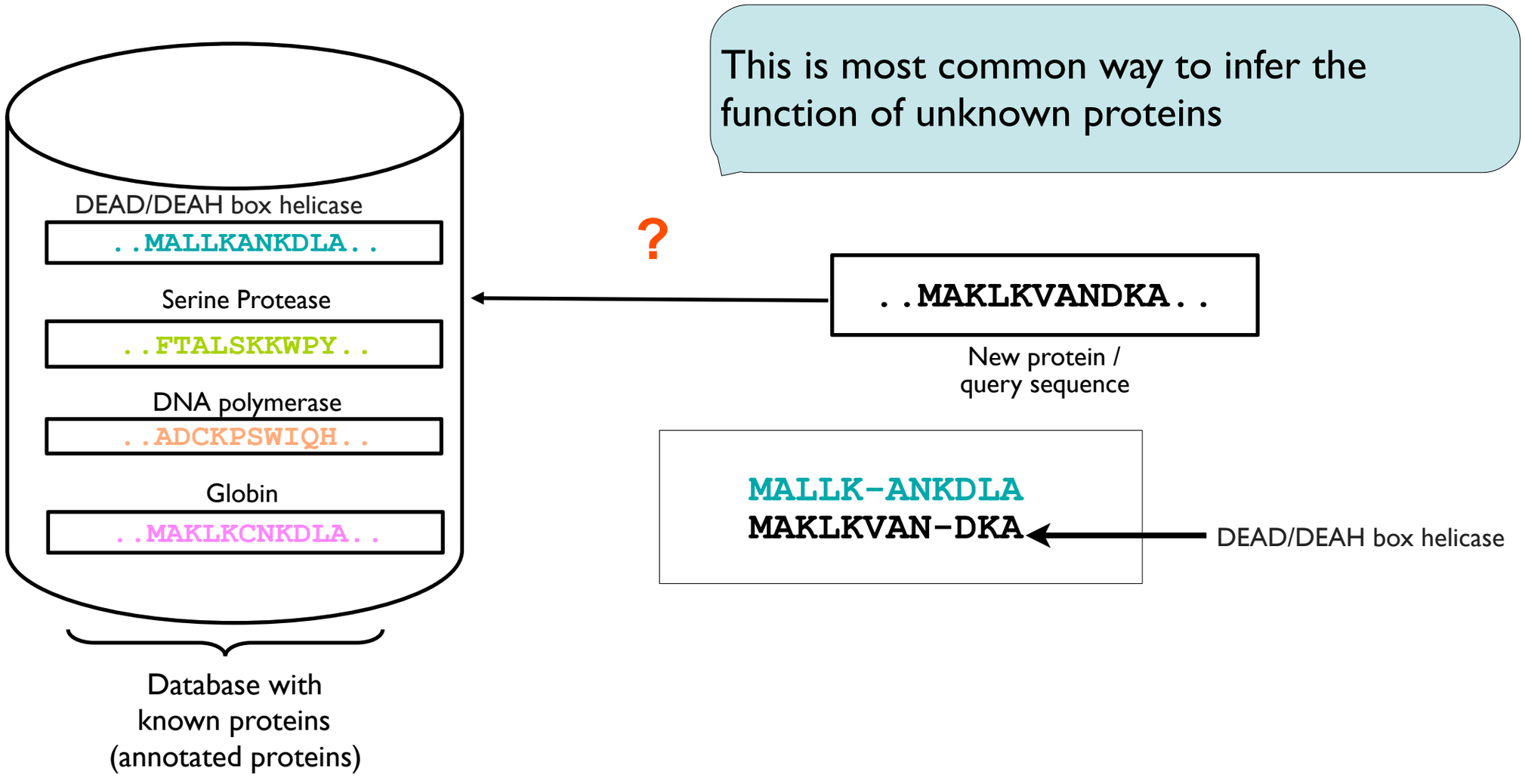
LI323: Statistique en Bioinformatique : Analyse statistique d'une famille de proteines

