Public Databases in Health and Life Sciences

“the potential to translate big data into big discovery”

Academic year 2023-2024

Public Databases in Health and Life Sciences

Topic 4. Secondary and derived protein databases.

- Databases containing data derived from the analysis or treatment of primary data.
- Sequence based prediction of protein function.
- Database of protein motif, domains, families and functional sites.
- Conserved domain databases and domain classification.
- Databases for protein structural domains.

Practical sessions #4 and #5

Primary vs. Secondary Databases

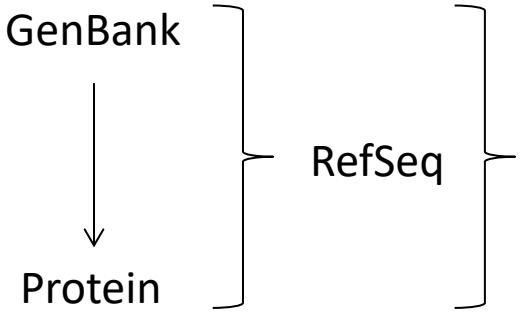
Primary db

- Contains experimentally-derived data such as nucleotide sequences, protein sequences or macromolecular structures.
- Experimental results are submitted directly into the database by researchers, and the data are essentially archival in nature.

Secondary/Derived db

- Data derived from the analysis or treatment of primary data.

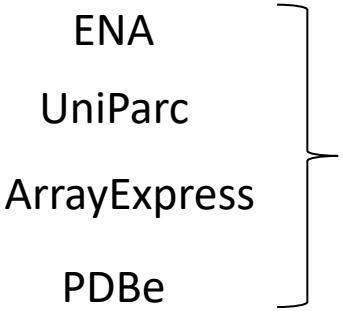
NCBI: Derivative Databases



Nucleotide derived
Example: GENE, Genomes, etc
Protein-derived
Example: CDD

https://www.ncbi.nlm.nih.gov/

EMBL-EBI: Derivative Databases



Nucleotide derived
Example: Ensembl, Expresion Atlas
Protein-derived
Example: Pfam, InterPro, IntAct, UniProtKB?

https://www.ebi.ac.uk/

Determining protein function

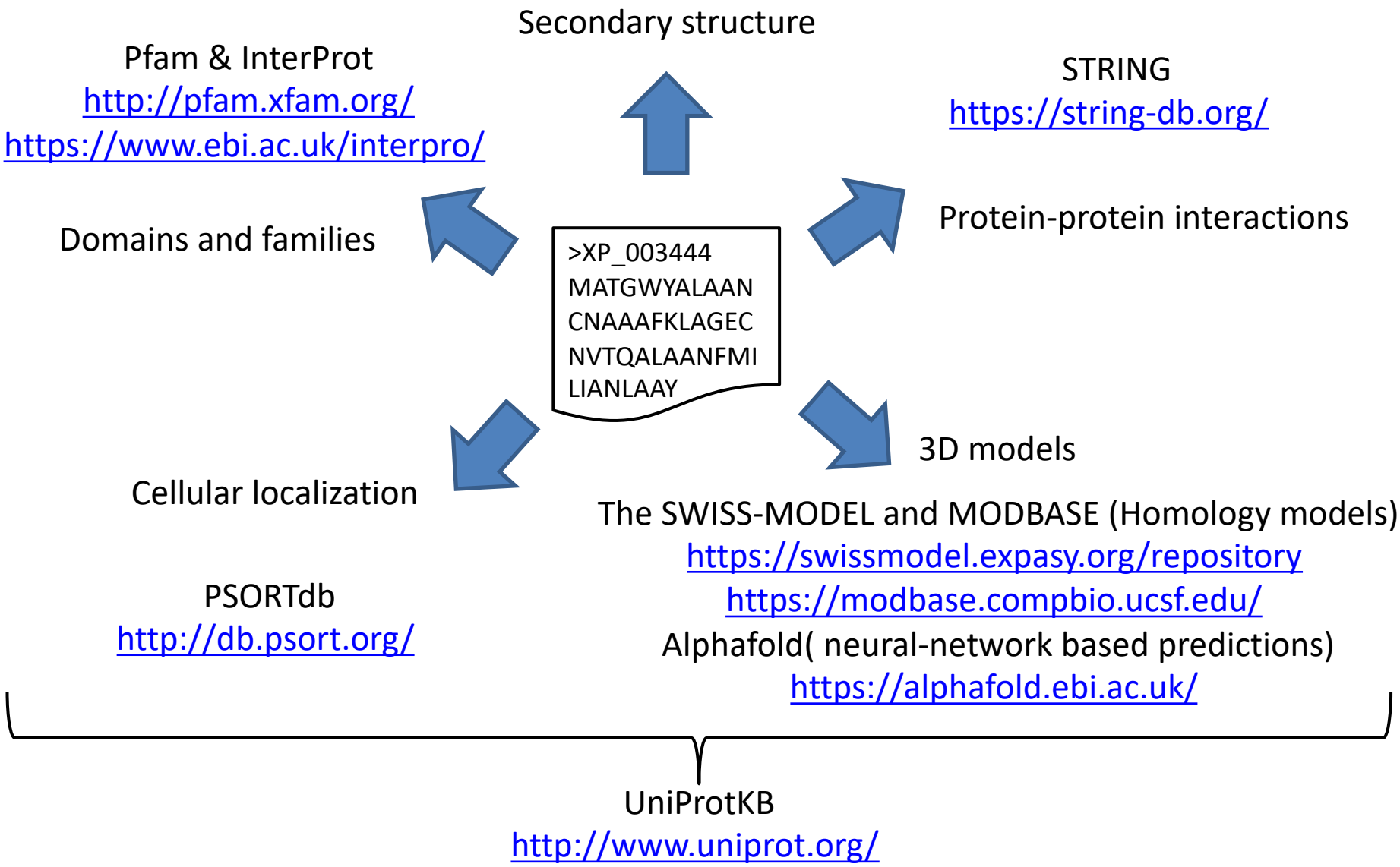
- Experimentally determining function (time consuming and expensive)
    - Functional assays
    - 3D structure
    - Co-localization
    - Protein-protein interaction (experimental approaches)
    - Microarray experiments

- **Predictive methods**
    - Transfer of function based on homology
        - BLAST
    - Based on the amino acid sequence (predicted protein secondary structure, transmembrane helices, subcellular localization, and posttranslational modifications)
    - 3D modeling
    - Motif and Domains

  Sequence Based Predictions at http://www.expasy.org/tools/

- Combination of all above

Protein function: predictive methods and associated databases



Protein classification

Proteins can be classified into different groups based on:

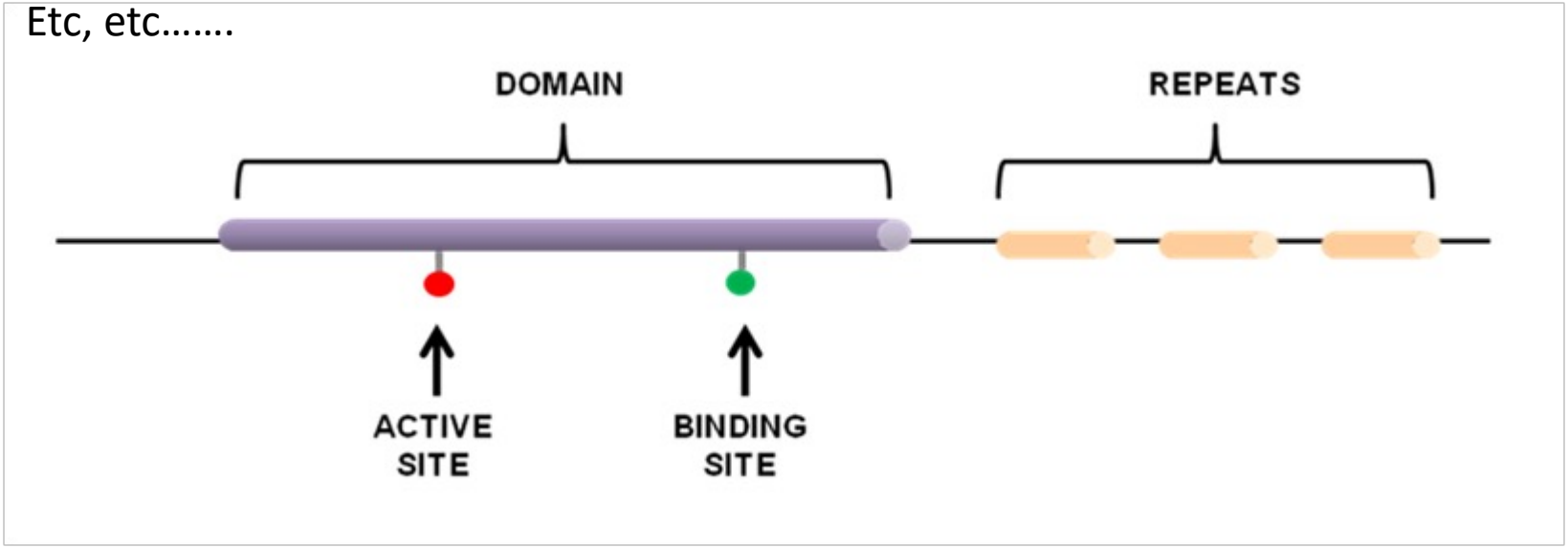
- the **SEQUENCE FEATURES** they possess
  - Groups of amino acids that confer certain characteristics to a protein
- the **DOMAINS** and **MOTIFS** they contain
  - Distinct functional and/or structural units in a protein.
- the **FAMILIES** to which they belong
  - A group of proteins that share a common evolutionary origin.

Protein classification: An introduction to EMBL-EBI resources
At https://www.ebi.ac.uk/training/online/

What are sequence features?

Sequence features are groups of amino acids that confer certain characteristics upon a protein and may be important for its overall function. Such features include:

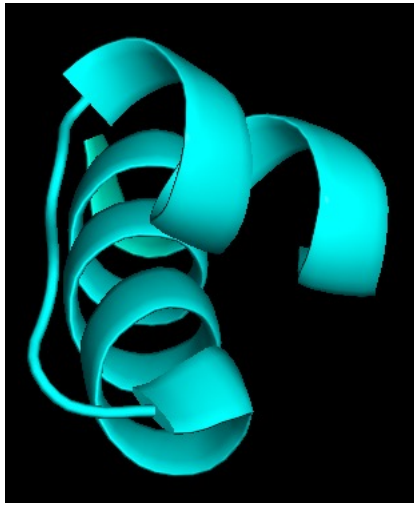
- Active sites
- Binding sites
- Post-translational modification sites
- Short amino acid repeats
- Signal peptides
- Etc, etc…….



