

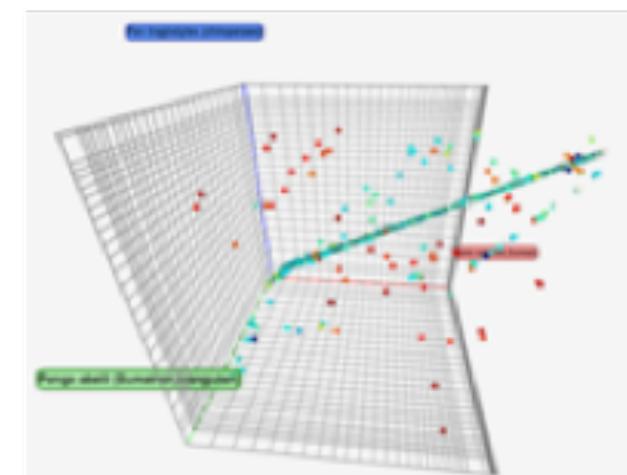
# Computational Genomics

# Introduction To Databases

## Finding:

### Genes, and Genomes

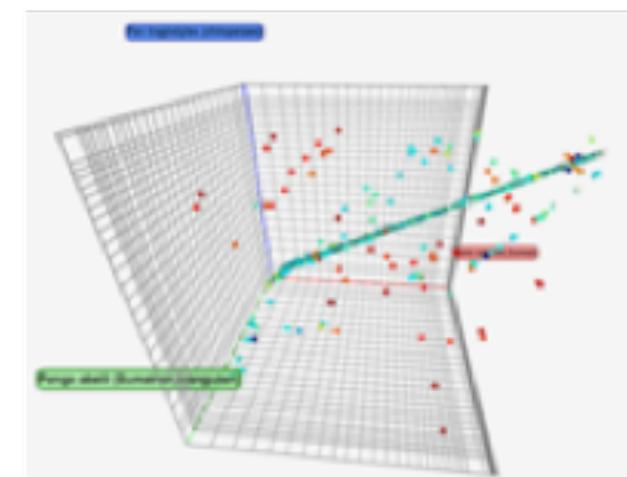
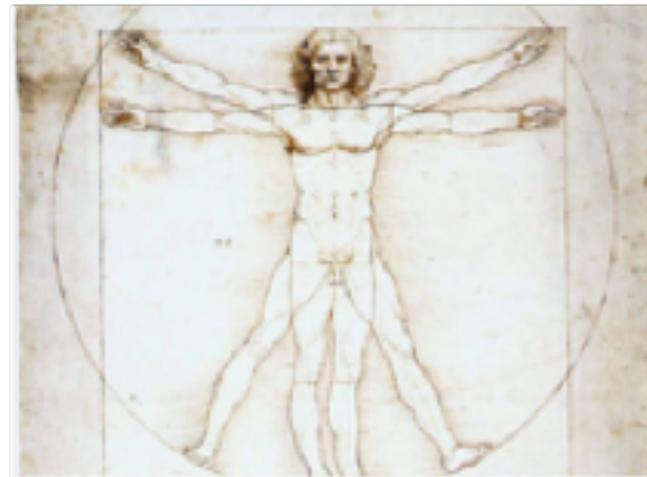
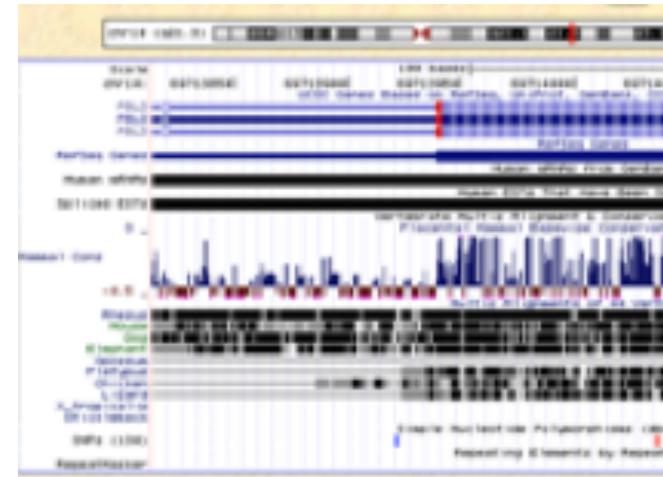
### Proteins and Proteomes



# Computational Genomics

## Introduction To Databases

### Protein Databases



# Protein Databases



Swiss Institute of  
Bioinformatics

**Expsy**  
Swiss Bioinformatics Resource Portal

e.g. BLAST, UniProt, MSH6, Albumin...

**SIB Resources**

- Genes & Genomes
  - Genomics
  - Metagenomics
  - Transcriptomics
- Proteins & Proteomes
  - UniProtKB/Swiss-Prot
  - ASAP
  - Cellosaurus
- Evolution & Phylogeny
  - Evolution biology
  - Population genetics
- Structural Biology
  - Drug design
  - Medicinal chemistry
  - Structural analysis
- Systems Biology
  - Glycomics
  - Lipidomics
  - Metabolomics
- Text mining & Machine learning

Home About SIB News Contact

Search

**InterPro** Classification of protein families

Home Search Browse Results Release notes Download Help About

InterPro 87.0 17 November 2021

## Classification of protein families

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.

