Protein classification: An introduction to EMBL-EBI resources

Amaia Sangrador [1]

- Proteins
- Beginner
- 0.5 hour

This course will provide an introduction to protein classification and basic concepts, such as proteins families, domains and sequence features.

Learning objectives:

- Understanding the importance of classifying proteins
- Understanding how protein families, domains and sequence features can be defined, and how these can be used to classify proteins
- Becoming familiar with the different predictive methods you can use to help classify proteins: patterns, profiles, fingerprints and hidden Markov models (HMMs).
- Knowing which resources for classifying proteins according to family, domain and sequence features are available at the EBI

Protein Classification

Proteins are the macromolecules responsible for the biological processes in the cell. They consist at their most basic level of a chain of amino acids, determined by the sequence of nucleotides in a gene. Depending on the amino acid [2] sequence (different amino acids have different biochemical properties) and interactions with their environment, proteins fold into a three-dimensional structure, which allows them to interact with other proteins and molecules and perform their function (see Figure 1 below).

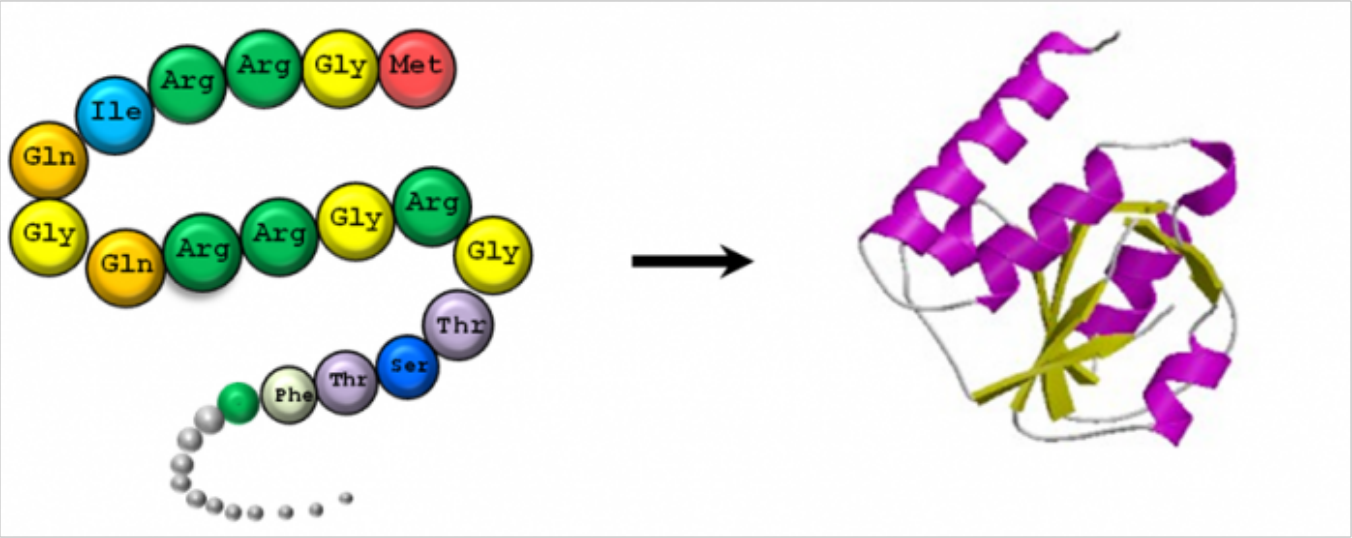

Figure 1 Proteins consist of one or more polypeptides. A polypeptide is a chain of amino acids. The polypeptide chains fold into their final three-dimensional structure to constitute a functional protein. The amino acid sequence and structure in this example correspond to ribosomal protein L2.

Proteins that have diverged from a common ancestral gene are known as homologous [3]. Proteins with similar sequences are assumed to be homologous and usually (within certain limits) have similar structures and functions.

Why classify proteins?

Proteins can be classified into groups according to sequence or structural similarity. These groups often contain well characterised proteins whose function is known. Thus, when a novel protein is identified, its functional properties can be proposed based on the group to which it is predicted to belong.

In this tutorial we will explain how families, domains and sequence features can be defined and used for protein classification. Although these terms are widely used in the biological literature, you will find that their definitions may vary depending on the source. Taking this into account, lets see how proteins can be classified into different groups based on:

- the FAMILIES to which they belong
- the DOMAINS they contain
- the SEQUENCE FEATURES they possess

What are protein families?

What are protein families?

A protein family [4] is a group of proteins that share a common evolutionary origin, reflected by their related functions and similarities in sequence or structure.

Protein families are often arranged into hierarchies, with proteins that share a common ancestor [5] subdivided into smaller, more closely related groups. The terms superfamily (describing a large group of distantly related proteins) and subfamily (describing a small group of closely related proteins) are sometimes used in this context. A hypothetical protein family hierarchy is illustrated in Figure 2 below.

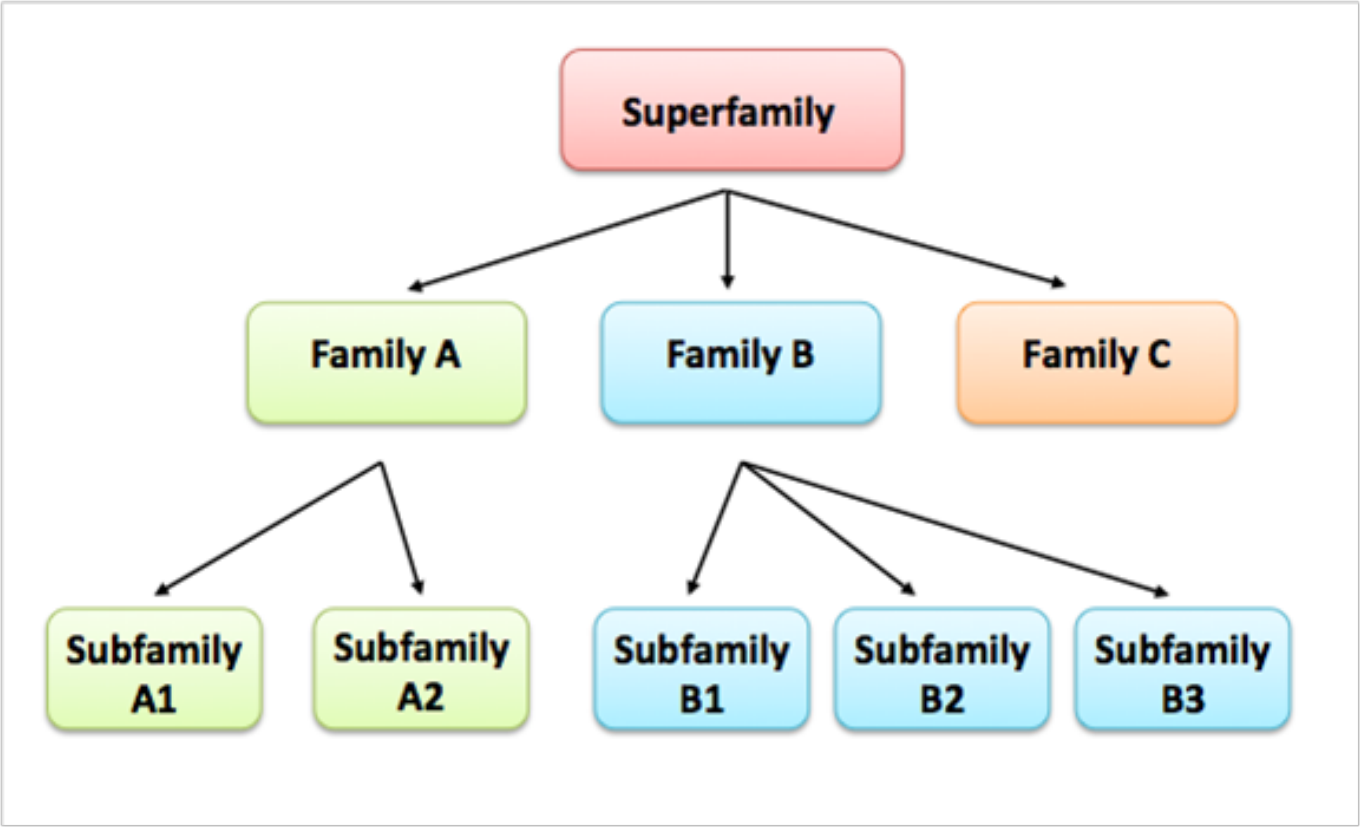

Figure 2 A hypothetical protein family hierarchy showing the relationships between superfamily, family and subfamily members. Directional arrows indicate that one group is a subgroup of another.

...

What are protein families?

One set of proteins that comprise a superfamily are the G protein-coupled receptors (GPCRs [6]). These are a large and diverse group of proteins that are involved in many biological processes, including photoreception, regulation of the immune system, and nervous system transmission. At the superfamily level, GPCRs share two common properties – they have seven transmembrane domains, and interact with specialised proteins (called G proteins) to influence intracellular pathways after binding extracellular [7] signals (you can visit this GPCR webpage [8] for more information).