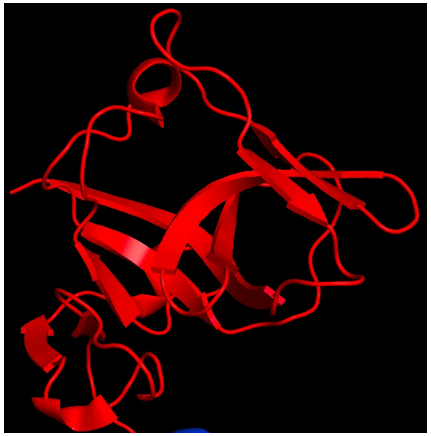
Protein classification: An introduction to EMBL-EBI resources
At https://www.ebi.ac.uk/training/online/

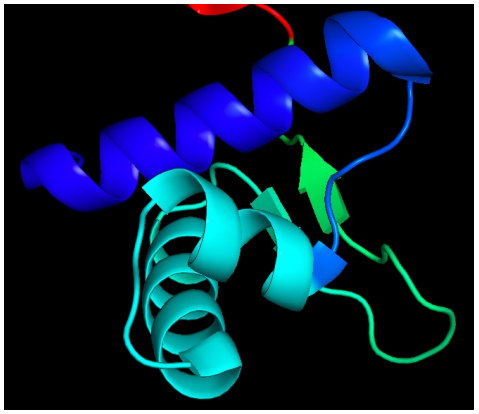
Motifs (AKA: fold) and Domains are units within the protein

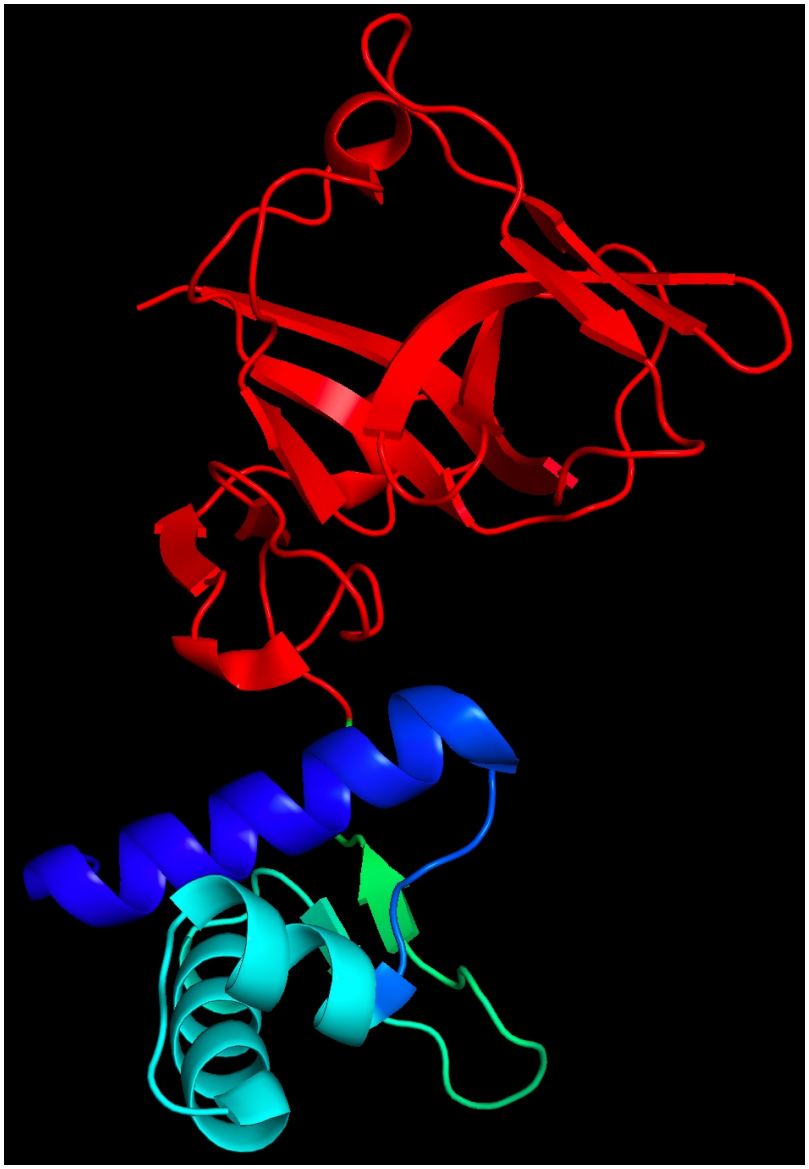
Motif



Domains







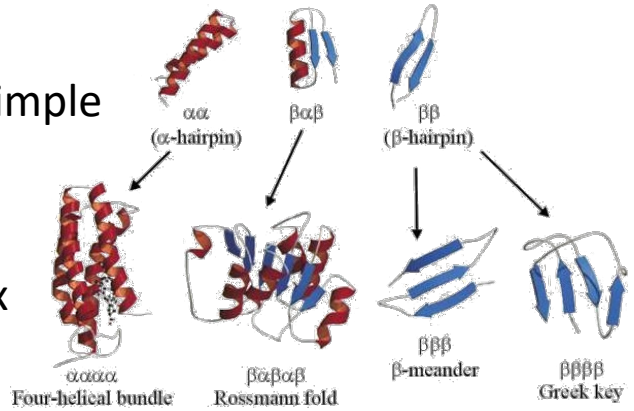
Motif AKA Fold (and rarely Supersecondary Structure):

Motif or fold is a recognizable folding pattern involving two or more elements of secondary structure and the connection(s) between them.
Generally spiking, we use the word “motif” when we talk about simple elements and the word fold for more complex elements

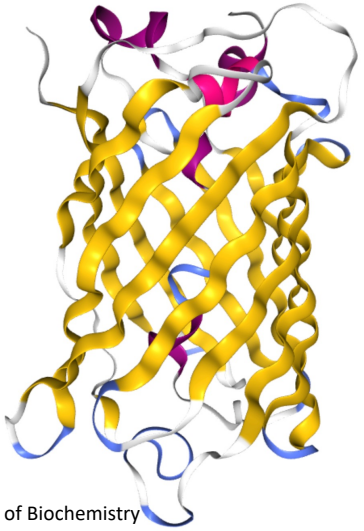
https://qph.fs.quoracdn.net/main-qimg-028bcdcfaba956d3df3cea5316688054-c

They can be very simple

A bit more complex



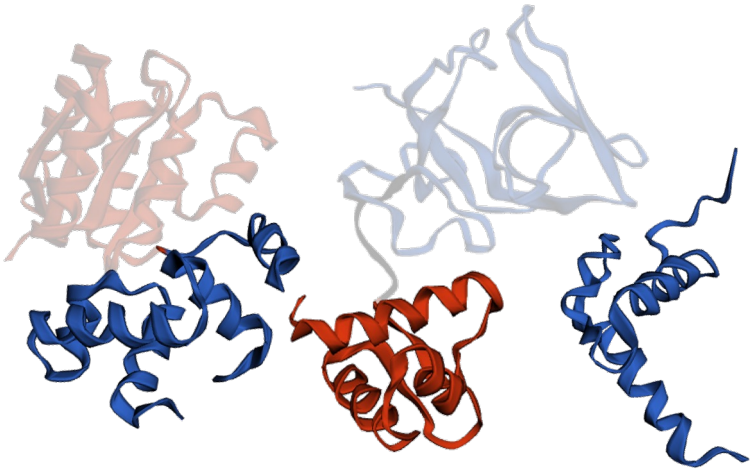
Much more complex



A motif may or may not be independently stable and it may or may not be related to a specific function.

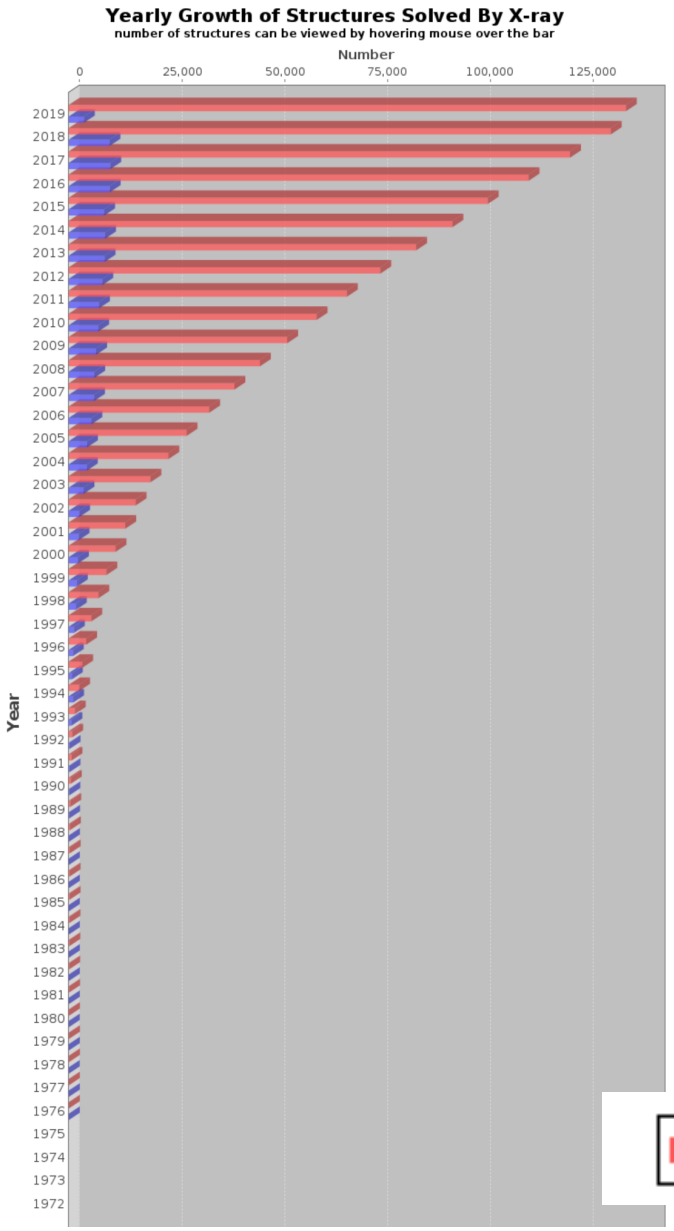
A motif **is not** a hierarchical structural element falling between secondary and tertiary structure.