► Citing InterPro

## Pfam 35.0 (November 2021, 19632 entries)

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 ENTRY** View Pfam annotation and alignments

**VIEW A CLAN** See groups of related entries

**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**  **Go** **Example**

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

Or view the [help](#) pages for more information

**UniProtKB**  
UniProt Knowledgebase

Swiss-Prot (565,928)  
Manually annotated and reviewed.  
Records with information extracted from literature and curator-evaluated computational analysis.

TiEMBL (225,013,025)  
Automatically annotated and not reviewed.  
Records that await full manual annotation.

**UniRef**

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

**UniParc**

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

**Proteomes**

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

Supporting data

Literature citations  
Cross-ref. databases

Taxonomy  
Diseases

Subcellular locations  
Keywords

# **EMBL-EBI Training**

Delivering world-class training in data-driven life sciences.

# Protein Databases

## Expsy



Swiss Institute of  
Bioinformatics

Home

About

SIB News

Contact

# Expsy

Swiss Bioinformatics Resource Portal



e.g. [BLAST](#), [UniProt](#), [MSH6](#), [Albumin](#)...

## SIB Resources ⓘ



### Glyco@Expsy

Zooming in on web-based glycoinformatics resources.



### SwissOrthology

One-stop shop for orthologs



### neXtProt

Human protein knowledgebase



### Nextstrain

Impact of pathogen genome data on science and public health



### UniProtKB/Swiss-Prot

Protein knowledgebase



### SWISS-MODEL

Protein structure homology-modelling



### Rhea

Expert-curated database of biochemical reactions.



### V-pipe

Viral genomics pipeline



### ASAP

Web-based, cooperative portal for single-cell data analyses



### SwissRegulon Portal

Tools and data for regulatory genomics



### SwissLipids

Knowledge resource for lipids



### EPD

Eukaryotic Promoter Database



### Cellosaurus

Knowledge resource on cell lines



### SwissDrugDesign

Widening access to computer-aided drug design



### STRING

Protein-protein interaction networks and enrichment analysis



### Bgee

Gene expression expertise

## Genes & Genomes

- Genomics
- Metagenomics
- Transcriptomics

## Proteins & Proteomes

- Evolution biology
- Population genetics

## Structural Biology

- Drug design
- Medicinal chemistry
- Structural analysis

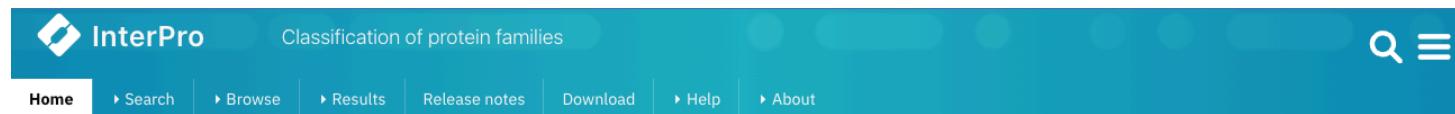
## Systems Biology

- Glycomics
- Lipidomics
- Metabolomics

## Text mining & Machine learning

# Protein Databases

## InterPro



The header features the InterPro logo with a blue diamond icon, followed by the text "InterPro" and "Classification of protein families". A search bar with a magnifying glass icon and a menu icon are on the right.

Home    ▶ Search    ▶ Browse    ▶ Results    Release notes    Download    ▶ Help    ▶ About

### In the spotlight



#### InterPro 87.0 New And Updated Features

By Typhaine Paysan-Lafosse

The latest release of InterPro comes with a number of improvements to the website and API. The list of changes is

[Read more →](#)



#### AlphaFold, A Novel Tool Sheding Light On Alzheimer's Disease Proteins

By Sara Chuguransky

Alzheimer's disease (AD) is the most common type of dementia, affecting around 50 million patients worldwide.

[Read more →](#)



#### AlphaFold Structure Predictions Available In InterPro

By Matthias Blum

AlphaFold 2.0 has revolutionised structure prediction enabling the rapid creation of high quality models across many model

[Read more →](#)



#### InterPro 86.0 Key Features And Bug Fixes

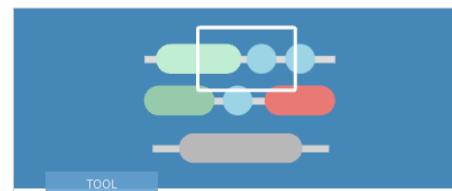
By Typhaine Paysan-Lafosse

The latest release of InterPro comes with a number of improvements to the website. The list of changes is detailed

[Read more →](#)

[Read all articles](#)

### Tools & libraries



#### InterProScan

InterProScan is the software package that allows sequences (protein and nucleic) to be scanned against InterPro's signatures. Signatures are predictive models, provided by several



[Read more →](#)

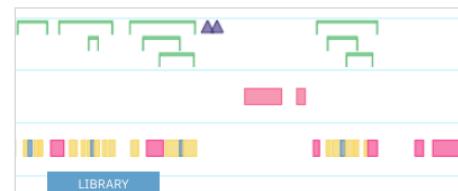


#### A new API for InterPro

You can now skip URL and use this JSON interface to work with your data directly. Currently there are 6 main endpoints: entry, protein, structure, taxonomy, proteome and set.



