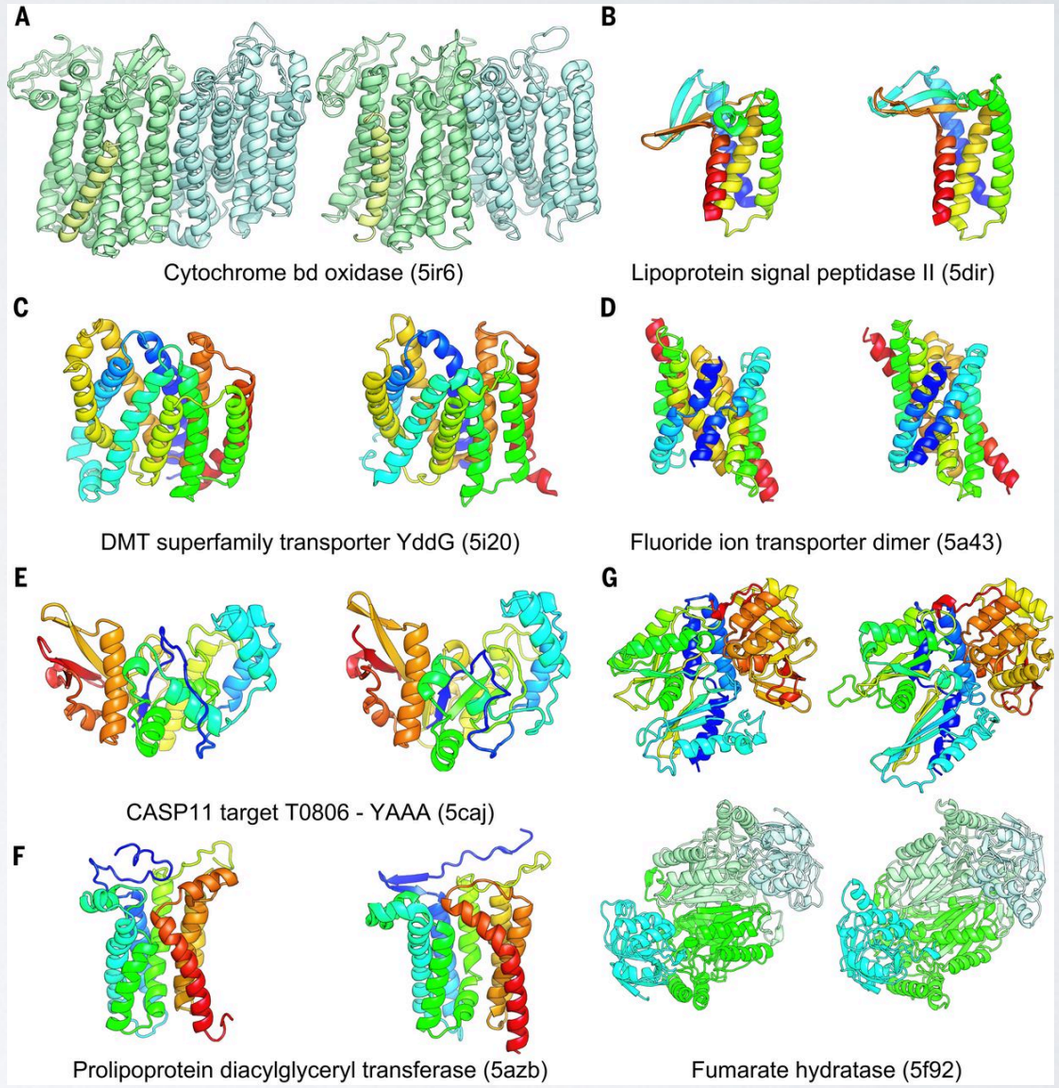
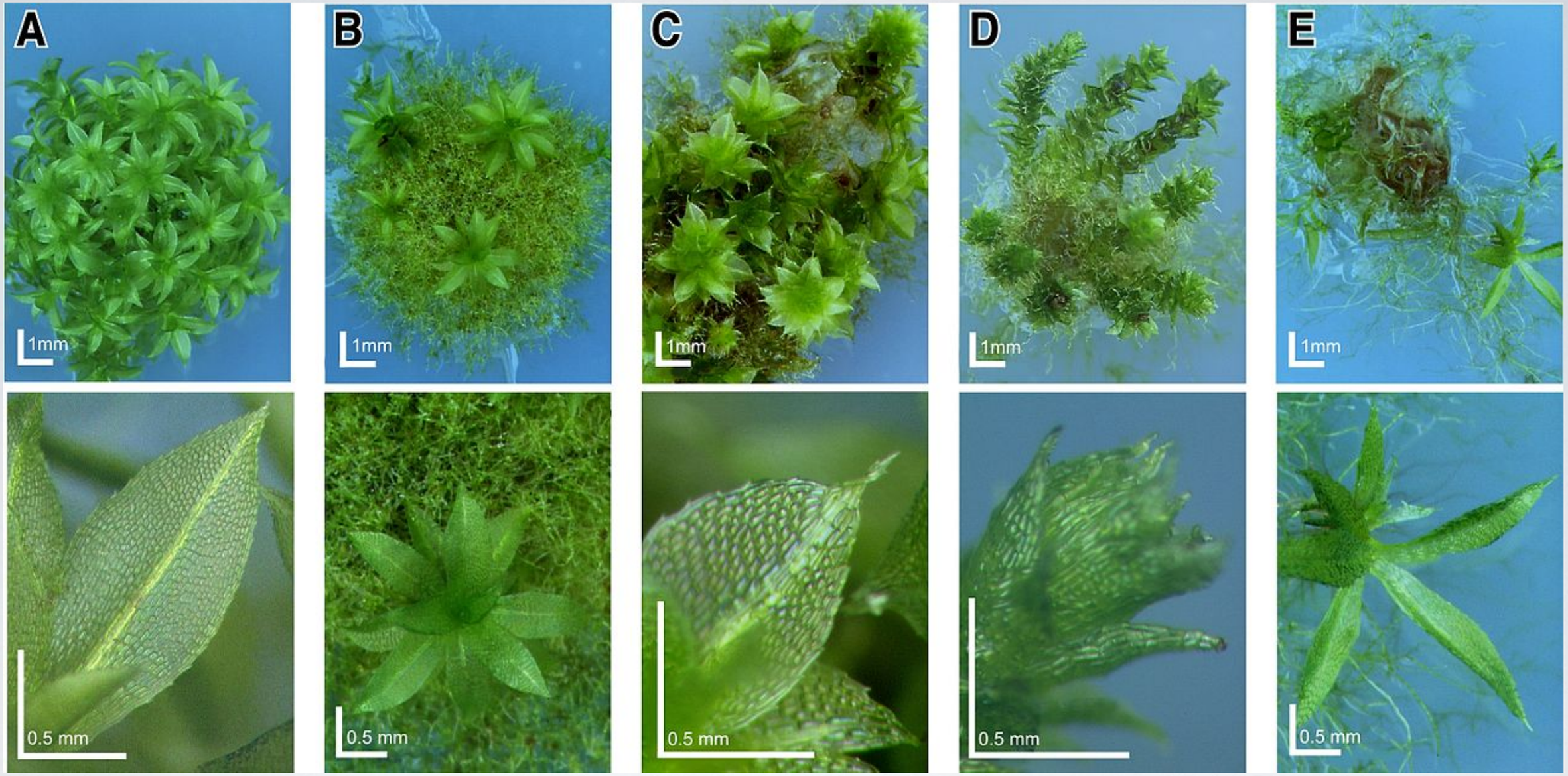
PROTEIN DOMAINS, HMMS & MOTIFS

CLASSIFYING PROTEINS BY FUNCTION

- Important to be able to classify proteins as to what functions they perform
- This information is taken from experimental studies
- Genes have function determined from mutant



https://science.sciencemag.org/content/355/6322/294



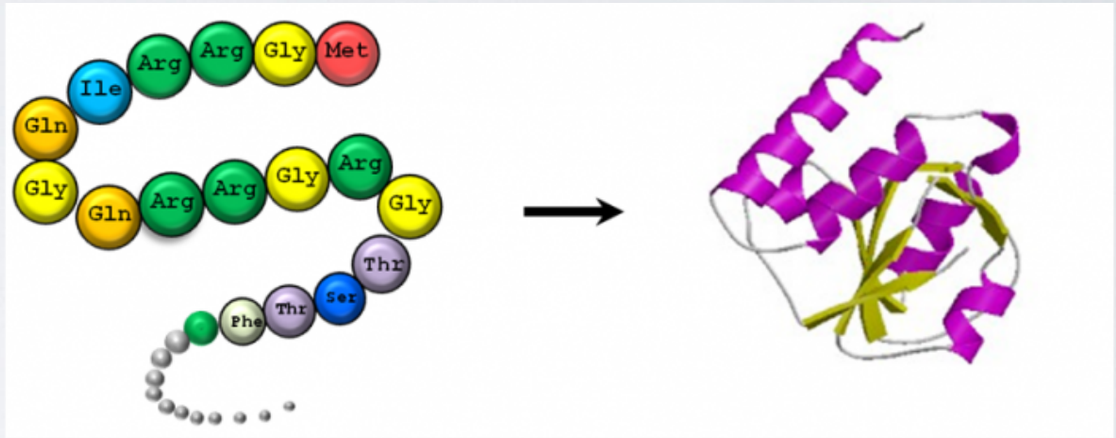
https://en.wikipedia.org/wiki/Reverse_genetics

