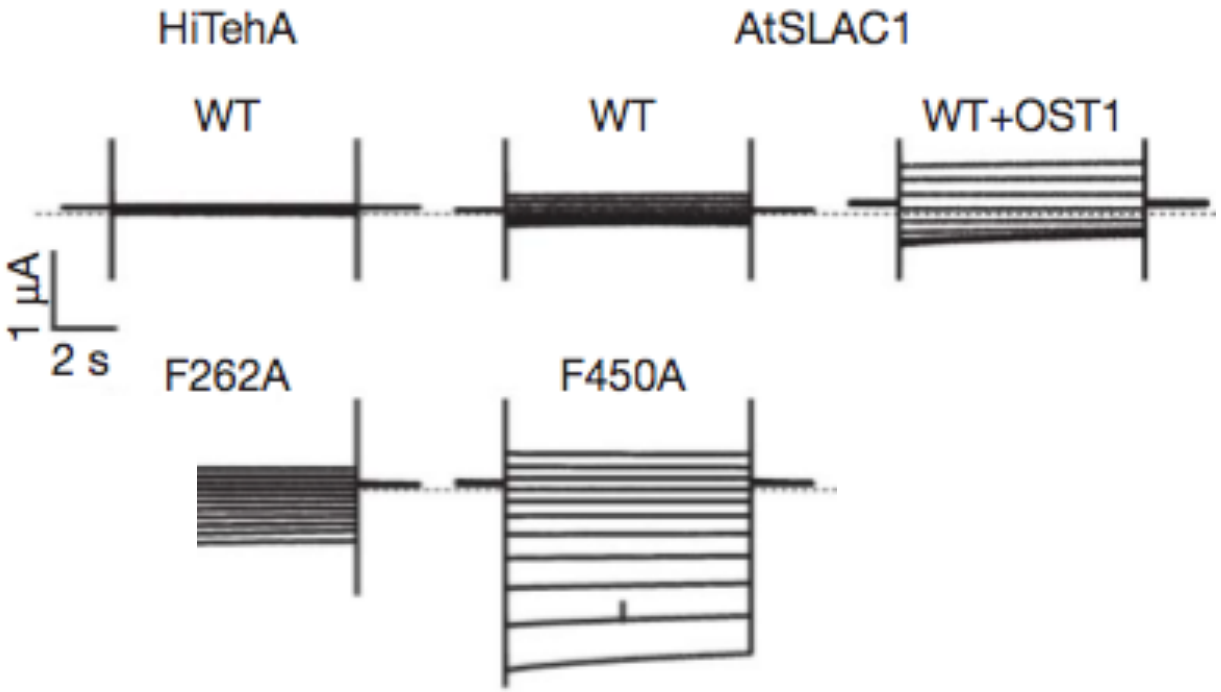

out
out
out
out

1. Actions-> Ribbon -> hide

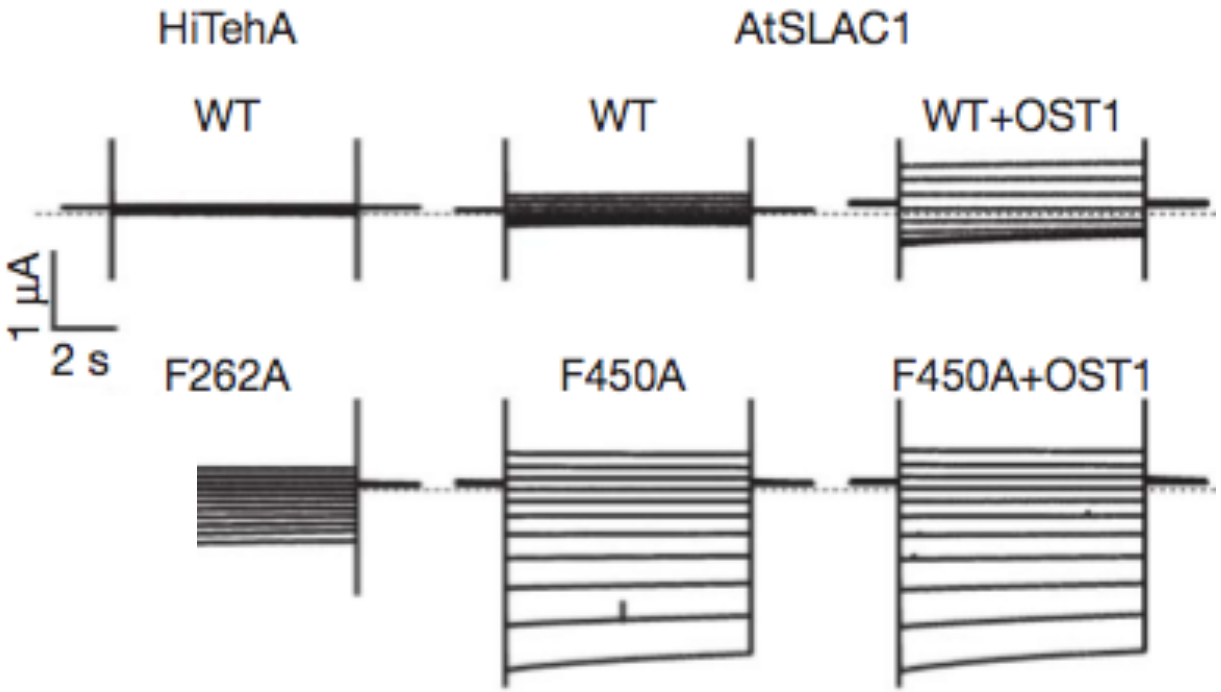




Chen et al. *Nature* 467 (2010)

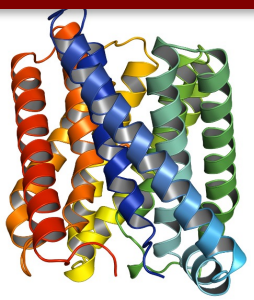


Chen et al. *Nature* 467 (2010)

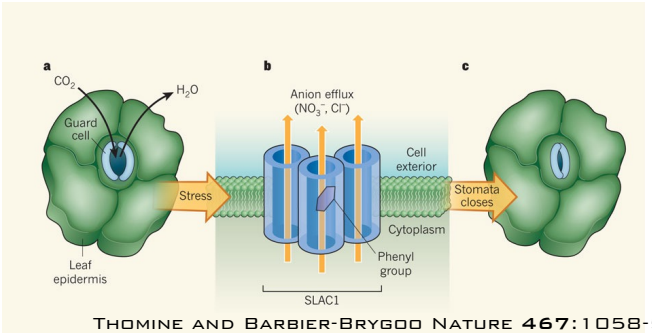


Chen et al. *Nature* 467 (2010)

H. influenzae protein structure

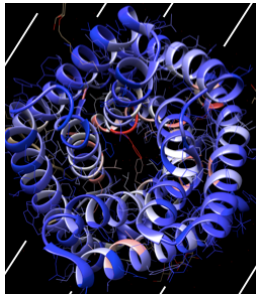


Functional hypothesis via homology to SLAC1

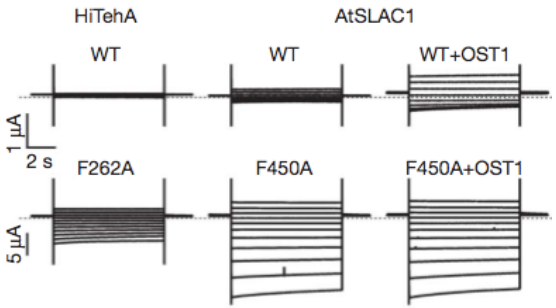


THOMINE AND BARBIER-BRYGOO NATURE 467:1058-59 (201

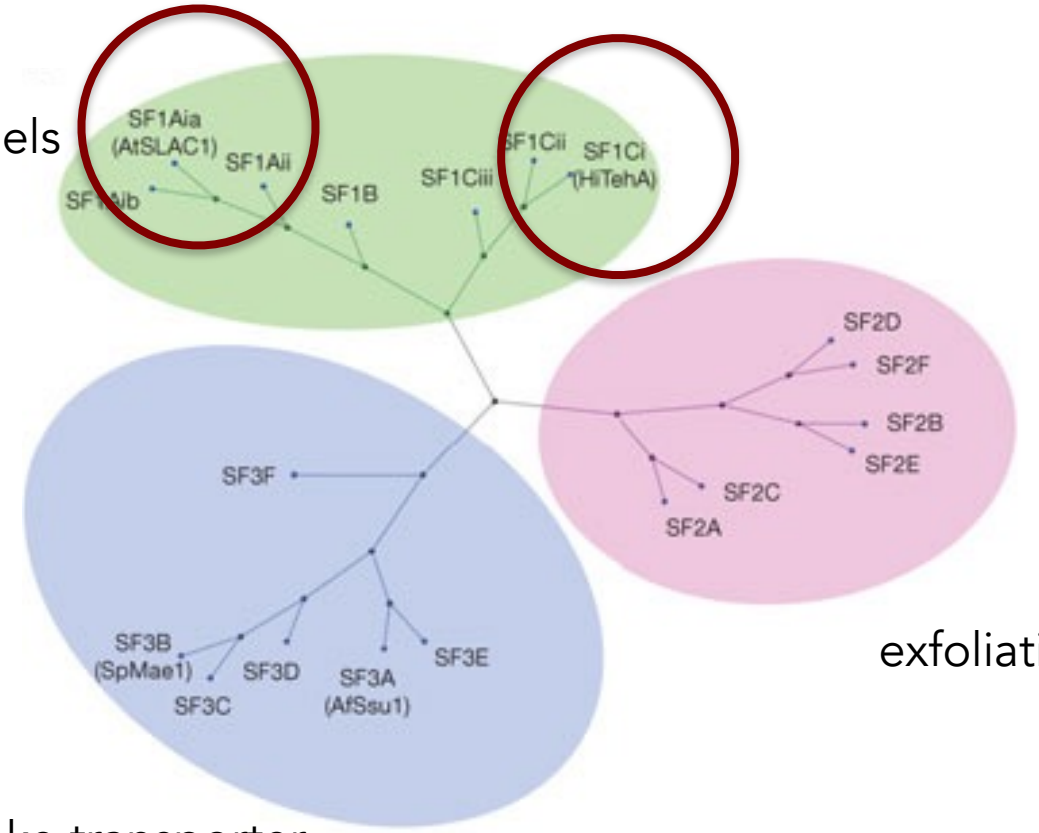
Identification potential functional residues using sequence conservation across the family and structural knowledge



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

```
> more Exercise_2/blast-2-seqs.link
```

>cGMP-gated cation channel alpha-1
MKLSMKNNIINTQQSFVTMPNVIVPDIEKEIRRMENGACSSFSEDDDSASTSEESENENP
HARGSFSYKSLRKGGPSQREQYLPGAIALFNVNNSSNKDQEPEEKKKKKEKKSKSDDKN
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
```

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
```

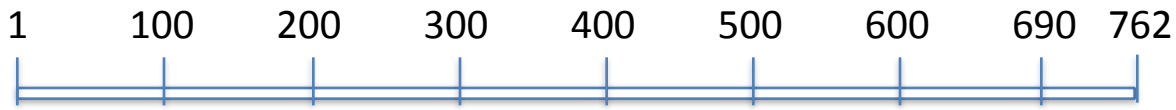
Range 3: 593 to 649 Graphics ▼ Next Match ▲ Previous Match ▲ First Match

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

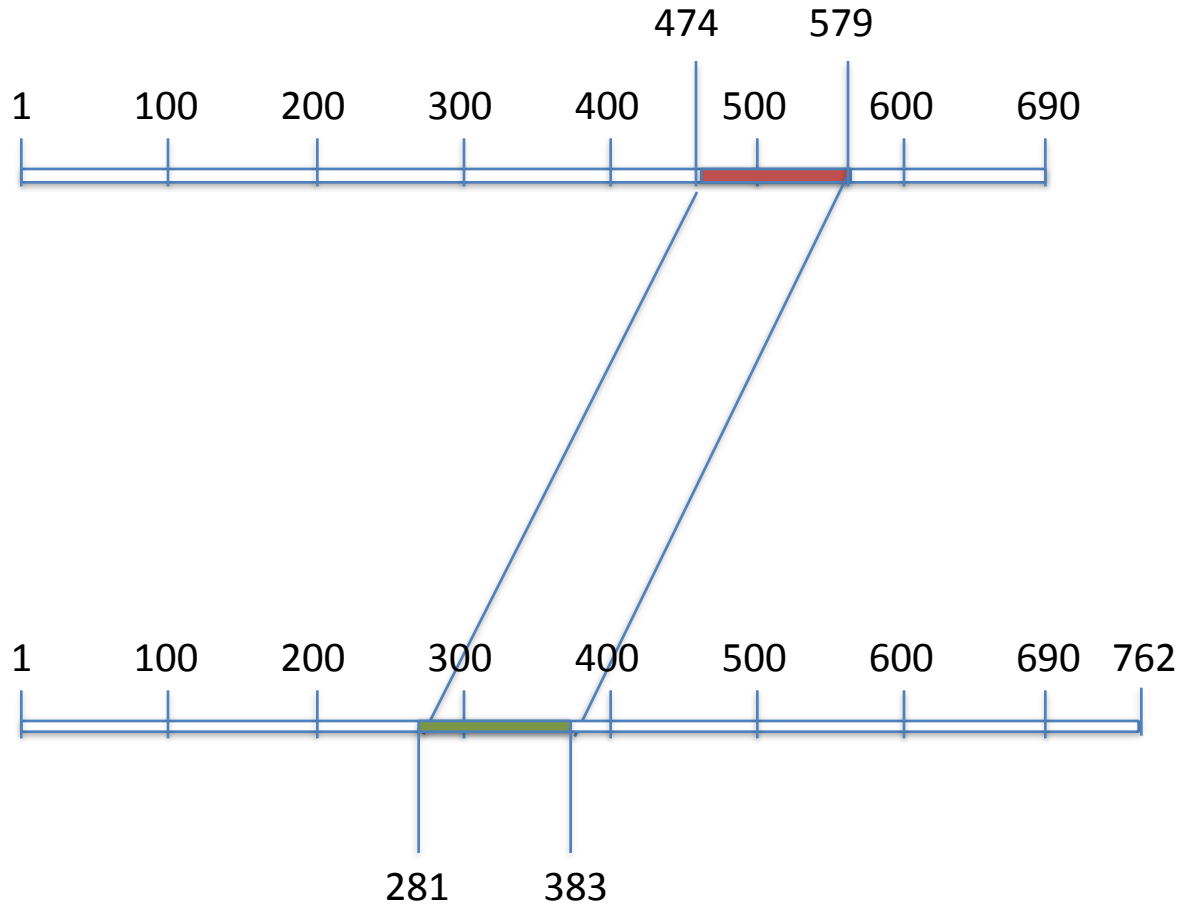
```
Query  317  VFYSISKAIGFGNDTWVY---PDINDPEFGRLARKYVYSL-YWS--TLTLTTIGETPP  368
            V +  +K IG G  TW +   P+   PE   L + + +S+ +WS   L    +   PP
Sbjct  593  VDFGFAKKIGSGQKTWTFCGTPEYVAPEV-ILNKGHDFSVDFWSLGILVYELLTGNPP  649
```

Mystery protein is a cGMP-gated cation channel?

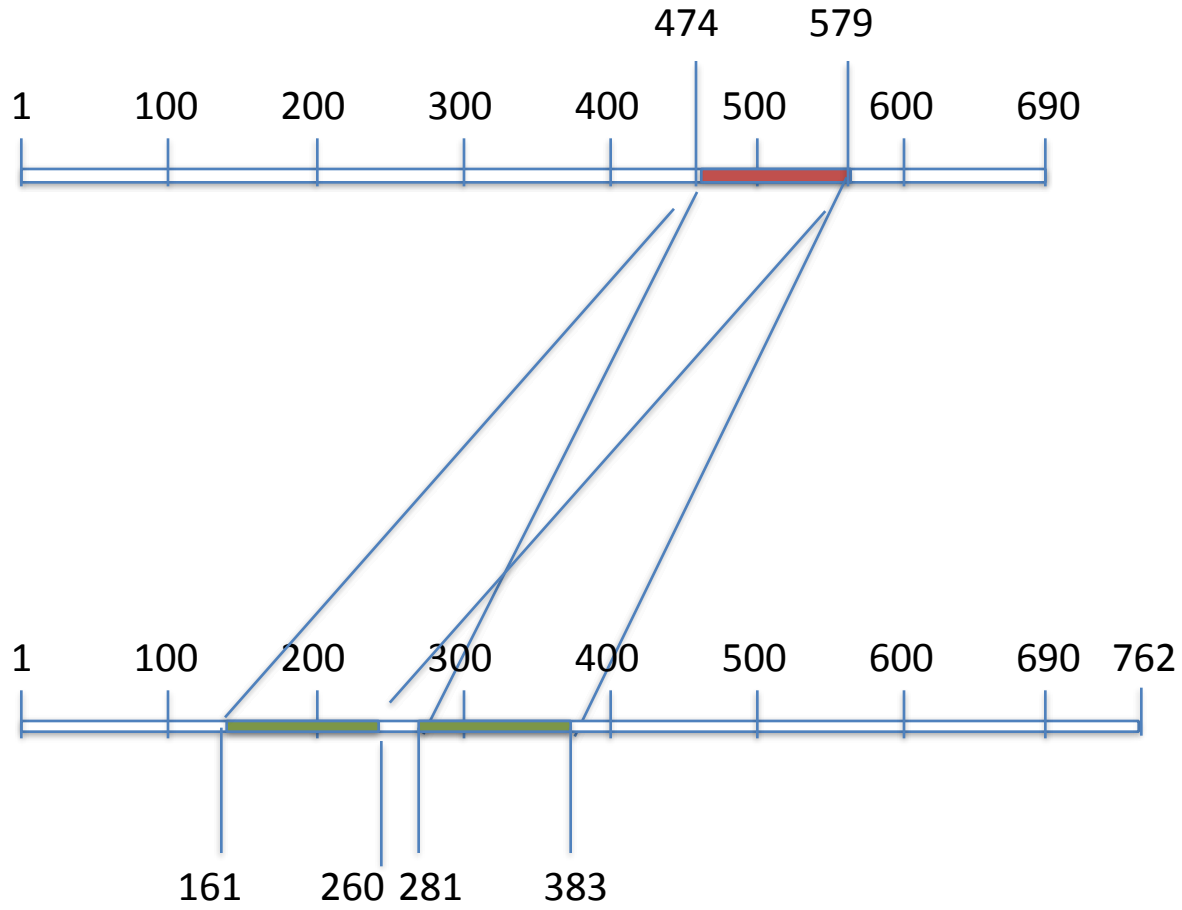
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)



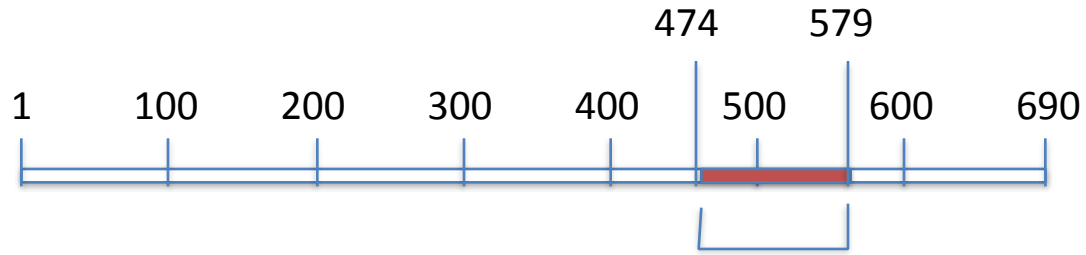
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)



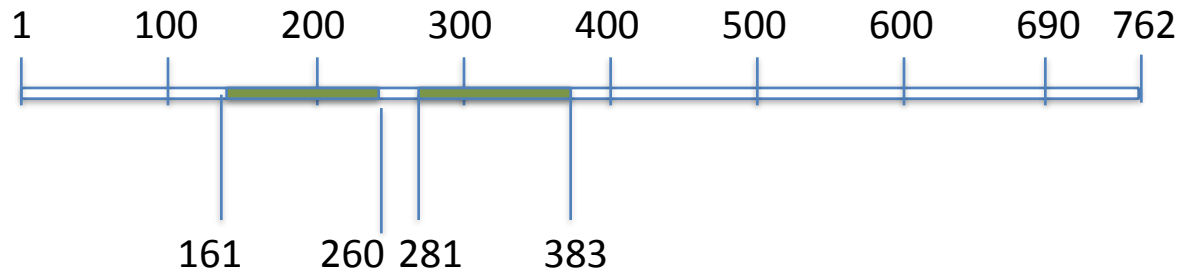
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)



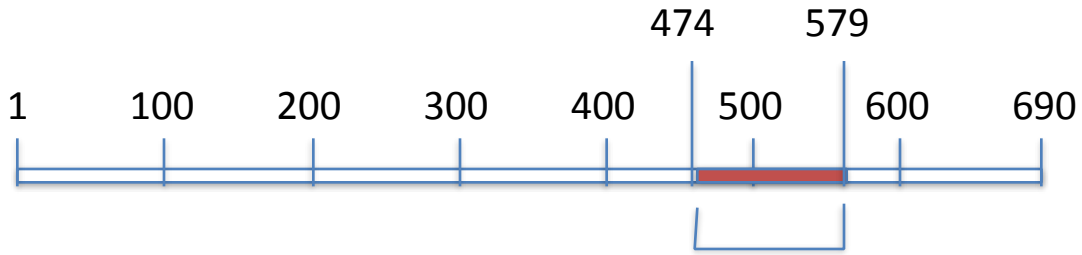
cGMP-gated cation channel alpha-1 P29973 (CNGA1_HUMAN)



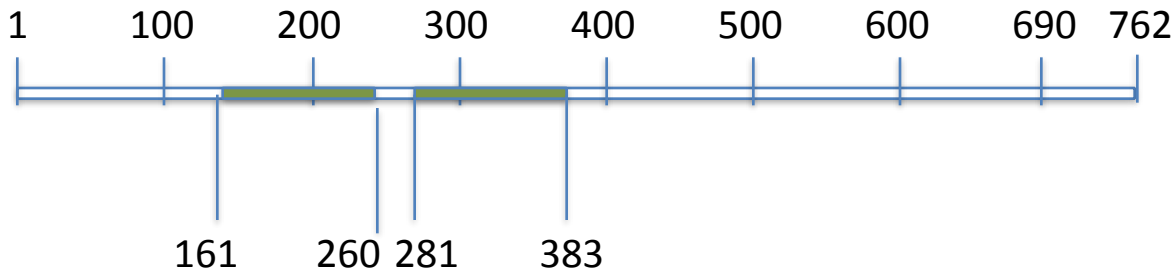
cGMP binding domain



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



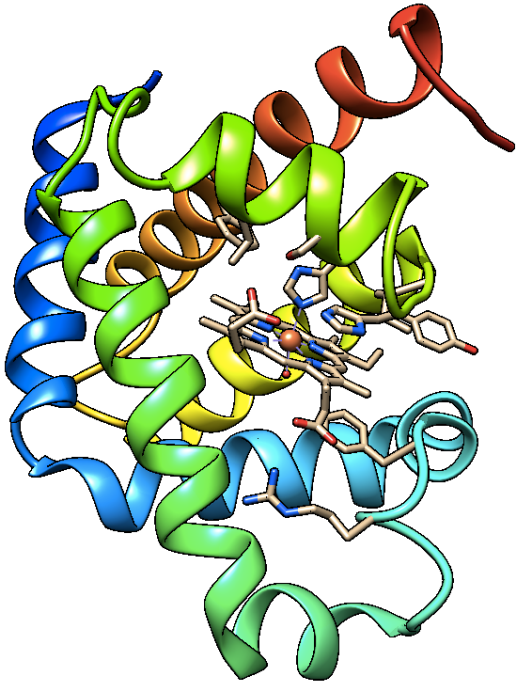
cGMP binding domain



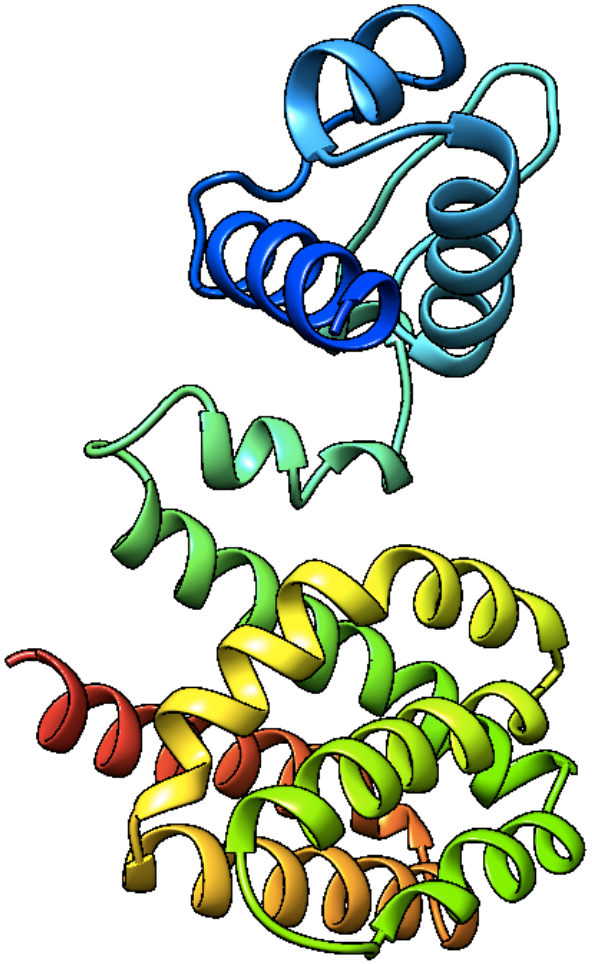
cGMP binding domains?