[Read more →](#)



#### Nightingale

Nightingale is a monorepo containing visualisation web components, including the formerly known Protvista, a powerful and blazing-fast tool for handling protein sequence



[Read more →](#)

# Protein Databases

## InterPro

Where does the data come from?

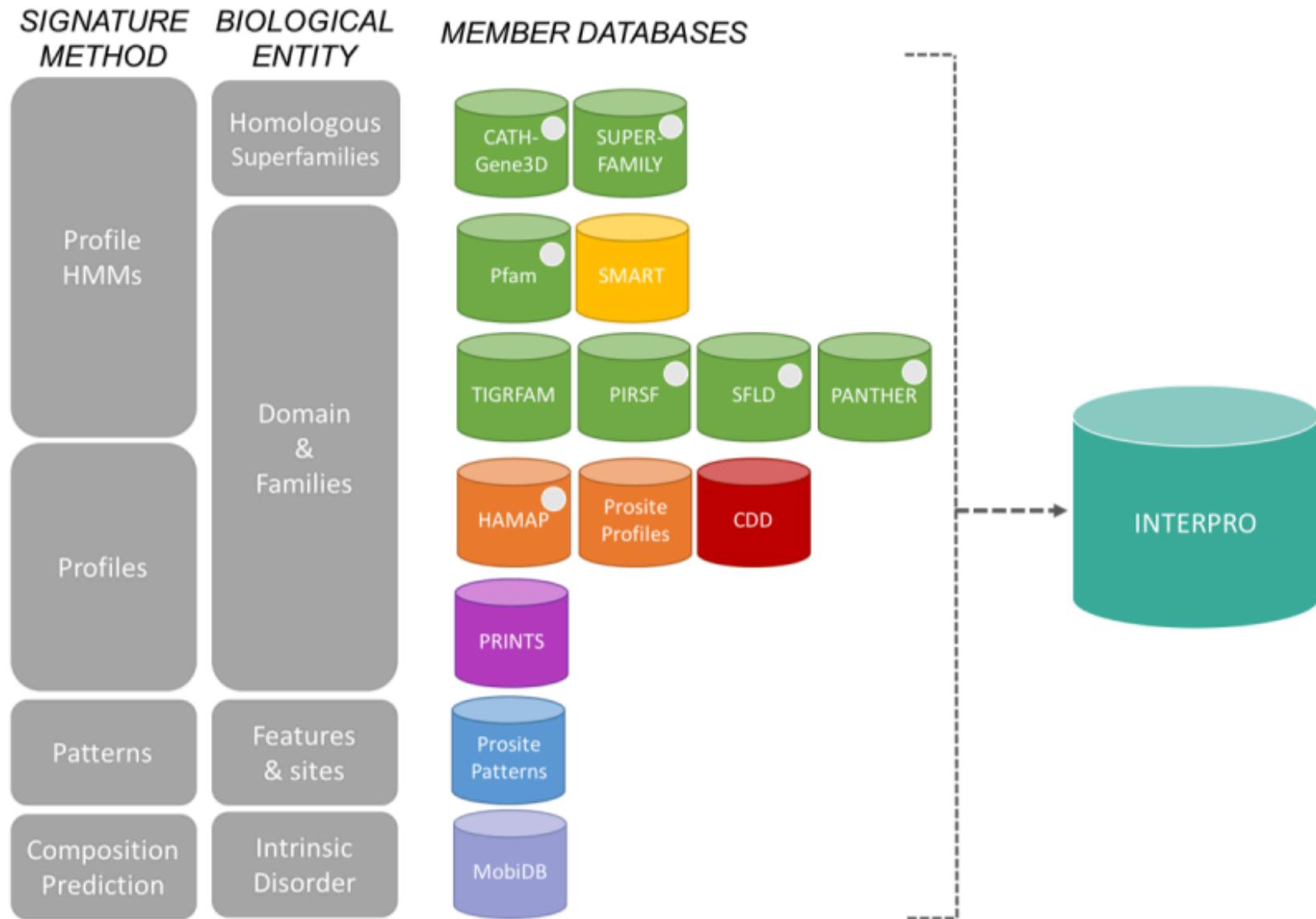
In order to classify proteins into families and to predict the presence of important domains or sequence features, we require computational tools. One such set of tools are predictive models known as protein signatures.

Signatures are built by the member databases in the InterPro consortium.

- Different member databases use different methods to construct their signatures, and they have their own particular focus of interest: structural and/or functional domains, protein families, or protein features such as active sites or binding sites.

# Protein Databases

## InterPro



# **Protein Databases**

## **InterPro**

### Why do we need InterPro?

Protein signature databases have become vital tools for classification of protein sequences in order to infer their function. Many protein signature-based resources are now available, each with their own strengths and weaknesses.