GENETICS TO FUNCTION

By seeing which mutations break a protein can determine what functional role the protein plays

PROTEIN CLASSIFICATION

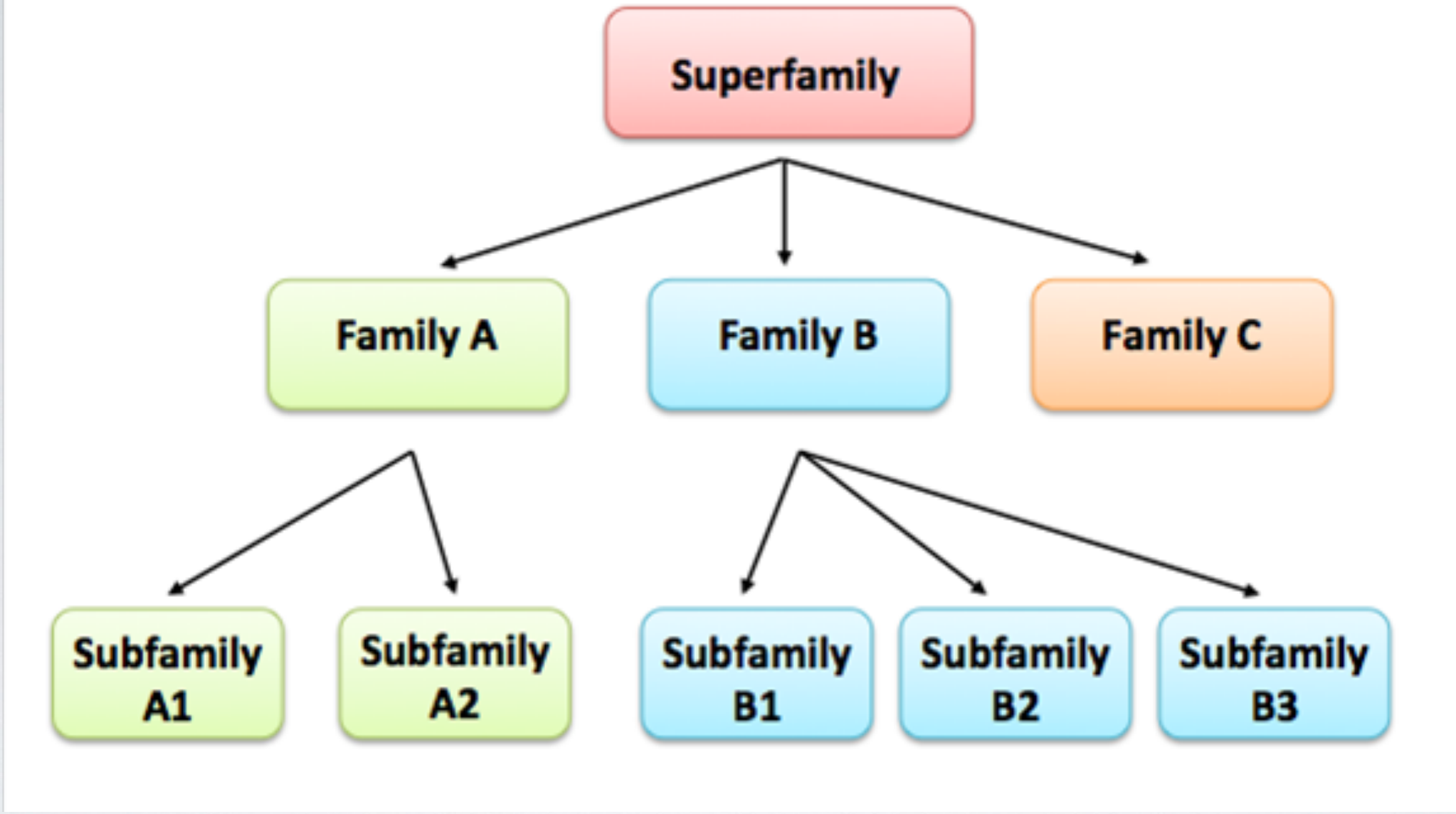
- Many many (!) proteins if consider all the types found in all organisms
- Proteins can be classified into **Families**
- Families can be classified into **Domains**
  - This can be discrete (eg DNA binding) or part of an enzyme
- Sequence can have **Features**



Protein sequence

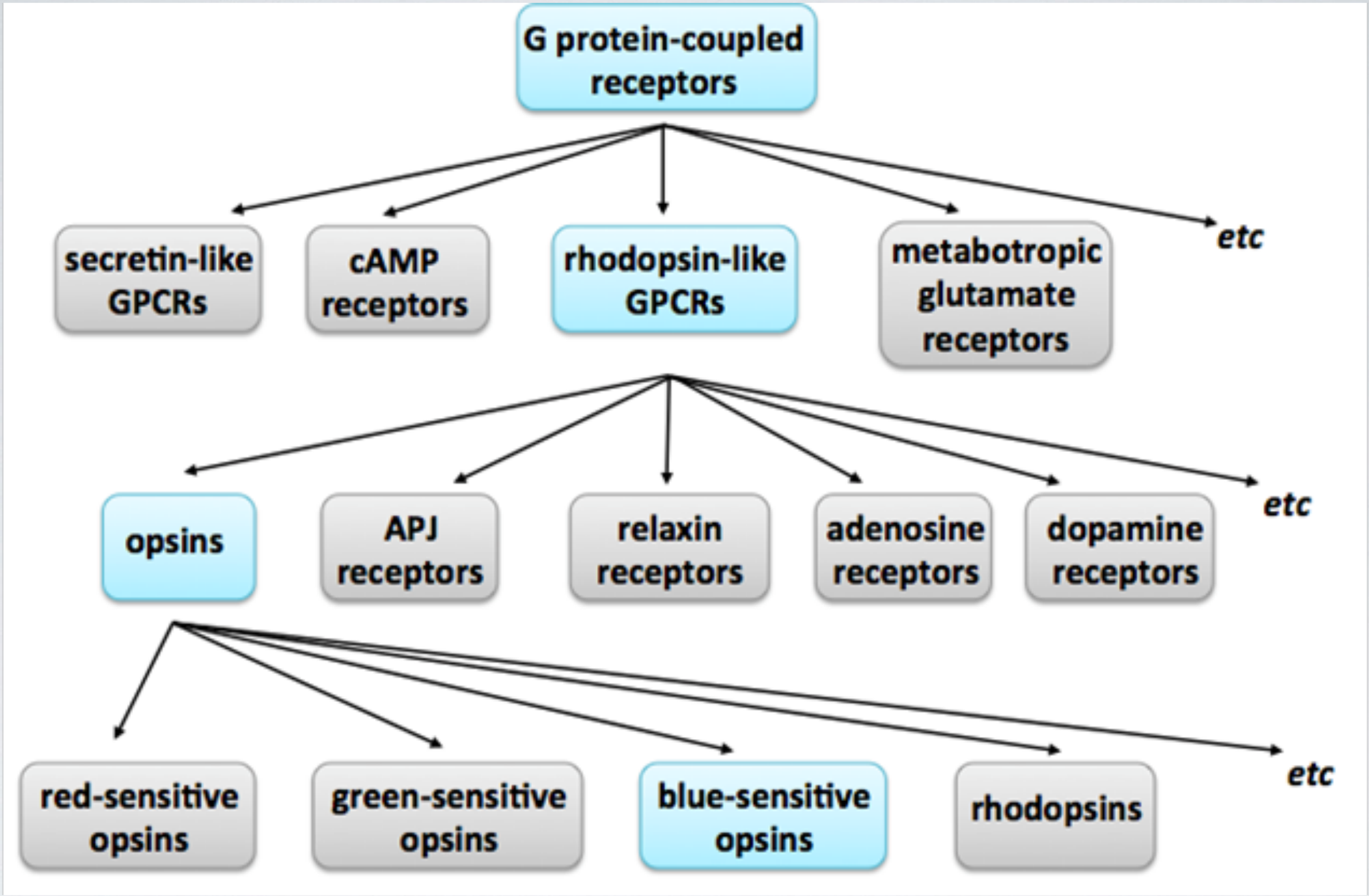
Protein Structure and Folds

https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/protein-classification



PROTEIN CLASSIFICATION

Classification of domains requires recognition of regions in protein which are evolutionary conserved and function as a unit



GPCR SUPERFAMILY

Sub-types of a family - top to bottom this is a classification that is general to specific