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

Marco Punta



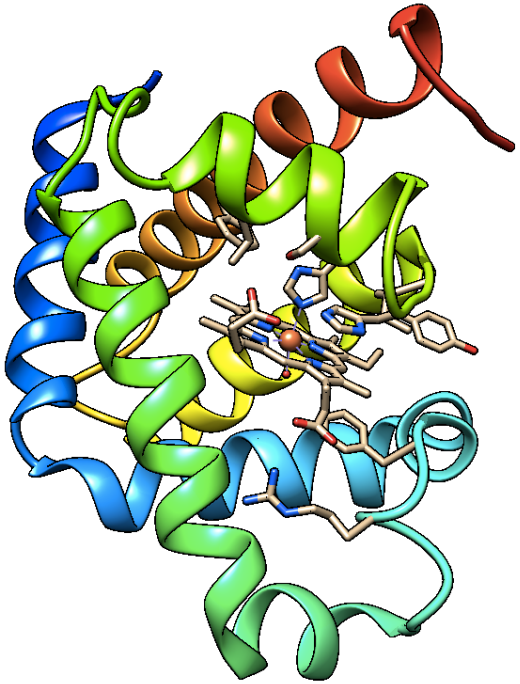
1MBN



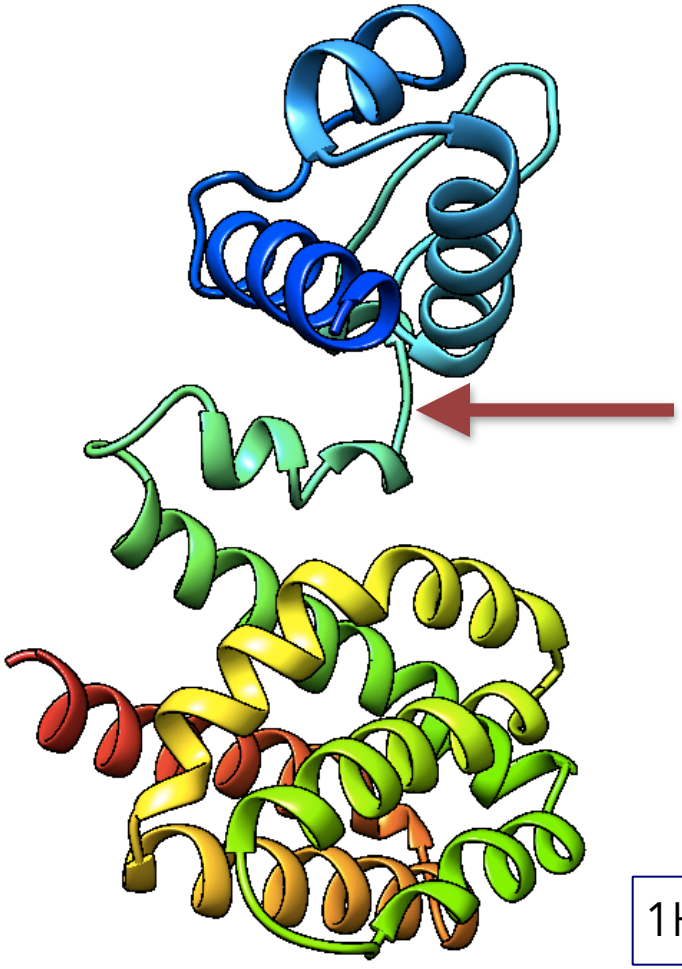
1HW2

Colour Scheme:

N’ **C’**



1MBN

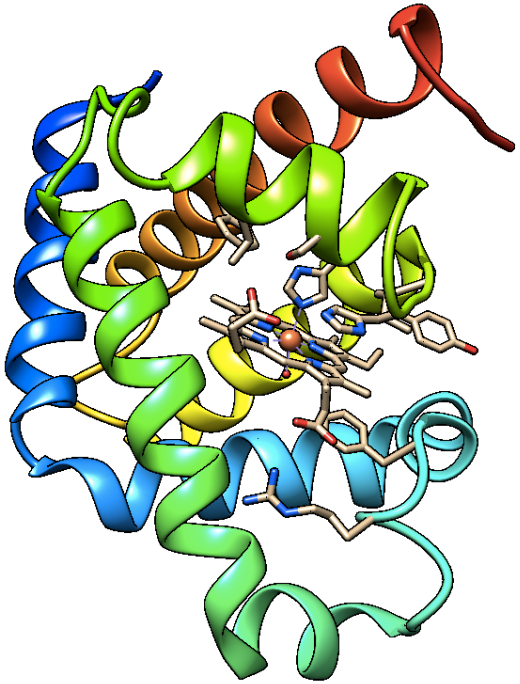


1HW2

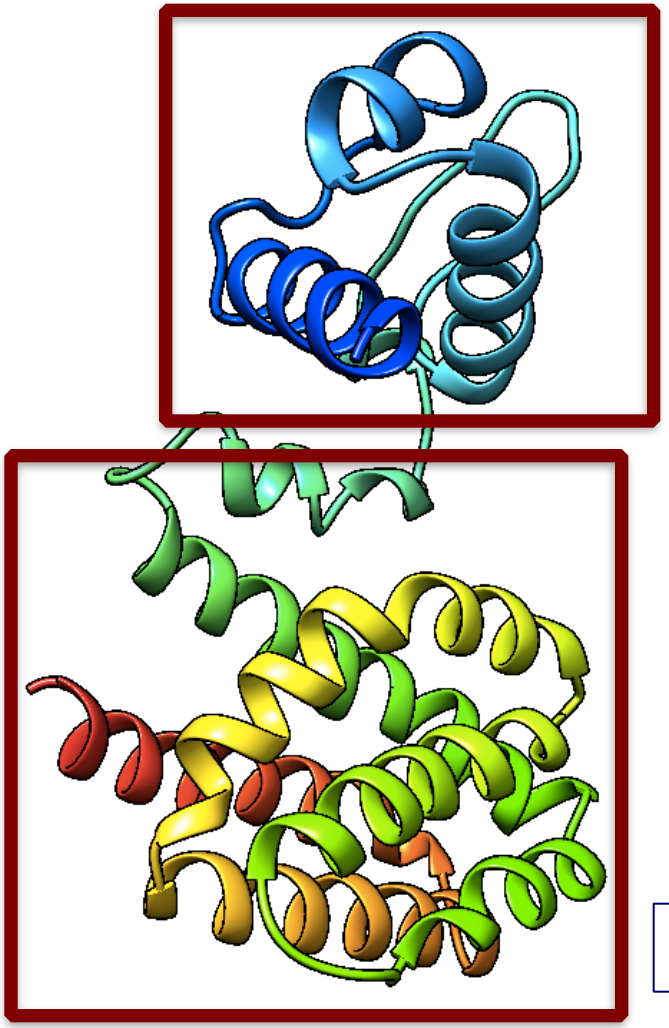
Colour Scheme:

N’ **C’**

Marco Punta



1MBN

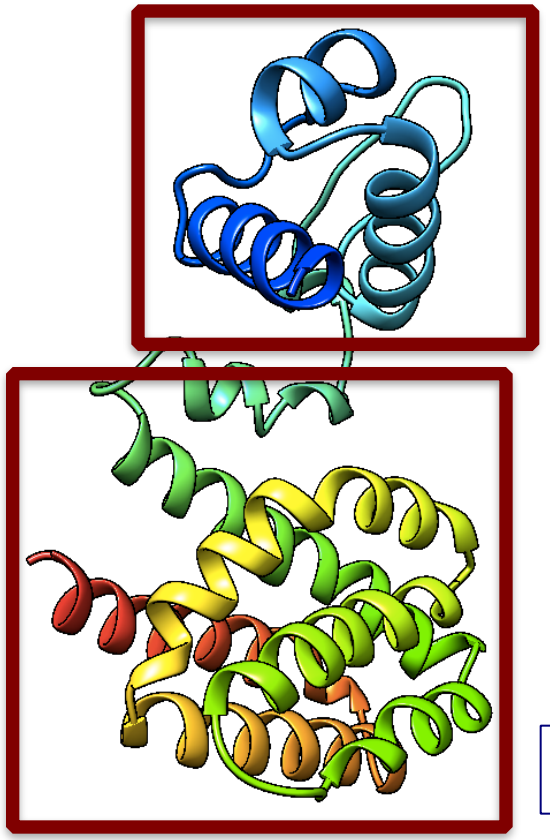


1HW2

Colour Scheme:

N' **C'**

Definition (Wikipedia):
A protein domain is a conserved part of a given protein sequence and structure that can evolve, function, and exist independently of the rest of the protein chain. A domain forms a compact three-dimensional structure and often can be independently stable and folded.
(Marco): in proteins individual domains can be combined to perform complex functions.



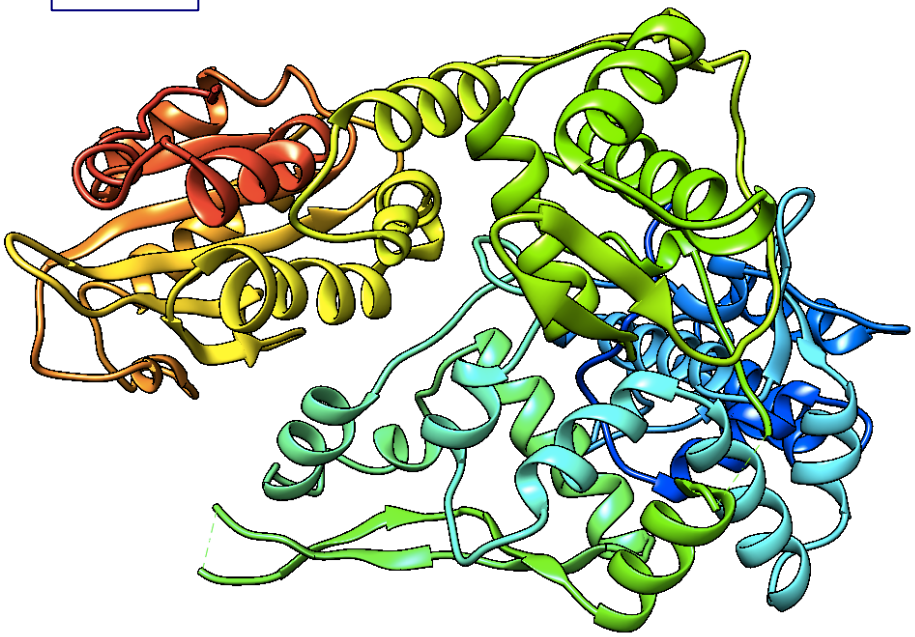
1HW2

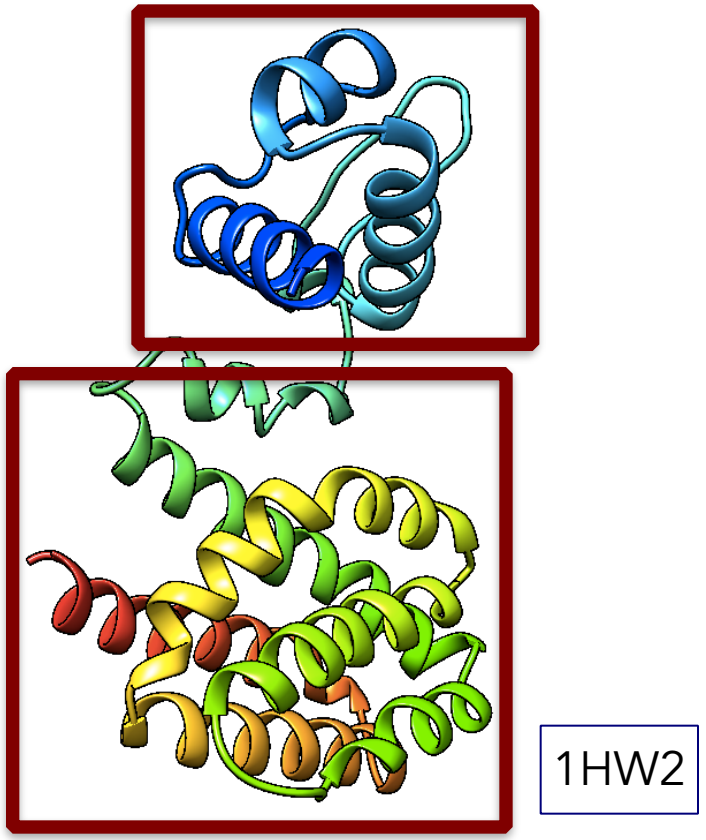
Colour Scheme:

N’ **C’**

Marco Punta

1FOK





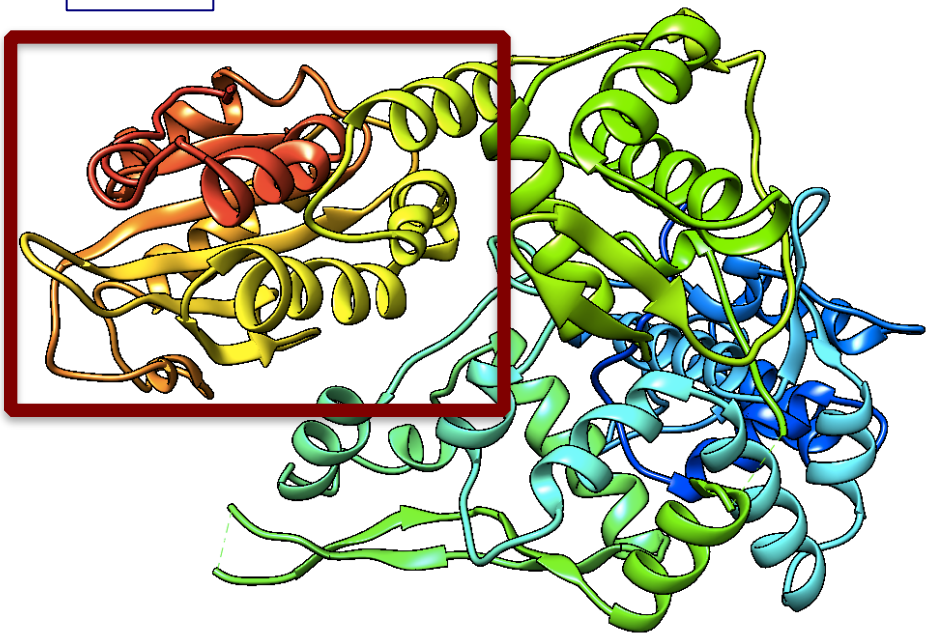
1HW2

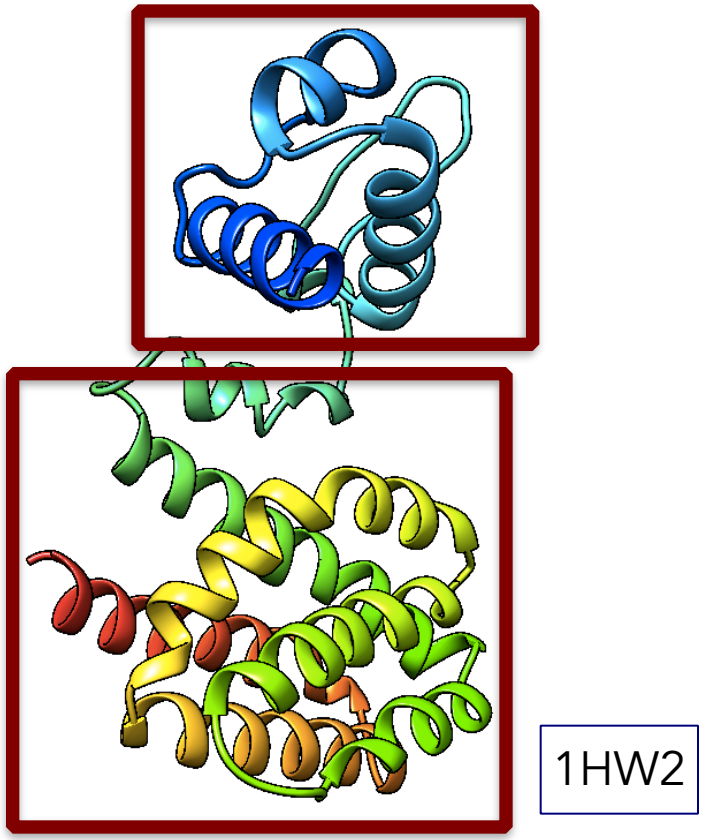
Colour Scheme:

N’ **C’**

Marco Punta

1FOK





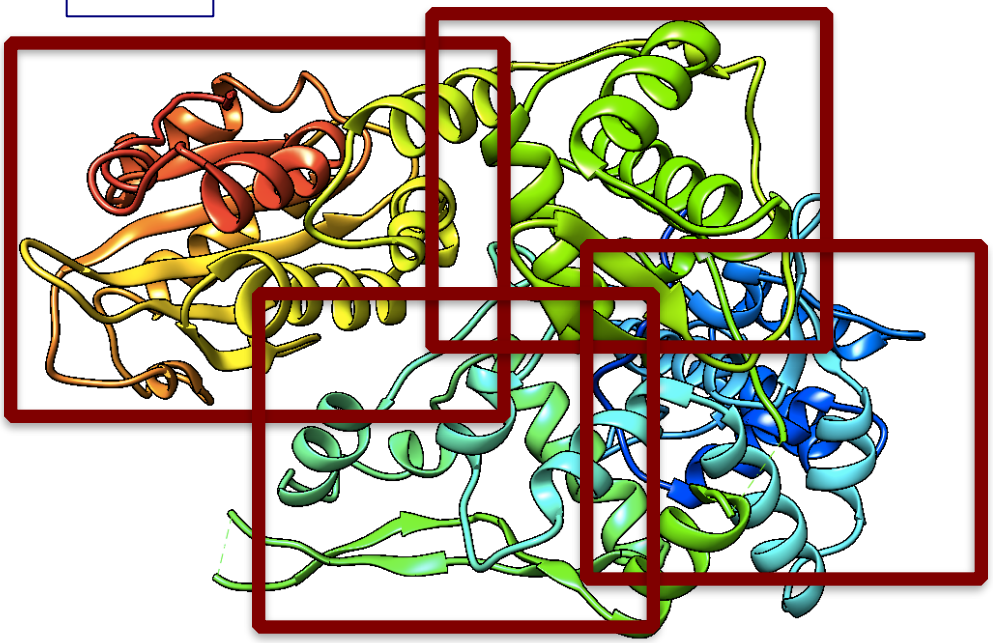
1HW2

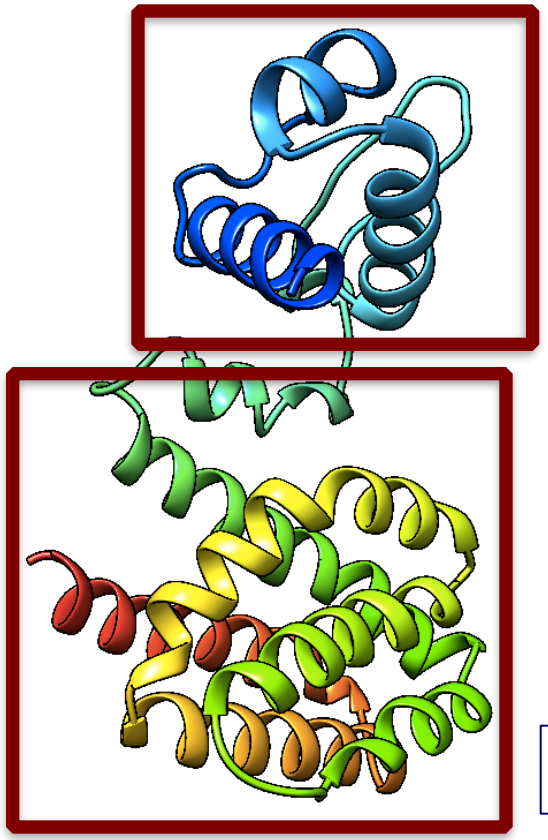
Colour Scheme:

N' **C'**

Marco Punta

1FOK



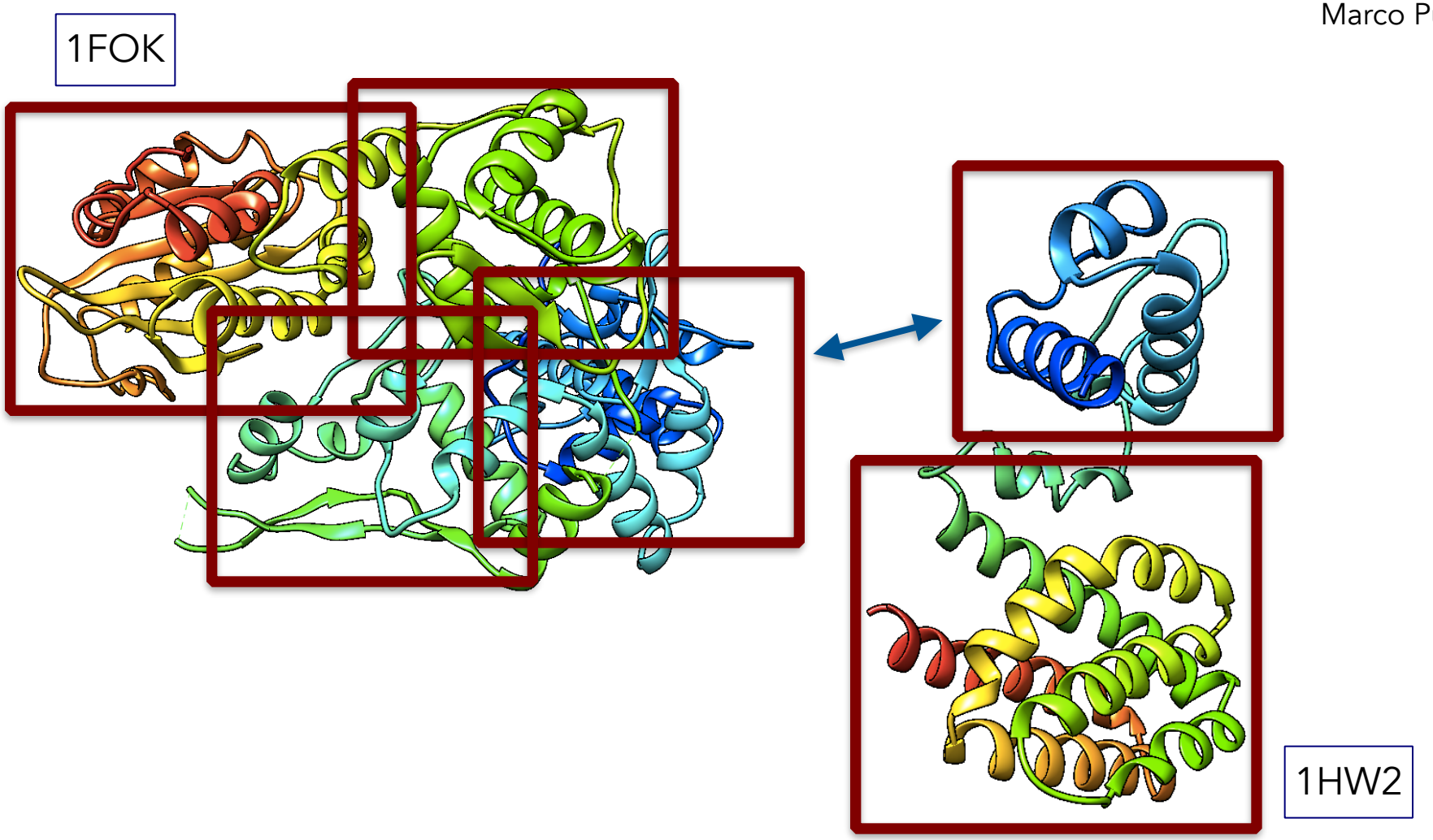


1HW2

Colour Scheme:

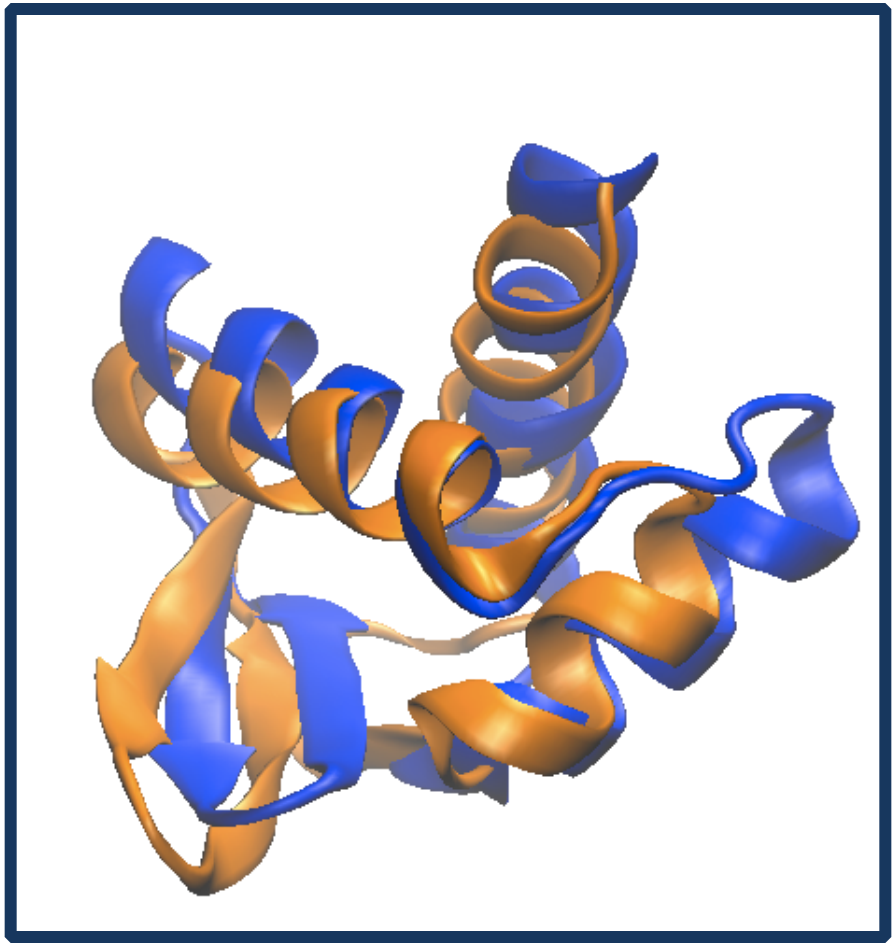
N' **C'**

Marco Punta



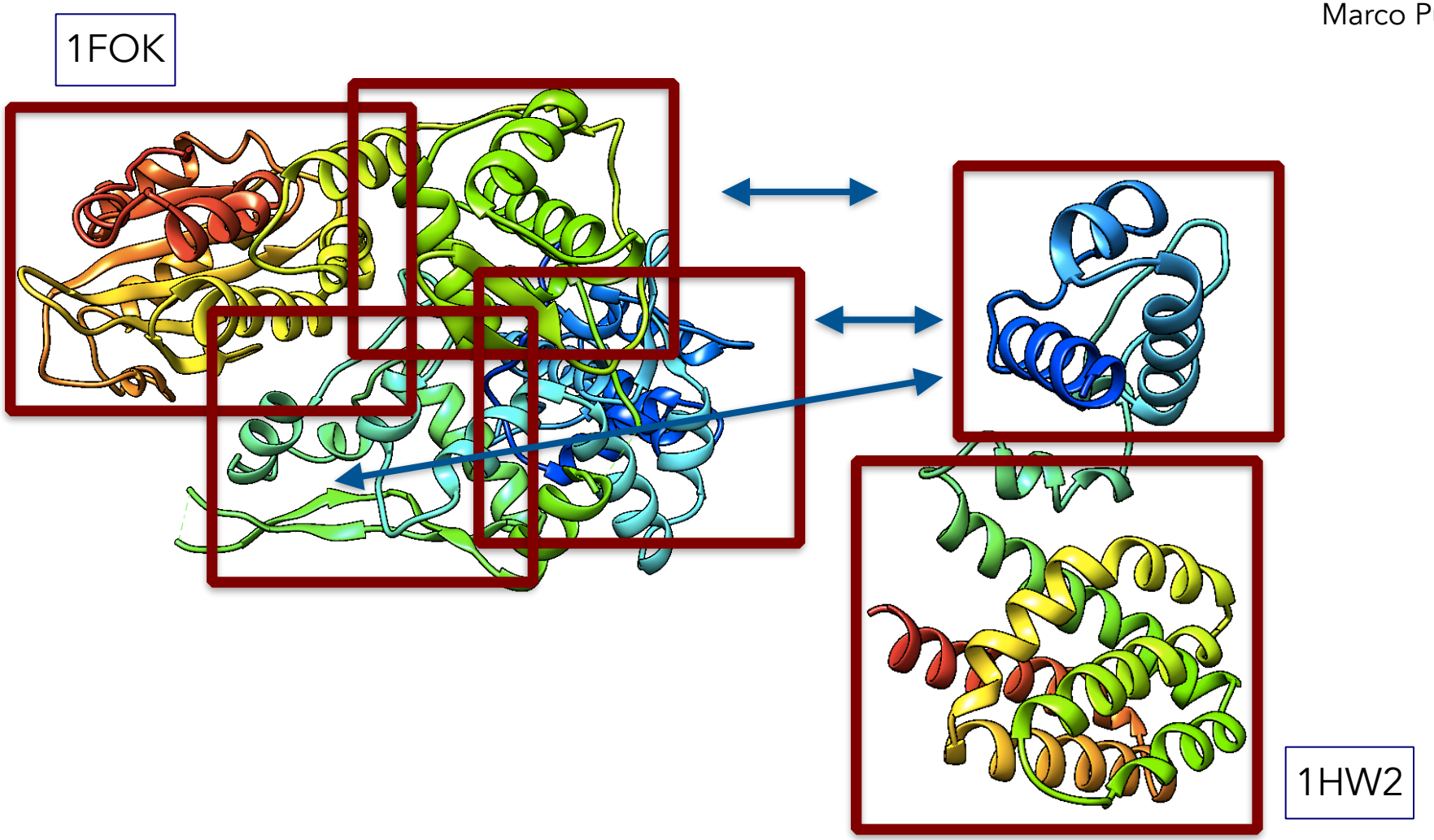
Colour Scheme:

Winged helix domain (WHD)

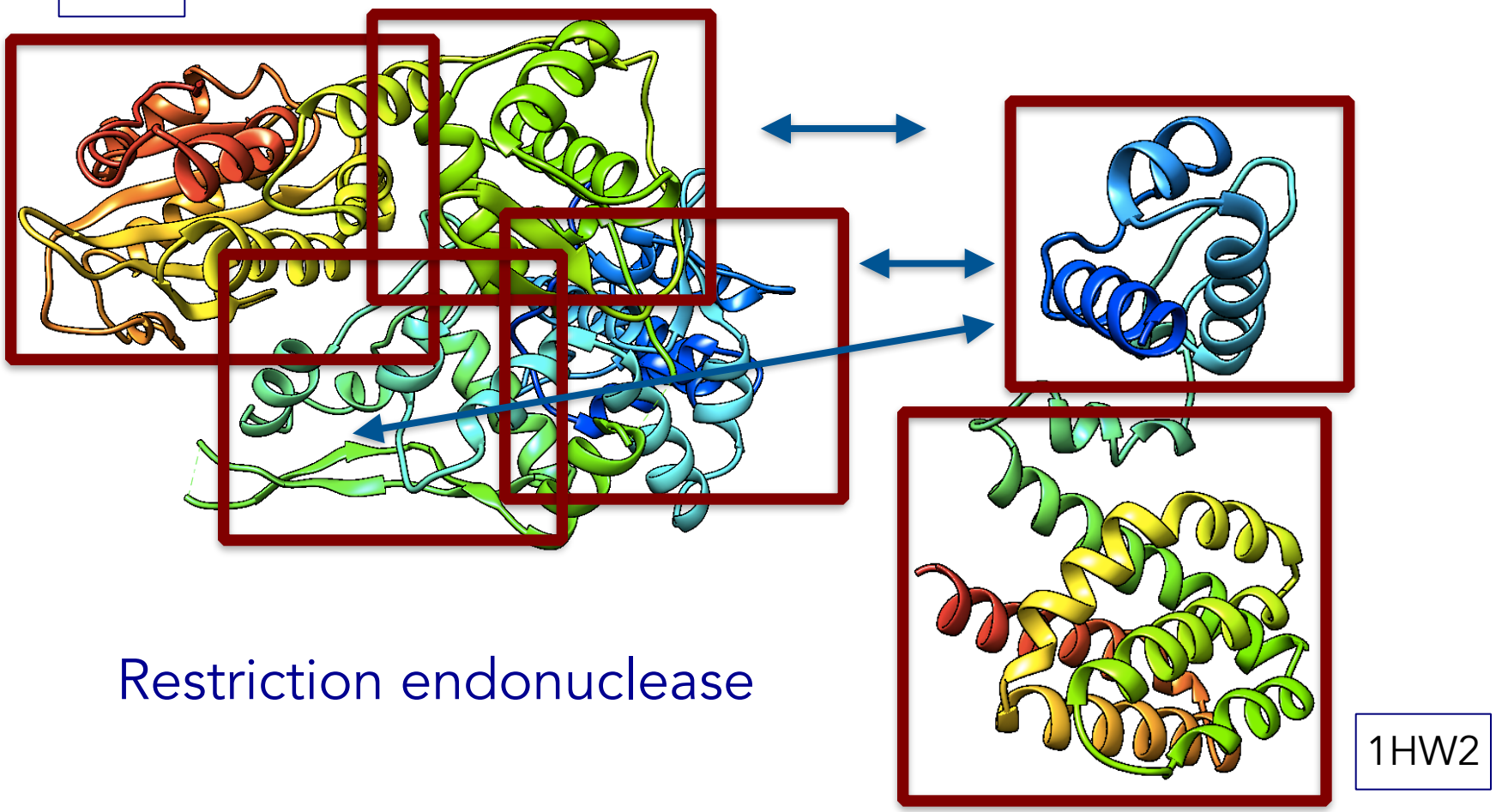


Z-score = 4.0
%id = 8%
RMSD = 2.7Å

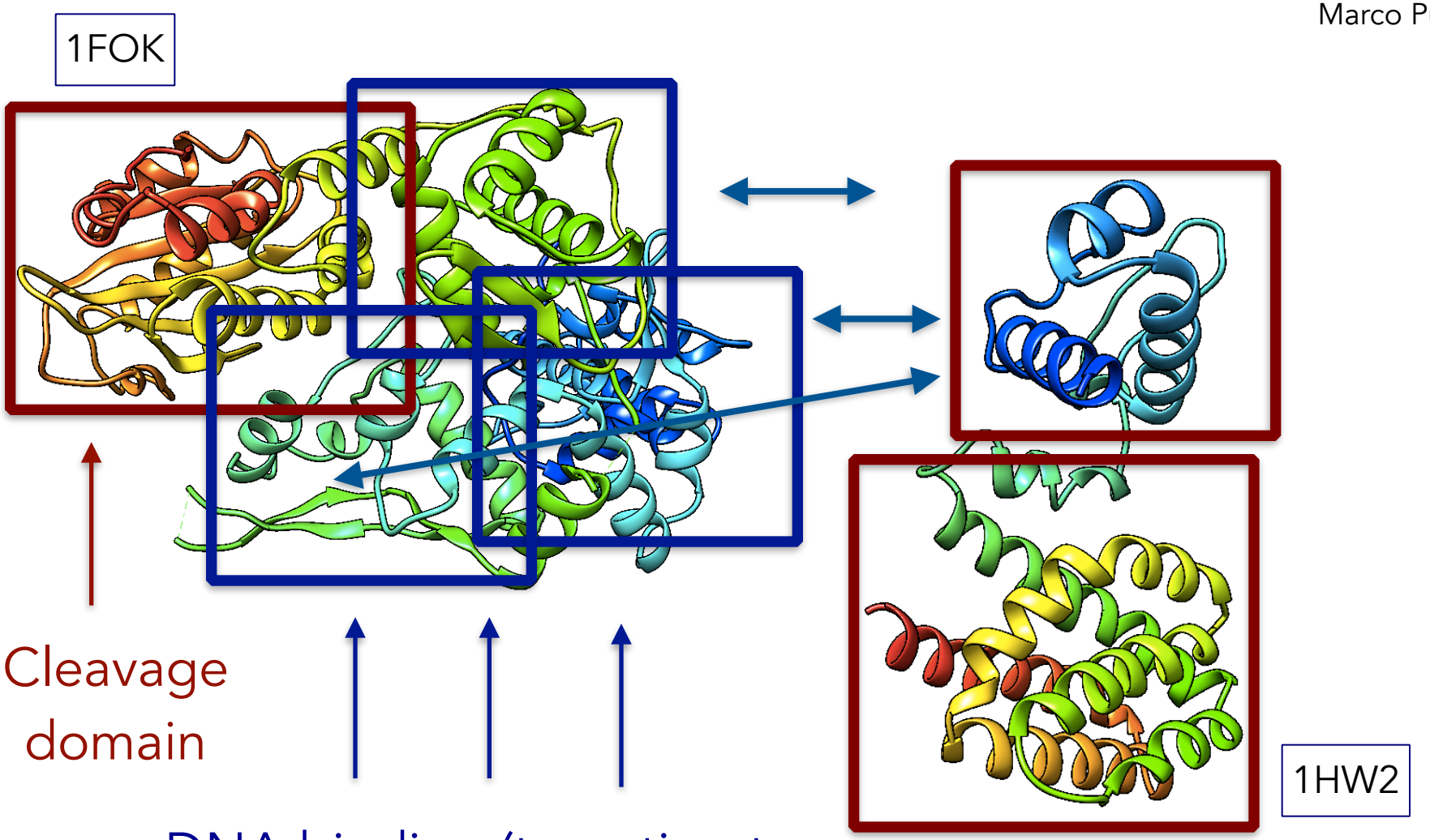
Marco Punta

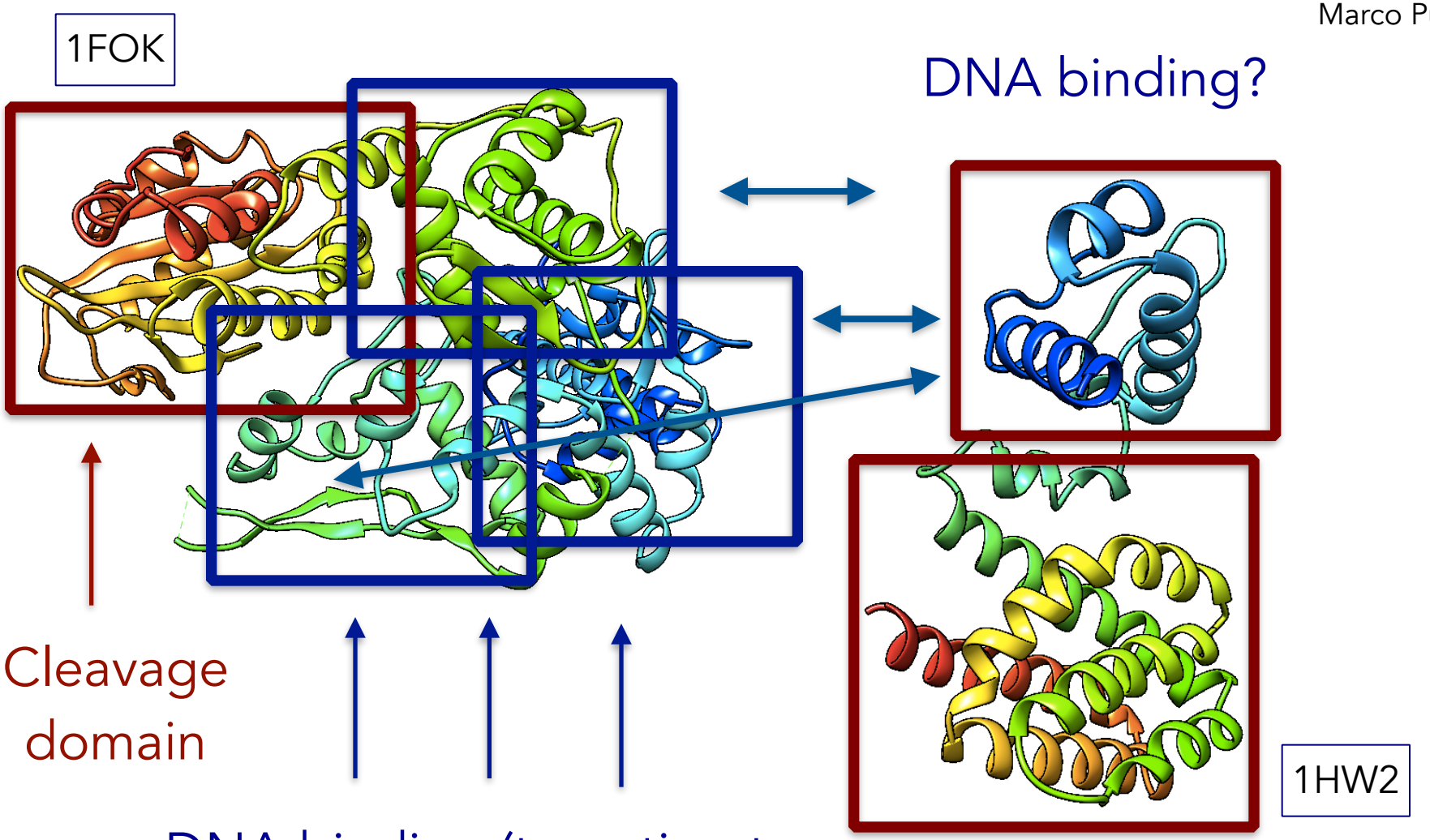


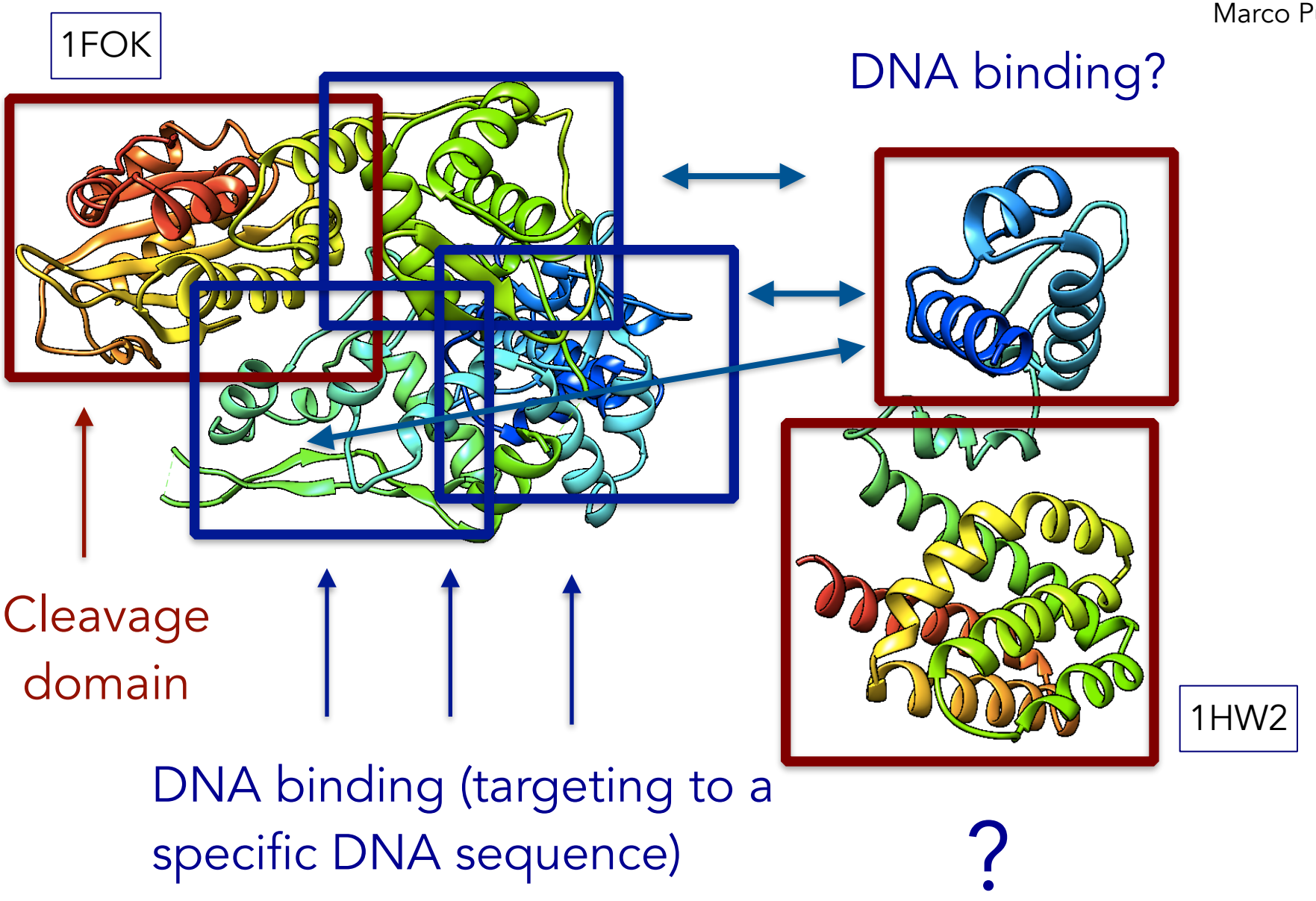
Colour Scheme:

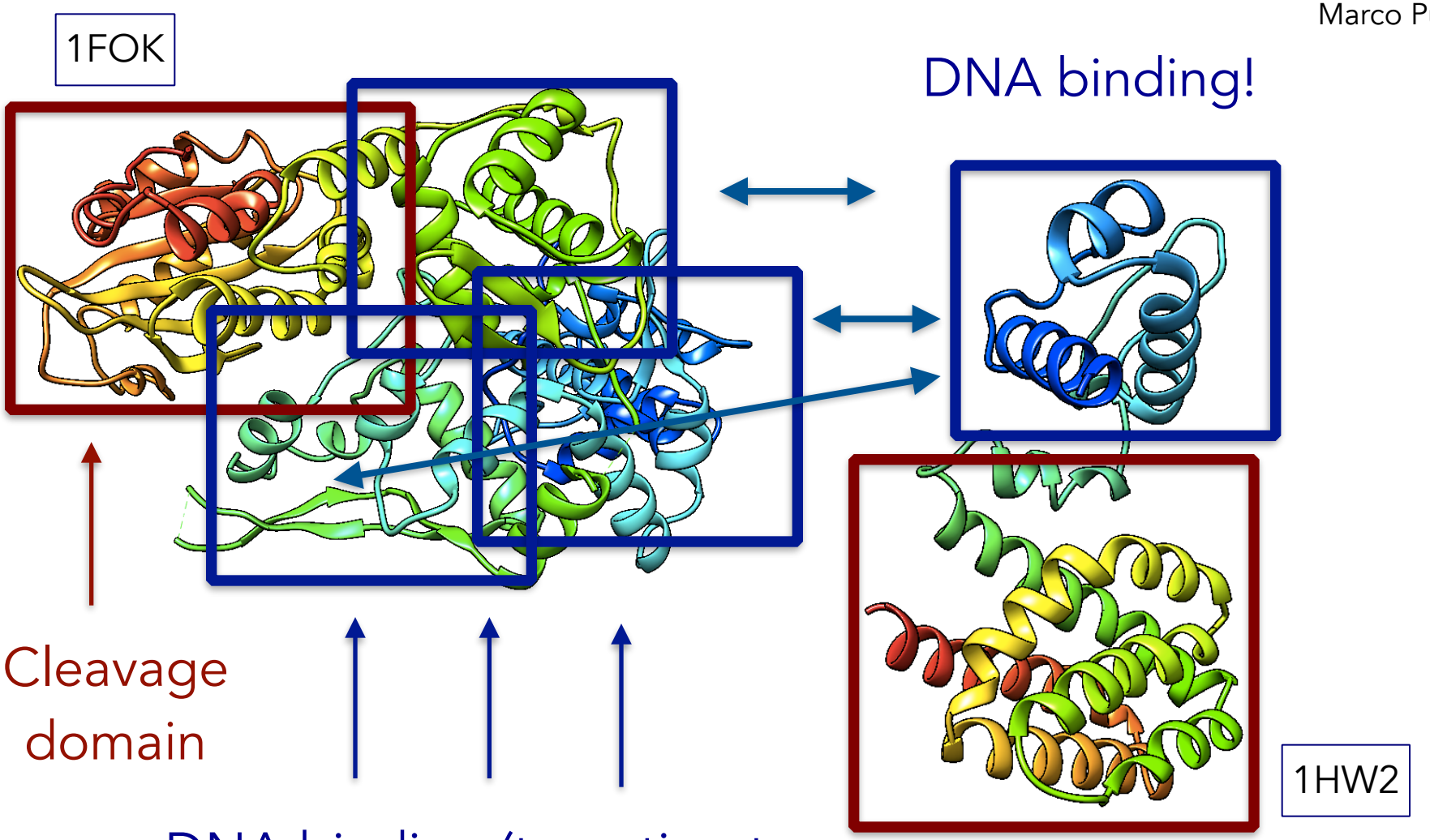


Restriction endonuclease


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

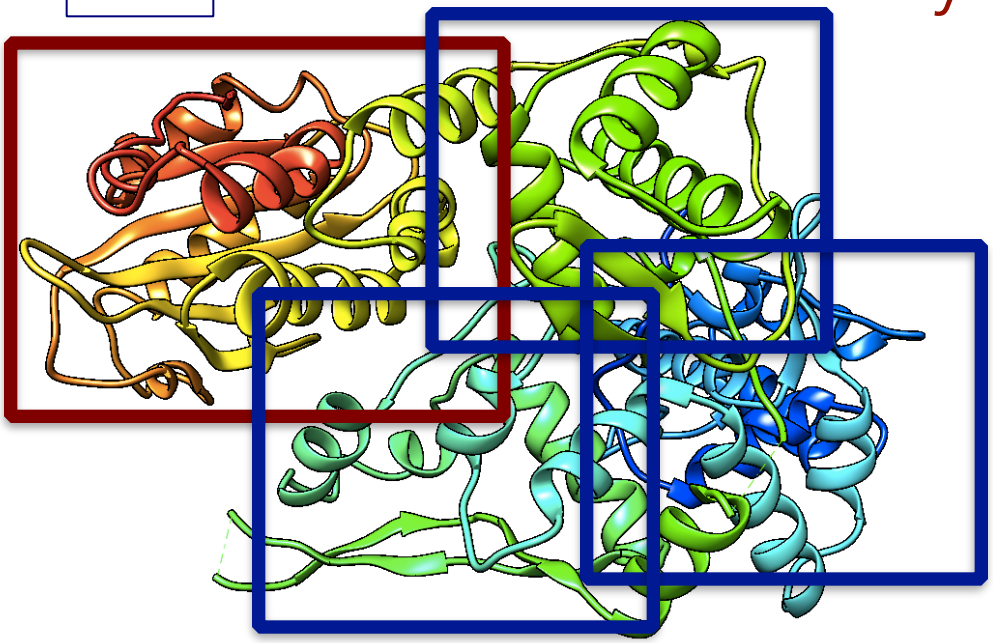

1FOK
DNA binding?
Cleavage domain
DNA binding (targeting to a specific DNA sequence)
1HW2


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

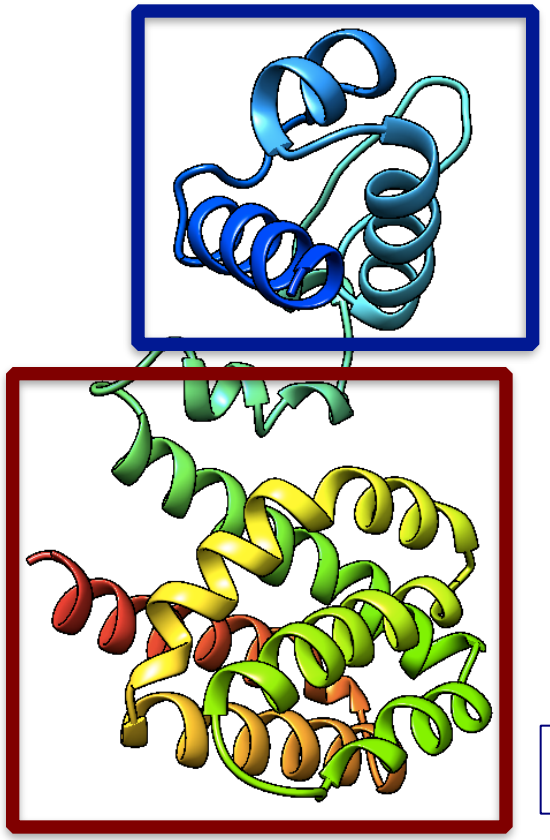

1FOK
DNA binding!
Cleavage domain
DNA binding (targeting to a specific DNA sequence)
1HW2
acyl-CoA binding domain controls affinity

1FOK

"syntactical change"