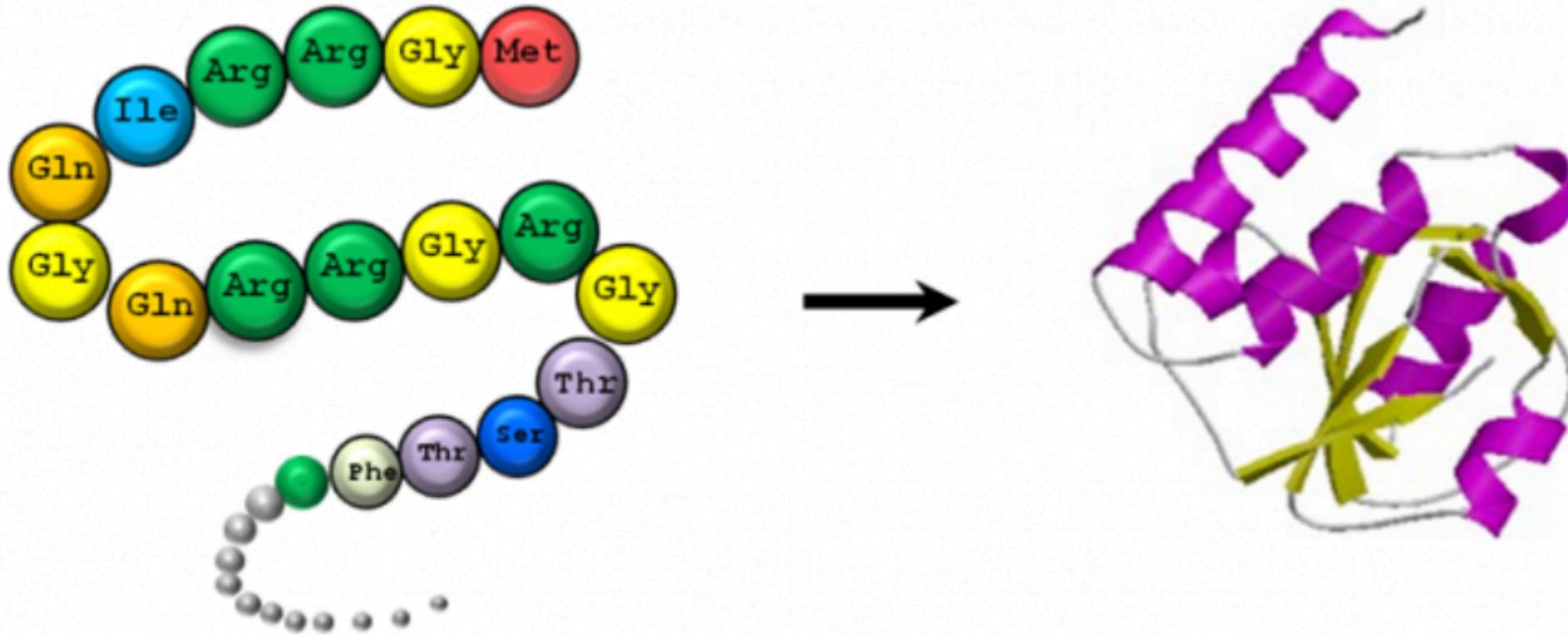
We need InterPro because it:

1. Reduces redundancy and simplifies protein sequence analysis by integrating signatures from different member databases that represent the same protein family, domain or site.
2. Unites the member databases, capitalising on their individual strengths to produce a powerful classification tool.
3. Provides a single convenient searchable location, allowing simultaneous querying of all member databases.
4. Adds information (including descriptive abstracts and Gene Ontology terms) to the signatures, which may be used to annotate the proteins they match.

# Protein Databases

## InterPro

### Protein Classification



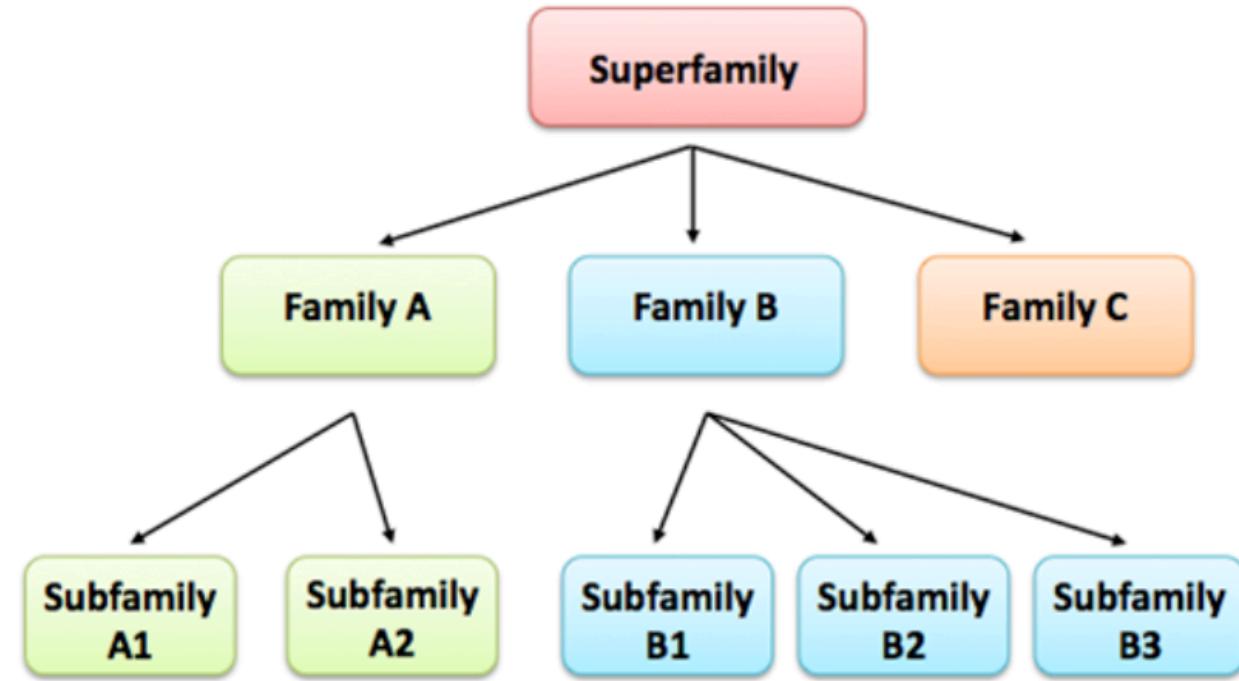
**Figure 1** Proteins consist of one or more polypeptides. A polypeptide is a chain of amino acids. The polypeptide chains fold into their final three-dimensional structure to constitute a functional protein. The amino acid sequence and structure in this example correspond to ribosomal protein L2.