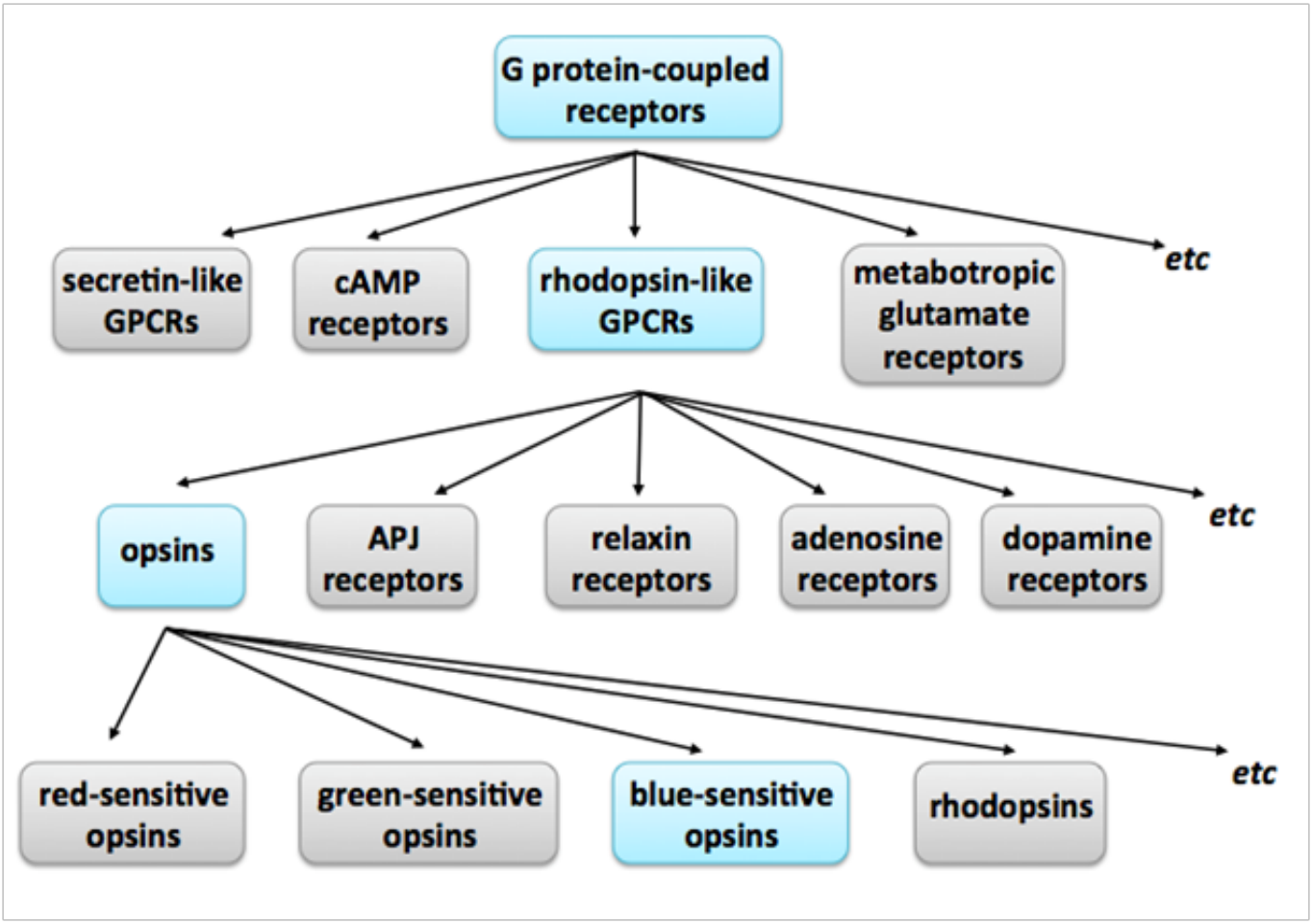

Figure 3 The GPCR superfamily hierarchy. Families and subfamilies to which the short-wave-sensitive opsin 1 protein belongs are highlighted in blue.

As we group the GPCRs [6] into smaller families, the individual groups have more properties in common. For example, the protein short-wave-sensitive opsin 1 belongs to a specialised family, known as the rhodopsin-like GPCRs. The rhodopsin-like GPCRs themselves can be further broken down into smaller families that respond to different signals. Short-wave-sensitive opsin 1 proteins belong to the opsin family (opsins being the photoreceptors of animal retinas), but more specifically, they are members of the blue-sensitive opsin subfamily, all of which are activated by a particular wavelength of light. This protein family [4] hierarchy is illustrated in Figure 3 above.

As can be seen from this example, when classifying proteins into hierarchical families, the level at which we can place a protein in the hierarchy is vital, since it determines the amount of specific functional information that we can infer.

What are protein domains?

What are protein domains?

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 role of a protein. Domains may exist in a variety of biological contexts, where similar domains can be found in proteins with different functions.

For example, Src homology 3 (SH3) domains are small domains of around 50 amino acid [2] residues that are involved in protein-protein interactions. SH3 domains have a characteristic 3D structure (see Figure 4). They occur

in a diverse range of proteins with different functions, including adaptor proteins, phosphatidylinositol 3-kinases, phospholipases and myosins.

Figure 4 Structure of the SH3 domain.

An example of a protein that contains *multiple* SH3 domains is the cytoplasmic protein Nck. Nck belongs to the adaptor family of proteins and it is involved in transducing signals from growth factor receptor tyrosine kinases to downstream signal recipients. The domain [9] composition of Nck is illustrated in Figure 5 below.

Figure 5 Domain composition of Nck. Nck contains three SH3 domains plus another domain known as SH2 (Src homology 2). Both SH3 and SH2 domains are usually found in proteins that interact with other proteins and mediate assembly of protein complexes. SH3 domains typically bind to proline-rich peptides in their respective binding partners, while SH2 domains interact with phosphotyrosine-containing target peptides.

What are protein domains?

As we have just seen with Nck, proteins can be composed of multiple domains. Often the individual domains have specific functions, such as binding a particular molecule or catalysing a given reaction, and together these contribute to the overall role of the protein (see, for example, the domain composition of the enzyme phospholipase D1 in Figure 6 below).

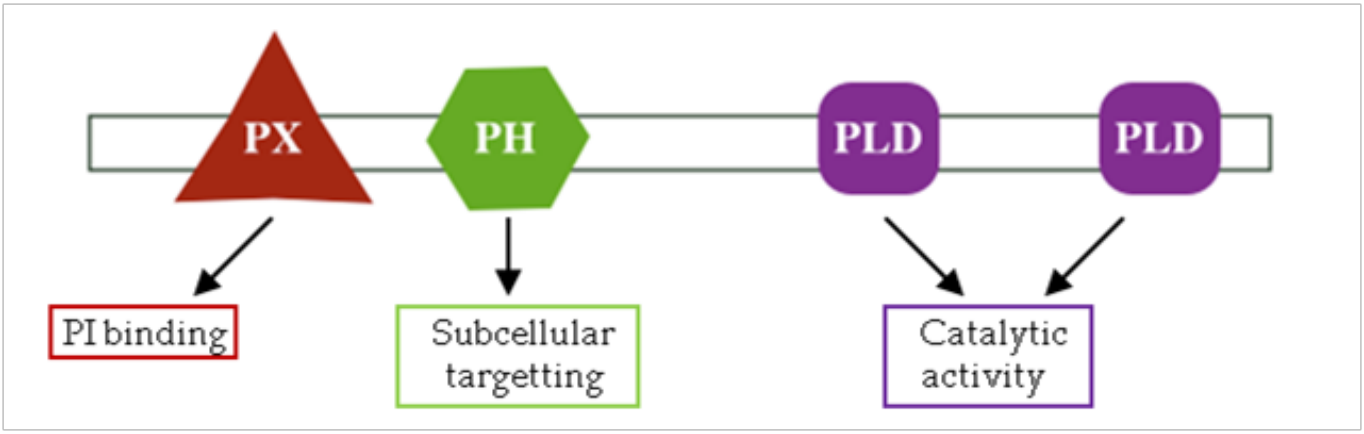

Figure 6 Domain composition of phospholipase D1, which is an enzyme that breaks down phosphatidylcholine.

The protein contains a PX (phox) domain that is involved in binding phosphatidylinositol, a PH (pleckstrin homology) domain that has a role in targeting the enzyme to particular locations within the cell, and two PLD (phospholipase D) domains responsible for the protein's catalytic activity.

Family- and domain-based protein classification

Family- and domain [9]-based classifications are not always straightforward and can overlap, since proteins are sometimes assigned to families by virtue of the domain(s) they contain. An example of this kind of complexity is outlined below.

Protein families and domain composition – an example

Regulator of G-protein signalling (RGS) domains are protein structural units that activate GTPases [10]. They are found in sequences that belong to the RGS protein family [4], which are multi-functional GTPase-accelerating proteins. All RGS protein family members contain an RGS domain, but while some (such as RGS1) consist of little more than the domain, others (such as RGS3 and RGS6) contain additional domains that confer further functions, such as DEP domains which are involved in membrane targeting.

RGS domains are also found in proteins belonging to *other families*, such as beta-adrenergic receptor kinases, axins, and some members of the sorting nexin family. The family groupings and domain composition of some of these proteins is summarised in Figure 7 below.