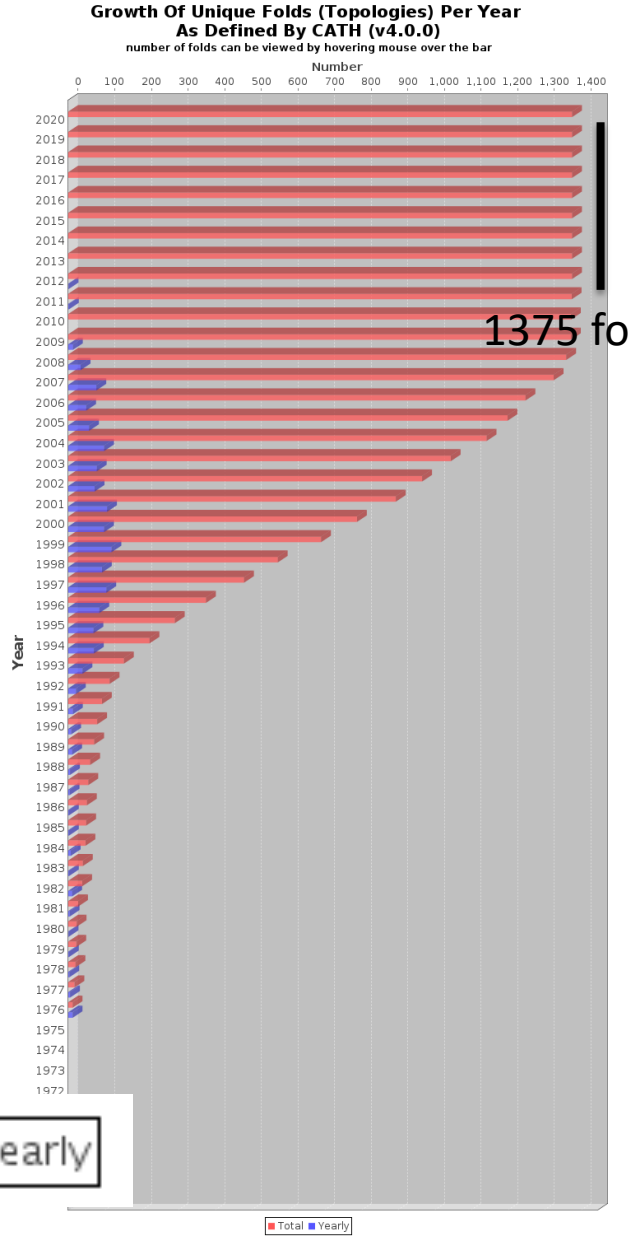
The same motif can have several types (which do not necessary coincide with their classification in Databases). See below three types of helix-turn-helix fold



Some text has been borrowed from Lehninger Principles of Biochemistry

The number of folds is finite





1375 folds

Motifs (AKA: fold) and Domains are units within the protein

- Motifs (AKA: “fold”) are used to describe what can be called a supersecondary structure of the protein. This is the combination of alpha-helices and beta-structures in the protein. It may not necessarily have a biological functions nor be independently stable.

- Domains are distinct functional and/or structural units in a protein. Usually, they are responsible for a particular function or interaction, contributing to the overall roll of the protein. They can fold by themselves and are stable once they are folded. They are supposed to be individual proteins in origin that while evolving they have joined others and are expressed toghether in order to be more efficient.

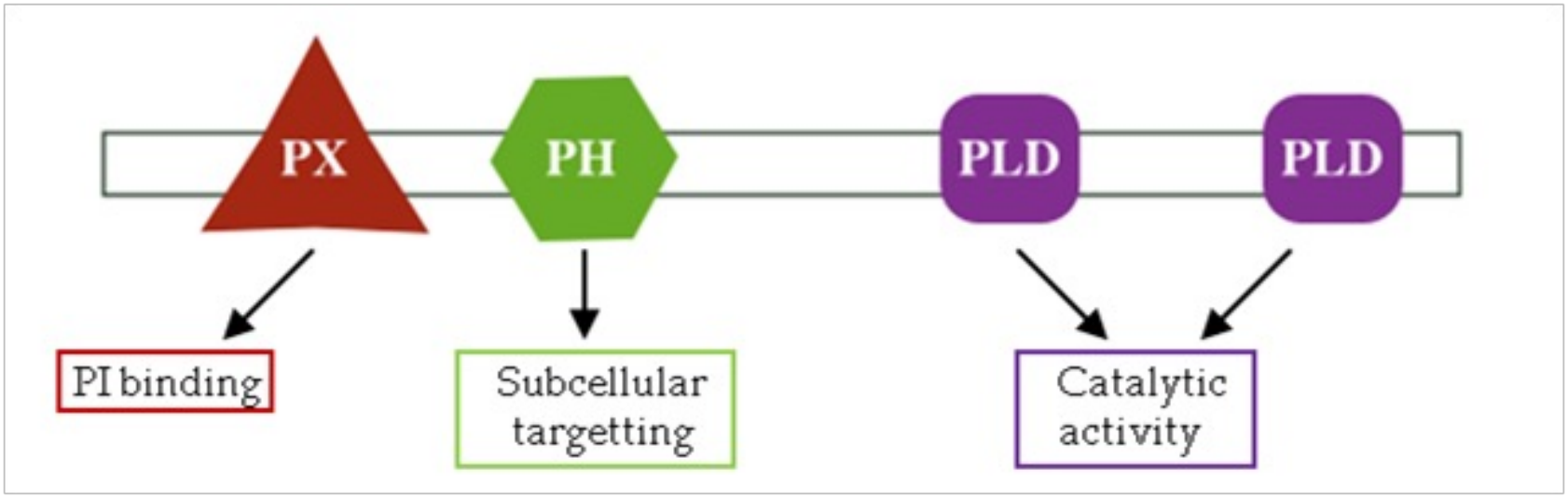
- Domains are a sequence of amino acids that are conserved between many different organisms.

- They are often used to hypothesize the overall function of uncharacterized proteins because the sequence of amino acids usually carries a similar function.

- They are especially important in classifying protein families.

Domains are distinct functional and/or structural units in a protein

Proteins can be composed of several domains.

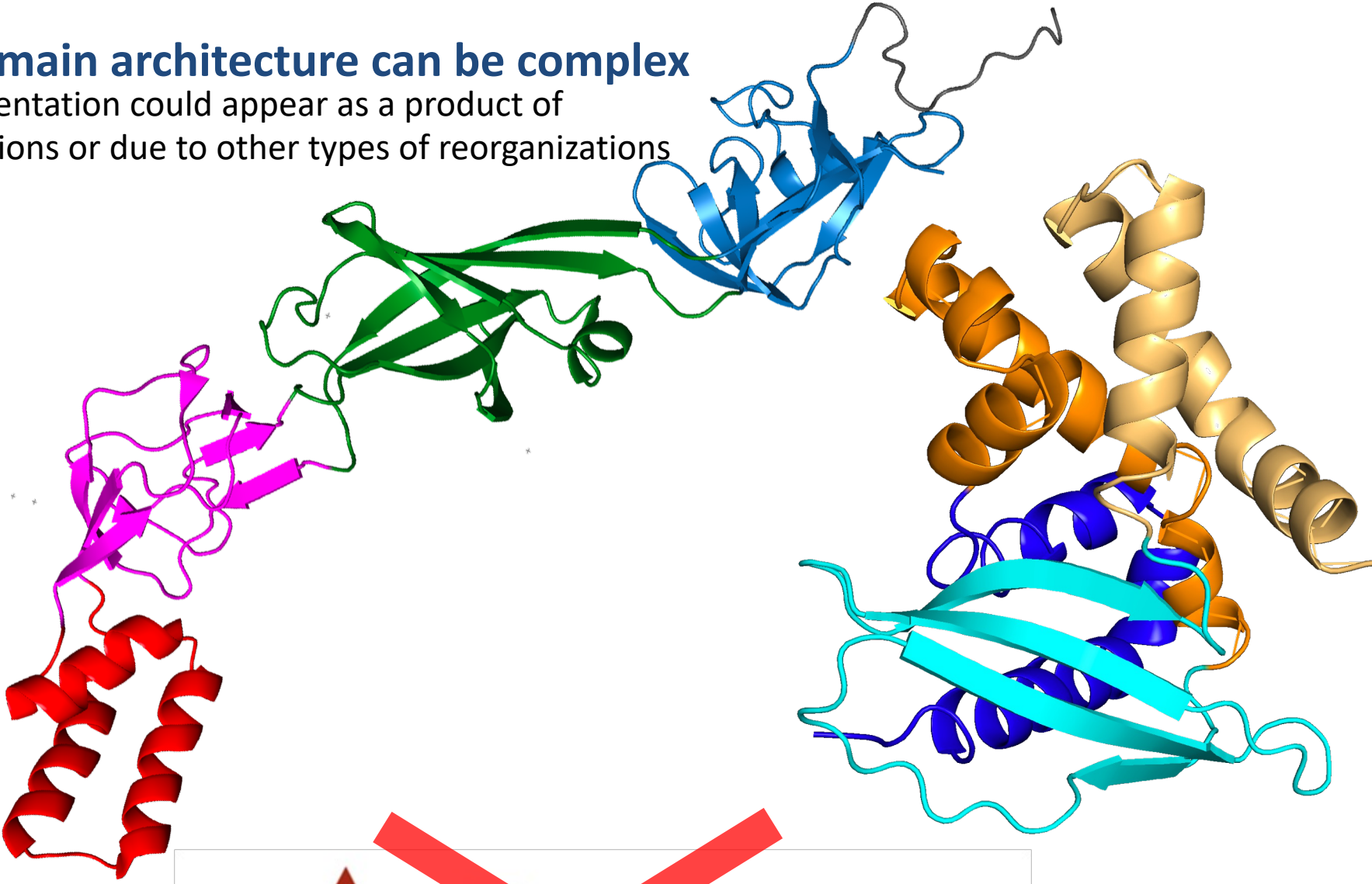


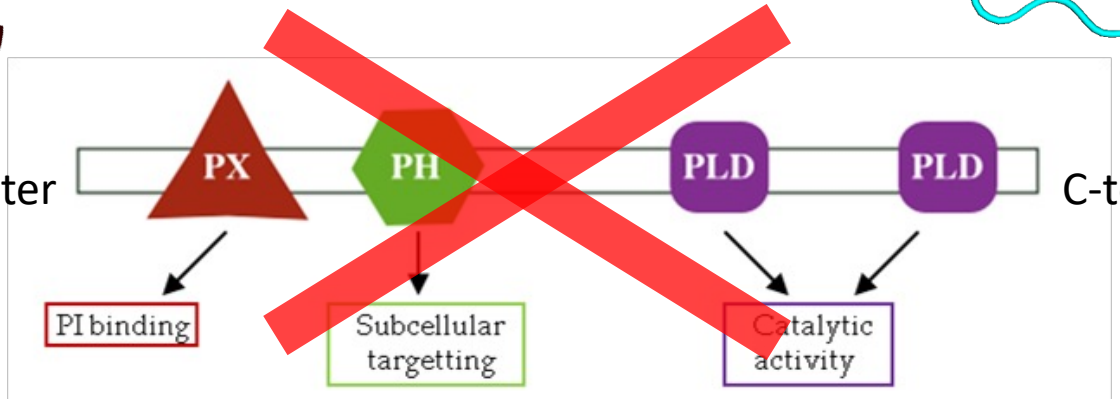
Phospholipase D1 Domains

Protein classification: An introduction to EMBL-EBI resources
At https://www.ebi.ac.uk/training/online/

Domain architecture can be complex

Segmentation could appear as a product of insertions or due to other types of reorganizations





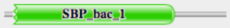
Sequence vs structural domains:

Periplasmic oligogalacturonide binding protein from Yersinia enterocolitica

Sequence search results

Show the detailed description of this results page.

