

# 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

# Homology

# Definition:

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

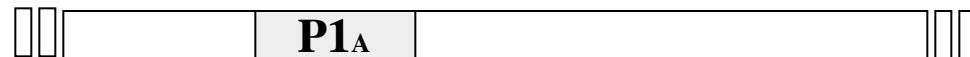
# Origins of homology in proteins

# Origin of homology in proteins

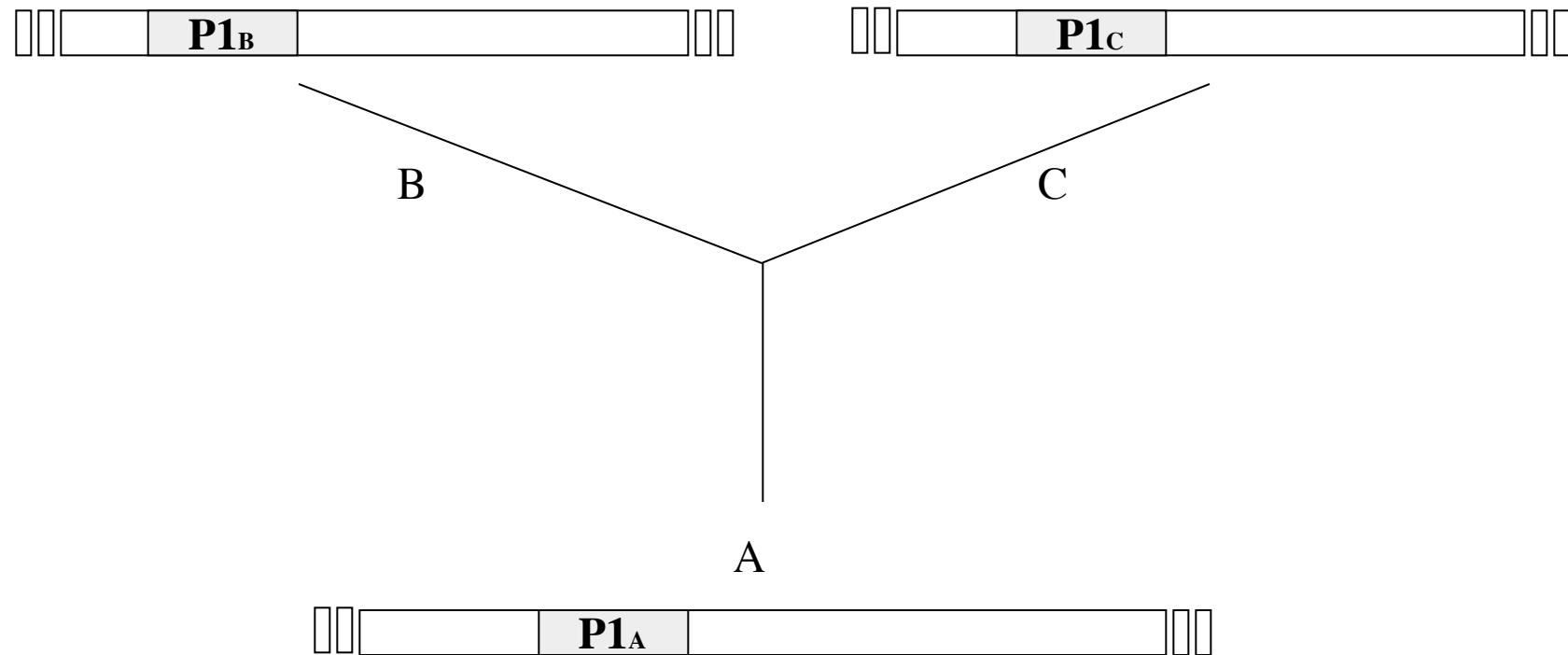
- Speciation (orthology)
- Gene duplication (paralogy)
- Horizontal gene transfer (xenology)
- Whole genome duplication (ohnology)
- Gametology, Synology

# Origin of homology in proteins

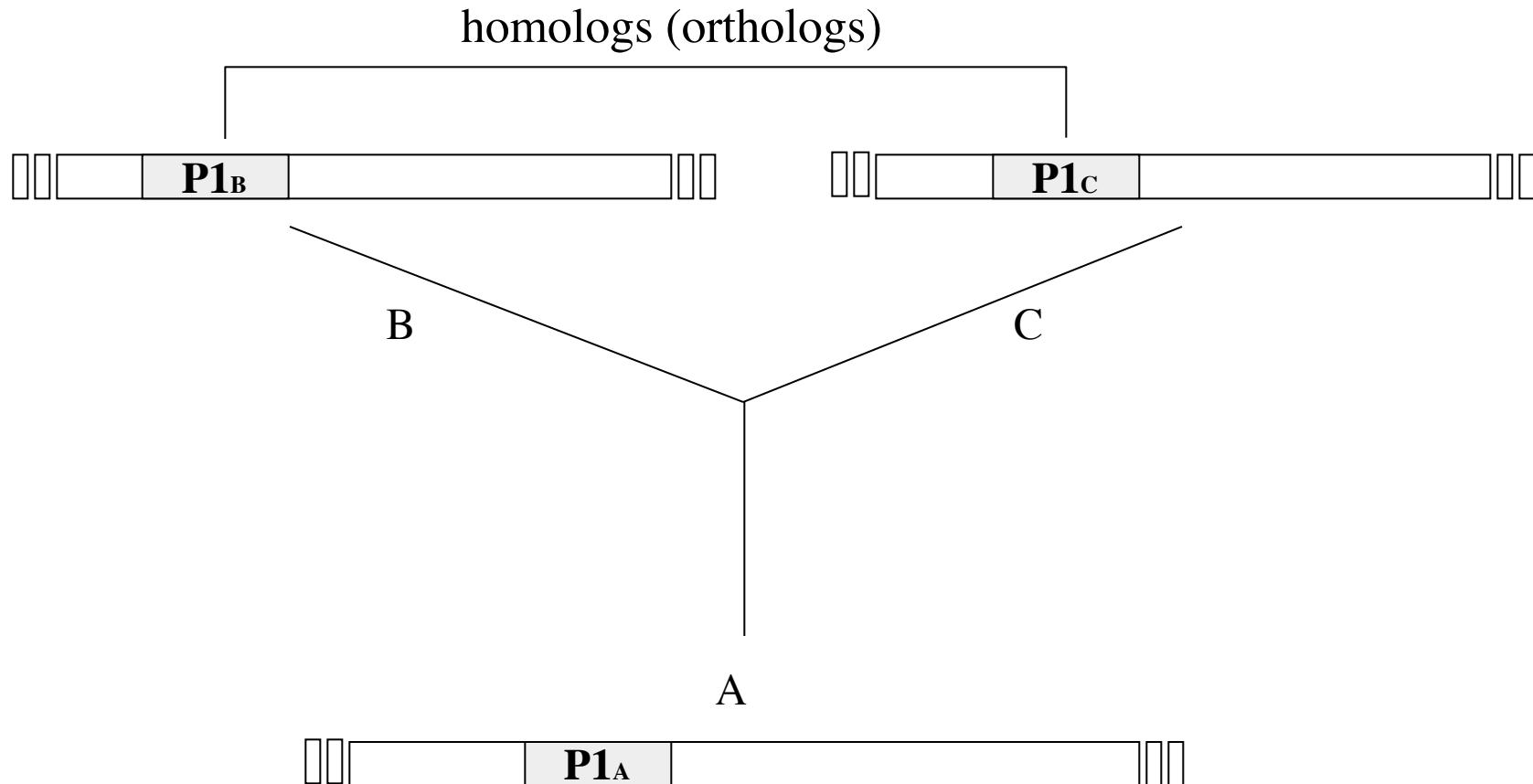
A



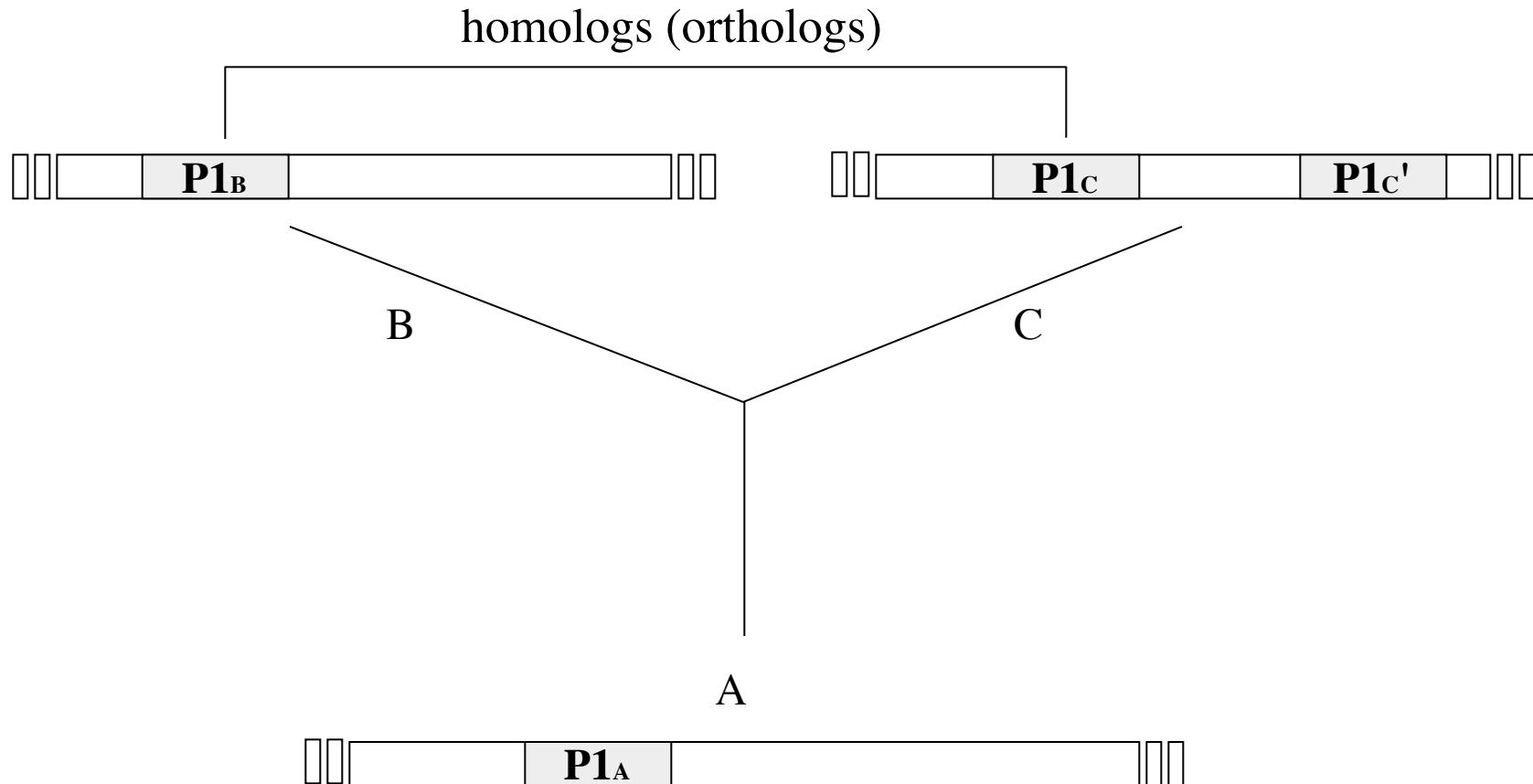
# Origin of homology in proteins



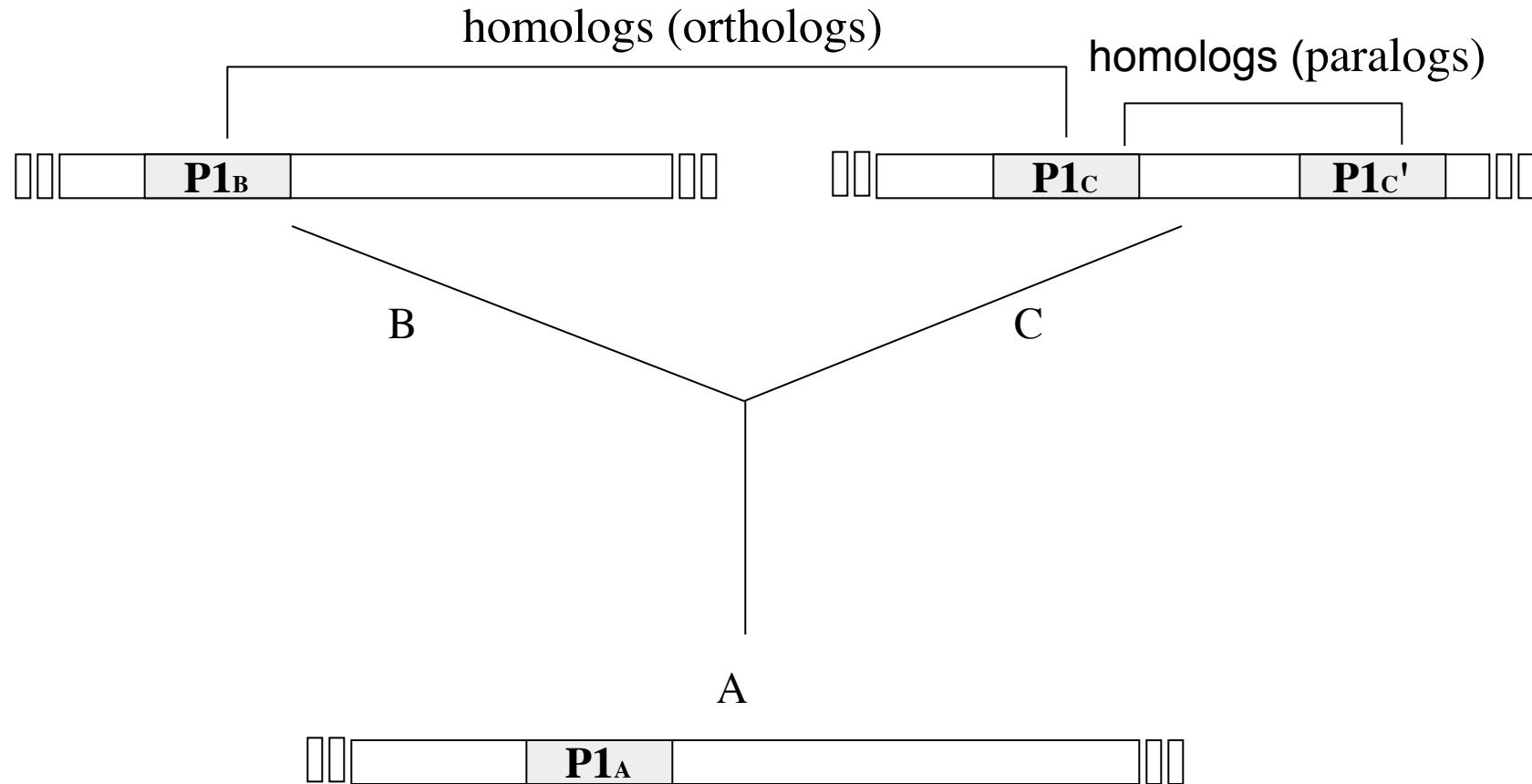
# Origin of homology in proteins



# Origin of homology in proteins

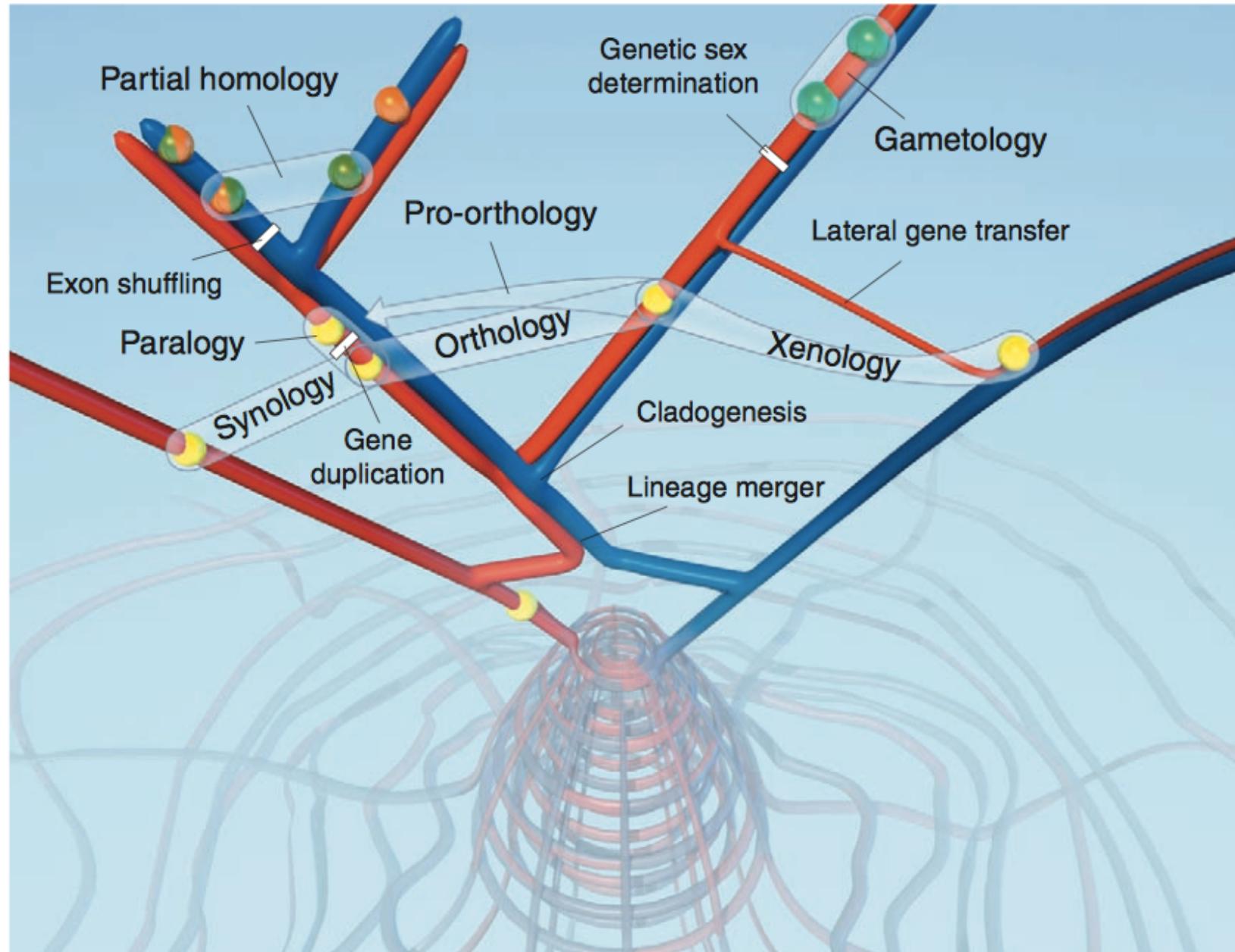


# Origin of homology in proteins



# Origin of homology in proteins

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



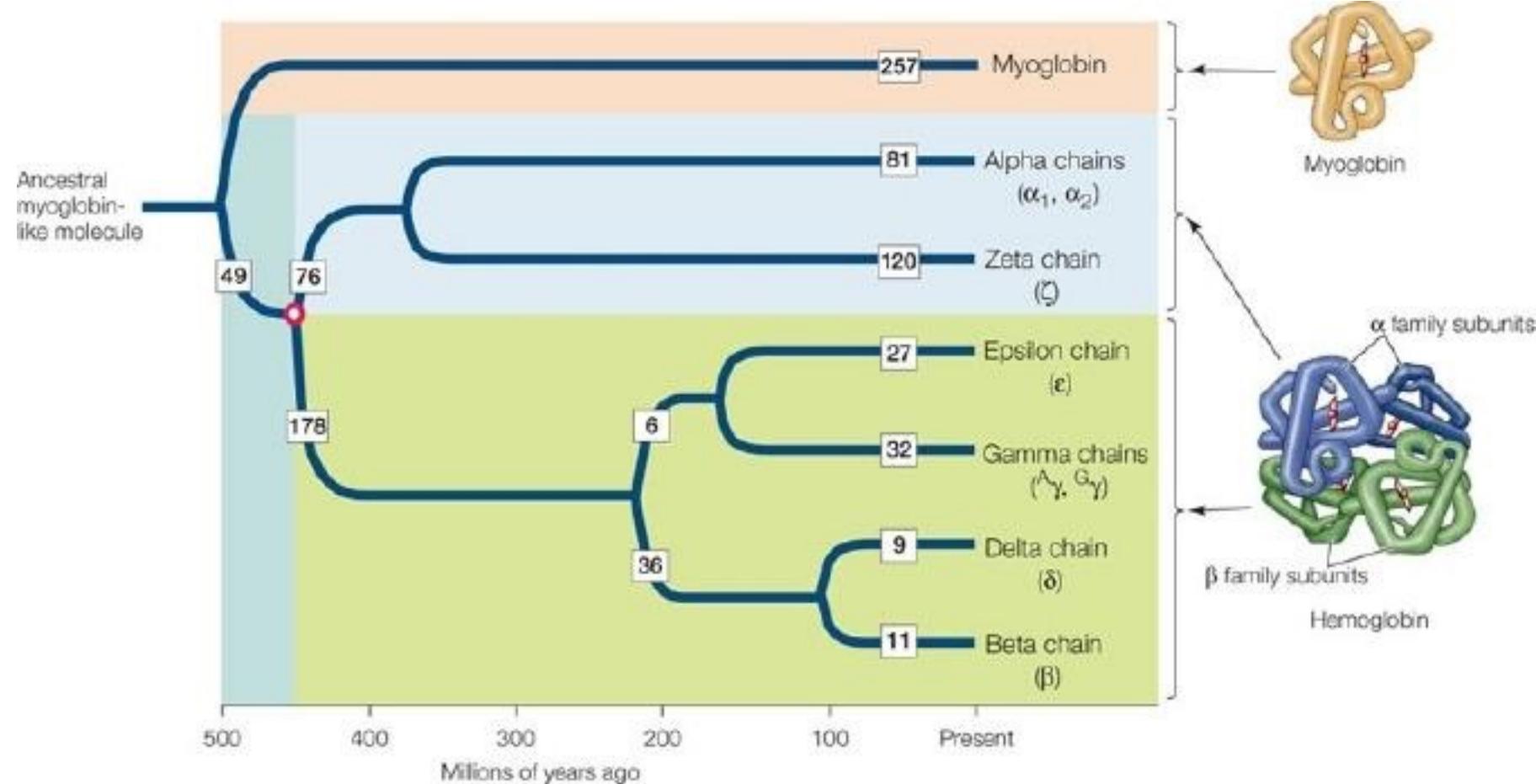
Mindell and Meyer Trends in Ecology and Evolution 2001

# Protein Families

# Definition:

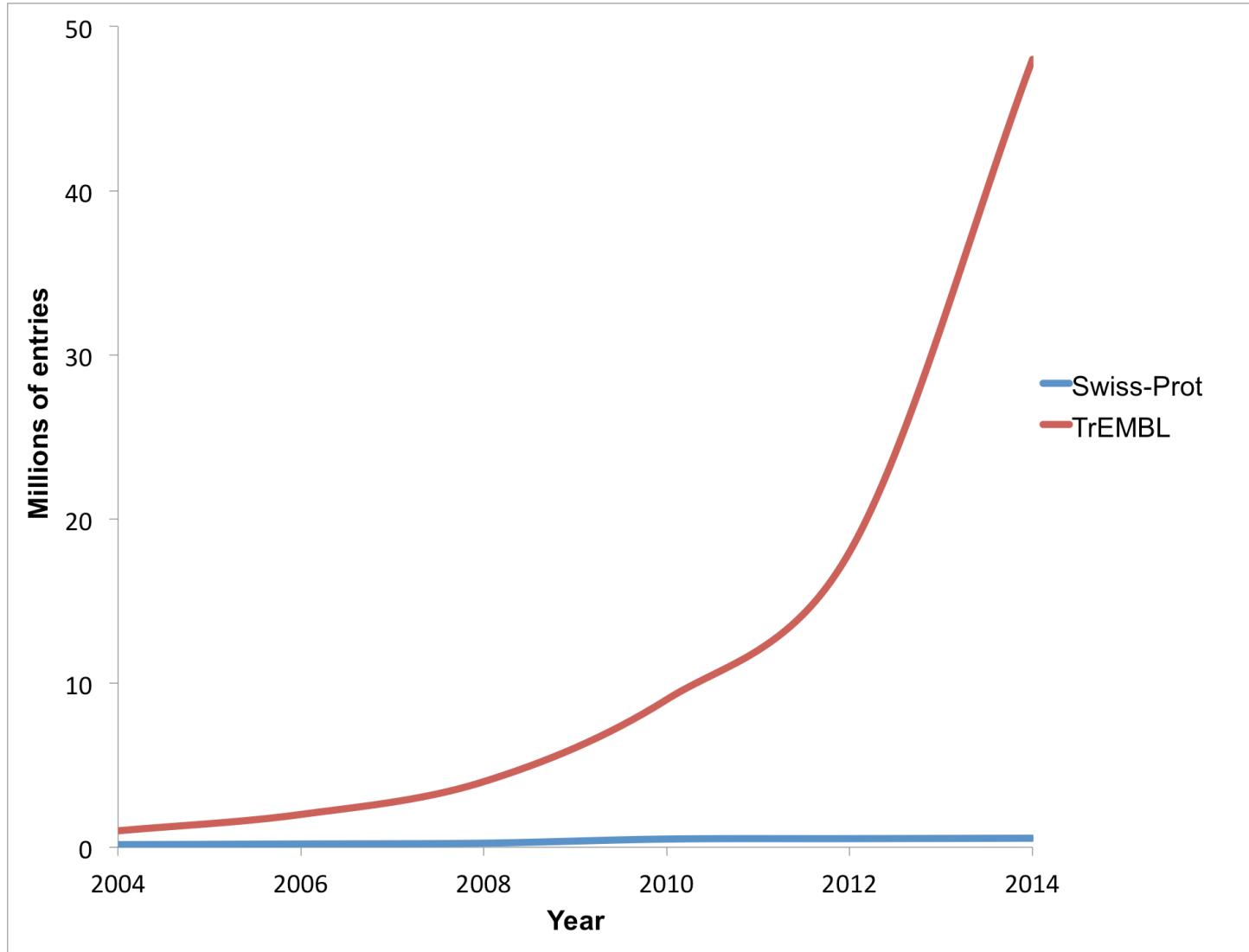
We call 'family' a group of evolutionary related proteins and/or protein regions

# Globins in Human



# Homology: why bother?

# Mind the gap!



# Mind the gap!

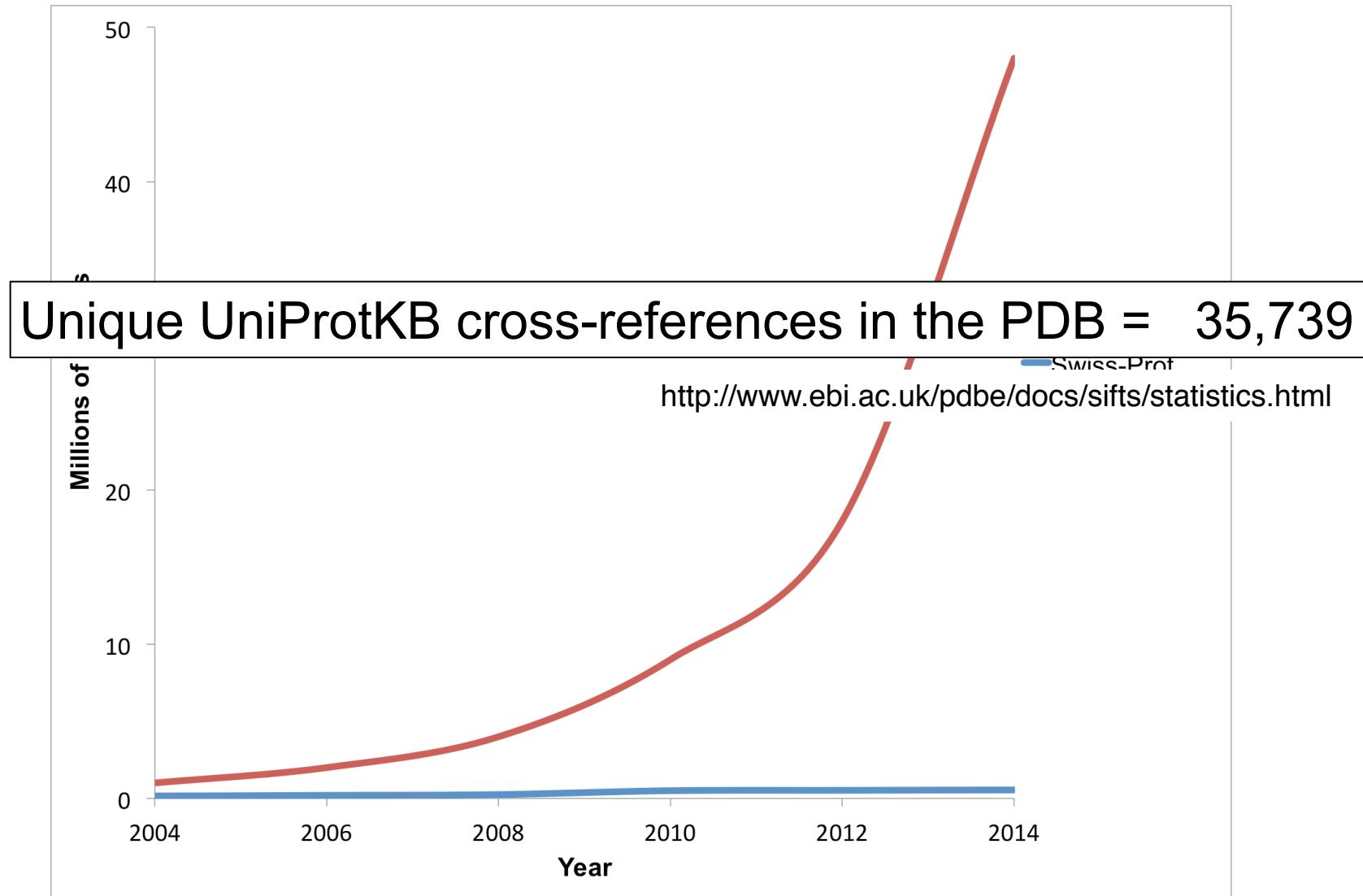


Figure courtesy of Alex Mitchell (EMBL-EBI)

EMBO Workshop, Norwich, 2015

Homologous protein regions have a  
similar (core) structure!

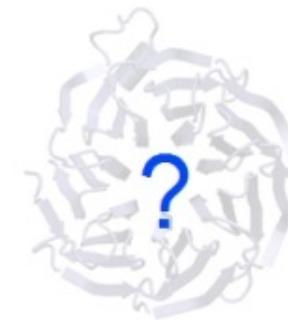
Chotia and Lesk *EMBO J* (1986)

# Homology Modelling

**Template**

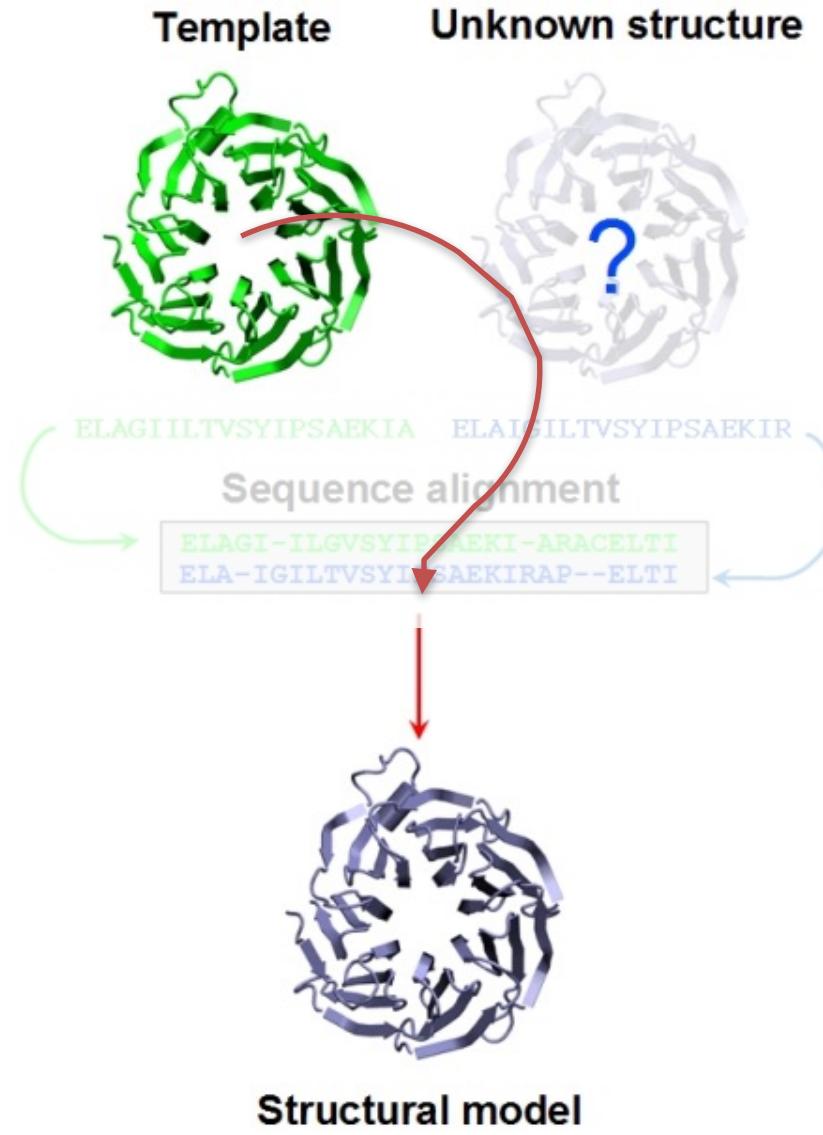


**Unknown structure**



# Homology Modelling

Marco Punta



# Homology: why bother?

Marco Punta

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<sup>1</sup>, Keith S. Wilson<sup>1</sup>, Evelyne Raux<sup>2</sup>, Sarah C. Woodcock<sup>2</sup> and Martin J. Warren<sup>2</sup>

Biosynthesis of the corrin ring of vitamin B<sub>12</sub> 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.

---

## articles

---

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

Heidi L. Schubert<sup>1</sup>, Keith S. Wilson<sup>1</sup>, Evelyne Raux<sup>2</sup>, Sarah C. Woodcock<sup>2</sup> and Martin J. Warren<sup>2</sup>

Biosynthesis of the corrin ring of vitamin B<sub>12</sub> 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.

---

## articles

---

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

Heidi L. Schubert<sup>1</sup>, Keith S. Wilson<sup>1</sup>, Evelyne Raux<sup>2</sup>, Sarah C. Woodcock<sup>2</sup> and Martin J. Warren<sup>2</sup>

Biosynthesis of the corrin ring of vitamin B<sub>12</sub> 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.

---

## articles

---

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

Heidi L. Schubert<sup>1</sup>, Keith S. Wilson<sup>1</sup>, Evelyne Raux<sup>2</sup>, Sarah C. Woodcock<sup>2</sup> and Martin J. Warren<sup>2</sup>

Biosynthesis of the corrin ring of vitamin B<sub>12</sub> 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.

---

## articles

---

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

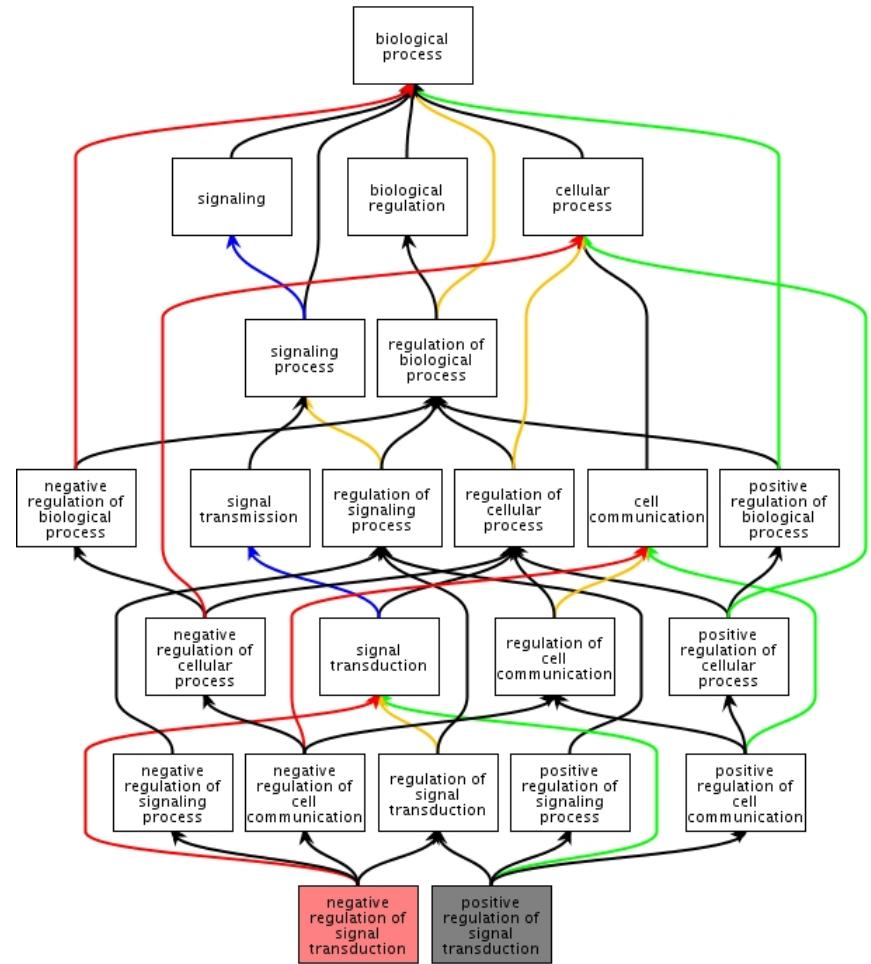
Heidi L. Schubert<sup>1</sup>, Keith S. Wilson<sup>1</sup>, Evelyne Raux<sup>2</sup>, Sarah C. Woodcock<sup>2</sup> and Martin J. Warren<sup>2</sup>

Biosynthesis of the corrin ring of vitamin B<sub>12</sub> 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.