Restriction endonuclease

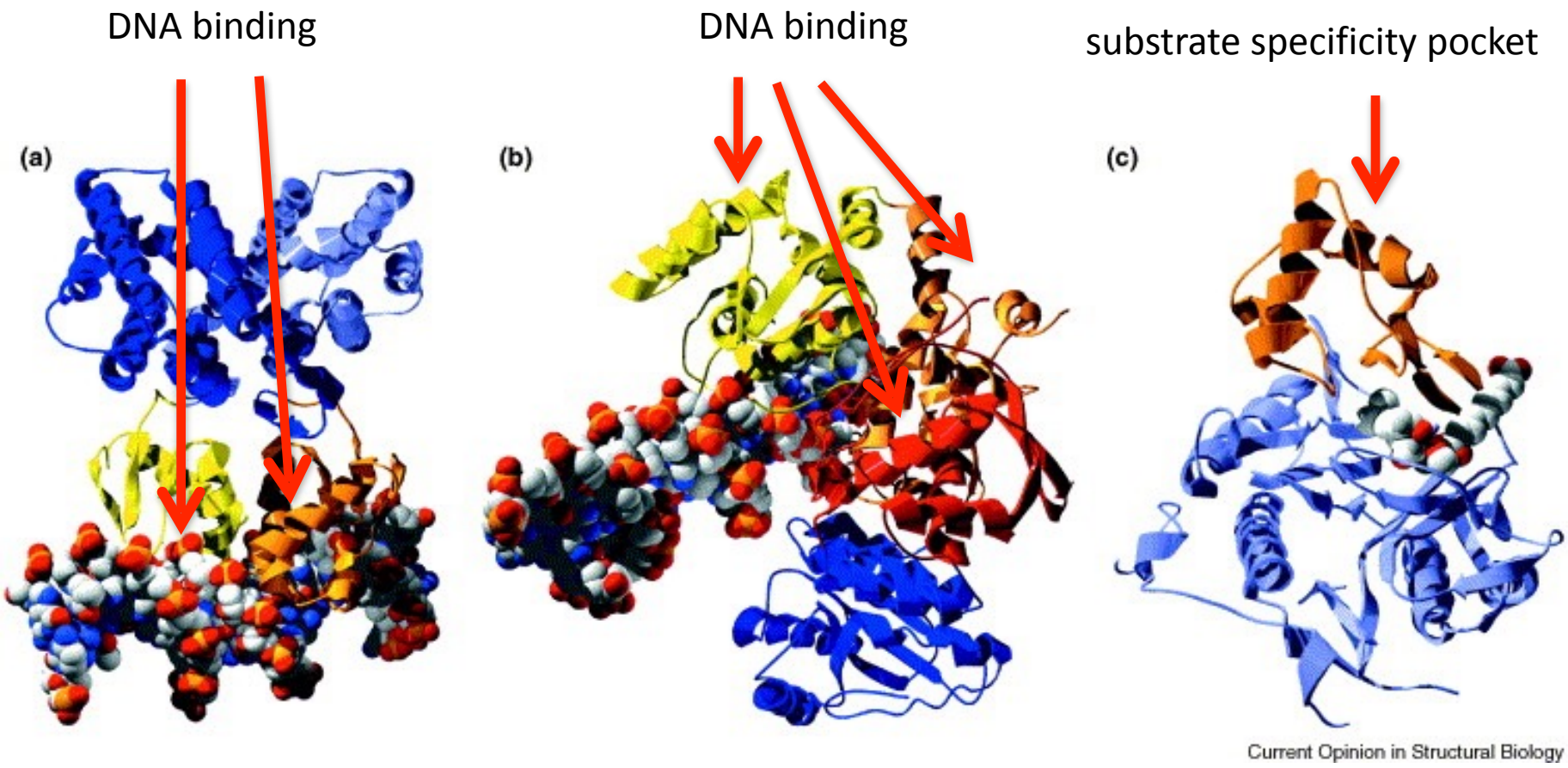


1HW2

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

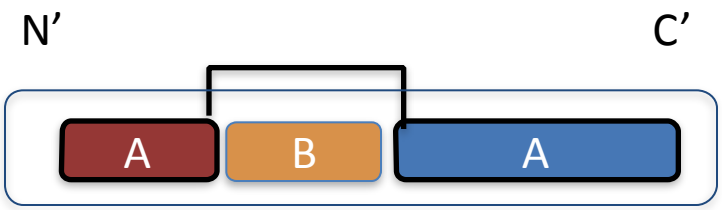
Restriction endonuclease

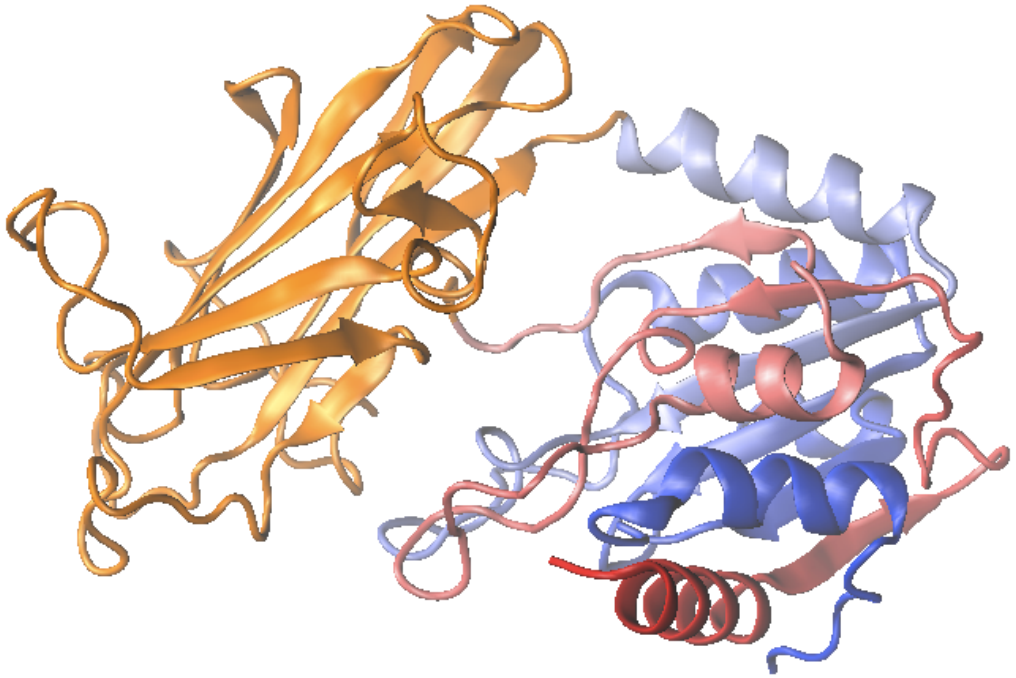
5′-GGATG-3′

Transcription factor

5′-TGGNNNNNCCA-3′

"Nested" domains





3ABZ-truncated

Function annotation transfer by homology

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

- functional drift can lead to radically different functions

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

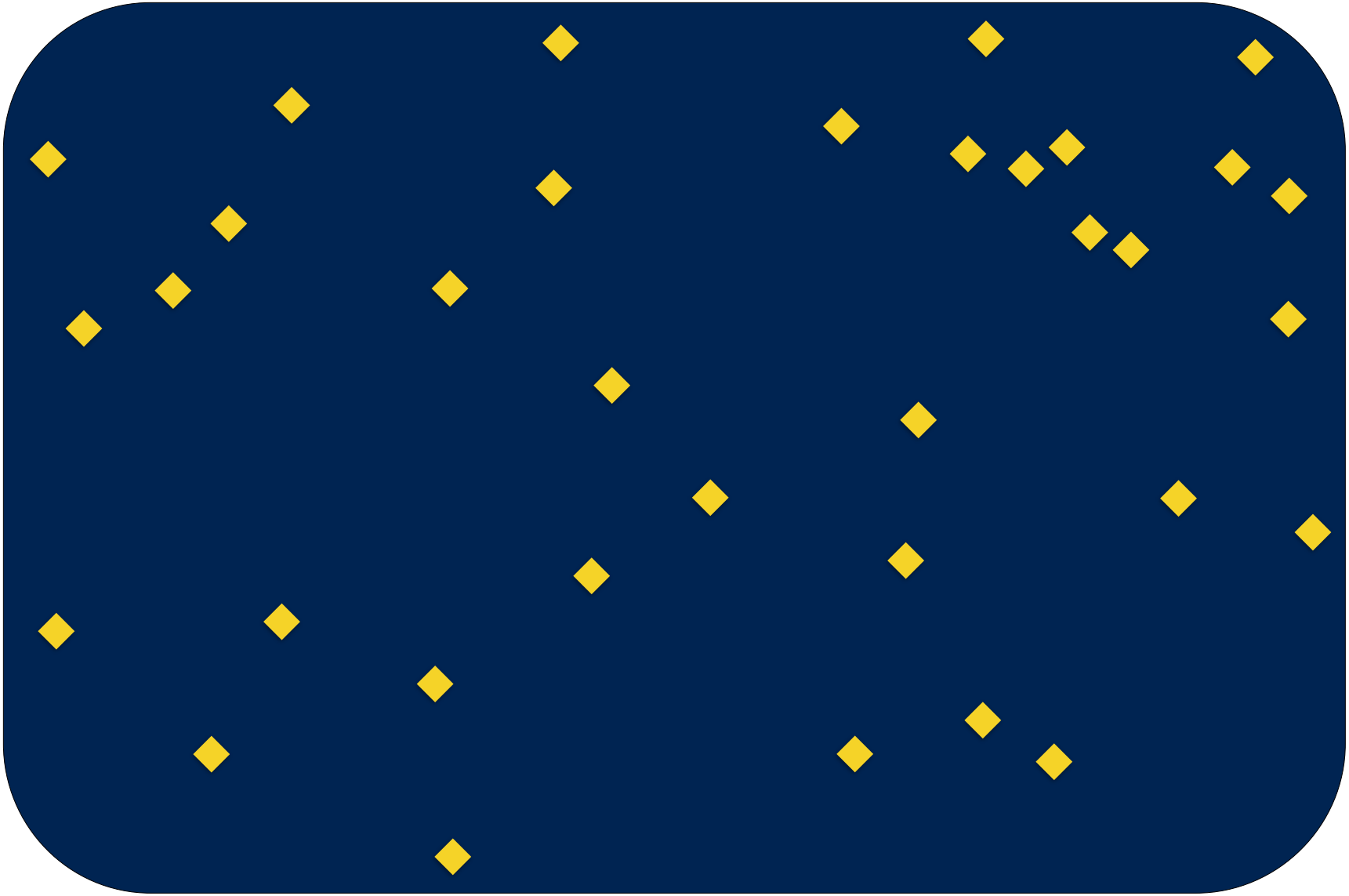
- if more than one functional domain is present annotation transfer can be attempted only between domains that are homologous and NOT for the full-length protein function

Protein families

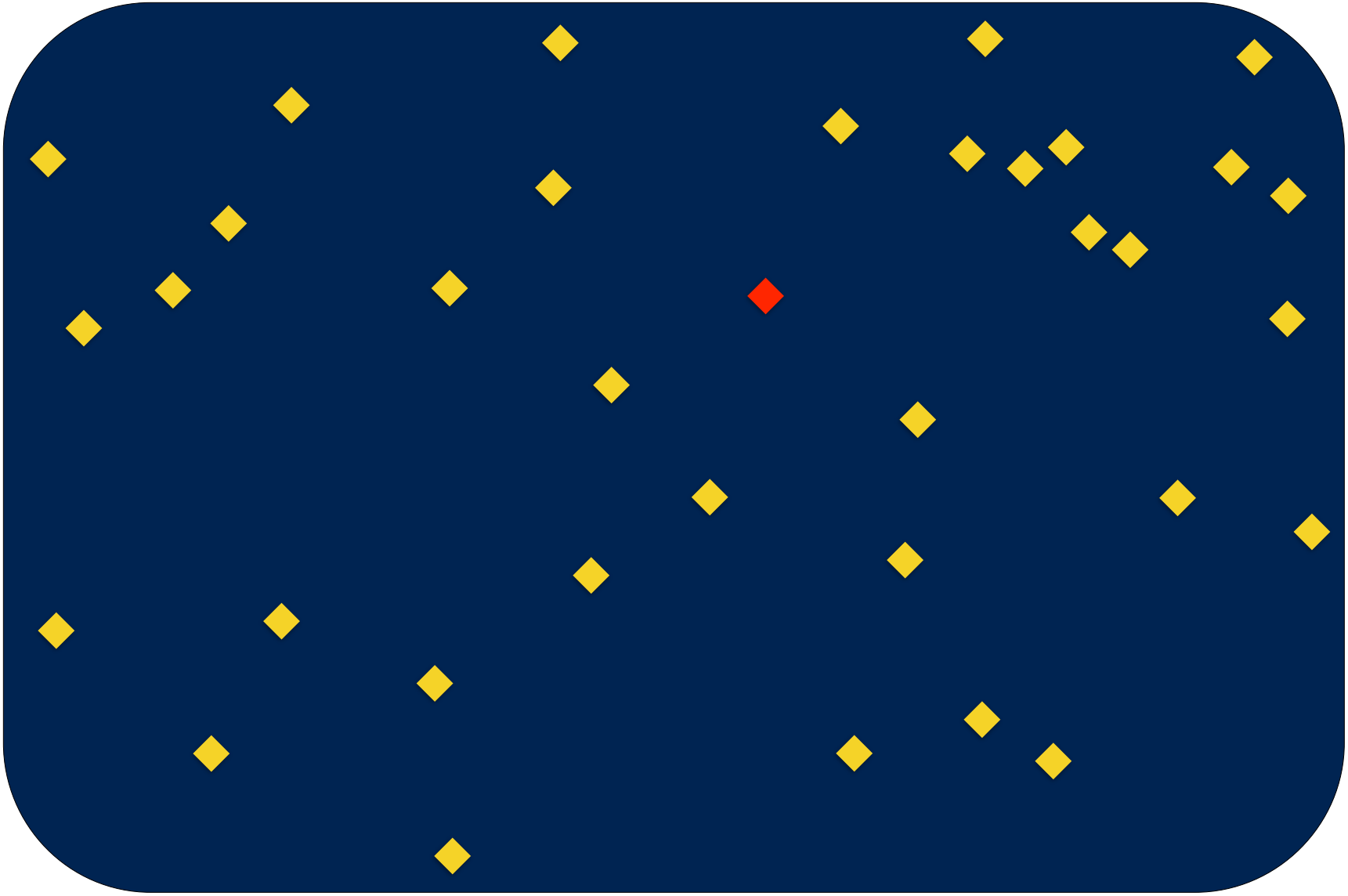
- Members will be structurally similar
- Members may share aspects of function
- Also, the whole set of members may reveal elements of protein and organism evolution

The sequence space

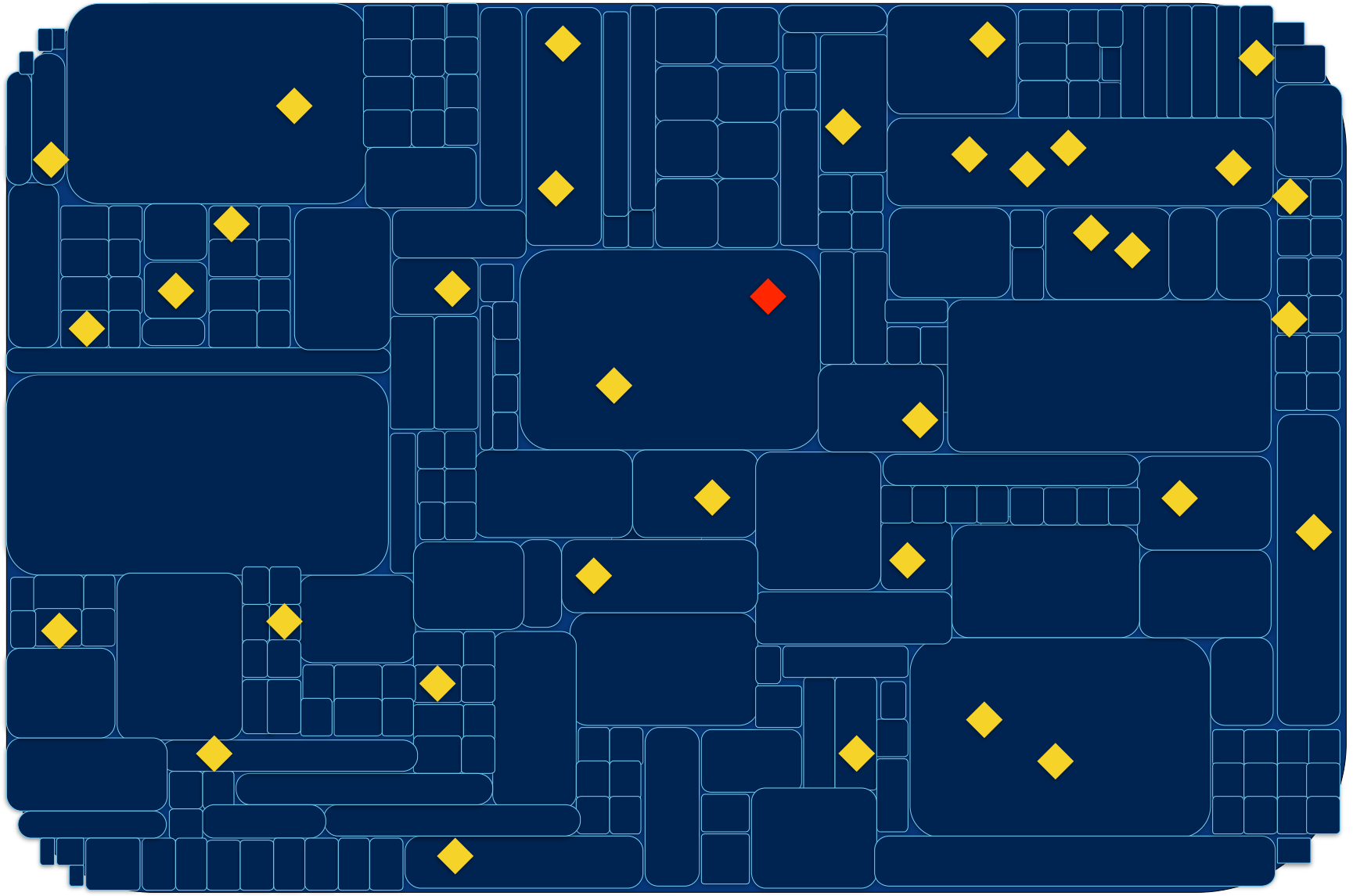
The sequence space and annotated proteins



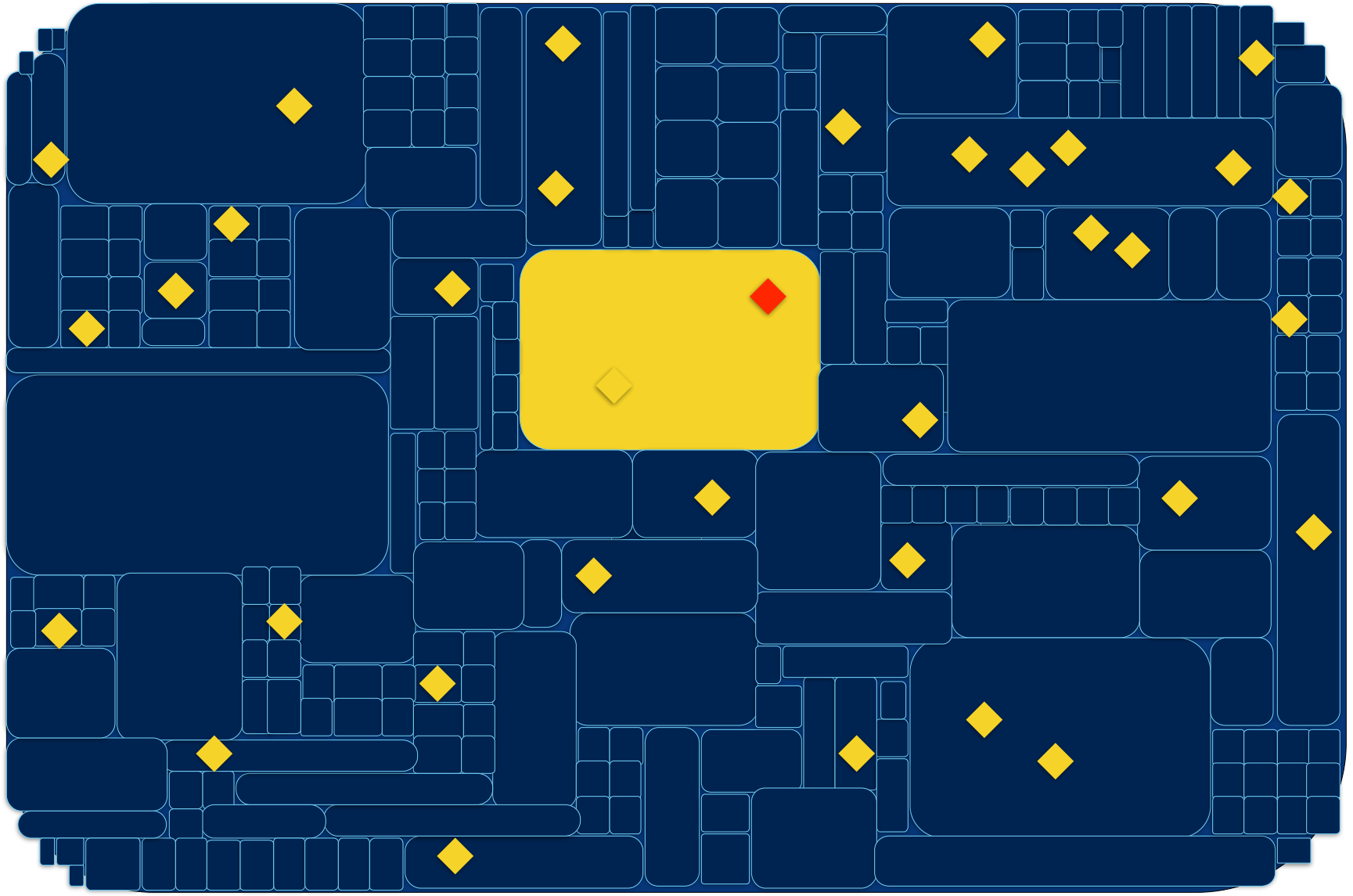
The sequence space and annotated proteins



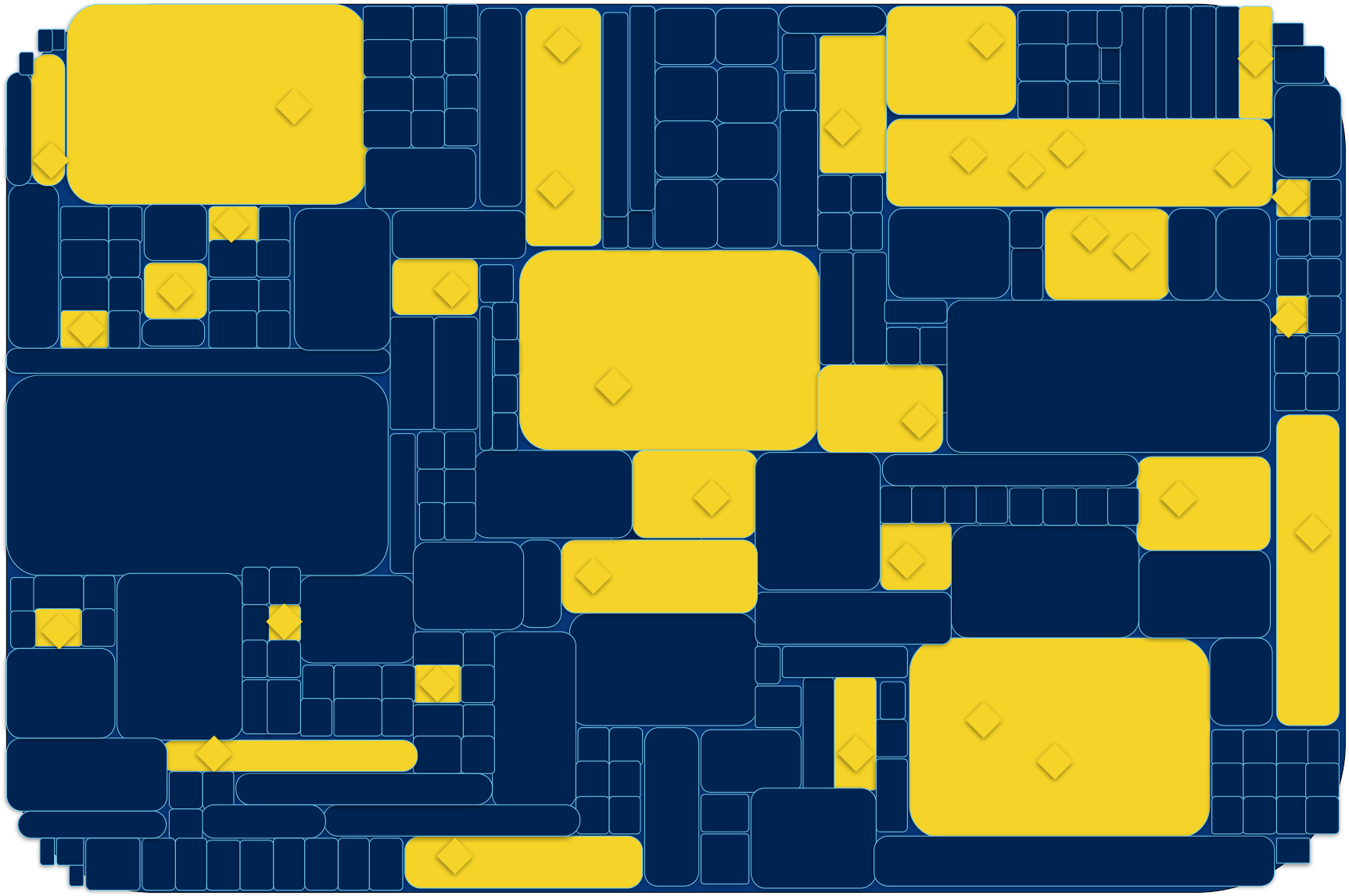
Protein families



Annotation transfer by homology



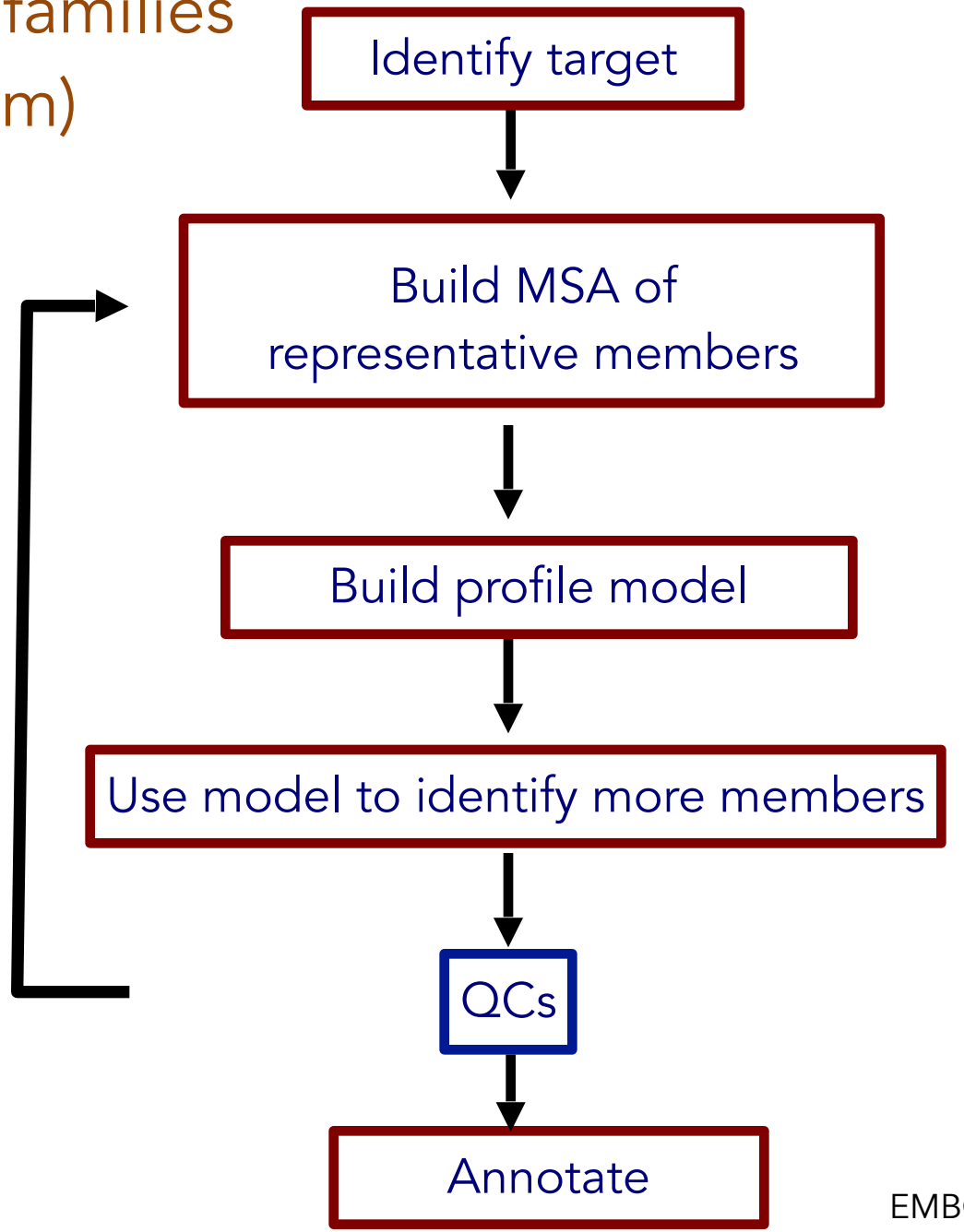
Annotation transfer by homology



Protein family databases

Building families (Pfam)