PROTEIN DOMAINS

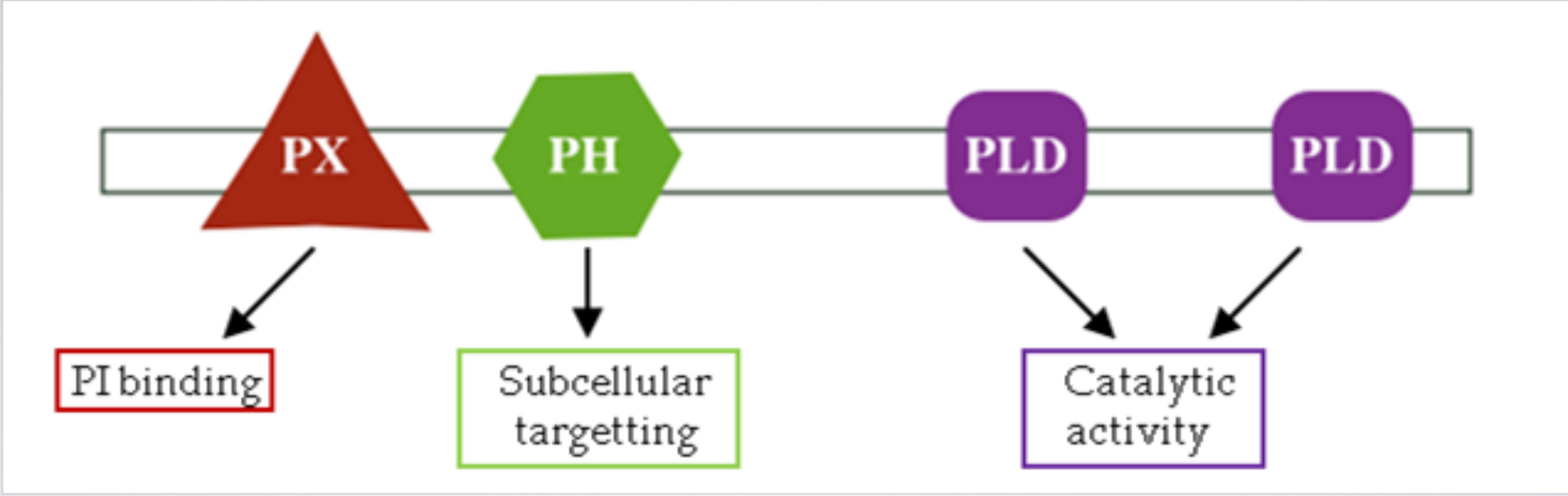
- Distinct functional or structural units of proteins