# 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



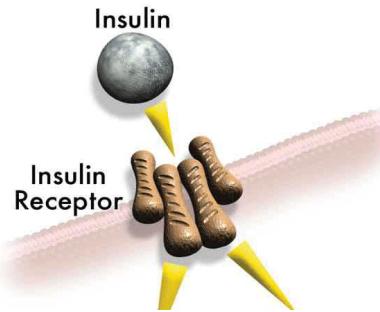
[www.ebi.ac.uk/QuickGO](http://www.ebi.ac.uk/QuickGO)

# GO: 3 ontologies in 1

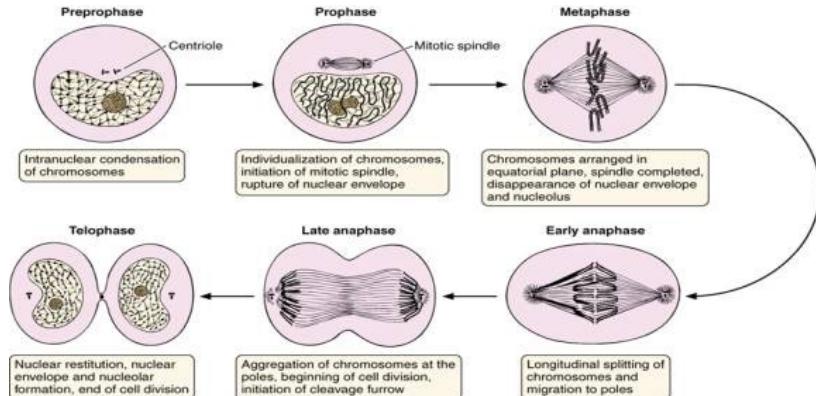
Marco Punta

## 1. Molecular Function

An elemental activity or task or job



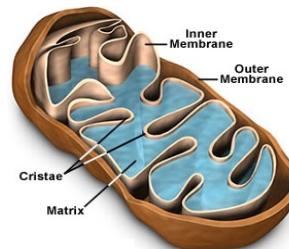
- protein kinase activity
- insulin receptor activity



## 2. Biological Process

A commonly recognised series of events

- cell division



- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

## 3. Cellular Component

Where a gene product is located

# 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
<b>Process</b>													
UniProtKB	O87696	cbiF		<a href="#">GO:0006779</a>	porphyrin-containing compound biosynthetic process	P	IEA	InterPro2GO	InterPro:IPR003043		1404	20150919	InterPro
UniProtKB	O87696	cbiF		<a href="#">GO:0008152</a>	metabolic process	P	IEA	InterPro2GO	InterPro:IPR000878 InterPro:IPR014776 InterPro:IPR014777		1404	20150919	InterPro
UniProtKB	O87696	cbiF		<a href="#">GO:0009236</a>	cobalamin biosynthetic process	P	IEA	InterPro2GO	InterPro:IPR006362		1404	20150919	InterPro
UniProtKB	O87696	cbiF		<a href="#">GO:0009236</a>	cobalamin biosynthetic process	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0169		1404	20150919	UniProt
UniProtKB	O87696	cbiF		<a href="#">GO:0009236</a>	cobalamin biosynthetic process	P	IEA	UniPathway2GO	UniPathway:UPA00148		1404	20150912	UniProt
UniProtKB	O87696	cbiF		<a href="#">GO:0032259</a>	methylation	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0489		1404	20150919	UniProt
UniProtKB	O87696	cbiF		<a href="#">GO:0055114</a>	oxidation-reduction process	P	IEA	InterPro2GO	InterPro:IPR003043		1404	20150919	InterPro
<b>Function</b>													
UniProtKB	O87696	cbiF		<a href="#">GO:0008168</a>	methyltransferase activity	F	IEA	InterPro2GO	InterPro:IPR000878 InterPro:IPR003043 InterPro:IPR014776 InterPro:IPR014777	1404	20150919	InterPro	
UniProtKB	O87696	cbiF		<a href="#">GO:0008168</a>	methyltransferase activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0489		1404	20150919	UniProt
UniProtKB	O87696	cbiF		<a href="#">GO:0016740</a>	transferase activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0808		1404	20150919	UniProt
UniProtKB	O87696	cbiF		<a href="#">GO:0043115</a>	precorrin-2 dehydrogenase activity	F	IEA	InterPro2GO	InterPro:IPR003043		1404	20150919	InterPro
UniProtKB	O87696	cbiF		<a href="#">GO:0046026</a>	precorrin-4 C11-methyltransferase activity	F	IEA	InterPro2GO	InterPro:IPR006362		1404	20150919	InterPro

Database	Gene Product Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference
UniProtKB	P02144	MB	<a href="#">GO:0001666</a>	response to hypoxia	P	IEA	Ensembl
UniProtKB	P02144	MB	<a href="#">GO:0006810</a>	transport	P	IEA	UniProt
UniProtKB	P02144	MB	<a href="#">GO:0007507</a>	heart development	P	IEA	Ensembl
UniProtKB	P02144	MB	<a href="#">GO:0009725</a>	response to hormone	P	IEA	Ensembl
UniProtKB	P02144	MB	<a href="#">GO:0015671</a>	oxygen transport	P	IEA	InterPro
UniProtKB	P02144	MB	<a href="#">GO:0015671</a>	oxygen transport	P	IEA	Ensembl
UniProtKB	P02144	MB	<a href="#">GO:0015671</a>	oxygen transport	P	IEA	UniProt
UniProtKB	P02144	MB	<a href="#">GO:0031444</a>	slow-twitch skeletal muscle fiber contraction	P	IEA	Ensembl
UniProtKB	P02144	MB	<a href="#">GO:0042542</a>	response to hydrogen peroxide	P	IEA	Ensembl
UniProtKB	P02144	MB	<a href="#">GO:0043353</a>	enucleate erythrocyte differentiation	P	IEA	Ensembl
UniProtKB	P02144	MB	<a href="#">GO:0050873</a>	brown fat cell differentiation	P	IEA	Ensembl
UniProtKB	P02144	MB	<a href="#">GO:0005344</a>	oxygen transporter activity	F	IEA	UniProt
UniProtKB	P02144	MB	<a href="#">GO:0005506</a>	iron ion binding	F	IEA	InterPro
UniProtKB	P02144	MB	<a href="#">GO:0019825</a>	oxygen binding	F	IEA	InterPro
UniProtKB	P02144	MB	<a href="#">GO:0019825</a>	oxygen binding	F	IEA	Ensembl
UniProtKB	P02144	MB	<a href="#">GO:0020037</a>	heme binding	F	IEA	InterPro
UniProtKB	P02144	MB	<a href="#">GO:0046872</a>	metal ion binding	F	IEA	UniProt
UniProtKB	P02144	MB	<a href="#">GO:0070062</a>	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
<b>Process</b>													
JniProtKB	P02144	MB		<a href="#">GO:0001666</a> response to hypoxia	P	IEA	Ensembl Compara		Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
JniProtKB	P02144	MB		<a href="#">GO:0006810</a> transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)		UniProtKB-KW:KW-0813	9606	20140913	UniProt	
JniProtKB	P02144	MB		<a href="#">GO:0007507</a> heart development	P	IEA	Ensembl Compara		Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
JniProtKB	P02144	MB		<a href="#">GO:0009725</a> response to hormone	P	IEA	Ensembl Compara		Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
JniProtKB	P02144	MB		<a href="#">GO:0015671</a> oxygen transport	P	IEA	InterPro2GO		InterPro:IPR002335  InterPro:IPR012292	9606	20140913	InterPro	
JniProtKB	P02144	MB		<a href="#">GO:0015671</a> oxygen transport	P	IEA	Ensembl Compara		Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
JniProtKB	P02144	MB		<a href="#">GO:0015671</a> oxygen transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)		UniProtKB-KW:KW-0561	9606	20140913	UniProt	
JniProtKB	P02144	MB		<a href="#">GO:0031444</a> slow-twitch skeletal muscle fiber contraction	P	IEA	Ensembl Compara		Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
JniProtKB	P02144	MB		<a href="#">GO:0042542</a> response to hydrogen peroxide	P	IEA	Ensembl Compara		Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
JniProtKB	P02144	MB		<a href="#">GO:0043353</a> enucleate erythrocyte differentiation	P	IEA	Ensembl Compara		Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
JniProtKB	P02144	MB		<a href="#">GO:0050873</a> brown fat cell differentiation	P	IEA	Ensembl Compara		Ensembl:ENSMUSP00000125995	9606	20140913	Ensembl	
<b>Function</b>													
JniProtKB	P02144	MB		<a href="#">GO:0005344</a> oxygen transporter activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)		UniProtKB-KW:KW-0561	9606	20140913	UniProt	
JniProtKB	P02144	MB		<a href="#">GO:0005506</a> iron ion binding	F	IEA	InterPro2GO		InterPro:IPR000971  InterPro:IPR012292	9606	20140913	InterPro	
JniProtKB	P02144	MB		<a href="#">GO:0019825</a> oxygen binding	F	IEA	InterPro2GO		InterPro:IPR002335  InterPro:IPR012292	9606	20140913	InterPro	
JniProtKB	P02144	MB		<a href="#">GO:0019825</a> oxygen binding	F	IEA	Ensembl Compara		Ensembl:ENSRNOP00000006184	9606	20140913	Ensembl	
JniProtKB	P02144	MB		<a href="#">GO:0020037</a> heme binding	F	IEA	InterPro2GO		InterPro:IPR000971  InterPro:IPR002335  InterPro:IPR012292	9606	20140913	InterPro	
JniProtKB	P02144	MB		<a href="#">GO:0046872</a> metal ion binding	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)		UniProtKB-KW:KW-0479	9606	20140913	UniProt	
<b>Component</b>													
JniProtKB	P02144	MB		<a href="#">GO:0070062</a> 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
<b>Process</b>													
UniProtKB	P02008	HBZ		<a href="#">GO:0000122</a> negative regulation of transcription from RNA polymerase II promoter	P	IEA	Ensembl Compara		Ensembl:ENSMUSP00000020531	9606	20140913	Ensembl	
UniProtKB	P02008	HBZ		<a href="#">GO:0006810</a> transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)		UniProtKB-KW:KW-0813	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		<a href="#">GO:0015671</a> oxygen transport	P	IEA	InterPro2GO		InterPro:IPR002338  InterPro:IPR002340  InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		<a href="#">GO:0015671</a> oxygen transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)		UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		<a href="#">GO:0043249</a> erythrocyte maturation	P	IEA	Ensembl Compara		Ensembl:ENSMUSP00000020531	9606	20140913	Ensembl	
UniProtKB	P02008	HBZ		<a href="#">GO:0005344</a> oxygen transporter activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)		UniProtKB-KW:KW-0561	9606	20140913	UniProt	
UniProtKB	P02008	HBZ		<a href="#">GO:0005344</a> oxygen transporter activity	F	TAS	PMID:7555018			9606	20030904	PINC	
UniProtKB	P02008	HBZ		<a href="#">GO:0005506</a> iron ion binding	F	IEA	InterPro2GO		InterPro:IPR000971  InterPro:IPR002338  InterPro:IPR002340  InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		<a href="#">GO:0005515</a> protein binding	F	IPI	PMID:11159543		UniProtKB:P68871	9606	20140914	IntAct	
UniProtKB	P02008	HBZ		<a href="#">GO:0005515</a> protein binding	F	IPI	PMID:6683087		UniProtKB:P68871	9606	20140914	IntAct	
UniProtKB	P02008	HBZ		<a href="#">GO:0019825</a> oxygen binding	F	IEA	InterPro2GO		InterPro:IPR002338  InterPro:IPR002340  InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		<a href="#">GO:0020037</a> heme binding	F	IEA	InterPro2GO		InterPro:IPR000971  InterPro:IPR002338  InterPro:IPR002340  InterPro:IPR012292	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		<a href="#">GO:0046872</a> metal ion binding	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)		UniProtKB-KW:KW-0479	9606	20140913	UniProt	
<b>Component</b>													
UniProtKB	P02008	HBZ		<a href="#">GO:0005833</a> hemoglobin complex	C	IEA	InterPro2GO		InterPro:IPR002338  InterPro:IPR002340	9606	20140913	InterPro	
UniProtKB	P02008	HBZ		<a href="#">GO:0005833</a> hemoglobin complex	C	TAS	PMID:7555018			9606	20030904	PINC	
UniProtKB	P02008	HBZ		<a href="#">GO:0070062</a> extracellular vesicular exosome	C	IDA	PMID:23533145			9606	20140714	UniProt	

Database	Gene Product ID	Symbol Qualifier	GO Identifier	GO Term Name	Aspect Evidence Reference			With	Taxon	Date	Assigned Product By	Form ID
<b>Process</b>												
UniProtKB P02144	MB	<a href="#">GO:0001666</a>	response to hypoxia	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995		9606	20140913	Ensembl	
UniProtKB P02144	MB	<a href="#">GO:0006810</a>	transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0813		9606	20140913	UniProt	
UniProtKB P02144	MB	<a href="#">GO:0007507</a>	heart development	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995		9606	20140913	Ensembl	
UniProtKB P02144	MB	<a href="#">GO:0009725</a>	response to hormone	P	IEA	Ensembl Compara	Ensembl:ENSRNOP0000006184		9606	20140913	Ensembl	
UniProtKB P02144	MB	<a href="#">GO:0015671</a>	oxygen transport	P	IEA	InterPro2GO	InterPro:IPR002335  InterPro:IPR012292		9606	20140913	InterPro	
UniProtKB P02144	MB	<a href="#">GO:0015671</a>	oxygen transport	P	IEA	Ensembl Compara	Ensembl:ENSRNOP0000006184		9606	20140913	Ensembl	
UniProtKB P02144	MB	<a href="#">GO:0015671</a>	oxygen transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561		9606	20140913	UniProt	
UniProtKB P02144	MB	<a href="#">GO:0031444</a>	slow-twitch skeletal muscle fiber contraction	P	IEA	Ensembl Compara	Ensembl:ENSRNOP0000006184		9606	20140913	Ensembl	
UniProtKB P02144	MB	<a href="#">GO:0042542</a>	response to hydrogen peroxide	P	IEA	Ensembl Compara	Ensembl:ENSRNOP0000006184		9606	20140913	Ensembl	
UniProtKB P02144	MB	<a href="#">GO:0043353</a>	enucleate erythrocyte differentiation	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995		9606	20140913	Ensembl	
UniProtKB P02144	MB	<a href="#">GO:0050873</a>	brown fat cell differentiation	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000125995		9606	20140913	Ensembl	
<b>Function</b>												
UniProtKB P02144	MB	<a href="#">GO:0005344</a>	oxygen transporter activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561		9606	20140913	UniProt	
UniProtKB P02144	MB	<a href="#">GO:0005506</a>	iron ion binding	F	IEA	InterPro2GO	InterPro:IPR000971  InterPro:IPR012292		9606	20140913	InterPro	
UniProtKB P02144	MB	<a href="#">GO:0019825</a>	oxygen binding	F	IEA	InterPro2GO	InterPro:IPR002335  InterPro:IPR012292		9606	20140913	InterPro	
UniProtKB P02144	MB	<a href="#">GO:0019825</a>	oxygen binding	F	IEA	Ensembl Compara	Ensembl:ENSRNOP0000006184		9606	20140913	Ensembl	
UniProtKB P02144	MB	<a href="#">GO:0020037</a>	heme binding	F	IEA	InterPro2GO	InterPro:IPR000971  InterPro:IPR002335  InterPro:IPR012292		9606	20140913	InterPro	
UniProtKB P02144	MB	<a href="#">GO:0046872</a>	metal ion binding	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0479		9606	20140913	UniProt	
<b>Component</b>												
UniProtKB P02144	MB	<a href="#">GO:0070062</a>	extracellular vesicular exosome	C	IDA	PMID:23533145			9606	20140714	UniProt	

Database	Gene Product ID	Symbol Qualifier	GO Identifier	GO Term Name	Aspect Evidence Reference			With	Taxon	Date	Assigned Product By	Form ID
<b>Process</b>												
UniProtKB P02008	HBZ	<a href="#">GO:0000122</a>	negative regulation of transcription from RNA polymerase II promoter	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000020531		9606	20140913	Ensembl	
UniProtKB P02008	HBZ	<a href="#">GO:0006810</a>	transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0813		9606	20140913	UniProt	
UniProtKB P02008	HBZ	<a href="#">GO:0015671</a>	oxygen transport	P	IEA	InterPro2GO	InterPro:IPR002338  InterPro:IPR002340  InterPro:IPR012292		9606	20140913	InterPro	
UniProtKB P02008	HBZ	<a href="#">GO:0015671</a>	oxygen transport	P	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561		9606	20140913	UniProt	
UniProtKB P02008	HBZ	<a href="#">GO:0043249</a>	erythrocyte maturation	P	IEA	Ensembl Compara	Ensembl:ENSMUSP00000020531		9606	20140913	Ensembl	
UniProtKB P02008	HBZ	<a href="#">GO:0005344</a>	oxygen transporter activity	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0561		9606	20140913	UniProt	
UniProtKB P02008	HBZ	<a href="#">GO:0005344</a>	oxygen transporter activity	F	TAS	PMID:7555018			9606	20030904	PINC	
UniProtKB P02008	HBZ	<a href="#">GO:0005506</a>	iron ion binding	F	IEA	InterPro2GO	InterPro:IPR000971  InterPro:IPR002338  InterPro:IPR002340  InterPro:IPR012292		9606	20140913	InterPro	
UniProtKB P02008	HBZ	<a href="#">GO:0005515</a>	protein binding	F	IPI	PMID:1159843	InterPro:IPR000971  InterPro:IPR002338  InterPro:IPR002340  InterPro:IPR012292		9606	20140914	IntAct	
UniProtKB P02008	HBZ	<a href="#">GO:0005515</a>	protein binding	F	IPI	PMID:67871	InterPro:IPR000971  InterPro:IPR002338  InterPro:IPR002340  InterPro:IPR012292		9606	20140914	IntAct	
UniProtKB P02008	HBZ	<a href="#">GO:0019825</a>	oxygen binding	F	IEA	InterPro2GO	InterPro:IPR002338  InterPro:IPR002340  InterPro:IPR012292		9606	20140913	InterPro	
UniProtKB P02008	HBZ	<a href="#">GO:0020037</a>	heme binding	F	IEA	InterPro2GO	InterPro:IPR000971  InterPro:IPR002338  InterPro:IPR002340  InterPro:IPR012292		9606	20140913	InterPro	
UniProtKB P02008	HBZ	<a href="#">GO:0046872</a>	metal ion binding	F	IEA	UniProt Keywords2GO (UniProtKB/Swiss-Prot entries)	UniProtKB-KW:KW-0479		9606	20140913	UniProt	
<b>Component</b>												
UniProtKB P02008	HBZ	<a href="#">GO:0005833</a>	hemoglobin complex	C	IEA	InterPro2GO	InterPro:IPR002338  InterPro:IPR002340		9606	20140913	InterPro	
UniProtKB P02008	HBZ	<a href="#">GO:0005833</a>	hemoglobin complex	C	TAS	PMID:7555018			9606	20030904	PINC	
UniProtKB P02008	HBZ	<a href="#">GO:0070062</a>	extracellular vesicular exosome	C	IDA	PMID:23533145			9606	20140714	UniProt	

# 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:

Marco Punta

- Functional motifs
- Conservation of functional residues
- Genomic context (mostly in bacteria)

If structure available:

- structural motifs
- Electrostatic, cavities, etc.

# Detecting homology

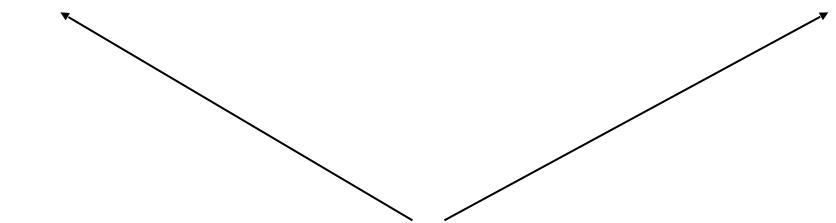
# From sequence

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

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

ALHWRAA**L**AATVLLVIVLLAGS**W**LAVLAE

ALHW**K**AAGAATVLLVIVLLAGSYLAVLAE



ALHWRAAGAATVLLVIVLLAGSYLAVLAE

Human: 1 MGLSDGEWQLVLNWGKVEADIPGHQEVLIIRLFKGHPETLEKFDKFHLKSEDEMKA 60  
MGLSDGEWQLVLNWGKVEAD GHGQEVLI LFK HPETL KFDKFK LKSE MK SE

Mouse: 1 MGLSDGEWQLVLNWGKVEADLAGHGQEVLIAGLKFTHPETLDKFDKFNLKSEEDMKG 60

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

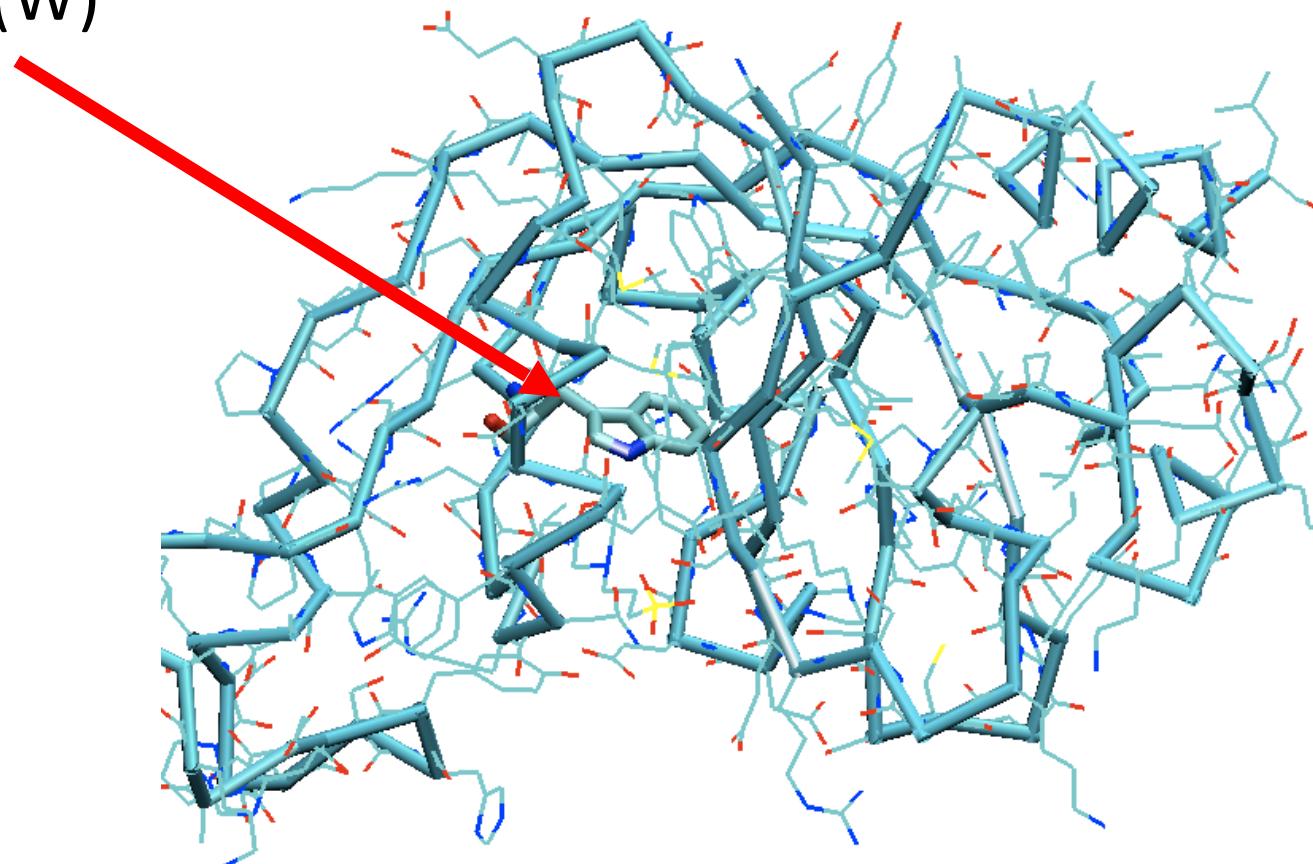
Mouse: 61 DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIEVLKKRH 120

Human: 121 PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154  
GDFGADAQGAM KALELFR D A YKELGFQG

Mouse: 121 SGDFGADAQGAMSKALELFRNDIAAKYKELGFQG 154

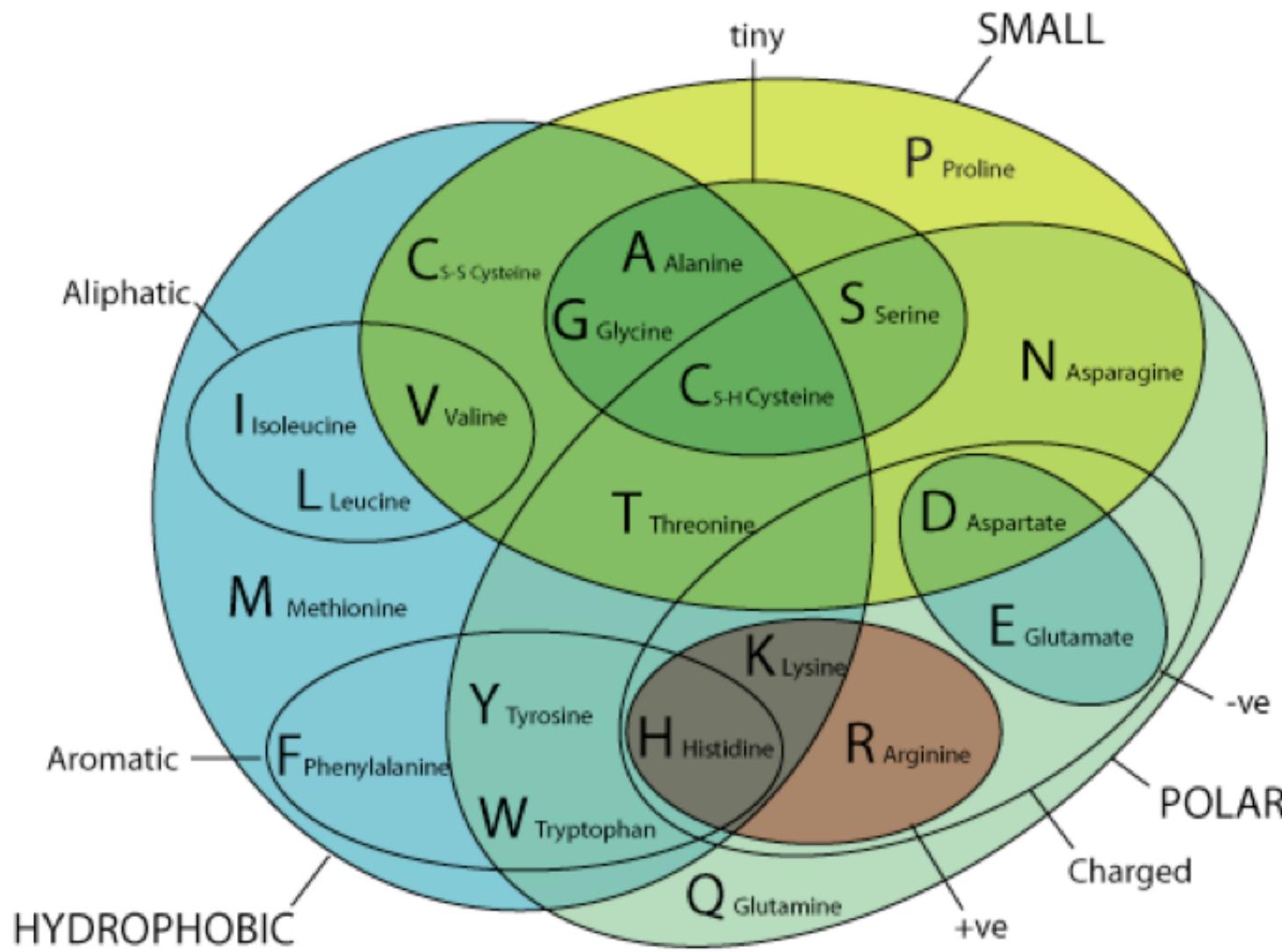
# Protein structural and functional constraints

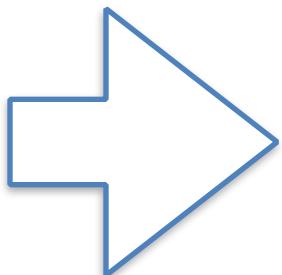
Trp (W)



# aa physico-chemical properties

Marco Punta





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

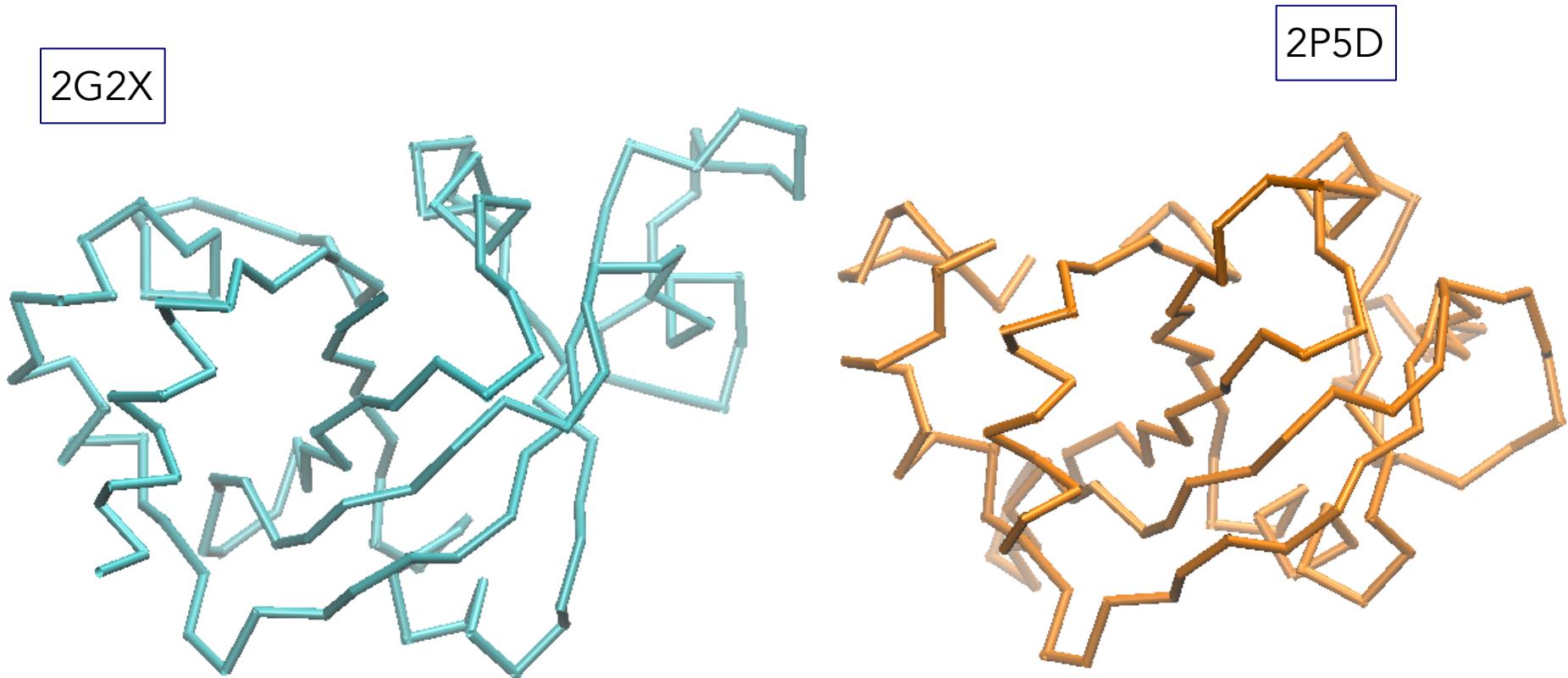
# BLOSUM62 matrix

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

# 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, ...

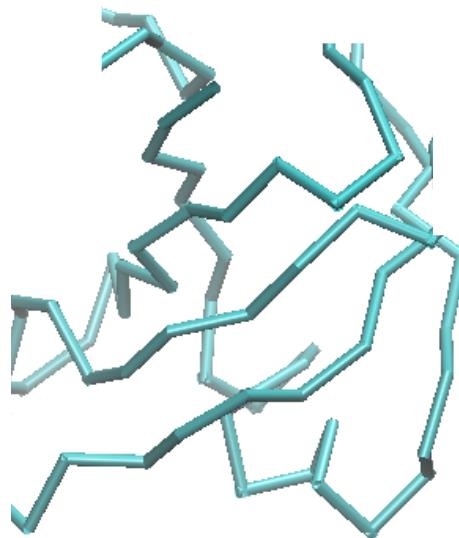
# From structure



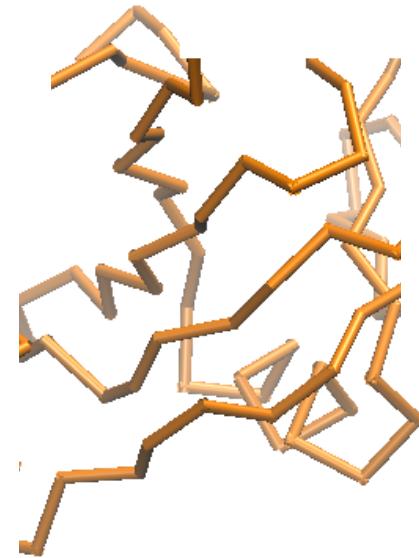
# Structural similarity

Marco Punta

2G2X



2P5D

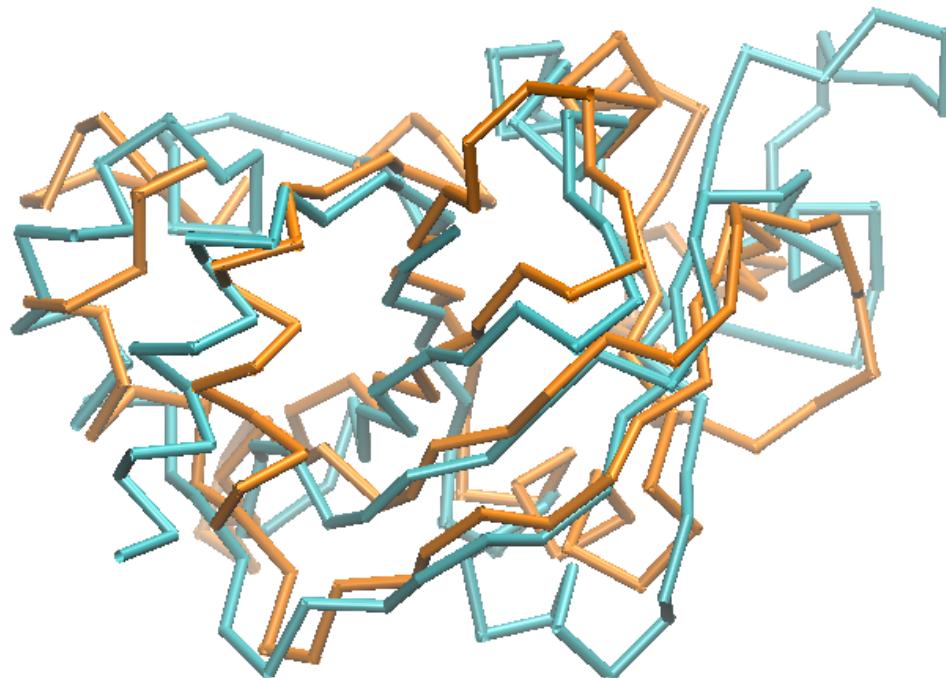


# Structural similarity

Marco Punta

2G2X

2P5D



Z-score = 12.2

RMSD = 2.9

Lali = 122

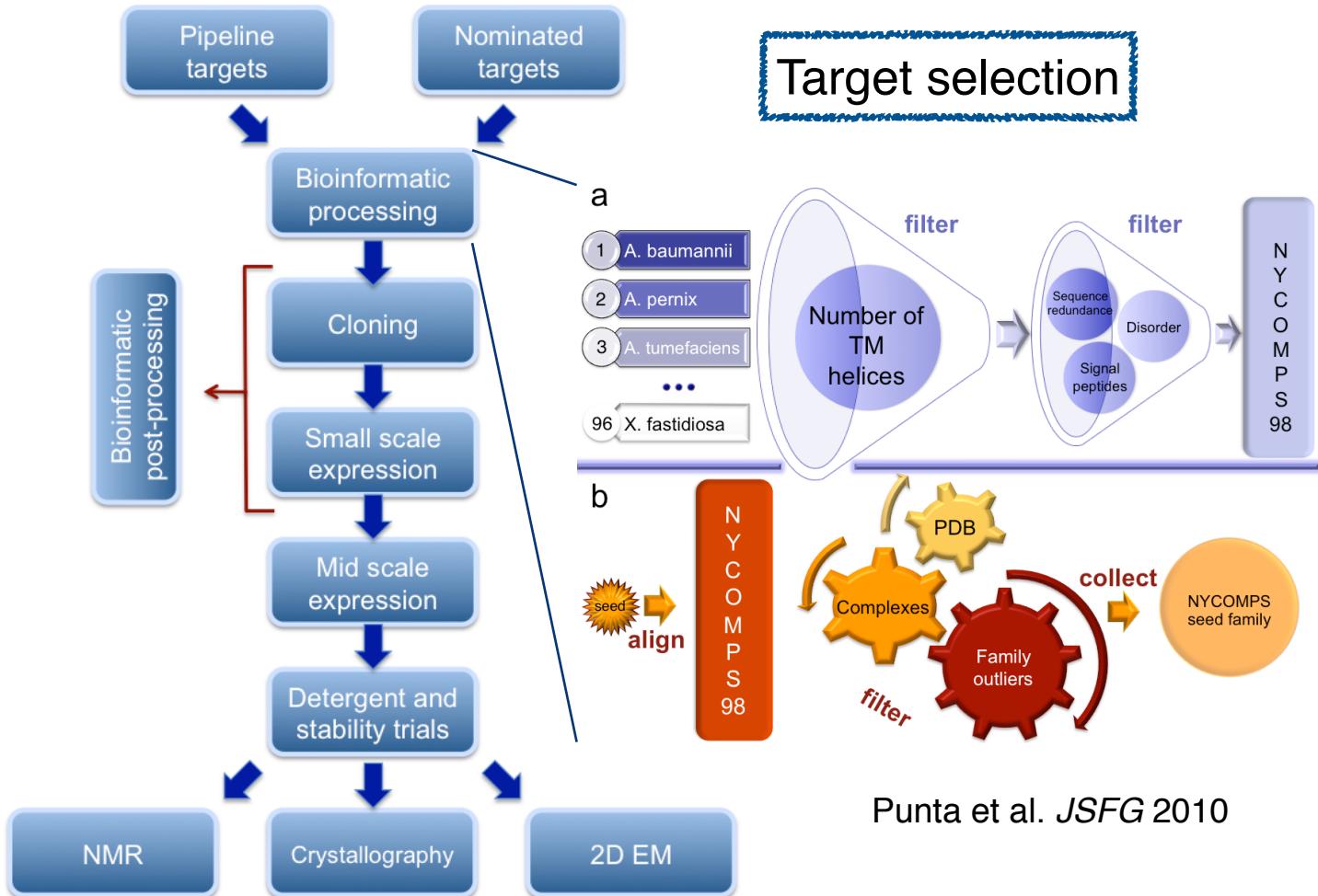
%id = 20

DALI: [http://ekhidna.biocenter.helsinki.fi/dali\\_lite/start](http://ekhidna.biocenter.helsinki.fi/dali_lite/start)

# Exercise

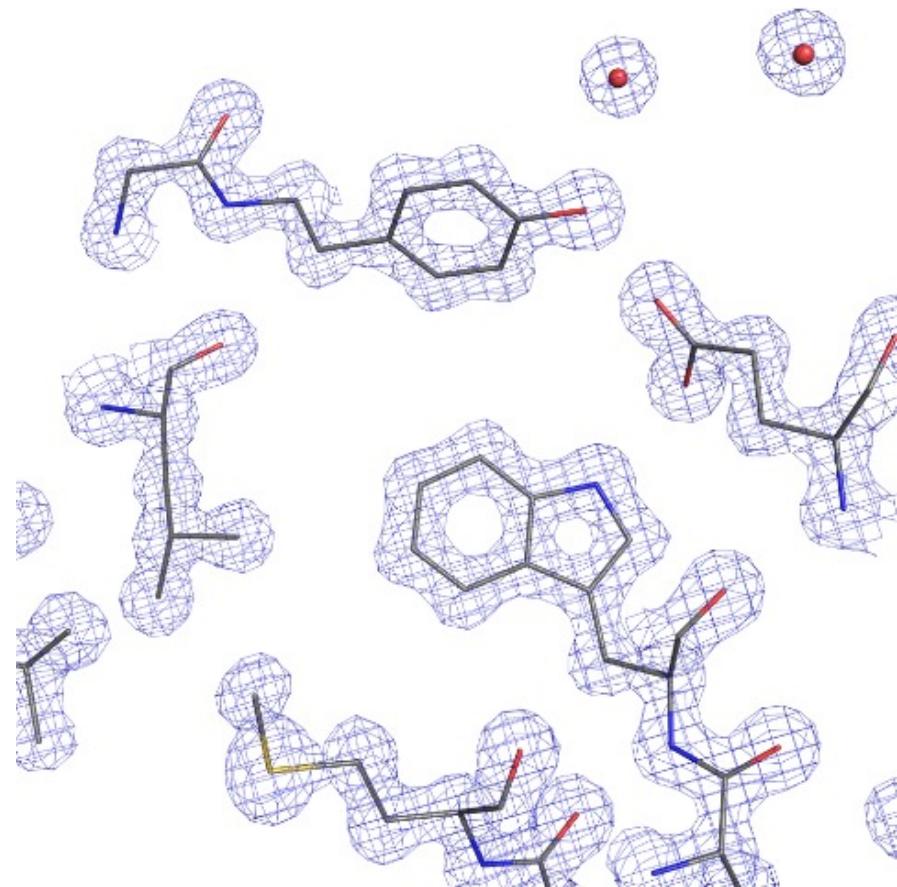
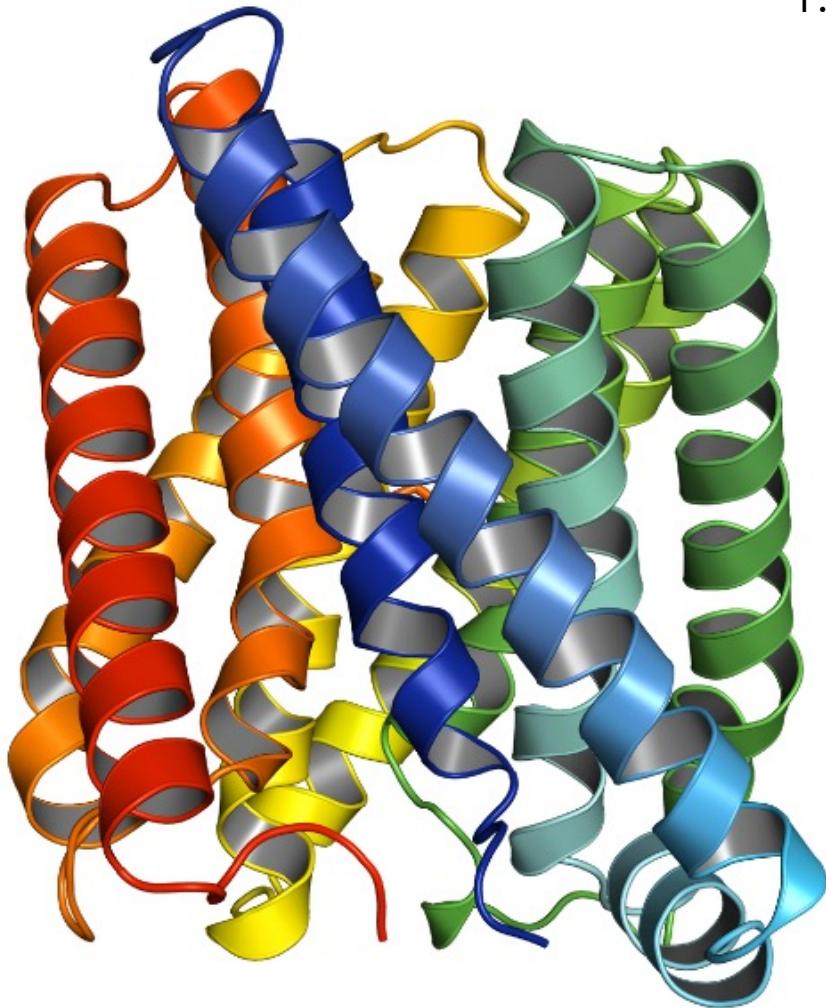
Homology-based function annotation transfer #1

# NYCOMPS pipeline

Punta et al. *JSFG* 2010Love et al. *JSFG* 2010

H. influenzae protein [3M71] ← PDB id  
1.20 Å

Marco Punta



Chen et al. *Nature* 467 (2010)

## Alignment

Q9LD83 SLAC1\_ARATH - Guard cell S-type anion channel SLA... - Arabidopsis thal...