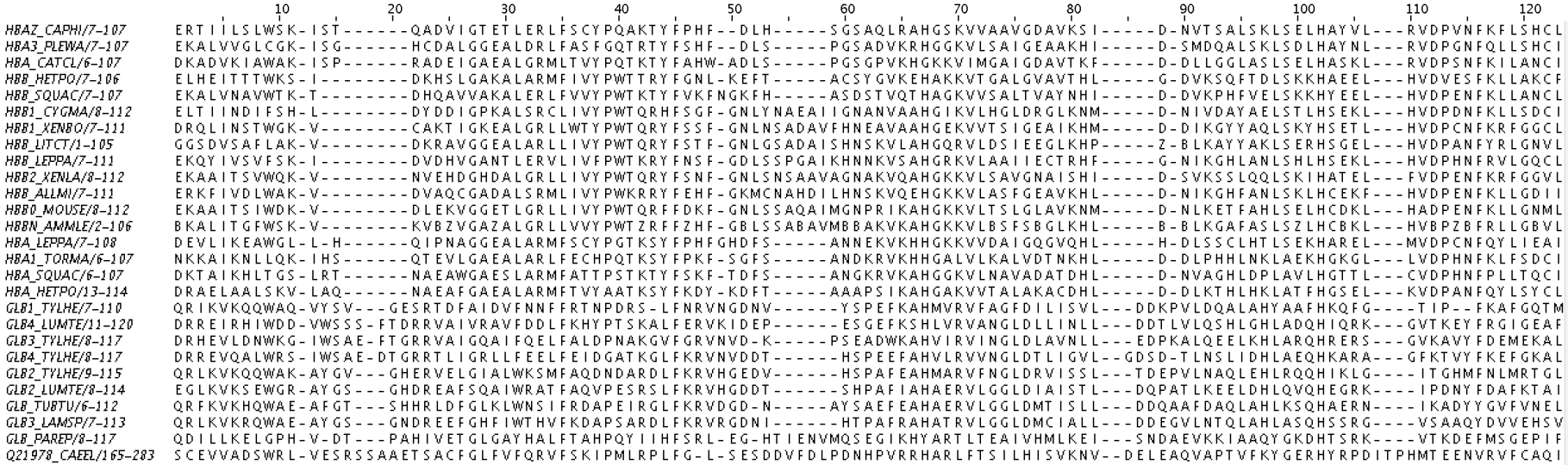
Marco Punta



```
Human: 1   MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60

Human: 61  DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120

Human: 121 PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
```

Marco Punta

Family power



Marco Punta

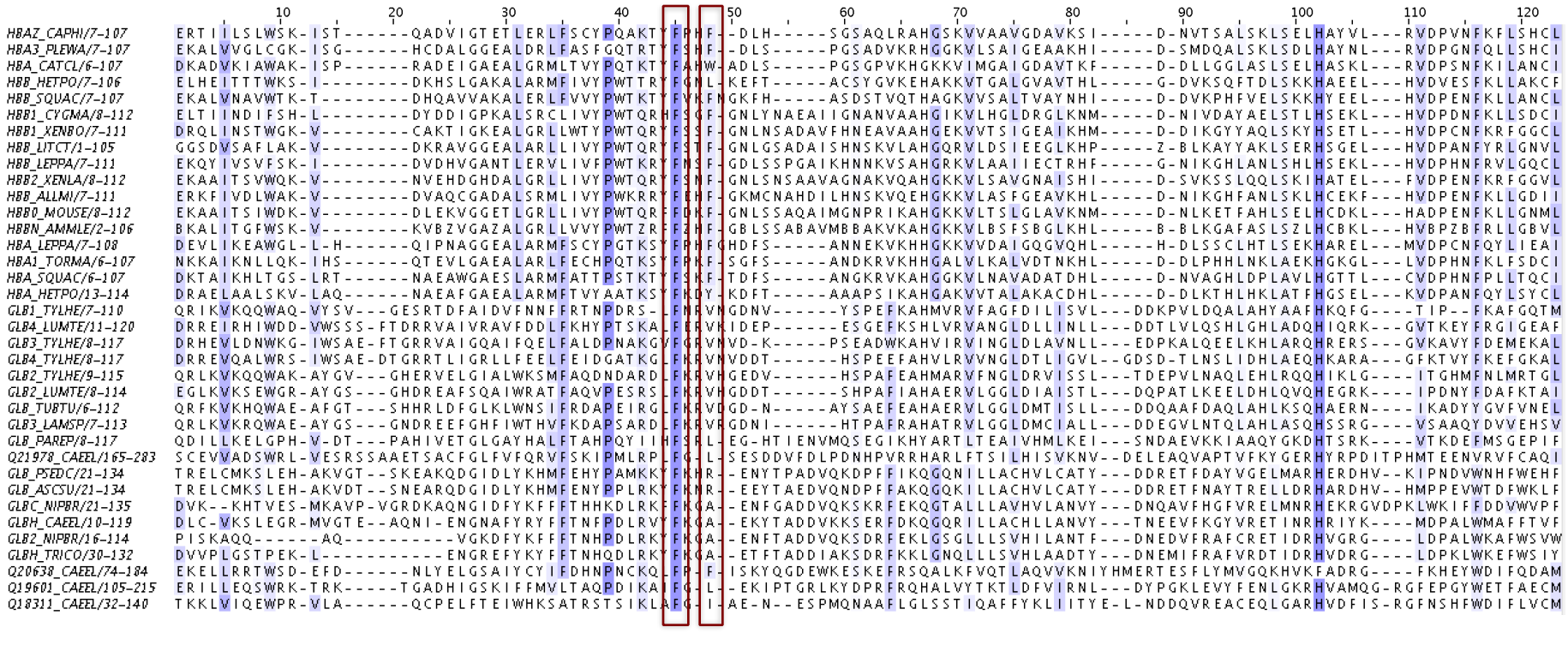
Family power

Family power

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

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

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

Marco Punta

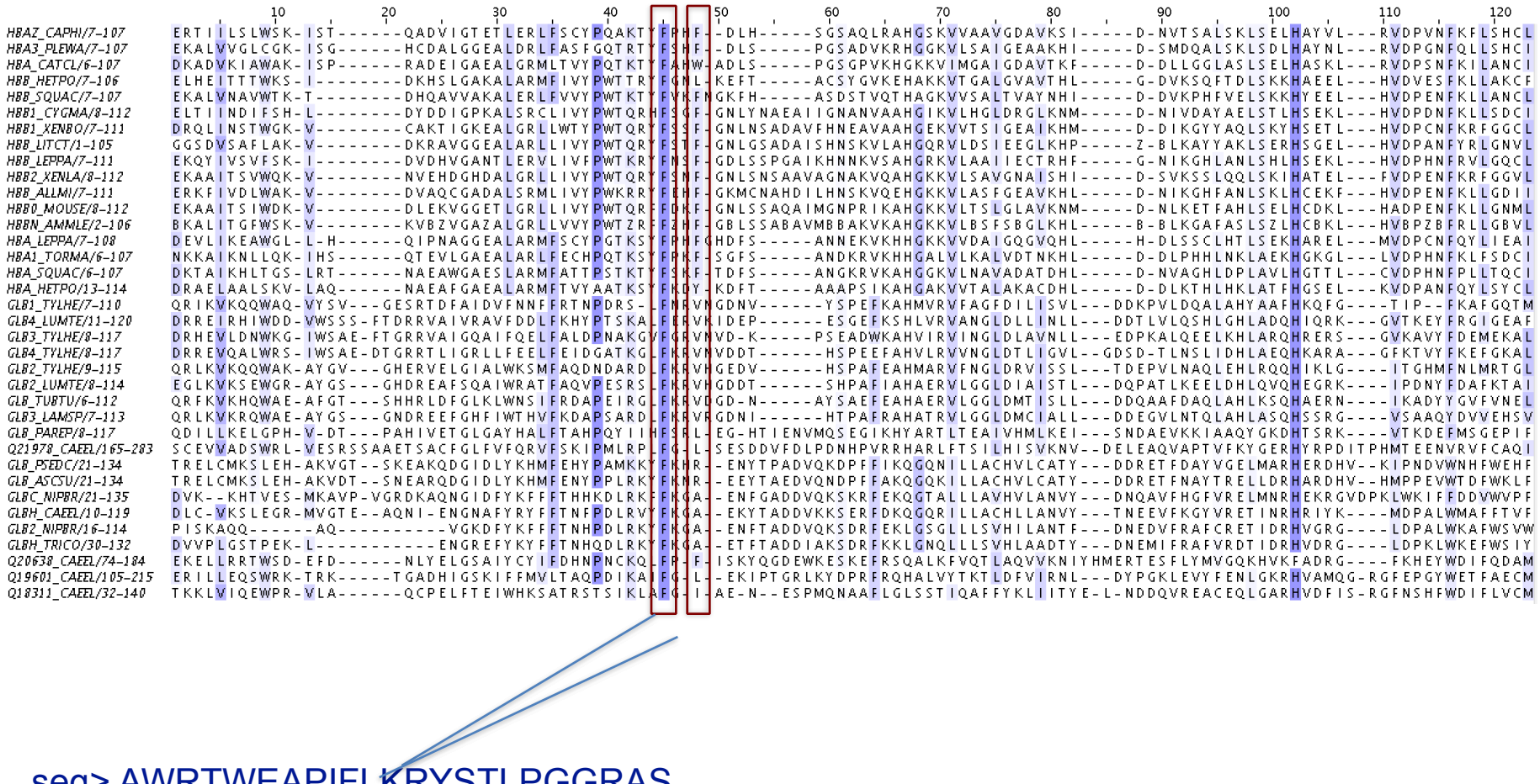
Family power



seq> AWRTWEAPIFLKRYSTLPGGRAS…

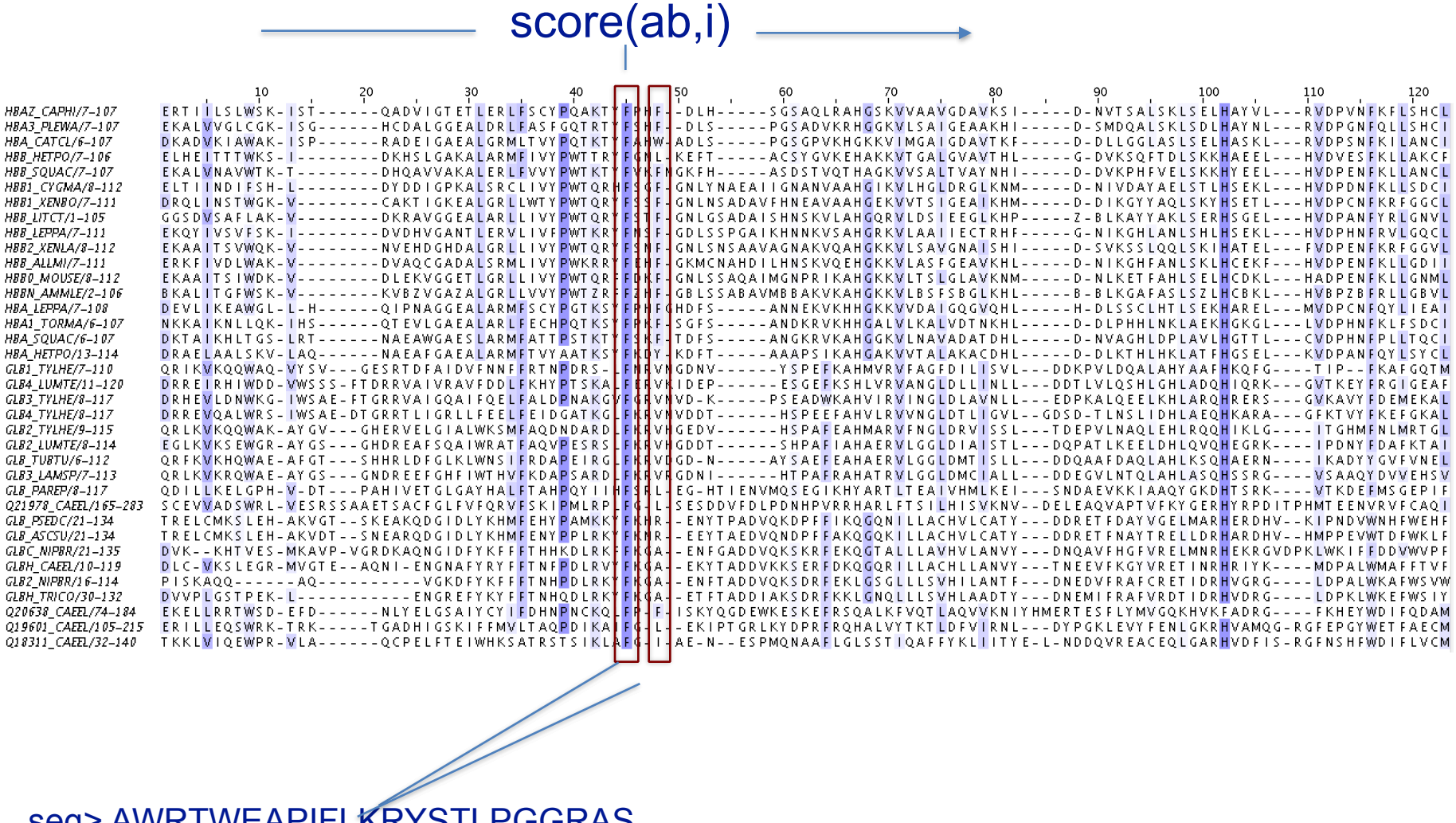
Marco Punta

Family power



Marco Punta

Family power

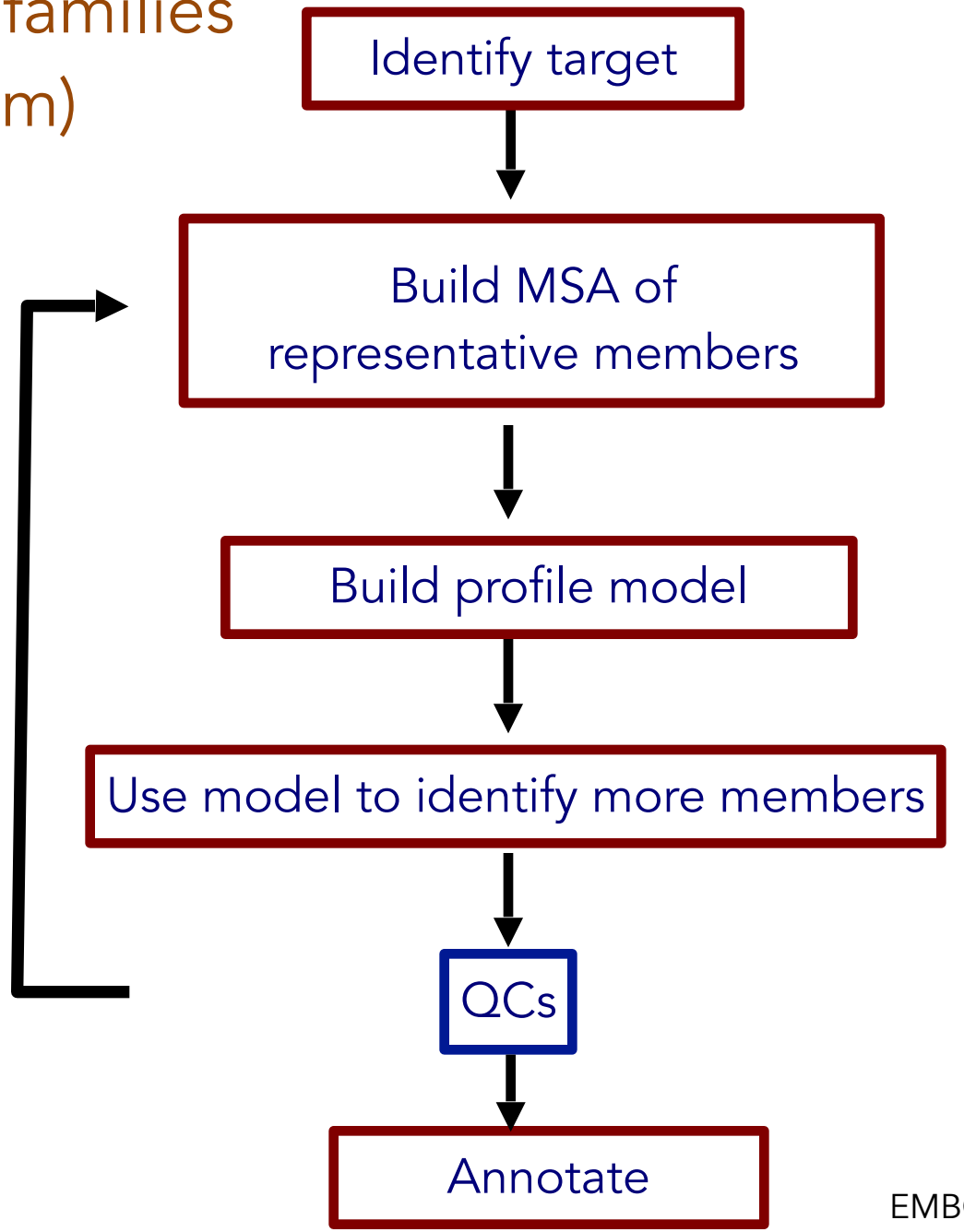


Sequence-profile alignments

- Position specific substitution matrices
- profile-hidden Markov models

Building families (Pfam)

Marco Punta



Functions, organisms, structures

Marco Punta

Prokaryotes

TIGRFAMs

~4400 families

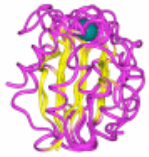
Signalling, extracellular and chromatin-associated proteins

SMART

~1000 domains

Superfamily 1.75

HMM library and genome assignments server



Structural domains from SCOP

Gene3D

Structural domains from CATH

No limits, domains

~14000 families

No limits, full-length proteins

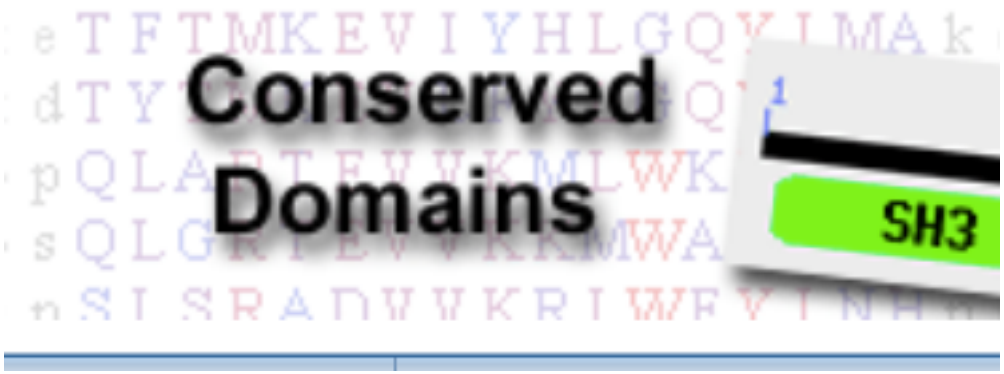
Marco Punta

>7000 families, >50000 subfamilies

~2000 families

Integration

CDD

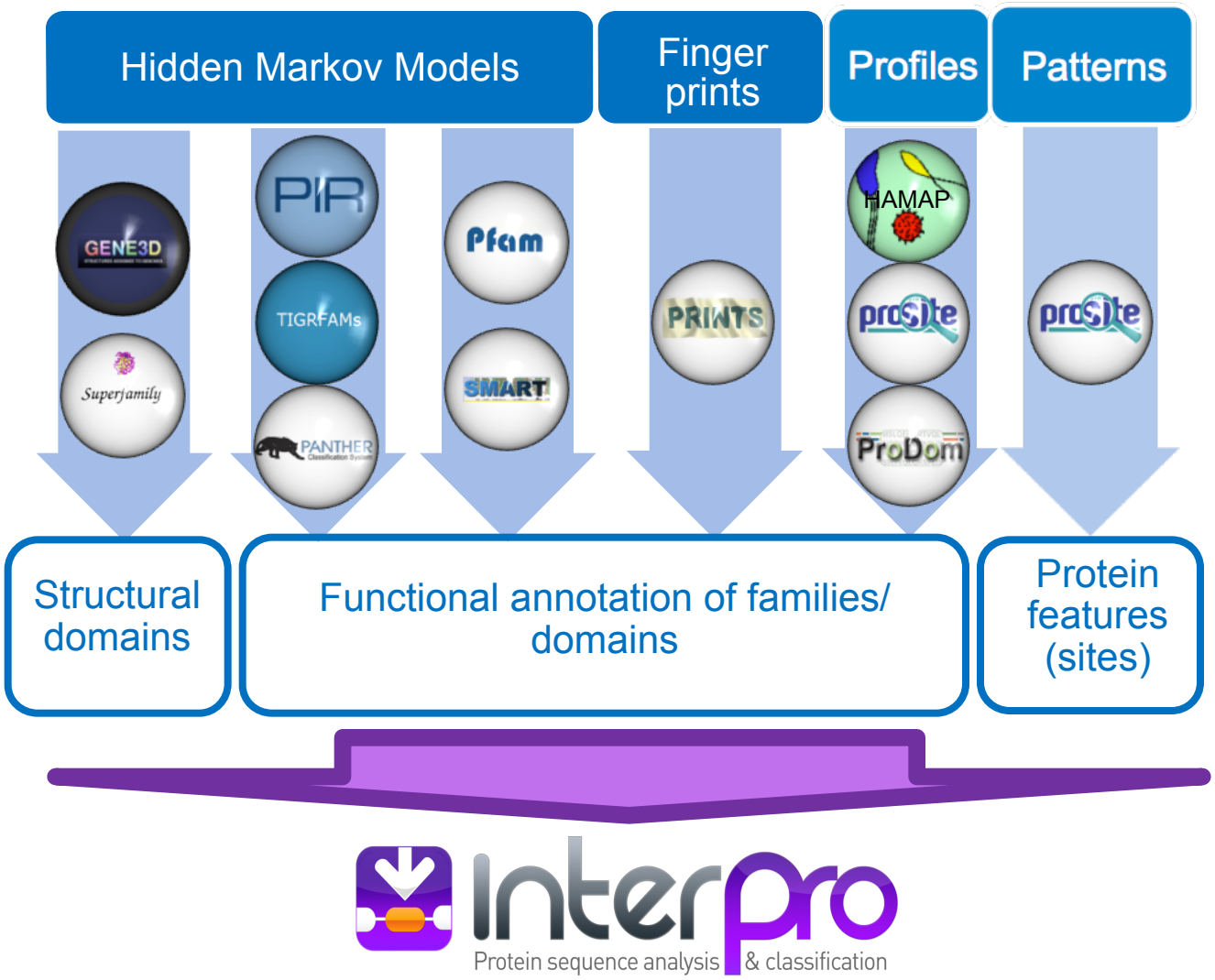


Uses RPS-BLAST

Integration

Marco Punta

Member databases



Slide courtesy of Alex Mitchell (EMBL-EBI)

Overview
Similar proteins (2905)
Structures

Filter view on

Entry type
Family
Domains
Repeats
Site

Status
Unintegrated

Colour by help
domain relationship
source database

Export FASTA

Protein

Phosphotransferase RcsD (P39838)

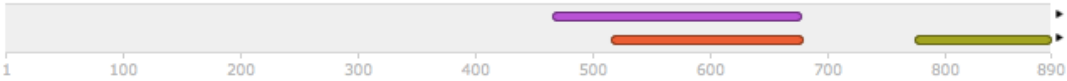
Accession P39838 (RCSD_ECOLI)
Species Escherichia coli (strain K12)
Length 890 amino acids (complete)

Source: UniProtKB

Protein family membership

None predicted.

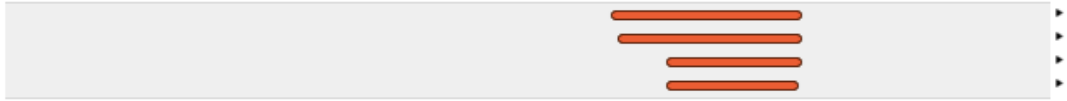
Domains and repeats



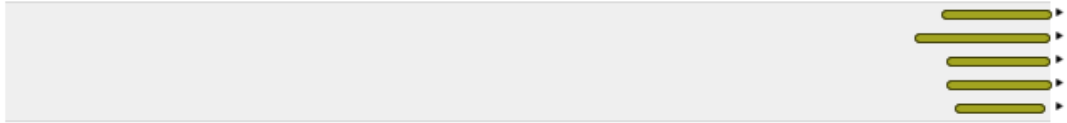
Detailed signature matches

IPR005467 Signal transduction histidine kinase, core

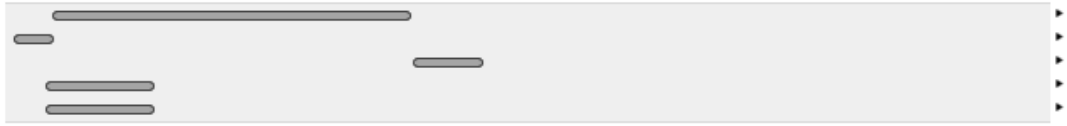
IPR003594 Histidine kinase-like ATPase, C-terminal domain



IPR008207 Signal transduction histidine kinase, phosphotransfer (Hpt) domain



no IPR Unintegrated signatures



Overview
Similar proteins (2905)
Structures

Filter view on

Entry type
Family
Domains
Repeats
Site

Status
Unintegrated

Colour by help
domain relationship
source database

Protein

Phosphotransferase RcsD (P39838)

Export FASTA

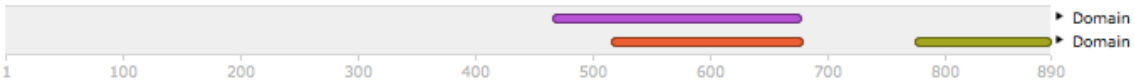
Accession P39838 (RCSD_ECOLI)
Species Escherichia coli (strain K12)
Length 890 amino acids (complete)

Source: UniProtKB

Protein family membership

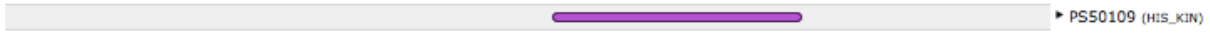
None predicted.