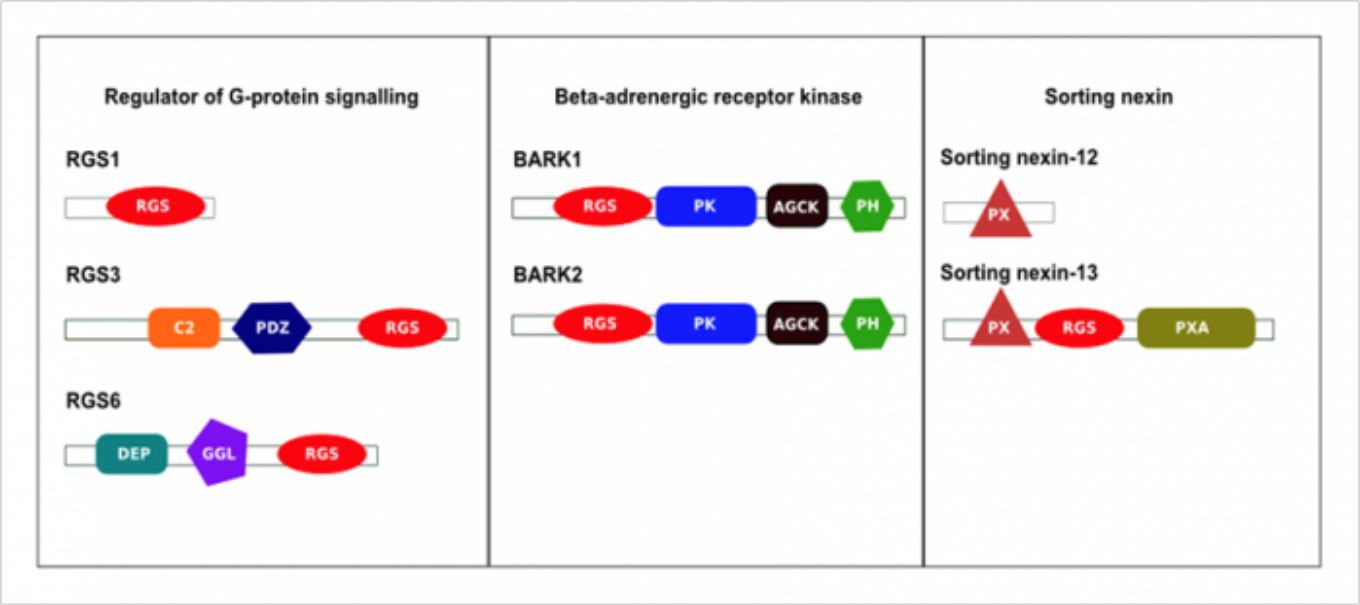

Figure 7 Family groupings and domain composition of some RGS domain-containing proteins.

What are sequence features?

What are sequence features?

Sequences 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, which contain amino acids involved in catalytic activity. For example, the enzyme lipase, which catalyses the formation and hydrolysis of fats, has two amino acid [2] residues (a histidine followed by a glycine) that are essential for its catalytic activity.
- binding sites, containing amino acids that are directly involved in binding molecules or ions, like the iron-binding site [11] of haemoglobin.
- post-translational modification (PTM) sites, which contain residues known to be chemically modified (phosphorylated, palmitoylated, acetylated, *etc*) after the process of protein translation [12].
- repeats, which are typically short amino acid sequences that are repeated within a protein, and may confer binding or structural properties upon it.

Sequence features differ from domains in that they are usually quite small (often only a few amino acids long), whereas domains represent entire structural or functional units of the protein (see Figure 8). Sequence features are often nested within domains – a protein kinase domain [9], for example, usually contains a protein kinase active site [13].

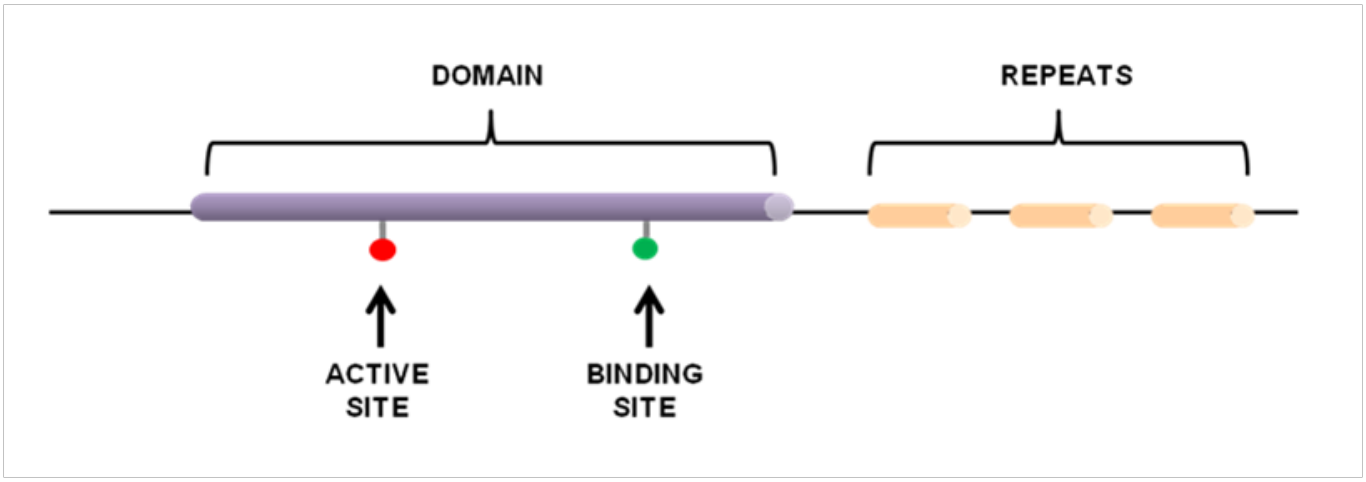

Figure 8 Graphical representation of repeats, domains and sites on a protein sequence.

What are sequence features?

Proteins can also be classified according to the sequence features they contain. For example, ferredoxins are sulphur-iron proteins that mediate electron transfer in a variety of biological redox reactions, including the photosynthetic process. They can be divided into several groups according to the nature of their iron-sulphur cluster (you can find out more information about ferredoxins here [14]).

In the 2Fe-2S ferredoxins (which bind a cluster of two iron (Fe) and two sulphur (S) atoms), there are 4 cysteines residues involved in iron-sulphur binding. The 2Fe-2S binding site [11] is shown on the ferredoxin 3D-structure in Figure 9 below.

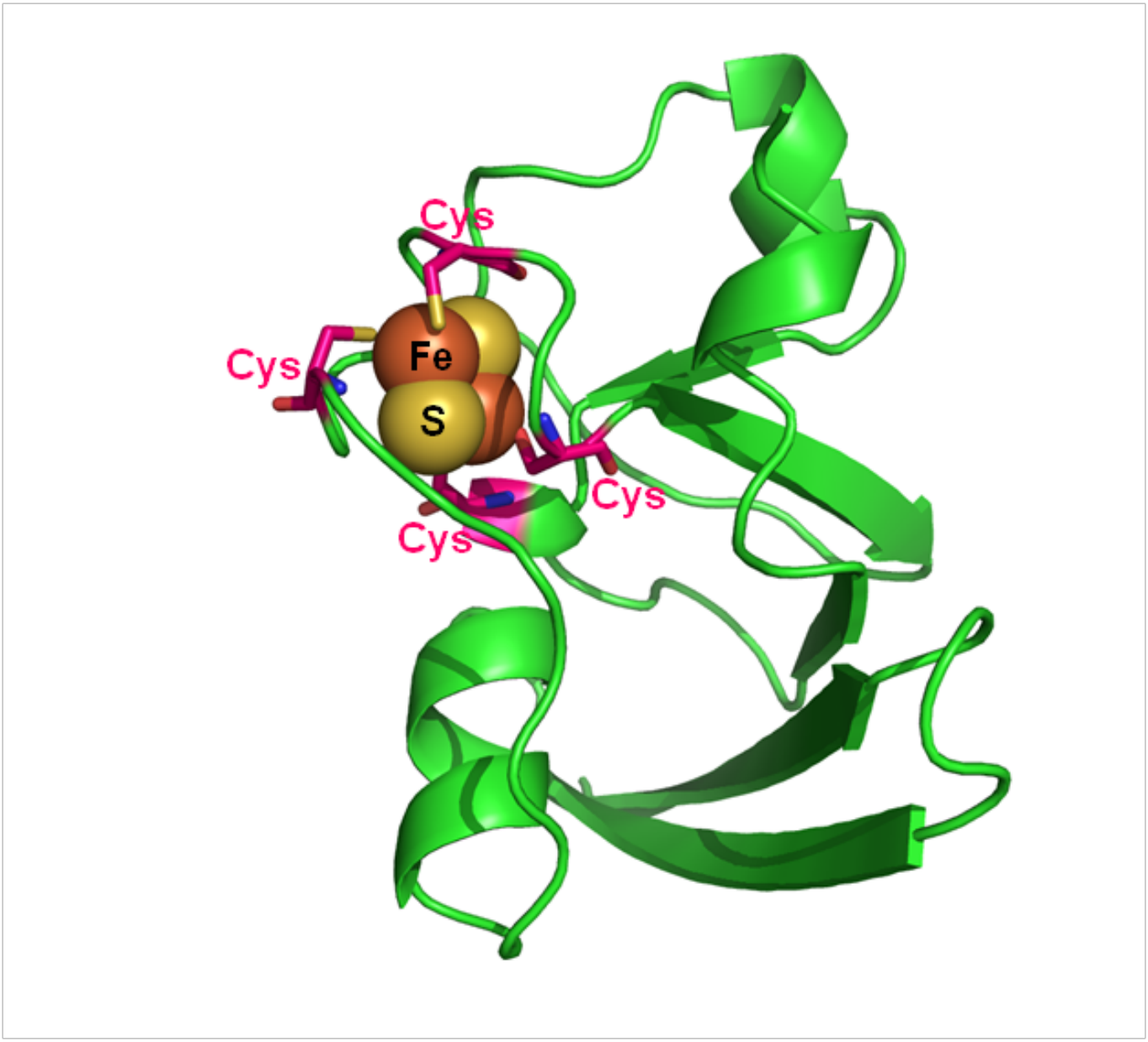

Figure 9 3D-structure of a plant-type ferredoxin with its 2Fe-2S cluster. The conserved cysteine (Cys) residues that help form the binding site are highlighted in red. The iron and sulphur atoms bound to the cysteines are displayed as spheres.