Martin Weigt,
Juliana Silva Bernardes
<juliana.silva_bernardes@upmc.fr>

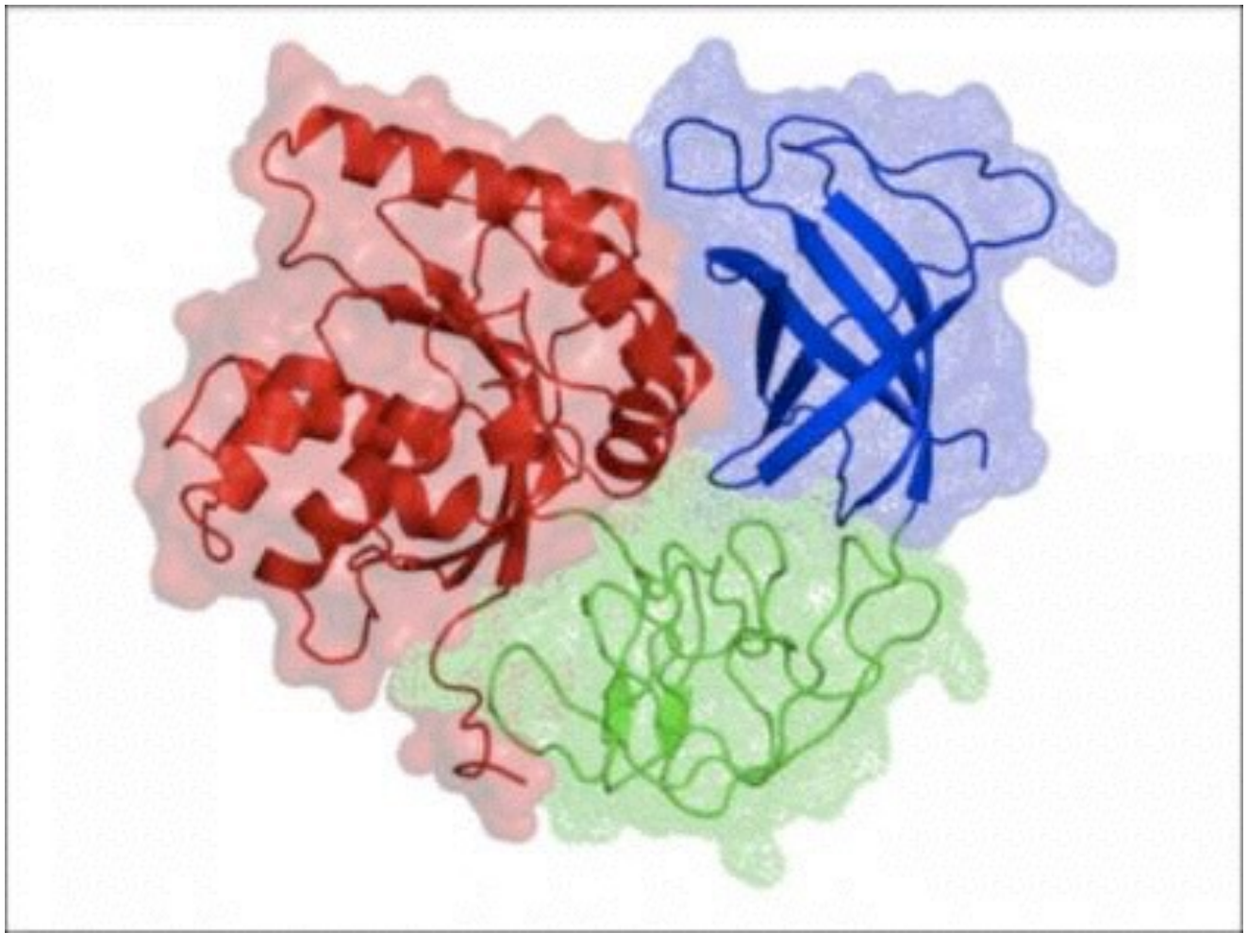
Functional Transfer Annotation

➡Homology detection is essential for functional transfer annotation

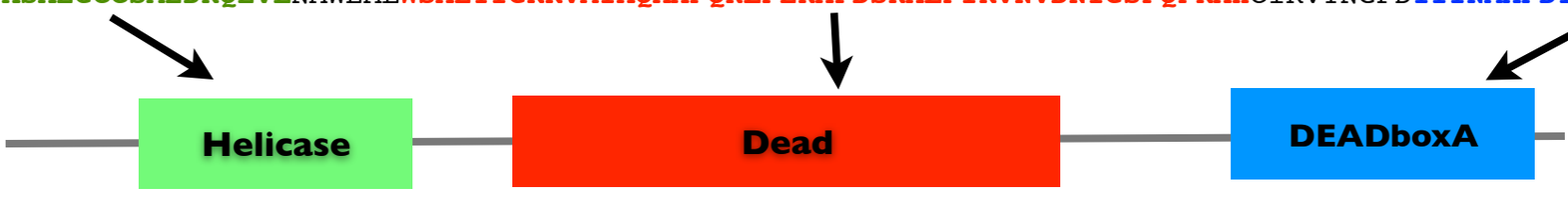


Protein Domains

- Domains are the building blocks of proteins.



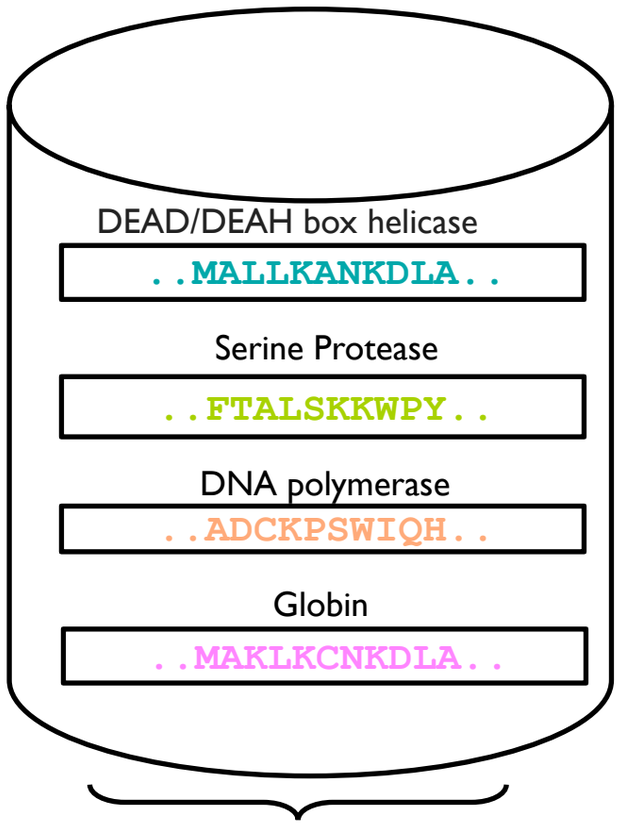
MFRFALLCAFVADASAEGCCSMEDRQEVLNAWEALWSAEYTGRRVMIAQAAFQKLFEKAPDSKALFTRVNVDNIGSPQFRAHCIRVTNGFDTIINMAFDTDVLEELLTHLGNQHTKYQGMRAA