- SH3 structure shows the 3D folds of the protein when modeled

- Multiple domains can be found within a protein

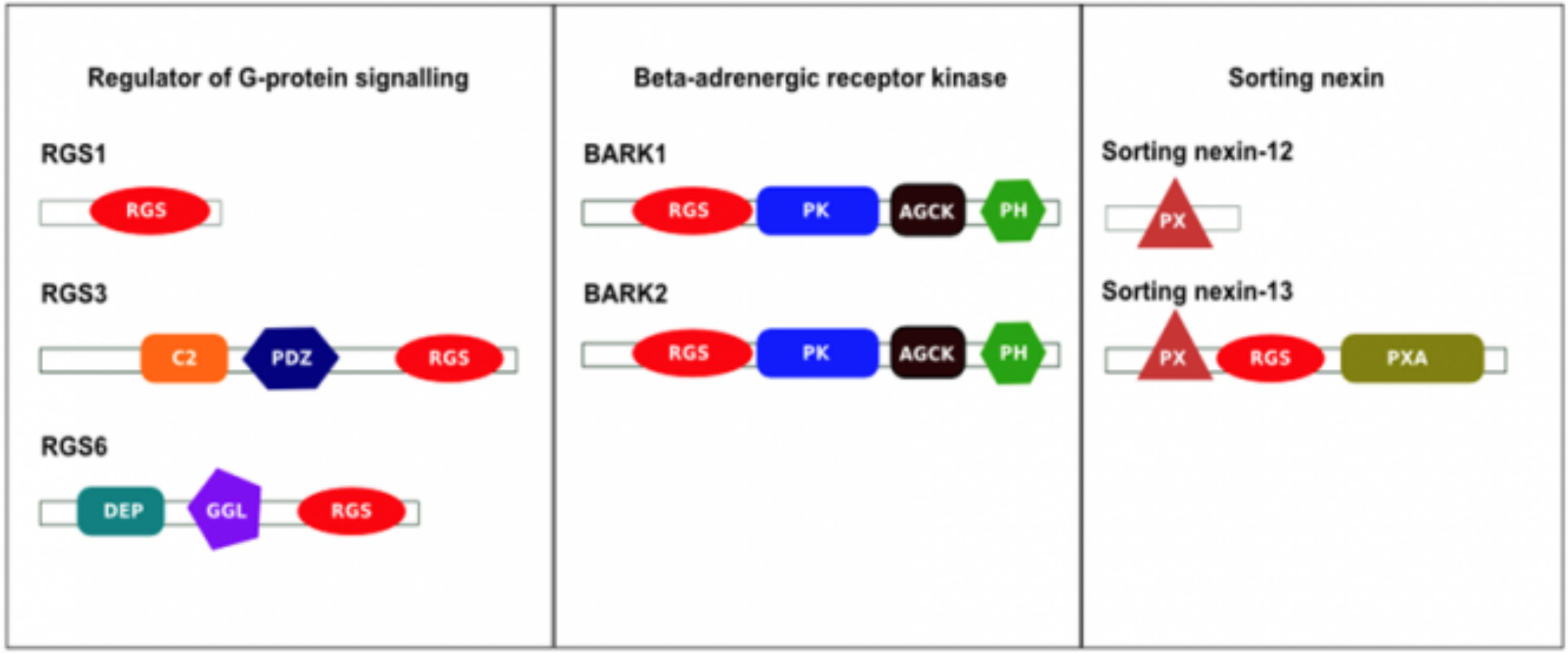
SH3-structure

Multidomain protein schematic



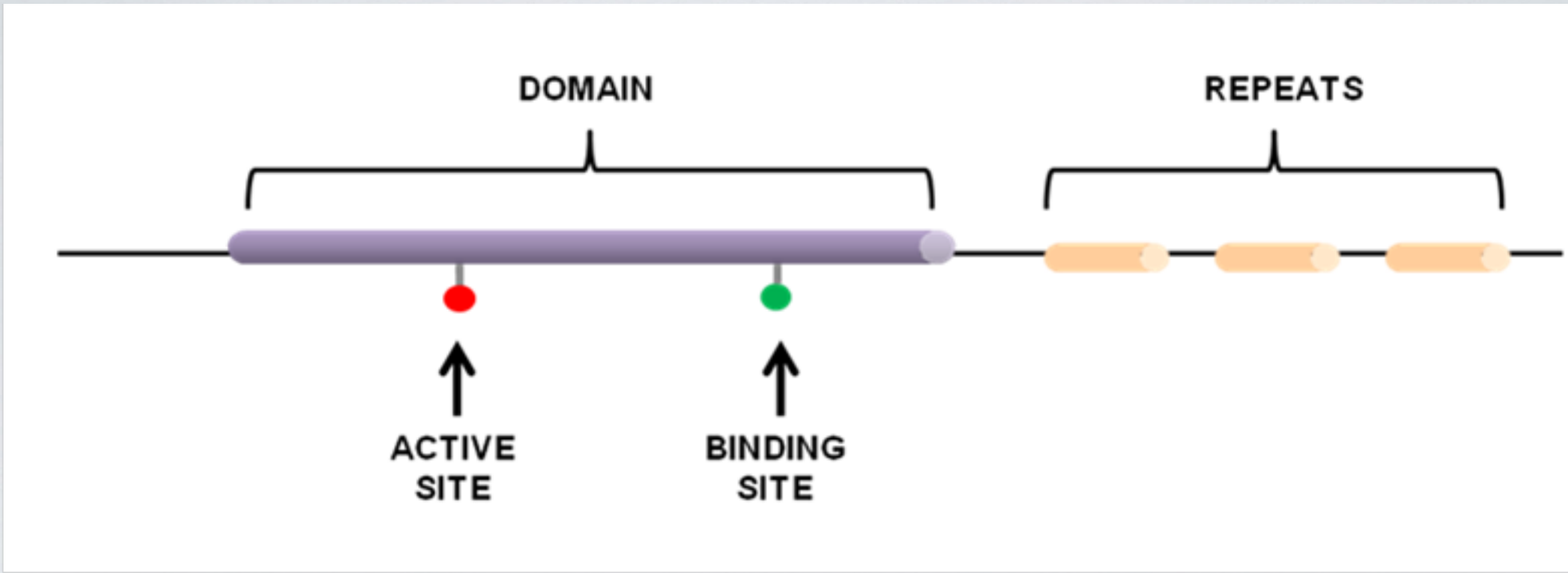
PROTEIN DOMAINS

These domains can have specific functions based



SIGNATURES OF DOMAINS CAN BE CHARACTERISTIC

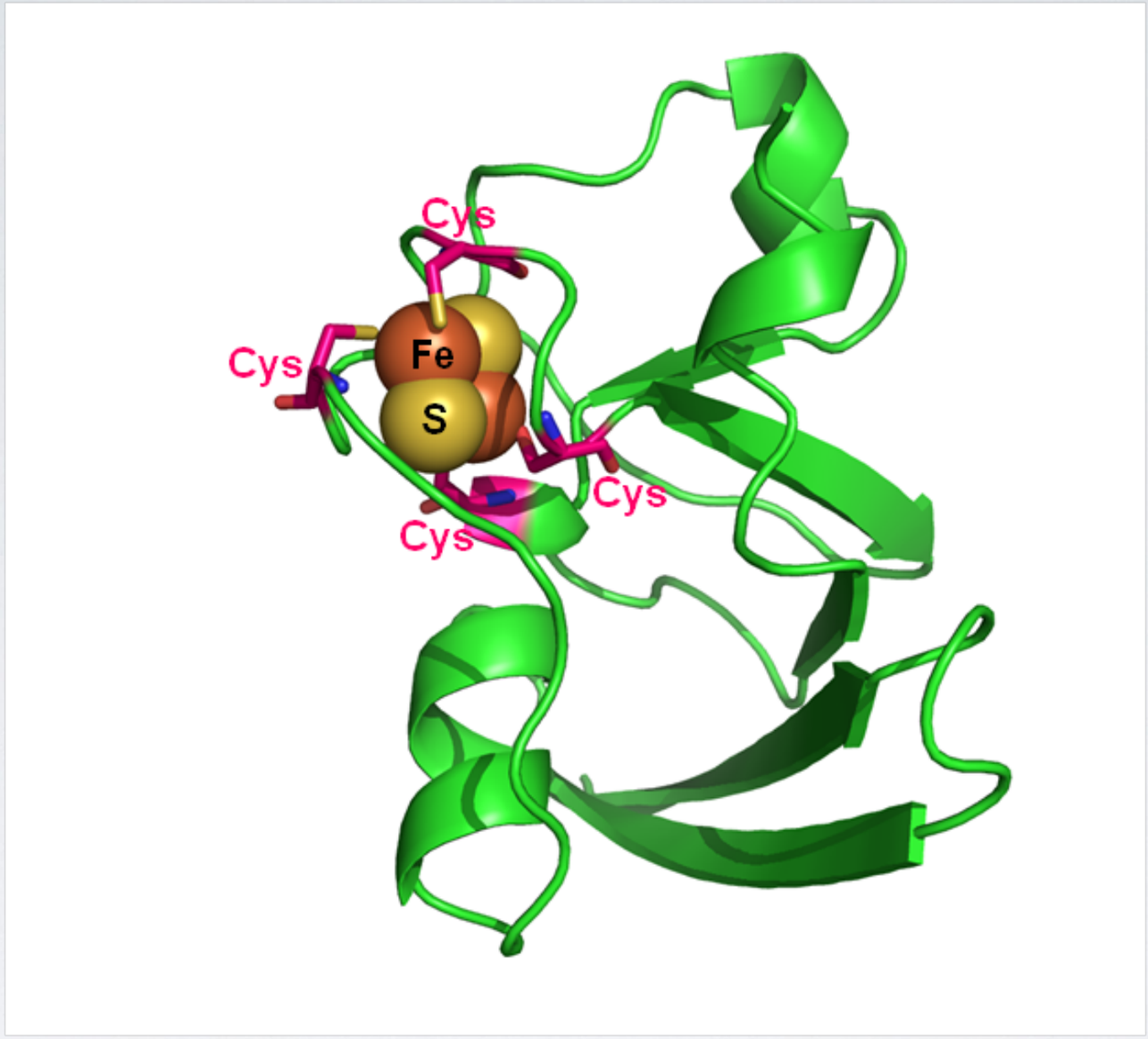
the order and domain content of a protein can be a signature of the type of function



SEQUENCE FEATURES

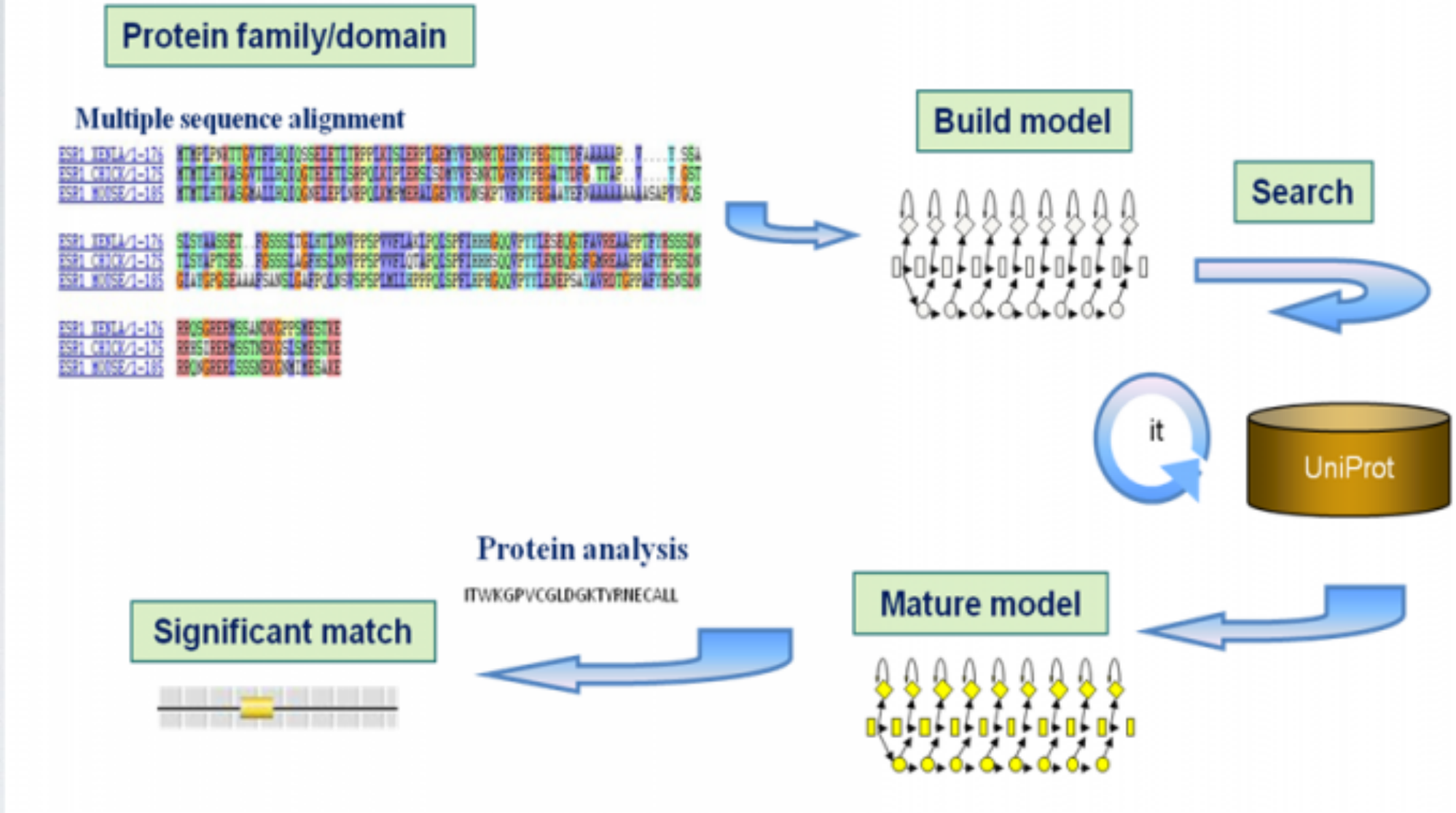
Sequence features are groups of amino acids which confer certain characteristics

- Could be active site with particular function in enzyme
- binding site for protein-DNA, protein-RNA, protein-protein interactions
- post translational modification site
- repeats within a protein (eg short motifs that repeat)