# Protein Databases

## InterPro

### What are protein families?

- A protein family 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 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.



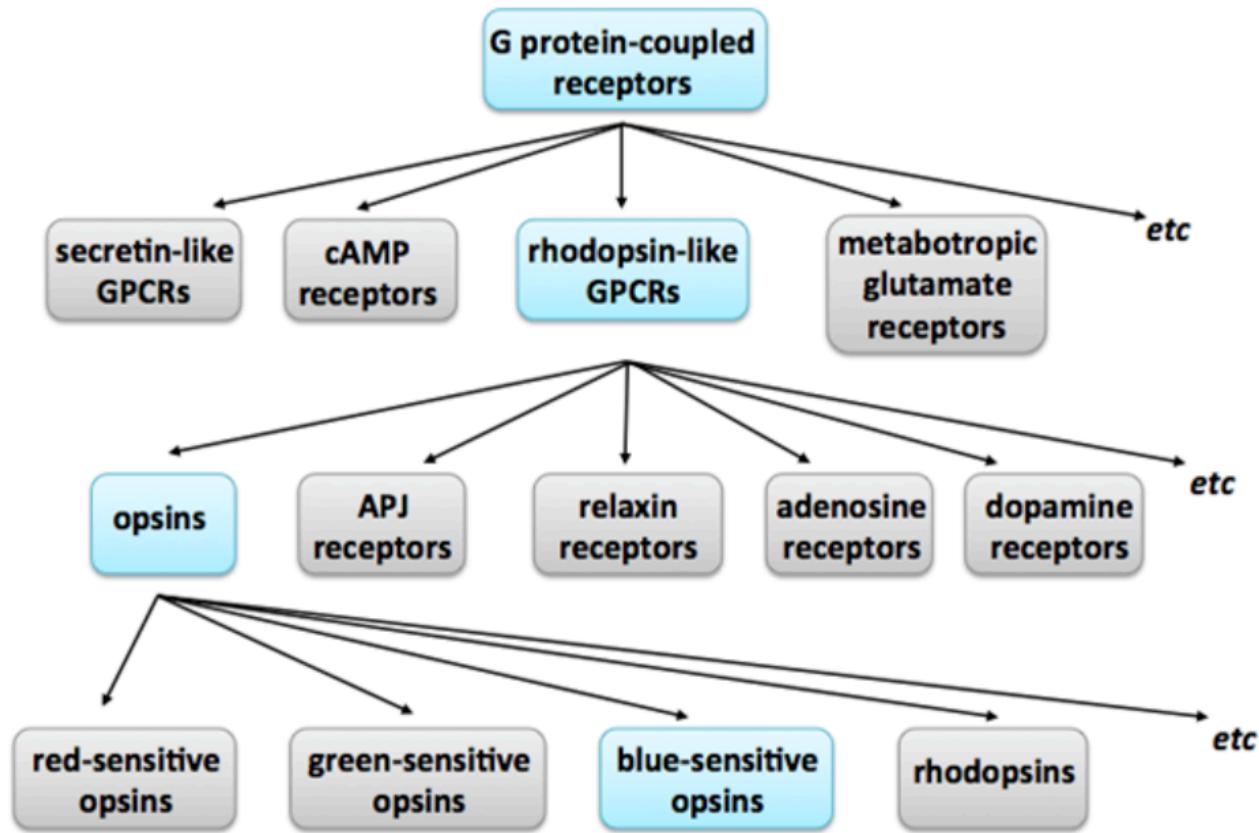
**Figure 2** A hypothetical protein family hierarchy showing the relationships between superfamily, family and subfamily members. Directional arrows indicate that one group is a subgroup of another.

# Protein Databases

## InterPro

### What are protein families?

One set of proteins that comprise a superfamily are the G protein-coupled receptors (GPCRs). 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 specialized proteins (called G proteins) to influence intracellular pathways after binding extracellular signals (you can visit this GPCR webpage 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.