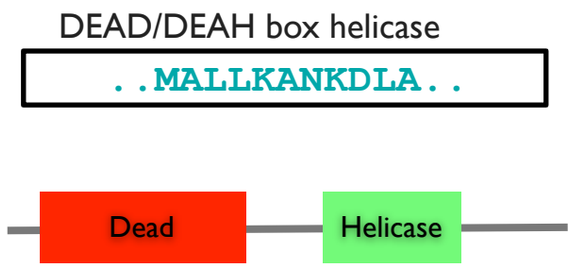
Protein Classification

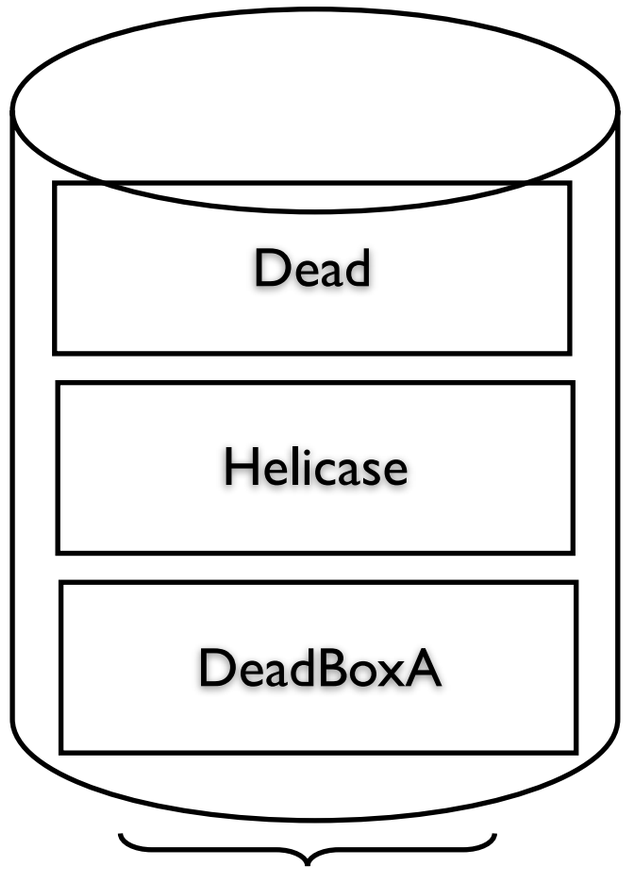
➡To improve functional transfer annotation we can classify known protein sequences according to their functional regions (**domains**).

- ➡ Proteins are generally comprised of one or more **domains**.



Database with
known proteins
(annotated proteins)