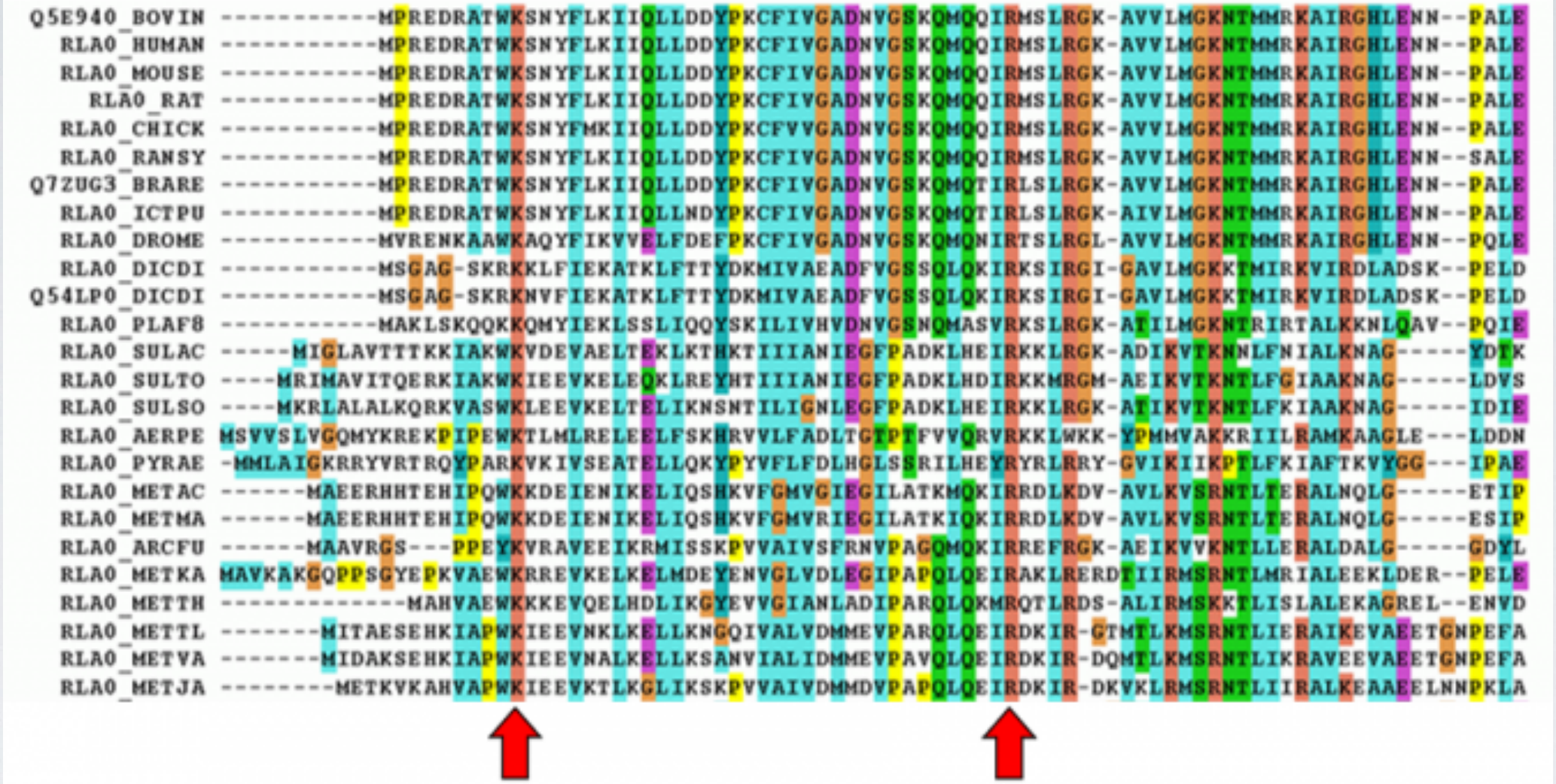
PROTEINS CLASSIFIED BY FEATURES

Sequence features like the type of iron- and sulfur-binding residues are used to classify a protein - this is a 2Fe-2S ferredoxin



CLASSIFYING PROTEINS

Start with known proteins which are similar and determined to be homologous (BLASTP)

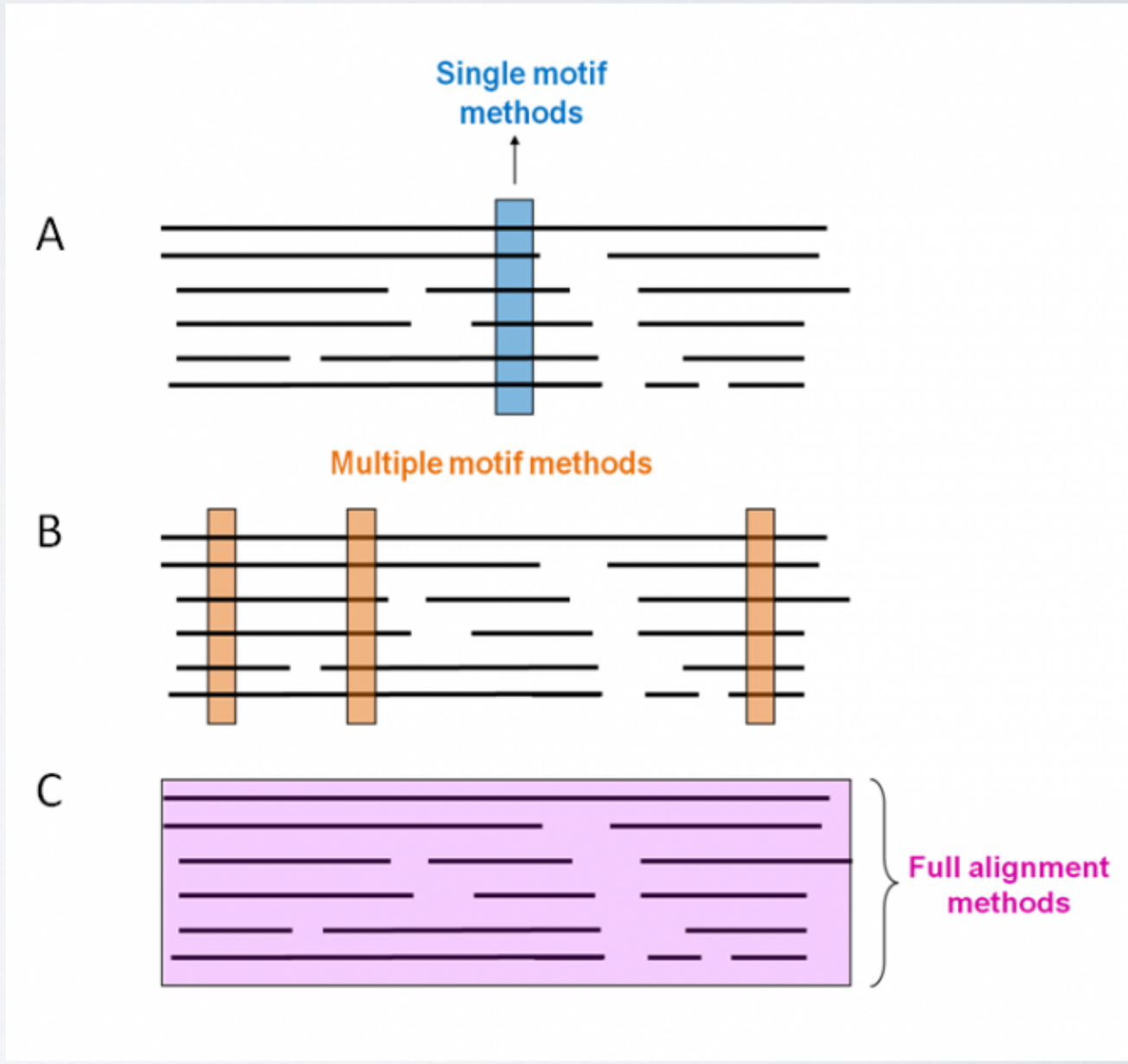


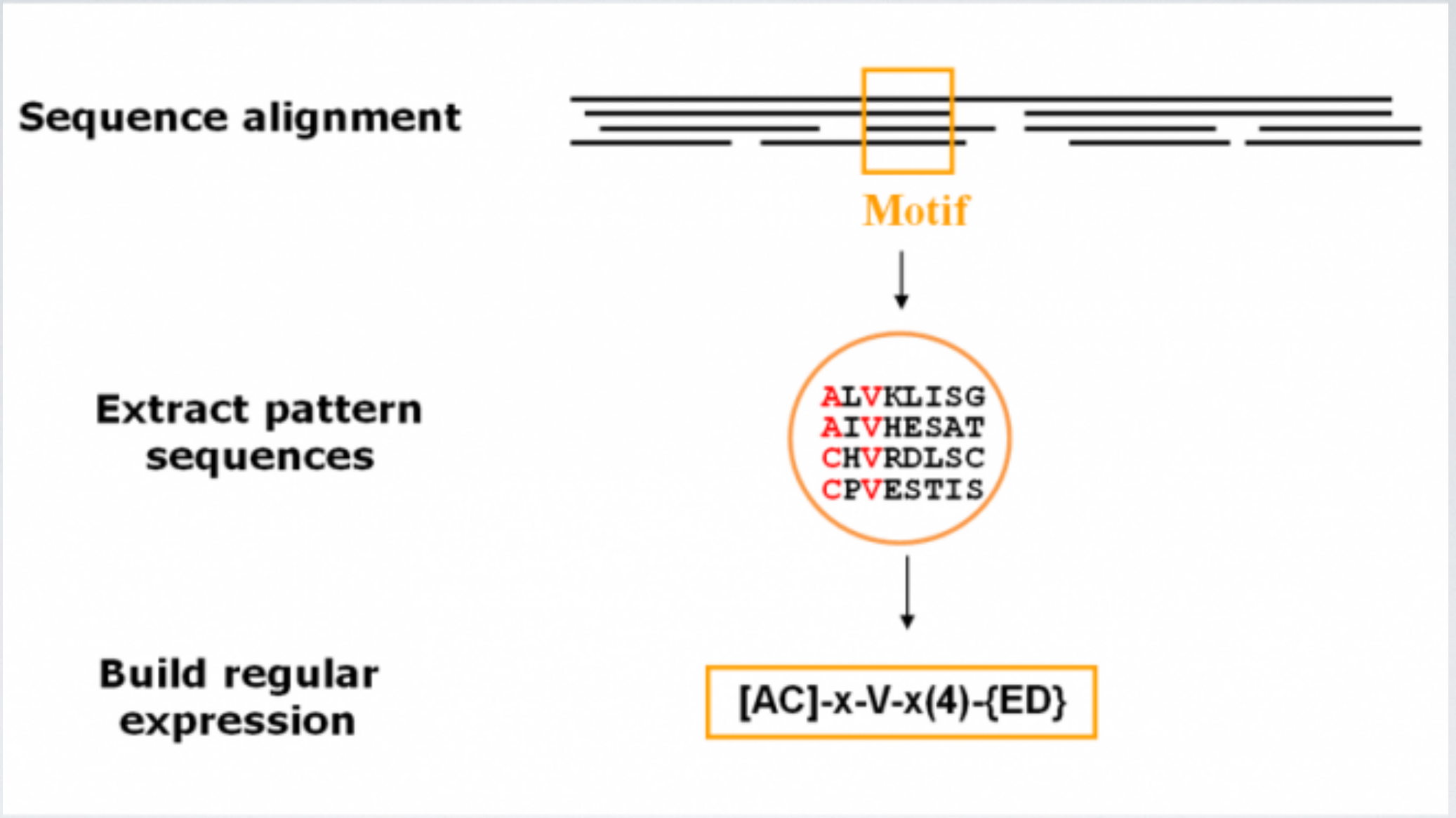
PROTEIN ALIGNMENT

Considering multiple sequences in the alignment so is more sensitive than BLAST. Only some residues are informative to classify the sequence
This is revealed through the multiple alignment.