**E-value:** 3e-10

**Score:** 160

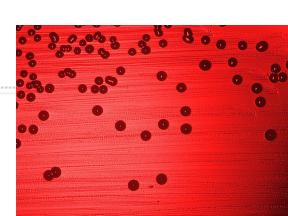
**Ident.:** 22.0%

**Positives :** 41.0%

**Query Length:** 328

**Match Length:** 556

			
P44741	20	PFPL--PTGYFGIPLGLAALSLAWFHLE-----NLFPAAARMVSDVLGIVASAVWILFILM	72
		PF L P G FGI LGL++ ++ W L N +++ V+ + + V +	
Q9LD83	183	PFLLRFPICGCFGICLGLSSQAVLWLALAKSPATNFLHITPLINLVVWLFLSVLVSFSFT	242
P44741	73	YAYKLRYYYFEEVRAEYHSPVRFSFIALIPITMLVG---DILYRWNPLIAEVLIWIGTIG	129
		Y K +YFE V+ EY PVR +F + M + ++ N IW +G	
Q9LD83	243	YILKCIFYFEAVKREYFHPVRVNFFFAPWVVCMFLAISVPPMFSPNRKYLHPAIWCVFMG	302
P44741	130	QLLFSTLRLRSELWQGGVFEQ--KSTHPSFYLPAVAANFTSASSLALLGYHDLGYLFFGAG	187
		F L++ W G + K +PS +L +V NF A + +G+ ++ + G	
Q9LD83	303	PYFFLELKIYGQWLGGKRRRLCKVANPSSH-SVVGNFVGAILASKVGDEVAKFLWAVG	361
P44741	188	MIAWIIFEPVLLQHLRISSLEPQFRATMGIVLAPAFVCVSAYLSINHGEVDTLAKILWGY	247
		+++ L Q L S P+ + + A S + +G+ D ++ +	
Q9LD83	362	FAHYLVVFTLYQRRLPTSEALPKELHPVYSMFIAAPSAASIAWNTIYQQFDGCSRCCFFI	421
P44741	248	GFLQLFFLRLFPWIVEKGLNIGLWAWSFGLASMANSATAFY----HGNVLQGVSIFAFV	303
		L+ + ++ W+++F + + A+ AT Y G + +++	
Q9LD83	422	ALFLYISLVARINFFTGFKFSAWWSYTFPMTT-ASVATIKYAEAVPGYPSRALALTSF	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%

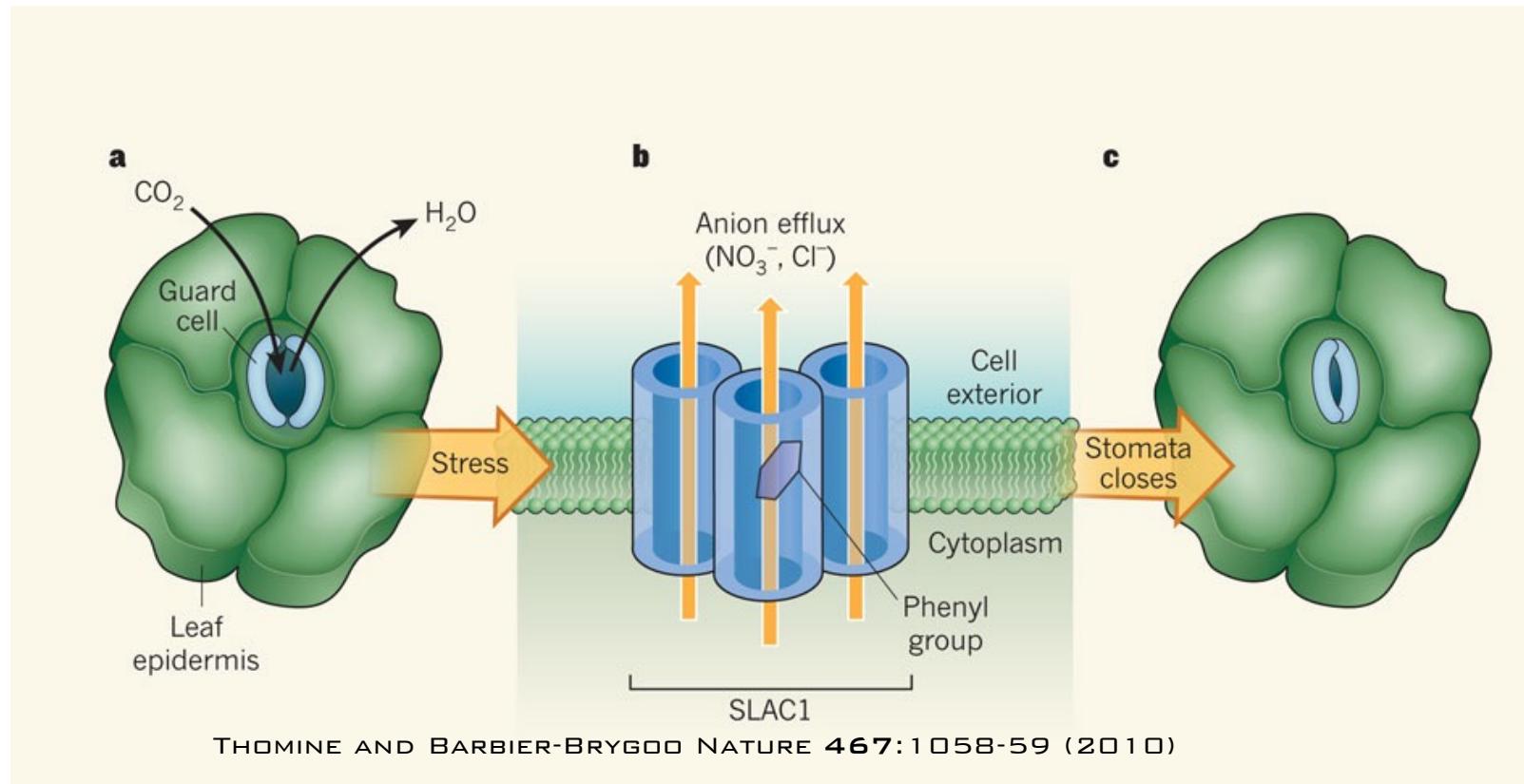
**Positives :** 41.0%

**Query Length:** 328

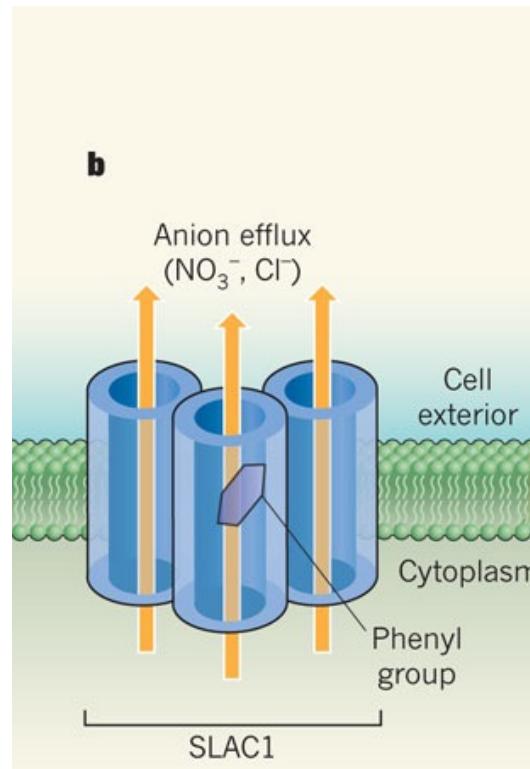
**Match Length:** 556

			
P44741	20	PFPL--PTGYFGIPLGLAALSLAWFHLE-----NLFPAAARMVSDVLGIVASAVWILFILM	72
		PF L P G FGI LGL++ ++ W L N +++ V+ + + V +	
Q9LD83	183	PFLLRFPICGCFGICLGLSSQAVLWLALAKSPATNFLHITPLINLVVWLFLSVLVSFSFT	242
P44741	73	YAYKLRYYYFEEVRAEYHSPVRFSFIALIPITMLVG---DILYRNPLIAEVLIWIGTIG	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	PYFFLELKIYGQWLGGKRRRLCKVANPSSH-SVVGNFVGAILASKVGDEVAKFLWAVG	361
P44741	188	MIAWIIFEPVLLQHLRISSLEPQFRATMGIVLAPAFVCVSAYLSINHGEVDTLAKILWGY	247
		+++ L Q L S P+ + + A S + +G+ D ++ +	
Q9LD83	362	FAHYLVVFTLYQRRLPTSEALPKELHPVYSMFIAAPSAASIAWNTIYQQFDGCSRCCFFI	421
P44741	248	GFLQLFFLLRLFPWIVEKGLNIGLWAWSFGLASMANSATAFY----HGNVLQGVSIFAFV	303
		L+ + ++ W+++F + + A+ AT Y G + +++	
Q9LD83	422	ALFLYISLVARINFFTGFKFSAWWSYTFPMTT-ASVATIKYAEAVPGYPSRALALTSF	480
P44741	304	FSNVMIIGLLVLMTI 317	
		S M+ +L + T+	
Q9LD83	481	ISTAMVCVLFVSTL 494	



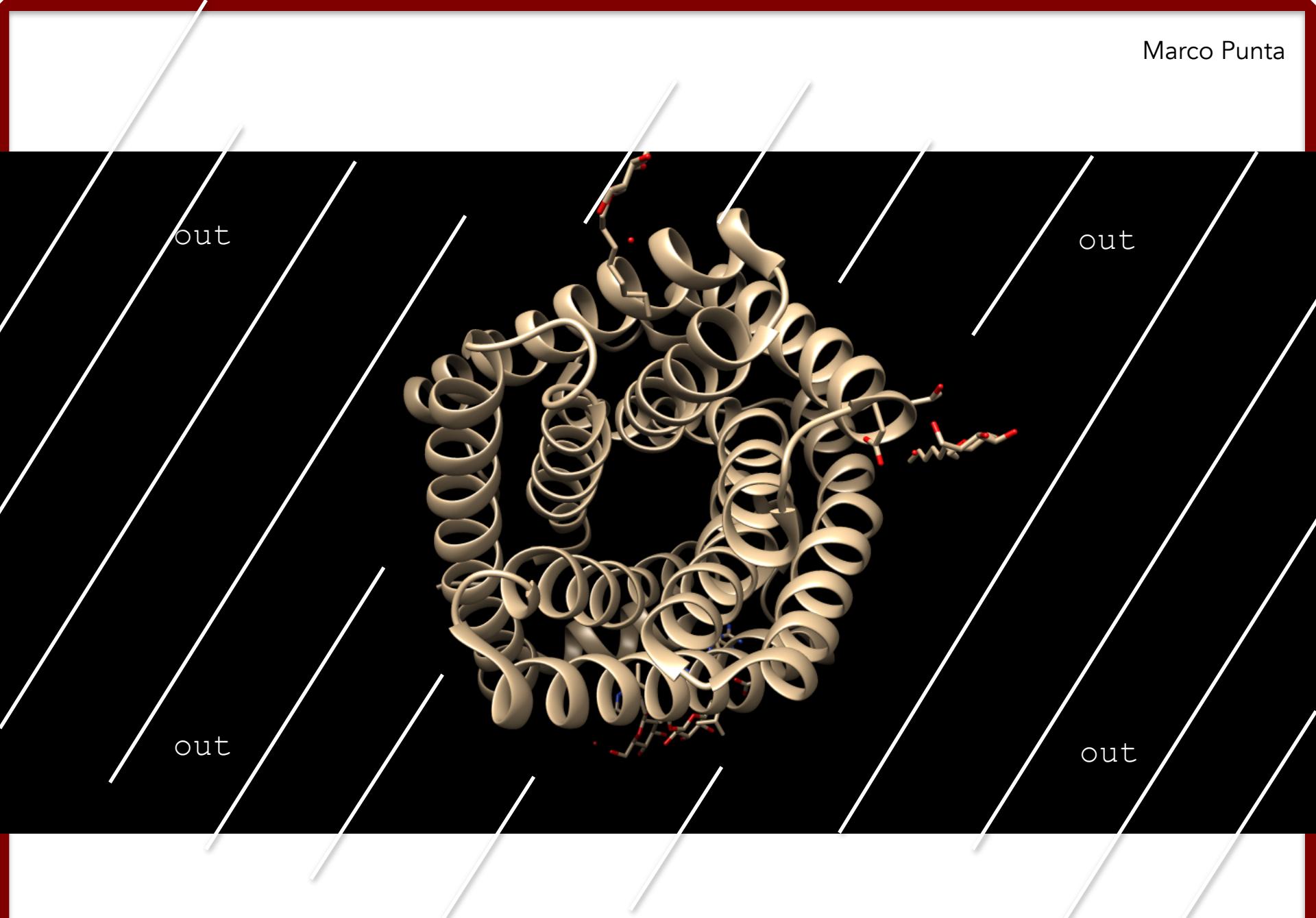


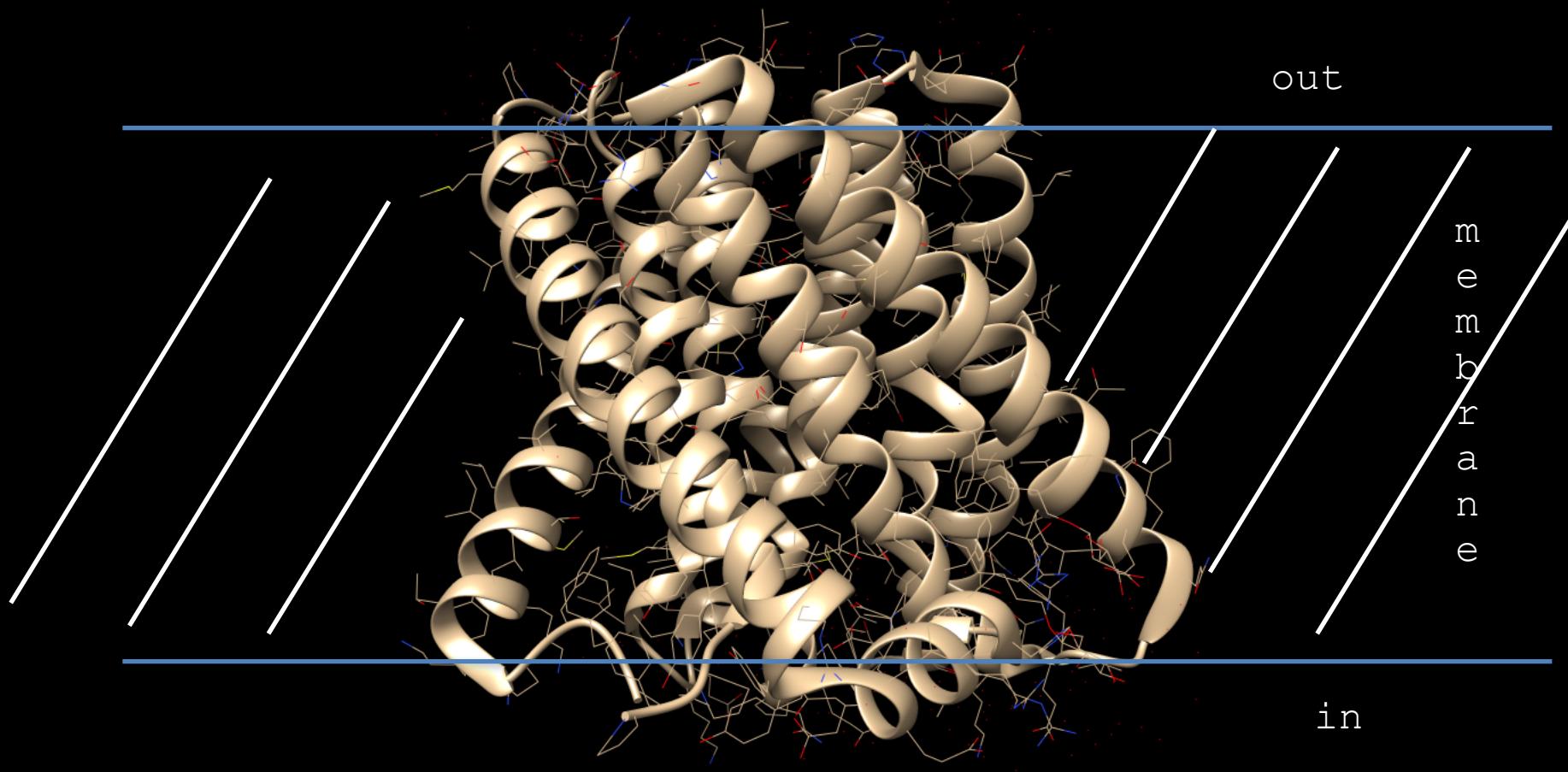
THOMINE AND BARBIER-BRYGGO NATURE 467:1058-59 (2010)

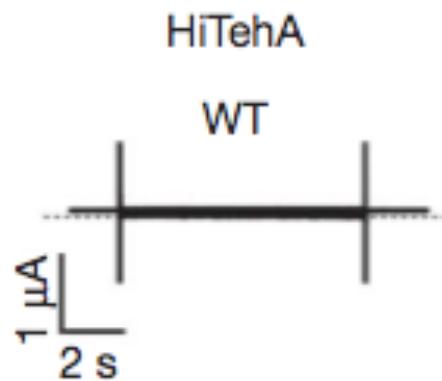


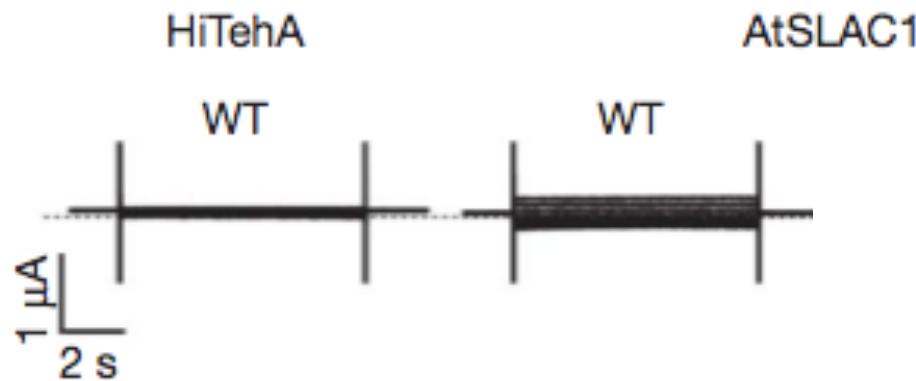
1. OPEN Chimera
2. File -> Open "3M71.pdb"

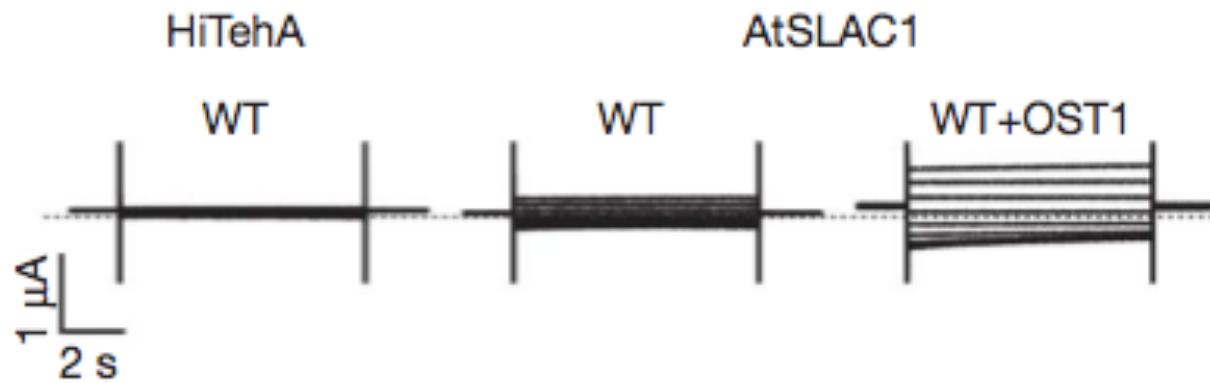












1. Actions -> Atoms/Bonds -> wire
2. Actions -> Atoms/Bonds -> show

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

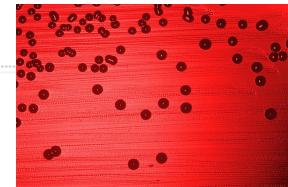
**Ident.:** 22.0%

**Positives :** 41.0%

**Query Length:** 328

**Match Length:** 556

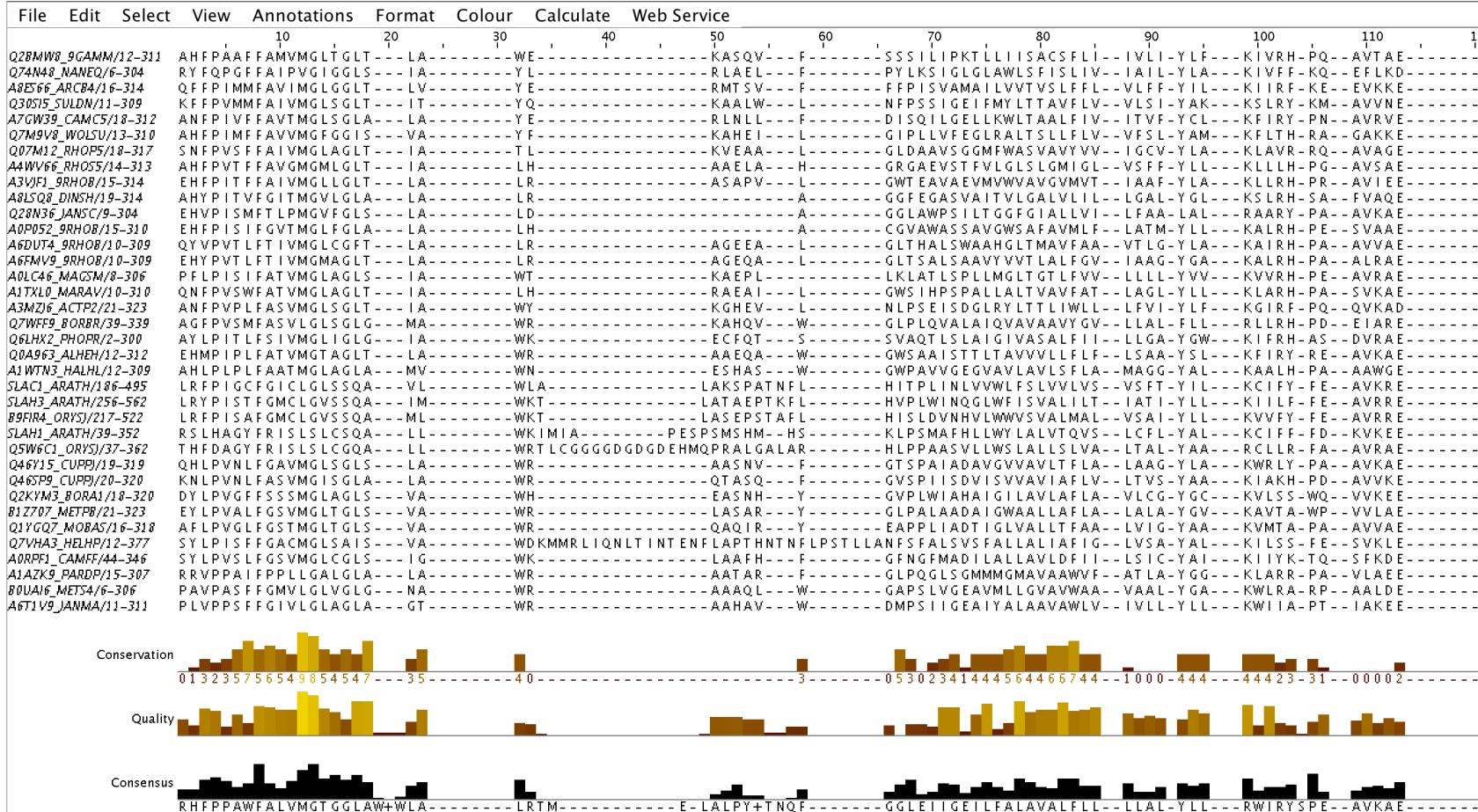
			
P44741	20	PFPL--PTGYFGIPLGLAALSLAWFHLE-----NLFPAAARMVSDVLGIVASAVWILFILM	72
		PF L P G FGI LGL++ ++ W L N +++ V+ + + V +	
Q9LD83	183	PFLLRFPICGCFGICLGLSSQAVLWLALAKSPATNFLHITPLINLVVWLFSLVVLVSVSFT	242
P44741	73	YAYKLRYYYFEEVRAEYHSPVRFSFIALIPITMLVG---DILYRNPLIAEVLIWIGTIG	129
		Y K +YFE V+ EY PVR +F + M + ++ 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	PYFFLELKIYGQWLGGKRRRLCKVANPSSH-SVVGNFVGAILASKVGDEVAKFLWAVG	361
P44741	188	MIAWIIFEPVLLQHLRISSLEPQFRATMGIVLAPAFVCVSAYLSINHGEVDTLAKILWGY	247
		+++ L Q L S P+ + + A S + +G+ D ++ +	
Q9LD83	362	FAHYLVVVFVTLYQRRLPTSEALPKELHPVYSMFIAAPSAASIAWNTIYQQFDGCSRCCFI	421
P44741	248	GFLQLFFLRLFPWIVEKGLNIGLWAFFLASMANSATAFY----HGNVLQGVSIIFAV	303
		L+ + ++ W++F + + A+ AT Y G + +++	
Q9LD83	422	ALFLYISLVARINFFTGFKFSAWWSYTFIMTT-ASVATIKYAEAVPGYPSRALALTSF	480
P44741	304	FSNVMIIGLLVLMTI 317	
		S M+ +L + T+	
Q9LD83	481	ISTAMVCVLFVSTL 494	



# 1. OPEN Jalview

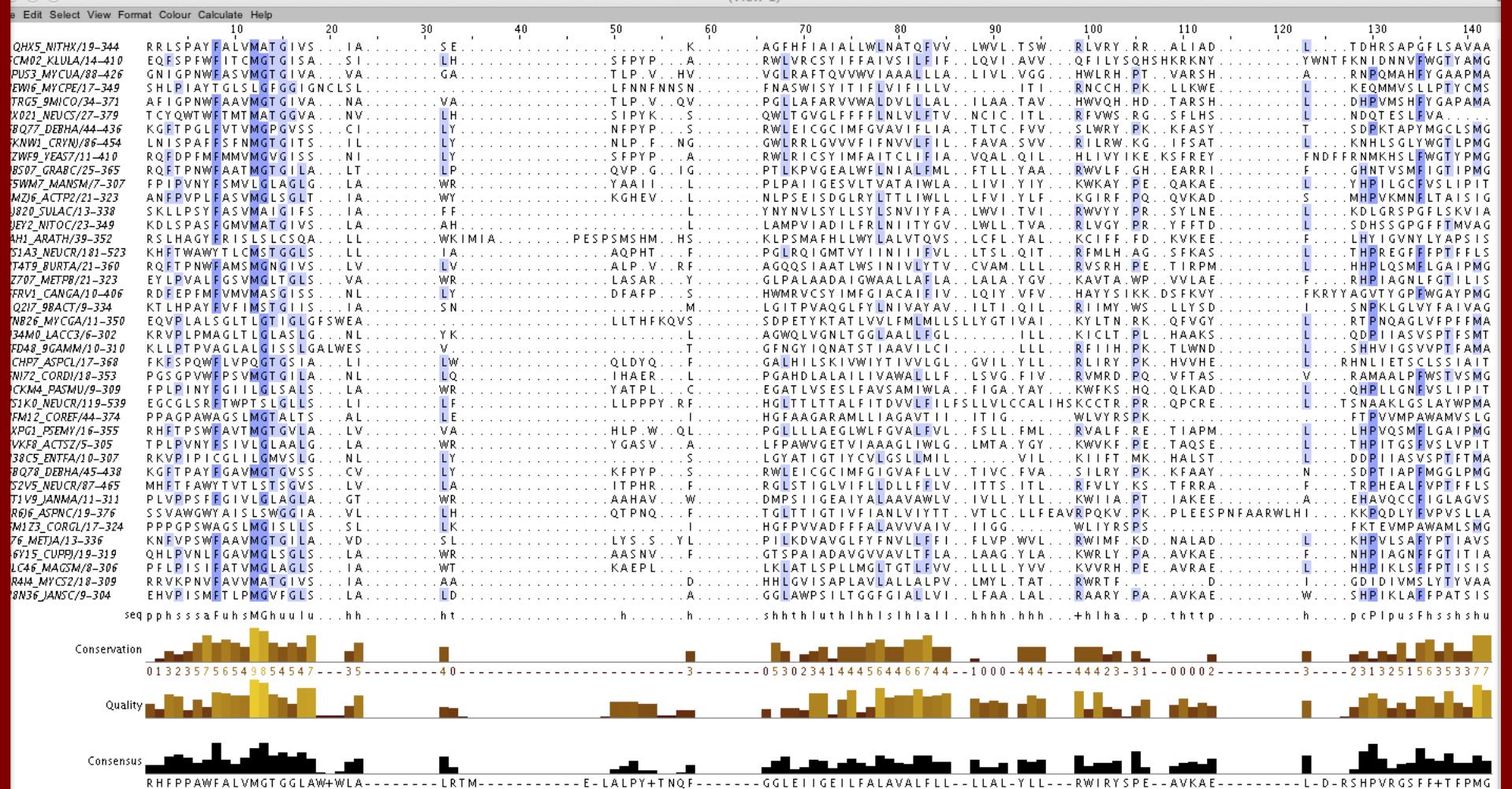
Marco Punta

## 2. File -> Input Alignment -> From File "PF03595\_seed.txt"

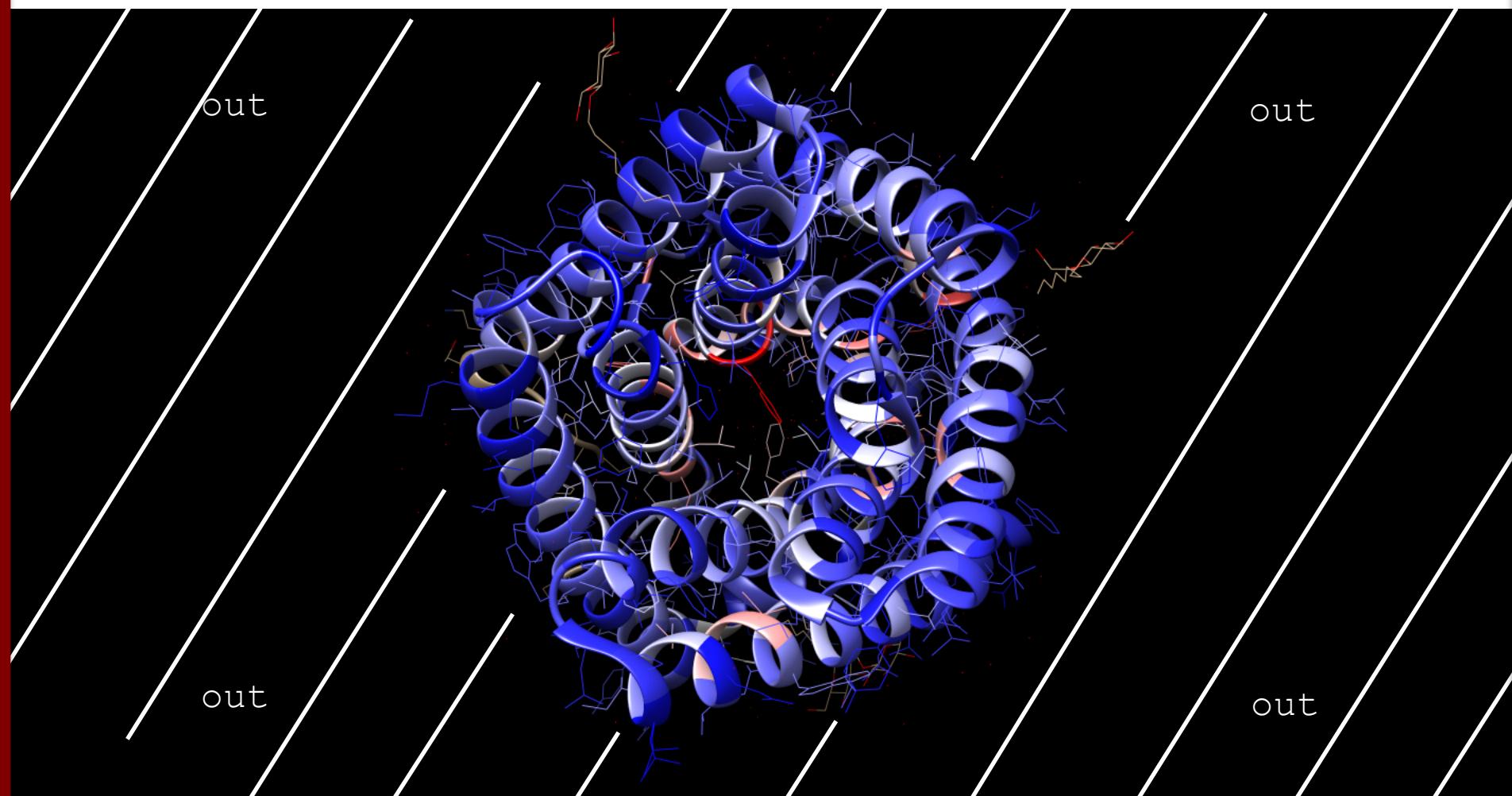


## 1. Colour -&gt; BLOSUM62

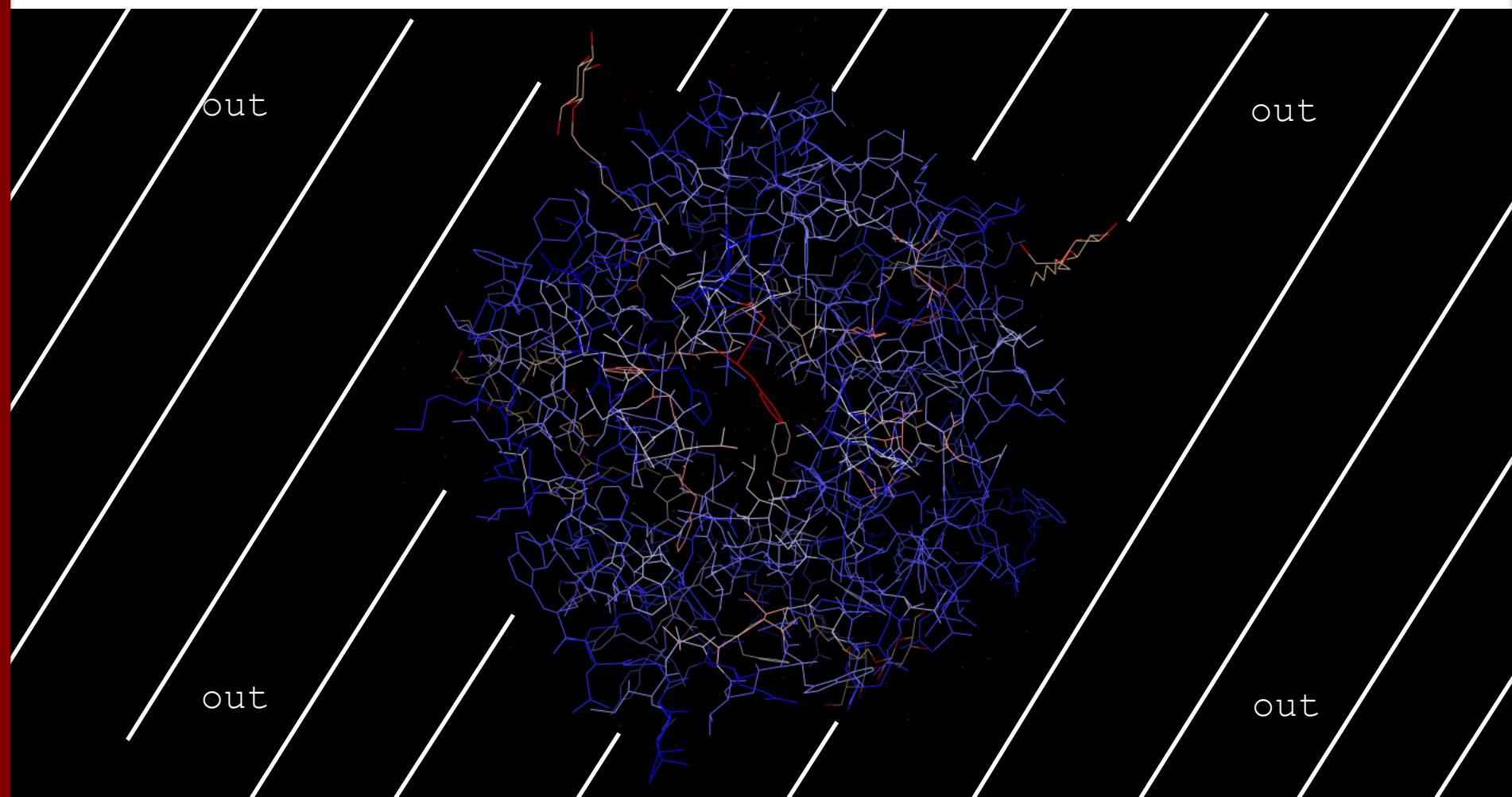
(View 1)

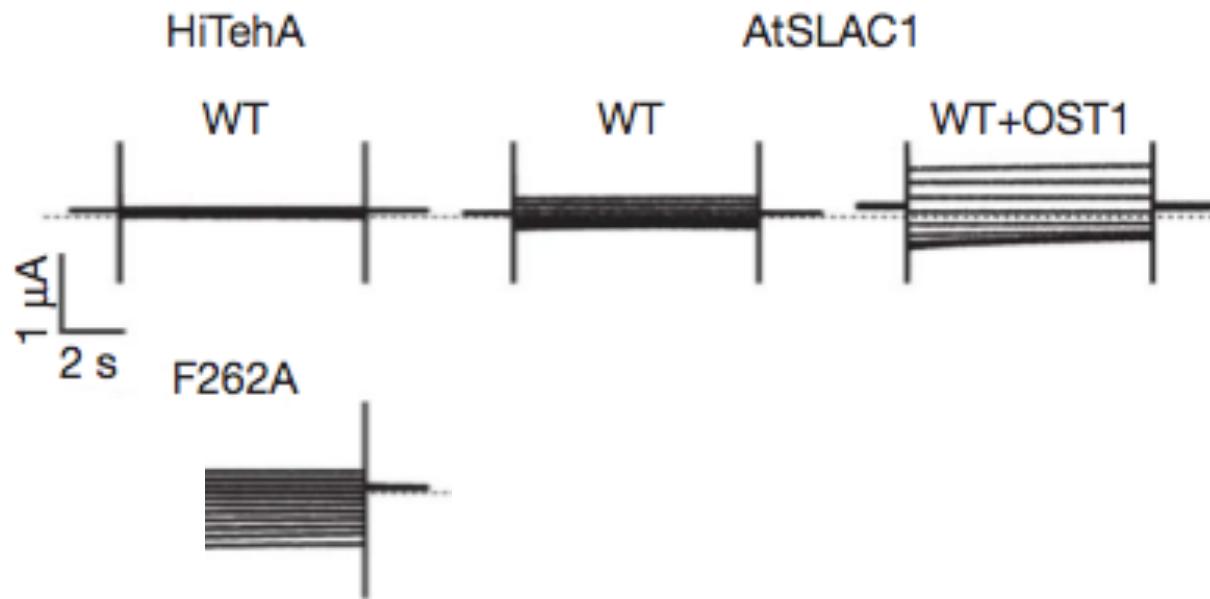


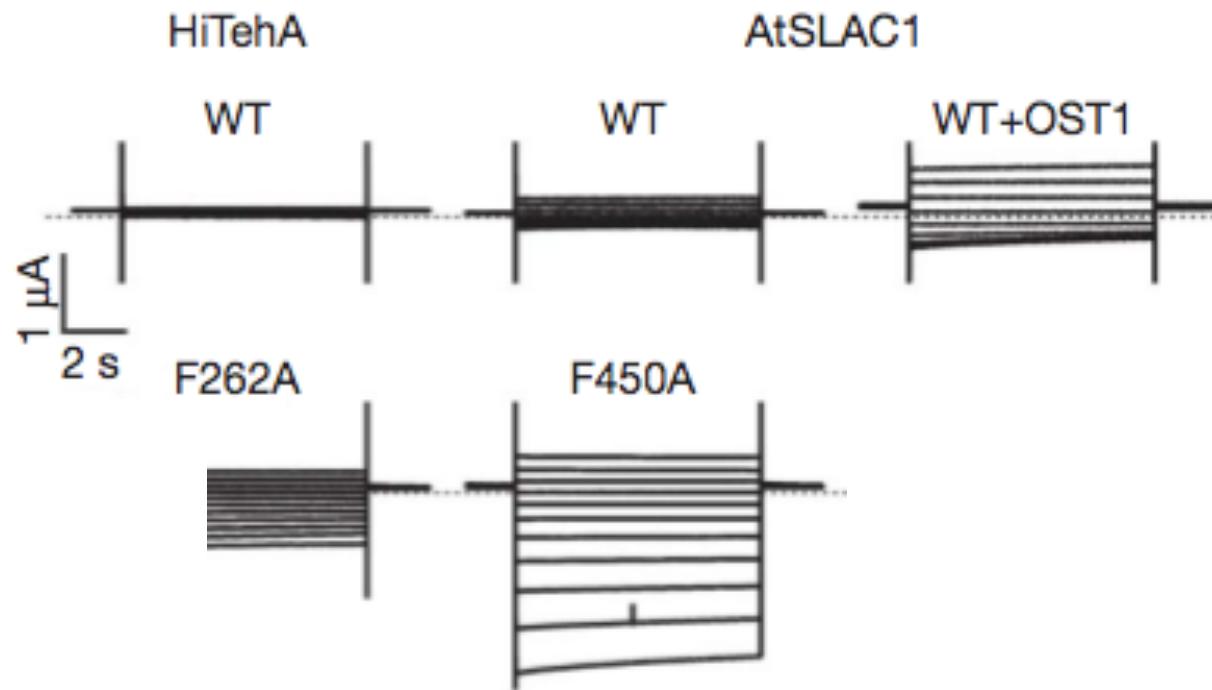
1. Tools-> Sequence -> Multialign Viewer
2. Choose "PF03595\_seed.txt"
3. Select Aligned FASTA
4. Structure -> Render by Conservation

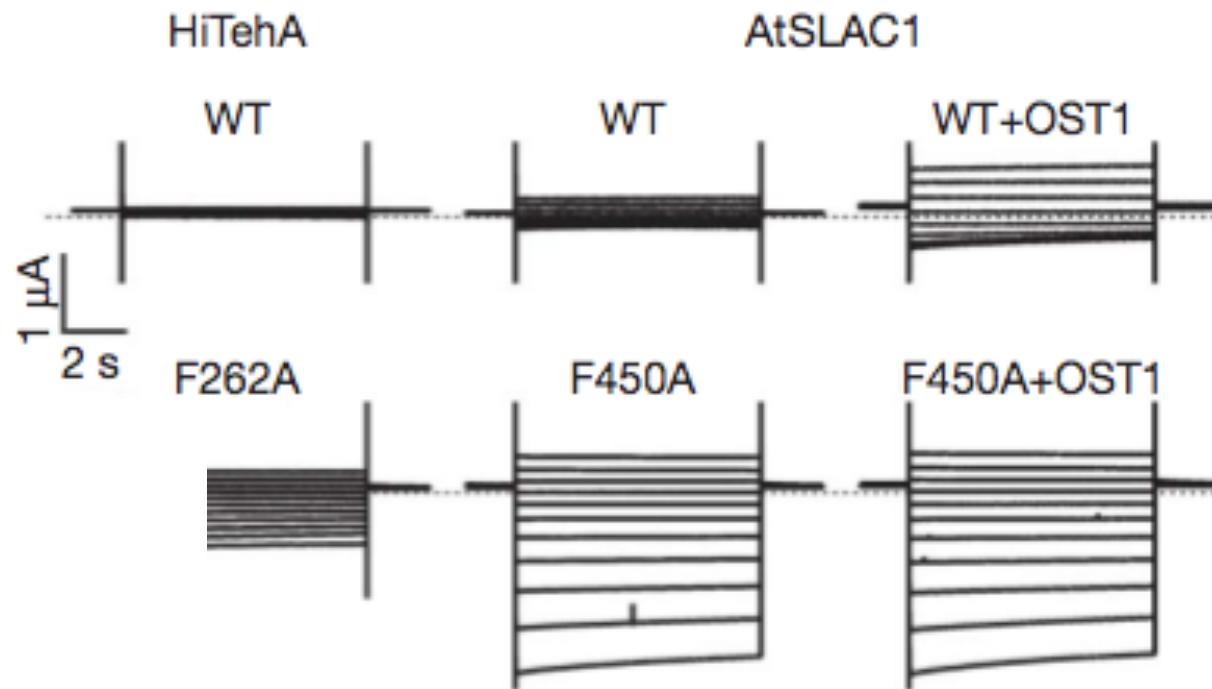


1. Actions-> Ribbon -> hide

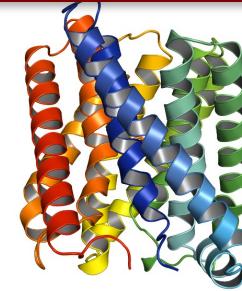




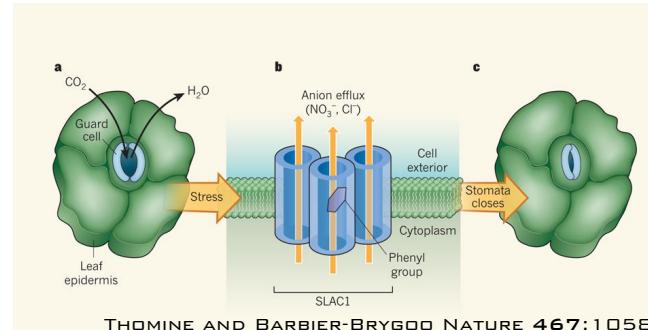




*H. influenzae* protein structure



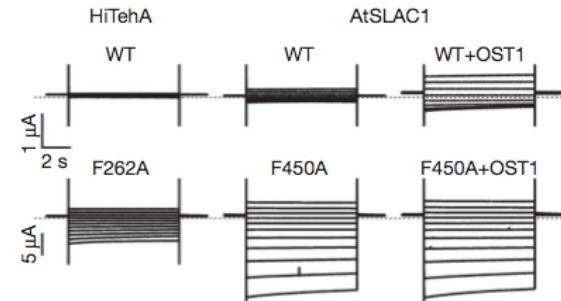
Functional hypothesis via homology to SLAC1



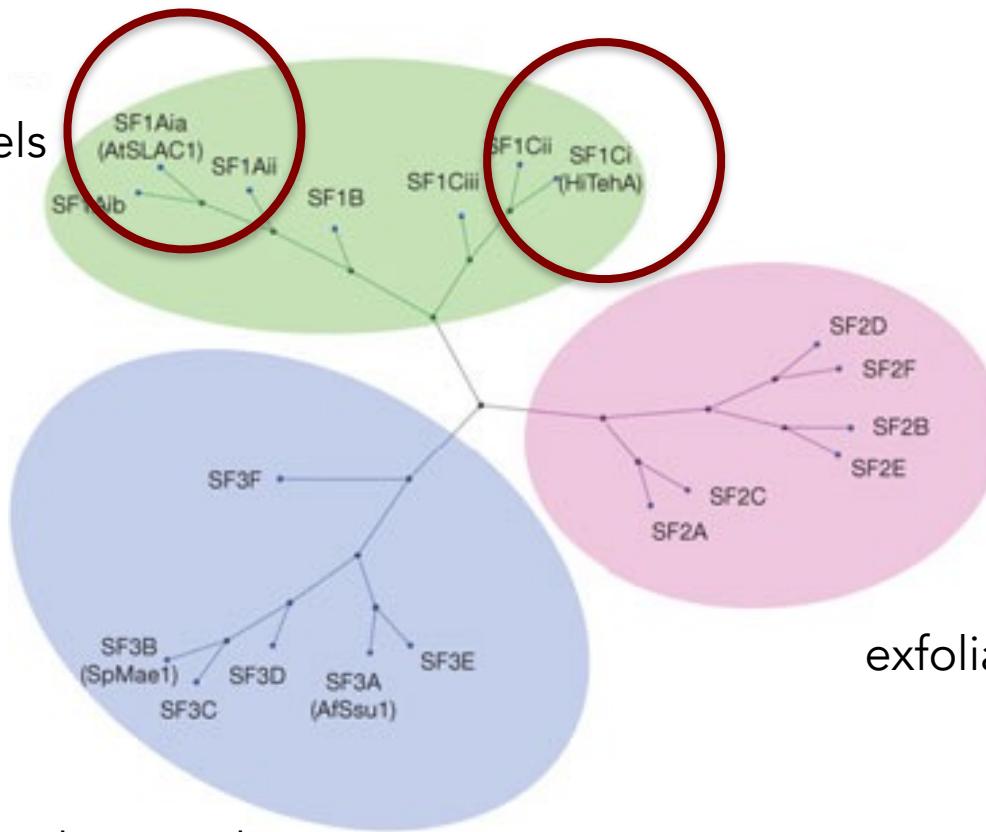
Identification potential functional residues using sequence conservation across the family and structural knowledge