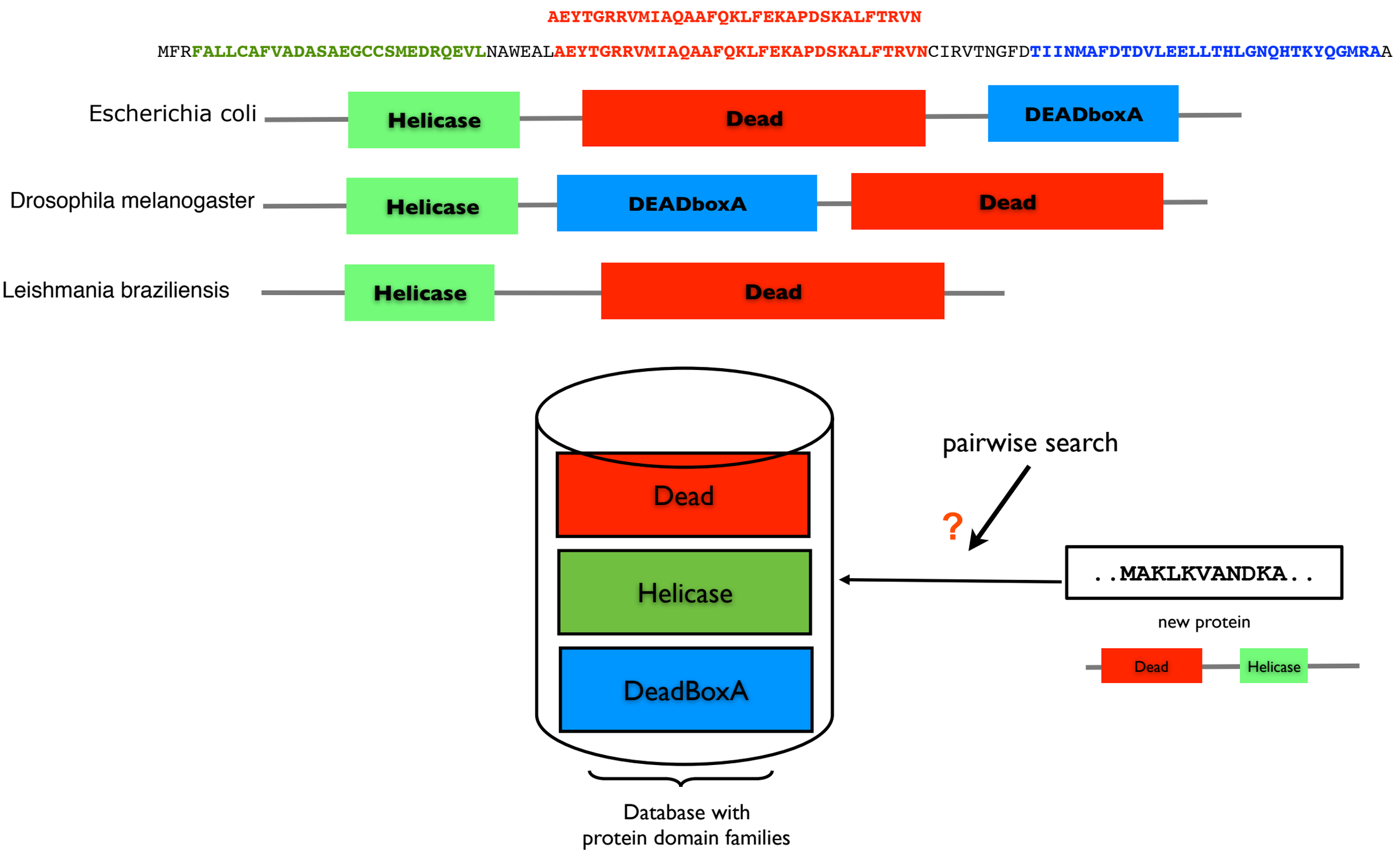




Database with
protein domain families

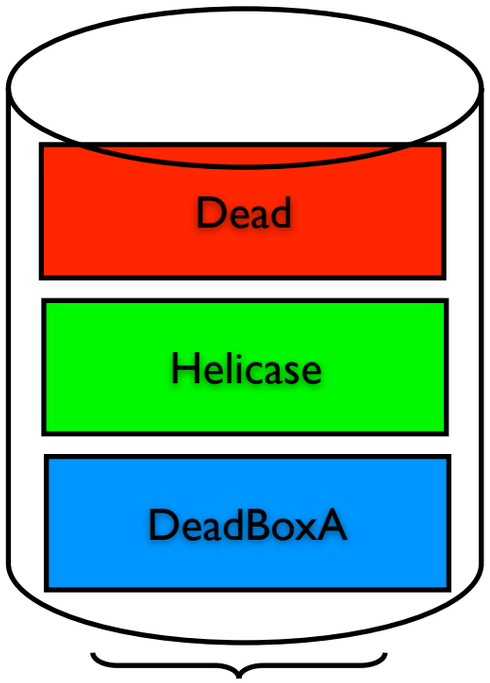
Domain Recognition

- Identifying domains can help to determine protein function.



Domain Recognition Tools

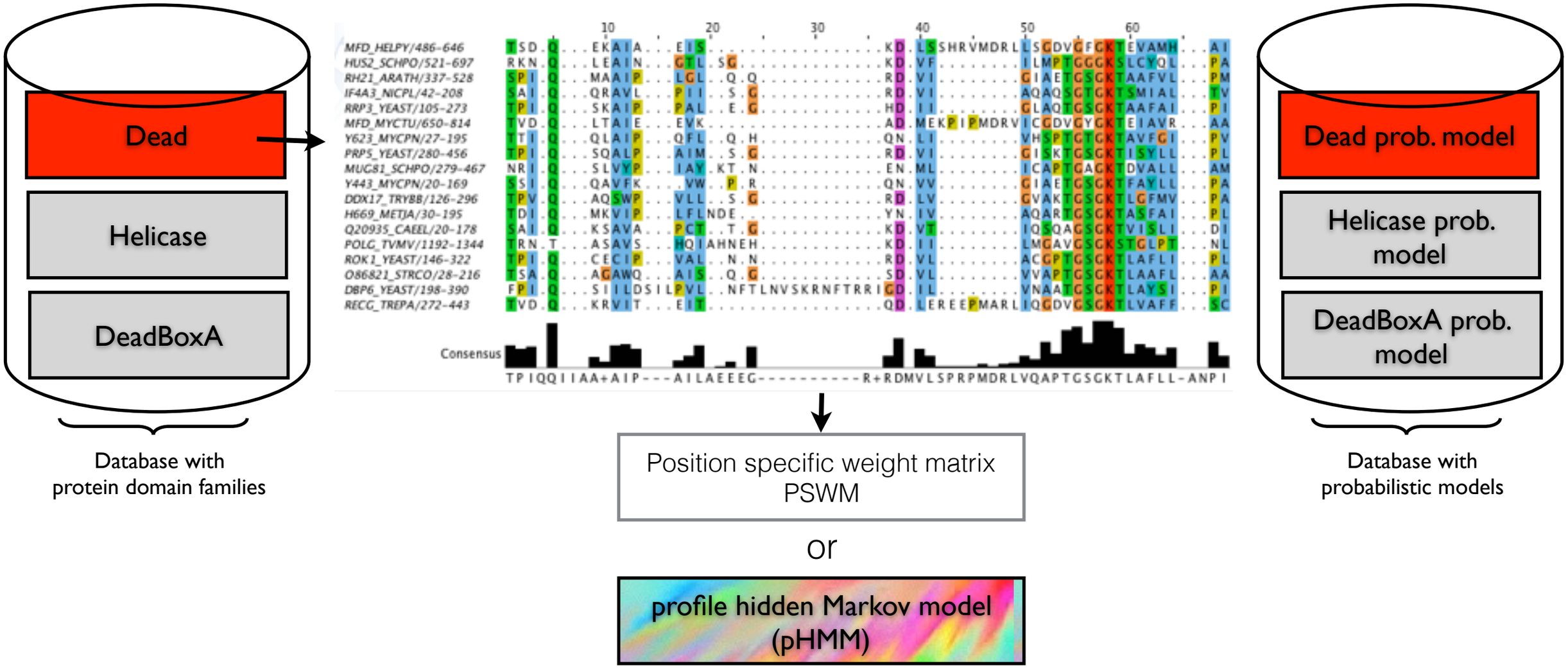
- Known domains are described with probabilistic models representing the consensus among domain sequences



Database with
protein domain families

Domain Recognition Tools

- Known domains are described with probabilistic models representing the consensus among domain sequences



Domain Databases

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Pfam 27.0 (March 2013, 14831 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	**YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...**
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY	View Pfam family annotation and alignments
VIEW A CLAN	See groups of related families
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID Go Example

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

PFAM Database

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Pfam 27.0 (March 2013, 14831 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS

SEQUENCE SEARCH

VIEW A PFAM FAMILY

VIEW A CLAN

VIEW A SEQUENCE

VIEW A STRUCTURE

KEYWORD SEARCH

JUMP TO

VIEW PFAM FAMILY ANNOTATION AND ALIGNMENTS

Enter a family identifier (e.g. *Piwi*) or accession (e.g. *PF02171*) to see all data for that family.