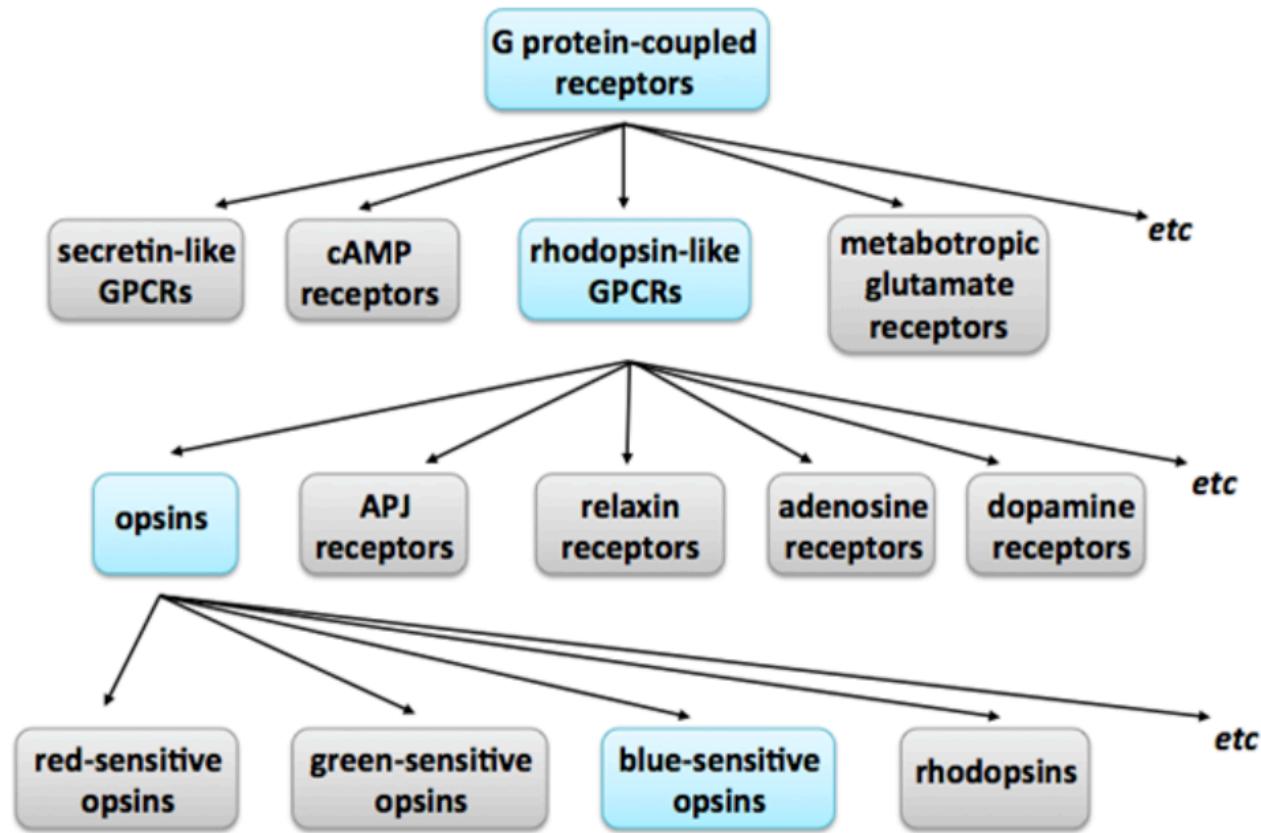
# Protein Databases

## InterPro

### What are protein families?

As we group the GPCRs 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 hierarchy is illustrated in Figure 3.

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.



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

# Protein Databases

## InterPro

### 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 residues that are involved in protein-protein interactions. SH3 domains have a characteristic 3D structure (Figure 4). They occur in a diverse range of proteins with different functions, including adaptor proteins, phosphatidylinositol 3-kinases, phospholipases and myosins.



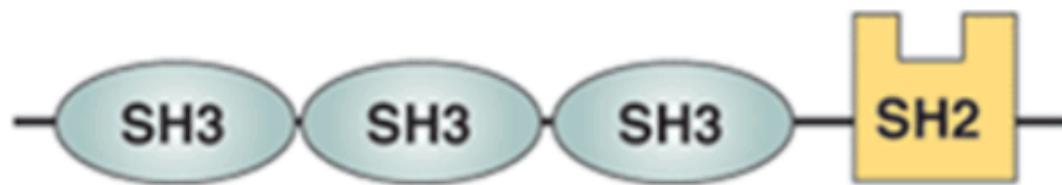
**Figure 4** Structure of the SH3 domain.

# Protein Databases

## InterPro

### What are protein domains?

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 composition of Nck is illustrated in Figure 5.