Suggested experiments to test functional hypothesis



Anion channels



exfoliative toxins

malate uptake transporter  
sulphite efflux pump

# Exercise

Homology-based function annotation transfer #2

> more Exercise\_2/blast-2-seqs.link

>cGMP-gated cation channel alpha-1

MKLSMKNNIINTQQSFVTMPNVIVPDIEKEIRRMENGACSSFSEDDDSASTSEESENENP  
HARGSF SYKSLRKGGPSQREQYLPGAI ALFNVNNSSNKDQEPEEEKKKKKEKKS KSDDKN  
ENKNDPEKKKKKKDKEKKKKEEKSKDKKEEEKKEV VVIDPSGNTYYNWLFCITLPVMYNW  
TMVIARACFDELQSDYLEYWLILDYVSDIVYLIDMFVRTRTGYLEQGLLVKEELLINKY  
KSNLQFKLDVLSLIPTDLLYFKLGWNYPEIRLNRLRFSRMFEFFQRTETRTNYPNIFRI  
SNLVMYIVIIHWNACVFYSISKAIGFGNDTWVYPDINDPEFGRLARKYVYSLYWSTLTL  
TTIGETPPPV RDSEYVFVVVDLIGVLIFATIVGNIGSMISNMNAARA EFQARIDA IKQY  
MHFRNVSKDMEKRVIKWFDYLWTNKKTVD EKEVLKYLPDKLRAEIAINVHLDTLKKVRIF  
ADCEAGLLVELVLKLQPQVYSPGDYICKKG DIGREMYIIKEGKLAVVADDGVTQFVVLSD  
GSYFGEISILNIKGSKAGNRRTANIKSIGYSDLFCLSKDDLMEALTEY PDAKTMLEEKGK  
QILMKDGLLDLNIANAGSDPKDLEEKVTRMEGSVDLLQTRFARILA EYESMQQKLKQRLT  
KVEKFLKPLIDTEFSSIEGPGAESGPIDST

>mystery protein

MGNGSVKPKHSKHPDGHSGNLTTDALRNKVTELEREELRRKDAEIQEREYHLKELREQLSK  
QTVAIAELTEELQNKCICQLNKLQDVVHMQGGSPHQASPDKVPLEVHRKTSGLVSLHSRRG  
AKAGVSAEPTTRTYDLNKPPEFSFEKARVRKDSSEKKLITDALKNQFLKRLDPQQIKDM  
VECMYGRNYQQGSYIIKQGEPGNHIFVLAEGRLEVFGQGEKLLSSIPMWTTFGEIAYNC  
TRTASVKAITNVKTWALDREVFNIMRRTAQARDEQYRNFLRSVSLLKNLPEDKLTKIID  
CLEVEYYDKGDYIIREGEEGSTFFILAKGKVKTQSTEHDQPLQLIKTLQKGEYFGEKAL  
ISDDVRSANIIAEENDVACLVIDRETFNQTVGTFEELQKYLEGYVANLNRDDEKRHAKRS  
MSNWKLSKALSLEMIQLKEKVARFSSSSPFQNLEIIATLGVGGFGRVELVKVNENVAFA  
MKCIRKKHVDTKQQEHVYSEKRILEELCSPFIVKLYRTFKDNKYVYMLLEACLGGELWS  
ILRDRGSFDEPTSKFCVACVTEAFDYLHRLGIYRDLK PENLILDAEGYLKLVDGFACK  
IGSGQKTWTFCGTPEYVAPEVILNKGHDFSVDFWSLGLILVYELLTGNNPPFSGVDQMMTYN  
LILKGIEKMDFPRKITRRPEDLIRRLCRQNPTERLGNLKNGINDIKKHRWLNGFNWEGLK  
ARSLPSPLQRELKGPIDHSDYFDKYPPEKGMPPELSGWDKDF

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](#)Query subrange 

```
TTIGETPPPVRDSEYVFVVVDFLIGVLIFATIVGNIGSMISNMNAARAEGQARIDAIKQY  
MHFRNVSKDMEKRVIKWFDYLWTNKKTVDEKEVLKLYLPDKLRAEIAINVHLDTLKKVRIF  
ADCEAGLLVELVLKLQPQVYSPGDIYICKKGDIGREMYIIKEGKLAVVADDGVTQFVVLSD  
GSYFGEISILNIKGSKAGNRRTANIKSIGYSDLFCLSKDDLMEALTEYPDAKTMLEEKKGK  
QILMKDGLLDLNIANAGSDPKDLEEVKTRMEGSVDLLQTRFARILAYESMQQKLQRFLT  
KVEKFLKPLIDTEFSSIEGPGAESGPIDST
```

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](#)Subject subrange 

```
MSNWKLSKALSLEMIQLKEKVARFSSSSPFQNLEIIATLGVGGFGRVELVKVNENVAFA  
MKCIRKKHVDTKQQEHVYSEKRILEELCSPFIVKLYRTFKDNKYVYMLLEACLGGELWS  
ILRDRGSFDEPTSKFCVACVTEAFDYLHRLGIYRDLKPENLILDAEGYLKLVDFGFAKK  
IGSGQKTWTFCGTPEYVAPEVILNKGHDFSVDFWSLGLILVYELLTGNNPPSGVDQMNTYN  
LILKGIEKMDFPRKITRRPEDLIRRRLCRQNPTERLGNLNGINDIKHRWLNGFNWEGLK  
ARSLPSPLQRELKGPIDHSYFDKYPPEKGMPDELSCWDKDF
```

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 + + Y GDYI ++G+ G +I+ +GK+ V		
Sbjct	281	LRSVSLNKNLNPEDKLTKIIDCLEVEYYDKDYIIIREGEEGSTFFILAKGKVKTQSTEGR		340
Query	530	DGVTQFVVLSDGSYFGEISILNIKGSKAGNRRTANIKSIGYSSDLFCLSKD	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	DALNKNQFLKRLDPQQIKDMVEMYGRNYQQGSYIIIKQGEPGNHIFVLAEGRLEVFGKEK	220	
Query	532	VTQFVVLSDGSYFGEISILNIKGSKAGNRRTANIKSIGYSSDLFCLSKD	579	
		+ + + + FGE++IL RTA++K+I + L ++		
Sbjct	221	LLSSIPM--WTTFGEYLAIL-----YNCTRTASVKAITNVKTWALDR		260

Range 3: 593 to 649 [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#) [▲ First Match](#)

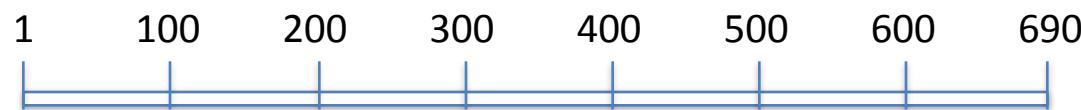
Score	Expect	Method	Identities	Positives	Gaps
22.3 bits(46)	0.081	Compositional matrix adjust.	17/58(29%)	27/58(46%)	7/58(12%)

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

Mystery protein is a cGMP-gated cation channel?

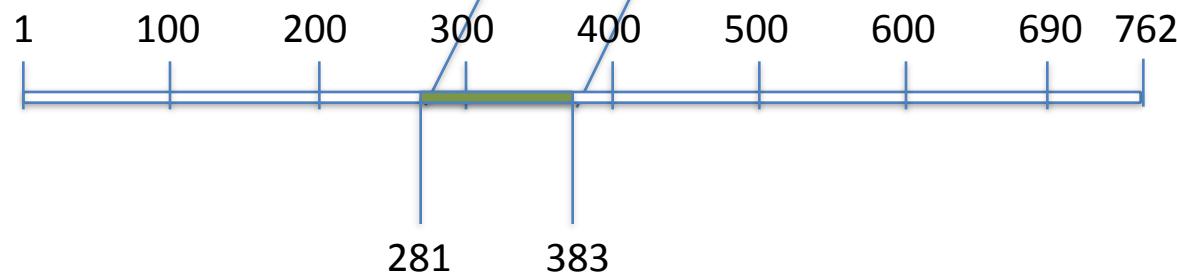
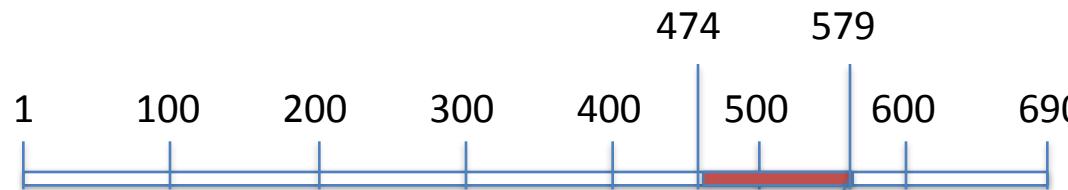
cGMP-gated cation channel alpha-1

P29973 (CNGA1\_HUMAN)



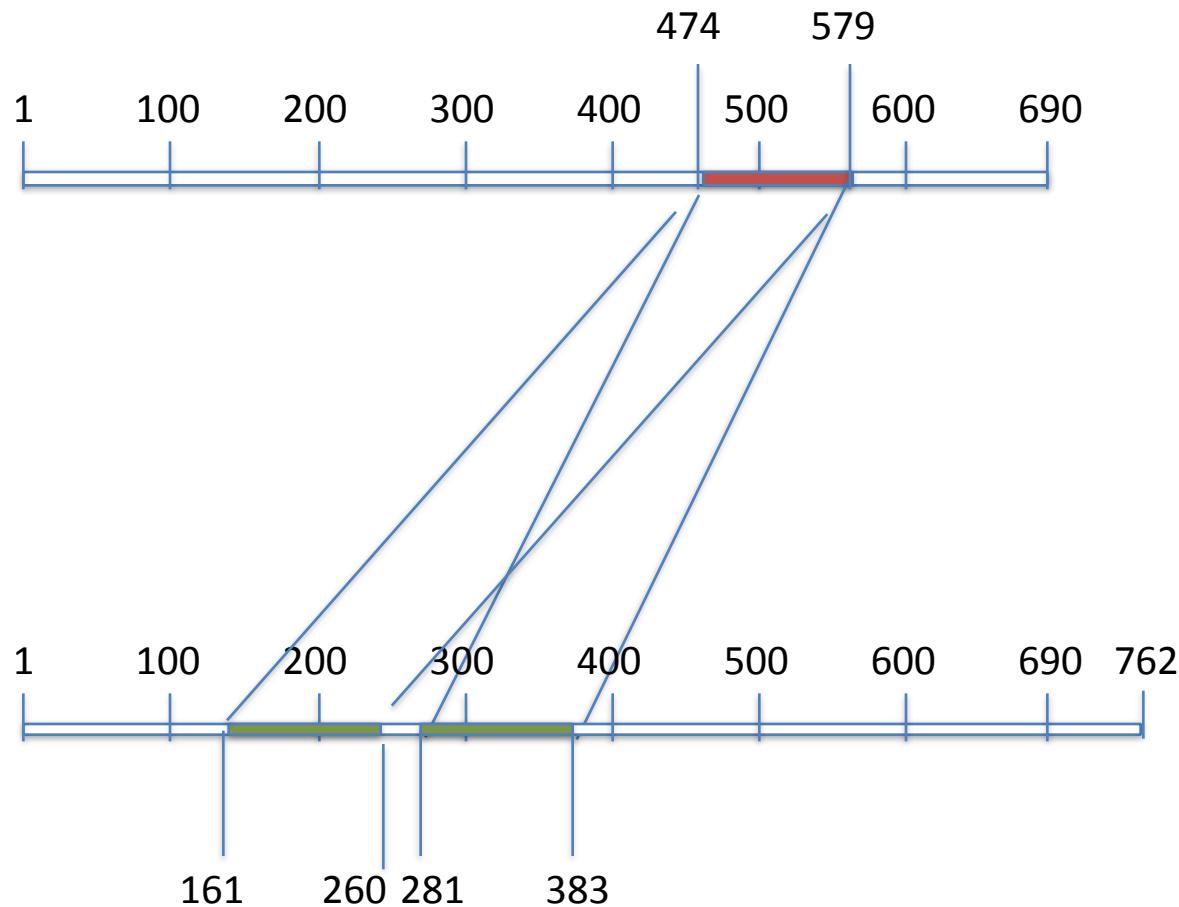
cGMP-gated cation channel alpha-1

P29973 (CNGA1\_HUMAN)



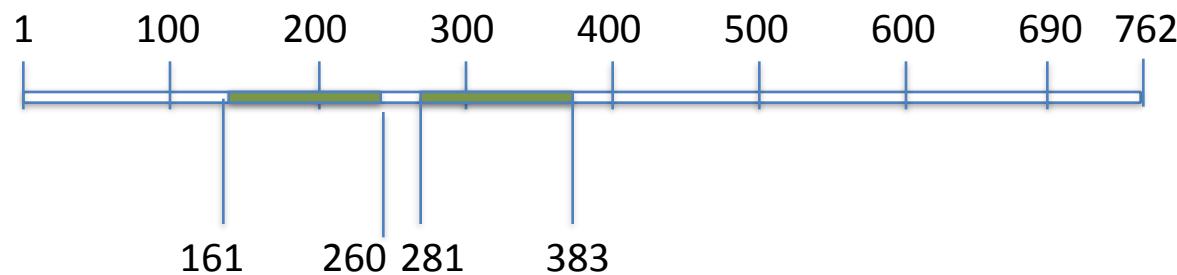
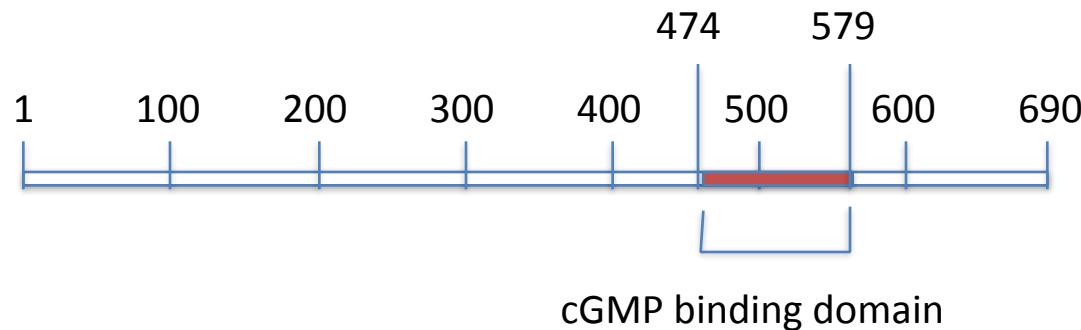
cGMP-gated cation channel alpha-1

P29973 (CNGA1\_HUMAN)



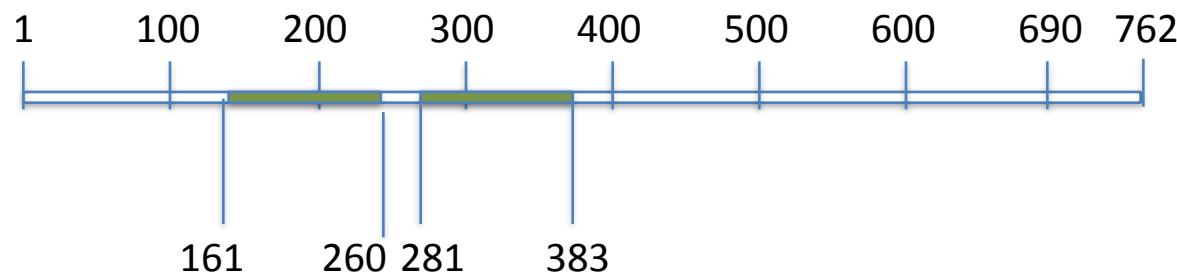
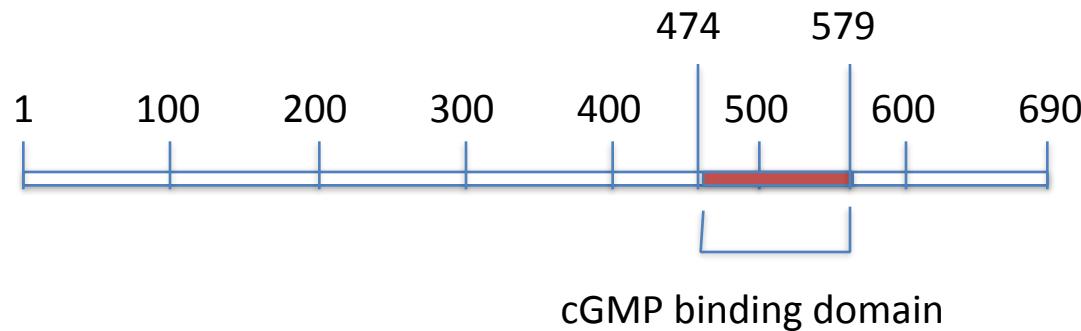
cGMP-gated cation channel alpha-1

P29973 (CNGA1\_HUMAN)

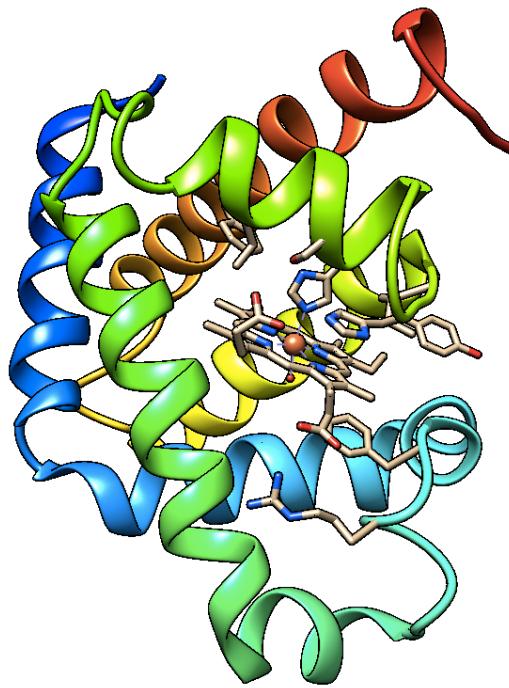


cGMP-gated cation channel alpha-1

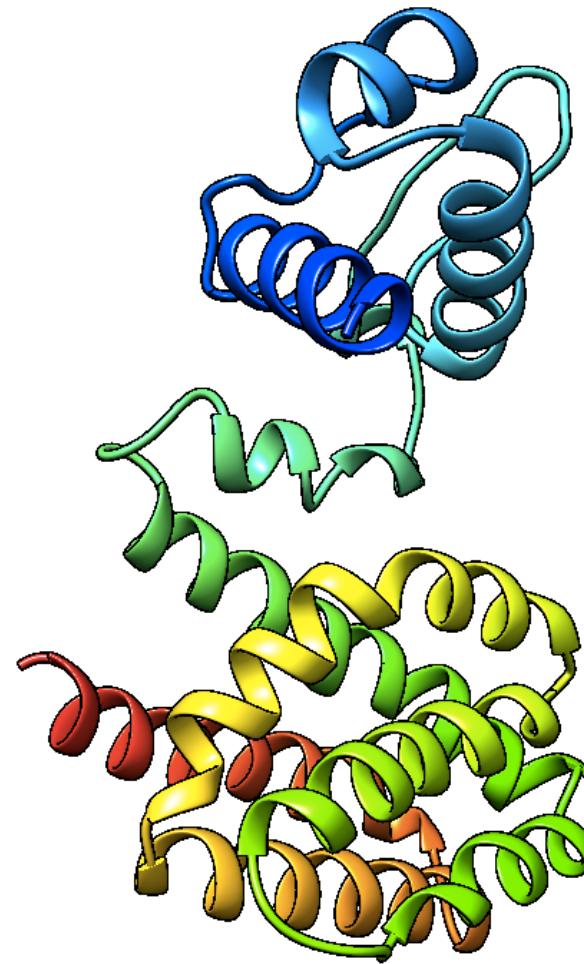
P29973 (CNGA1\_HUMAN)



Mystery protein is a cGMP-dependent protein kinase 2  
Q13237 (KGP2\_HUMAN)



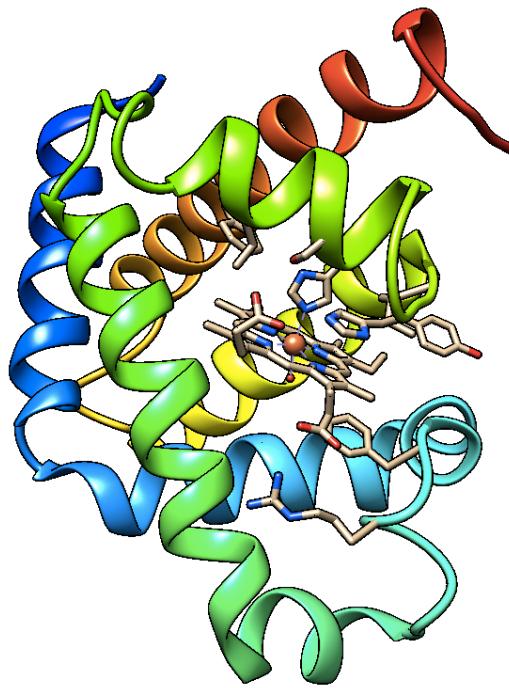
1MBN



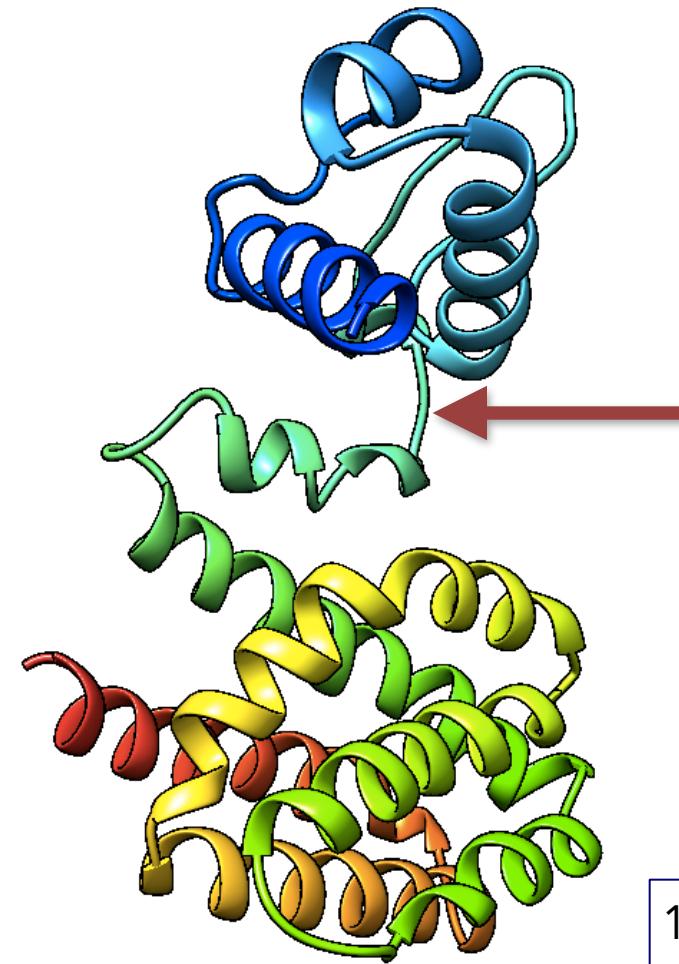
1HW2

**Colour Scheme:**





1MBN



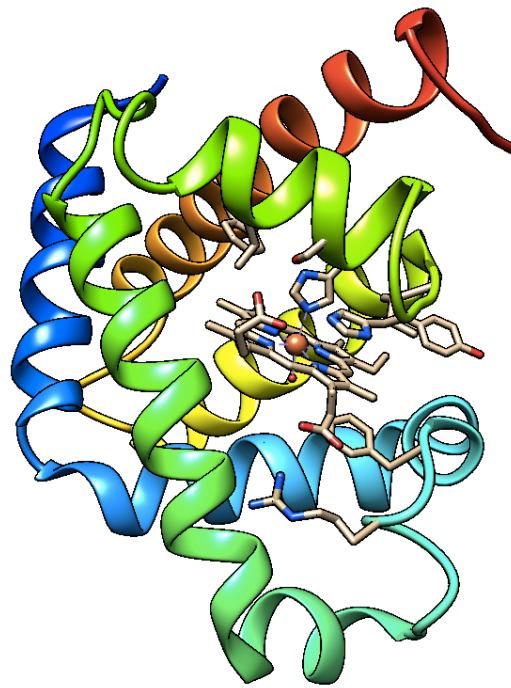
1HW2

**Colour Scheme:**

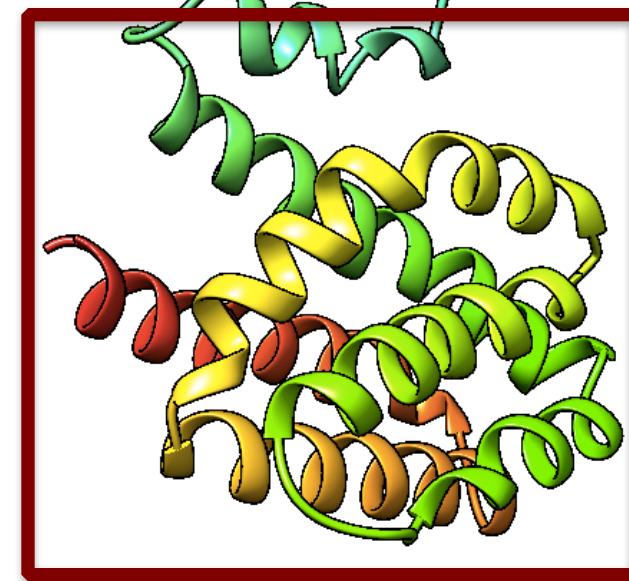
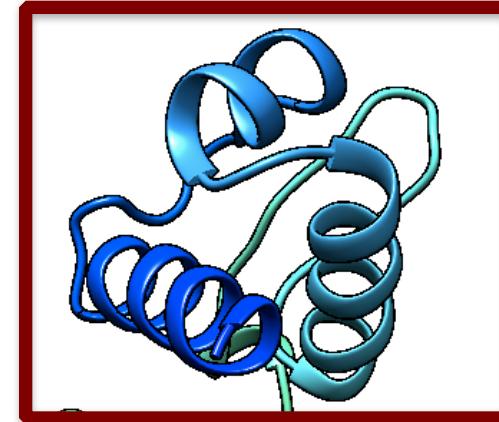


N'

C'



1MBN



1HW2

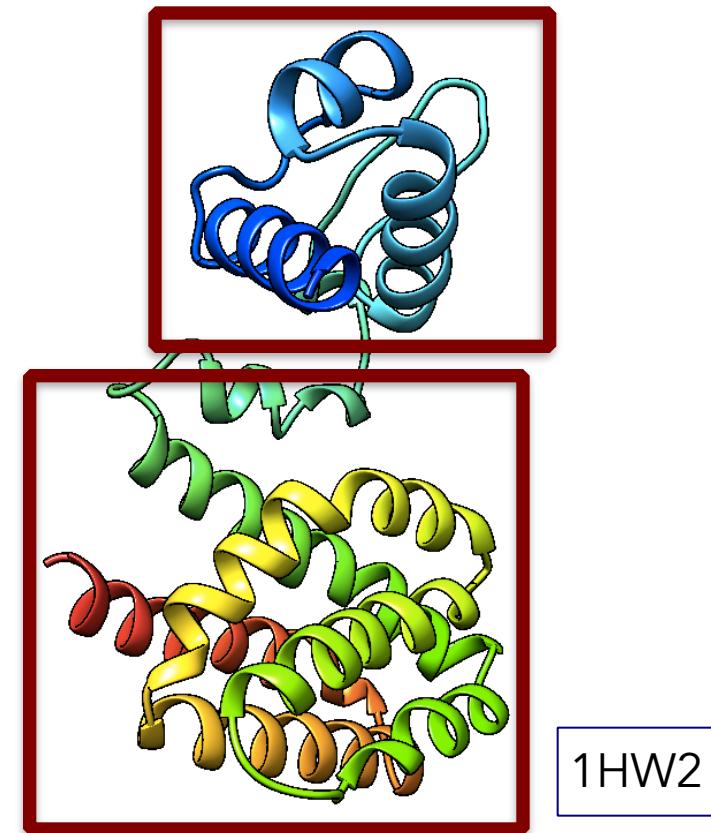
**Colour Scheme:**



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.

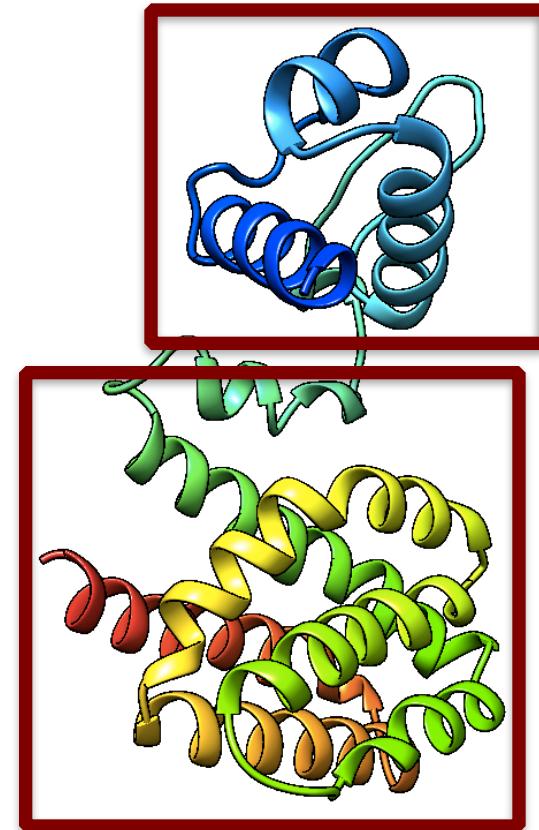
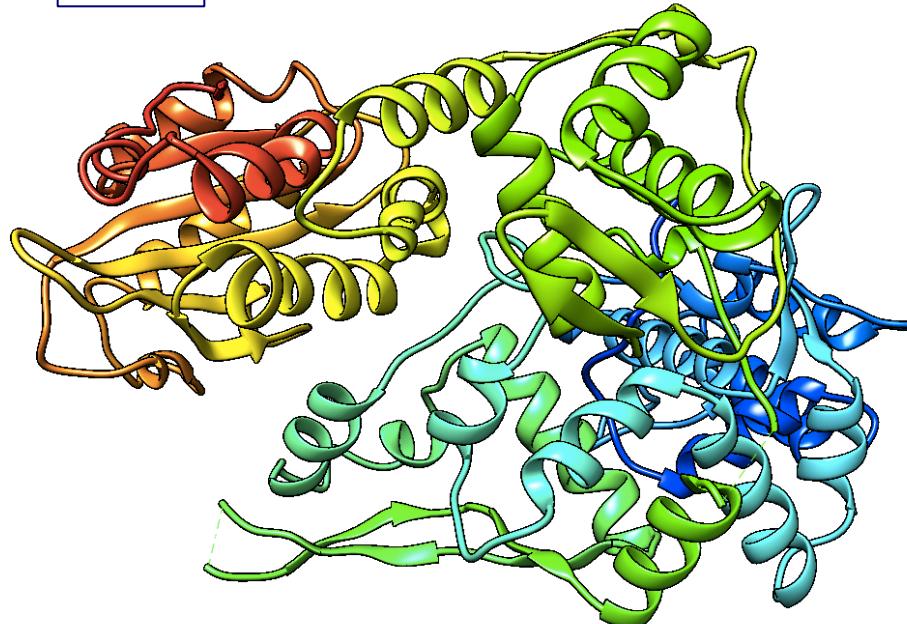
(Marco): in proteins individual domains can be combined to perform complex functions.



**Colour Scheme:**



1FOK

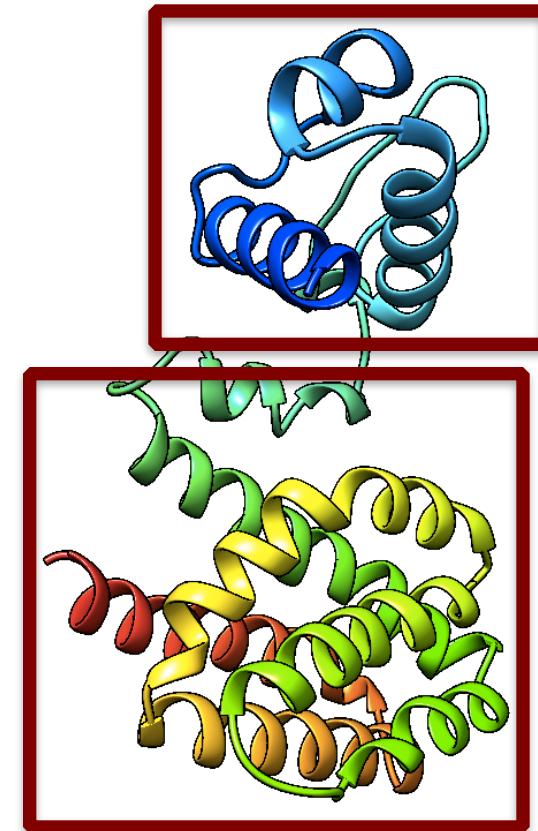
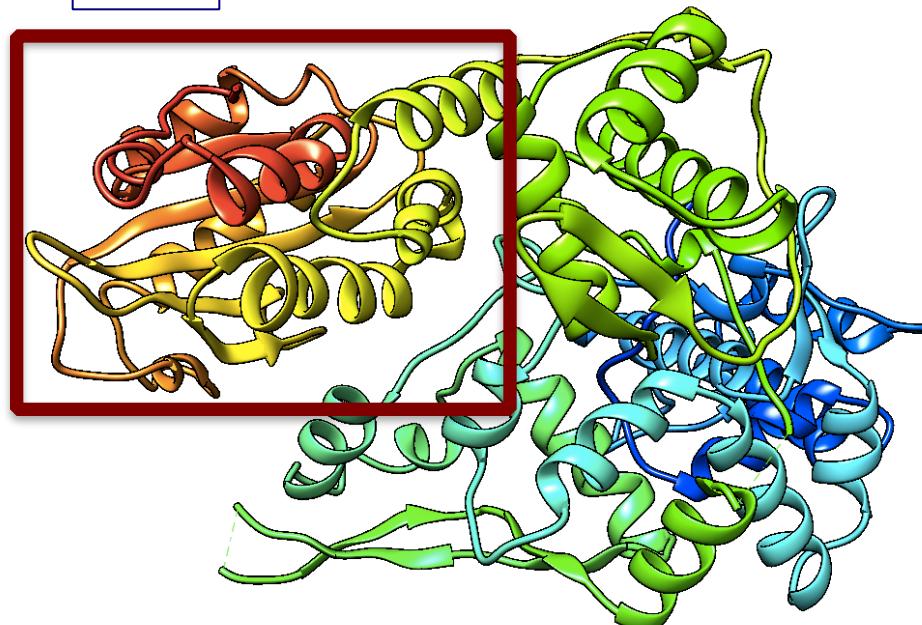


1HW2

**Colour Scheme:**



1FOK

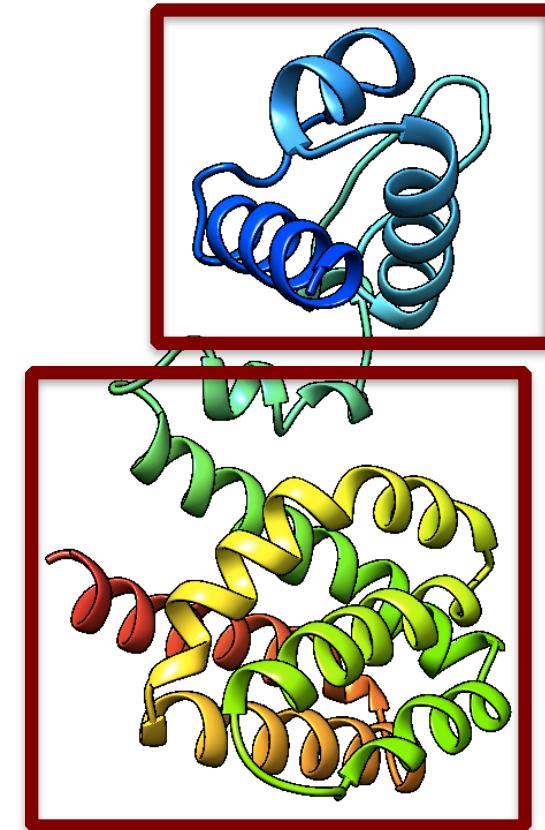
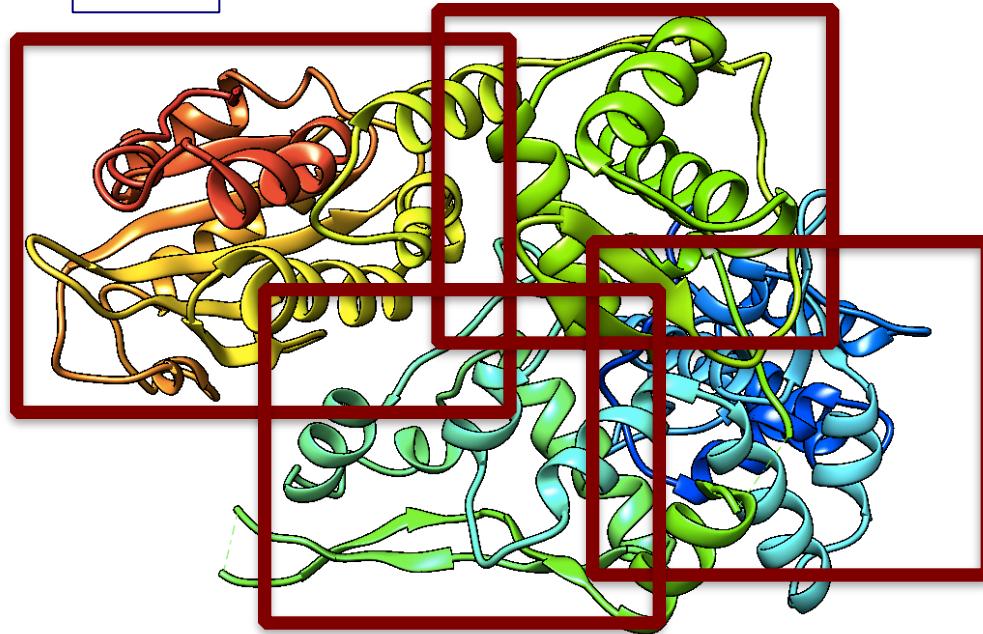


1HW2

**Colour Scheme:**



1FOK

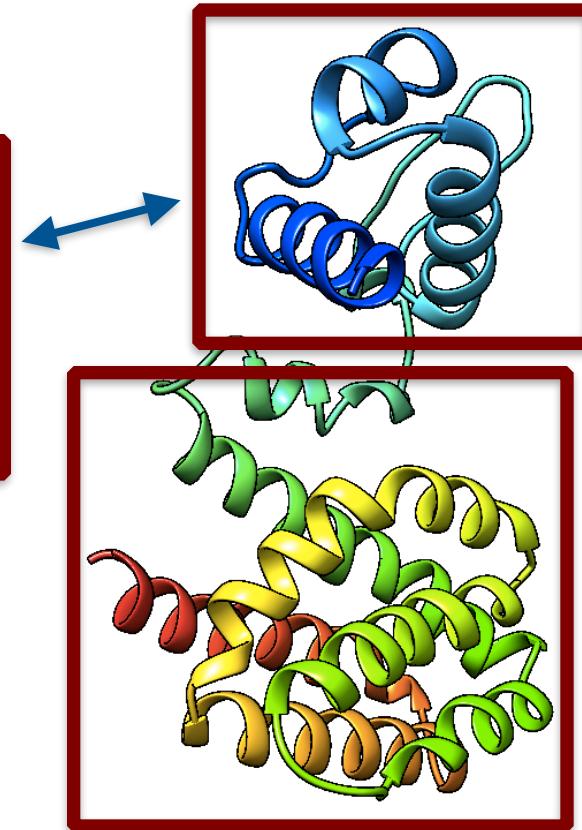
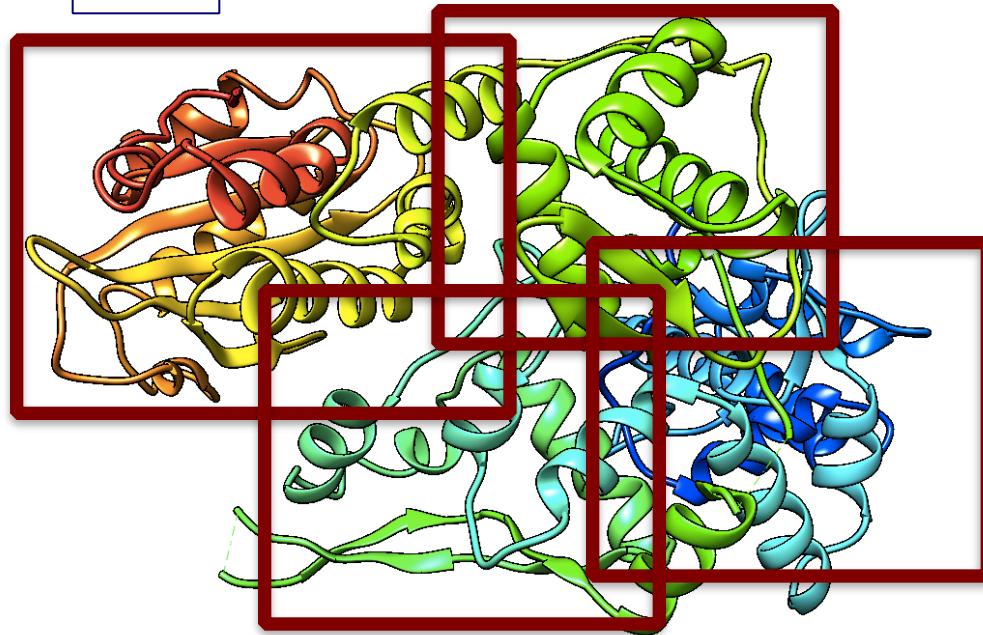


1HW2

**Colour Scheme:**



1FOK

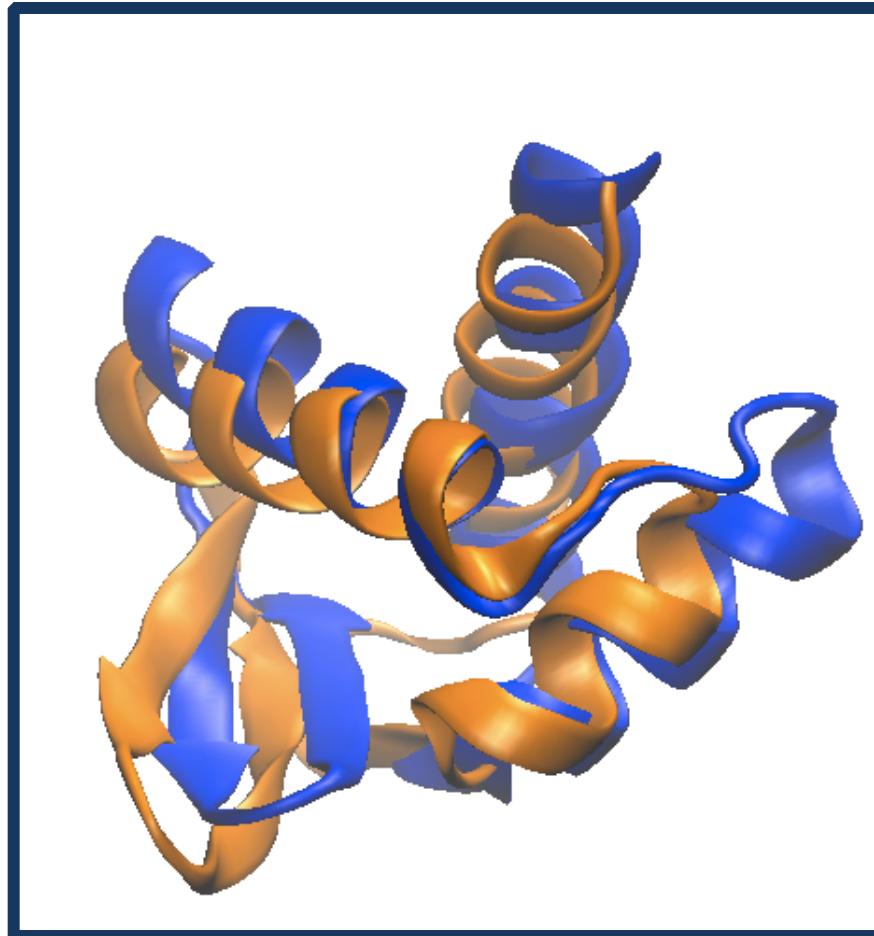


1HW2

Colour Scheme:

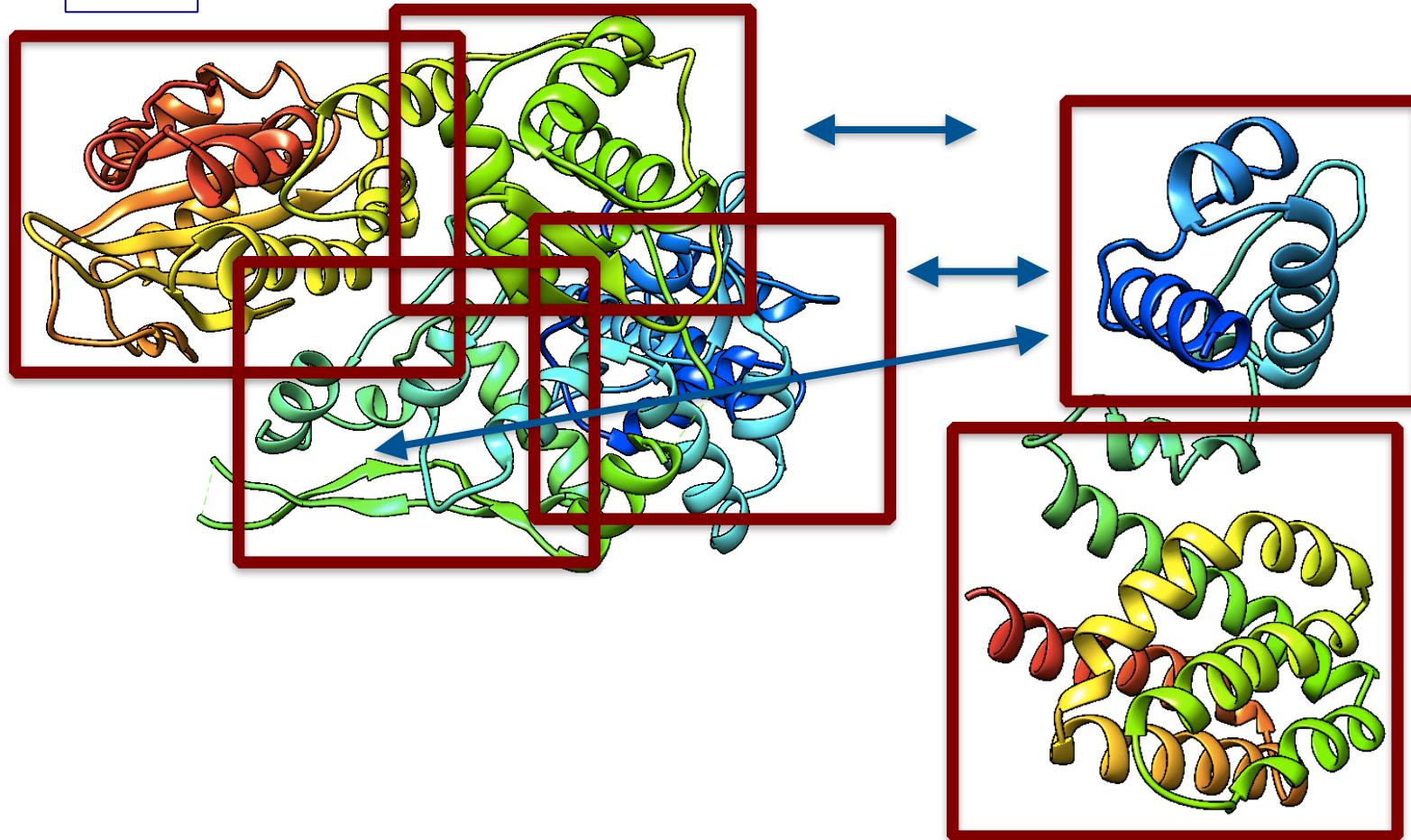


## Winged helix domain (WHD)



Z-score = 4.0  
%id = 8%  
RMSD = 2.7 Å

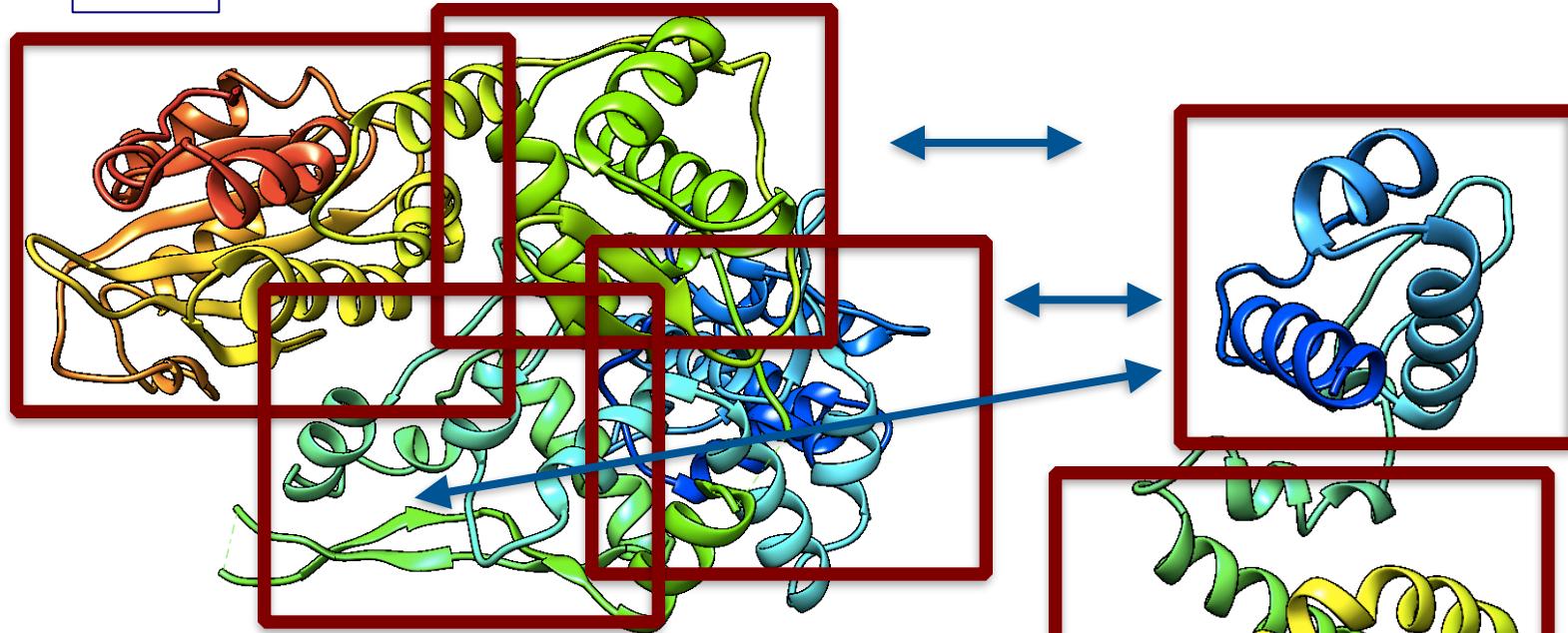
1FOK



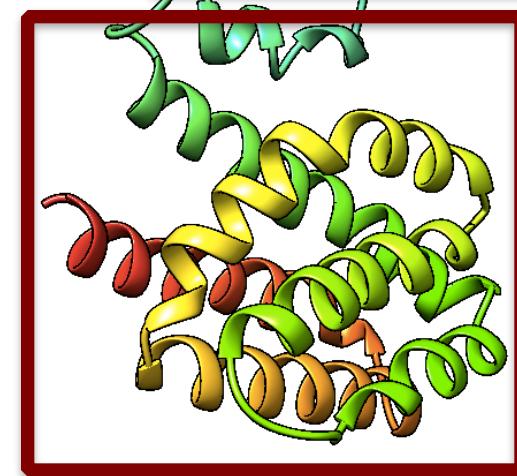
1HW2

**Colour Scheme:**

1FOK

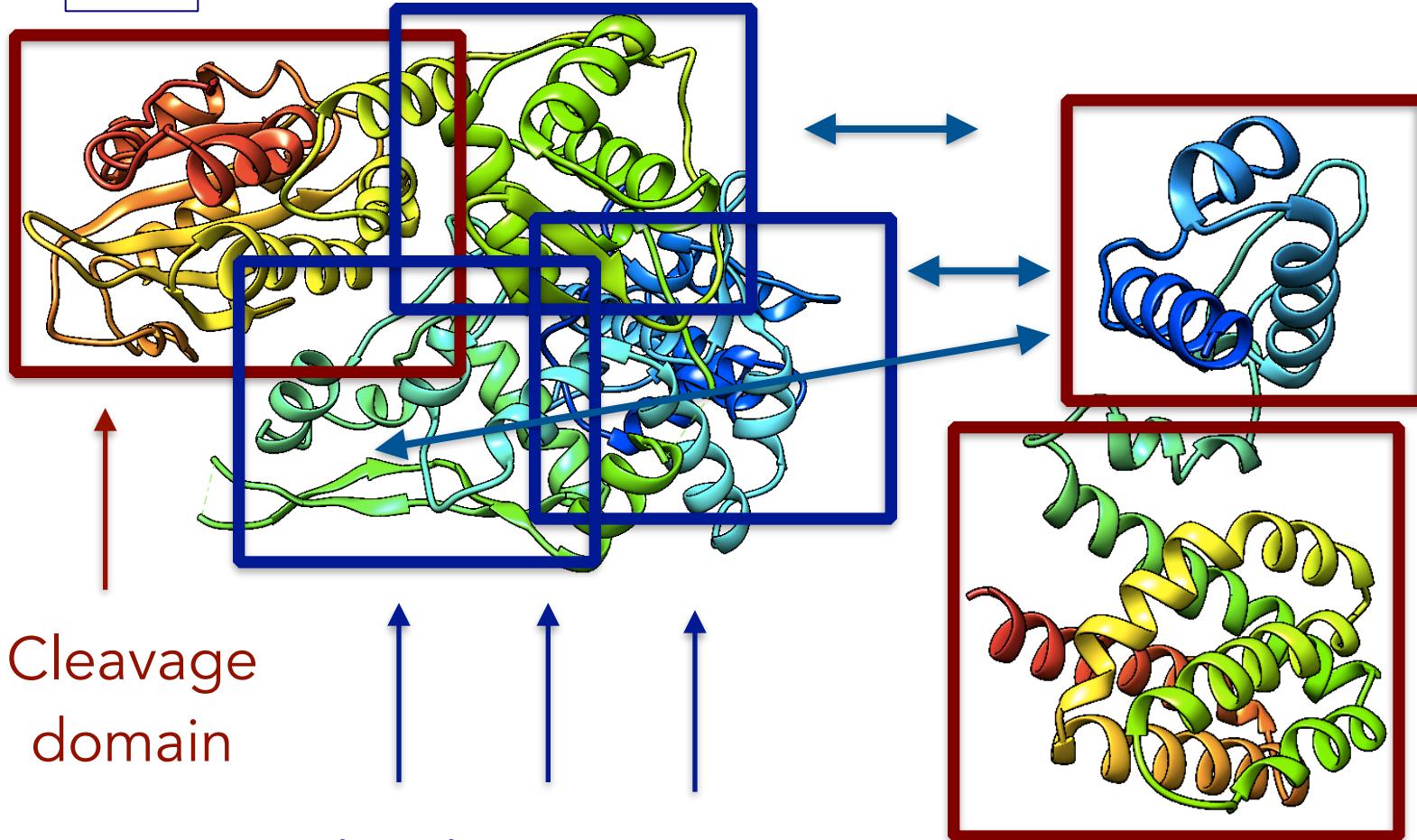


Restriction endonuclease



1HW2

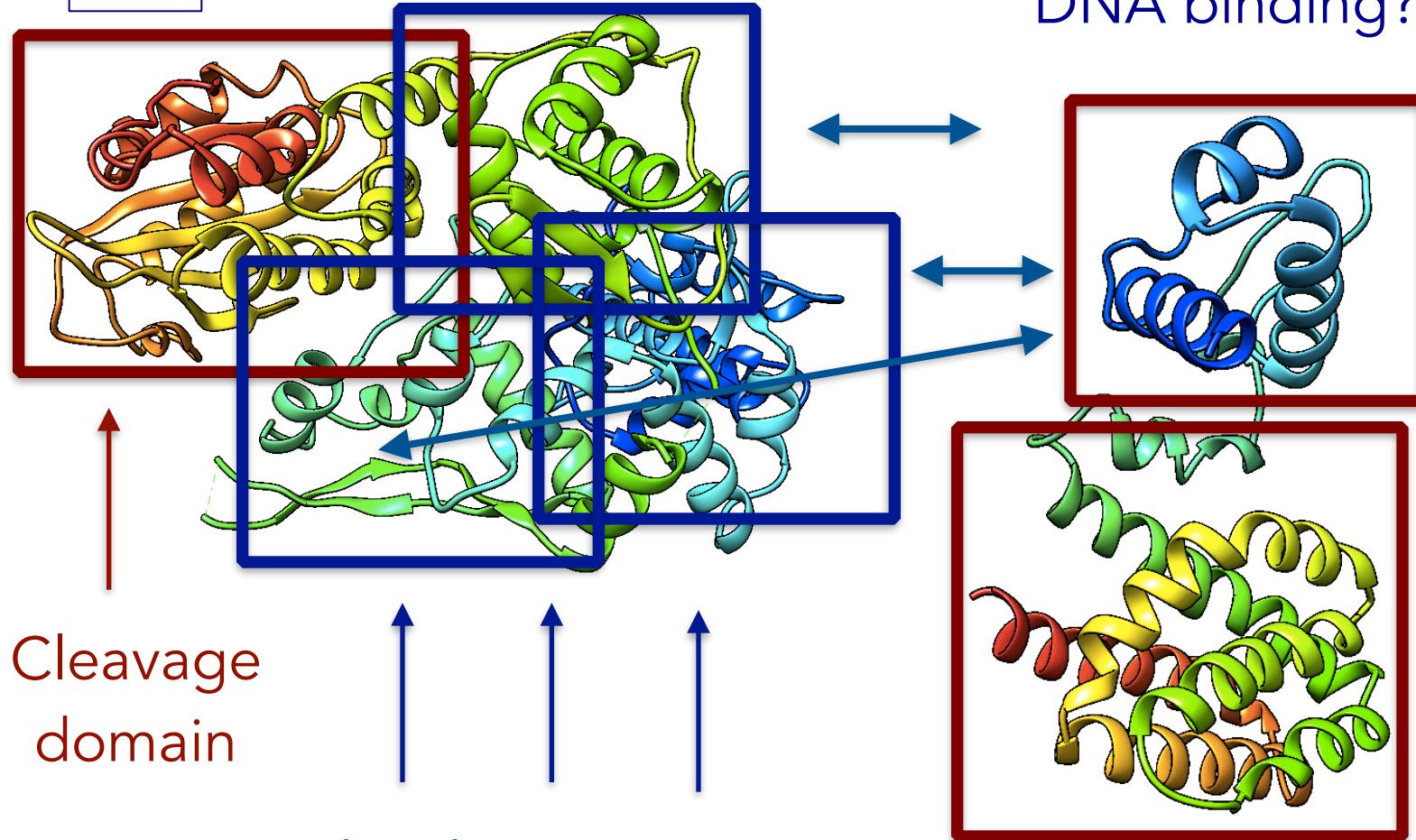
1FOK



1HW2

DNA binding (targeting to a  
specific DNA sequence)

1FOK

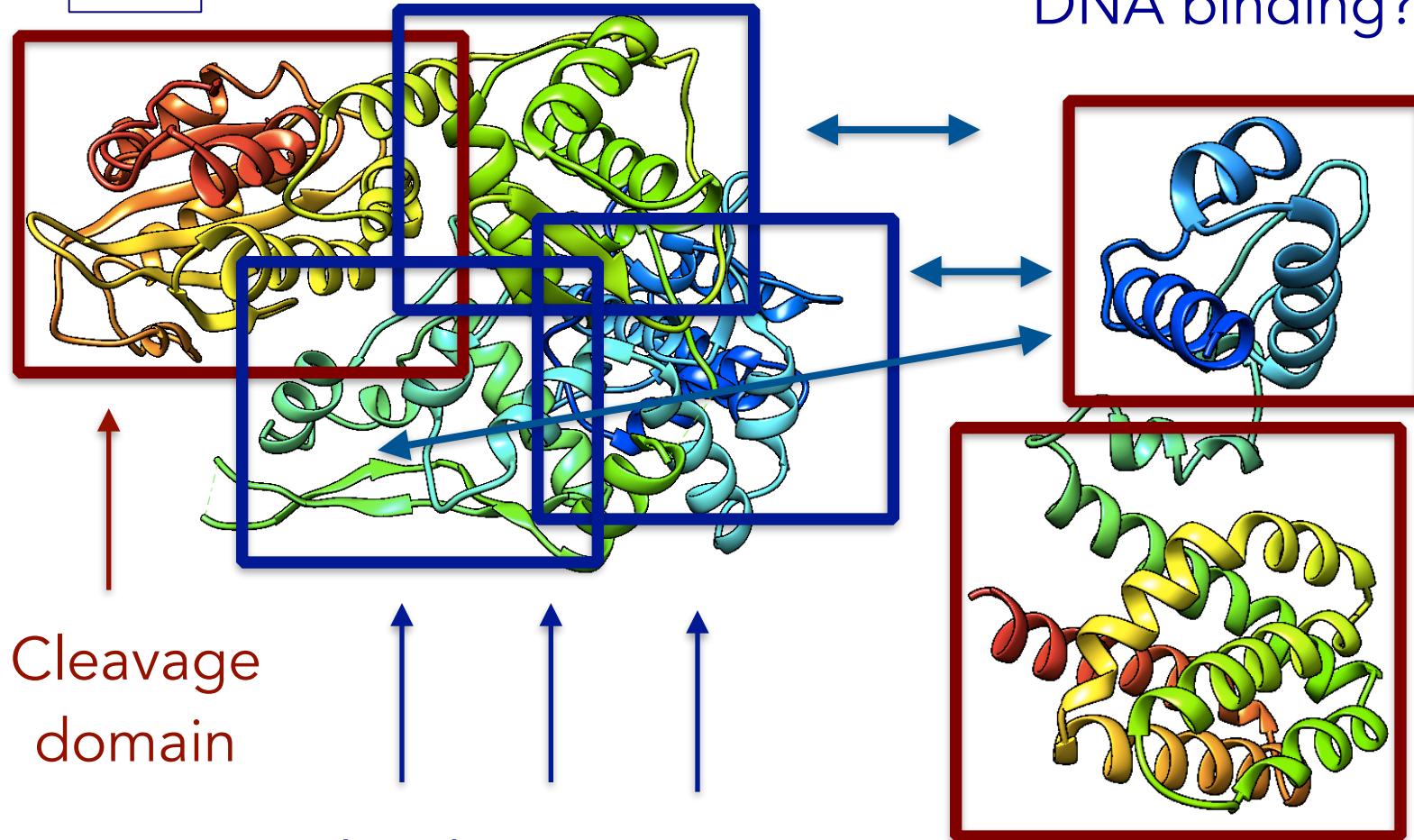


DNA binding?

1HW2

DNA binding (targeting to a  
specific DNA sequence)

1FOK

Cleavage  
domain

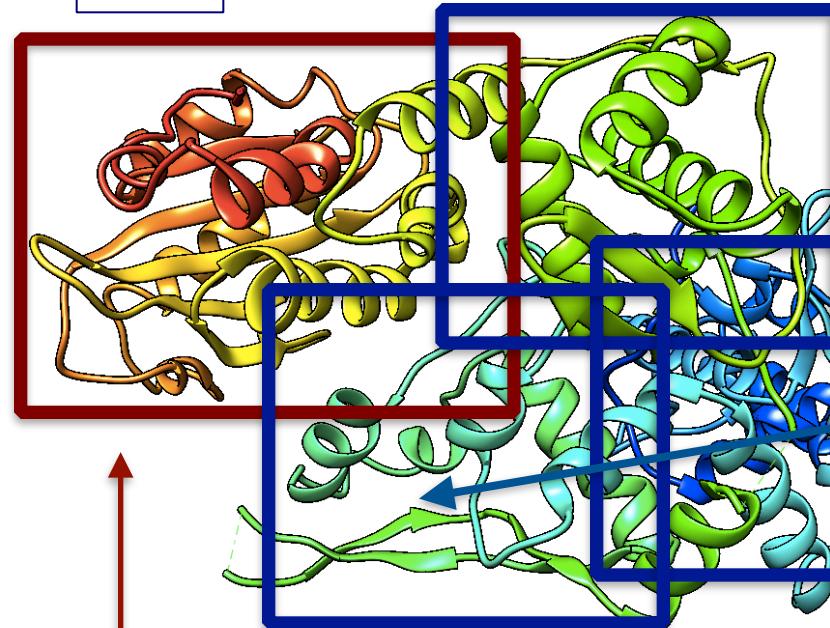
DNA binding (targeting to a  
specific DNA sequence)

DNA binding?

1HW2

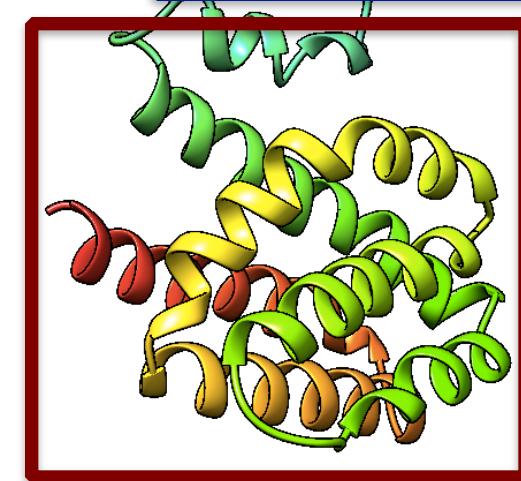
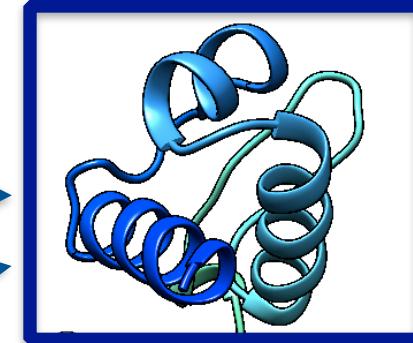
?

1FOK

Cleavage  
domain

DNA binding (targeting to a  
specific DNA sequence)

DNA binding!

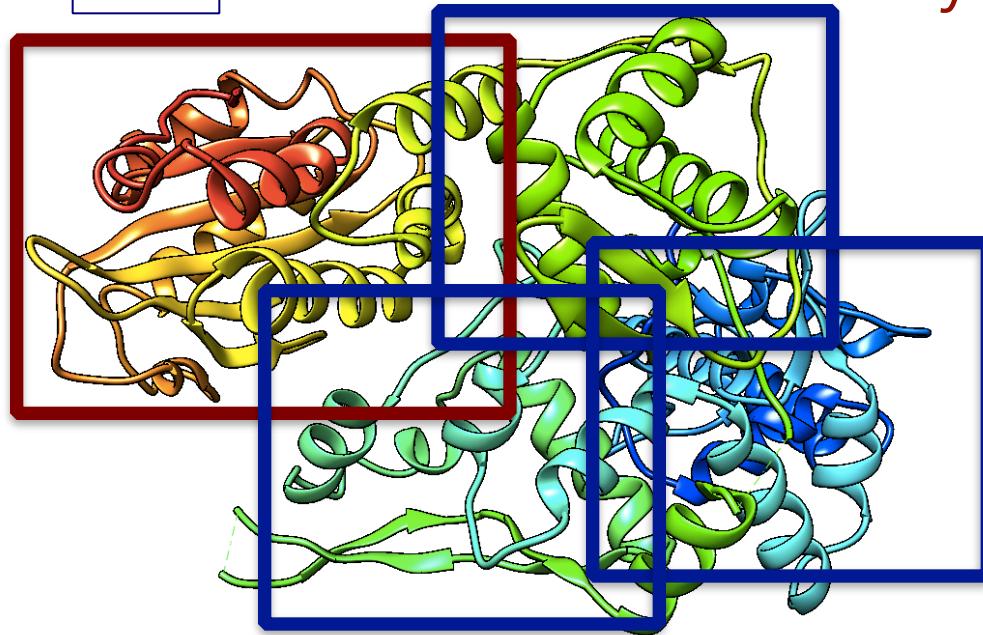


1HW2

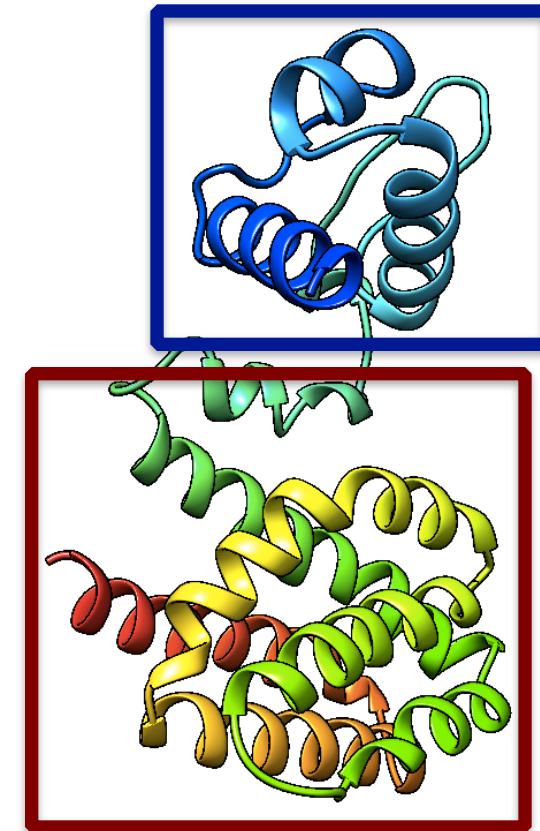
acyl-CoA  
binding domain  
controls affinity

1FOK

“syntactical change”



Restriction endonuclease



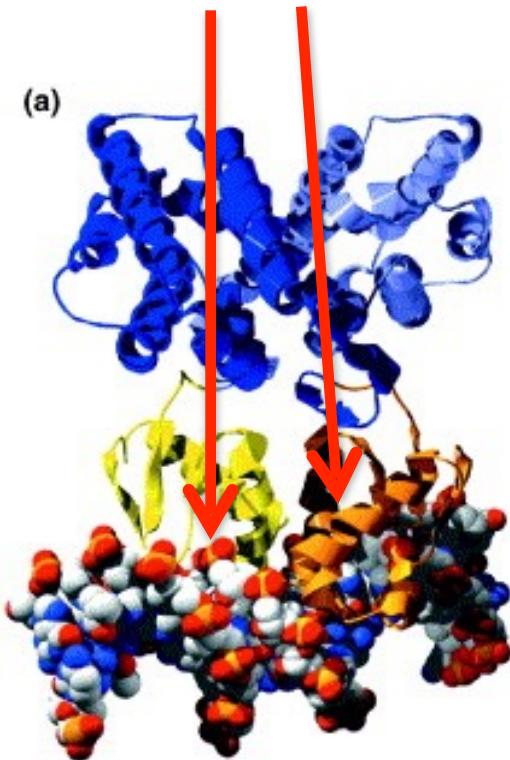
1HW2

Transcription factor

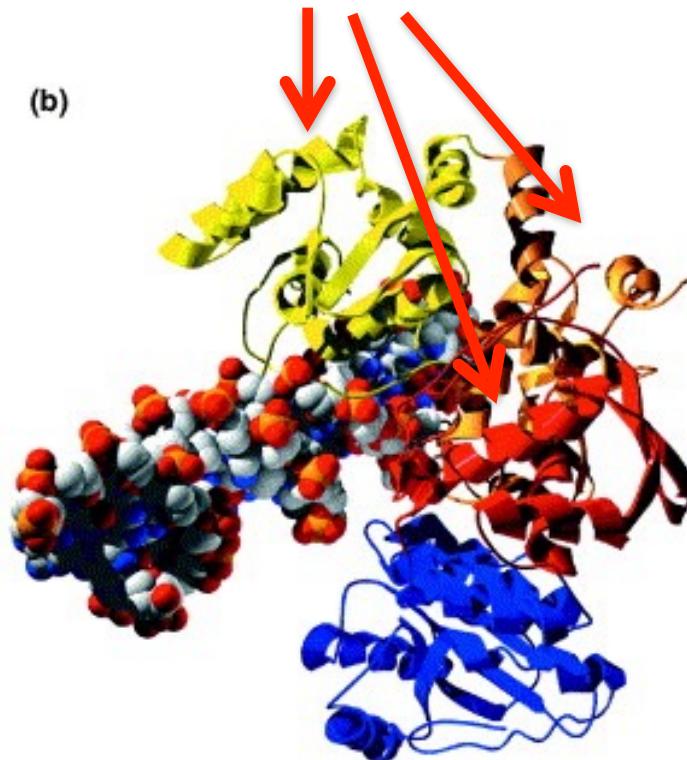
# Semantic change

Marco Punta

DNA binding

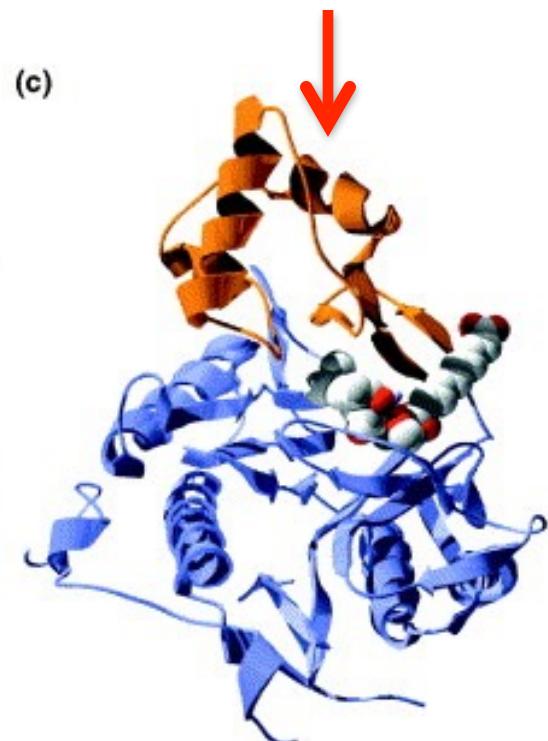


DNA binding



substrate specificity pocket

(c)



Transcription  
factor

Restriction  
endonuclease

Human methionine  
aminopeptidase 2

## “syntactical change”

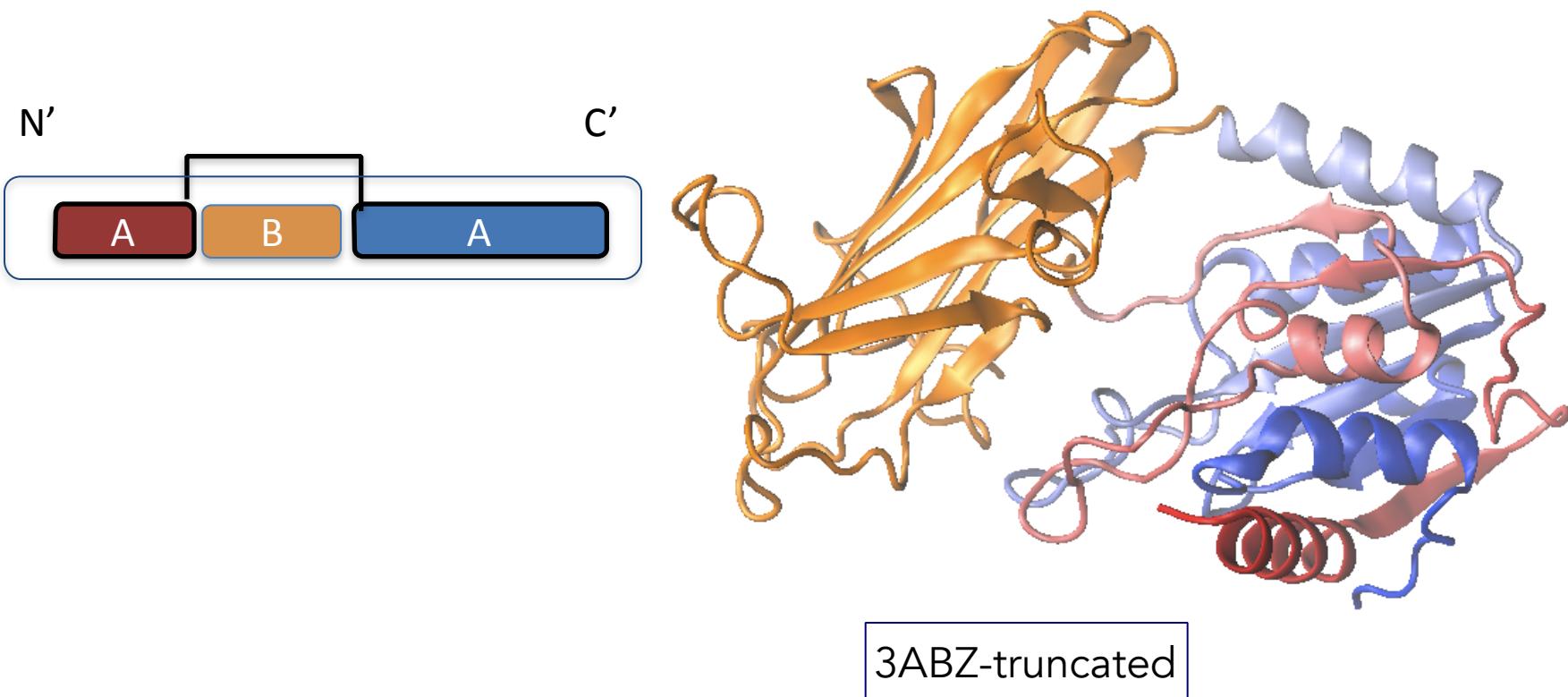
DNA sequence  
recognised



Restriction endonuclease  
 $5'-GGATG-3'$

Transcription factor  
 $5'-TGGNNNNNCCA-3'$

# “Nested” domains



# 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

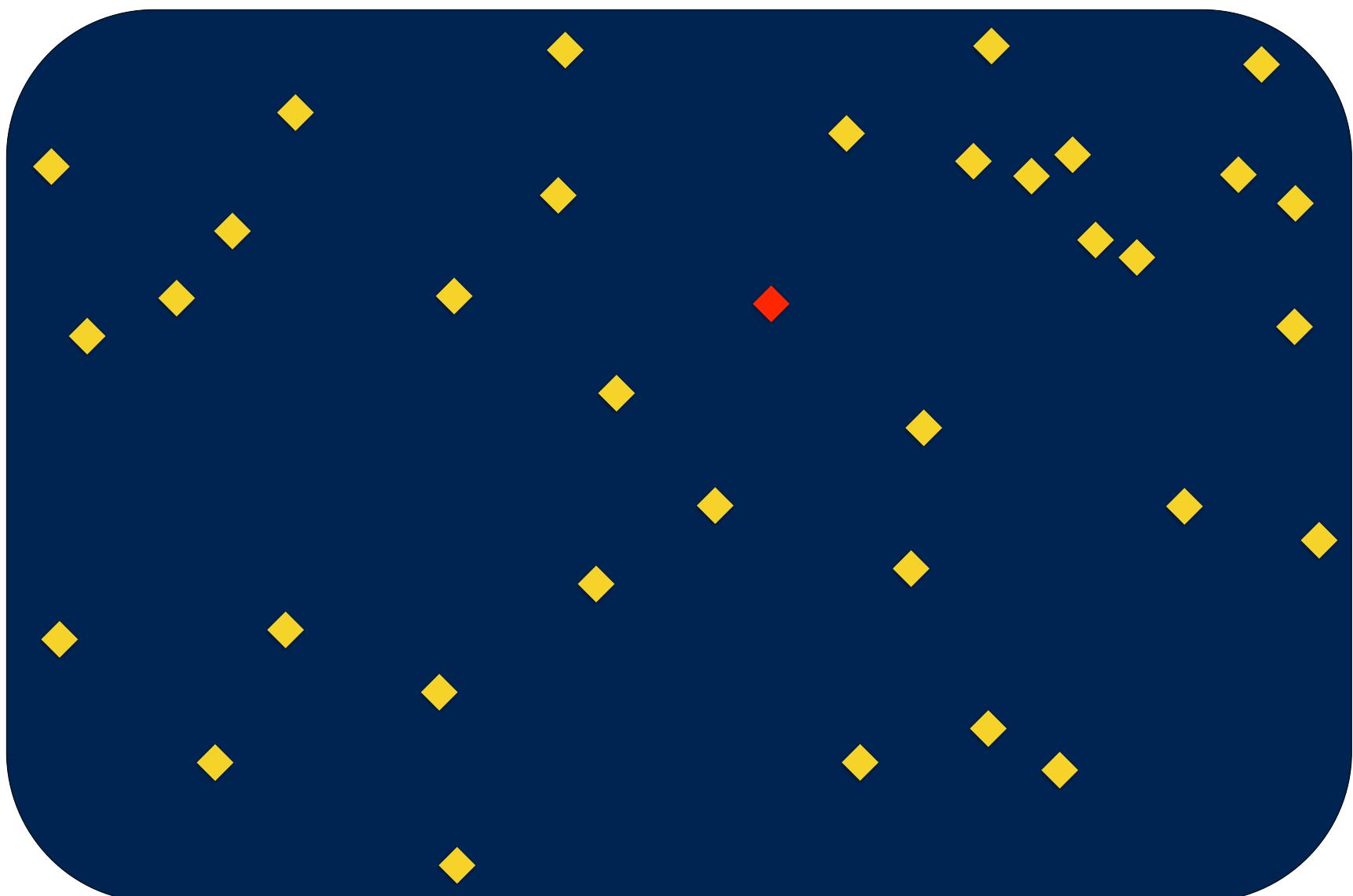
- Members will be structurally similar
- Members may share aspects of function
- Also, the whole set of members may reveal elements of protein and organism evolution

## The sequence space

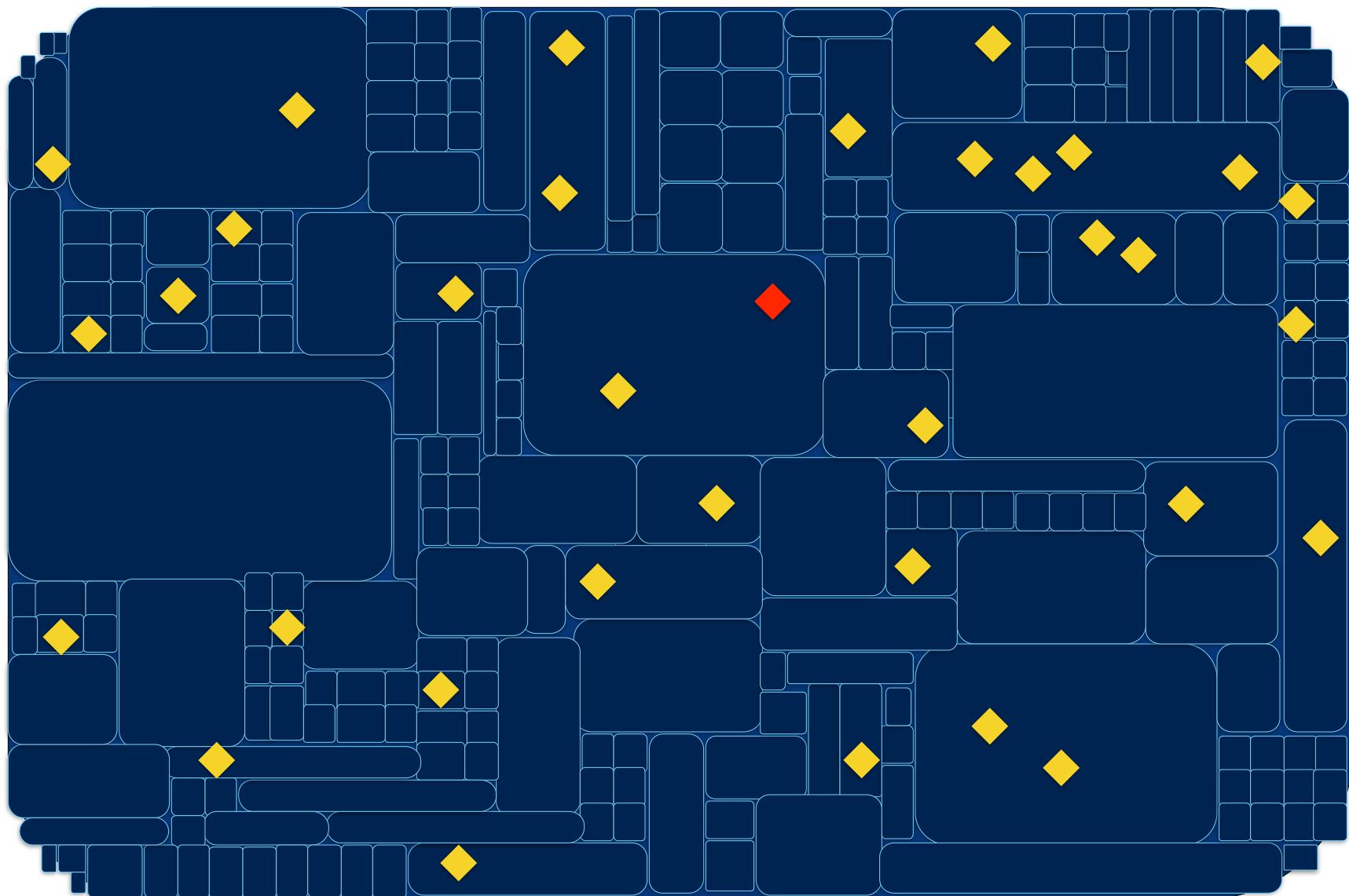
## The sequence space and annotated proteins



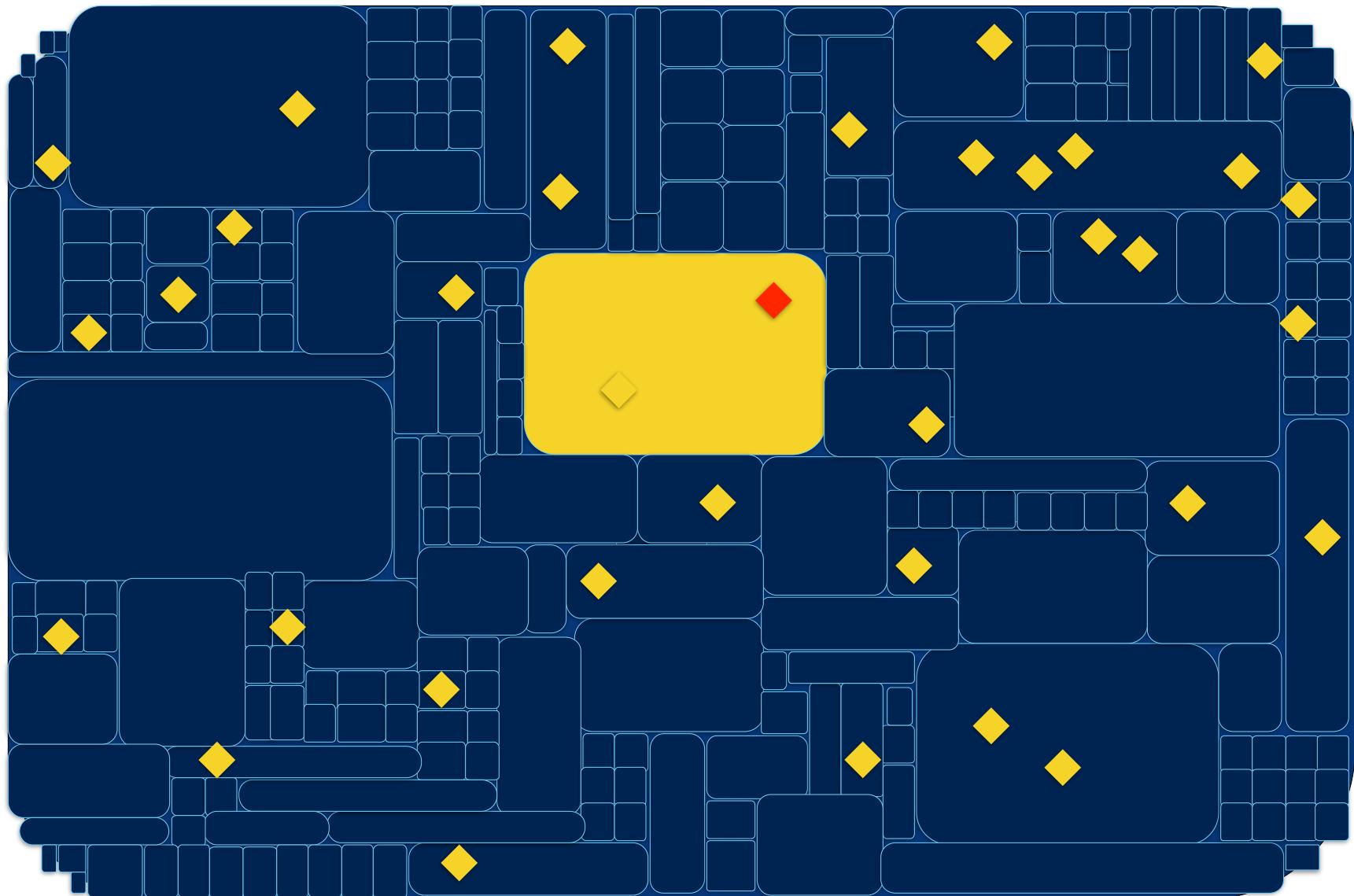
## The sequence space and annotated proteins