PF00018 Go Example

You can also browse through the list of all Pfam families.

Family: *SH3_1* (PF00018)

695 architectures | 10749 sequences | 11 interactions | 444 species | 373 structures

Summary
Domain organisation
Clan
Alignments
HMM logo
Trees
Curation & model
Species
Interactions
Structures

Jump to...
enter ID/acc Go

Summary: SH3 domain

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: SH3 domain | Pfam | InterPro

This is the Wikipedia entry entitled "SH3 domain". More...

SH3 domain

Edit Wikipedia article

The **SRC Homology 3 Domain** (or **SH3 domain**) is a small protein domain of about 60 amino acids residues first identified as a conserved sequence in the viral adaptor protein v-Crk and the non-catalytic parts of enzymes such as phospholipase and several cytoplasmic tyrosine kinases such as Abl and Src.[1][2] It has also been identified in several other protein families such as: PI3 Kinase, Ras GTPase-activating protein, CDC24 and cdc25.[3][4][5] SH3 domains are found in proteins of signaling pathways regulating the cytoskeleton, the Ras protein, and the Src kinase and many others. They also regulate the activity state of adaptor proteins and other tyrosine kinases and are thought to increase the substrate specificity of some tyrosine kinases by binding far away from the active site of the kinase. Approximately 300 SH3 domains are found in proteins encoded in the human genome.

Contents [hide]

1 Structure
2 Peptide binding
3 Proteins with SH3 domain
4 See also
5 References
6 External links

SH3 domain

Ribbon diagram of the SH3 domain, alpha spectrin, from chicken (PDB accession code 1SHG), colored from blue (N-terminus) to red (C-terminus).

Identifiers	
Symbol	SH3_1
Pfam	PF00018
Pfam clan	CL0010
InterPro	IPR001452
SMART	SM00326
PROSITE	PS50002
SCOP	1shf
SUPERFAMILY	1shf
CDD	cd00174
Available protein structures: [show]	

Structure

The SH3 domain has a characteristic beta-barrel fold that consists of five or six β-strands arranged as two tightly packed anti-parallel β sheets. The linker regions may contain short helices. The SH3-type fold is an ancient fold found in eukaryotes as well as prokaryotes.[6]

Peptide binding

The classical SH3 domain is usually found in proteins that interact with other proteins and mediate assembly of specific protein complexes, typically via binding to proline-rich peptides in their respective binding partner. Classical SH3 domains are restricted in humans to intracellular proteins, although the small human MIA family of extracellular proteins also contain a domain with an SH3-like fold.

Many SH3-binding epitopes of proteins have a consensus sequence that can be represented as a regular expression or Short linear motif:

PFAM Database

Family: *SH3_1* (PF00018)

695 architectures | 10749 sequences | 11 interactions | 444 species | 373 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to... enter ID/acc Go

Alignments

We store a range of different sequence alignments for families. As well as the seed alignment from which the family is built, we provide the full alignment, generated by searching the sequence database using the family HMM. We also generate alignments using four representative proteomes (RP) sets, the NCBI sequence database, and our metagenomics sequence database. **More...**

View options

We make a range of alignments for each Pfam-A family. You can see a description of each **above**. You can view these alignments in various ways but please note that some types of alignment are never generated while others may not be available for all families, most commonly because the alignments are too large to handle.

	Seed (61)	Full (10749)	Representative proteomes				NCBI (20245)	Meta (89)
			RP15 (1639)	RP35 (2410)	RP55 (4041)	RP75 (5929)		
Jalview	✓	✓	✓	✓	✓	✓	✓	✓
HTML	✓	–	✓	✓	✓	–	×	×
PP/heatmap	×[1]	–	✓	✓	✓	–	×	×
Pfam viewer	✓	✓	×	×	×	×	×	×