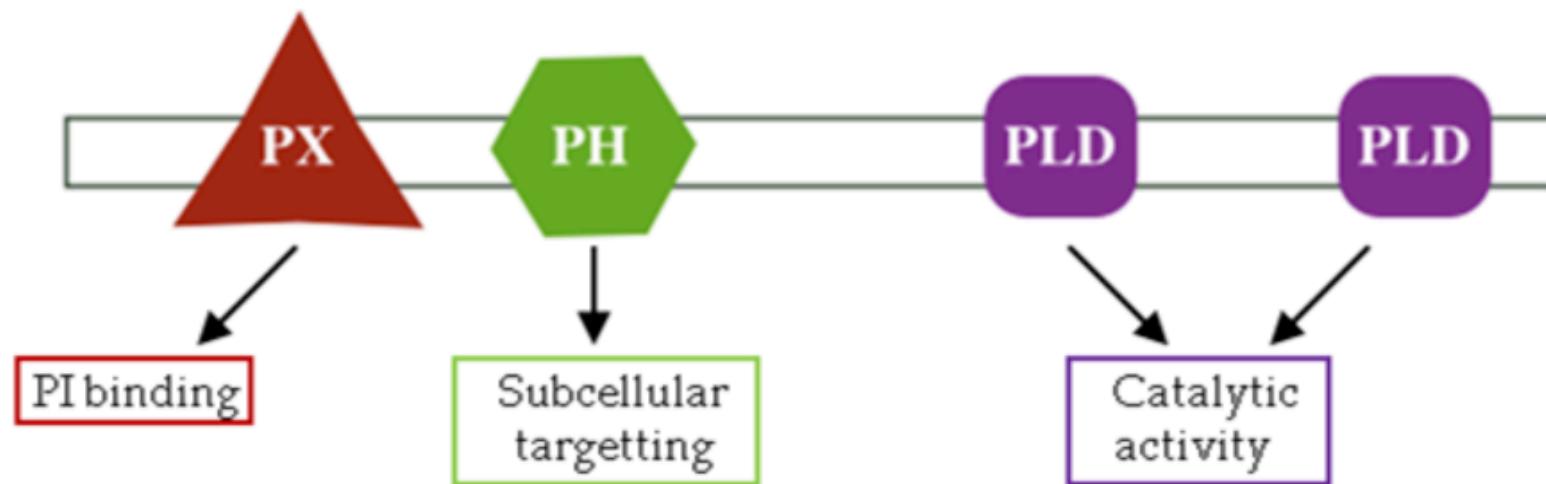
**Figure 5** Domain composition of Nck. Nck contains three SH3 domains plus another domain known as SH2 (Src homology 2). Both SH3 and SH2 domains are usually found in proteins that interact with other proteins and mediate assembly of protein complexes. SH3 domains typically bind to proline-rich peptides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

# Protein Databases

## InterPro

### 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).
- 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.



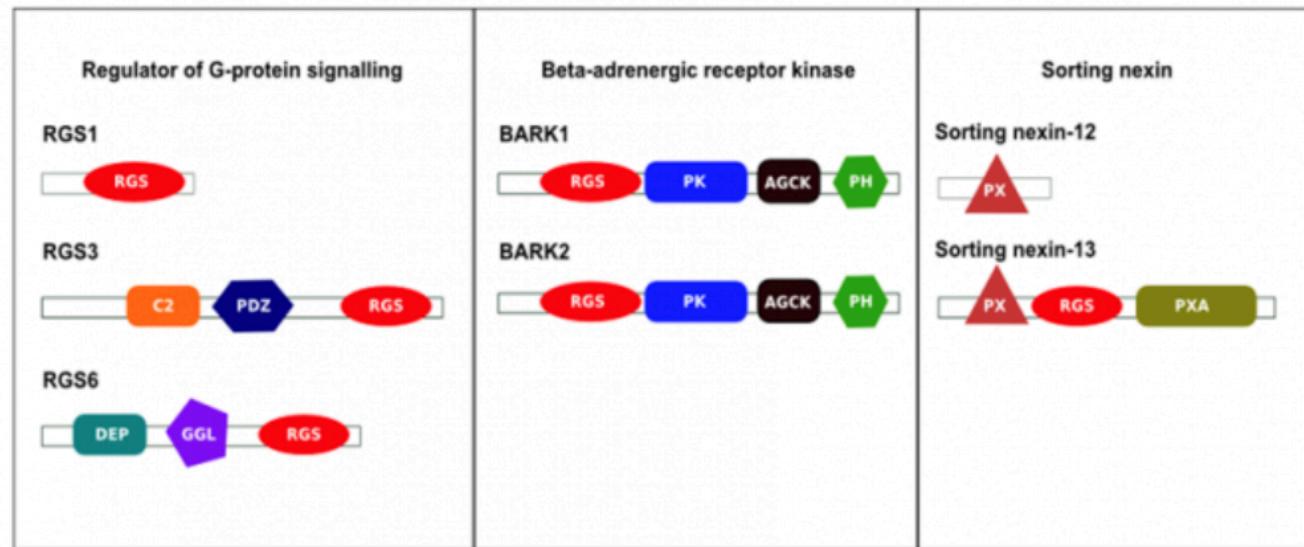
**Figure 6** Domain composition of phospholipase D1, which is an enzyme that breaks down phosphatidylcholine.

# Protein Databases

## InterPro

### Family- and domain-based protein classification

- Family- and domain-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.
- Regulator of G-protein signalling (RGS) domains are protein structural units that activate GTPases. They are found in sequences that belong to the RGS protein family, 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.