HOW TO CLASSIFY

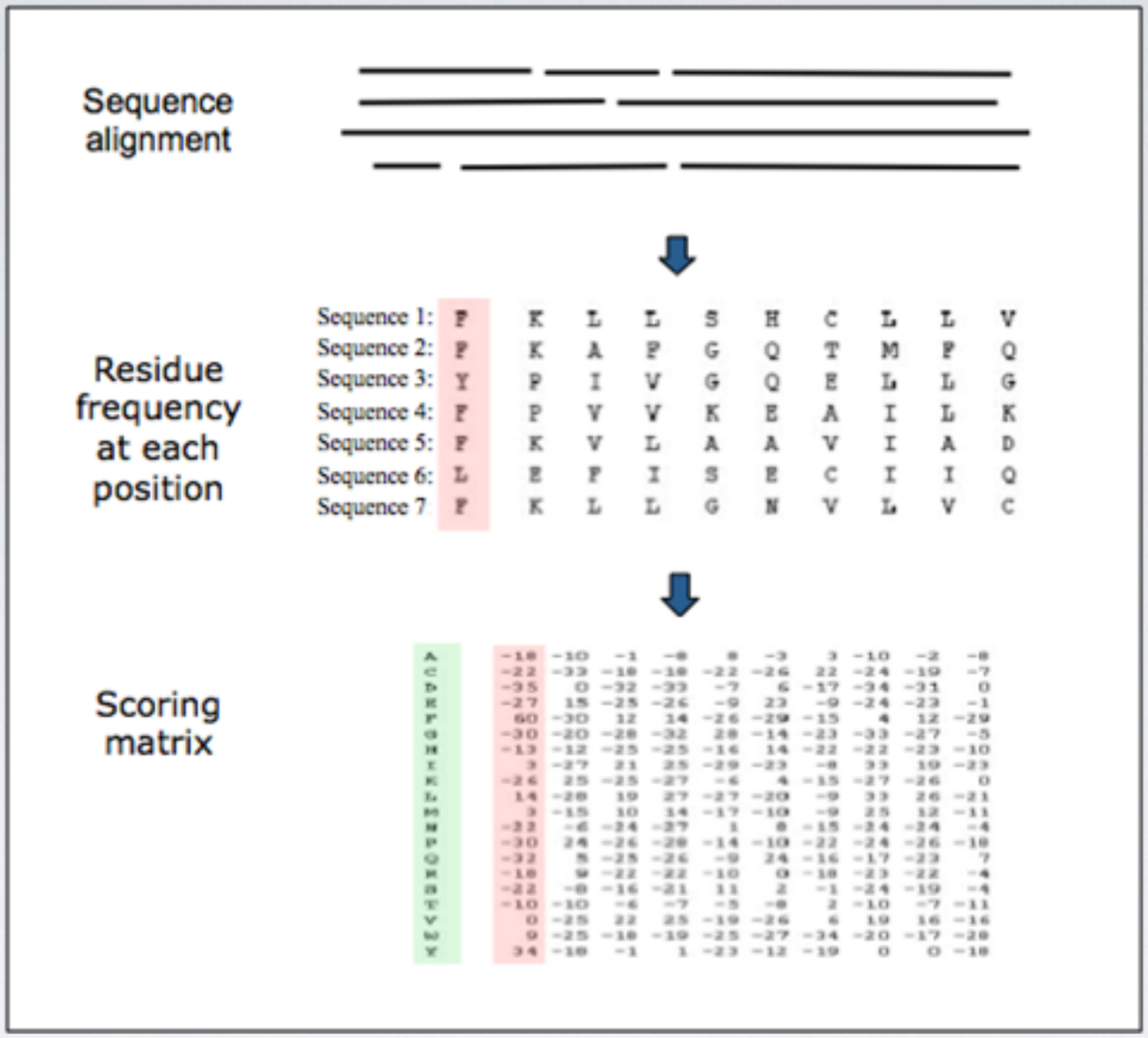
- Motif pattern
- Multiple Motifs - interspersed
  - Profile
  - Fingerprint
- Full alignment method with Hidden Markov Models





MOTIF DEFINED BY A PATTERN

Can write down the pattern with a series of letters and then logic called a Regular Expression



PROFILE

Created by converting a multiple alignment into a Position Specific Scoring Matrix - PSSM
Amino acids at each position in the alignment are scored according to the frequency with which they occur

SCORING A PROFILE

	1	2	3
A	**0.01**	0.04	0.02
G	0.02	0.02	0.03
C	0.02	0.93	0.93
T	0.95	0.01	0.02

Probability Matrix

`log2(0.01/0.25) = -4.6`

	1	2	3
A	**-4.6**	-2.6	-3.6
G	-3.6	-3.6	-3.0
C	-3.6	1.8	1.8
T	1.9	-4.6	-3.6

Position Weight Matrix

```
...TCC...
...TCG...
...TCC...
...TAC...
...GCC...
...TCC...
```

SCORING A PROFILE

	1	2	3
A	**0.01**	0.04	0.02
G	0.02	0.02	0.03
C	0.02	0.93	0.93
T	0.95	0.01	0.02

`log2(0.01/0.25) = -4.6`

Let's score the sequence
AGATCCTGCTCG

	1	2	3
A	**-4.6**	-2.6	-3.6
G	-3.6	-3.6	-3.0
C	-3.6	1.8	1.8
T	1.9	-4.6	-3.6

Position Weight Matrix

```
                        (A,1)  (G,2)  (A,3)
AGATCCTGCTCG  Score =  -4.6  + -3.6  + -3.6  =-11.8

                        (G,1)  (A,2)  (T,3)
AGATCCTGCTCG  Score =  -3.6  + -2.6  + -3.6  =-9.8

                        (T,1)  (C,2)  (C,3)
AGATCCTGCTCG  Score =   1.9  + 1.8  + 1.8    =5.5
```

Score above 0 is a good score!