We found **1** Pfam-A match to your search sequence (**all** significant)



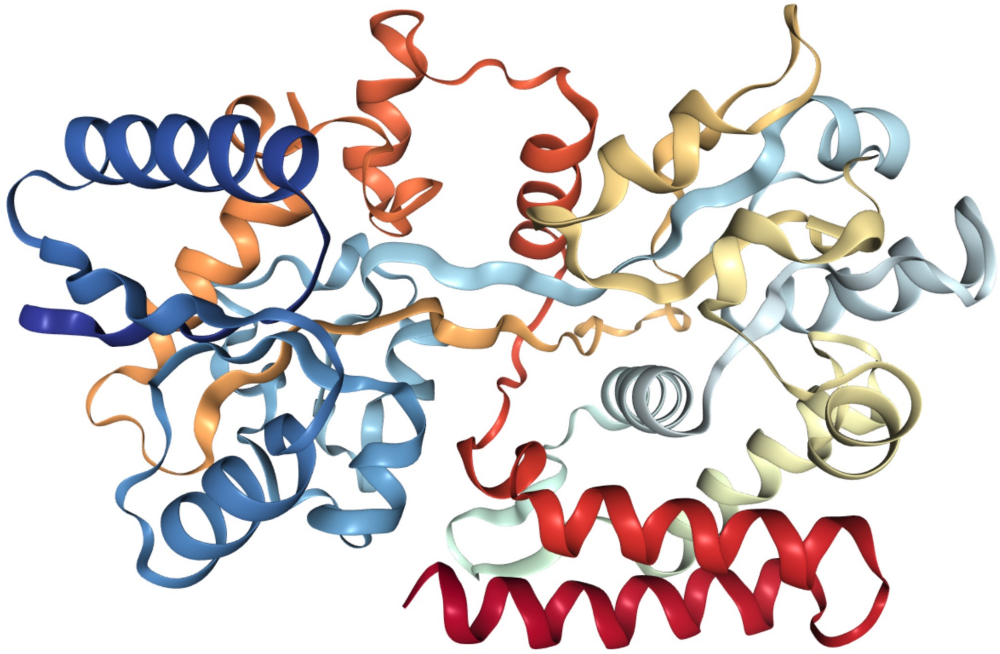
Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
SBP_bac_1	Bacterial extracellular solute-binding p ...	Family	CL0177	19	314	25	311	**13**	**312**	315	87.5	1.5e-24	n/a	Show



PDB Information

PDB	
Method	
Host Organism	
Gene Source	
Primary Citation	
Header	Sugar-Binding Protein
Released	2007-03-10
Resolution	2.200
CATH Insert Date	24 May, 2007

PDB Images (4)

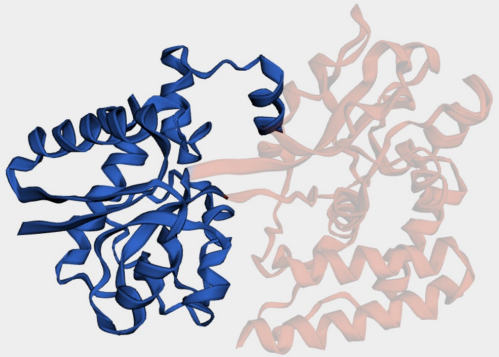
PDB Prints

PDB Chains (1)

Chain ID	Date inserted into CATH	CATH Status
A	24 May, 2007	Chopped

CATH Domains (2)

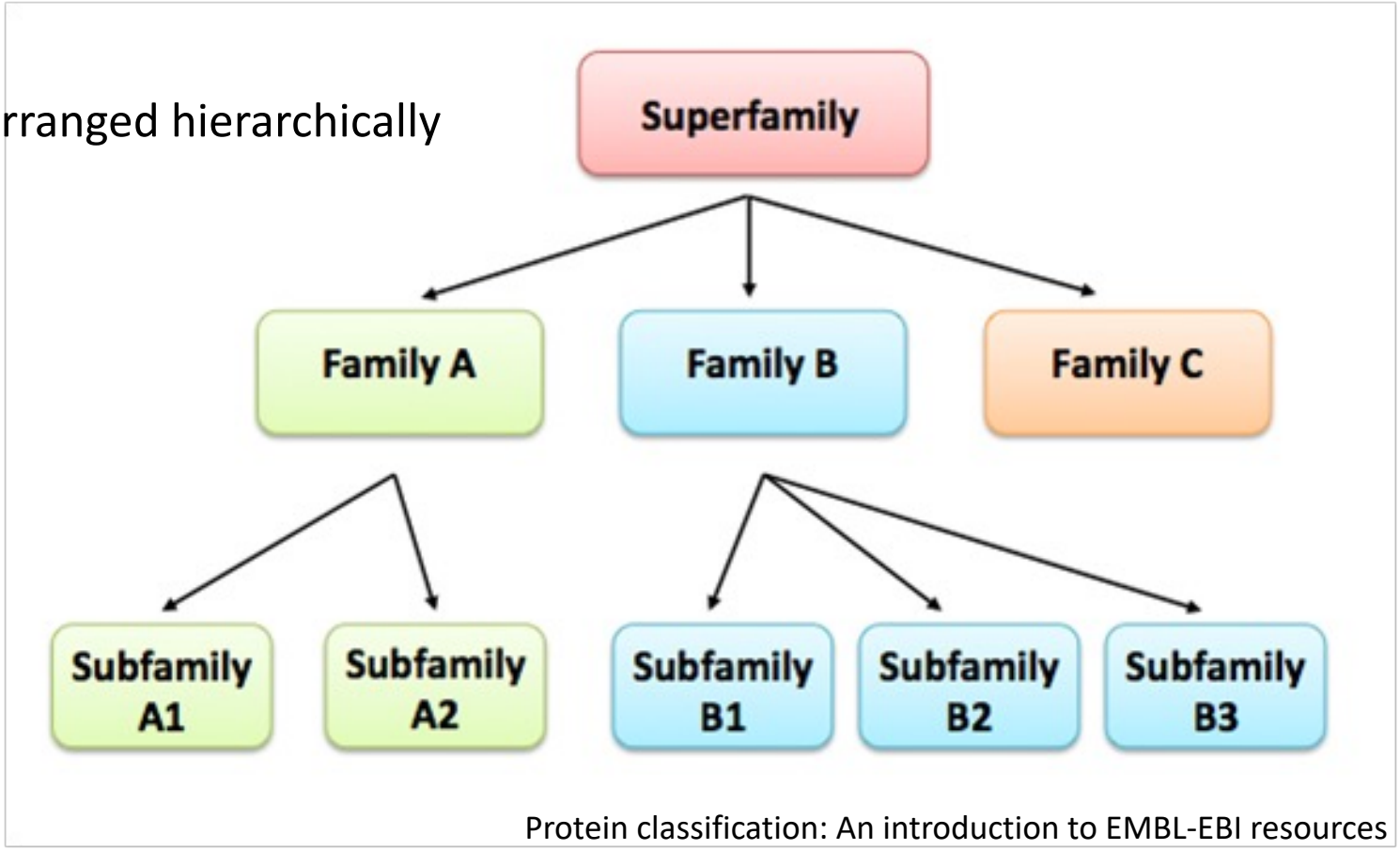
Domain ID	Date inserted into CATH	Superfamily	CATH Status
2uvgA01	09 Sep, 2013	3.40.190.10	Assigned
2uvgA02	09 Sep, 2013	3.40.190.10	Assigned



2uvg (A)

Protein families

- A group of proteins that share a common evolutionary origin.
- Proteins are sometimes assigned to families by virtue of the domain(s) they contain.
- They are arranged hierarchically



Protein classification: An introduction to EMBL-EBI resources
At https://www.ebi.ac.uk/training/online/