[1]Cannot generate PP/Heatmap alignments for seeds; no PP data available

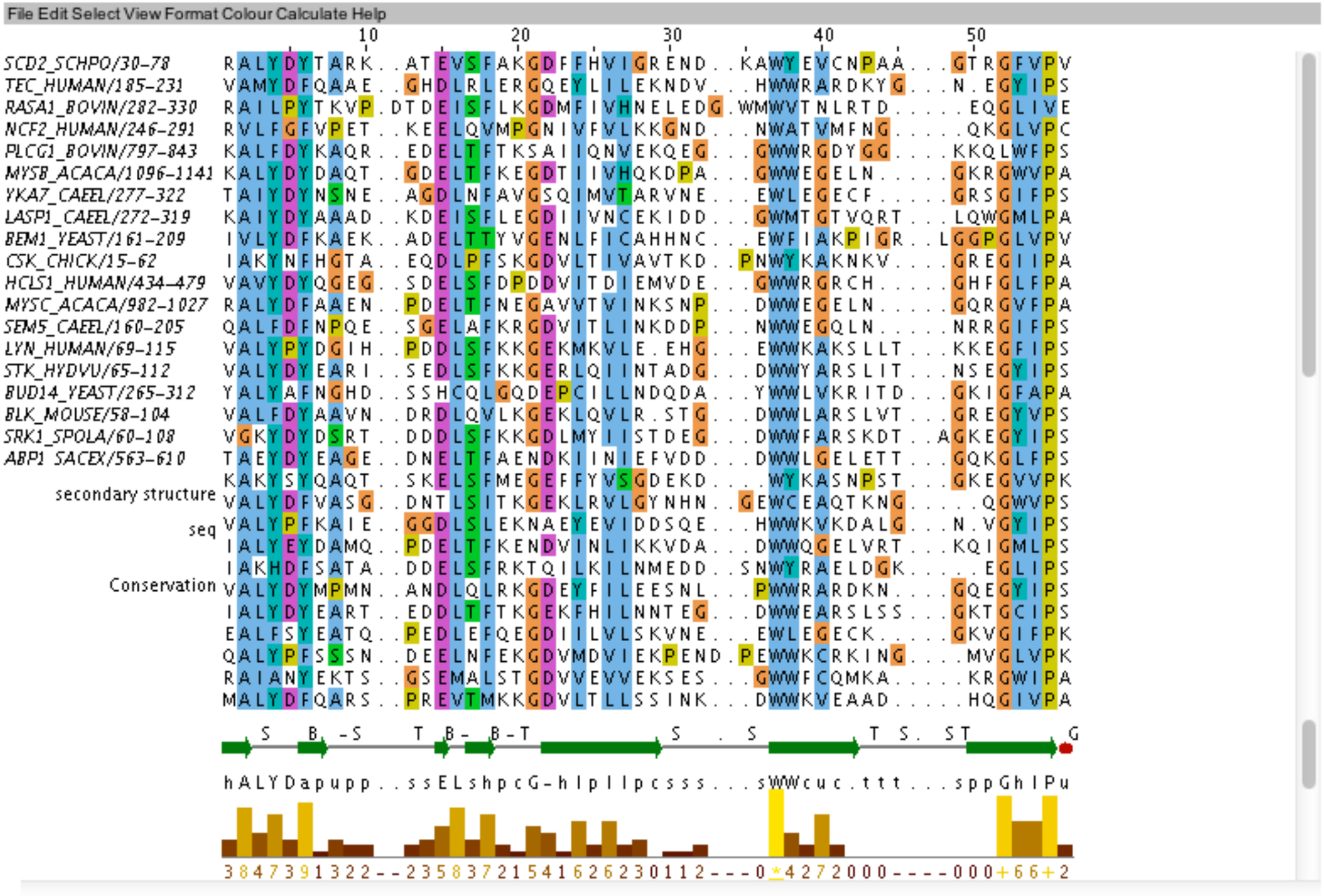
Key: ✓ available, × not generated, – not available.

Format an alignment

PFAM Database

EMBL-EBI

View seed alignment for *PF00018* using Jalview



Successfully loaded file http://pfam.xfam.org/family/PF00018/alignment/seed

How to use aligned sequences to detect homology?

Training set

A	B	B	C	A	C
A	-	B	B	A	C
A	B	-	B	A	C
C	B	B	C	A	C

1 2 3 4 5 6

Domain D

$\mathcal{A}$ ={A,B,C,-}

How to use aligned sequences to detect homology?

Training set

A	B	B	C	A	C
A	-	B	B	A	C
A	B	-	B	A	C
C	B	B	C	A	C
0	1	2	3	4	5

Domain D

$\mathcal{A}$ ={A,B,C,-}

Test sequence ABCAB

The test sequence matches the domain D ?

How to use aligned sequences to detect homology?

Training set

A	B	B	C	A	C
A	-	B	B	A	C
A	B	-	B	A	C
C	B	B	C	A	C

0 1 2 3 4 5

$\mathcal{A}$ ={A,B,C,-}

Test sequence ABCAB

$n_i(a)$ = frequency of amino acid a in the position i

$$D_{train} = \begin{pmatrix} a_0^1 & a_2^1 & \dots & a_{L-1}^1 \\ a_0^2 & a_2^2 & \dots & a_{L-1}^2 \\ & & \dots & \\ a_0^M & a_2^M & \dots & a_{L-1}^M \end{pmatrix} .$$

$$\sum_{a \in \mathcal{A}} n_i(a) = M$$

Where $a_i^m \in \mathcal{A}$ and $i \in \{0, ..., L-1\}$

Computing PSWM

Training set

A	B	B	C	A	C
A	-	B	B	A	C
A	B	-	B	A	C
C	B	B	C	A	C
0	1	2	3	4	5

L=? M=? i={…}

L=6 M=4 i={0, 1, …,5}

$n_0(A) = 3$ $n_0(B) = 0$ $n_0(C) = 1$ $n_0(-) = 0$

$n_0(A) + n_0(B) + n_0(C) + n_0(-) = 4$

$$D_{train} = \begin{pmatrix} a_0^1 & a_2^1 & \dots & a_{L-1}^1 \\ a_0^2 & a_2^2 & \dots & a_{L-1}^2 \\ & & \dots & \\ a_0^M & a_2^M & \dots & a_{L-1}^M \end{pmatrix} .$$

Your turn : compute all $n_i(a)$

Where $a_i^m \in \mathcal{A}$ and $i \in \{0, ..., L-1\}$

$n_i(a)$ = frequency of amino acid a in the position i

$$\sum_{a \in \mathcal{A}} n_i(a) = M$$

Computing PSWM

Training set

L=6 M=4 i={0, 1, …,5} $\mathcal{A}$ ={A,B,C,-}

A	B	B	C	A	C
A	-	B	B	A	C
A	B	-	B	A	C
C	B	B	C	A	C
0	1	2	3	4	5

	0	1	2	3	4	5
A	3					
B	0					
C	1					
-	0					
	4					

Computing PSWM

L=6 M=4 i={0, 1, …,5} $\mathcal{A}$ ={A,B,C,-}

Training set

A	B	B	C	A	C
A	-	B	B	A	C
A	B	-	B	A	C
C	B	B	C	A	C
0	1	2	3	4	5

Frequency matrix

	0	1	2	3	4	5
A	3	0	0	0	4	0
B	0	3	3	2	0	0
C	1	0	0	2	0	4
-	0	1	1	0	0	0
Total	4	4	4	4	4	4