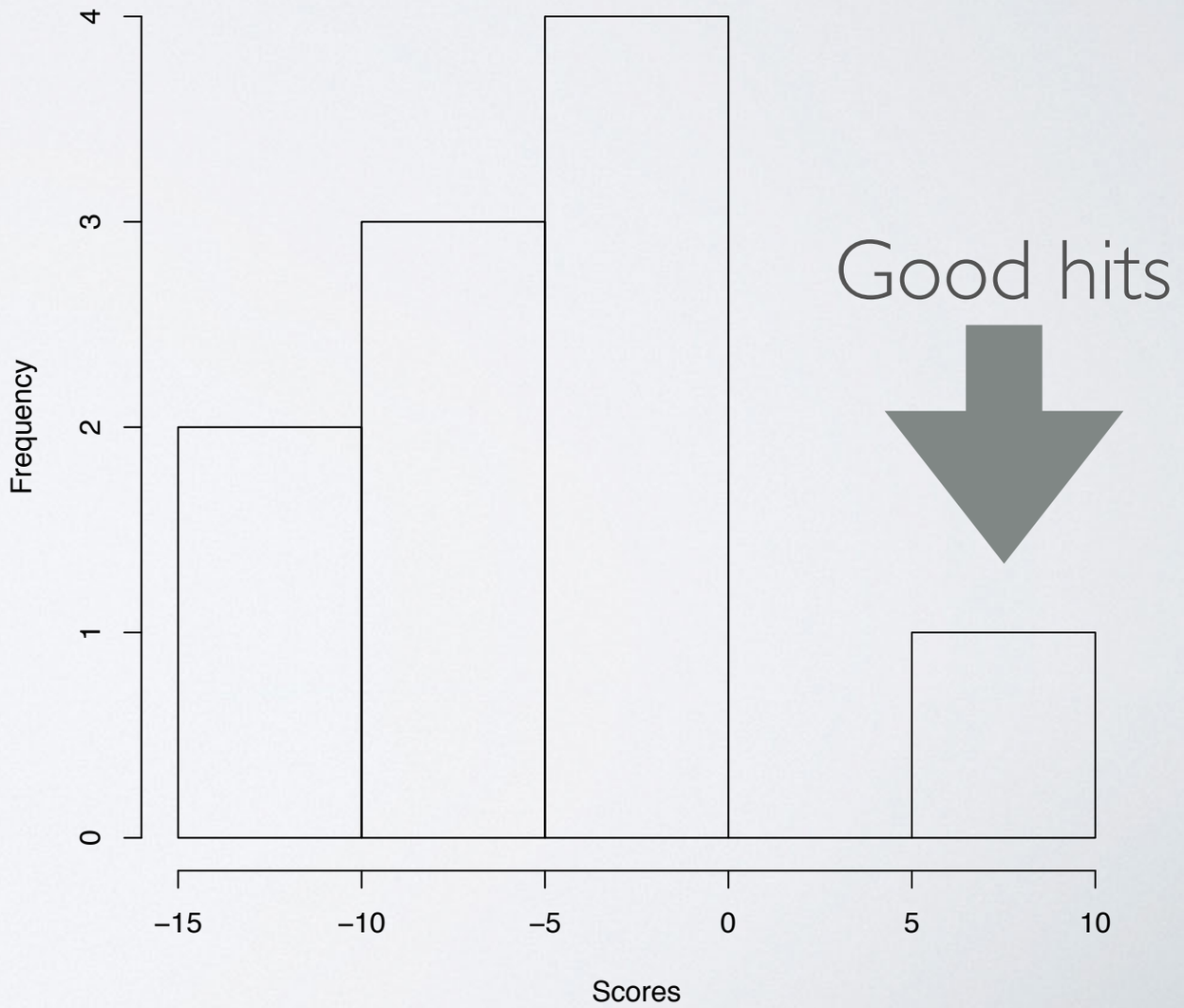
SCORING A PROFILE

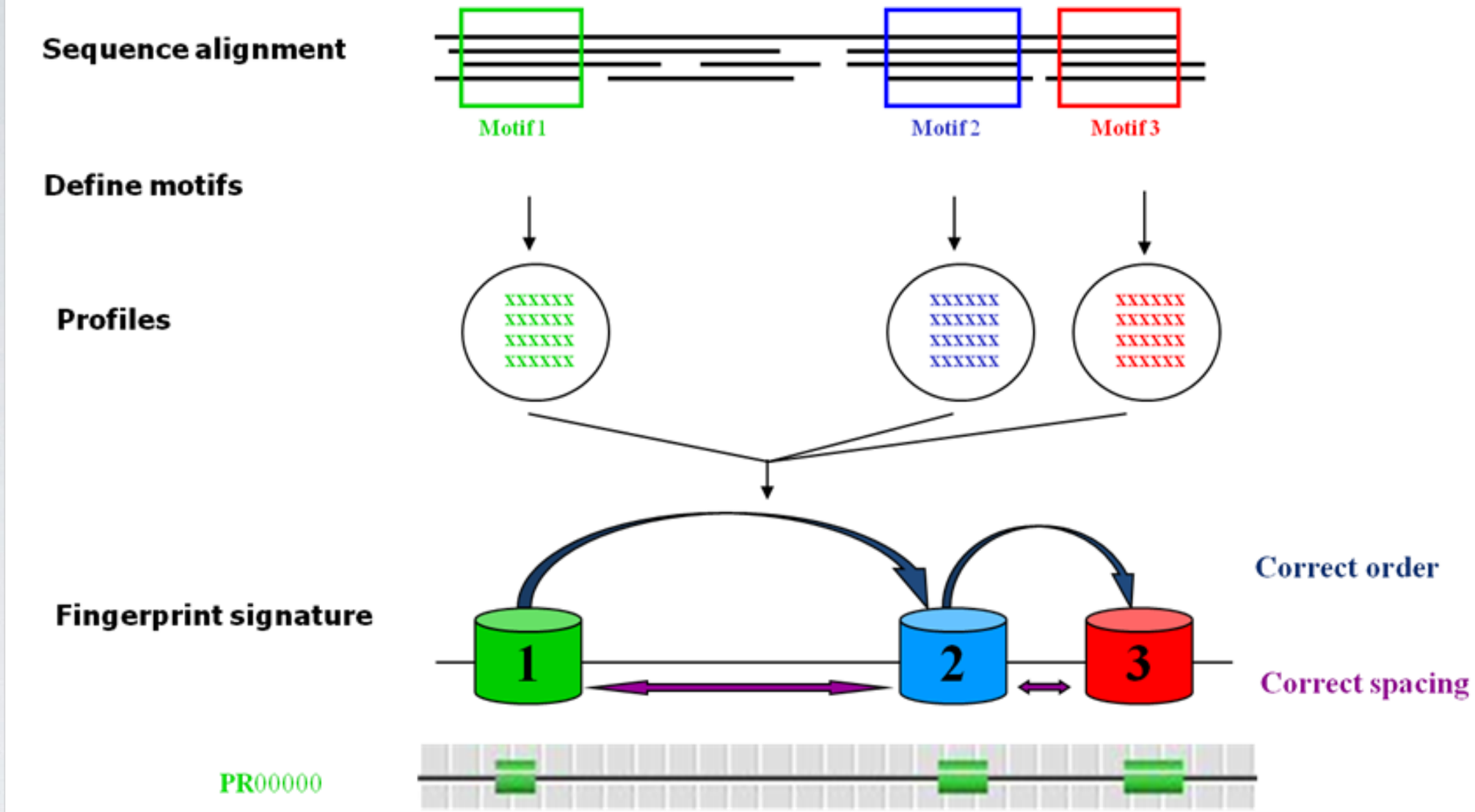
	1	2	3
A	-4.6	-2.6	-3.6
G	-3.6	-3.6	-3.0
C	-3.6	1.8	1.8
T	1.9	-4.6	-3.6

Consider the distribution of scores across the whole sequence to evaluate if there is a significance as well.

AGATCTTGCTCG

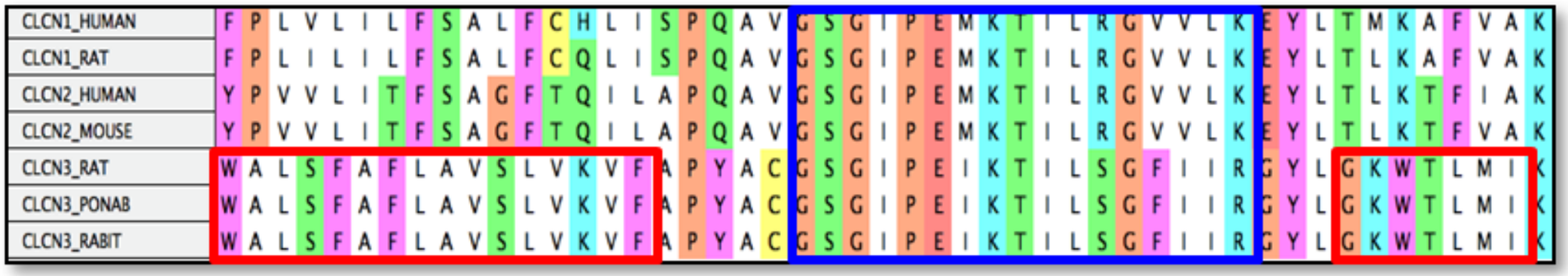
-11.8, -9.8, 5.5, etc





FINGERPRINT

Combination of motif or profile into a signature



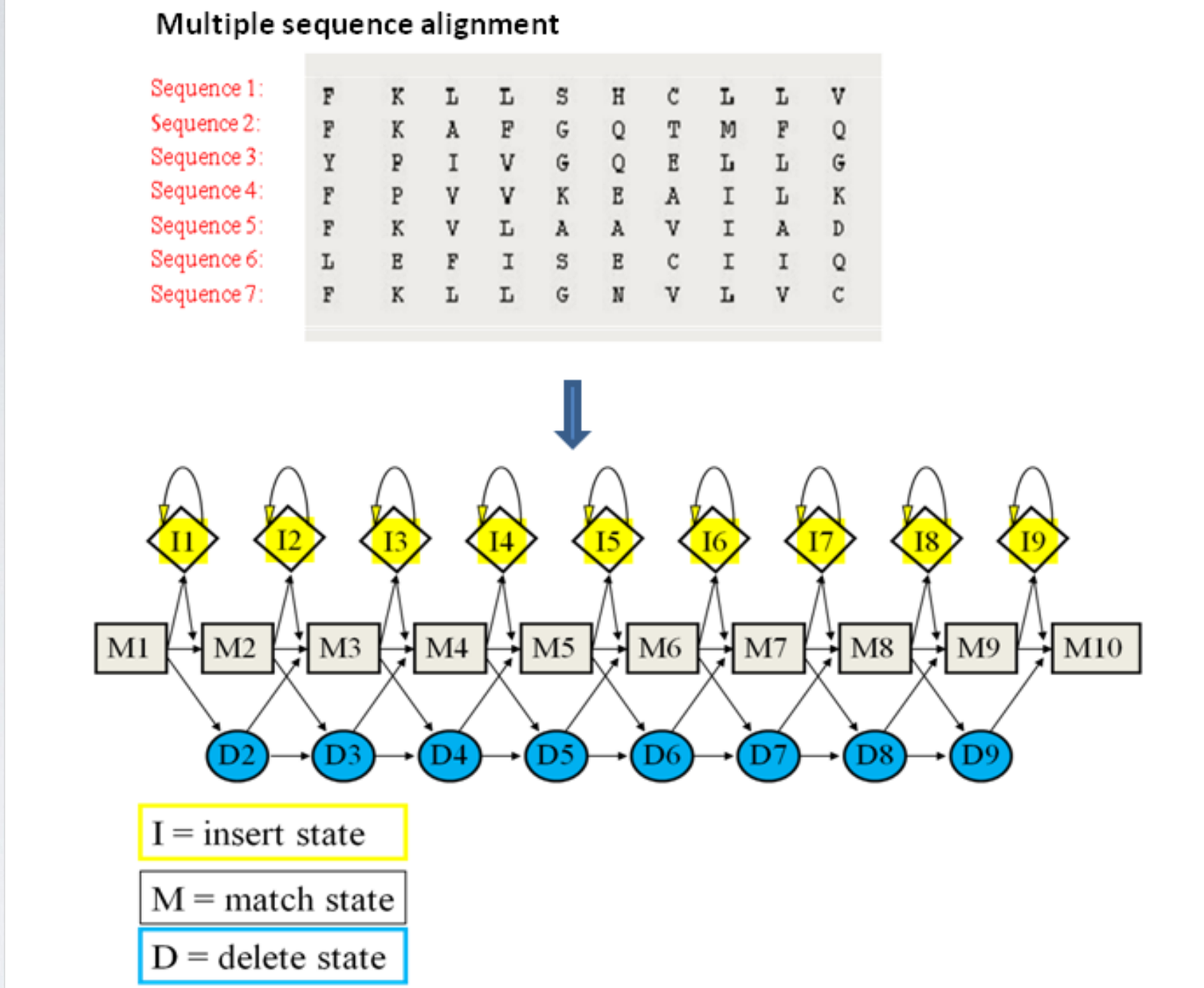
Amino acids relatively well conserved across all chloride channel protein family members