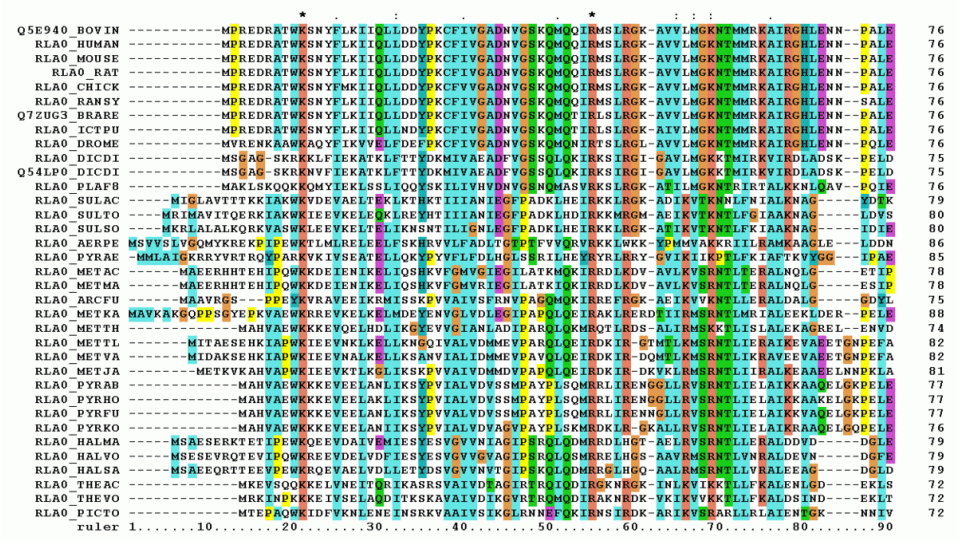
Domain Classification

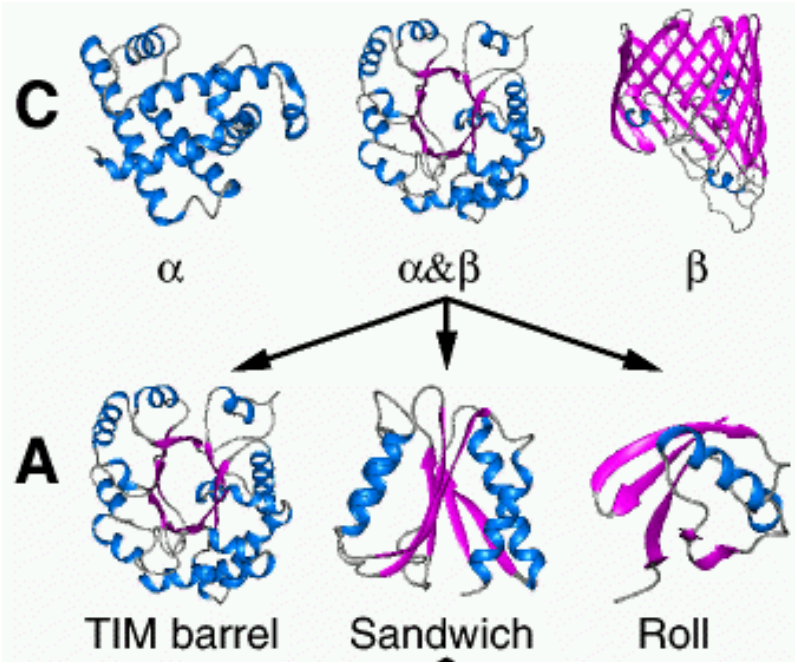
Sequence based (alignments)

Families and superfamilies

Structure based

The arrange of secondary structures





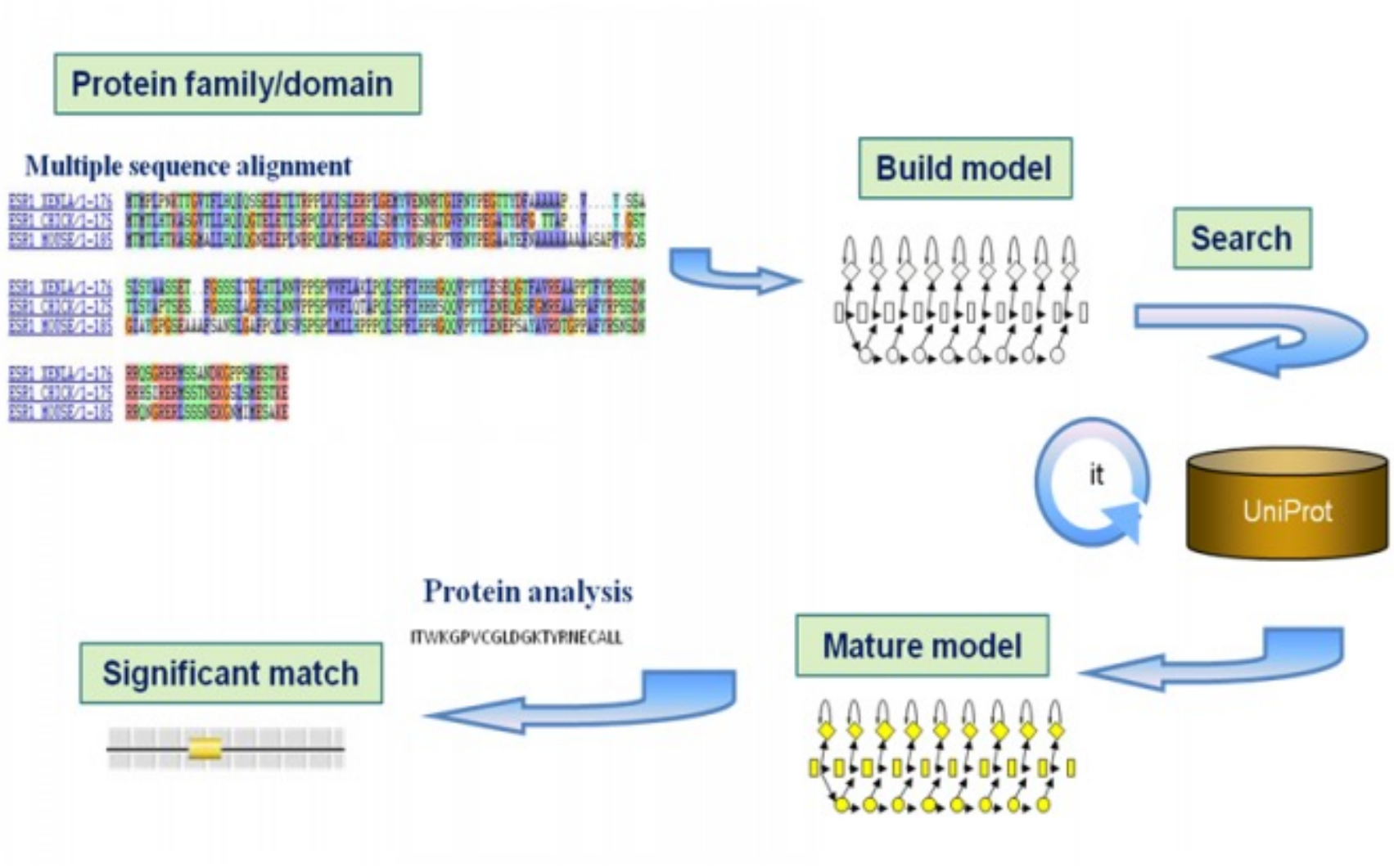
What are protein signatures?

Signature sequences are contiguous patterns of amino acids that are associated with a particular structure or function in proteins.

There are different types of signatures, built using different computational approaches. These include:

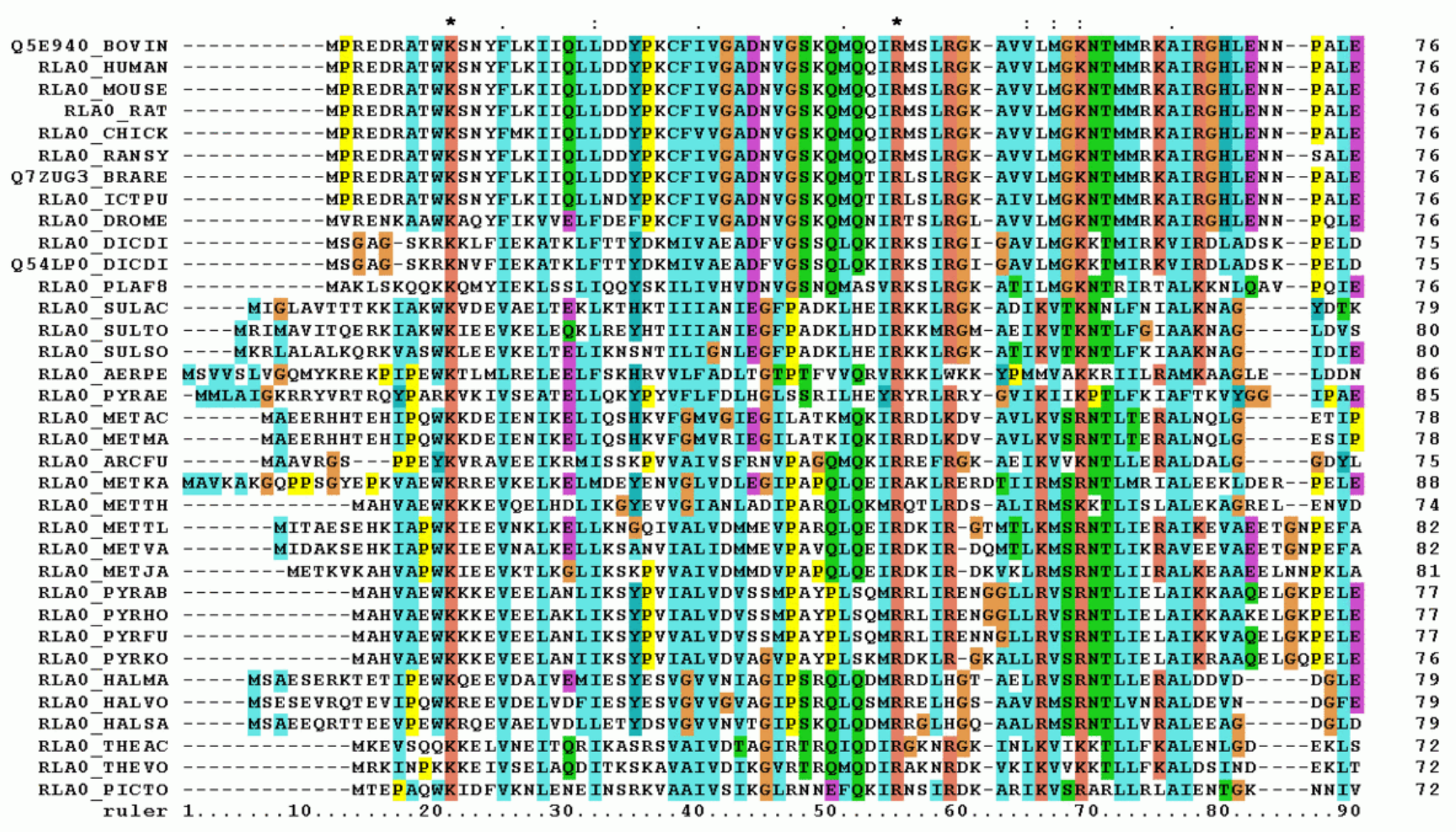
- patterns (motifs)
- profiles (pfam is a profile database)
- fingerprints
- hidden Markov models (HMMs) (a type of profiles)

The process of building a protein signature



Protein classification: An introduction to EMBL-EBI resources
At https://www.ebi.ac.uk/training/online/