What are protein signatures?

In order to classify proteins into families and to predict the presence of important domains or sequence features, we require computational tools. One set of such tools are the predictive models known as *protein signatures*.

There are different types of signatures, built using different computational approaches. However, their common starting point is a multiple sequence alignment of proteins sharing a set of characteristics (e.g. belonging to the same family or sharing a domain [9]) (see Figure 10 below). When building the initial model, the level of amino acid [2] conservation at different positions in the alignment is taken into account. The model is then used to search a protein database in an iterative manner, refining the model as more distantly related sequences in the database are identified. Once the model is mature, the signature is ready and can be used for protein sequence analysis.

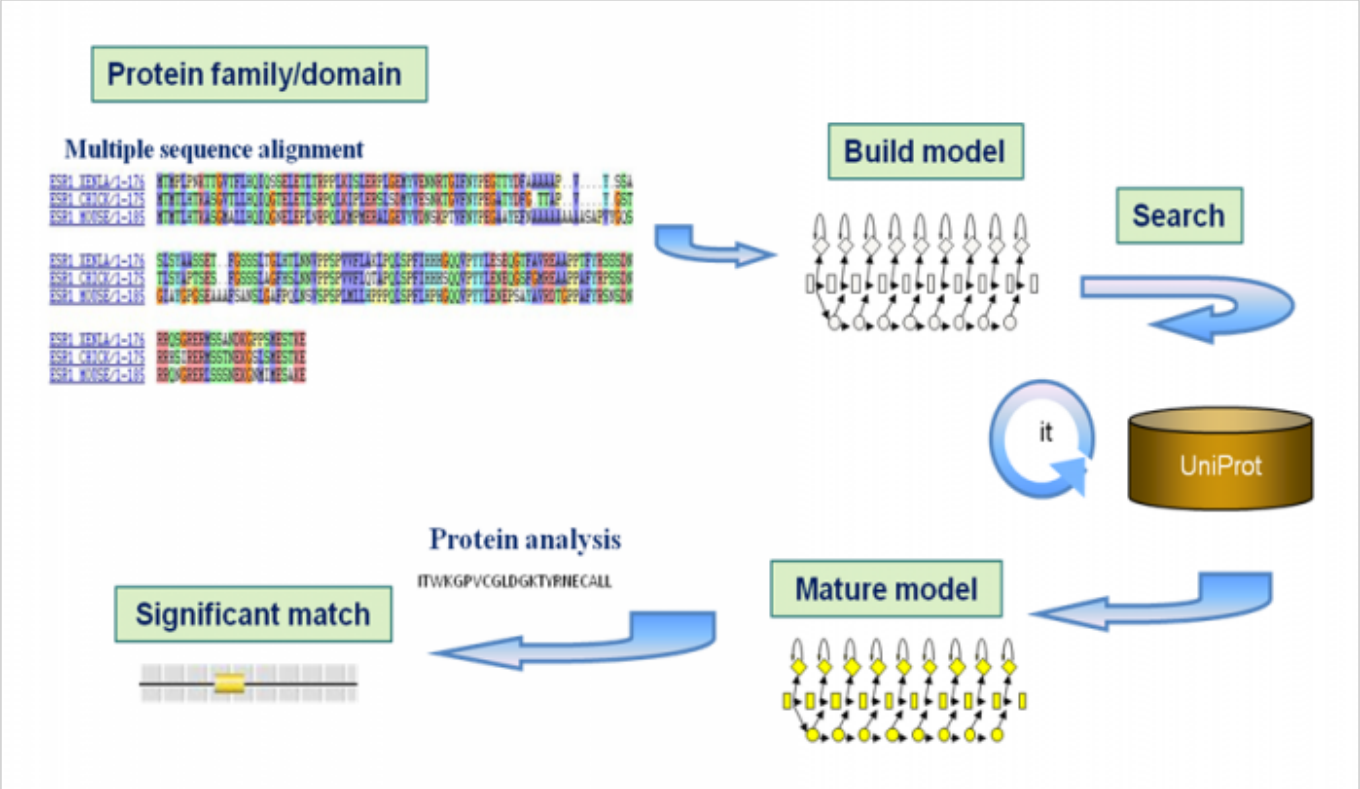

Figure 10 The process of building a protein signature starts with a multiple sequence alignment, which is used to build a predictive model. By searching a protein database in an iterative way, more distantly related sequences can be identified. This information is used to create a final mature model.

How do protein signatures compare to other ways of classifying proteins?

Multiple sequence alignments can provide us with valuable information for protein classification since they allow us to identify the (often few) amino acid [2] residues that are conserved in distantly related proteins (see Figure 11). It is not possible to identify such important residues with pairwise alignment techniques, such as BLAST [15]. As a consequence, protein signatures built from multiple sequence alignments are usually better at detecting divergent homologues than pairwise comparison methods.

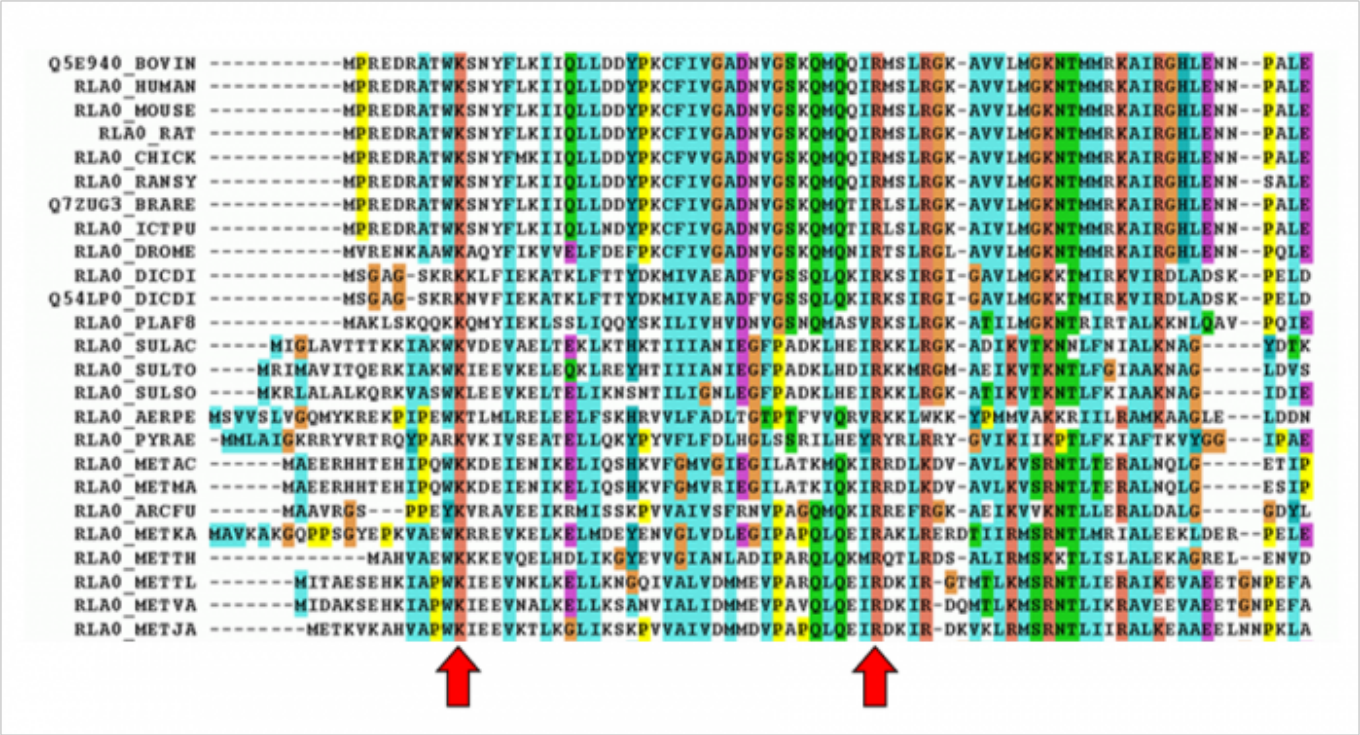

Figure 11 Multiple sequence alignment for 60S acidic ribosomal protein P0 from different organisms (eukaryota and archaea). There are two amino acids indicated by red arrows, lysine (K) and arginine (R), that are conserved in all sequences. Multiple sequence alignment methods are important for identifying highly conserved residues that are essential for stability or function of the protein.

Signature Types

Different approaches can be used to generate signatures. These include:

- patterns
- profiles [16]
- fingerprints
- hidden Markov models (HMMs)

Each approach starts with a protein multiple sequence alignment, and can focus on a single conserved sequence region (known as a motif [17]), multiple conserved motifs, or the full alignment of the entire protein or a particular domain [9] (see Figure 12).