The position specific weight matrix is computed by $\omega_i(a) = \frac{n_i(a)+1}{M+q}$

Where q is the size of $\mathcal{A}$

Computing PSWM

L=6 M=4 i={0, 1, …,5} $\mathcal{A}$ ={A,B,C,-}

Frequency matrix

	0	1	2	3	4	5
A	3	0	0	0	4	0
B	0	3	3	2	0	0
C	1	0	0	2	0	4
-	0	1	1	0	0	0
Sum	4	4	4	4	4	4

PSWM

	0	1	2	3	4	5
A	4/8	1/8	1/8	1/8	5/8	1/8
B	1/8	4/8	4/8	3/8	1/8	1/8
C	2/8	1/8	1/8	3/8	1/8	5/8
-	1/8	2/8	2/8	1/8	1/8	1/8
Sum	1	1	1	1	1	1

The position specific weight matrix is computed by $\omega_i(a) = \frac{n_i(a)+1}{M+q}$ Where q is the size of $\mathcal{A}$

$w_0(A) = (n_0(A) + 1)/ (M + q) = (3 +1)/(4+4) = 4/8=1/2$

$w_0(B) = (n_0(B) + 1)/ (M + q) = (0 +1)/(4+4) = 1/8$

$w_0(C) = (n_0(C) + 1)/ (M + q) = (1 +1)/(4+4) = 2/8$

$w_0(-) = (n_0(-) + 1)/ (M + q) = (0 +1)/(4+4) = 1/8$

How to determine conserved positions

- We are looking for conserved positions (higher weights)
- To find them we compute the entropy S_i of each position

$$S_i = \log_2(q) + \sum_{a \in \mathcal{A}} \omega_i(a) \cdot \log_2 [\omega_i(a)]$$

A	B	B	C	A	C
A	-	B	B	A	C
A	B	-	B	A	C
C	B	B	C	A	C
0	1	2	3	4	5

PSWM

	0	1	2	3	4	5
A	4/8	1/8	1/8	1/8	5/8	1/8
B	1/8	4/8	4/8	3/8	1/8	1/8
C	2/8	1/8	1/8	3/8	1/8	5/8
-	1/8	2/8	2/8	1/8	1/8	1/8
Sum	1	1	1	1	1	1

Entropy vector

0	1	2	3	4	5

$S_0 = \log_2(q) + [\, w_0(A).\log_2(w_0(A)) + w_0(B).\log_2(w_0(B)) + w_0(C).\log_2(w_0(C)) + w_0(-).\log_2(w_0(-))]$

$S_0 = \log_2(4) + [\, 4/8.\log_2(4/8) + 1/8.\log_2(1/8) + 2/8.\log_2(2/8) + 1/8.\log_2(1/8)]$

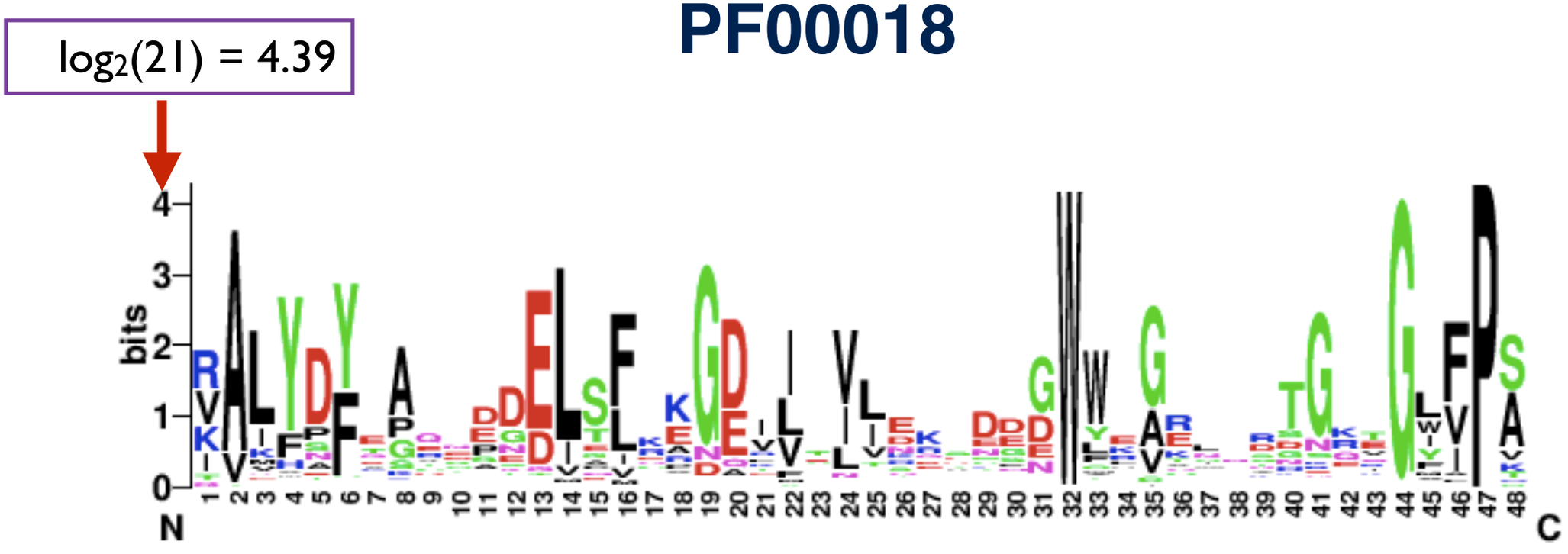
S_i will be a value between 0 (all aa are differents) and $\log_2(4)=2$ (all aa are equals)

The most conserved aa is given by $a_i^\star = \mathrm{argmax}_{a \in \mathcal{A}} \omega_i(a)$.