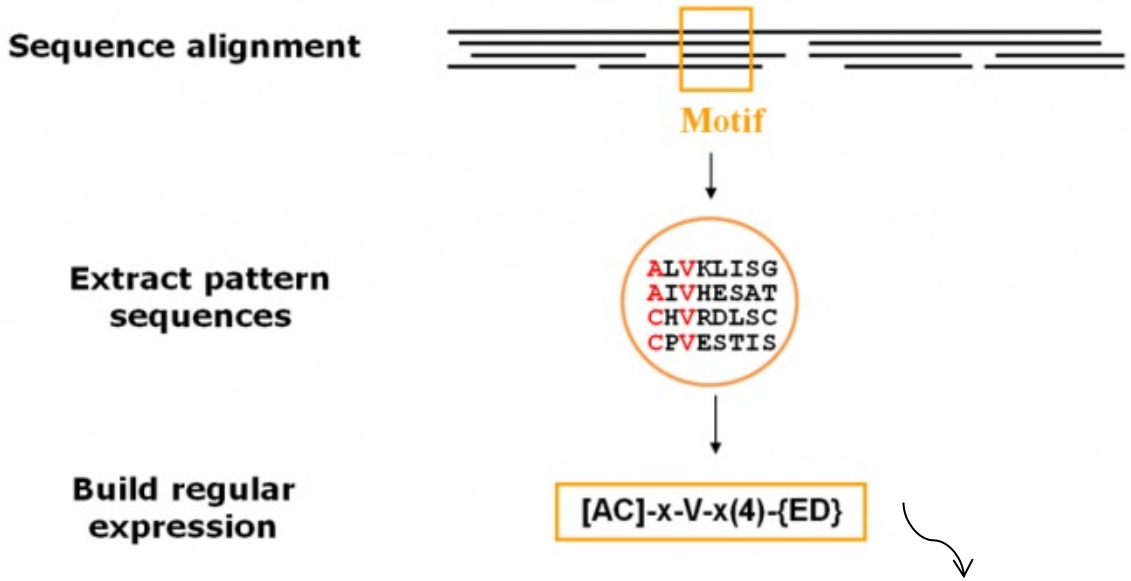
Multiple sequence alignments can provide us with valuable information for protein classification since they allow us to identify the (often few) amino acid residues that are conserved in distantly related proteins.



Protein classification: An introduction to EMBL-EBI resources
At https://www.ebi.ac.uk/training/online/

What are patterns (AKA: motifs => BEWARE!!! Same name different concept)?

Many important sequence features, such as binding sites or the active sites of enzymes, consist of only a few amino acids that are essential for protein function. When creating patterns, a conserved motif is used to build a regular expression.



The pattern illustrated here is translated as:
[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}.

Databases that use patterns:
Prosite http://prosite.expasy.org/

Pattern syntax at PROSITE for pattern (motif) searching

Pattern	Explanation
[AC]-x-V-x(4)-{ED}	[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}
<A-x-[ST](2)-x(0,1)-V	Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val at the N-terminal of the sequence
<{C}*>	No Cys from the N-terminal to the C-terminal i.e. All sequences that do not contain any Cys.
IIRIFHLRNI	Ile-Ile-Arg-Ils-Phe-His-Leu-Arg-Asn-Ile

ScanProsite Results Viewer

Ouput format: Graphical view - this view shows ScanProsite results together with ProRule-based predicted intra-domain features [help].

include splice variants (Swiss-Prot)

Hits for USERPAT1{P-x(2)-G-E-S-G(2)-[AS]} motif on all UniProtKB/Swiss-Prot (release 2017_09 of 27-Sep-17: 555594 entries) database sequences :

found: 11 hits in 11 sequences

http://prosite.expasy.org/scanprosite/scanprosite_doc.html

What are profiles?

They are built by converting multiple sequence alignments (MSAs) into position-specific scoring systems. Amino acids at each position in the alignment are scored according to the frequency with which they occur.

PWMs are often derived from a set of aligned sequences that are thought to be functionally related and have become an important part of many software tools for computational motif discovery.

A position weight matrix (PWM), also known as a position-specific weight matrix (PSWM) or position-specific scoring matrix (PSSM), is a commonly used representation of motifs (patterns) in biological sequences.

Databases that use profiles to classify proteins:

CDD https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
HAMAP http://hamap.expasy.org/
PROSITE http://prosite.expasy.org/
PRODOM http://prodom.prabi.fr/prodom/current/html/home.php

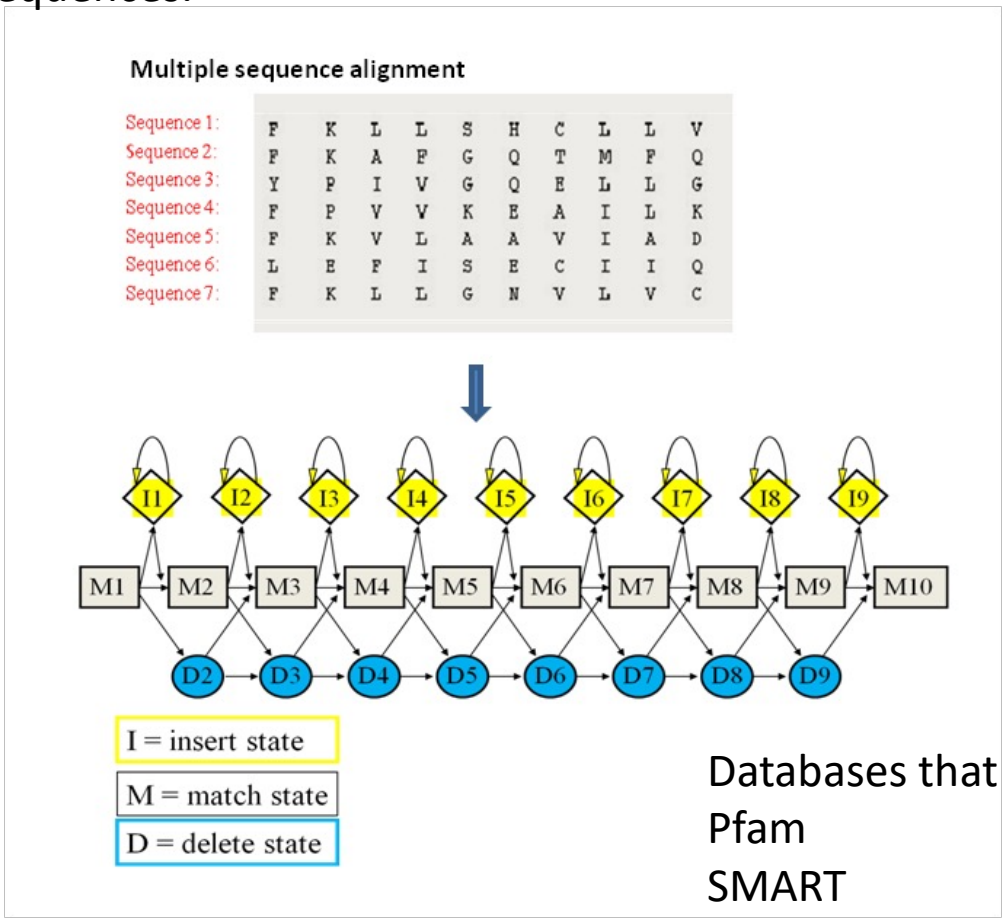
http://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi

P - consensus sequence position **C** - consensus sequence residue

P	C	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	M	-1	-3	1	2	1	6	0	-2	-3	-1	-2	-1	-1	-2	0	-2	-1	-1	-3	-2
2	I	-1	-4	4	1	3	3	-1	-3	-3	-1	-2	-1	-1	-3	-2	-3	-3	-3	-3	-3
3	K	-1	-2	-3	-3	-2	-2	-3	-4	-2	-3	0	2	-3	0	0	-1	2	3	4	0
4	N	-2	1	-3	-3	-3	-2	1	-3	-2	-3	-1	-1	0	3	3	5	-1	-1	2	0
5	N	-2	-2	0	0	-1	1	-2	-4	-2	-3	0	0	-3	3	-1	-1	-1	-2	4	1
6	L	1	-2	1	1	1	0	-2	-3	4	-2	0	0	-2	-3	-2	-3	-2	-3	-3	-2
7	G	-1	5	-4	-4	-4	-3	-4	-3	-3	2	1	-2	-3	-1	4	-2	-1	-2	-2	-1
8	V	-2	-4	4	-1	4	-1	-2	-4	-4	-2	-2	2	-2	-3	-3	0	-3	-1	-4	-3
9	A	-1	-4	3	2	3	1	-2	-3	-3	-2	-2	2	-2	-3	-3	-4	-3	-3	-4	-3
10	V	0	-3	-1	-1	1	0	-2	-3	0	-3	-1	2	0	0	2	2	0	0	1	0
P	C	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	I	-1	-5	6	2	3	0	-1	-3	-4	-3	-3	-2	1	-5	-4	-4	-4	-4	-5	-4
12	G	0	3	-4	-3	-3	-3	-4	-4	-3	-4	-1	-2	-3	4	1	2	-1	-2	3	1
13	S	-2	4	-5	-4	-4	-3	-4	-4	-3	-4	0	-2	-4	0	-1	-2	3	3	-1	-2
14	K	-2	-3	-2	-3	-2	-2	-3	-4	-3	-4	-1	0	-1	1	1	4	4	3	-2	0
15	Q	-2	-1	-3	-3	-2	-2	-3	-3	0	-3	0	2	0	-1	2	-1	3	3	0	2
16	Y	-4	-5	-3	-1	-3	-2	6	3	-5	-4	-4	-4	7	-5	-4	-1	-4	-4	-5	-5
17	A	4	-3	0	-2	4	-2	-3	-4	-3	-2	-1	-2	-3	-4	-3	-4	-3	-4	-4	-3
18	V	1	-1	1	-1	3	-1	-3	-4	-3	5	2	1	-3	-3	-3	-3	-3	-3	-3	-3
19	N	-1	5	-5	-5	-4	-4	-5	-5	-3	-4	2	-2	-4	4	-2	-2	-2	-3	0	-2
20	L	-3	-6	1	6	-1	0	-1	-4	-5	-3	-4	-3	-3	-5	-4	-5	-4	-4	-6	-5

What are HMMs?

Hidden Markov models (HMMs) are used by many databases. Like profiles, they can be used to convert multiple sequence alignments into position-specific scoring systems. HMMs are adept at representing amino acid insertions and deletions, meaning that they can model entire alignments, including divergent regions. They are sophisticated and powerful statistical models, very well suited to searching databases for homologous sequences.