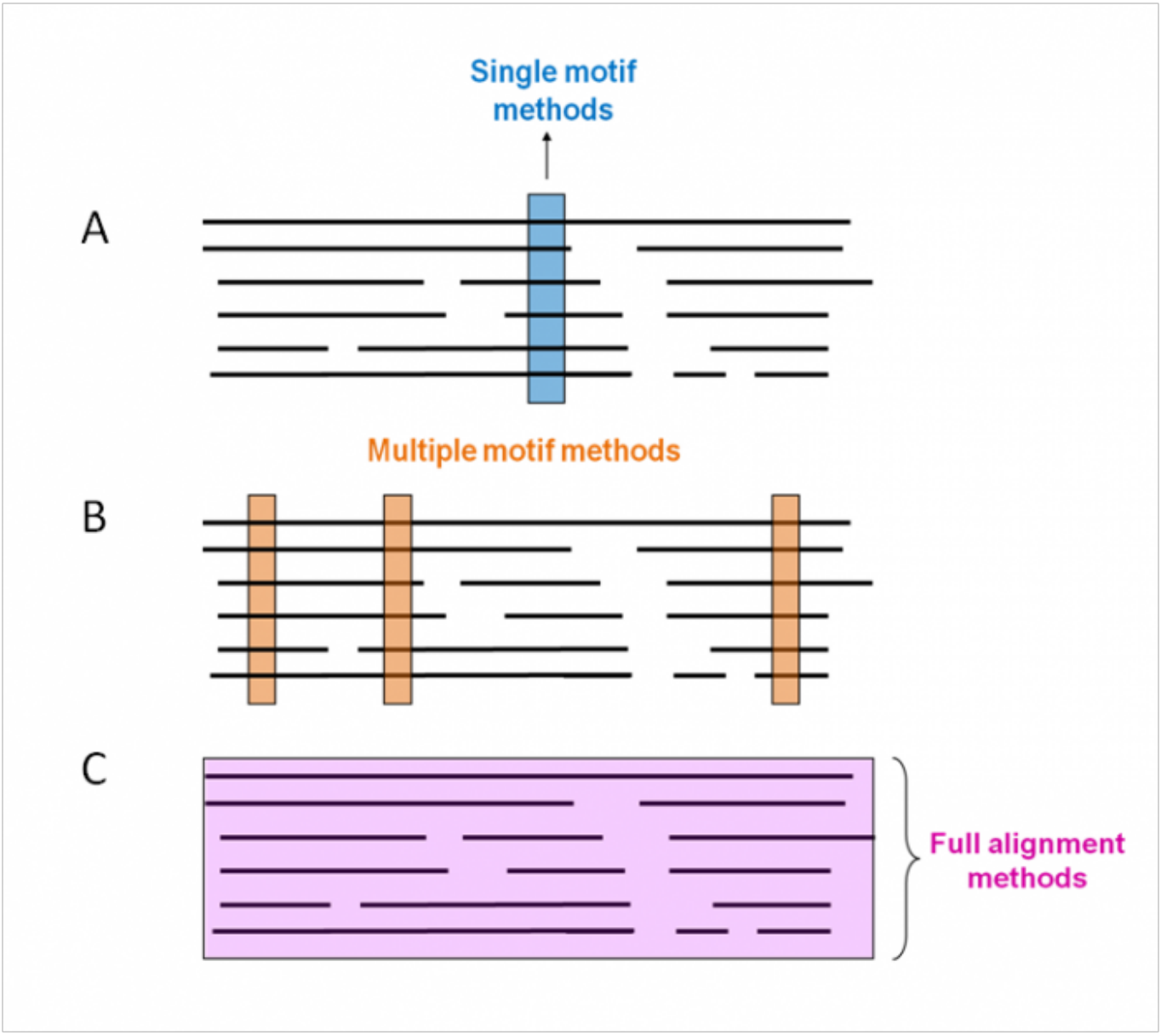

Figure 12 Representation of different strategies for building signatures. A) single motif methods, B) multiple motif methods; and C) full alignment methods. Protein multiple sequence alignments are represented by black lines and the conserved regions used to build the signatures are indicated by coloured boxes.

What are patterns?

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. Patterns are very good at recognising such features. They are built by identifying these regions in multiple sequence alignments. The pattern [18] of conservation within the sequence feature is then modelled as a regular expression [19], as is indicated in Figure 13.

Learn more about regular expressions here [20].

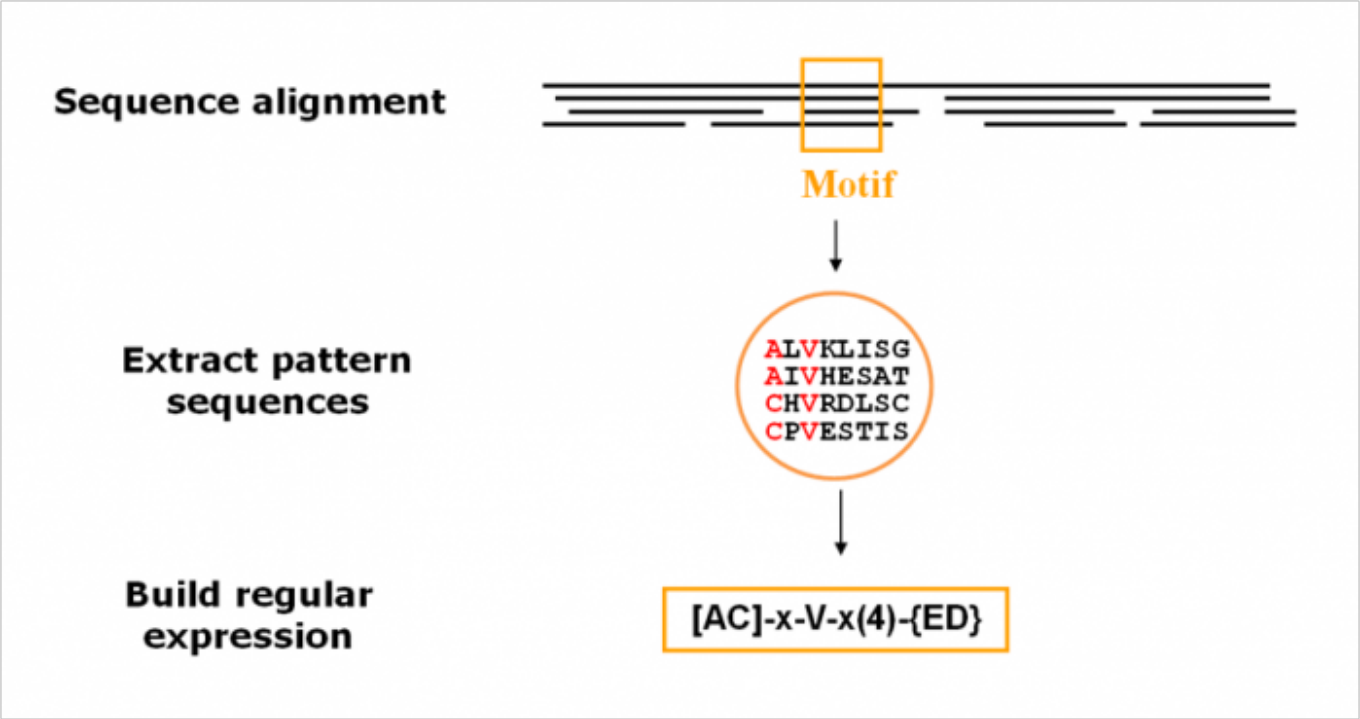

Figure 13 When creating patterns, a conserved motif is used to build a regular expression. The pattern illustrated here is translated as: [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}.

An example of a database that uses patterns is PROSITE [21] (Bairoch, A. 1991 [22]).

What are profiles?

Profiles [16] are used to model protein families and domains. They are built by converting multiple sequence alignments into position-specific scoring systems (PSSMs). Amino acids at each position in the alignment are scored according to the frequency with which they occur, as represented in Figure 14. Substitution matrices (such as BLOSUM matrices [23]) can be used to add evolutionary distance weighting these scores.

Learn more about PSSMs [24] and substitution matrices [25].