Domains and repeats

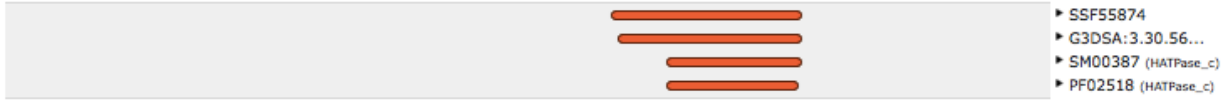


Detailed signature matches

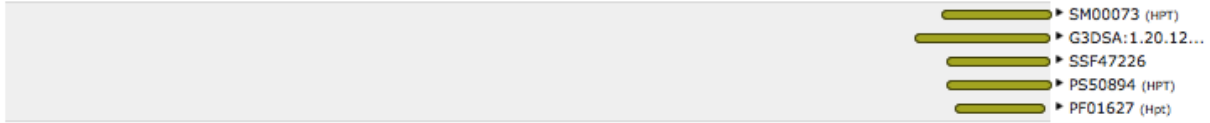
IPR005467 Signal transduction histidine kinase, core



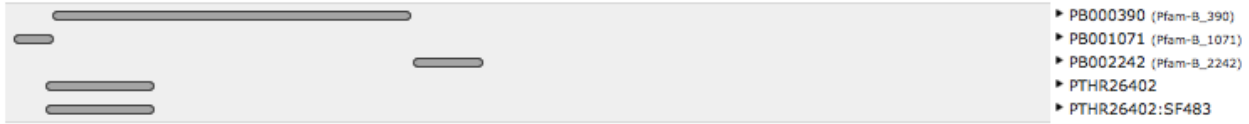
IPR003594 Histidine kinase-like ATPase, C-terminal domain



IPR008207 Signal transduction histidine kinase, phosphotransfer (Hpt) domain



no IPR Unintegrated signatures



Overview
Similar proteins (2905)
Structures

Filter view on

Entry type
- Family
- Domains
- Repeats
- Site

Status
- Unintegrated

Colour by help
- domain relationship
- source database

Export FASTA

Protein

Phosphotransferase RcsD (P39838)

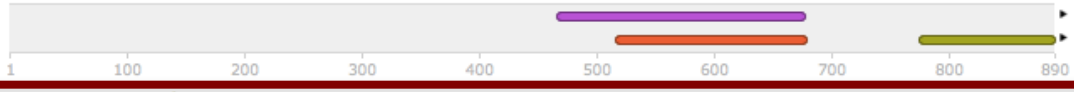
Accession P39838 (RCSD_ECOLI)
Species Escherichia coli (strain K12)
Length 890 amino acids (complete)

Source: UniProtKB

Protein family membership

None predicted.

Domains and repeats



Detailed signature matches

IPR005467 Signal transduction histidine kinase, core
- PS50109 (HIS_KIN)

IPR003594 Histidine kinase-like ATPase, C-terminal domain
- SSF55874
- G3DSA:3.30.56...
- SM00387 (HATPase_c)
- PF02518 (HATPase_c)

IPR008207 Signal transduction histidine kinase, phosphotransfer (Hpt) domain
- SM00073 (HPT)
- G3DSA:1.20.12...
- SSF47226
- PS50894 (HPT)
- PF01627 (Hpt)

no IPR Unintegrated signatures
- PB000390 (Pfam-B_390)
- PB001071 (Pfam-B_1071)
- PB002242 (Pfam-B_2242)
- PTHR26402
- PTHR26402:SF483

Overview
Similar proteins (2905)
Structures

Filter view on

Entry type
- Family
- Domains
- Repeats
- Site

Status
- Unintegrated

Colour by help
- domain relationship
- source database

Protein

Phosphotransferase RcsD (P39838)

Export FASTA

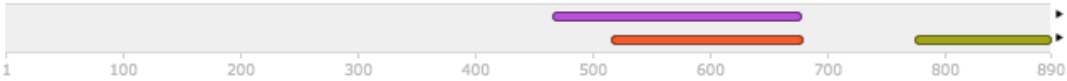
Accession P39838 (RCSD_ECOLI)
Species Escherichia coli (strain K12)
Length 890 amino acids (complete)

Source: UniProtKB

Protein family membership

None predicted.

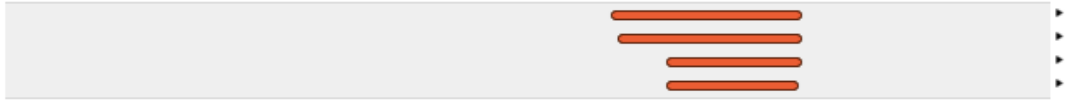
Domains and repeats



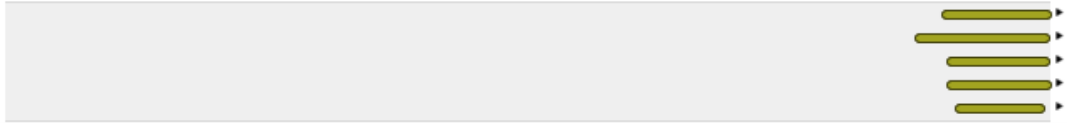
Detailed signature matches

IPR005467 Signal transduction histidine kinase, core

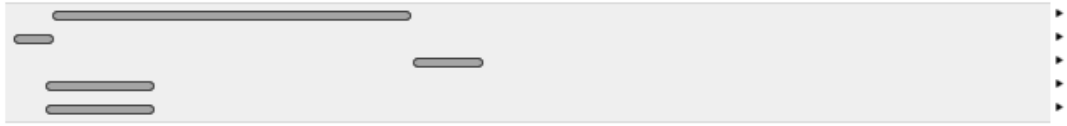
IPR003594 Histidine kinase-like ATPase, C-terminal domain



IPR008207 Signal transduction histidine kinase, phosphotransfer (Hpt) domain



no IPR Unintegrated signatures



Domain

Signal transduction histidine kinase, core (IPR005467)

Short name: Sig_transdc_His_kinase_core

Domain relationships

None.

Description

Most prokaryotic signal-transduction systems and a few eukaryotic pathways use phosphotransfer schemes involving two conserved components, a histidine protein kinase (HK) and a response regulator protein (RR). The HK, which is regulated by environmental stimuli, autophosphorylates at a histidine residue, creating a high-energy phosphoryl group that is subsequently transferred to an aspartate residue in the RR domain. Phosphorylation induces a conformational change in RR that results in activation of an associated domain that effects the response.

Both prokaryotic and eukaryotic HKs contain the same basic signaling components, namely a diverse sensing domain and a highly conserved kinase core that has a unique fold, distinct from that of the Ser/Thr/Tyr kinase superfamily. The overall activity of the kinase is modulated by input signals to the sensing domain. HKs undergo an ATP-dependent autophosphorylation at a conserved His residue in the kinase core. Autophosphorylation is a bimolecular reaction between homodimers, in which one HK monomer catalyzes the phosphorylation of the conserved His residue in the second monomer.

The sensing domains are variable in sequence, reflective of the many different environmental signals to which HKs are responsive, whereas the about 250-residue kinase core is more conserved. The kinase core is composed of a dimerization domain and an ATP/ADP-binding phosphotransfer or catalytic domain and can be identified by five conserved primary sequence motifs present in both eukaryotic and procaryotic HKs. These motifs have been termed the H, N, G1, F and G2 boxes. The conserved His substrate is the central feature in the H box, whereas the N, G1, F and G2 boxes define the nucleotide binding cleft. In most HKs, the H box is part of the dimerization domain. However, for some proteins, like CheA, the conserved His is located at the far N terminus of the protein in a separate HPt domain. The N, G1, F and G2 boxes are usually contiguous, but the spacing between these motifs is somewhat varied. The catalytic core forms an alpha-beta sandwich consisting of five antiparallel beta strands and three alpha helices [PMID: 10966457 PMID: 11406410 PMID: 11369279]

Add your annotation

Contributing signatures

Signatures from InterPro member databases are used to construct an entry.

PROSITE profiles

PS50109 (HIS_KIN)

Orthologous families, trees

OMA browser

OrthoDB

COGs
Phylogenetic classification of proteins encoded in complete genomes

eggNOG version 3.0

Treefam

InParanoid8
Ortholog Groups with Inparalogs

Pros:

Better prediction of protein function (in principle, ortholog conjecture)

Gene history

Species trees

Caveats:

Lateral gene transfer difficult to model/recognise -> bacteria difficult

Gene loss difficult to account for, may lead to wrong ortho-para assignment

Large families difficult to model

Marco Punta

Team Exercise

Building a new Pfam family

EMBL-EBI

HMMER

Home | Search | Results | Software | Help | About

The HMMER web server: fast a
This site has been designed to provide n
coupled with **intuitive and in**

Quicksta

QUICK SEARCH

Paste in your sequence or use the example

search against

Reference Proteomes | UniProtKB | SwissProt | Pfam

Submit Reset

Alternative Search Options

Blog News

August, 2015

hmmer.org is updating

hmmer.org is moving off of Janelia tonight, into the great cloud. You may see some flakiness as DNS nameservers update.

Download HMMER

Get the latest version

v3.1b2

Recent

HMMER web
R.D. Finn, J.
F. Schreiber,
Nucleic Aci

EMBL-EBI Services

HMMER

Home | Search | Results | Software | Help | About

phmmer | hmmscan | hmmsearch | jackhmmer

protein sequence vs protein sequence database

Paste a Sequence | Upload a File | Accession Search

Paste in your sequence or use the example

Submit Reset

▼ Sequence Database

Frequently used databases
- (•) Reference Proteomes
- ( ) UniProtKB
- ( ) SwissProt
- ( ) PDB

Representative Sets (UniProt)
- ( ) rp75
- ( ) rp55
- ( ) rp35
- ( ) rp15

Other databases
- ( ) QfO
- ( ) Pfamseq

▶ **Restrict by Taxonomy**

EMBL-EBI Services Re

HMMER

Home | Search | Results | Software | Help | About

phmmer hmmscan hmmsearch jackhmmer

protein sequence vs protein sequence database

Paste a Sequence | Upload a File | Accession Search

Paste in your sequence or use the example

HEAIGSGDLDLRSAFRRTSLAGAGRRTSDSHEDAGTLDFSSLLKKRD
SFRRDSKLEAPAEEDVWEILRQAPPSEYERIAFQHGVTDLRGMLKRL
KGMKQDEKK

Submit Reset

▼ **Sequence Database**

Frequently used databases	Representative Sets (UniProt)	Other databases
◉ Reference Proteomes ○ UniProtKB ○ SwissProt ○ PDB	○ rp75 ○ rp55 ○ rp35 ○ rp15	○ QfO ○ Pfamseq

▶ **Restrict by Taxonomy**

EMBL-EBI

HMMER

Home | Search | Results | Software | Help | About

phmmer | **hmmscan** | hmmsearch | jackhmmer

protein sequence vs profile-HMM database

Paste a Sequence | Upload a File | Accession Search

Paste in your sequence or use the example

HEAIGSGDLDLRSAFRRTSLAGAGRRTSDSHEDAGTLDFSSLLKKRD
SFRRDSKLEAPAEEDVWEILRQAPPSEYERIAFQHGVTDLRGMLKRL
KGMKQDEKK

Submit Reset

▼ HMM Database

Protein Families

☑ Pfam ☑ TIGRFAM ☑ Gene3D ☑ Superfamily ☑ PIRSF

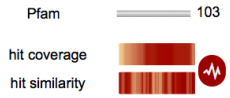
(select all)

PHMMER Results

Score Taxonomy Domain Download

Search Again

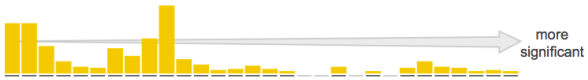
Sequence Matches and Features



✔ disorder ↻ coiled-coil ✔ tm & signal peptide

Show hit details

Distribution of Significant Hits



Bacteria Eukaryota Archaea Viruses Unclassified Sequences Other Sequences

« First « Previous **Page** 1 **of 4** Next » Last »

Significant Query Matches (330) in *uniprotrefprot* (v.2015-06-24) Customize

	Target	Species	E-value
>	Q3UIK0_MOUSE	Mus musculus	1.3e-60
>	E9Q9T8_MOUSE	Mus musculus	3.0e-58
>	Q3TF37_MOUSE	Mus musculus	1.0e-57
>	MYPC_RAT	Rattus norvegicus	2.0e-56
>	M3XYE3_MUSPF	Mustela putorius furo	9.1e-55

UniProtKB - Q3UIK0 (Q3UIK0_MOUSE)

Basket

Protein: Submitted name: **Myosin-binding protein C, cardiac-type**

Gene: **Mybpc3**

Organism: *Mus musculus (Mouse)*

Status: Unreviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

Display None

- Function
- Names & Taxonomy
- Subcell. location
- Pathol./Biotech

BLAST | Align | Format | Add to basket | History

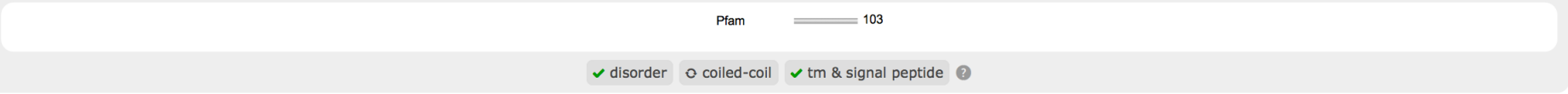
Help video | Other tutorials and videos | Feedback

Function[i]

GO - Molecular function[i]

- identical protein binding — Source: MGI
- myosin binding — Source: MGI
- myosin heavy chain binding — Source: MGI
- structural constituent of cytoskeleton — Source: MGI

Sequence Matches and Features



No hits were found for your query.

TIGRFAM Matches

Advanced

Family		Description	Start	End	Domain E-values	
Id	Accession				Ind.	Cond.

No hits were found for your query.

Your search took: 0.05 seconds

Gene3D Matches

Advanced

Family		Description	Region	Start	End	Domain E-values	
Id	Accession					Ind.	Cond.

No hits were found for your query.

Your search took: 0.05 seconds

* These hmmscan results have been modified by the Gene3D DomainFinder post processing program.

Superfamily Matches

Superfamily			Family			Region	Model Match								
							Start	End	Alignment		Model			Bit Score	Domain E-values	
Accession	Description	E-value	Accession	Description	E-value				Start	End	Start	End	Length		Ind	Cond
No hits were found for your query.																

Your search took: 0.05 seconds

* These hmmscan results have been modified by Superfamily post-processing and family assignment code.

PIRSF Matches Advanced

Family			Subfamily			Region
Accession	Description	E-value	Accession	Description	E-value	
No hits were found for your query.						

Your search took: 0.04 seconds

* These hmmscan results have been modified by PIRSF post-processing and family assignment code.

Overview
Similar proteins
Structures

Filter view on

Entry type
- Family
- Domains
- Repeats
- Site

Status
- Unintegrated

Colour by help
- domain relationship
- source database

Protein

Export Select format

Submitted

Length 103 amino acids

Protein family membership

None predicted.

Domains and repeats

None predicted.

Detailed signature matches

no IPR

Unintegrated signatures

GO term prediction

Biological Process

None predicted.

UniProtKB - Q3UIK0 (Q3UIK0_MOUSE)

Basket

Protein: Submitted name: **Myosin-binding protein C, cardiac-type**

Gene: **Mybpc3**

Organism: *Mus musculus (Mouse)*

Status: Unreviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

Display None

- Function
- Names & Taxonomy
- Subcell. location
- Pathol./Biotech

BLAST | Align | Format | Add to basket | History

Help video | Other tutorials and videos | Feedback

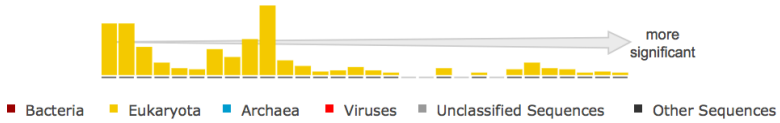
Function[i]

GO - Molecular function[i]

- identical protein binding — Source: MGI
- myosin binding — Source: MGI
- myosin heavy chain binding — Source: MGI
- structural constituent of cytoskeleton — Source: MGI

Show hit details

Distribution of Significant Hits



Significant Query Matches (330) in *uniprotrefprot* (v.2015-06-24) Customize

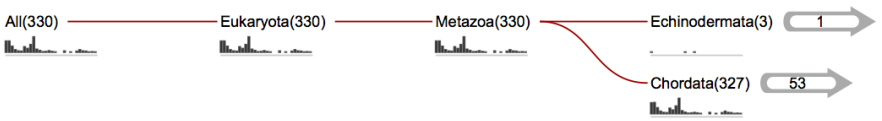
	Target	Species	E-value
v	Q3UIK0_MOUSE	Mus musculus	1.3e-60

Query		Target Envelope		Target Alignment		Bias	Accuracy	% Identity (count)	% Similarity (count)	Bit Score	E-value	
start	end	start	end	start	end						Ind.	Cond.
1	103	263	365	263	365	1.99	1.00	100.0 (103)	100.0 (103)	209.4	3.1e-60	1.1e-64

```
                .........*.........*.........*.........*.........*.........*.........*.........*
Query         1 heaigsgdldlrsafrrtslagagrrtsdshedagtldfssllkkrdsfrrdskleapaeedvweilrqappseyeriaf 80
                heaigsgdldlrsafrrtslagagrrtsdshedagtldfssllkkrdsfrrdskleapaeedvweilrqappseyeriaf
Target      263 HEAIGSGDLDLRSAFRRTSLAGAGRRTSDSHEDAGTLDFSSLLKKRDSFRRDSKLEAPAEEDVWEILRQAPPSEYERIAF 342
PP              9*******************************************************************************

                .........*.........*...
Query        81 qhgvtdlrgmlkrlkgmkqdekk 103
                qhgvtdlrgmlkrlkgmkqdekk
Target      343 QHGVTDLRGMLKRLKGMKQDEKK 365
PP              **********************98
```

PHMMER Results

Taxonomic distribution of all search hits



Species Distribution

Species	Count	View
Takifugu rubripes	22	Show
Danio rerio	19	Show
Callithrix jacchus	18	Show
Mus musculus	14	Show
Oreochromis niloticus	12	Show
Homo sapiens	11	Show
Gasterosteus aculeatus	11	Show
Canis lupus familiaris	10	Show
Macaca mulatta	10	Show
Astyanax mexicanus	9	Show
Gorilla gorilla gorilla	8	Show
Tetraodon nigroviridis	8	Show

PHMMER Results

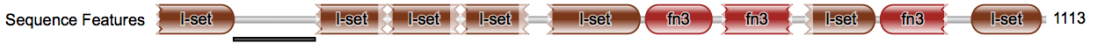
Score Taxonomy **Domain** Download

Jump to the exact match for your query architecture

Domain Architectures

94 SEQUENCES — Show All

with domain architecture: **I-set, I-set, I-set, I-set, I-set, fn3, fn3, I-set, fn3, I-set**, *example:*Q3TF37_MOUSE

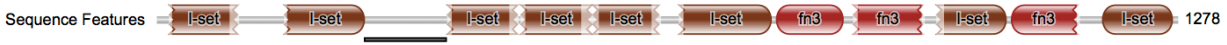


35 SEQUENCES — Show All

with domain architecture: **I-set, I-set, I-set, I-set, fn3, fn3, I-set, fn3, I-set**, *example:*F6ZHP7_HORSE

34 SEQUENCES — Show All

with domain architecture: **I-set, I-set, I-set, I-set, I-set, I-set, fn3, fn3, I-set, fn3, I-set**, *example:*Q3UIK0_MOUSE



19 SEQUENCES — Show All

with domain architecture: **I-set, I-set, I-set, fn3, fn3, fn3, I-set, fn3, I-set**, *example:*G1Q885_MYOLU

12 SEQUENCES — Show All

with domain architecture: **I-set, I-set, I-set, I-set, fn3, fn3, fn3, I-set, fn3, I-set**, *example:*W5MUP3_LEPOC

10 with domain architecture: **I-set, I-set, I-set**, *example:*F7CWG3_CALJA

PHMMER Results

- **Job:** 5B650320-65F7-11E5-8E90-C8CCF69F8818.1
- **Started:** 2015-09-28 16:41:18
- **Algorithm:** phmmer
- **HMMER Options:** -E 1 --domE 1 --incE 0.01 --incdomE 0.03 --mx BLOSUM62 --pextend 0.4 --popen 0.02 --seqdb uniprotrefprot

▼ Format

Text
A plain text file containing the hit alignments and scores.

Tab Delimited
A tab delimited text file containing the hit information. No alignments.

XML
An XML file formated for machine parsing of the data.

JSON
All the results information encoded as a single json string.

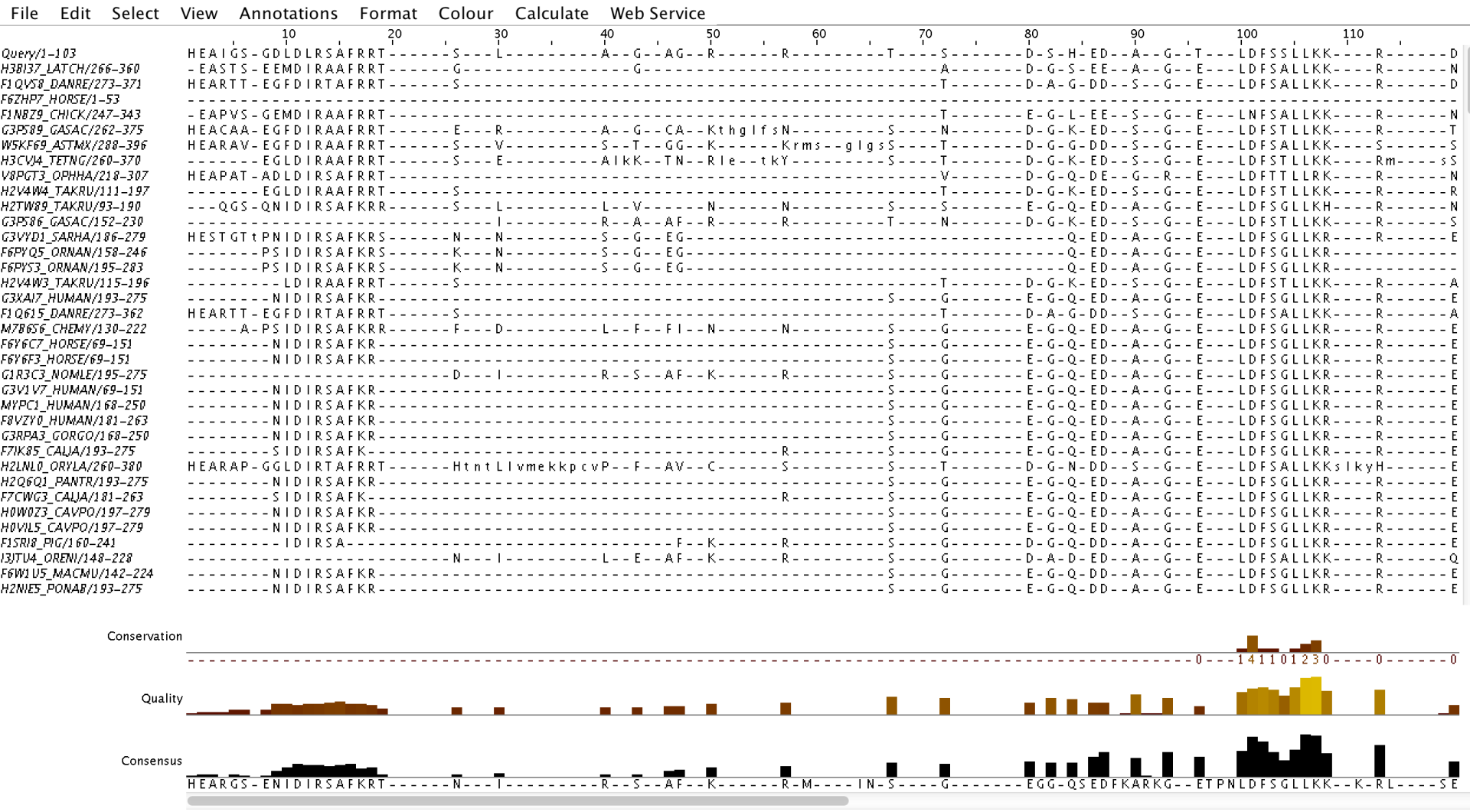
FASTA
Download the significant hits from your search as a gzipped FASTA file.

Full length FASTA
A gzipped file containing the full length sequences for significant search hits.

Aligned FASTA
A gzipped file containing aligned significant search hits in FASTA format.

STOCKHOLM
Download an alignment of significant hits as a gzipped STOCKHOLM file.