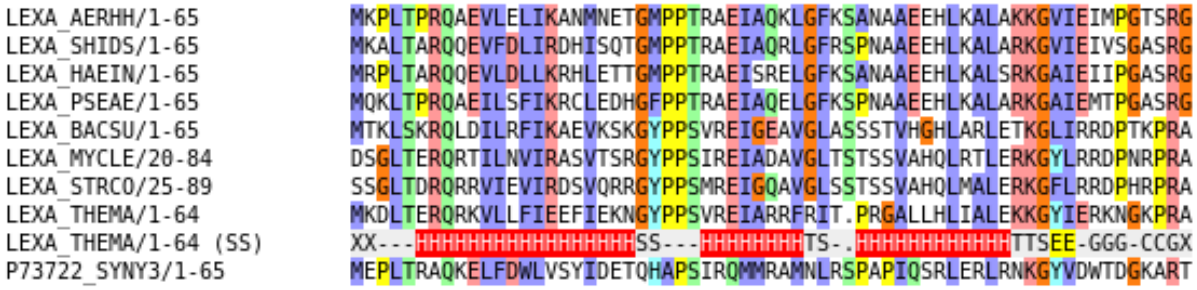
Databases that use HMM to classify proteins:

Pfam http://pfam.xfam.org/

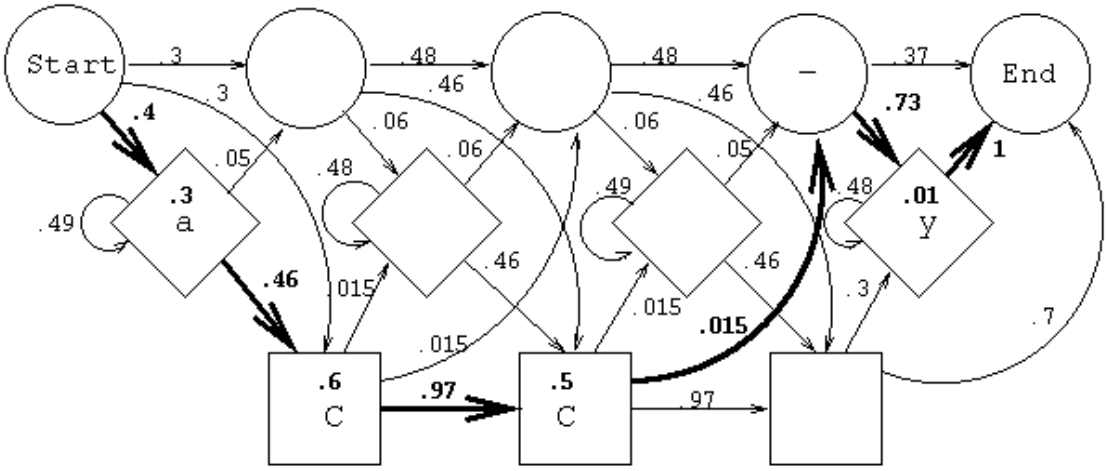
SMART http://smart.embl-heidelberg.de/

The Pfam database is a large collection of protein families

It stores profiles that contain all information inside a multiple alignment



These profiles are stored in the form of a Hidden Markov model (HMM)

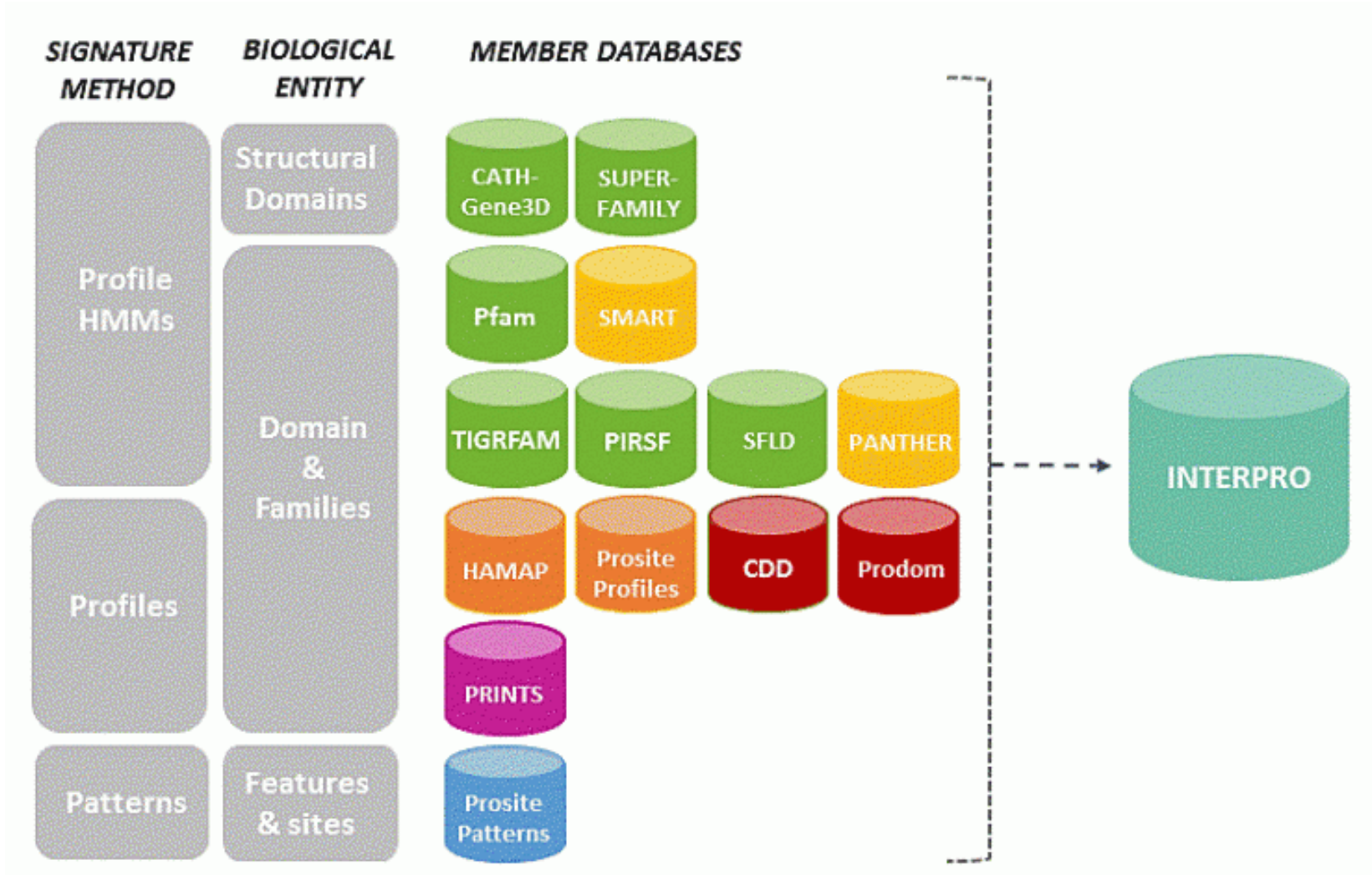


http://pfam.xfam.org/

InterPro: Databases that combine patterns, profiles and HMMs to classify proteins

https://www.ebi.ac.uk/interpro/

InterPro is the main resource for protein classification at the EBI.

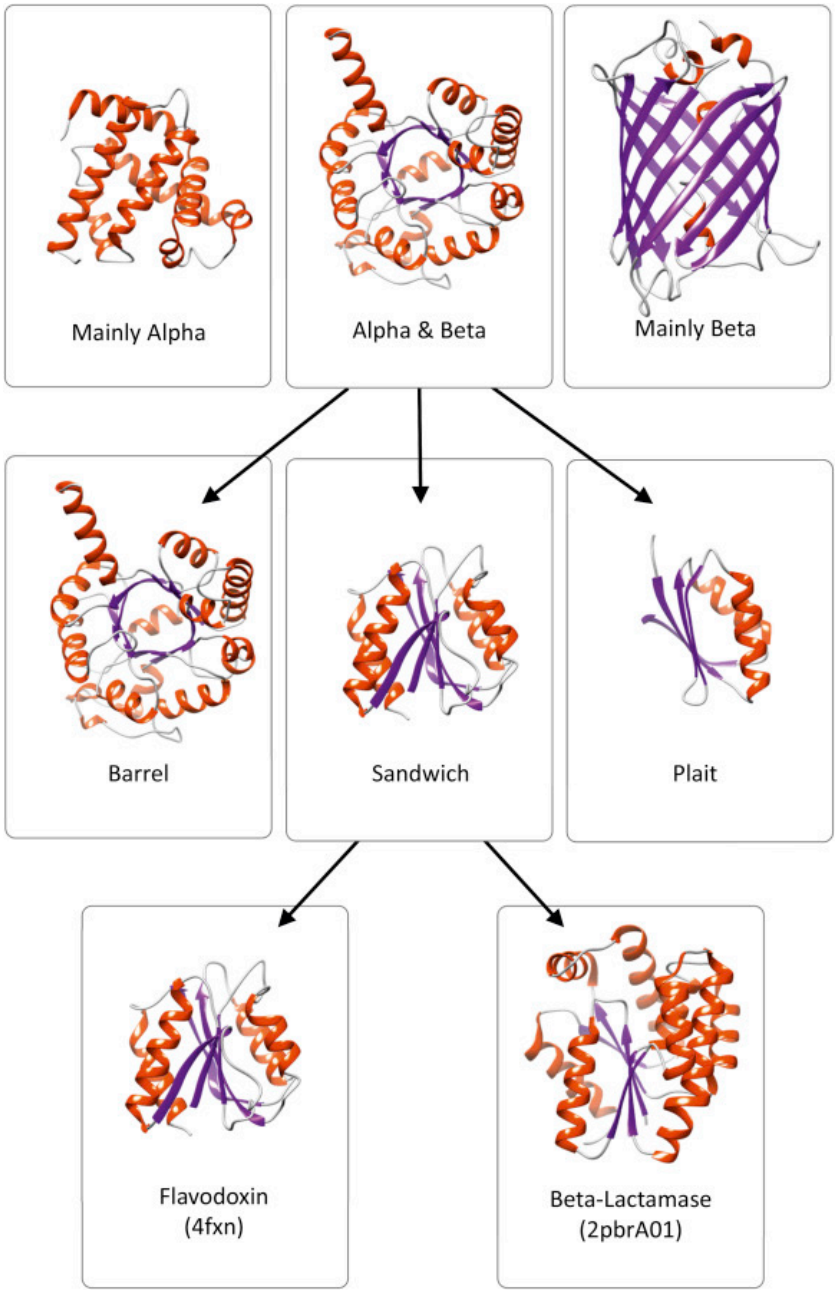


CATH: Protein Structure Classification Database

http://www.cathdb.info/

The first three levels of the CATH structure classification hierarchy:

1. Class (based on secondary structure content).
2. Architecture (based on gross spatial arrangement of secondary structures).
3. Topology or Fold (similar folding arrangement of secondary structures).