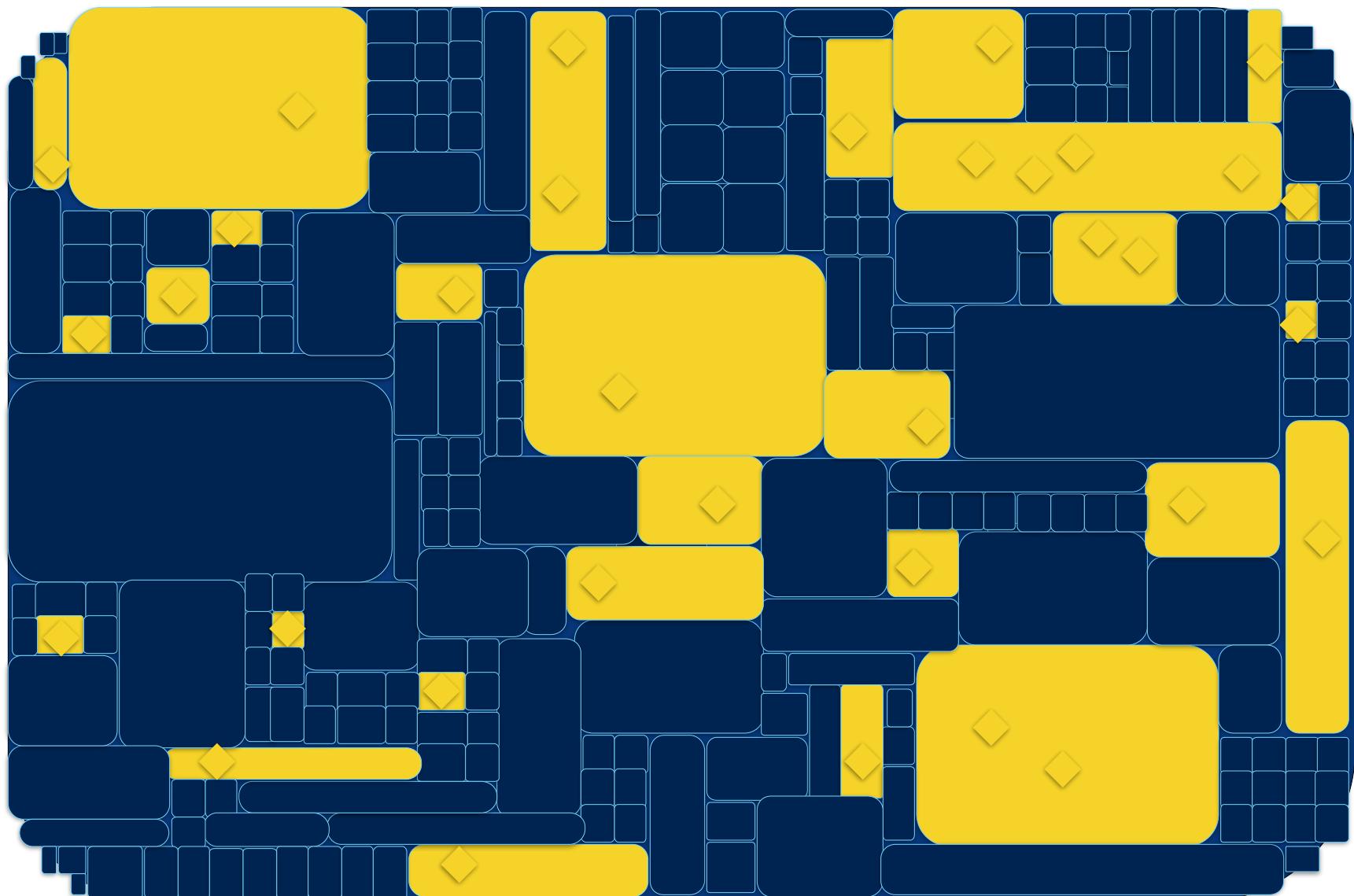
## Protein families



## Annotation transfer by homology



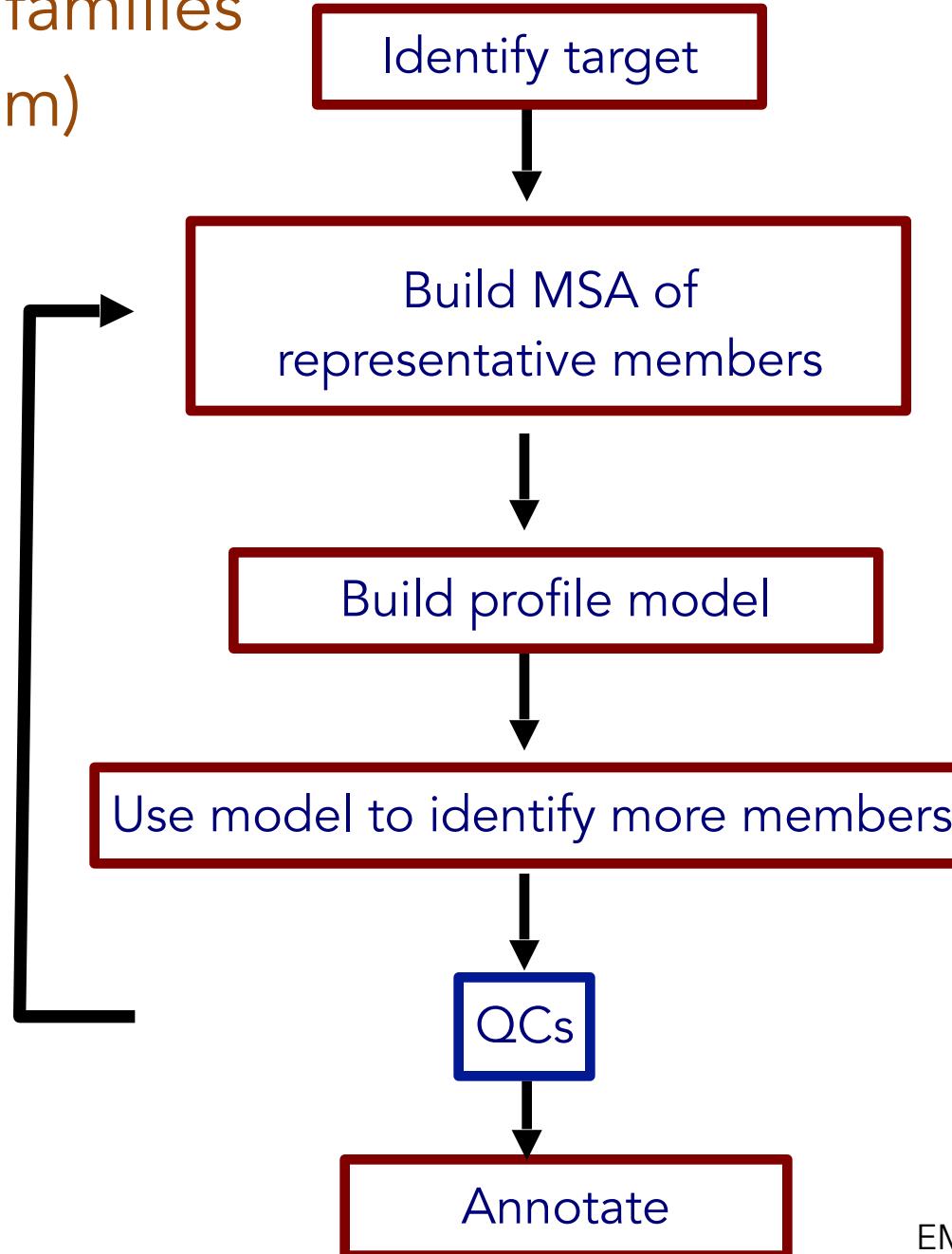
# Annotation transfer by homology



# Protein family databases

# Building families (Pfam)

Marco Punta



Human: 1 MGLSDGEWQLVLNWWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFHLKSEDEMKA  
60

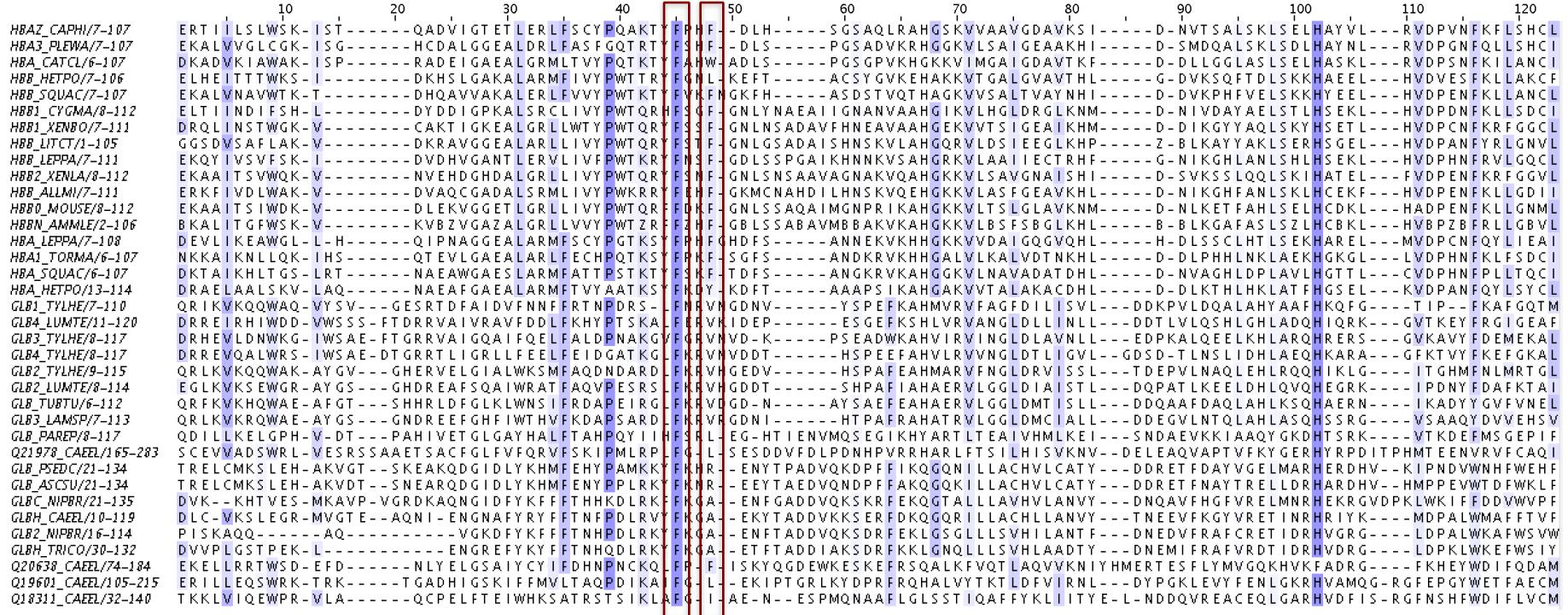
Human: 61 DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH  
120

Human: 121 PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154

# Family power

	10	20	30	40	50	60	70	80	90	100	110	120
HBAZ_CAPH/7-107	ERT I ILS LWSK- I ST-----QAD VIG T ETLER L FSC YPQAKTY FPHE- -DLH-----SGS AQL RAHS GKVVA AVGDAVKS I-----D- NVT SALS KLS ELHAY VL-- -RVD P VNFKL SHCL											
HBA3_PLEWA/7-107	EKALVVGLCGK- I SG-----HCD ALGGEA LDRLF AS FG QRTY FSHF- -DLS-----PGS ADV KR HGG KVVL SAIG EA AKH I-----D- SMD QALS KLS D L HAY NL-- -RVD PG NFQ LL SHCI											
HBA_CATCL/6-107	DKAD VKIA WAK- I SP-----RADE I GAE AL GRMLT V PQT K TY FAHW- ADLS-----PGS GPV KH GKK VIM GAIG DAVTKF-----D- DL LG GLAS L SEL HASKL-- -RVD PSN FK LL ANCI											
HBB_HETPO/7-106	ELHE ITT TWKS- I-----DK HS LGAKA LARMF I VY PWT TRY FGNL- KEFT-----AC S GVK E HAKK VT GAL GVAV THL-----G- DV KS QFTD LS KK HAEE EL-- -HVD VES FK LL AKCF											
HBB_SQUAC/7-107	EKA LVNAV WT K-T-----DH QAVV VAKA LER L FV VY PWK T K TY FVK FNG KFH-----ASD STV QT HAG KVVS ALT VAY NH I-----D- DV KPH FM ELS KK HY EEL-- -HVD P EN FK LL ANCI											
HBB1_CYGMA/8-112	ELT I IND I FSH- L-----DY DD I GPKA LSR C L I VY PWK T Q RHF S G-----GN LYNAE AII I GN ANA VAA HG I KV L HGL DR GL K NM-----D- NIV DAY AEL ST LH S EKL-- -HVD PDN FK LL SD CI											
HBB1_XENBO/7-111	DR QL I NSTW GK- V-----CA KT I GKEA L G R L L WTY PW T QRY FSS F-----GN LNS A DA VF HNE A VAA HG E KV V TS I GEA I KHM-----D- DI KG YY A QL SKY HSE TL-- -HVD PCN FK R FGG CL											
HBB_LITCT/1-105	GG SD VS A FLAK- V-----DK RAV GGEA L R L L I VY PW T QRY FST F-----G N LGS ADA I SH NS KV L VH GQ RV L D S I E EG L KHP-----Z- BL KAY Y AKL S ER H S G E L-- -HVD PAN FY RL GN VL											
HBB_ LEPPA/7-111	EK QY I VSV FS K- I-----DV DH VG A N T L E R V L I V F P W T K R Y F N S F-----G D L S P G A I K H N N K V S A H G R K V L A A I I E CTR HF-----G- NI KG HL AN L S H L S E KL-- -HVD PHN FR VL GQ CL											
HBB2_XENLA/8-112	EKA AIT S VVW QK- V-----NVE HD GHD A L G R L L I V Y PW T QRY F S N F-----G N L S N S A A V A G N A K V Q A H G K K V L S A V G N A I SH I-----D- SV K S S L Q Q L S K I H A T E L-- -F V D P E N FK R F G G V L											
HBB_B_ALLMI/7-111	ER K F IV D LWAK- V-----D V A Q C G A D A L S R M L I V Y PW K R R Y F E H F-----G K M C N A H D I L H N S KV Q E H G K K V L S F G E A V K H L-----D- NI KG H F A N L S K L H C E K F-- -HVD P E N FK LL G D I I											
HBB_D_MOUSE/8-112	EKA AIT S I WDK- V-----D LE KV GG ET L G R L L I V Y PW T Q R F D K F-----G N L S S A Q A I M G N P R I K A H G K K V L T S L G L A V K N M-----D- NL KET FA H F L S E L H C D K L-- -HAD P E N FK LL G N M L											
HBB_N_AMMLE/2-106	BK AL IT G F W S K- V-----KVB Z V G A Z A L G R L L V V Y PW T Z R F Z H E-----GB L S S A B A V M B B A V K V A H G K K V L B S F S B G L K H L-----B- BL K G A F S L Z L H C B K L-- -HVB P Z B F R L L G B V L											
HBA_ LEPPA/7-108	DEV L I K E A W G L- L- H-----Q I P NAG GE A L A R M F S C Y P G T K S Y F P H F G H D F S-----AN N E K V K H H G K V V D A I Q G Q V Q H L-----H- D L S S C L H T L S E K H A R E-----M V D P C N F Q Y L I E A I											
HBA1_TORMA/6-107	NKKA I K N L L Q K- I H S-----Q T E V L G A E A L R L F E C H P Q T K S Y F P K F- -S G F S-----A N D K R V K H H G K V L K A L V D T N K H L-----D- D L P H L N K L A E K H G K G L-- -L V D P H N F K L F S D C I											
HBA_HETPO/13-114	D K T A I K H L T G S- L R T-----N A E A W G A E A L A R M F A T T P S T K T Y F S K F- -T D F S-----A N G K R V K H A G G K V L S A V A D A T D H L-----D- N V A G H L D P L V A L H G T T L-- -C V D P H N F P L L T Q C I											
GLB1_TYLHE/7-110	D R A E L A A L S K V- L A Q-----N A E A F G A E A L A R M F T V Y A A T K S Y F K D Y- -K D F T-----A A A P S I K A H G A K V V T A L A K A C D H L-----D- D L K T H L H K L A T F H G S E L-- -K V D P A N F Q Y L S Y C L											
GLB4_LUMTE/11-120	Q R I K V K Q Q W A O- V Y S V-----G E S R T D F A I D F V N M F F R T N P D R S- -L F N R V N G D N V-----Y S P E F K A H M V R V F A G F D I L S V L-----D D K P V L D Q L A H Y A A F H K Q F G-----T I P-----F K A F Q T M											
GLB3_TYLHE/8-117	D R R E I R H I W D D- V W S S S-----F T D R R V A I V R A V E F D D L F K H P T S K A L F E R V K I D E P-----E S G E F K S H L V R V A N G L D L L I N L L-----D D T L V L Q S H L G H L A D Q H I Q R K-----G V T K E Y F R G I G E A F											
GLB3_TYLHE/8-117	D R H E V L D N W K G- I W S A E-----F T G R R V A I Q G Q A I Q E L F A L D P N A K G V F G R V N V D-----K-----P S E A D W K A H V I R V I N G L D L A V N L L-----E D P K A L Q E E L K H L A R Q H R E S-----G V K A V Y F D E M E K A L											
GLB2_TYLHE/8-117	D R R E V Q A L W R S- I W S A E-----D T G R R T L I G R L L F E E L F E I D G A T K G L F K R V N V D D T-----H S P E F A H V L R V V N G L D T L I G V L-----G D S D -T L N S L I D H L A E Q H K A R A-----G F K T V Y F K E F G K A L											
GLB2_TYLHE/9-115	Q R L K V K Q Q W A K- A Y G Y-----G H E R V E G L A W K S M F A Q D N D A R D L F K R V H G E D V-----H S P A F E A H M A R V F G N G L D R V I S S L-----T D E P V L N A Q L E H L R Q Q H I K L G-----I T G H M F N L M R T G L											
GLB2_LUMTE/8-114	E G L K V K S E W G R- A Y G S-----G H D R E A F S Q A I W R A T F A Q V P E S R S L F K R V H G D D T-----S H P A F I A H E A R V L G N G L D I A I S T L-----D Q P A T L K E E D L H Q L V Q H E G R K-----I P D N Y F D A F K T A I											
GLB2_TUBTU/6-112	Q R F K V K H Q W A E- A Y G T-----S H H R L D F G L K L W N S I F R D A P E I R G L F K R V D G D- N-----A Y S A F E F A H E A R V L G L D M T I S L L-----D D Q Q A F D Q A L H L K S Q H A E R N-----I K A D Y Y G F V E N E L											
GLB3_LAMSP/7-113	Q R L K V K R Q W A E- A Y G S-----G N D R E E F G H F I W T H V F K D A P S A R D L F K R V R G D N-----H T P A F R A H A T R V L G G L D M C I A L L-----D D E G V L N T Q L A H L A Q H S S R G-----V S A A Q Y D V V E H S V											
GLB_PAREP/8-117	Q D I L L K E L G P H- V- D T-----P A H I V E T G L G A Y H A L F T A H P Q Y I I H F S R L-----E G- H T I E N V M Q S E G I K H Y A R T L T E A I V H M L K E I-----S N D A E V K K I A A Q Y G K D H T S R K-----V T K D E F M S G E P I F											
Q21978_CAEEL/165-283	S C E V V A D S W R L- V E S R S S A A E T S A C F G L F V F Q R V F S K I P M L R P L F G-----L- S E S D D V F D L P D N H P V R R H A R L T S I L H I S V K N V-----D E L E A Q V P A T V F K Y G E R H Y R P D I T P H M T E E N V R V F C A Q I											
GLB_PSED/C/21-134	T R E L C M K S L E H- A K V G T-----S K E A K Q D G I D L Y K H M F E H Y P A M K K Y F K H R-----E N Y T P A D V Q K D P F F I K Q G Q N I L L A C H V L C A T Y-----D D R E T F D A Y V G E L M A R H E R D H V-----K I P N D V W W H F E W E H F											
GLB_A_SCSU/21-134	T R E L C M K S L E H- A K V D T-----S N E A R Q D G I D L Y K H M F E N Y P P L R K Y F K N R-----E E Y T A E D V Q N D P F F A K Q Q Q K I L L A C H V L C A T Y-----D D R E T F N A Y T R E L L D R H A R D H V-----H M P P E V W W T D F W K L F											
GLB_C_NPBR/21-135	D V K- -K T V E S- M K A V P-----V G R D K A Q N G I D E Y K F F T H H K D L R K F F K G A-----E N F G A D D V Q K S K R F E K Q T A L L L A V H V L A N V Y-----D N Q A V F H G F V R E L M N R H E K R G V D P K L W K I F F D D V W V P F											
GLB_H_CAEEL/10-119	D L C- V K S L E G R- M V G T E-----A Q N I-----E N G N A F Y R Y F F T N F P D L R V Y F K G A-----E K Y T A D D V K K S E R F D K Q G Q R I L L A C H L L A N V Y-----T N E E V F K G Y V R E T I N R H R I Y K-----M D P A L W M A F F T V F											
GLB2_NPBR/16-114	P I S K A Q Q-----A Q-----V G K D F Y K F F F T N H P D L R K Y F K G A-----E N F T A D D V Q K S D R F E K L G S G L L L S V H I L A N T F-----D N E D V F R A F C R E T I D R H V G R G-----L D P A L W K A F W S V W											
GLB_H_TRICO/30-132	D V V P L G S T P E K- L-----E N G E F Y K Y F F T N H Q D L R K Y F K G A-----E T F T A D D I A K S D R F K K L G N Q L L L S V H L A A D T Y-----D N E M I F R A F V R D T I D R H V D R G-----L D P K L W K E F W S I Y											
Q20638_CAEEL/74-184	E K E L L R R T W S D- E F D-----N L Y E L G S A I Y C Y F F D H N P N C K Q L F P-----F- I S K Y Q G D E W K E S K E F R S Q A L K F V Q T L A Q V V K N I Y H M E R T E S F L Y M V G Q K H V K F A D R G-----F K H E Y W D I F Q D A M											
Q19601_CAEEL/105-215	E R I L L E Q S W R K- T R K-----T G A D H I G S K I F F M V L T A Q P D I K A I F G- L-----E K I P T G R L K Y D P R F R O H A L V Y T K T L D F V I R N L-----D Y P G K L E V Y F E N L G R H V A M Q G- R G F E P G Y W E T F A E C M											
Q18311_CAEEL/32-140	T K K L V I Q E W P R- V L A-----Q C P E L F T E I W H K S A T R S T S I K L A F G- I-----A E- N-----E S P M Q N A A F G L G L S S T I Q A F F Y K L I I T Y E- L- N D D Q V R E A C E Q L G A R H V D F I S-----R G F N S H F W D I F L V C M											

# Family power



# Family power

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

Mouse: 1 MGLSDGEWQLVNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEDEMKGSE 60



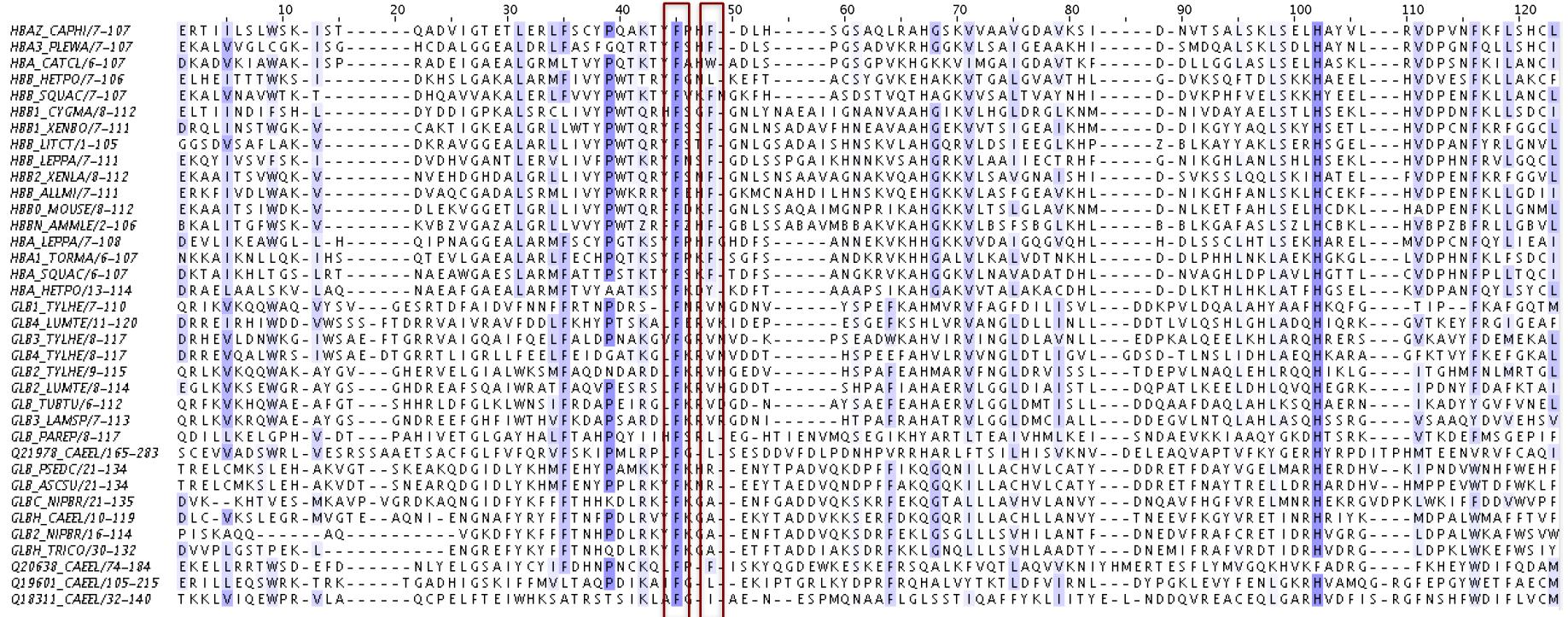
Human: 61 DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120  
DLKKHG TVLTALG ILKKKG H AEI PLAQSHATKHKIPVKYLEFISE II VL H

Mouse: 61 DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIIEVLKKRH 120

Human: 121 PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154  
GDFGADAQGAM KALELFR D A YKELGFGQG

Mouse: 121 SGDFGADAQGAMS KALELFRNDIAAKYKELGFQG 154

# Family power



seq> AWRTWEAPIFLKRYSTLPGGRAS...

# Family power

Sequence alignment of the HBAZ family. The sequence is shown from position 10 to 120. Red boxes highlight conserved regions across the family. A blue line connects the sequence to the title "seq> AWRTWEAPIFLKRYSTLPGGRAS...".

```

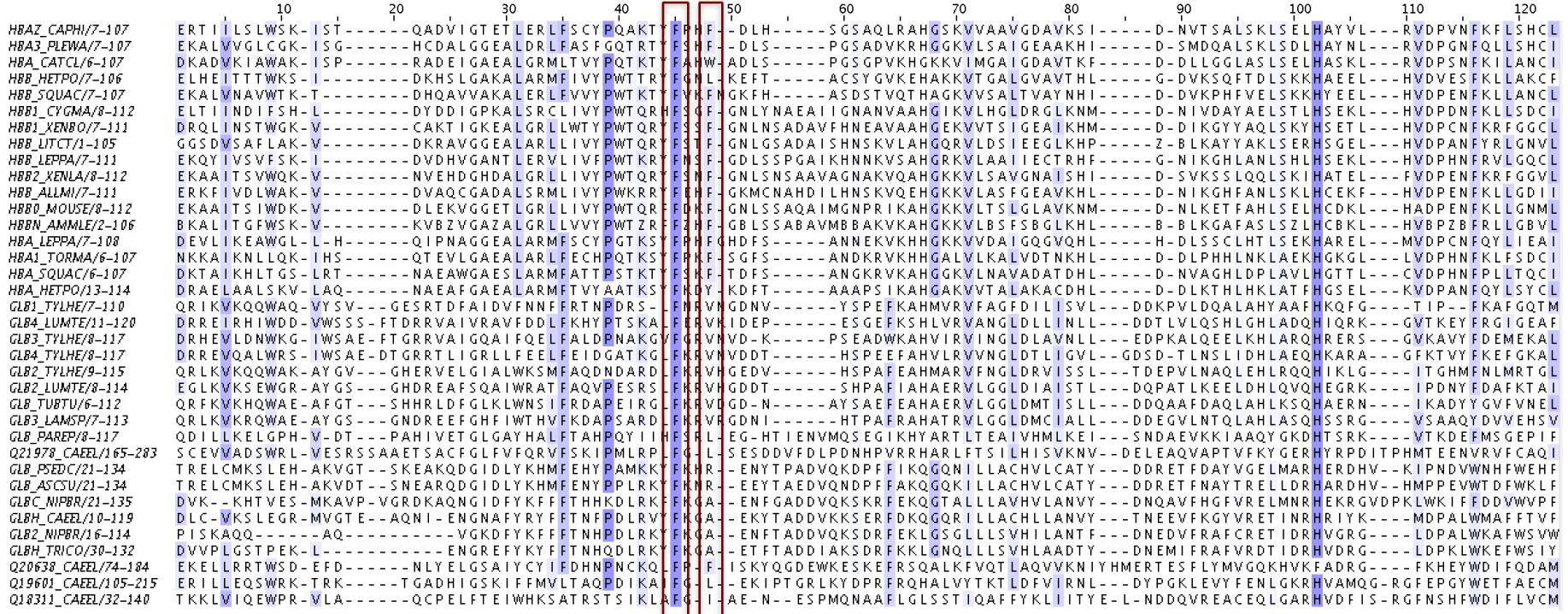
ERT I I L S L W S K - I S T - - - Q A D V I G T E T L E R L F S C Y P P A K T Y F P H F - - - D L H - - - S G S A Q L R A H G S K V V A A V G D A V K S I - - - D - N V T S A L S K L S L E L H A Y V L - - - R V D P V N F K F L S H C L
E K A L V V G L C G K - I S G - - - H C D A L G G E A L D R L F A S F G Q T R T Y F S H - - - D L S - - - P G S A D V K R H G G K V L S A I G E A A K H I - - - D - S M D Q A L S K L S D L H A Y N L - - - R V D P G N F Q L L S H C I
D K A D V I A W A K - I S P - - - R A D E I G A E A L G R M L T V Y P Q T K T Y F A H W - A D L S - - - P G S G P V K H G K K V I M G A I G D A V T K F - - - D - D L L G G L A S L S E L H A S K L - - - R V D P S N F K I L A N C I
E L H E I T T W K S - I - - - D K H S L G A K A L A R M F I V Y P W T T R Y F G N L - K E F T - - - A C S Y G V K E H A K K V T G A L G V A V T H L - - - G - D V K S Q F T D L S K K H A E E L - - - H V D V E S F K L L A K C F
E K A L V N A V W T K - T - - - D H Q A V V V A K A L E R L F V V Y P W T K T Y F Y K F M G K F H - - - A S D S T V Q T H A G K V V S A L T V A Y N H I - - - D - D V K P H F V E L S K K H Y E E L - - - H V D P E N F K L L A N C L
E L T I I N D I F S H - L - - - D Y D D I G P K A L S R C L I V Y P W T Q R H E S G F - G N L Y N A E A I I G N A N V A A H G I K V L H G L D R G L K N M - - - D - N I V D A Y A E L S T L H S E K L - - - H V D P D N F K L L A N C L
H B B 1 _ X E N B O / 7 - 1 1 1 D R Q L I N S T W G K - V - - - C A K T I G K E A L G R L L W T Y P W T Q R Y F S S F - G N L N S A D A V F H N E A V A A H G E K V V T S I G E A I K H M - - - D - D I K G Y Y A Q L S K Y H S E T L - - - H V D P C N F K R F G G C L
H B B _ L I C T / 1 - 1 0 5 G G S D M S A F L A K - V - - - D K R A V G G E A L A R L L I V Y P W T Q R Y F S T F - G N L G S A D A I S H N S K V L A H G Q R V L D S I E E G L K H P - - - Z - B L K A Y Y A K L S R E H S G E L - - - H V D P A N F Y R L G N V L
H B B _ L E P P A / 7 - 1 1 1 E K Q Y I V S V F S K - I - - - D V D H V G A N T L E R V L I V F P W T K R Y F S M F - G D L L S P G A I K H N N K V S A H G R K V L A A I I E C T R H F - - - G - N I K G H L A N L S H L S E K L - - - H V D P H N F R V L Q G C L
H B B 2 _ X E N L A / 8 - 1 1 2 E K A A I T S V W Q K - V - - - N V E H D G H D A L G R L L I V Y P W T Q R Y F S M F - G N L S N S A A V A G N A K V O A H G K K V L S A V G N A I S H I - - - D - S V K S S L Q Q L S K I H A T E L - - - F V D P E N F K R F G G V L
H B B 2 _ A L U M / 7 - 1 1 1 E R K F I V D L W Q A K - V - - - D V A Q C G A D A L S R M L I V Y P W K R K Y F E H F - G K M C N A H D I L H N S K V Y Q E H G K K V L S F G E A V K H L - - - D - N I K G H F A N L S K I H C E K F - - - H V D P E N F K L L G D I I
H B B 3 _ M O U S E / 8 - 1 1 2 E K A A I T S I W D K - V - - - D L E K V G G E T L G R L L I V Y P W T Q R Y F D K F - G N L S S A Q A I M G N P R I K A H G K K V L T S L G L A V K N M - - - D - N L K E T F A H L S E L H C D K L - - - H A D P E N F K L L G N M L
H B B N _ A M M L E / 2 - 1 0 6 B K A L I T G F W S K - V - - - K V B Z V G A Z A L G R L L L V V Y P W T Z R F E Z H F - G B L S S A B A V M B B A K V K A H G K K V L B S F S B G L K H L - - - B - B L K G A F A S L S Z L H C B K L - - - H V B P Z B F R L L G B V L
H B A _ L E P P A / 7 - 1 0 8 D E V L I K E A W G L - L - H - - - Q I P N A G E A L A R M F S C Y P G T K S F P R H F C O H D F S - - - A N N E K V Y H K H G K V V D A I G Q G V Q H L - - - H - D L S S C L H T L S E K H A R E S - - - M V D P C N F Q Y L I E A I
H B A 1 _ T O R M A / 6 - 1 0 7 N K K A I K N L L Q Q K - I H S - - - Q T E V L G A E A L A R L F E C H P Q T K S F P R H F - S G F S - - - A N D K R V K H H G A L V L K A L V D T N K H L - - - D - D L P H L N K L A E K H G K G L - - - L V D P H N F K L F S D C I
H B A _ S Q U A C / 6 - 1 0 7 D K T A I K H L T G S - L R T - - - N A E A W G A E S L A R M F A T T P S T K T Y F S H F - T D F S - - - A N G K R V K H A G G K V L N A V A D A T D H L - - - D - N V A G H L D P L V A L H G T T L - - - C V D P H N F P L L T Q C I
H B A _ H E T P O / 1 3 - 1 1 4 D R A E L A L A S K V - L A Q - - - N A E A F G A E A L A R M F T V Y A A T K S Y F K D Y - K D F T - - - A A P S I I K A H G K V V T A L A K A C D H L - - - D - D L K T H L H K L A T F H G S E L - - - K V D P A N F Q Y L S Y C L
G L B 1 _ T Y L H E / 7 - 1 1 0 Q R I K M K Q Q W A Q - M Y S V - - - G E S R T D F A I D V F N N F F R T N P D R S - F N R V M G D N V - - - Y S P E F K A H M V R V F A G F D I L I S V L - - - D D K P V L D Q A L A H Y A A F H K Q F G - - - T I P - - F K A F G Q T M
G L B 4 _ L U M T E / 1 1 - 1 2 0 D R R E I R H I W W D - D V W S S - F T D R R V A I V R A V F D D L F K H Y P T S K A - F F R V H I D E P - - - E S G E F K S H L V R V M A N G L D L I N L L - - - D D T L V L Q S H L G H L A D Q H I Q R K - - - G V T K E Y F R G I E A F
G L B 3 _ T Y L H E / 8 - 1 1 7 D R H E V L D N W K G - I W S A E - F T G R R V A I Q A I F Q E L F A L D P N A K G Y F G R V M V D - K - - - P S E A D W K A H V I R M V G L D L A V N L L - - E D P K A L Q E E L K H L A R Q H R E S - - - G V K A V Y F D E M E K A L
G L B 4 _ T Y L H E / 8 - 1 1 7 D R R E V Q A L W R S - I W S A E - D T G R R T L I G R L L F E L F E I D G A T K G - F H R V M V D D T - - - H S P F E F A H V L R V V N G L D T L I G V L - - G D S D - T L N S L I D H L A E Q H K A R A - - - G F K T V Y F K E F G K A L
G L B 2 _ T Y L H E / 9 - 1 1 5 Q R L K M K Q Q W A K - A Y G V - - - G H E R V E L G I A L W K S M F A Q D N D A R D - F H R V H G E D V - - - H S P A F E A H M A R V F N G L D R V I S S L - - - T D E P V L N A Q L E H L R Q Q H I K L G - - - I T G H M F N L M R T G L
G L B 2 _ L U M T E / 8 - 1 1 4 E G L K V K S E W G R - A Y G S - - - G H D R E A F S Q A I W R A T F A Q V P E S R S - F H R V H G D D T - - - S H P A F I A H E A R V L G G L D I A I S T L - - - D Q P A T L K E E D L H Q V G H E R G K - - - I P D N Y D F A F K T A I
G L B 2 _ T U B T U / 6 - 1 1 2 Q R F L K M K H Q W A E - A F G T - - - S H H R L D F G L K L W N S I F R D A P E I R G - F H R V D G D - N - - - A Y S A E F E A H A E R V L G G L D M T - I S L L - - - D D Q A A F D A Q L A H L K S Q H A E R N - - - I K A D Y Y G V F V N E L
G L B 3 _ L A M S P / 7 - 1 1 3 Q R L K M K R Q W A E - A Y G S - - - G N D R E E F G H F I W T H V F K D A P S A R D - F H R V P G D N I - - - H T P A F R A H A T R V L G G L D M C I A L L - - - D D E G V L N T Q L A H L A S Q H S S R G - - - V S A A Q Y D V V E H S V
Q 2 1 9 7 8 _ C A E E L / 1 6 5 - 2 8 3 Q D I L L K E L G P H - V - D T - - - P A H I V E T G L G A Y H A L F T A H P Q Y I I H E S P R L - E G - H T I E N V M Q S E G I K H Y A R T L T E A I V H M L K E I - - - S N D A E V K K I A A Q Y G K D H T S R K - - - V T K D E F M S G E P I F
G L B _ P S E D C / 2 1 - 1 3 4 S C E V V A D S W R L - - V E S R S S A A E T S A C I G L F V F Q R V F S K I P M L R P L F G I L - - S E S D D V F D L P D N H P V R R H A R L F T S I I H I S V K N V - - D E L E A Q V P A T V F K Y G E R H Y R P D I T P H M T E E N V R V F C A Q I
G L B _ A S C V U / 2 1 - 1 3 4 T R E L C M K S L E H - A K V G T - - - S K E A K Q D G I D L Y K H M F E H Y P A M K K Y F K H R - - E N Y T P A D V Q K D P F F I K Q Q G N I L L A C H V L C A T Y - - - D D R E T F D A Y V G E L M A R H E R D H V - - K I P N D V W N H F W E H F
G L B C _ N P B R / 2 1 - 1 3 5 T R E L C M K S L E H - A K V D T - - - S N E A R Q D G I D L Y K H M F E N Y P P L R K Y F K H R - - E E Y T A E D V Q N D P F F A K Q Q G Q K I L L A C H V L C A T Y - - - D D R E T F N A Y T R E L L D R H A R D H V - - H M P P E V V W T D F W K L F
D V K - - K H T V E S - M K A V P - V G R D K A Q N G I D F Y K F F F T H H K D L R K F H G A - - E N F G A D D V Q K S K R F E K Q O G T A L L L A V H V L A N V Y - - D N Q A V F H G F V R E L M N R H E K R G V D P K L W K I F F D D V V W V P F
G L B H _ C A E E L / 1 0 - 1 1 9 D L C - M K S L E G R - M V G T E - A Q N I - E N G N A F Y R Y F F T N F P D L R V Y F H G A - - E K Y T A D D V K K S E R F D K Q Q G R I L L A C H L L A N V Y - - T N E E V F K G Y V R E T I N R H R I Y K - - - M D P A L W M A F F T V F
G L B H _ N P B R / 1 6 - 1 1 4 P I S K A Q - - - A Q - - - V G K D F Y K F F F T N H P D L R K Y F H G A - - E N F T A D D V Q K S D R F E K L G S G L L L S V H I L A N T F - - - D N E D V F R A F C R E T I D R H V G R G - - - L D P A L W K A F W S V W
G L B H _ T R I C O / 3 0 - 1 3 2 D V V P L G S T P E K - L - - - E N G R E F Y K Y F F T N H Q D L R K Y F H G A - - E T F T A D D I A K S D R F K K L G N Q L L S V H L A A D T Y - - - D N E M I F R A F V R D T I D R H V D R G - - - L D P K L W K E F W S I Y
Q 2 0 6 3 8 _ C A E E L / 7 4 - 1 8 4 E K E L L R R T W S D - E F D - - - N L Y E L G S A I Y C Y I F D H N P N C K Q - F F - I S K Y Q G D E W K E S K E F R S Q A L K F V Q T L A Q V V K N I Y H M E R T E S F L Y M V G Q K H V K F A D R G - - - F K H E Y W D I F Q D A M
Q 1 8 6 0 1 _ C A E E L / 1 0 5 - 2 1 5 E R I L L E Q S W R K - T R K - - - T G A D H I G S K I F F M V L T A Q P D I K A - F G - L - - E K I P T G R L K Y D P R F R Q H A L V Y T K T L D F V I R N L - - - D Y P G K L E V Y F E L G K R R H V A M Q G - R G F E P G Y W E T F A E C M
Q 1 8 3 1 1 _ C A E E L / 3 2 - 1 4 0 T K K L V I Q E W P R - V L A - - - Q C P E L F T E I W H K S A T R S T S I K L A F G - I - A E - N - - E S P M Q N A A F L G L S S T I Q A F F Y K L I I T Y E - L - N D D Q V R E A C E Q L G A R H V D F I S - R G F N S H F W D I F F L V C M

```

seq> AWRTWEAPIFLKRYSTLPGGRAS...