How to determine conserved positions

$$S_i = \log_2(q) + \sum_{a \in \mathcal{A}} \omega_i(a) \cdot \log_2 \left[\omega_i(a)\right]$$

PF00018



How to evaluate a test sequence?

Training set

A	B	B	C	A	C
A	-	B	B	A	C
A	B	-	B	A	C
C	B	B	C	A	C
0	1	2	3	4	5

PSWM

	0	1	2	3	4	5
A	4/8	1/8	1/8	1/8	5/8	1/8
B	1/8	4/8	4/8	3/8	1/8	1/8
C	2/8	1/8	1/8	3/8	1/8	5/8
-	1/8	2/8	2/8	1/8	1/8	1/8
Sum	1	1	1	1	1	1
	w_0	w_1	w_2	w_3	w_4	w_5

Test sequence ABBAAC

$$P(b_0, ..., b_{L-1}|\omega) = \prod_{i=0}^{L-1} \omega_i(b_i)$$

The test sequence matches the domain D ?

P(ABBACC|w) = w_0(A) * w_1(B) * w_2(B) * w_3(A) * w_4(A) * w_5(C)

P(ABBACC|w) = 4/8* 4/8 * 4/8 * 1/8 * 5/8 * 5/8 = 1600/262144 = 0,0061

How to know if this value is good?

How to evaluate a test sequence?

- We have to compare it to the null model that is not specific to a given position

$$P^{(0)}(b_0, ..., b_{L-1}) = \prod_{i=0}^{L-1} f^{(0)}(b_i) \qquad \text{where} \qquad f^{(0)}(b) = \frac{1}{L} \sum_{i=0}^{L-1} \omega_i(b) \ ,$$

Training set

A	B	B	C	A	C
A	-	B	B	A	C
A	B	-	B	A	C
C	B	B	C	A	C
0	1	2	3	4	5

PSWM

	0	1	2	3	4	5
A	4/8	1/8	1/8	1/8	5/8	1/8
B	1/8	4/8	4/8	3/8	1/8	1/8
C	2/8	1/8	1/8	3/8	1/8	5/8
-	1/8	2/8	2/8	1/8	1/8	1/8
Sum	1	1	1	1	1	1

$f^{(0)}(A) = 0.2708$

$f^{(0)}(B) = 0.2916$

$f^{(0)}(C) = 0.2708$

$f^{(0)}(-) = 0.1666$

$f^{(0)}(A) = (w_0(A) + w_1(A) + w_2(A) + w_3(A) + w_4(A) + w_5(A)) / 6$

$f^{(0)}(A) = (4/8 + 1/8 + 1/8 + 1/8 + 5/8 + 1/8) / 6 = (13/8) / 6 = 0.2708$

$f^{(0)}(B) = (1/8 + 4/8 + 4/8 + 3/8 + 1/8 + 1/8) / 6 = (14/8) / 6 = 0.2916$

$f^{(0)}(C) = (2/8 + 1/8 + 1/8 + 3/8 + 1/8 + 5/8) / 6 = (13/8) / 6 = 0.2708$

$f^{(0)}(-) = (1/8 + 2/8 + 2/8 + 1/8 + 1/8 + 1/8) / 6 = (8/8) / 6 = 0.1666$

How to evaluate a test sequence?

- To compare PSWM and Null model we use the log likelihood

$$\ell(b_0, ..., b_{L-1}) = \log_2 \frac{P(b_0, ..., b_{L-1}|\omega)}{P^{(0)}(b_0, ..., b_{L-1})} = \sum_{i=0}^{L-1} \log_2 \frac{\omega_i(b_i)}{f^{(0)}(b_i)}$$

Training set

A	B	B	C	A	C
A	-	B	B	A	C
A	B	-	B	A	C
C	B	B	C	A	C
0	1	2	3	4	5

PSWM

	0	1	2	3	4	5
A	4/8	1/8	1/8	1/8	5/8	1/8
B	1/8	4/8	4/8	3/8	1/8	1/8
C	2/8	1/8	1/8	3/8	1/8	5/8
-	1/8	2/8	2/8	1/8	1/8	1/8
Sum	1	1	1	1	1	1

$f^{(0)}(A) = 0.2708$

$f^{(0)}(B) = 0.2916$

$f^{(0)}(C) = 0.2708$

$f^{(0)}(-) = 0.1666$

$l(ABBACC) = \log_2 P(ABBAAC|w)/P^0(ABBAAC) = 0.0061/(\ldots)$

$P^0(ABBACC) = f^{(0)}(A)*f^{(0)}(B)*f^{(0)}(B)*f^{(0)}(A)*f^{(0)}(A)*f^{(0)}(C)$

$P^0(ABBACC) = 0.2708* 0.2916* 0.2916* 0.2708* 0.2708* 0.2708 = 0.00046$

$l(ABBACC) = \log_2 P(ABBAAC|w)/P^0(ABBAAC) = \log_2 0.0061/(0.00046) = 3.73$