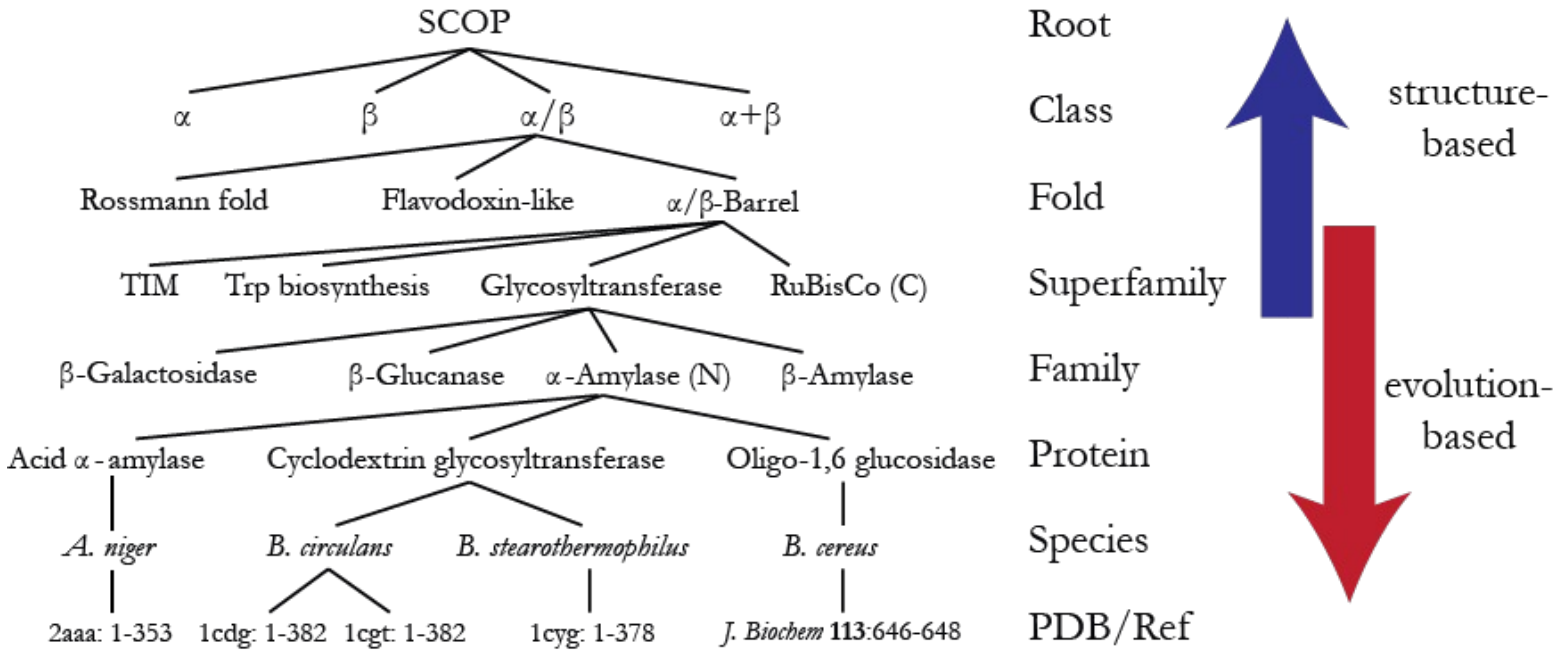
Biochimie. Volume 119, December 2015, Pages 209-217.

SCOP2: Structural Classification of Proteins database

Manually (visually) classified

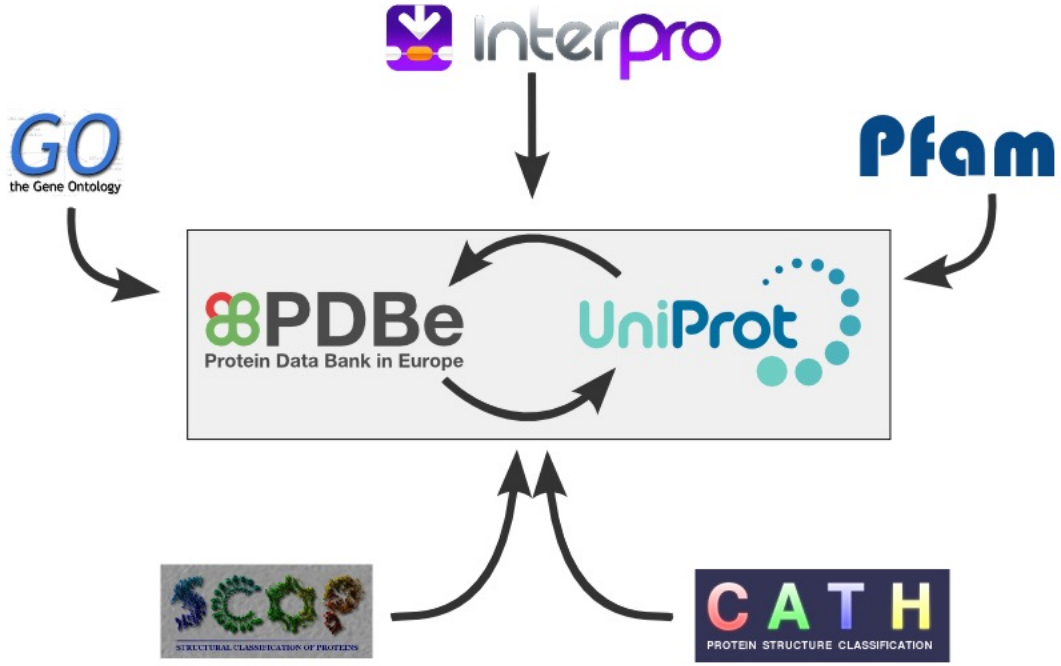
- All alpha
- All beta
- Alpha and beta (a/b) proteins with alternating alpha-helices and beta-strands
- Alpha + beta (a+b) proteins with segregated alpha-helices and beta-strands
- Small proteins little or no secondary structures

http://scop2.mrc-lmb.cam.ac.uk/



Murzin A. G., et al. (1995). J. Mol. Biol. 247, 536-540.

SIFTS: Structure integration with function, taxonomy and sequence



https://www.ebi.ac.uk/pdbe/docs/sifts/index.html

pdb_chain_uniprot.tsv.gz pdb_chain_uniprot.csv.gz	A summary of the PDBe to UniProt residue level mapping, showing the start and end residues of the mapping using SEQRES, PDB sequence and UniProt numbering.
pdb_chain_taxonomy.tsv.gz pdb_chain_taxonomy.csv.gz	A summary of the NCBI tax_id(s),scientific_name(s) and chain type for each PDB chain that has been processed.
pdb_pubmed.tsv.gz pdb_pubmed.csv.gz	A summary of the Pubmed id(s) associated with each PDB entry, together with an ordinal number.
pdb_chain_enzyme.tsv.gz pdb_chain_enzyme.csv.gz	A summary of the EC number(s) for each PDB chain that has been processed.
pdb_chain_go.tsv.gz pdb_chain_go.csv.gz	A summary of the GO identifier(s) for each PDB chain that has been processed.
pdb_chain_interpro.tsv.gz pdb_chain_interpro.csv.gz	A summary of the InterPro identifier(s) for each PDB chain that has been processed.
pdb_chain_pfam.tsv.gz pdb_chain_pfam.csv.gz	A summary of the Pfam domain identifier(s)(derived via the UniProt mapping) for each PDB chain that has been processed.
pdb_chain_cath_uniprot.tsv.gz pdb_chain_cath_uniprot.csv.gz	A summary of the CATH identifier(s) and UniProt primary accession number(s) for each PDB chain that has been processed.
pdb_chain_scop_uniprot.tsv.gz pdb_chain_scop_uniprot.csv.gz	A summary of the SCOP identifier(s) and UniProt primary accession number(s) for each PDB chain that has been processed.
uniprot_pdb.tsv.gz uniprot_pdb.csv.gz	A summary of the UniProt to PDB mappings showing the UniProt accession followed by a semicolon-separated list of PDB four letter codes.
pdb_chain_ensembl.tsv.gz pdb_chain_ensembl.csv.gz	A summary of the Ensembl identifier(s) and UniProt primary accession number(s) for each PDB chain that has been processed.
pdb_chain_hmmer.tsv.gz pdb_chain_hmmer.csv.gz	A summary of the Pfam domain identifier(s) (automatically calculated using HMMER) for each PDB chain that has been processed.
uniprot_segments_observed.tsv.gz uniprot_segments_observed.csv.gz	A summary of the UniProt to PDBe residue level mapping (observed residues only), showing the start and end residues of the mapping using SEQRES, PDB sequence and UniProt numbering.

Specialized Sequence Databases in Health and Life Science

Ensembl automatic annotation databases for human, mouse, other vertebrate and eukaryote genomes.
SILVA, a database of ribosomal RNAs
VectorBase The NIAID Bioinformatics Resource Center for Invertebrate Vectors of Human Pathogens
Wormbase, genome of the model organism *Caenorhabditis elegans*
PATRIC, the PathoSystems Resource Integration Center
Flybase, genome of the model organism *Drosophila melanogaster*
EcoCyc genome and the biochemical machinery of the model organism *E. coli K-12*
ENZYME (Enzyme database at ExPASy, CH)
BRENDA (The comprehensive enzyme information system, DE)
ReLiBase (database system for analysing receptor/ligand complexes deposited in the Protein Data Bank at EBI, UK)
HGMD disease-causing mutations (HGMD Human Gene Mutation Database)
HUGO (Official Human Genome Database: HUGO Gene Nomenclature Committee)
IEDB Immune Epitope Database
OrthoMCL Ortholog Groups of Protein Sequences from Multiple Genomes.
DrugBank combines detailed drug data with comprehensive drug target information
PRIDE Archive - proteomics data repository
EST database is a collection of short single-read transcript sequences from GenBank.
……

See the List of biological databases at Wikipedia