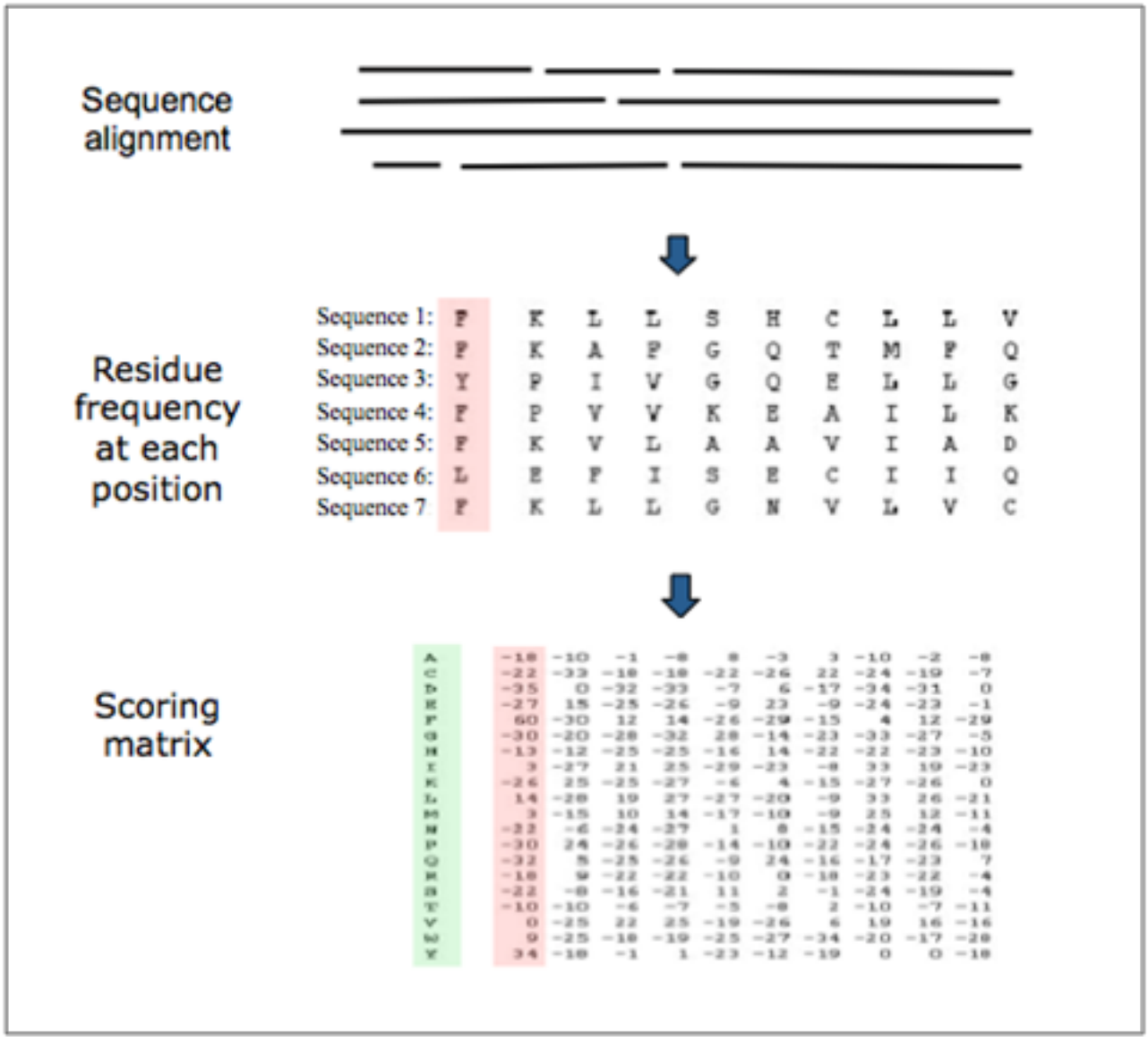

Figure 14 Representation of a scoring matrix based on a multiple sequence alignment. Each of the 20 amino acids commonly found in proteins is given a score for each position in the sequence according to the frequency with which they occur in the original alignment. Other factors, such as evolutionary distances can also be considered.

Examples of databases that use profiles to classify proteins include CDD (Marchler-Bauer A. et al. 2015 [22]), HAMAP [26] (Lima. T, et al. 2009 [22]) and PROSITE [21] (which produces profiles as well as patterns. Sigrist. CJ, et al. 2010 [22]). The PRODOM [27] (Servant. F, et al. 2002 [22]) database also uses a related approach, using PSI-BLAST [28] to create its profiles. You can find out more about profiles by reading Gribskov M. et al. 1987. [22]

What are fingerprints?

What are fingerprints?

While single motif [17] methods are good at identifying features in a protein, most protein families are characterised not by one, but by several conserved regions, which occur in a certain order. Identifying these regions is the principle behind fingerprints. Fingerprints are composed of multiple short conserved motifs, which are drawn from sequence alignments, as illustrated in Figure 15. Each motif is then converted into an individual profile (as described in the previous section) to create a fingerprint signature.

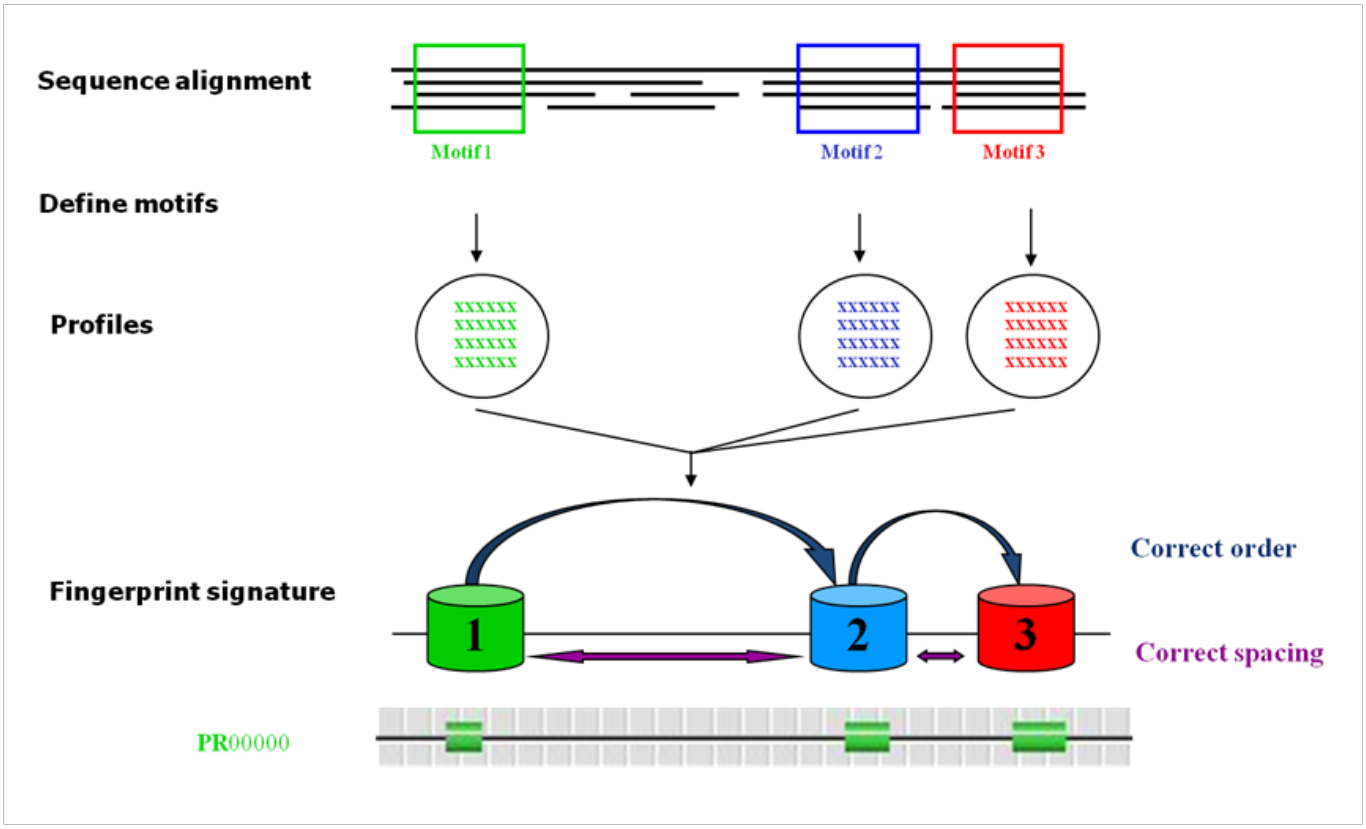

Figure 15 Representation of the steps involved in creating a fingerprint signature.

Fingerprints are used by the PRINTS [29] database, you can find out more about PRINTS in this book [22] (Attwood. TK, et al. 2006).

What are fingerprints?

Fingerprints are very good at modeling the often small differences between closely related proteins, as illustrated in the example in Figure 16 below.

This means fingerprints can distinguish individual subfamilies within protein families. This allows functional characterisation of sequences at a high level of specificity [30] (identifying individual cellular pathways in which a protein might be involved, the ligand [31] it might bind, the exact reaction it may catalyse, and so on).

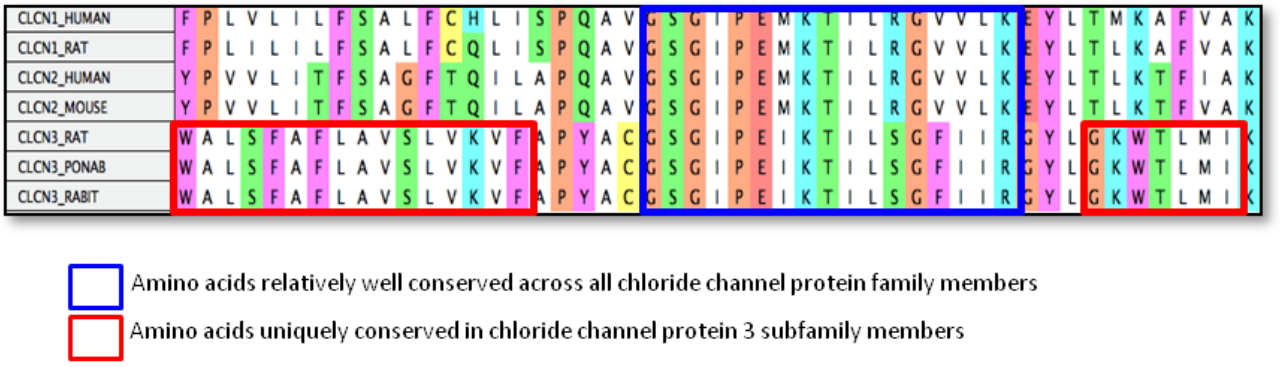

Figure 16 Multiple sequence alignment showing amino acid conservation across chloride channel protein family members. By using multiple short conserved motifs, fingerprints are able to distinguish closely related subfamilies from each other.