# Family power

score(ab,i)



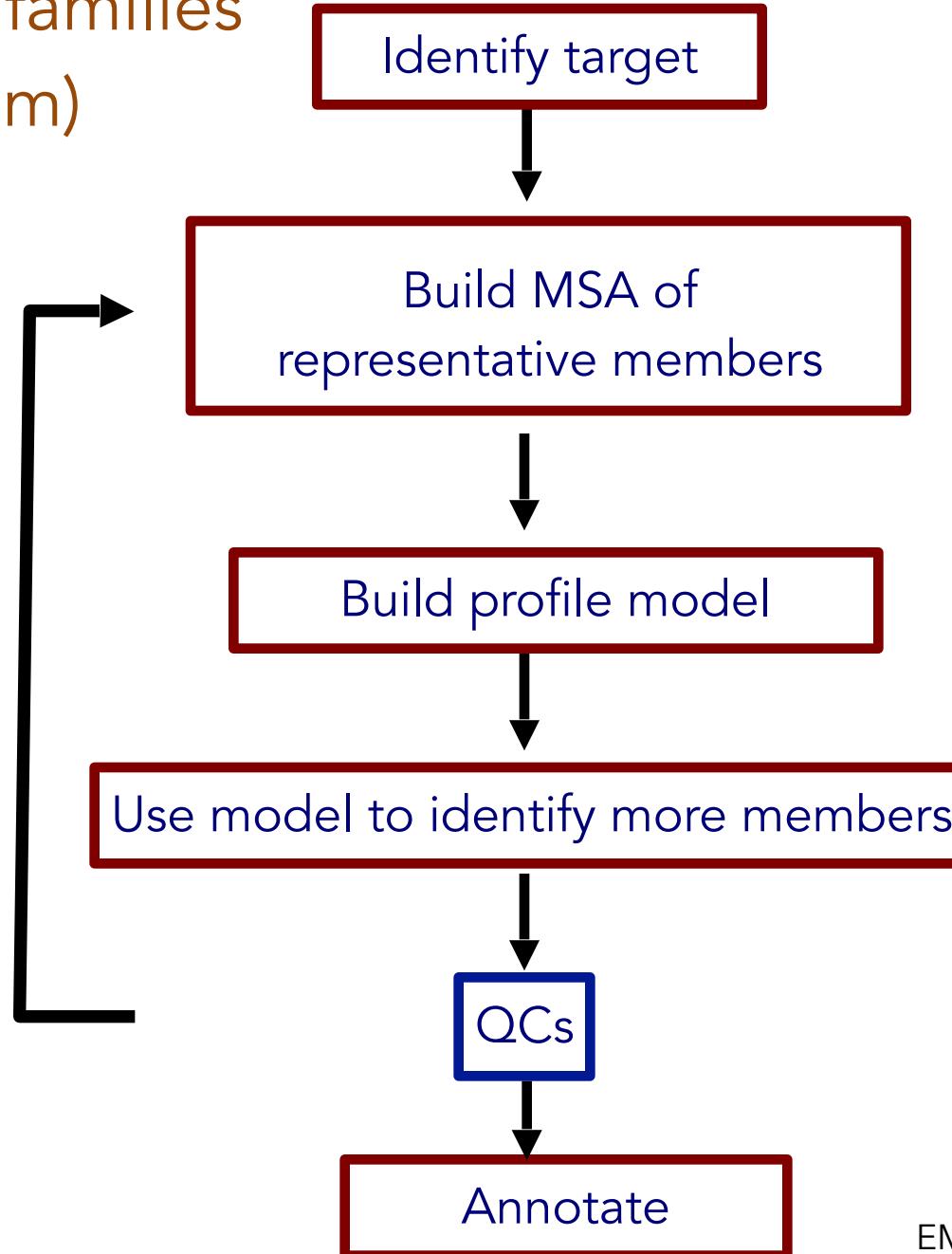
seq> AWRTWEAPIFLKRYSTLPGGRAS...

# Sequence-profile alignments

- Position specific substitution matrices
- profile-hidden Markov models

# Building families (Pfam)

Marco Punta



# Functions, organisms, structures

Marco Punta

Prokaryotes

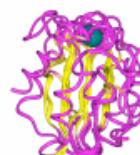


~4400 families

Signalling, extracellular and chromatin-associated proteins



~1000 domains



Gene3D  
Structural domains from CATH

Structural domains from SCOP

# No limits, domains

Marco Punta



# No limits, full-length proteins

Marco Punta

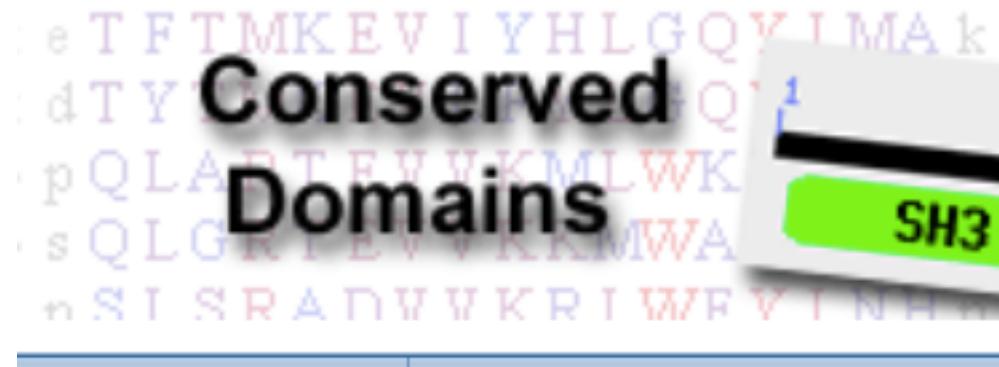


>7000 families, >50000 subfamilies



~2000 families

## CDD



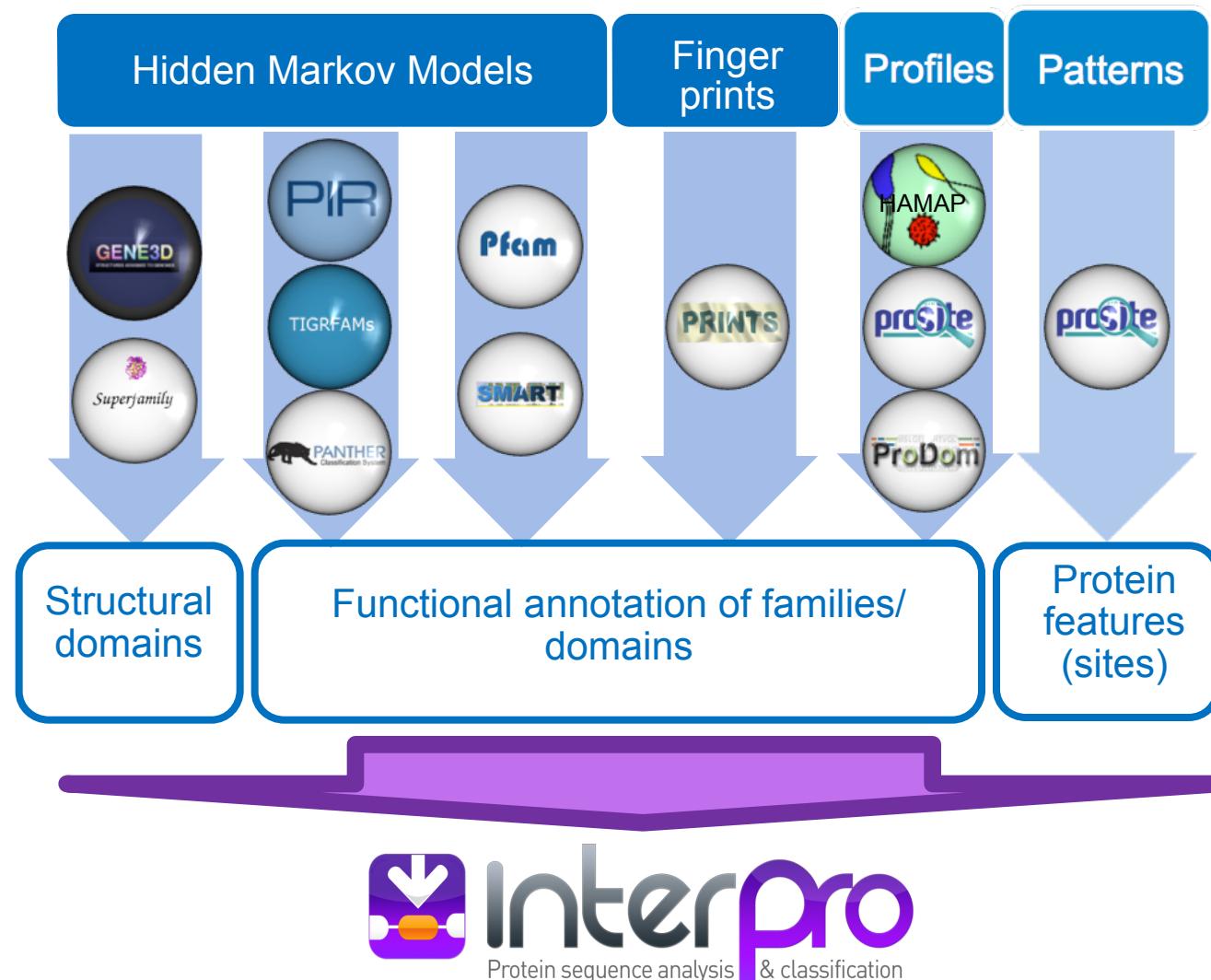
Uses RPS-BLAST

# Integration

Marco Punta



# Member databases



**Overview**

Similar proteins (2905)

Structures

**Filter view on****Entry type**

-  Family
-  Domains
-  Repeats
-  Site

**Status**

-  Unintegrated

**Colour by**

- domain relationship
- source database

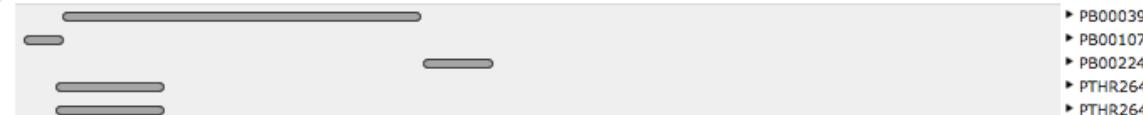
**P Protein**[Export FASTA](#)**Phosphotransferase RcsD (P39838)**

Source: UniProtKB

**Accession** [P39838 \(RCSD\\_ECOLI\)](#)**Species** Escherichia coli (strain K12)**Length** 890 amino acids (complete)**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**

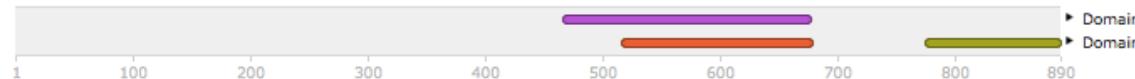
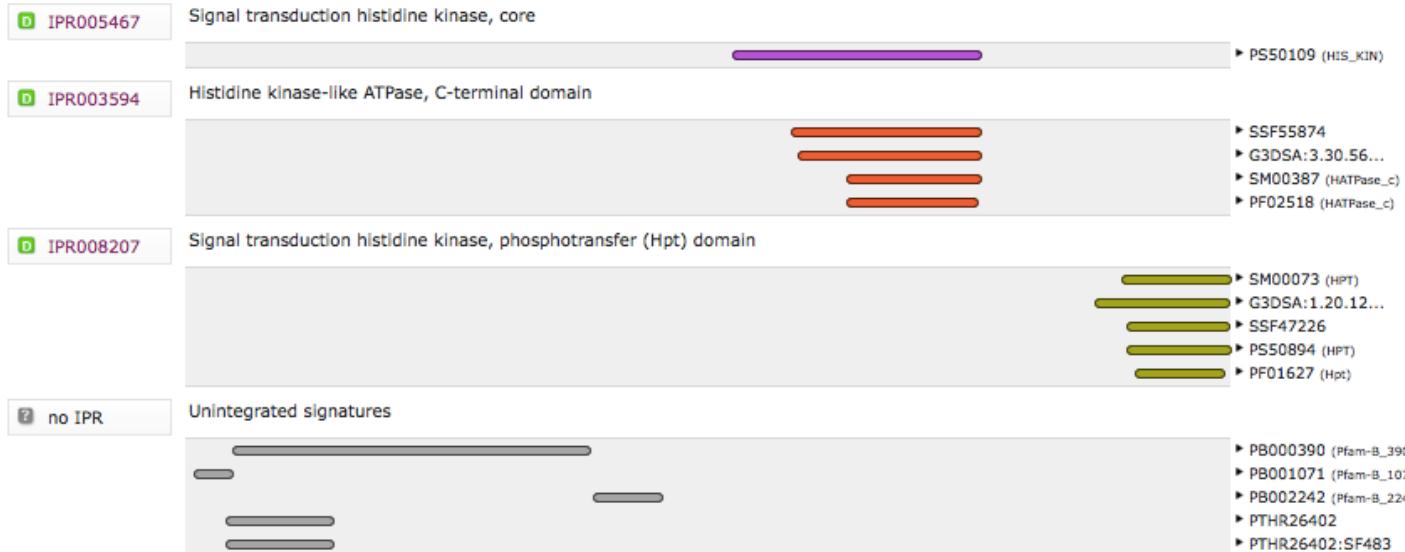
- domain relationship
- source database

**P 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**

**Overview**

Similar proteins (2905)

Structures

**Filter view on****Entry type**

-  Family
-  Domains
-  Repeats
-  Site

**Status**

-  Unintegrated

**Colour by**

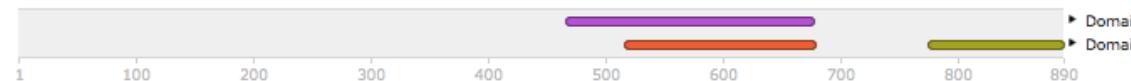
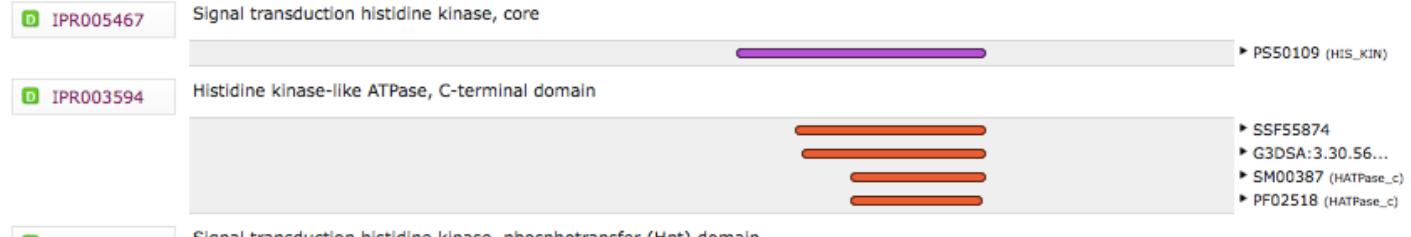
- domain relationship
- source database

**P Protein**[Export FASTA](#)**Phosphotransferase RcsD (P39838)****Accession** [P39838 \(RCSD\\_ECOLI\)](#)**Species** Escherichia coli (strain K12)**Length** 890 amino acids (complete)

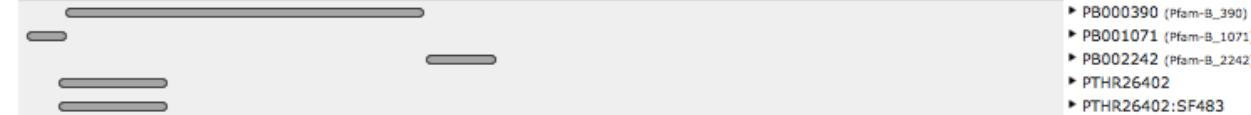
Source: UniProtKB

**Protein family membership**

None predicted.

**Domains and repeats****Detailed signature matches**

no IPR

**Unintegrated signatures**

**Overview**

Similar proteins (2905)

Structures

**Filter view on****Entry type**

-  Family
-  Domains
-  Repeats
-  Site

**Status**

-  Unintegrated

**Colour by**

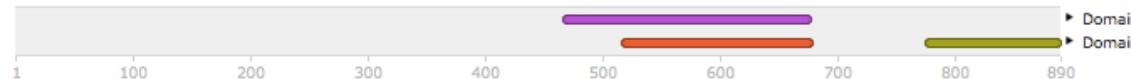
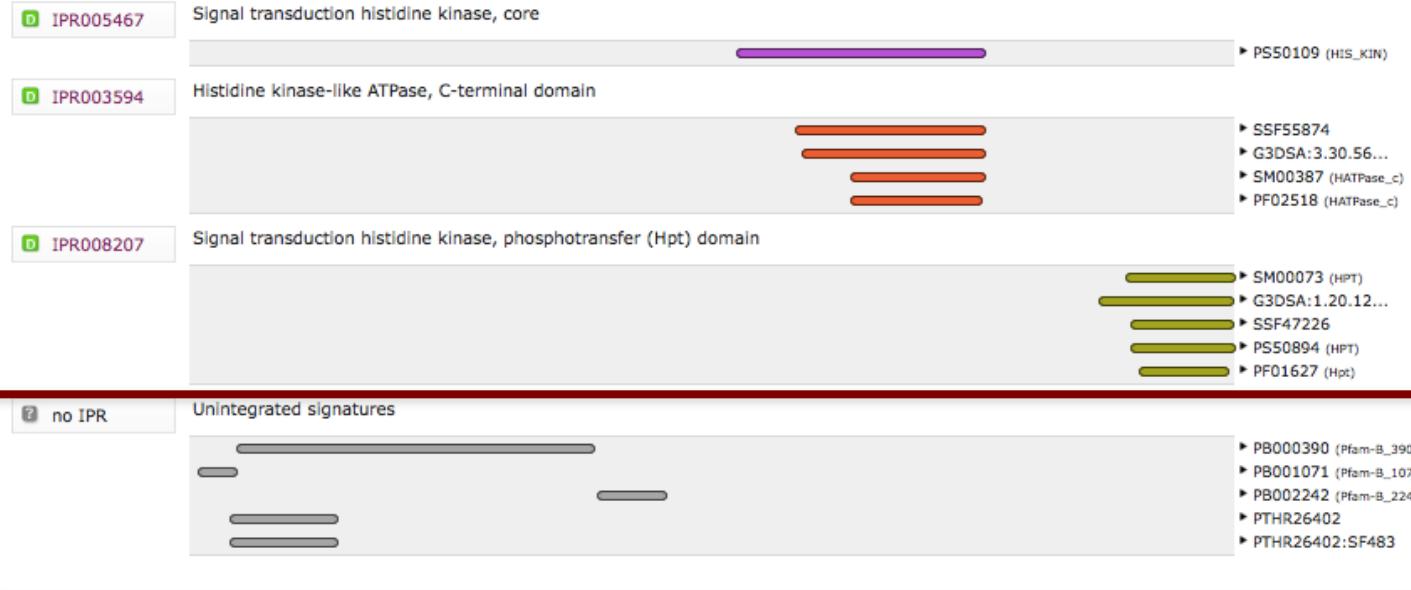
- domain relationship
- source database

[Export FASTA](#)**P 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**

P39838

Examples: IPR020405, kinase, P51587, PF02932, GO:0007165

Search

[Home](#) | [Search](#) | [Release notes](#) | [Download](#) | [About InterPro](#) | [Help](#) | [Contact](#)**Overview**

Proteins matched (402223)

Domain organisations (17400)

Pathways &amp; interactions

Species

Structures

Literature (21)

Cross-references (2)

**D Domain****Signal transduction histidine kinase, core (IPR005467)**Short name: *Sig\_transdc\_His\_kinase\_core***Domain relationships**

None.

 Add your annotation**Contributing signatures**

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

 PROSITE profiles PSS0109 (HIS\_KIN)**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 prokaryotic 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  PMTD: 10966457  PMTD: 11406410  PMTD: 113692791

# Orthologous families, trees



**COGs**  
Phylogenetic classification of proteins encoded in complete genomes



## 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

# Team Exercise

## Building a new Pfam family



## QUICK SEARCH

Paste in your sequence or use the [example](#)

### search against

- Reference Proteomes  UniProtKB  SwissProt  Pfam

[Alternative Search Options](#)

### Blog News

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.

August, 2015

### Download HMMER

Get the latest version

v3.1b2

### Recent

[HMMER web](#)

R.D. Finn, J.

F. Schreiber,

Nucleic Ac

Ac 2015



Home | **Search** | Results | Software | Help | About

phmmmer    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

[Home](#) [Search](#) [Results](#) [Software](#) [Help](#) [About](#)[phmmmer](#)[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](#) ?

```
HEAIGSGDLDLRSAFRRRTSLAGAGRRTSDSHEDAGTLDFSSLLKKRD  
SFRRDSKLEAPAEEDVWEILRQAPPSEYERIAFQHGVTDLRGMLKRL  
KGMKQDEKK
```

[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 ?



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](#) ?

```
HEAIGSGDLDLRSAFRRRTSLAGAGRRTSDSHEDAGTLDFSSLLKKRD  
SFRRDSDKLEAPAEEDVWEILRQAPPSEYERIAFQHGVTDLRGMLKRL  
KGMKQDEKK
```

**Submit**

Reset

**▼ HMM Database** ?

## Protein Families

- Pfam  TIGRFAM  Gene3D  Superfamily  PIRSF  
[\(select all\)](#)



Score Taxonomy Domain Download

## PHMMER Results

[Search Again](#)Sequence Matches and Features [?](#)

Pfam 103

hit coverage  
hit similarity 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 - Q3UIKO (Q3UIKO\_MOUSE)

Protein | Submitted name: **Myosin-binding protein C, cardiac-type**Gene | **Mybpc3**Organism | *Mus musculus (Mouse)*Status |  Unreviewed - Annotation score:  - Experimental evidence at protein level<sup>i</sup>

Display None

 BLAST Align Format Add to basket History Help video Other tutorials and videos FeedbackFunction<sup>i</sup>GO - Molecular function<sup>i</sup>

- [identical protein binding](#)  Source: MGI
- [myosin binding](#)  Source: MGI
- [myosin heavy chain binding](#)  Source: MGI ▾
- [structural constituent of cytoskeleton](#)  Source: MGI ▾

 Function Names & Taxonomy Subcell. location Pathol./Biotech

## Sequence Matches and Features [?](#)

Pfam  103

disorder  coiled-coil  tm & signal peptide [?](#)

No hits were found for your query.

[Advanced](#)

### TIGRFAM Matches

Family		Description	Start ▾	End ▾	Domain E-values				
Id ▾	Accession				Ind. ▾	Cond. ▾			
No hits were found for your query.									
Your search took: 0.05 seconds									

[Advanced](#)

### Gene3D Matches

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									
Accession	Description	E-value	Accession	Description	E-value		Start	End	Alignment		Start	End	Length	Model	Ind	Cond
									Start	End	Start	End	Length	Bit Score		
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.

Search InterPro...

Search

Examples: IPR020405, kinase, P51587, PF02932, GO:0007165

Home | Search | Release notes | Download | About InterPro | Help | Contact

**Overview**

Similar proteins

Structures

**Filter view on****Entry type**  Family  Domains  Repeats  Site**Status**  Unintegrated**Colour by**[help](#) domain relationship source database**P 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

► PTHR19900 (MYOSIN-B...)

► PTHR19900:SF52 (MYO...)

**GO term prediction**

## Biological Process

None predicted.

## UniProtKB - Q3UIKO (Q3UIKO\_MOUSE)

Protein | Submitted name: **Myosin-binding protein C, cardiac-type**Gene | **Mybpc3**Organism | *Mus musculus (Mouse)*Status |  Unreviewed - Annotation score:  - Experimental evidence at protein level<sup>i</sup>

Display None

 BLAST Align Format Add to basket History Help video Other tutorials and videos FeedbackFunction<sup>i</sup>GO - Molecular function<sup>i</sup>

- [identical protein binding](#)  Source: MGI
- [myosin binding](#)  Source: MGI
- [myosin heavy chain binding](#)  Source: MGI ▾
- [structural constituent of cytoskeleton](#)  Source: MGI ▾

 Function Names & Taxonomy Subcell. location Pathol./Biotech

[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](#)

V	Target						Species				E-value	
	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 .....\*.....\*.....\*.....\*.....\*.....\*.....\*.....\*.....\*  
heraigsgdldlrslarrrtslagagrrtsdshedagtlpdfssllkkrrdsfrrdskleapaaedvweilrqappseyeriaf 80  
heraigsgdldlrslarrrtslagagrrtsdshedagtlpdfssllkkrrdsfrrdskleapaaedvweilrqappseyeriaf  
Target 263 HEAIGSGDLDLRLSARFRRTSLAGAGRRTSDSHEDAGTLDFSSLKKRDRSFRRDSKLEAPAAEDVWEILRQAPPSEYERIAF 342  
PP 9\*\*\*\*\*  
9\*\*\*\*\*

Query 81 qhgvtndlrgmlkr1kgmkqdekk 103  
qhgvtndlrgmlkr1kgmkqdekk  
Target 343 QHGVTDLRGMLKRLKGKQDEKK 365  
PP \*\*\*\*\*98



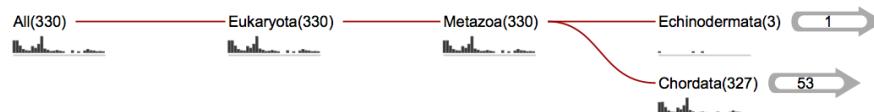
Home Search Results Software Help About

Score Taxonomy Domain Download

## PHMMER Results

Search

## 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

Jump to the exact match for your query architecture

## Domain Architectures

<< First << Previous Page  of 294  
SEQUENCESwith domain architecture: **I-set, I-set, I-set, I-set, I-set, fn3, fn3, I-set, fn3, I-set**, example: Q3TF37\_MOUSE 

Sequence Features

[Show All](#)35  
SEQUENCESwith domain architecture: **I-set, I-set, I-set, I-set, fn3, fn3, I-set, fn3, I-set**, example: F6ZHP7\_HORSE 

Sequence Features

[Show All](#)34  
SEQUENCESwith domain architecture: **I-set, I-set, I-set, I-set, I-set, fn3, fn3, I-set, fn3, I-set**, example: Q3UIK0\_MOUSE 

Sequence Features

[Show All](#)19  
SEQUENCESwith domain architecture: **I-set, I-set, I-set, fn3, fn3, I-set, fn3, I-set**, example: G1Q885\_MYOLU 

Sequence Features

[Show All](#)12  
SEQUENCESwith domain architecture: **I-set, I-set, I-set, I-set, fn3, fn3, fn3, I-set, fn3, I-set**, example: W5MUP3\_LEPOC 

Sequence Features

[Show All](#)

10

with domain architecture: **I-set, I-set, I-set**, example: F7CWG3\_CAI\_1A



## PHMMER Results

- **Job:** 5B650320-65F7-11E5-8E90-C8CCF69F8818.1
- **Started:** 2015-09-28 16:41:18
- **Algorithm:** phmmmer
- **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.

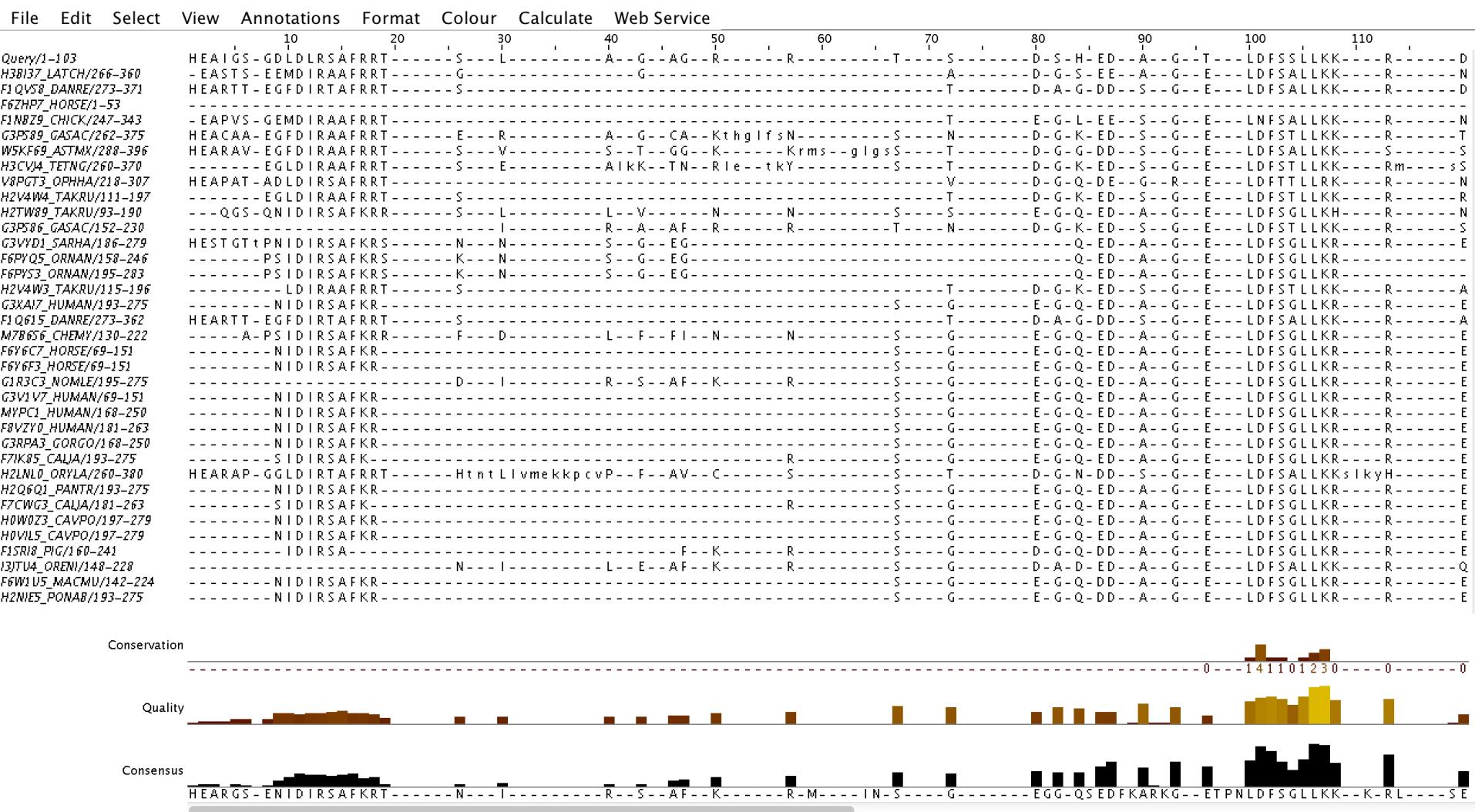


#### STOCKHOLM

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



# 1. File -> Input alignment -> From file

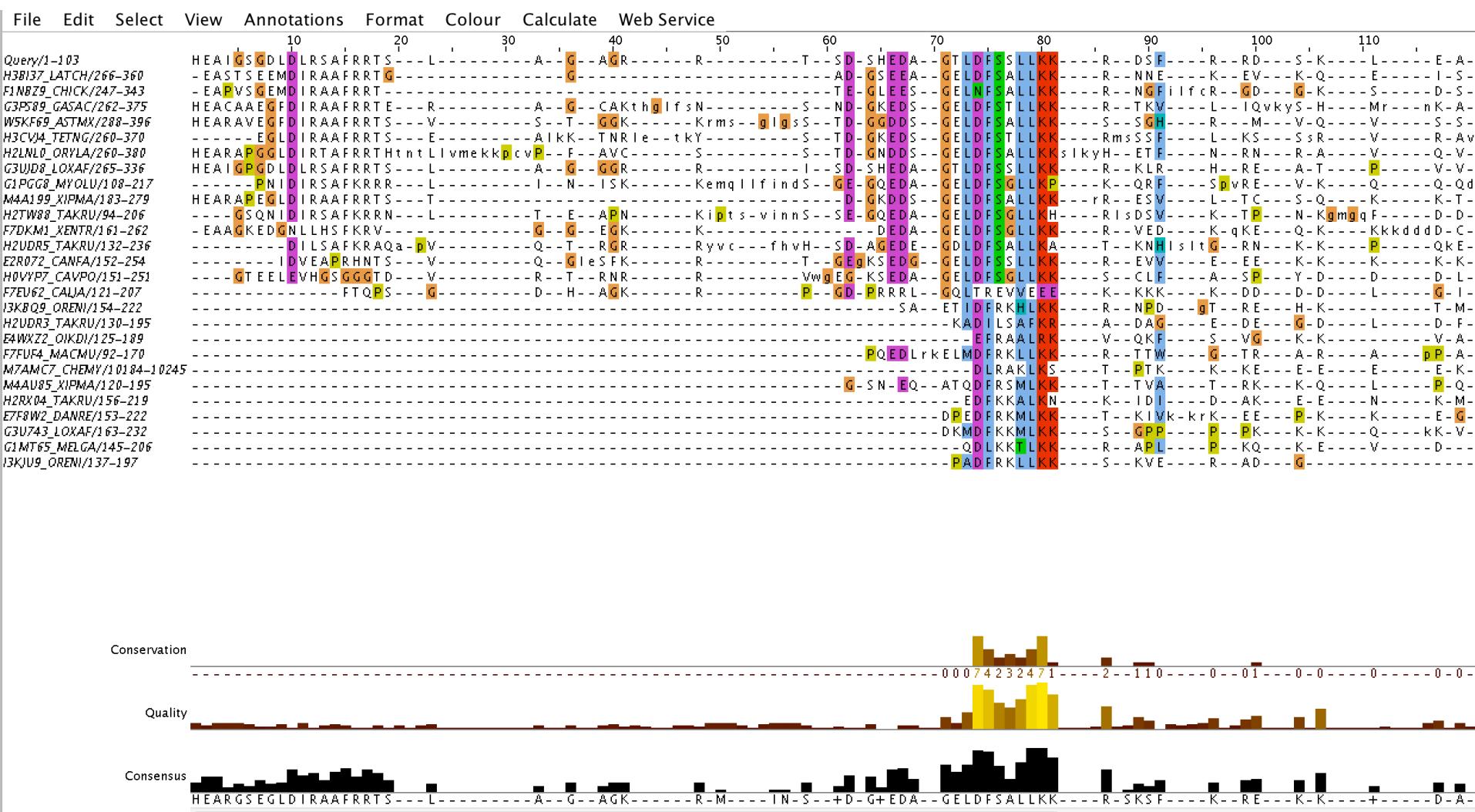


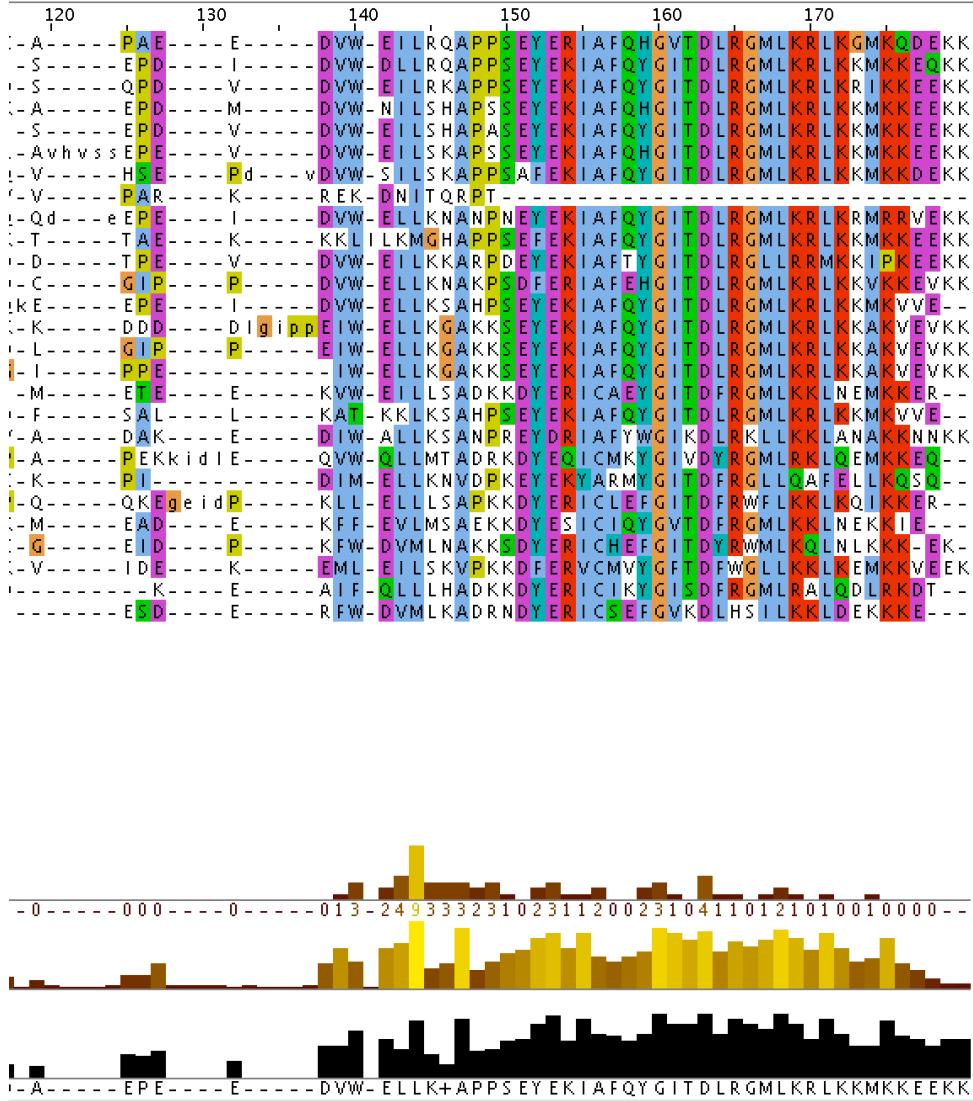
1. Edit -> Remove redundancy

2. Select 90% and Remove

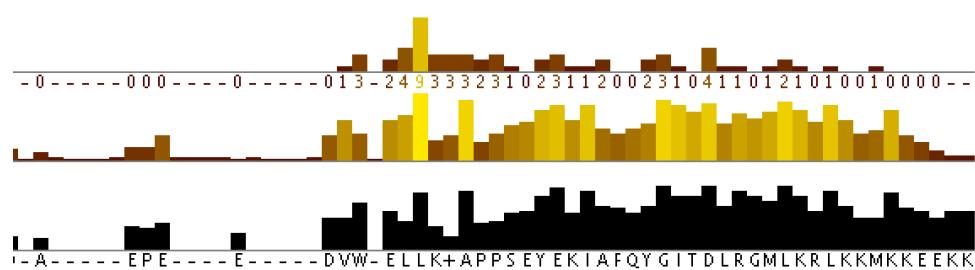
1. Edit -> Remove redundancy
2. Select 90% and Remove
3. Edit -> Remove empty columns

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

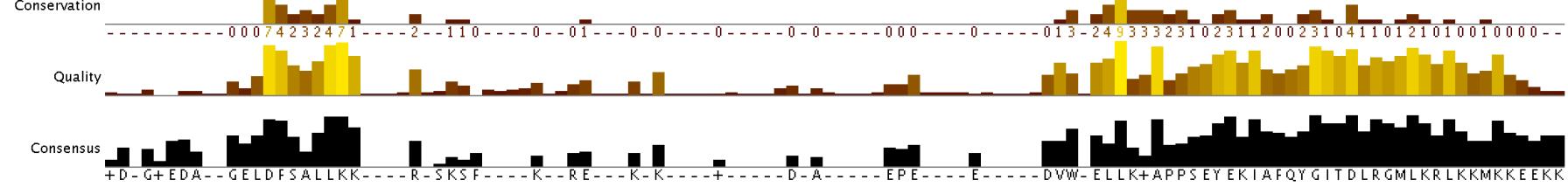




120            130            140            150            160            170  
 - A - P A E - E - D V W - E I L R Q A P P S E Y E K I A F Q H G V T D L R G M L K R L K G M K Q D E K K  
 - S - E P D - I - D V W - D L L R Q A P P S E Y E K I A F Q Y G I T D L R G M L K R L K K M K K E Q K K  
 - S - Q P D - V - D V W - E I L R K A P P S E Y E K I A F Q Y G I T D L R G M L K R L K R I K K E E K K  
 - A - E P D - M - D V W - N I L S H A P S S E Y E K I A F Q H G I T D L R G M L K R L K K M K K E E K K  
 - S - E P D - V - D V W - E I L S H A P A S S E Y E K I A F Q Y G I T D L R G M L K R L K K M K K E E K K  
 - A v h v s s E P E - V - D V W - E I L S K A P S S E Y E K I A F Q H G I T D L R G M L K R L K K M K K E E K K  
 - V - H E E D - D - D V W - C I L S K A P P S A F E K I A F Q Y G I T D I D P C M I L K D I L K K M K K E E K K  
 - V - P A R - K - R E K - D N I T Q R P T -  
 - Q d - e E P E - I - D V W - E L L K N A N P N E T E N I A F Q Y G I T D L R G M L K R L K K M K K V E K K  
 - T - T A E - K - K K L I L K M G H A P P S S E F E K I A F Q Y G I T D L R G M L K R L K K M K K E E K K  
 - D - T P E - V - D V W - E I L K K A R P D E Y E K I A F T Y G I T D L R G L L R R M K K I P K E E K K  
 - C - G I P - P - D V W - E L L K N A K P S D F E R I A F E H G I T D L R G M L K R L K K V K K E V K K  
 - K E - E P E - I - D V W - E L L K S A H P S S E Y E K I A F Q Y G I T D L R G M L K R L K K M K V V E -  
 - K - D D D - D i g i p p E I W - E L L K G A K K S S E Y E K I A F Q Y G I T D L R G M L K R L K K A K V E V K K  
 - L - G I P - P - E I W - E L L K G A K K S S E Y E K I A F Q Y G I T D L R G M L K R L K K A K V E V K K  
 - I - P P E - I W - E L L K G A K K S S E Y E K I A F Q Y G I T D L R G M L K R L K K A K V E V K K  
 - M - E T E - E - K V W - E I L S A D K D Y E R I C A E Y G I T D P R G M L K K L N E M K K E R -  
 - F - S A L - L - K A T - K K L K S A H P S S E Y E K I A F Q Y G I T D L R G M L K R L K K M K V V E -  
 - A - D A K - E - D I W - A I L K S A N P R E Y D R I A F Y W G I K D L R K L L K K L A N A K K N N K K  
 - A - P E K K i d I E - Q V W - Q L L M T A D R K D Y E Q I C M K Y G I V D Y R G M L R K L Q E M K K E D -  
 - K - P I - D I M - E L L K N V D P K E Y E K Y A R M Y G I T D P R G L L Q A F E L L K Q S Q -  
 - Q - Q K E g e i d P - K L L - E L L L S A P K K D Y E R I C I L E F G I T D P R W F L K K L K Q I K K E R -  
 - M - E A D - E - K F F - E V L M S A E K K D Y E S I C I O Y G V T D P R G M L K K L N E K K I E -  
 - G - E I D - P - K F W - D V M L N A K K S D Y E R I C I H E F G I T D Y R W M L K Q L N L K K K - E K -  
 - V - I D E - K - E M L - E I L S K V P K K D F E R V C M V Y G F T D P W G L L K K L K E M M K V E K K  
 - K - A I F - Q L L H A D K K D Y E R I C I K Y G I S D P R G M L R A L Q O D L R K D T -  
 - E S D - E - R F W - D V M L K A D R N D Y E R I C S E F G V K D L H S I L K K L D E K K K E -