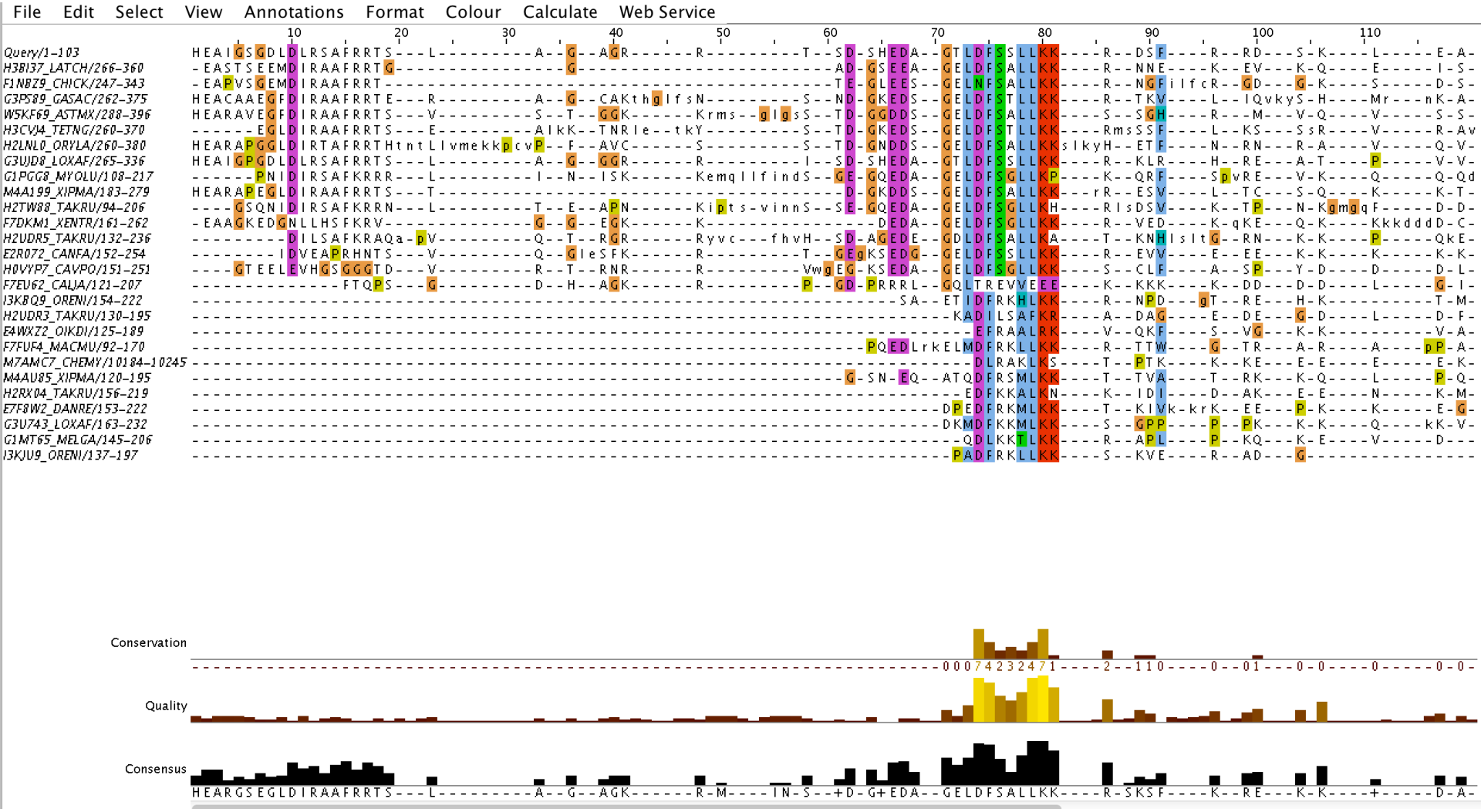
1. File -> Input alignment -> From file

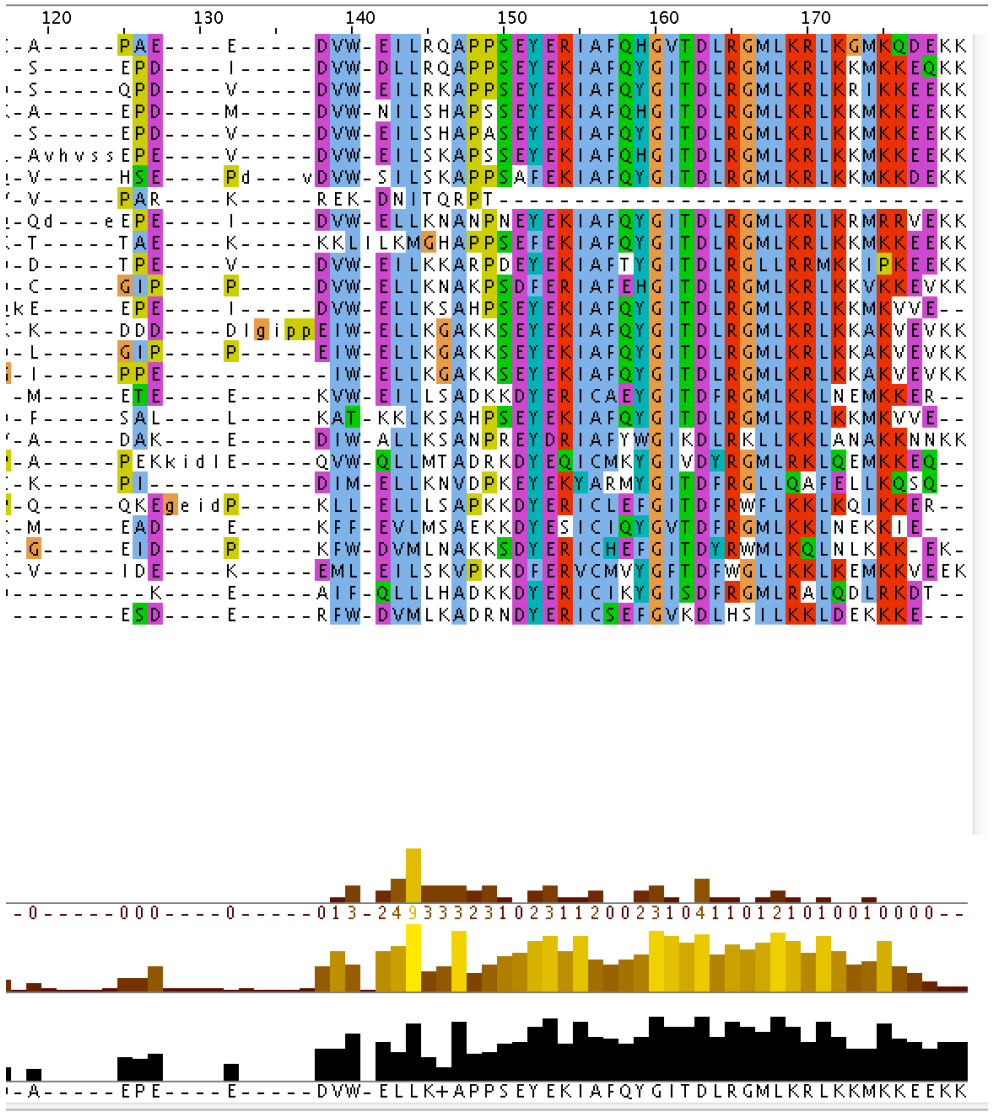


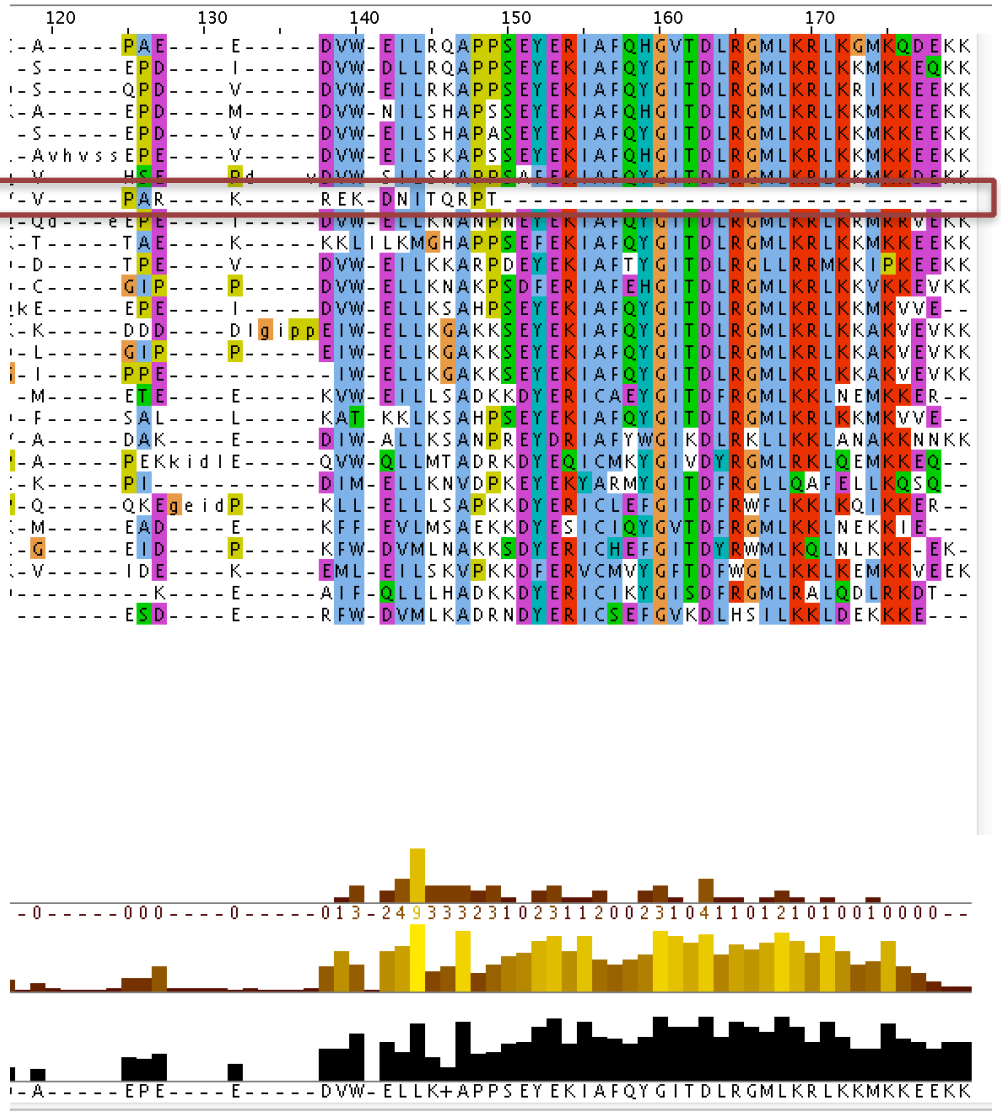
1. Edit -> Remove redundancy
2. Select 90% and Remove

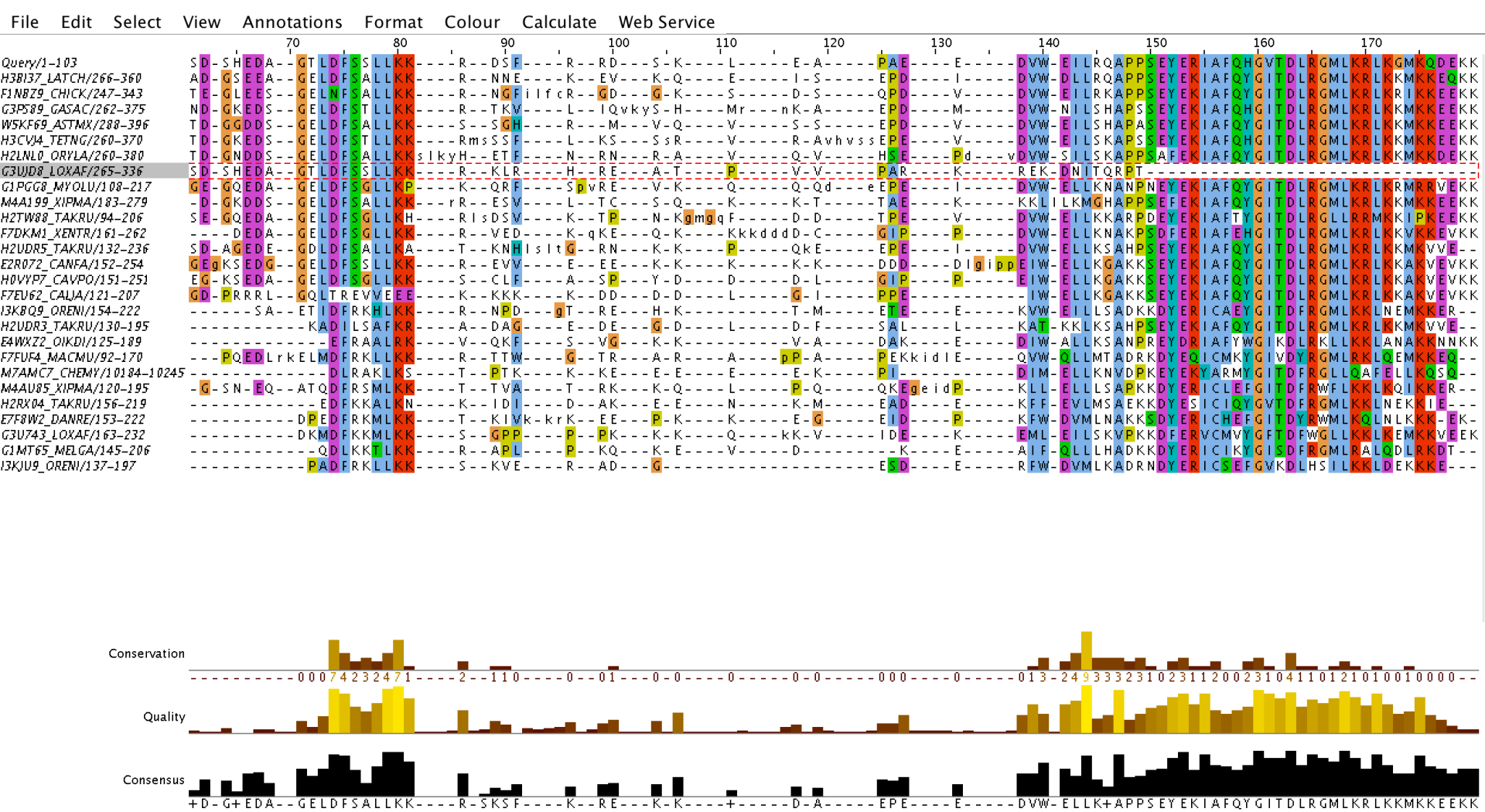
1. Edit -> Remove redundancy
2. Select 90% and Remove
3. Edit -> Remove empty columns

1. Edit -> Remove redundancy
2. Select 90% and Remove
3. Edit -> Remove empty columns
4. Colour -> Clustalx


File
Edit
Select
View
Annotations
Format
Colour
Calculate
Web Service
Conservation
Quality
Consensus

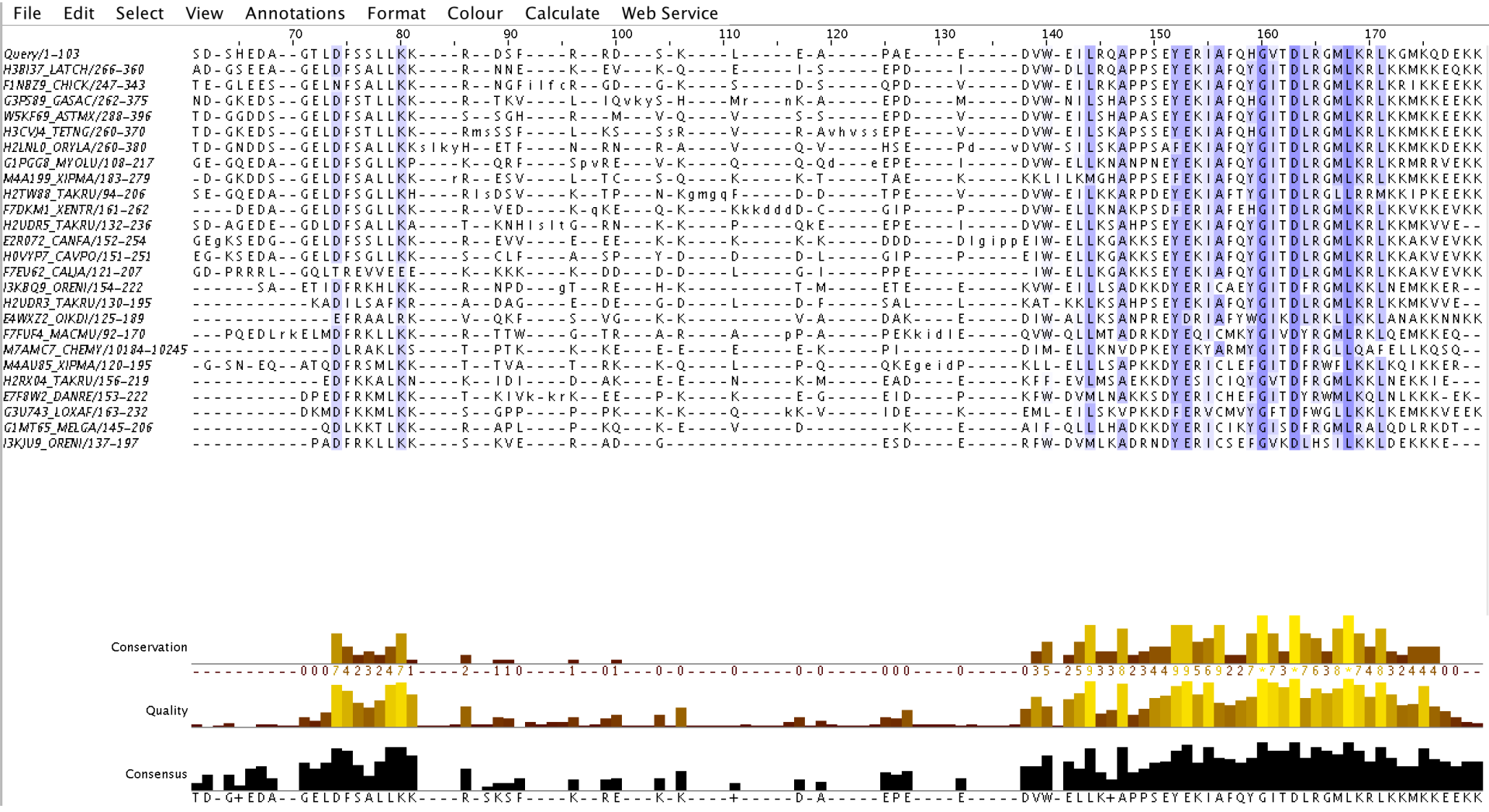

120
130
140
150
160
170
-0-----000----0-----013-24933323102311200231041101210100100000--
-A-----EPE----E-----DVW-ELLK+APPSEYEKIAFQYGITDLRGMLKRLKKMKKEEKK

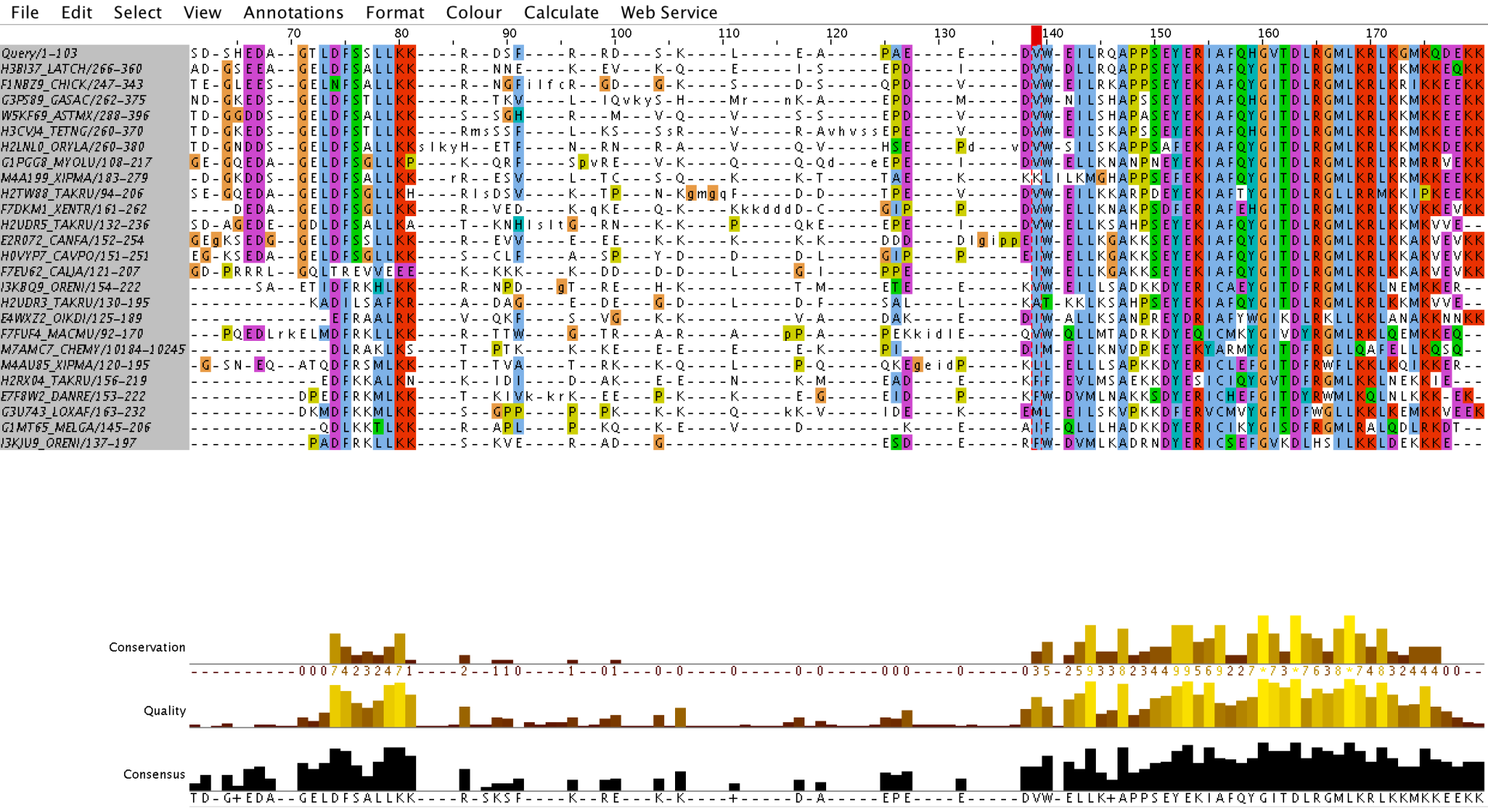

120
130
140
150
160
170
-A-----PAE----E-----DVW-EILRQAPPSEYERIAFQHGVTDLRGMLKRLKGMKQDEKK
-S-----EPD----I-----DVW-DLLRQAPPSEYEKIAFQYGITDLRGMLKRLKKMKKEQKK
-S-----QPD----V-----DVW-EILRKAPPSEYEKIAFQYGITDLRGMLKRLKRIKKEEKK
-A-----EPD----M-----DVW-NILSHAPSSEYEKIAFQHGITDLRGMLKRLKKMKKEEKK
-S-----EPD----V-----DVW-EILSHAPASEYEKIAFQYGITDLRGMLKRLKKMKKEEKK
-Avhvss EPE----V-----DVW-EILSKAPSSEYEKIAFQHGITDLRGMLKRLKKMKKEEKK
-V-----PAR----K-----REK-DNITQRPT----------
-T-----TAE----K-----KKLILKMGHAPPSEFEKIAFQYGITDLRGMLKRLKKMKKEEKK
-D-----TPE----V-----DVW-EILKKARPDEYEKIAFTYGITDLRGLLRRMKKIPKEEKK
-C-----GIP----P-----DVW-ELLKNAKPSDFERIAFEHGITDLRGMLKRLKKVKKEVKK
kE-----EPE----I-----DVW-ELLKSAHPSEYEKIAFQYGITDLRGMLKRLKKMKVVE--
-K-----DDD----DIgippEIW-ELLKGAKKSEYEKIAFQYGITDLRGMLKRLKKAKVEVKK
-L-----GIP----P-----EIW-ELLKGAKKSEYEKIAFQYGITDLRGMLKRLKKAKVEVKK
-I-----PPE----------IW-ELLKGAKKSEYEKIAFQYGITDLRGMLKRLKKAKVEVKK
-M-----ETE----E-----KVW-EILLSADKKDYERICAEYGITDFRGMLKKLNEMKKER--
-F-----SAL----L-----KAT-KKLKSAHPSEYEKIAFQYGITDLRGMLKRLKKMKVVE--
-A-----DAK----E-----DIW-ALLKSANPREYDRIAFYWGIKDLRKLLKKLANAKKNNKK
-A-----PEKkidlE-----QVW-QLLMTADRKDYEQICMKYGIVDYRGMLRKLQEMKKEQ--
-K-----PI----------DIM-ELLKNVDPKEYEKYARMYGITDFRGLLQAFELLKQSQ--
-Q-----QKEgeidP-----KLL-ELLLSAPKKDYERICLEFGITDFRWFLKKLKQIKKER--
-M-----EAD----E-----KFF-EVLMSAEKKDYESICIQYGVTDFRGMLKKLNEKKIE---
-G-----EID----P-----KFW-DVMLNAKKSDYERICHEFGITDYRWMLKQLNLKKK-EK-
-V-----IDE----K-----EML-EILSKVPKKDFERVCMVYGFTDFWGLLKKLKEMKKVEEK
---------K----E-----AIF-QLLLHADKKDYERICIKYGISDFRGMLRALQDLRKDT--
--------ESD----E-----RFW-DVMLKADRNDYERICSEFGVKDLHSILKKLDEKKKE---
-0-----000----0-----013-24933323102311200231041101210100100000--
-A-----EPE----E-----DVW-ELLK+APPSEYEKIAFQYGITDLRGMLKRLKKMKKEEKK