Amino acids uniquely conserved in chloride channel protein 3 subfamily members

WHY ARE FINGERPRINTS USEFUL?

Can capture and model small differences between sub-families can capture the individual differences.

In this example of a chloride channel protein family identified by blue box a subset can be further classified into channel 3 - subfamily



HIDDEN MARKOV MODELS

Can model the alignment by capturing insertion and deletions and probabilistically score sequence similarity.

DATABASES OF PROTEIN DOMAINS

- Pfam - Protein https://pfam.xfam.org/; Panther - http://www.pantherdb.org/; SMART - http://smart.embl-heidelberg.de/
  - databases of HMMs of domains
- Prosite - https://prosite.expasy.org/ - motifs
- Interpro - https://www.ebi.ac.uk/interpro/
  - HMMs + Profiles + fingerprints

PROTEINS TO FUNCTION

- Together these domains and classifications can provide ways to link an unknown sequence to function

- If domains that makeup the protein are known can make guess about the protein function even if a homolog does not have a known function in other species

- Domains can be shared among many types of proteins

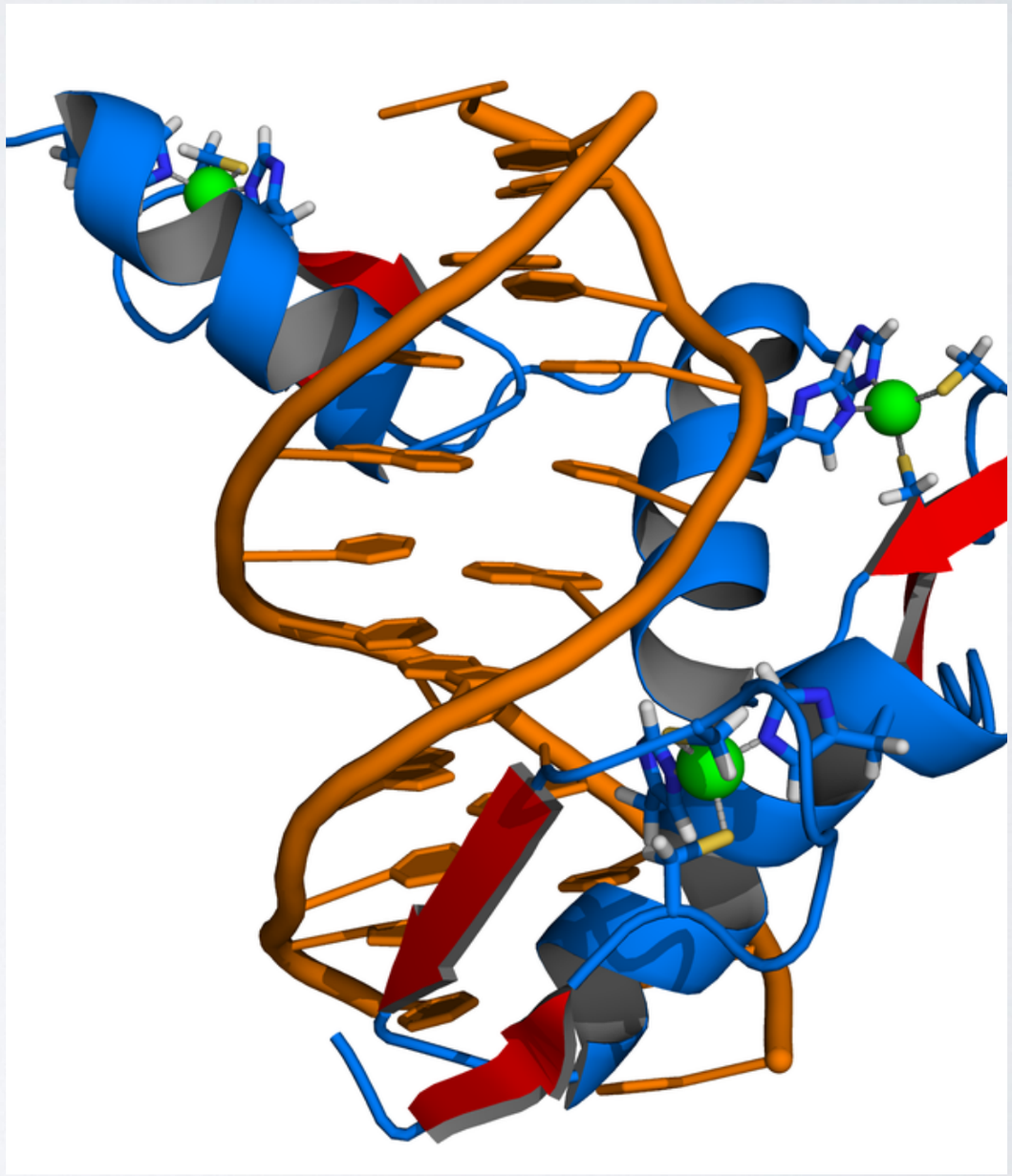
- Shuffling of domains and motifs can provide new function

SOME PROTEIN DOMAINS

- Zinc-finger - https://en.wikipedia.org/wiki/Zinc_finger

- Typically bind DNA, protein, RNA

- Often part of transcription factors



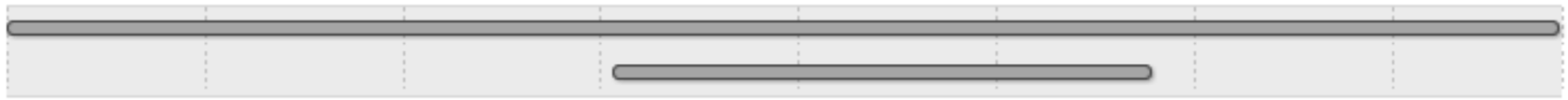
https://fungidb.org/fungidb/app/record/gene/YOR172W

Detailed signature matches

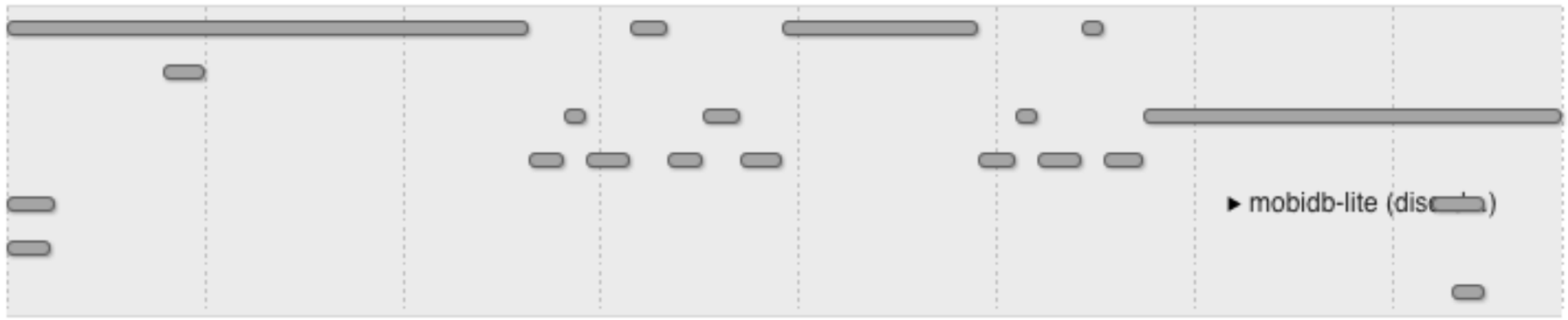
H IPR036864 Zn(2)-C6 fungal-type DNA-binding domain superfamily

D IPR001138 Zn(2)-C6 fungal-type DNA-binding domain

? no IPR Unintegrated signatures



Other features



Residue annotation

TAKE HOME POINTS

- Proteins can be classified by their similarity
- Parts of proteins can further be found to be conserved and functional
- Motifs, Profiles, Fingerprints, HMMs
- Domains discovery can be used to assign function