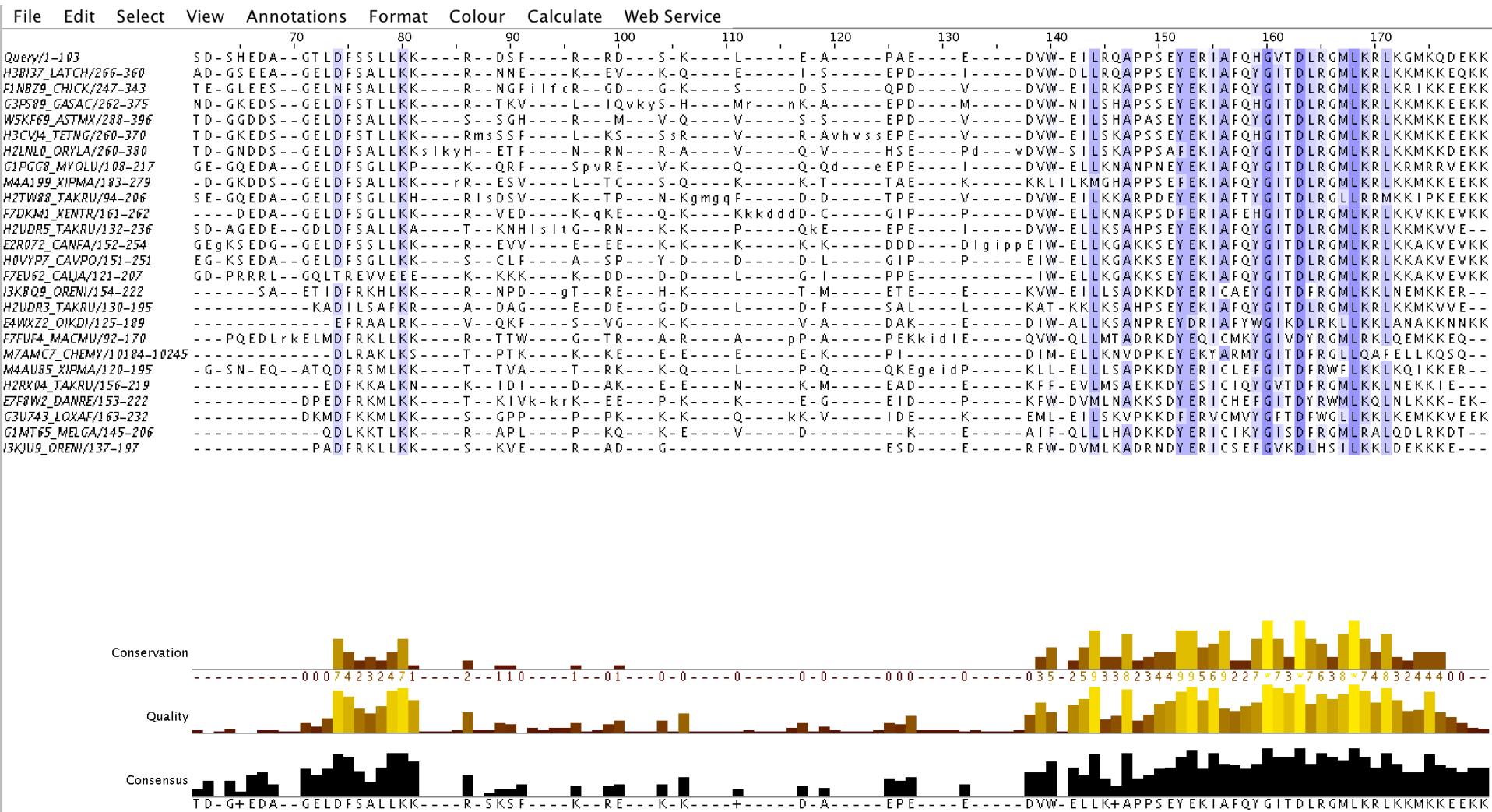
File	Edit	Select	View	Annotations	Format	Colour	Calculate	Web Service	70	80	90	100	110	120	130	140	150	160	170
Query/1-103									S-D-SHEDA-GTLDLSSLLKK--R--DS-E--R-RD--S-K-L-E-A-PAE-E--Dvw-EILRQAPPSEYERIAFQHGVITDLRGMLKRLKGMKKEDEKK										
H3B137_LATCH/266-360									AD-GSEEAA-GELDESALLKK--R-NNE-K-EV--K-Q-E-I-S-EPD-E--Dvw-DLLRQAPPSEYEKIAFQYGITDLRGMLKRLKGMKKEEKK										
F1NB79_CHICK/247-343									TE-GLEES-GELDNLSALLKK--R-NGEifcR-GD-G-K-S-D-S-QPD-V--Dvw-EILRKAPPSEYEKIAFQYGITDLRGMLKRLKGMKKEEKK										
G3PS89_GASAC/262-375									ND-GKEDS-GELDESTLLKK--R-TKV-L-IQvkyS-H-Mr-nK-A-EPD-M--Dvw-NILSHAPSSEYEKIAFQHGVITDLRGMLKRLKGMKKEEKK										
W5KF69_ASTMX/288-396									TD-GGDDDS-GELDESTLLKK--S-SGH-R-M-V-Q-V-S-S-EPD-V--Dvw-EILSHAPSEYEKIAFQYGITDLRGMLKRLKGMKKEEKK										
H3CWJ4_TETNG/260-370									TD-GKEDS-GELDESTLLKK--RmSSS-L-KS-SsR-V-R-AvhvssEPE-V--Dvw-EILSKAPPSEYEKIAFQHGVITDLRGMLKRLKGMKKEEKK										
H2LNLD_ORYLA/260-380									TD-GNDDDS-GELDESTLLKKsIkyH-ETP-N-RN-R-A-V-H-E-Pd-vDvw-SILSKAPPSEYEKIAFQYGVITDLRGMLKRLKGMKKEEKK										
GRUW88_LOKAF/265-336									SD-SHEDA-GTLDLSSLLKK--R-KLR-H-RF-A-T-P-Y-W-PAR-K-R-FEK-DNTQRP--K--Dvw-EILRKAPPSEYEKIAFQYGVITDLRGMLKRLKGMKKEEKK										
G1PGG8_MYOLU/108-217									GE-GQEDA-GELDESGLLKP--K-QR-SpvRE-V-K-Q-Q-Qd-eEPE--I-Dvw-ELLKNANPNEYEKIAFQYGVITDLRGMLKRLKRMRRVEKK										
M4A199_XIPMA/183-279									-D-GKDDs-GELDESALLKK--rR-ESV-L-TC-S-Q-K-K-T-TAE-K-KLILKMGHAPPSEFEKIAFQYGVITDLRGMLKRLKGMKKEEKK										
H2TW88_TAKRU/94-206									SE-GQEDA-GELDESGLLKH--R-isDSV-K-TP-N-Kmgqf-D-D-TPE-V-Dvw-EILKARPDEYEKIAFTYGVITDLRGMLLRRMNNKKTPKEEKK										
F2DKM1_XENTR/161-262									--DEDA-GELDESGLLKH--R-VED-K-qKE-Q-K-KkkdddD-C-GIP-P-Dvw-ELLKNAKPSDFERIAFEHGIFTDLRGMLKRLKVKVEVKK										
H2UDR5_TAKRU/132-236									SD-AGEDE-GDLDESALLKA--T-KNHisItG-RN-K-K-P-QKE-EPE--I-Dvw-ELLKSAHPSEYEKIAFQYGVITDLRGMLKRLKGMKVV--										
E2R072_CANFA/152-254									GEOKS-EDG-GELDESGLLKK--R-EVY-E-E-K-K-K-K-K-K-DDD-D10IPPEI-ELLKGAKKSSEYEKIAFQYGVITDLRGMLKRLKAKVEVKK										
H0VVP7_CAVPO/151-251									EG-KS-EDA-GELDESGLLKK--S-CLF-A-SP-Y-D-D-L-GIP-P-Dvw-ELLKGAKKSSEYEKIAFQYGVITDLRGMLKRLKAKVEVKK										
F7EV62_CALJA/121-207									GD-PRRRRL-GQLTREVVEEE--K-KKK-K-DD-D-D-L-G-I-PPE-IW-ELLKGAKKSSEYEKIAFQYGVITDLRGMLKRLKAKVEVKK										
I3KB09_ORENI/154-222									--SA-ETIDFRKHLKK--R-NP-DOT-R-E-H-K-T-M-E-E-E-KWV-ELLLSADKKDYERICAEYGVITDFRGMLKKLNEMKKER--										
H2UDR3_TAKRU/130-195									--KADILSAFKR--A-DAG-E-DE-G-D-L-D-F-SAL-L-KAT-KKLKSAHPSYEKIAFQYGVITDLRGMLKRLKGMKVV--										
E4WXZ2_OIKDI/125-189									--E-FRAALRK--V-QK-P-S-VG-K-K-V-A-DAK-E-DIW-ALLKSANPREYDRIAFYWGVITDLRKLLKKLANAKKNNKK										
E7FU4_MACMU/92-170									+PQFDLrKELMDFRKLLLKK--R-TTW-G-TR-A-R-A-PP-A-PEKKidIE-QVW-QLLMTADRKYDQEICMKYGVITDVRGMLRKLQEMKKFQ--										
M7AMC7_CHEMY/10184-10245									--DLRAKIKS--T-PTK-K-KE-E-E-E-E-K-P-DIM-ELLKNVDPKEYEKIARMIGITDFRGLLOAFELLKOSQ--										
M4AU85_XIPMA/120-195									-G-SN-EQ-ATQDFRSMLKK--T-TV-A-T-RK-K-Q-L-P-Q-QKFeidP-KLL-ELLLSAPKKDYERICCLEFGITDFRFWFLKKLNQIKKER--										
H2RX04_TAKRU/156-219									--EDFKKALKN--K-ID-D-AK-E-E-N-K-M-EAD-E-KFFF-EVLMSAEKKDYESIC1QYGVITDFRGMLKKLNEMKKER--										
E7FW2_DANRE/153-222									--DPFLERKMLKK--T-KIVK-kRK-EE-P-K-K-E-G-EID-P-KFW-DVMLNAKKSDYERICHEFFGITYDWRWMLKQFLNLKQQ-EK-										
G3U743_LOKAF/163-232									--DKMDKFKMLKK--S-GPP-P-P-K-K-Q-KK-V-ID-E-EMI-E-LLSKVPKKDDEVRCMVYGVITDFWGLLKKBLKEMKKVVEEK-										
G1MT65_MELGA/145-206									--QDLKK-LKK--R-APL-P-KQ-K-E-V-D-K-E-AIF-QLLHADKKDYERICIKYGVISDFRGMLRALD-LRKDT--										
I3KJU9_ORENI/137-197									--PADERKLLKK--S-KWE-R-AD-G-EID-E-RFW-DVMLKADRNDYERICSSEFGVKDLHSILKKLDKKKE--										



1. Edit -> Delete

2. Colour -> BLOSUM62 Score

Colour -> Percentage Identity

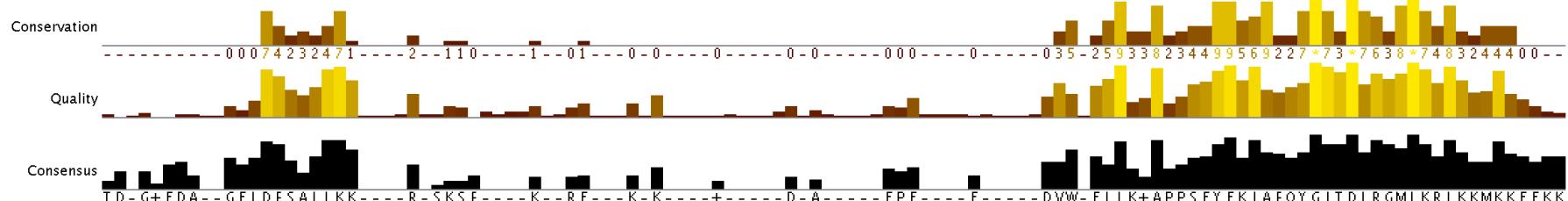


File Edit Select View Annotations Format Colour Calculate Web Service

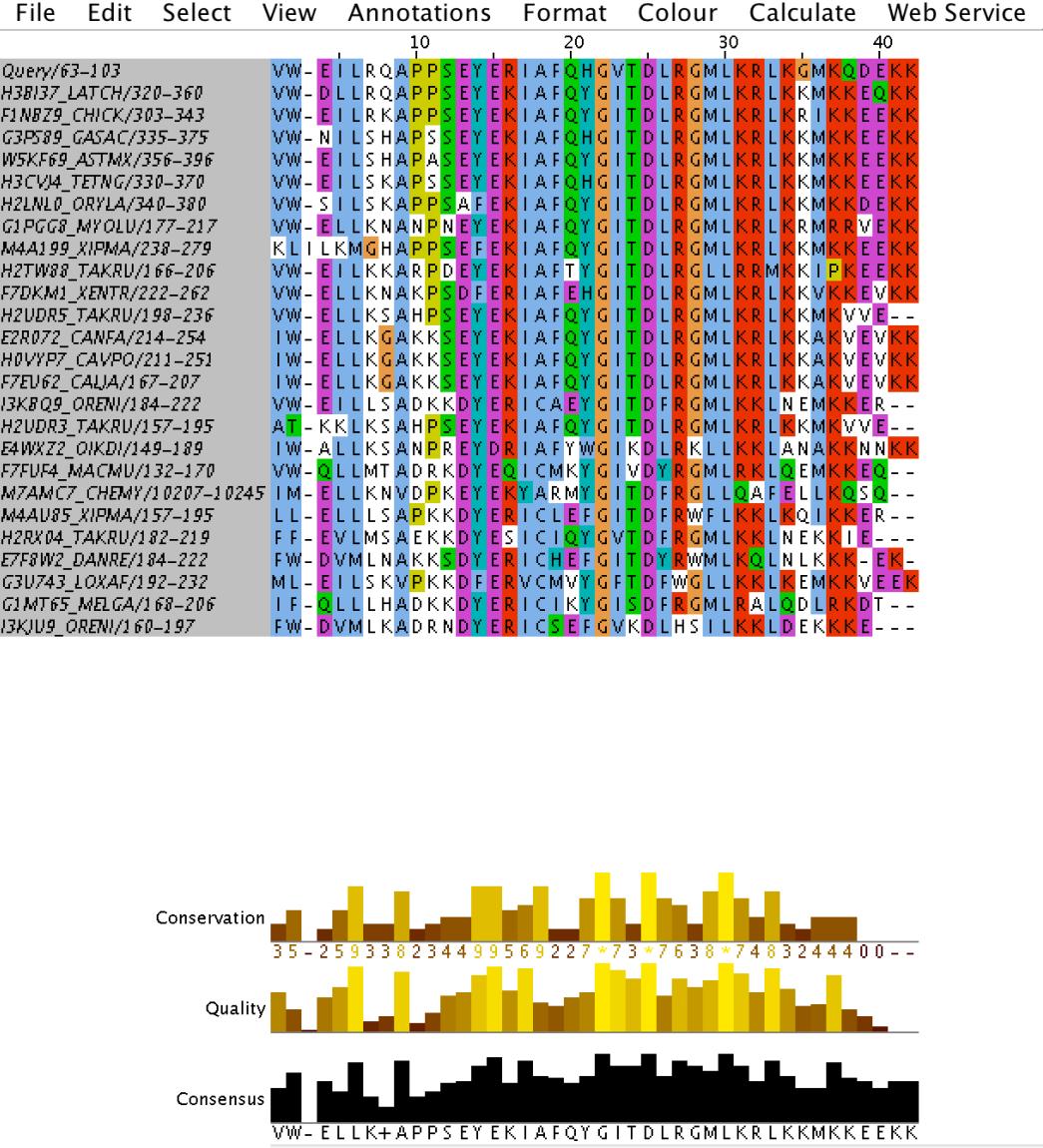
Query/1-103  
 F3NB137\_LATCH/266-360  
 F3NB29\_CHICK/247-343  
 G3PS89\_GASAC/262-375  
 W5KF69\_ASTMG/288-396  
 H3CV14\_TETNG/260-370  
 H2LNLD\_ORYLA/260-380  
 G1PGG8\_MYOLU/108-217  
 MAA199\_XIPMA/183-279  
 H27W88\_TAKRU/94-206  
 F7DKM1\_XENTR/161-262  
 H2UDR3\_TAKRU/132-236  
 E2R072\_CANFA/152-254  
 HWVYPT\_CAVPO/151-251  
 F7EU62\_CALJA/121-207  
 J3KB09\_ORENI/154-222  
 H2UDR3\_TAKRU/130-195  
 E4WXZ2\_OIKDI/125-189  
 F7FU4\_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  
 J3KJU9\_OREM/137-197

70 80 90 100 110 120 130 140 150 160 170

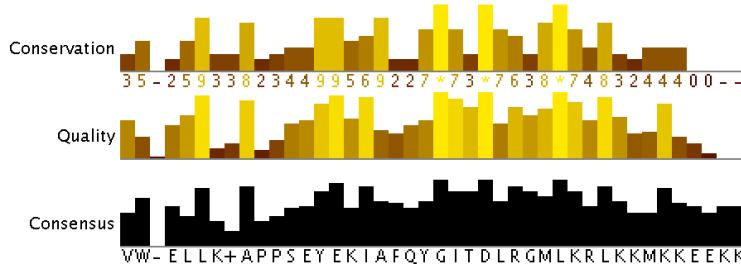
SD-SHEDA-GTLDLDSLKK-R-DSF-R-RD-S-K-L-E-A-PAE-E-DWW-EILRQAPPSEYERIAFQHGVTDLRGMLKRLLKGMKODEKK  
 AD-GSEEA-GELDFSGALLKK-R-NNE-K-EV--K-Q-E-I-S-EPD-I-DWW-DLLRQAPPSEYEKIAFQYGITDLRGMLKRLLKMKKEOKK  
 TE-GLEES-GELDFSGALLKK-R-NGFiIfcR-G-K-S-D-S-QPD-V-DWW-EILRKAPPSEYEKIAFQYGITDLRGMLKRLLKMKKEKK  
 ND-GKEDS-GELDFSTLLKK-R-TKM-L-IQvkyS-H-Mr-nK-A-EPD-M-DWW-NILSHAPSSEYEKIAFQHGITDLRGMLKRLLKMKKEKK  
 TD-GGDDS-GELDFSGALLKK-S-SGH-R-M-V-Q-V-S-S-EPD-V-DWW-EILSHAPASEYEKIAFQYGITDLRGMLKRLLKMKKEKK  
 TD-GKEDS-GELDFSGALLKK-RmSSF-L-KS-SsR-V-R-AvhvssEPE-V-DWW-EILSKAPSSEYEKIAFQHGITDLRGMLKRLLKMKKEKK  
 TD-GNDDs-GELDFSGALLKKsIkyH-ETF-N-RN-R-A-V-Q-V-HSE-Pd-v-DWW-SILSKAPSSEYEKIAFQYGITDLRGMLKRLLKMKKEDEKK  
 GE-GQEDA-GELDFSGLLKK-K-QRF-SpVRE-V-K-Q-Q-Qd-eEPE-I-DWW-ELLKNAKPSEFERIAFEGITDLRGMLKRLLKMKKEVKK  
 -D-GKDDs-GELDFSGALLKK-rR-ESV-L-TC-S-Q-K-T-TAE-K-KLILKGHAPPSEFERIAFQYGITDLRGMLKRLLKMKKEKK  
 SE-GQEDA-GELDFSGLLKK-R-isDSW-K-TP-N-KmOqf-D-D-TPE-V-DWW-EILKKARPDFEYKIAFTYGITDLRLGLLRRMKKIPKFEEKK  
 ---DEDA-GELDFSGLLKK-R-VED-K-qKE-Q-K-KkdddD-C-GIP-P-DWW-ELLKNAPSDFERIAFEGITDLRGMLKRLLKMKKEVKK  
 SD-AGEDE-GELDFSGALLKK-T-KNHSITG-RN-K-K-P-QKE-EPE-I-DWW-ELLKSAHPSEYEKIAFQYGITDLRGMLKRLLKMKVVVEKK  
 GE-GKSEDG-GELDFSGLLKK-R-EVW-E-EE-K-K-K-K-DDD-DIGippEW-ELLKGAKKSEYEKIAFQYGITDLRGMLKRLLKAKVVEKK  
 EG-KSEDA-GELDFSGLLKK-S-CLF-A-SP-Y-D-D-L-GIP-P-EIW-ELLKGAKKSEYEKIAFQYGITDLRGMLKRLLKAKVVEKK  
 GD-PRRL-GQLTREVMEEE-K-KKK-K-DD-D-D-L-G-I-PPE-IW-ELLKGAKKSEYEKIAFQYGITDLRGMLKRLLKAKVVEKK  
 -SA-ETIDFRKHKK-R-NPD-GT-R-E-H-K-T-M-E-E-KWW-EILSLADKKDYERICAEGITDLRGMLKKLNEMKKER-  
 -KADILSAFKR-A-DAG-E-DE-G-D-L-D-F-SAL-L-KAT-KKLKSAPHSSEYEKIAFQYGITDLRGMLKKLNEMKKER-  
 -EFRAALRK-V-QKF-S-VG-K-K-V-A-DAK-E-DIM-ALLKSANPREYDRIIAFYWGIKDLRKLLKKLANAKNNKK  
 PQEDLykELMDFRKLKK-R-TTW-G-TR-A-R-A-dP-A-PEKKidIE-QWW-0LLMTADRKYEQCMKIGIVDVRGMLRKUDDEMKEO-  
 -DLRAKLS-T-PTK-K-KE-E-E-E-K-Pi-DIM-ELLKNWDPKKEYEKYARMYGITDFRGLLCAFEELLKSO-  
 -G-SN-EQ-ATQDFRSMLKK-T-TVAA-T-RK-K-Q-L-P-Q-QKEgeidP-KLL-ELLSLAPKKDYERICICLEFGITDFRWFLKKQJIKER-  
 -EDFKKALRN-K-IDI-D-AK-E-E-N-K-M-EAD-E-KFF-EVLMSAEKKDYESICIOYGVTFDRGMLKKLNEMKKIE-  
 -DPEDFRKMLKK-T-KIMk-krK-EE-P-K-K-E-G-EID-P-KFW-DVMLNAKKDYERICHEFGITDWRMLKOLNLKK-EK-  
 -DKMDFKKMLKK-S-GPP-P-PK-K-K-Q-kK-V-IDE-K-EML-EILSKWPKKDFERVCMVYGFDFWGLLKKLEMKKVEEK-  
 -QDJKKKLKK-R-APL-P-KQ-K-E-V-D-K-E-AIF-LLLHADKKDYERICIKYGISDFRGMLRALDLRKDT-  
 -PADFRKLLKK-S-KVE-R-AD-G-ECD-E-RFW-DVMLKAQRNDYERICSEFGVKDLHSILKKLDEKKKE



1. Edit -> Remove left

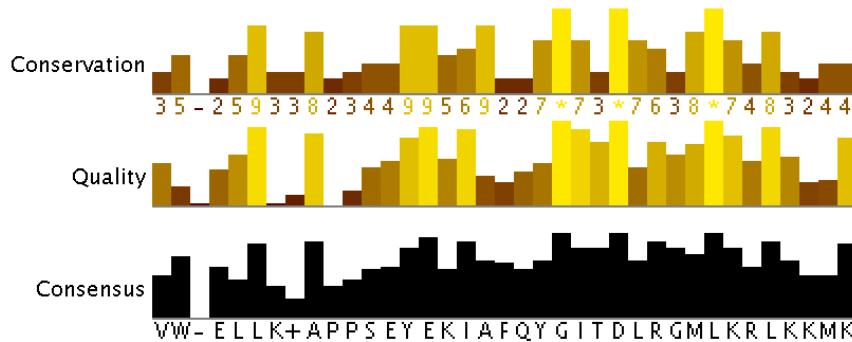


	10	20	30	40
Query/63-103	VW-	E I L R Q A P P S E Y E K I A F Q H G V T D L R G M L K R L K G M K Q D E K K		
H3B137_LATCH/320-360	VW-	D L L R Q A P P S E Y E K I A F Q Y G I T D L R G M L K R L K K M K K E E K K		
F1NBZ9_CHICK/303-343	VW-	E I L R K A P P S E Y E K I A F Q Y G I T D L R G M L K R L K R I K K E E K K		
G3PS89_GASAC/335-375	VW-	N I L S H A P P S E Y E K I A F Q H G I T D L R G M L K R L K K M K K E E K K		
WSKF69_ASTMX/356-396	VW-	E I L S H A P A S E Y E K I A F Q Y G I T D L R G M L K R L K K M K K E E K K		
H3CV44_TETNG/330-370	VW-	E I L S K A P P S E Y E K I A F Q H G I T D L R G M L K R L K K M K K E E K K		
H2LN02_ORYLA/340-380	VW-	S I L S K A P P S A F E K I A F Q Y G I T D L R G M L K R L K K M K K D E K K		
G1PGG8_MYOUL/177-217	VW-	E L L K N A N P N E Y E K I A F Q Y G I T D L R G M L K R L K R M R R V E K K		
M4A199_XIPMA/238-279	KL I L K M G H A P P S E F E K I A F Q Y G I T D L R G M L K R L K K M K K E E K K			
H2TW88_TAKRU/166-206	VW-	E I L K K A R P D E Y E K I A F T Y G I T D L R G L L R R M K K I P K E E K K		
F7DKM1_XENTR/222-262	VW-	E L L K N A K P S D F E R I A F E H G I T D L R G M L K R L K K V K K E V K K		
H2UDR5_TAKRU/198-236	VW-	E L L K S A H P S E Y E K I A F Q Y G I T D L R G M L K R L K K M K V V E --		
E2R072_CANFA/214-254	IW-	E L L K G A K K S E Y E K I A F Q Y G I T D L R G M L K R L K K A K V E V K K		
H0VYP7_CAVPO/211-251	IW-	E L L K G A K K S E Y E K I A F Q Y G I T D L R G M L K R L K K A K V E V K K		
F7EU62_CALJA/167-207	IW-	E L L K G A K K S E Y E K I A F Q Y G I T D L R G M L K R L K K A K V E V K K		
I3KBQ9_OREN/184-222	VW-	E I L S A D K K D Y E R I C A E Y G I T D F R G M L K K L N E M K K E R --		
H2UDR3_TAKRU/157-195	AT-	K K L K S A H P S E Y E K I A F Q Y G I T D L R G M L K R L K K M K V V E --		
E4WXZ2_OIKDI/149-189	IW-	A L L K S A N P R E Y D R I A F Y W G I K D L R K L L K K L A N A K R N N K K		
F7FU44_MACMV/132-170	VW-	Q L L M T A D R K D Y E Q I C M K Y G I V D Y F R G M L R K L Q E M K K E O --		
M7AMC7_CHEMY/10207-10245	I-	E L L K N V D P K E Y E K Y A R M Y G I T D F R G L L Q A F E L L K Q S O --		
M4AU85_XIPMA/157-195	LL-	E L L L S A P K K D Y E R I C L E F F G I T D F R W F L K K L K Q I K K E R --		
H2RX04_TAKRU/182-219	FF-	E V L M S A E K K D Y E S I C I Q Y G V T D F R G M L K K L N E K K I E --		
E7F8W2_DANRE/184-223	FW-	D V M L N A K K S D Y E R I C I C H E F F G I T D Y R W M L K Q L N L K K K - E K -		
G3U743_LOXAF/192-232	ML-	E I L S K V P K K D F E R V C M V Y G F T D F W G L L K K L K E M K K V E E K		
G1MT65_MELGA/168-206	I F-	Q L L L H A D K K D Y E R I C I C K Y G I S D F R G M L R A L Q D L R K D T --		
I3KJU9_OREN/160-197	FW-	D V M L K A D R N D Y E R I C I C S E F G W K D L H S I L K K L D E K K K E --		

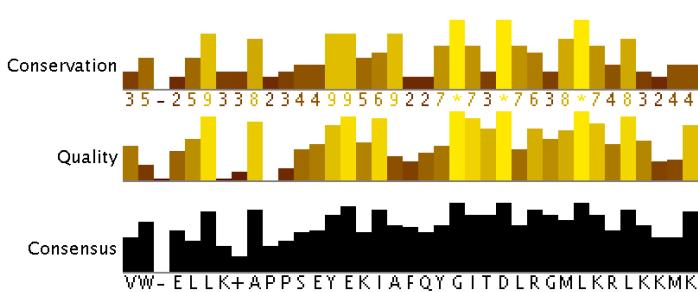


1. Edit -> Remove right

	10	20	30
Query/63-98	VW-EILRQAPPSEYERIAFQHGVTDLRGMLKRLKKMK		
H3B137_LATCH/320-355	VW-DLLRQAPPSEYEKIAFQYGITDLRGMLKRLKKMK		
F1NBZ9_CHICK/303-338	VW-EILRKAPPSEYEKIAFQYGITDLRGMLKRLKKMK		
G3PS89_GASAC/335-370	VW-NILSHAPSSEYEKIAFQHGVTDLRGMLKRLKKMK		
W5KF69_ASTMX/356-391	VW-EILSHAPASEYEKIAFQYGITDLRGMLKRLKKMK		
H3CVJ4_TETNG/330-365	VW-EILSKAPSSEYEKIAFQHGVTDLRGMLKRLKKMK		
H2LNLD_ORYLA/340-375	VW-SILSKAPPSAFEKIAFQYGITDLRGMLKRLKKMK		
G1PGG8_MYOLU/177-212	VW-ELLKNANPNEYEKIAFQYGITDLRGMLKRLKRM		
M4A199_XIPMA/238-274	KLILKMGHAPPSEFEKIAFQYGITDLRGMLKRLKKMK		
H2TW88_TAKRU/166-201	VW-EILKKARPDEYEKIAFTYGITDLRGLLRRMKKIP		
F7DKM1_XENTR/222-257	VW-ELLKNAKPSDFERIAFEHGITDLRGMLKRLKKVK		
H2UDR5_TAKRU/198-233	VW-ELLKSAHPSEYEKIAFQYGITDLRGMLKRLKKMK		
E2R072_CANFA/214-249	IW-ELLKGAKKSEYEKIAFQYGITDLRGMLKRLKKAK		
H0VYP7_CAVPO/211-246	IW-ELLKGAKKSEYEKIAFQYGITDLRGMLKRLKKAK		
F7EUV2_CALJA/167-202	IW-ELLKGAKKSEYEKIAFQYGITDLRGMLKRLKKAK		
I3KBQ9_ORENI/184-219	VW-EILLSADKKDYERICALEYGITDFRGMLKKLNEMK		
H2UDR3_TAKRU/157-192	AT-KKLKSAHPSEYEKIAFQYGITDLRGMLKRLKKMK		
E4WXZ2_OIKDI/149-184	IW-ALLKSANPREYDRIAFWGKIDLRKLKKLANAK		
F7FUF4_MACMU/132-167	VW-QLLMTADRKDYEQICMKYGINVDYRGMLRKLOEMK		
M7AMC7_CHEMY/10207-10242	IM-ELLKNVDPKEYEKYARMYGITDFRGGLLOAFELLK		
M4AU85_XIPMA/157-192	LL-ELLLSAPKKDYERICLLEFGITDFRWFLKKLKQIK		
H2RX04_TAKRU/182-217	FF-EVLMSAEKKDYESICCIQYGVTDFRGMLKKLNEK		
E7F8W2_DANRE/184-219	FW-DVMLNAKKSDYERICHEFGITDYRWMLKQLNLKK		
G3U743_LOXAF/192-227	ML-EILSKVPKKDFERVCMVYGFITDFFWGLLKKLKEMK		
G1MT65_MELGA/168-203	IF-QLLLHADKKDYERICIKYGISDFRGMLRALQDLR		
I3KJU9_ORENI/160-195	FW-DVMLKADRNDYERICSEFGVMKDLHSILKKLDEKK		



		10	20	30	
Query/63-98	VW-E	L R Q A P P S E Y E K I A F Q H G V T D L R G M L K R L K K M K			
H3B137_LATCH/320-355	VW-D	L R Q A P P S E Y E K I A F Q Y G I T D L R G M L K R L K K M K			
F1NB29_CHICK/303-338	VW-E	I L R K A P P S E Y E K I A F Q Y G I T D L R G M L K R L K K M K			
G3PS89_GASAC/335-370	VW-N	I L S H A P S S E Y E K I A F Q H G I T D L R G M L K R L K K M K			
W5KF69_ASTMX/356-391	VW-E	I L S H A P S S E Y E K I A F Q Y G I T D L R G M L K R L K K M K			
H3CVJ4_TETNG/330-365	VW-E	I L S K A P S S E Y E K I A F Q H G I T D L R G M L K R L K K M K			
H2LNLD_ORYLA/340-375	VW-S	I L S K A P P S A F E K I A F Q Y G I T D L R G M L K R L K K M K			
G1PGG6_MYOLU/177-212	VW-V	E L L K N A N P N E Y E K I A F Q Y G I T D L R G M L K R L K R M R			
M4A199_XIPMA/238-274	KL	I L K M G H A P P S E F E K I A F Q Y G I T D L R G M L K R L K K M K			
H2TW88_TAKRU/166-201	VW-E	I L L K K A R P D E Y E K I A F T Y G I T D L R G L L R R M K K I P			
F7DKM1_XENTR/222-257	VW-E	I L L K N A K P S D F E R I A F E H G I T D L R G M L K R L K K V N			
H2UDR5_TAKRU/198-233	VW-E	I L L K S A H P S E Y E K I A F Q Y G I T D L R G M L K R L K K M K			
EZR072_CANFA/214-249	IW-E	I L K G A K K S E Y E K I A F Q Y G I T D L R G M L K R L K K A K			
H0VYP7_CAVPO/211-246	IW-E	I L K G A K K S E Y E K I A F Q Y G I T D L R G M L K R L K K A K			
F7EU62_CALJA/167-202	IW-E	I L K G A K K S E Y E K I A F Q Y G I T D L R G M L K R L K K A K			
I3KBQ9_ORENI/184-219	VW-E	I L L S A D K K D Y E R I C A E Y G I T D F R G M L K K L N E M K			
H2UDR3_TAKRU/157-192	AT-K	K L K S A H P S E Y E K I A F Q Y G I T D L R G M L K R L K K M K			
E4WXZ2_OIKD/149-184	IW-A	L L K S A N P R E Y D R I A F Y W G I K D L R K L L K K L A N A K			
F7FUF4_MACMU/132-167	VW-Q	L L M T A D R K D Y E Q I C M K Y G I V D Y R G M L R K L Q E M K			
M7AMC7_CHENMY/10207-10242	IM-E	L L K N V D P K E Y E K Y A R M Y G I T D F R G L L Q A F E L L K			
M4AU85_XIPMA/157-192	LL-E	L L S A P K K D Y E R I C L E F G I T D F R W F L K K L K Q I K			
H2RX04_TAKRU/182-217	FF-F	E V L M S A E K K D Y E S I C I Q Y G V T D F R G M L K K L N E K K			
E7F8W2_DANRE/184-219	FW-F	D V M L N A K K S D Y E R I C H E F G I T D Y R W M L K Q L N L K			
G3U743_LOXAF/192-227	ML-E	I L S K V P K K D F E R V C M V Y G F T D F W G L L K K L K E M K			
G1MT65_MELGA/168-203	I F-Q	L L H A D K K D Y E R I C I K Y G I S D F R G M L R A L Q D L R			
I3KJU9_ORENI/160-195	FW-D	V M L K A D R N D Y E R I C S E F G V K D L H S I L K K L D E K K			



1. File -> Save as



Home | Search | Results | Software | Help | About

phmmmer 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 |

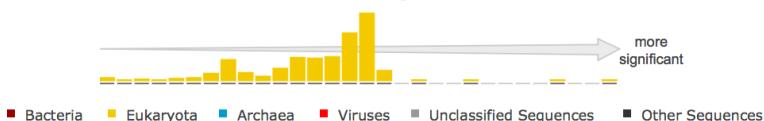


## HMMSEARCH Results

Score Taxonomy Domain Download

Search Again

## Distribution of Significant Hits



&lt; First &lt;&lt; Previous Page 1 of 4 Next &gt;&gt; Last &gt;

## 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 - Q3UIKO (Q3UIKO\_MOUSE)

Protein | Submitted name: **Myosin-binding protein C, cardiac-type**Gene | **Mybpc3**Organism | *Mus musculus (Mouse)*Status |  Unreviewed - Annotation score:  - Experimental evidence at protein level<sup>i</sup>

Display None

 BLAST Align Format Add to basket History Help video Other tutorials and videos FeedbackFunction<sup>i</sup>GO - Molecular function<sup>i</sup>

- [identical protein binding](#)  Source: MGI
- [myosin binding](#)  Source: MGI
- [myosin heavy chain binding](#)  Source: MGI ▾
- [structural constituent of cytoskeleton](#)  Source: MGI ▾

 Function Names & Taxonomy Subcell. location Pathol./Biotech

Function Names & Taxonomy Subcell. location Pathol./Biotech PTM / Pro Expression Interaction Structure Family & Domains Sequence Cross-references Publications Entry information Miscellaneous

# Function<sup>i</sup>

## GO - Molecular function<sup>i</sup>

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

Inferred from physical interaction<sup>i</sup>

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

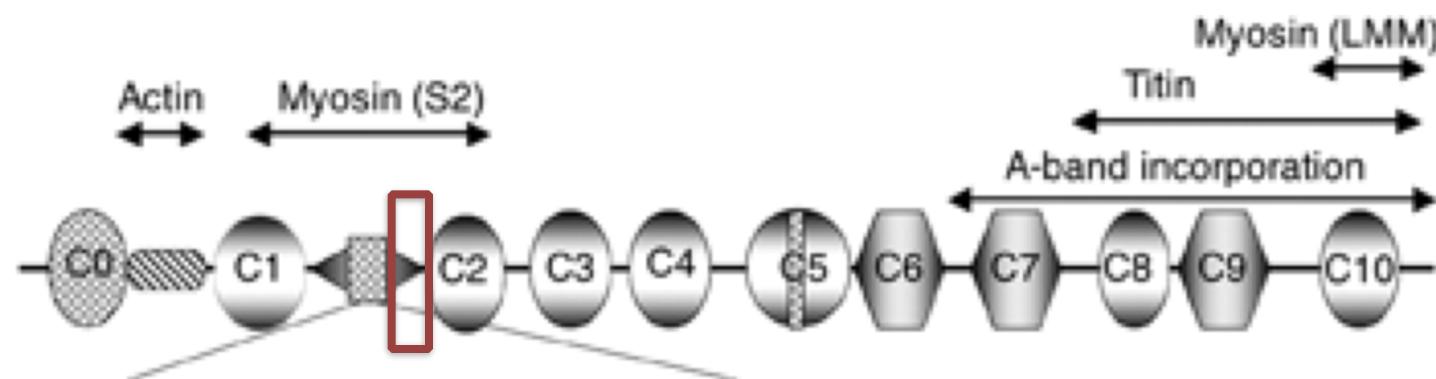
Format

Mapped (16)

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(AlIP+)). 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(AlIP+) 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(AlIP+:(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**

271-FRRRTSLAGGG RRISDSHEDT GILDFSSLLK KRDSFRTPRK-Human  
269-\*\*\*\*\*A\* \*\*T\*\*\*\*\*A \*TP\*\*\*\*\* RDS\*-Mouse  
269-\*\*\*\*D\*\*\*A\* \*\*TD\*\*\*\*\*A \*TP\*\*\*\*\* RDS\*-cMyBP-C<sup>AllP+</sup>

Cardiac-specific      IgG-like domains      Fibronectin type-III domains      cMyBP-C motif      Pro-Ala domain

[Home](#) [Search](#) [Results](#) [Software](#) [Help](#) [About](#)[phmmmer](#) [hmmscan](#) [hmmsearch](#) [jackhmmer](#)

## protein alignment/profile-HMM vs protein sequence database

[Paste a Sequence](#) | [Upload a File](#) | [Accession Search](#) jalview.fasta

### ▼ 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

[Home](#) | [Search](#) | [Results](#) | [Software](#) | [Help](#) | [About](#)**HMMSEARCH Results**[Score](#)   [Taxonomy](#)   [Domain](#)   [Download](#)[Search Again](#)**Distribution of Significant Hits** 

A legend showing color-coded boxes for different taxonomic groups: Bacteria (red), Eukaryota (yellow), Archaea (blue), Viruses (dark red), Unclassified Sequences (grey), and Other Sequences (black).

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

	Target	Species	E-value
>	<a href="#">2ihu_A</a>	<a href="#">Mus musculus</a>	5.8e-22
<a href="#">(show all) alignments</a>			

Your search took: 0.01 secs  
showing rows 1 - 1 of 1[Search Details](#)



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

## Customize

		Target				Species				E-value		
V	2Ihu_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	<u>LDFSSLLLKK</u>	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	.....*.....*.....*	36
Target	84	VWEILRQAPPSEYERIAFHGVTDLRGMLKLKGMM	119
PP	8	*****	*****

(show all) alignments

Your search took:0.01 secs  
showing rows 1 - 1 of 1



An Information Portal to  
112387 Biological  
Macromolecular Structures



Search by PDB ID, author, macromolecule, sequence, or ligands

Go

[Advanced Search](#) | [Browse by Annotations](#)

[Summary](#) [3D View](#) [Sequence](#) [Annotations](#) [Seq. Similarity](#) [3D Similarity](#) [Literature](#) [Biol. & Chem.](#) [Methods](#) [Links](#)

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

DOI:[10.22110/pdb2lhu/pdb](https://doi.org/10.22110/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](https://pubmed.ncbi.nlm.nih.gov/22235120/)

PubMedCentral: [PMC3318737](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3318737/)

DOI: [10.1074/jbc.M111.309591](https://doi.org/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.

Funnctionally, the m-domain is thought to directly interact with myosin, whereas phosphorylation of the m-domain has been shown to

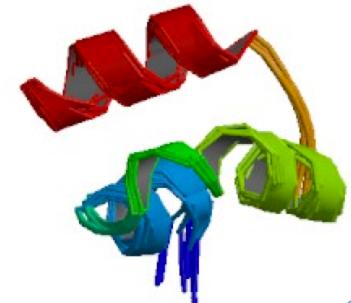
2LHU

[Display Files](#) ▾

[Download Files](#) ▾

[Download Citation](#) ▾

### 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 SSSLKKRDSFRRDSKLEAPAAEEDVWEILRQAPPSEYERIAFQHGVTDLRGMLKRLKGKQ  
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

Query/63-98  
H3B137\_LATCH/320-355  
F1NBZ9\_CHICK/303-338  
G3PS89\_GASAC/335-370  
W5KF69\_ASTMX/356-391  
H3CVJ4\_TETNG/330-365  
H2LN0L\_ORYLA/340-375  
G1PGG8\_MYOUL/177-212  
M4A199\_XIPMA/238-274  
H2TW88\_TAKRU/166-201  
F7DKM1\_XENTR/222-257  
H2UDR5\_TAKRU/198-233  
F7DKM1\_XENTR/222-257  
H2UDR5\_TAKRU/198-233

10	20	30
VW-EILRQAPPSEYERIAFQHGVTDLRGMLKRLKGKQ	VW-DLLRQAPPSEYEKIAFQYGITDLRGMLKRLKGKQ	VW-NILSHAPSSEYEKIAFQHGVTDLRGMLKRLKGKQ
VW-EILSHAPASEYEKIAFQYGITDLRGMLKRLKGKQ	VW-EILSKAPSSSEYEKIAFQHGVTDLRGMLKRLKGKQ	VW-SILSKAPPSAFEKIAFQYGITDLRGMLKRLKGKQ
VW-EILSKAPPSSEYEKIAFQHGVTDLRGMLKRLKGKQ	VW-ELLKNANPNEYEKIAFQYGITDLRGMLKRLKGKQ	VW-ELLKNAPSEYEKIAFQYGITDLRGMLKRLKGKQ
VW-ELLKNAPSEYEKIAFQYGITDLRGMLKRLKGKQ	VW-ELLKNGHAPPSFEFEKIAFQYGITDLRGMLKRLKGKQ	VW-ELLKSAHPSEYEKIAFQYGITDLRGMLKRLKGKQ
VW-ELLKNGHAPPSFEFEKIAFQYGITDLRGMLKRLKGKQ	VW-EILKKARPDEYEKIAFQYGITDLRGMLLRRMKKIP	VW-EILKSAHPSEYEKIAFQYGITDLRGMLKRLKGKQ
VW-EILKKARPDEYEKIAFQYGITDLRGMLLRRMKKIP	VW-ELLKNAKPSDFERIAFQYGITDLRGMLKRLKGKQ	VW-ELLKNAKPSDFERIAFQYGITDLRGMLKRLKGKQ
VW-ELLKNAKPSDFERIAFQYGITDLRGMLKRLKGKQ	VW-ELLKNAKPSDFERIAFQYGITDLRGMLKRLKGKQ	VW-ELLKNAKPSDFERIAFQYGITDLRGMLKRLKGKQ