What are HMMs?

Hidden Markov models (HMMs) are used by many databases. Like profiles [16], they can be used to convert multiple sequence alignments into position-specific scoring systems. HMMs are adept at representing amino acid [2] 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 [3] sequences.

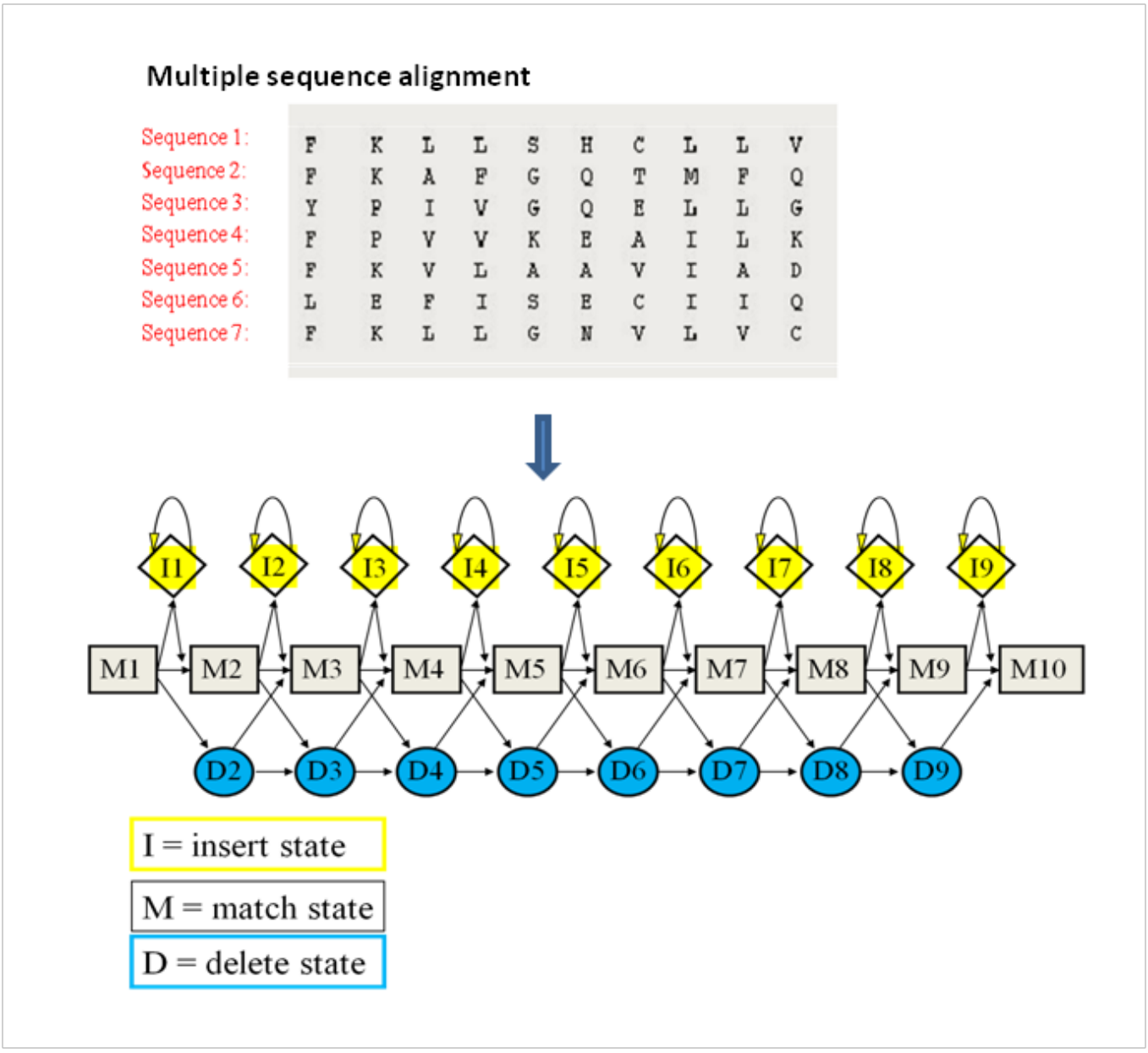

Figure 14 Representation of a Hidden Markov model based on a multiple sequence alignment. Amino acids are given a score at each position in the sequence alignment according to the frequency with which they occur. Transition probabilities (i.e., the likelihood [32] that one particular amino acid follows another particular amino acid) and insertion and deletion states are also modelled.

HMMs have wide utility, as is clear from the numerous databases that use this method for protein classification, including Pfam [33], SMART [34], TIGRFAM, PIRSF [35], PANTHER [36], SFLD [37], Superfamily [38] and Gene3D [39].

For more information on HMMs, see Profile hidden Markov models (Eddy. SR. 1998 [22]).

Protein classification resources at the EBI: InterPro

InterPro [40] is the main resource for protein classification at the EBI.

In InterPro, patterns, profiles [16], fingerprints and HMMs from a number of different databases are brought together into a single searchable resource, offering convenient access to their predictive capabilities without the need to visit the member databases individually (see Figure 18 for an overview of the databases used to construct InterPro).

By combining the different databases and signature types, InterPro capitalises on their individual strengths, producing a powerful tool for the prediction of protein function. InterPro aims to simplify and rationalise protein sequence analysis for the user by combining and organising information in a consistent manner, removing redundancy, and adding extensive annotation [41] and useful links about the signatures and the proteins they match.

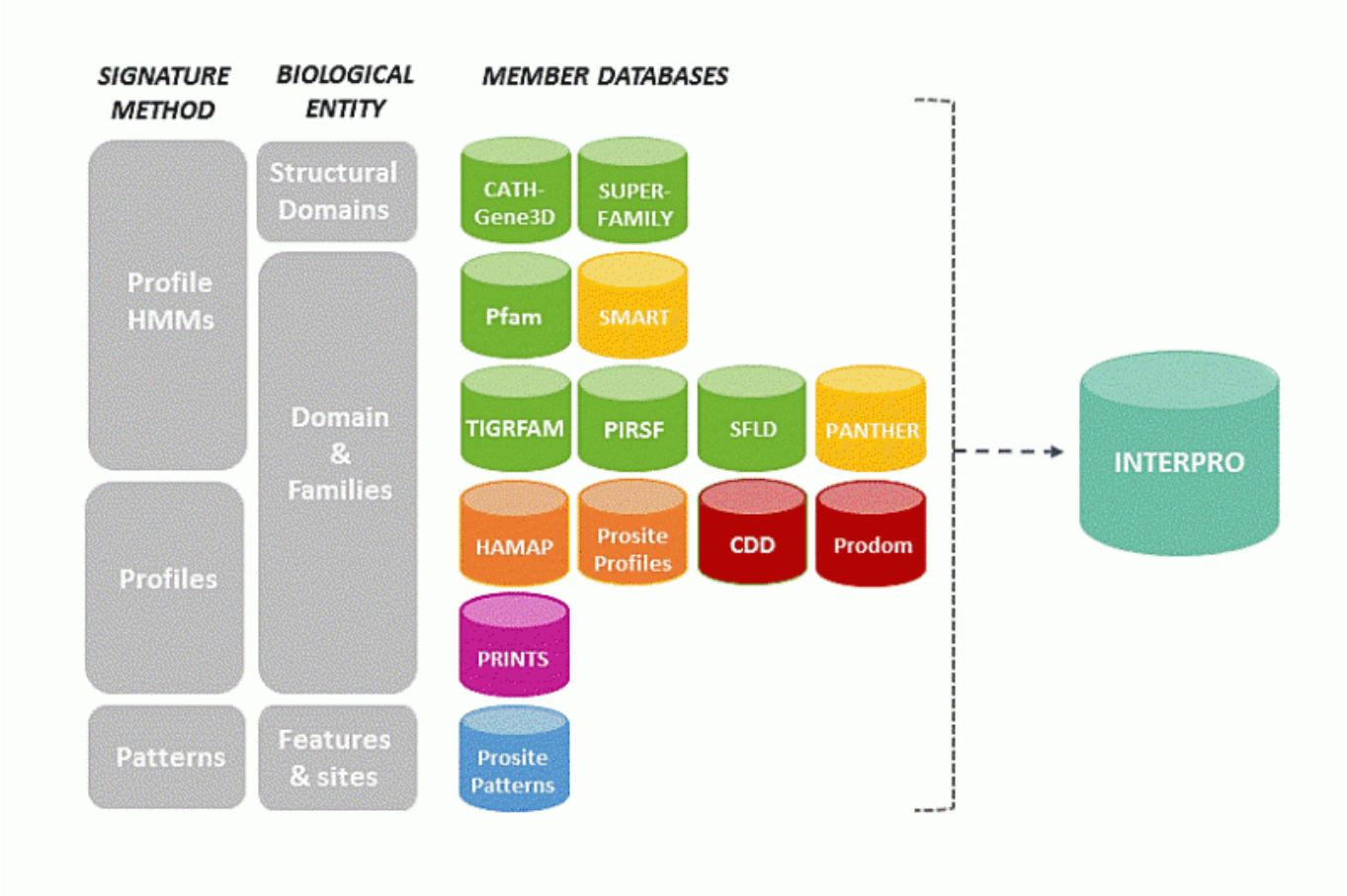

Figure 18 An overview of the different databases that are used to construct InterPro.

When to use InterPro

You can use InterPro [40] if you have an amino acid [2] sequence or set of sequences and you want to know:

- what they are, what family they belong to
- what is their function and how it can be explained in structural terms

You can also use InterPro for a variety of other purposes, such as examining the structural or functional predictions for any sequence already in the UniProt [42] database.