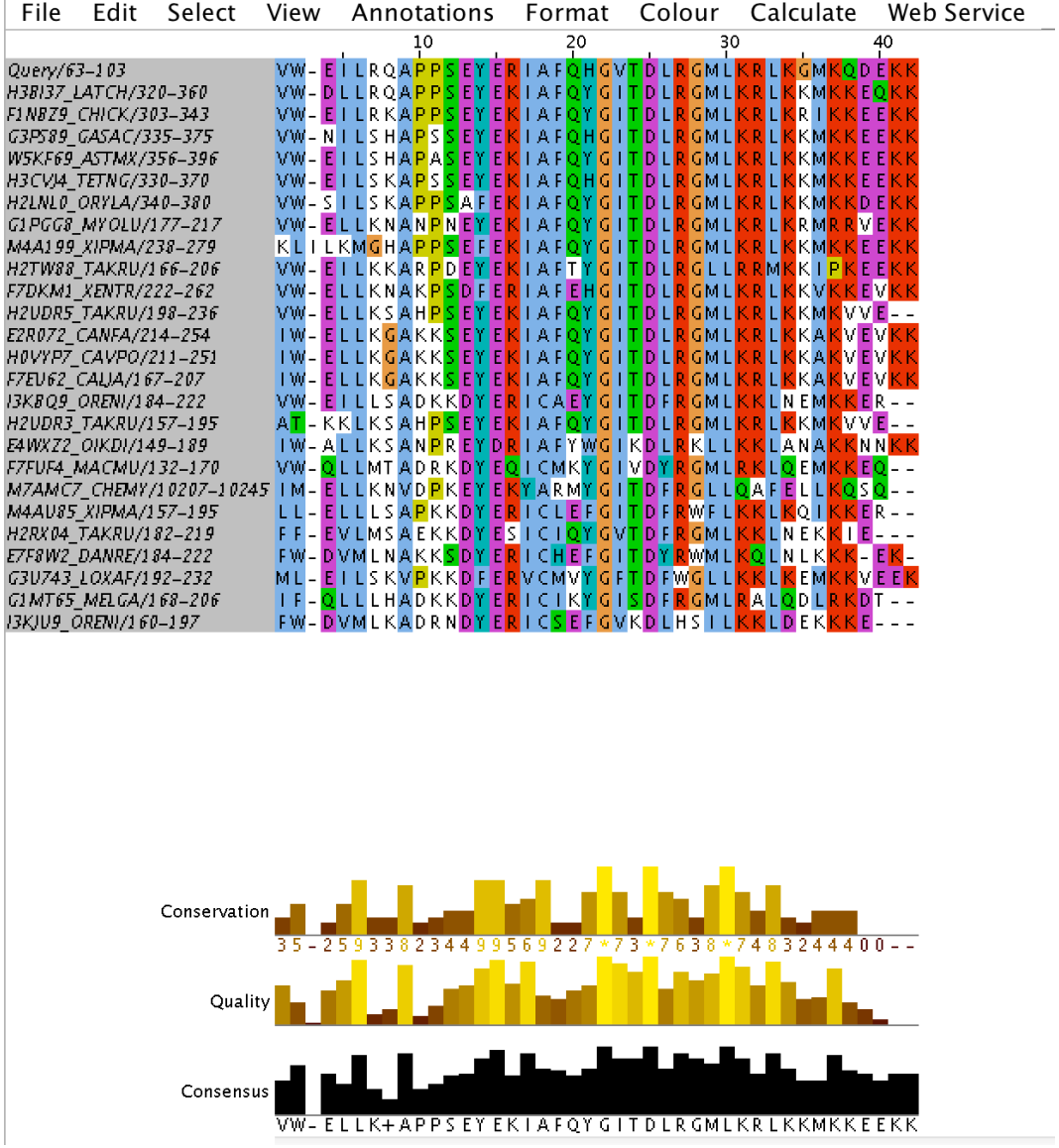
1. Edit -> Delete

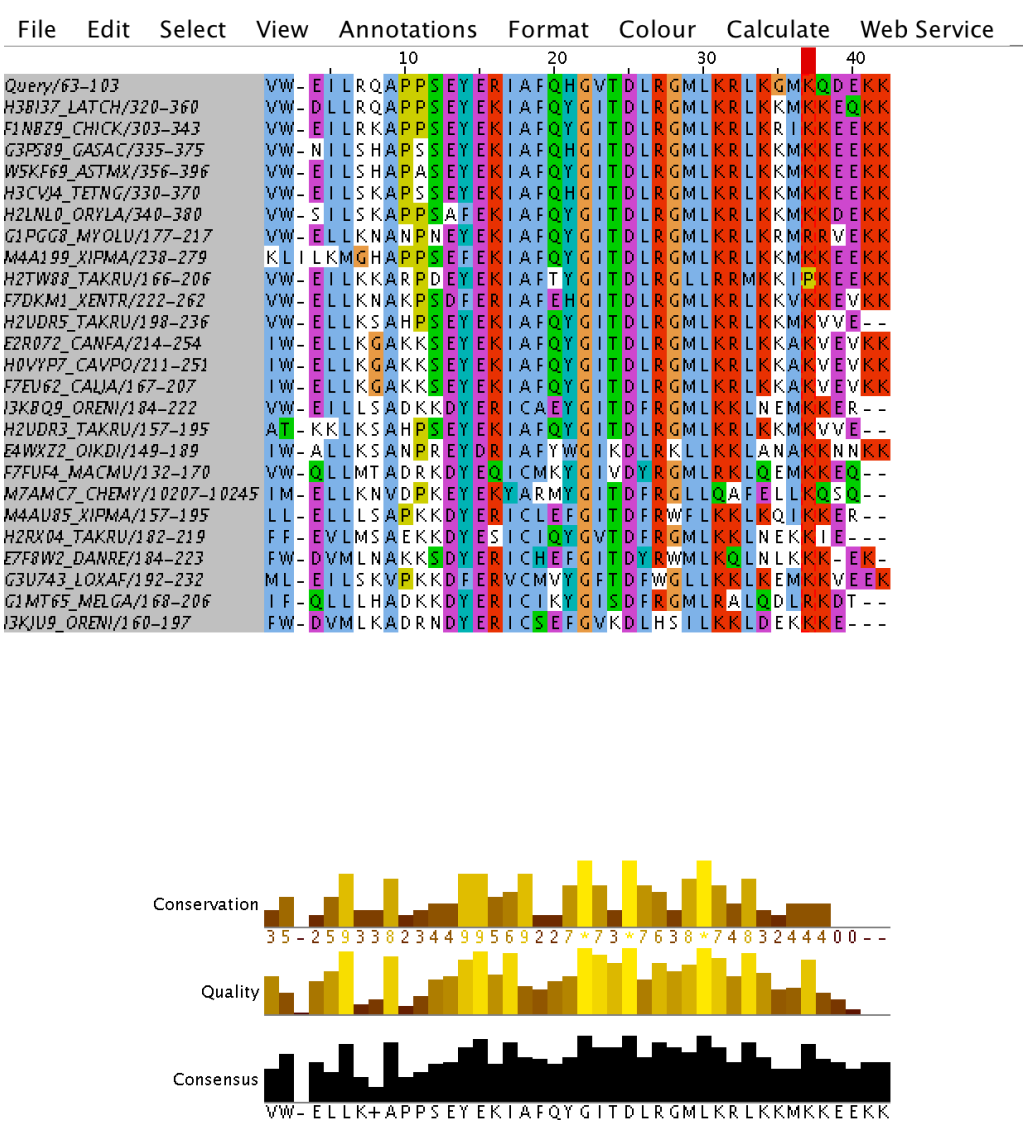
2. Colour -> BLOSUM62 Score / Colour -> Percentage Identity


File
Edit
Select
View
Annotations
Format
Colour
Calculate
Web Service
70
80
90
100
110
120
130
140
150
160
170
Query/1-103
H3BI37_LATCH/266-360
F1NBZ9_CHICK/247-343
G3PS89_GASAC/262-375
W5KF69_ASTMX/288-396
H3CVJ4_TETNG/260-370
H2LNL0_ORYLA/260-380
G1PGG8_MYOLU/108-217
M4A199_XIPMA/183-279
H2TW88_TAKRU/94-206
F7DKM1_XENTR/161-262
H2UDR5_TAKRU/132-236
E2R072_CANFA/152-254
H0VYP7_CAVPO/151-251
F7EU62_CALJA/121-207
I3KBQ9_ORENI/154-222
H2UDR3_TAKRU/130-195
E4WXZ2_OIKDI/125-189
F7FUF4_MACMU/92-170
M7AMC7_CHEMY/10184-10245
M4AU85_XIPMA/120-195
H2RX04_TAKRU/156-219
E7F8W2_DANRE/153-222
G3U743_LOXAF/163-232
G1MT65_MELGA/145-206
I3KJU9_ORENI/137-197
Conservation
Quality
Consensus



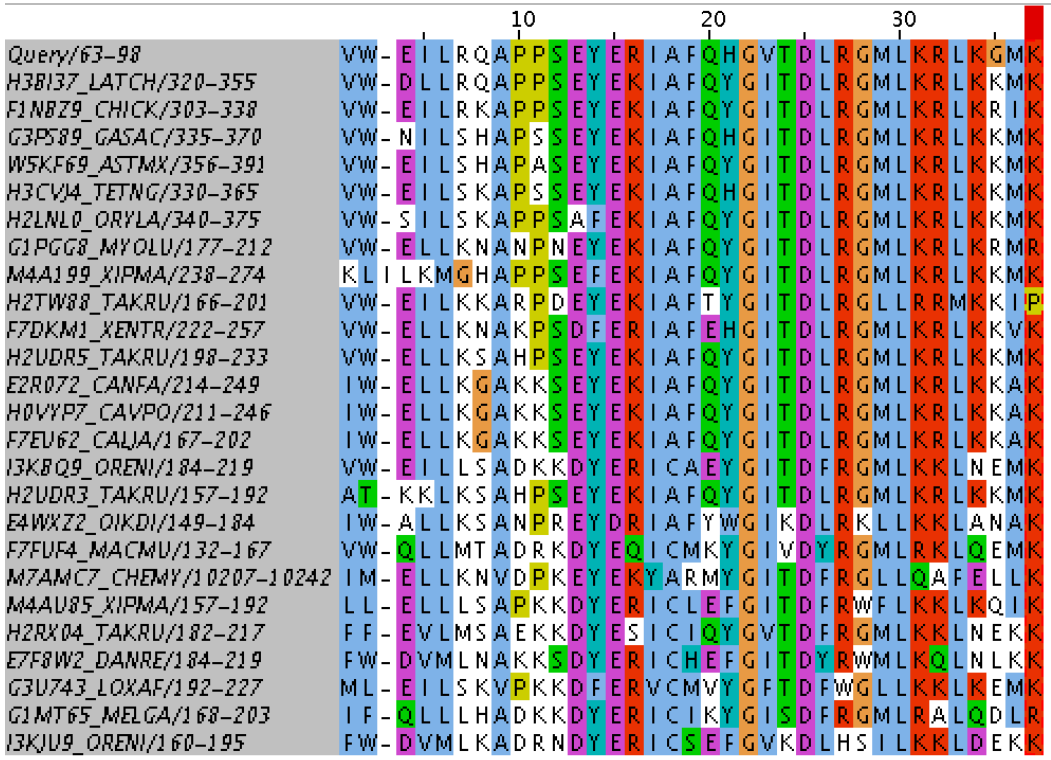
1. Edit -> Remove left

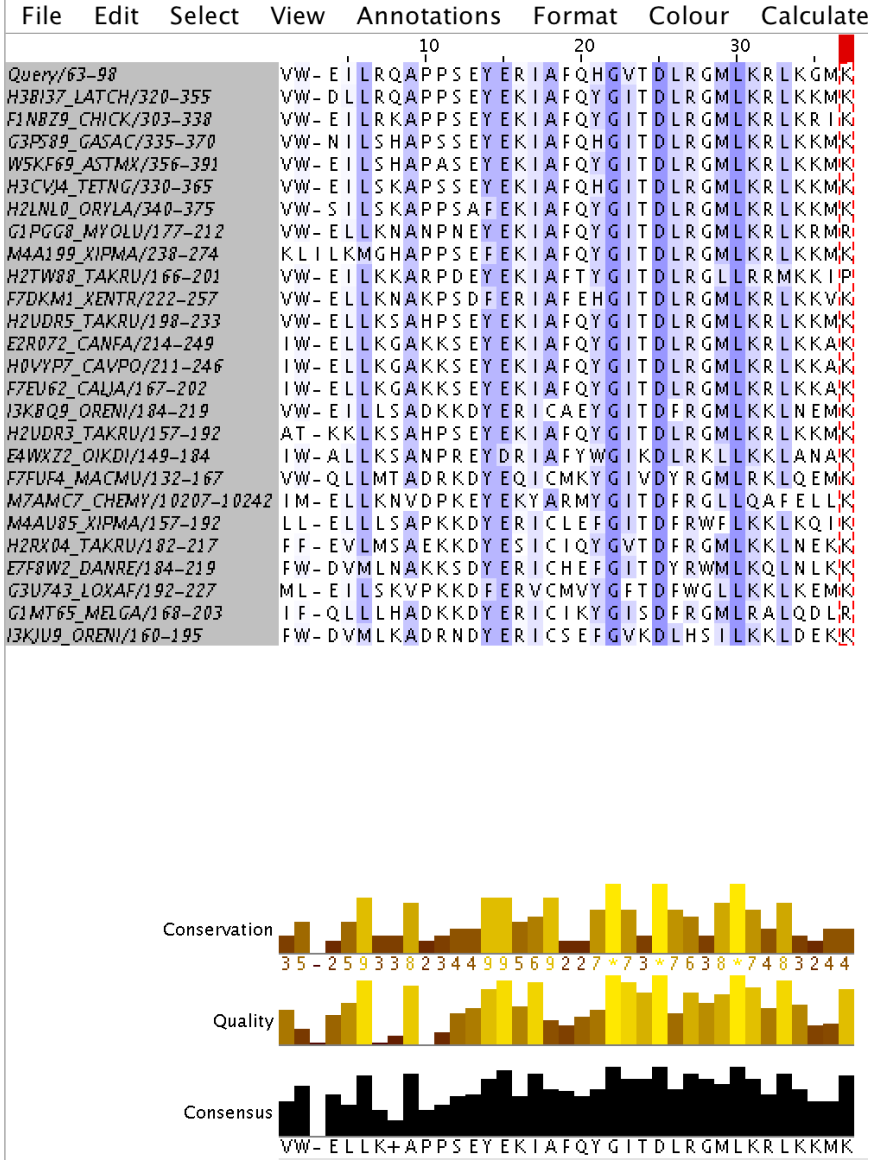

File Edit Select View Annotations Format Colour Calculate Web Service
10
20
30
40
Query/63-103
H3BI37_LATCH/320-360
F1NBZ9_CHICK/303-343
G3PS89_GASAC/335-375
W5KF69_ASTMX/356-396
H3CVJ4_TETNG/330-370
H2LNL0_ORYLA/340-380
G1PGG8_MYOLU/177-217
M4A199_XIPMA/238-279
H2TW88_TAKRU/166-206
F7DKM1_XENTR/222-262
H2UDR5_TAKRU/198-236
E2R072_CANFA/214-254
H0VYP7_CAVPO/211-251
F7EU62_CALJA/167-207
I3KBQ9_ORENI/184-222
H2UDR3_TAKRU/157-195
E4WXZ2_OIKDI/149-189
F7FUF4_MACMU/132-170
M7AMC7_CHEMY/10207-10245
M4AU85_XIPMA/157-195
H2RX04_TAKRU/182-219
E7F8W2_DANRE/184-222
G3U743_LOXAF/192-232
G1MT65_MELGA/168-206
I3KJU9_ORENI/160-197
Conservation
Quality
Consensus
VW-ELLK+APPSEYEKIAFQYGITDLRGMLKRLKKMKKEEKK



1. Edit -> Remove right

File Edit Select View Annotations Format Colour Calculate Web Service





1. File -> Save as

EMBL-EBI Services Re

HMMER

Home | Search | Results | Software | Help | About

phmmer hmmscan **hmmsearch** jackhmmer

protein alignment/profile-HMM vs protein sequence database

Paste a Sequence | **Upload a File** | Accession Search

Upload a file Browse... jalview.fasta

Submit Reset

▼ Sequence Database

Frequently used databases
- (•) Reference Proteomes
- ( ) UniProtKB
- ( ) SwissProt
- ( ) PDB

Representative Sets (UniProt)
- ( ) rp75
- ( ) rp55
- ( ) rp35
- ( ) rp15

Other databases
- ( ) QfO
- ( ) Pfamseq

▶ **Restrict by Taxonomy**

▼ Cut-Offs

(•) **E-value** ( ) Bit score

Significance E-values: Sequence 0.01 Hit 0.03

HMMSEARCH Results

Score Taxonomy Domain Download

Search Again

Distribution of Significant Hits

Bacteria Eukaryota Archaea Viruses Unclassified Sequences Other Sequences

« First « Previous **Page** 1 **of 4** Next » Last »

Significant Query Matches (362) in *uniprotrefprot* (v.2015-06-24) Customize

	Target	Species	E-value
>	W4Z1N0_STRPU	Strongylocentrotus purpuratus	1.8e-34
>	W4Y2S3_STRPU	Strongylocentrotus purpuratus	3.1e-31
>	M7AMC7_CHEMY	Chelonia mydas	1.1e-25
>	W5LEB4_ASTMX	Astyanax mexicanus	3.5e-22
>	F7CWG3_CALJA	Callithrix jacchus	1.1e-20
>	H2UDR6_TAKRU	Takifugu rubripes	1.3e-20
>	H2UDR5_TAKRU	Takifugu rubripes	1.3e-20
>	F6W1U5_MACMU	Macaca mulatta	1.5e-20
>	F1QVS8_DANRE	Danio rerio	1.5e-20
>	H2UDR4_TAKRU	Takifugu rubripes	2.6e-20
>	H2UDR8_TAKRU	Takifugu rubripes	2.7e-20
>	H2UDR1_TAKRU	Takifugu rubripes	2.7e-20
>	H2UDR2_TAKRU	Takifugu rubripes	2.7e-20
>	H2UDR7_TAKRU	Takifugu rubripes	2.7e-20
>	M3ZYX5_XIPMA	Xiphophorus maculatus	2.7e-20
>	F1Q615_DANRE	Danio rerio	3.5e-20

Annotation?

UniProtKB - Q3UIK0 (Q3UIK0_MOUSE)

Basket

Protein: Submitted name: **Myosin-binding protein C, cardiac-type**

Gene: **Mybpc3**

Organism: *Mus musculus (Mouse)*

Status: Unreviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

Display None

- Function
- Names & Taxonomy
- Subcell. location
- Pathol./Biotech

BLAST | Align | Format | Add to basket | History

Help video | Other tutorials and videos | Feedback

Function[i]

GO - Molecular function[i]

- identical protein binding — Source: MGI
- myosin binding — Source: MGI
- myosin heavy chain binding — Source: MGI
- structural constituent of cytoskeleton — Source: MGI

Function

GO - Molecular function

- identical protein binding Source: MGI
- myosin binding Source: MGI
- myosin heavy chain binding Source: MGI

Inferred from physical interaction

PubMed 17075052

- heart morphogenesis Source: MGI
- muscle contraction Source: MGI
- myosin filament assembly Source: MGI
- regulation of heart contraction Source: MGI
- regulation of heart rate Source: MGI
- regulation of striated muscle contraction Source: Ensembl
- sarcomere organization Source: MGI
- ventricular cardiac muscle tissue morphogenesis Source: MGI

Complete GO annotation...

Literature citation

Map to

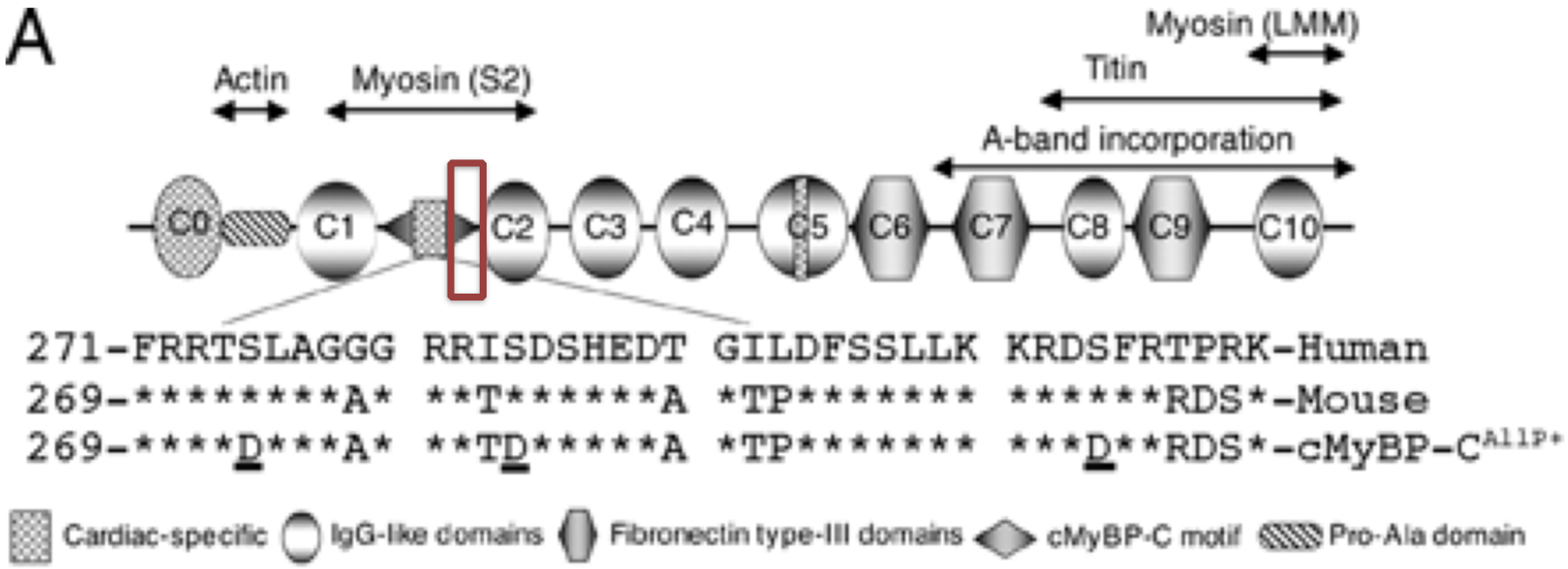
Mapped (16)

Format

Cardiac myosin binding protein C phosphorylation is cardioprotective.

Sadayappan S., Osinska H., Klevitsky R., Lorenz J.N., Sargent M., Molkentin J.D., Seidman C.E., Seidman J.G., Robbins J.

Cardiac myosin binding protein C (cMyBP-C) has three phosphorylatable serines at its N terminus (Ser-273, Ser-282, and Ser-302), and the residues' phosphorylation states may alter thick filament structure and function. To examine the effects of cMyBP-C phosphorylation, we generated transgenic mice with cardiac-specific expression of a cMyBP-C in which the three phosphorylation sites were mutated to aspartic acid, mimicking constitutive phosphorylation (cMyBP-C(AllP+)). The allele was bred into a cMyBP-C null background (cMyBP-C((t/t))) to ensure the absence of endogenous dephosphorylated cMyBP-C. cMyBP-C(AllP+) was incorporated normally into the cardiac sarcomere and restored normal cardiac function in the null background. However, subtle changes in sarcomere ultrastructure, characterized by increased distances between the thick filaments, indicated that phosphomimetic cMyBP-C affects thick-thin filament relationships, and yeast two-hybrid data and pull-down studies both showed that charged residues in these positions effectively prevented interaction with the myosin heavy chain. Confirming the physiological relevance of these data, the cMyBP-C(AllP+:(t/t)) hearts were resistant to ischemia-reperfusion injury. These data demonstrate that cMyBP-C phosphorylation functions in maintaining thick filament spacing and structure and can help protect the myocardium from ischemic injury.


A
Actin
Myosin (S2)
Myosin (LMM)
Titin
A-band incorporation
C0
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
271-FRRTSLAGGG RRISDSHEDT GILDFSSLLK KRDSFRTPRK-Human
269-********A* **T******A *TP******* ******RDS*-Mouse
269-****D***A* **TD*****A *TP******* ***D**RDS*-cMyBP-C^A11P+
Cardiac-specific
IgG-like domains
Fibronectin type-III domains
cMyBP-C motif
Pro-Ala domain

EMBL-EBI Services Research

HMMER

Home | Search | Results | Software | Help | About

phmmer hmmscan **hmmsearch** jackhmmer

protein alignment/profile-HMM vs protein sequence database

Paste a Sequence | **Upload a File** | Accession Search

Upload a file Browse... jalview.fasta

Submit Reset

▼ Sequence Database

Frequently used databases

○ Reference Proteomes ○ UniProtKB ○ SwissProt ● PDB

Representative Sets (UniProt)

○ rp75 ○ rp55 ○ rp35 ○ rp15

Other databases

○ QfO ○ Pfamseq

▶ Restrict by Taxonomy

▼ Cut-Offs

● **E-value** ○ Bit score

HMMSEARCH Results

Score Taxonomy Domain Download

Search Again

Distribution of Significant Hits

Bacteria Eukaryota Archaea Viruses Unclassified Sequences Other Sequences

Significant Query Matches (1) in *pdb* (v.2015-06-24) Customize

	Target	Species	E-value
>	2lhu_A	Mus musculus	5.8e-22

(show all) alignments

Your search took:0.01 secs

showing rows 1 - 1 of 1

Search Details

■ Bacteria ■ Eukaryota ■ Archaea ■ Viruses ■ Unclassified Sequences ■ Other Sequences

Significant Query Matches (1) in *pdb* (v.2015-06-24)

Customize

	Target	Species	E-value
v	2lhu_A	Mus musculus	5.8e-22

Query		Target Envelope		Target Alignment		Bias	Accuracy	% Identity (count)	% Similarity (count)	Bit Score	E-value	
start	end	start	end	start	end						Ind.	Cond.
23	31	56	69	58	66	0.06	0.85	33.3 (3)	77.8 (7)	0.5	7100	0.026

```
                .........
Query       23 tDlRgmLKr 31
                D+ ++LK+
Target      58 LDFSSLLKK 66
PP             69******8
```

Query		Target Envelope		Target Alignment		Bias	Accuracy	% Identity (count)	% Similarity (count)	Bit Score	E-value	
start	end	start	end	start	end						Ind.	Cond.
1	36	84	119	84	119	0.01	0.99	75.0 (27)	97.2 (35)	77.9	4.8e-21	1.8e-26

```
                .........*.........*.........*......
Query        1 vwelLkkAdkseYEkIafqYGItDlRgmLKrLkkmK 36
               vwe+L++A++seYE+Iafq+G+tDlRgmLKrLk mK
Target      84 VWEILRQAPPSEYERIAFQHGVTDLRGMLKRLKGMK 119
PP             8*********************************9
```

(show all) alignments

Your search took:0.01 secs

showing rows 1 - 1 of 1

Structural Insight into the Unique Cardiac Myosin Binding Protein-C Motif: A Partially Folded Domain

2LHU

Display Files
Download Files
Download Citation

DOI:10.2210/pdb2lhu/pdb

Primary Citation

Structural insight into unique cardiac myosin-binding protein-C motif: a partially folded domain.

Howarth, J.W., Ramisetti, S., Nolan, K., Sadayappan, S., Rosevear, P.R.

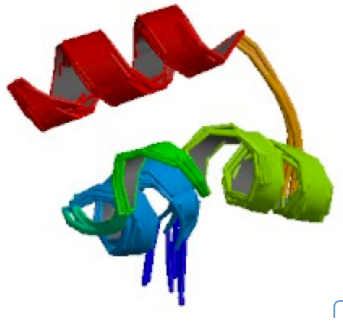
Journal: (2012) J.Biol.Chem. **287:** 8254-8262

PubMed: 22235120
PubMedCentral: PMC3318737
DOI: 10.1074/jbc.M111.309591
Search Related Articles in PubMed

PubMed Abstract:

The structural role of the unique myosin-binding motif (m-domain) of cardiac myosin-binding protein-C remains unclear. Functionally, the m-domain is thought to directly interact with myosin, whereas phosphorylation of the m-domain has been shown to

Structure Image



Currently viewing **unique chains** only. **show all chains**

Chain A : Mybpc3 protein

FASTA | **Sequence & DSSP** | **Image**

Polymer 1

Length: 124 residues

Chain Type: polypeptide(L)

Reference: **UniProtKB A9JR55**

Annotations *Add Annotations*

Select

Secondary Structure: **DSSP** 18% helical (4 helices; 23 residues)

[hide] **[reference]**

DSSP

PDB MGSSHHHHHHSSGLVPRGSHMHEAIGSGDLDLRSAFRRTSLAGAGRRTSDSHEDAGTLDF

PDB

DSSP

PDB SSLLKKRDSFRRDSKLEAPAEEDVWEILRQAPPSEYERIAFQHGVTDLRGMLKRLKGMKQ

PDB

315 320 330 340

DSSP

PDB DEKK

PDB

DSSP Legend

T: turn

empty: no secondary structure assigned

G: 3/10-helix

S: bend

File Edit Select View Annotations Format Colour Calculate

10 20 30

Query/63-98	VW-EILRQAPPSEYERIAFQHGVTDLRGMLKRLKGMK
H3BI37_LATCH/320-355	VW-DLLRQAPPSEYEKIAFQYGITDLRGMLKRLKKMK
F1NBZ9_CHICK/303-338	VW-EILRKAPPSEYEKIAFQYGITDLRGMLKRLKRIK
G3PS89_GASAC/335-370	VW-NILSHAPSSEYEKIAFQHGITDLRGMLKRLKKMK
W5KF69_ASTMX/356-391	VW-EILSHAPASEYEKIAFQYGITDLRGMLKRLKKMK
H3CVJ4_TETNG/330-365	VW-EILSKAPSSEYEKIAFQHGITDLRGMLKRLKKMK
H2LNL0_ORYLA/340-375	VW-SILSKAPPSAFEKIAFQYGITDLRGMLKRLKKMK
G1PGG8_MYOLU/177-212	VW-ELLKNANPNEYEKIAFQYGITDLRGMLKRLKRMR
M4A199_XIPMA/238-274	KLILKMGHAPPSEFEKIAFQYGITDLRGMLKRLKKMK
H2TW88_TAKRU/166-201	VW-EILKKARPDEYEKIAFTYGITDLRGLLRRMKKIP
F7DKM1_XENTR/222-257	VW-ELLKNAKPSDFERIAFEHGITDLRGMLKRLKKVK
H2UDR5_TAKRU/198-233	VW-ELLKSAHPSEYEKIAFQYGITDLRGMLKRLKKMK