InterPro cannot help you if:

- you want to perform structural alignment of protein sequences
- you have a genomic DNA sequence and are interested in gene annotation [43] (intron [44]/exon [45] predictions, identification of promoter regions, etc).

Find out more about InterPro in our courses InterPro: Quick tour [46] and InterPro: Functional and structural analysis of protein sequences [47]

Summary

Protein classification allows functional and structural properties to be inferred for novel proteins that have not been experimentally characterised.

Proteins can be classified according to the family to which they belong, and/or the domains and features they contain:

- A protein family [4] is a group of proteins that share a common evolutionary origin reflected by their related functions and similarities in sequence and/or structure.
- Domains are distinct functional and/or structural units in a protein that can exist in a variety of biological contexts.
- Sequence features include active sites, binding sites, post-translational modification sites and repeats.

Signatures are mathematical models constructed from multiple sequence alignments that can be used to classify proteins.

Using protein signatures is often a more sensitive way of identifying protein function than pairwise sequence similarity searches, such as BLAST [15].

Different types of signatures use different methods, focussing on single motifs (patterns), multiple motifs (fingerprints) or considering the whole alignment (profiles [16] and HMMs).They offer distinct advantages in terms of protein sequence analysis and can be used to classify proteins into families, or to identify domains or sequence features.

The EBI offers a resource for protein family classification and domain [9] and site prediction using protein signatures: InterPro [40]. InterPro combines signatures from multiple, diverse source databases into a single searchable resource.

Quiz: Introduction to Protein Classification at the EBI

Questions: *5*

Attempts allowed:	*Unlimited*
Available:	*Always*
Pass rate:	*75 %*
Backwards navigation:	*Allowed*

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Learn more

Find out more

Introduction to Bioinformatics. T.K Attwood and DJ Parry-Smith. Addison Wesley Longman, Harlow; 1999; ISBN 0 582 327881

Bioinformatics: Genes, Proteins & Computers .Orengo C.A.. et al.. Advanced Text, BIOS Scientific Publishers, 2003

Madera M. and Gough J. 2002. A comparison of profile hidden Markov model procedures for remote homology detection. [48] *Nucleic Acids Research*, 30 (19) 4321-4328

Thompson J.D., Plewniak F., Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. [49] *Nucleic Acids Research*, 27 (13) 2682-2690

Recommended courses

Online course: InterPro: Quick tour [50]

Online course: InterPro: Functional and structural analysis of protein sequences [47]

The EBI offers hands-on courses coveraing a range of subjects, including training on InterPro. An up-to-date list of training courses is available here [51].

Get help and support on InterPro

Contact

For all support enquiries, please contact the InterPro [40] helpdesk (interhelp [at] ebi.ac.uk).

Further help

On the InterPro website [52] you can find more help, including:

- Information on how to use InterPro [53]
- more about InterPro [54]
- InterPro documentation [55]

References

Bairoch, A. 1991 PROSITE: a dictionary of sites and patterns in proteins. [56] Nucleic Acids Res. 19, 2241–2245.

Eddy, S.R.1998. Profile hidden Markov models. [57] Bioinformatics. 14: 755-63

Gribskov M. et al. 1987. Profile analysis: detection of distantly related proteins. [58] Proc Natl Acad Sci U S A. 84(13): 4355-8.

Lima T, et al.2009. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. [59] Nucleic Acids Res. 37:D471-8.

Marchler-Bauer A. et al. 2015. CDD: NCBI's conserved domain database [60]. Nucleic Acids Res., 43, D222–226.

Servant F et al. 2002. ProDom: automated clustering of homologous domains. [61] Brief. Bioinformatics.3:246-51.

Sigrist CJ et al.2010. PROSITE, a protein domain database for functional characterization and annotation. [62] Nucleic Acids Res. 38:D161-6.

The PRINTS [29] protein fingerprint database: functional and evolutionary applications. Attwood, TK et al. Encyclopaedia of Genetics, Genomics, Proteomics [63] and Bioinformatics. 2006. John Wiley & Sons, Ltd.

Contributors

[64]

Alex Mitchell **[64]**

EMBL-EBI
Curation co-ordinator, InterPro and EBI Metagenomics

Alex Mitchell is co-ordinator for the InterPro and EBI Metagenomics databases at EMBL-EBI. He obtained his DPhil in pharmacology from the University of Oxford, and was previously employed as a molecular biologist at the Institute of Psychiatry. He moved to the University of Manchester to work on protein sequence analysis and functional classification, before joining EMBL-EBI in 2011.


[1]

[Amaia Sangrador](#) **[1]**

EMBL-EBI
Scientific curator, InterPro

Amaia Sangrador is a curator for InterPro at the European Bioinformatics Institute in Cambridge, UK. She joined EMBL-EBI in 2010. She has a PhD in Molecular Biology and has been involved in several research projects in the areas of immunology and comparative genomics.

Links
[1] https://www.ebi.ac.uk/training/online/trainers/a.sangrador
[2] https://www.ebi.ac.uk/training/online/glossary/amino-acid
[3] https://www.ebi.ac.uk/training/online/glossary/homologous
[4] https://www.ebi.ac.uk/training/online/glossary/protein-family
[5] https://www.ebi.ac.uk/training/online/glossary/ancestor
[6] https://www.ebi.ac.uk/training/online/glossary/gpcrs
[7] https://www.ebi.ac.uk/training/online/glossary/extracellular
[8] http://en.wikipedia.org/wiki/G_protein-coupled_receptor
[9] https://www.ebi.ac.uk/training/online/glossary/domain
[10] http://en.wikipedia.org/wiki/Regulator_of_G_protein_signalling
[11] https://www.ebi.ac.uk/training/online/glossary/binding-site
[12] https://www.ebi.ac.uk/training/online/glossary/translation
[13] https://www.ebi.ac.uk/training/online/glossary/active-site
[14] http://en.wikipedia.org/wiki/Ferredoxin
[15] https://www.ebi.ac.uk/training/online/glossary/blast
[16] https://www.ebi.ac.uk/training/online/glossary/profiles
[17] https://www.ebi.ac.uk/training/online/glossary/motif
[18] https://www.ebi.ac.uk/training/online/glossary/pattern
[19] https://www.ebi.ac.uk/training/online/glossary/regular-expression
[20] http://en.wikipedia.org/wiki/Regular_expression
[21] https://www.ebi.ac.uk/training/online/glossary/prosite
[22] https://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi/references
[23] https://www.ebi.ac.uk/training/online/glossary/blosum-matrices
[24] http://en.wikipedia.org/wiki/Position-Specific_Scoring_Matrix
[25] http://en.wikipedia.org/wiki/Substitution_matrix
[26] https://www.ebi.ac.uk/training/online/glossary/hamap
[27] https://www.ebi.ac.uk/training/online/glossary/prodom
[28] https://www.ebi.ac.uk/training/online/glossary/psi-blast
[29] https://www.ebi.ac.uk/training/online/glossary/prints
[30] https://www.ebi.ac.uk/training/online/glossary/specificity
[31] https://www.ebi.ac.uk/training/online/glossary/ligand
[32] https://www.ebi.ac.uk/training/online/glossary/likelihood
[33] https://www.ebi.ac.uk/training/online/glossary/pfam

[34] https://www.ebi.ac.uk/training/online/glossary/smart
[35] https://www.ebi.ac.uk/training/online/glossary/pirsf
[36] https://www.ebi.ac.uk/training/online/glossary/panther
[37] https://www.ebi.ac.uk/training/online/glossary/sfld
[38] https://www.ebi.ac.uk/training/online/glossary/superfamily
[39] https://www.ebi.ac.uk/training/online/glossary/gene3d
[40] https://www.ebi.ac.uk/training/online/glossary/interpro
[41] https://www.ebi.ac.uk/training/online/glossary/annotation
[42] https://www.ebi.ac.uk/training/online/glossary/uniprot
[43] https://www.ebi.ac.uk/training/online/glossary/gene-annotation
[44] https://www.ebi.ac.uk/training/online/glossary/intron
[45] https://www.ebi.ac.uk/training/online/glossary/exon
[46] https://www.ebi.ac.uk/training/online/course/interpro-quick-tour
[47] https://www.ebi.ac.uk/training/online/course/interpro-functional-and-structural-analysis-protei
[48] http://europepmc.org/abstract/MED/12364612
[49] http://europepmc.org/abstract/MED/10373585
[50] https://www.ebi.ac.uk/training/online/../../course/interpro-quick-tour
[51] http://www.ebi.ac.uk/training/handson/
[52] http://wwwdev.ebi.ac.uk/interpro/
[53] http://wwwdev.ebi.ac.uk/interpro/user_manual.html
[54] http://wwwdev.ebi.ac.uk/interpro/about.html
[55] http://wwwdev.ebi.ac.uk/interpro/documentation.html
[56] http://europepmc.org/abstract/MED/1598232
[57] http://europepmc.org/abstract/MED/9918945
[58] http://europepmc.org/abstract/MED/3474607
[59] http://europepmc.org/abstract/MED/18849571
[60] http://europepmc.org/abstract/MED/25414356
[61] http://europepmc.org/abstract/MED/12230033
[62] http://europepmc.org/abstract/MED/19858104
[63] https://www.ebi.ac.uk/training/online/glossary/proteomics
[64] https://www.ebi.ac.uk/training/online/trainers/mitchell