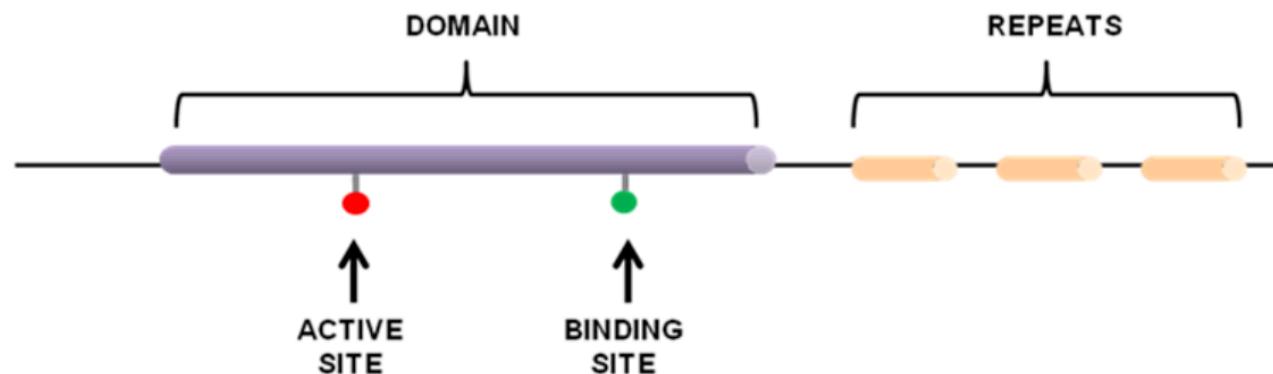
**Figure 7** Family groupings and domain composition of some RGS domain-containing proteins.

# Protein Databases

## InterPro

### 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, 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 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 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.
  - 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 (Figure 8). Sequence features are often nested within domains – a protein kinase domain, for example, usually contains a protein kinase active site.



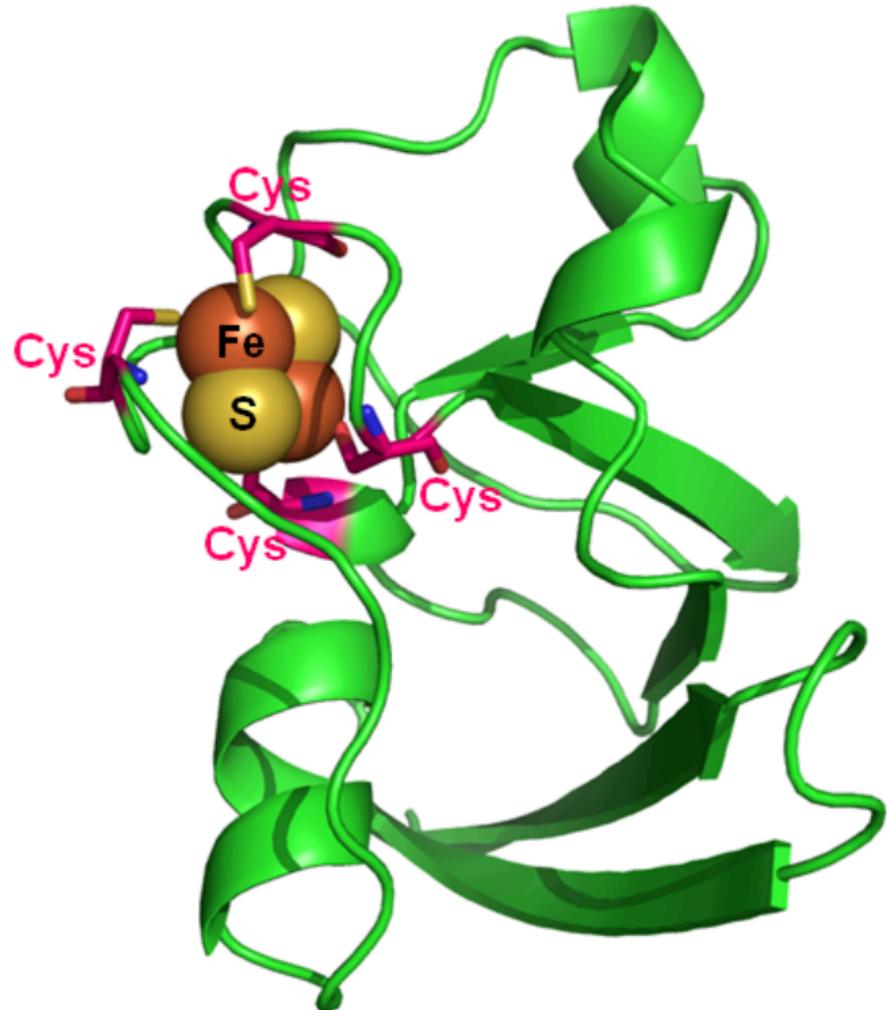
**Figure 8** Graphical representation of repeats, domains and sites on a protein sequence.

# Protein Databases

## InterPro

### 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](#)).
- 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 is shown on the ferredoxin 3D-structure in Figure 9.



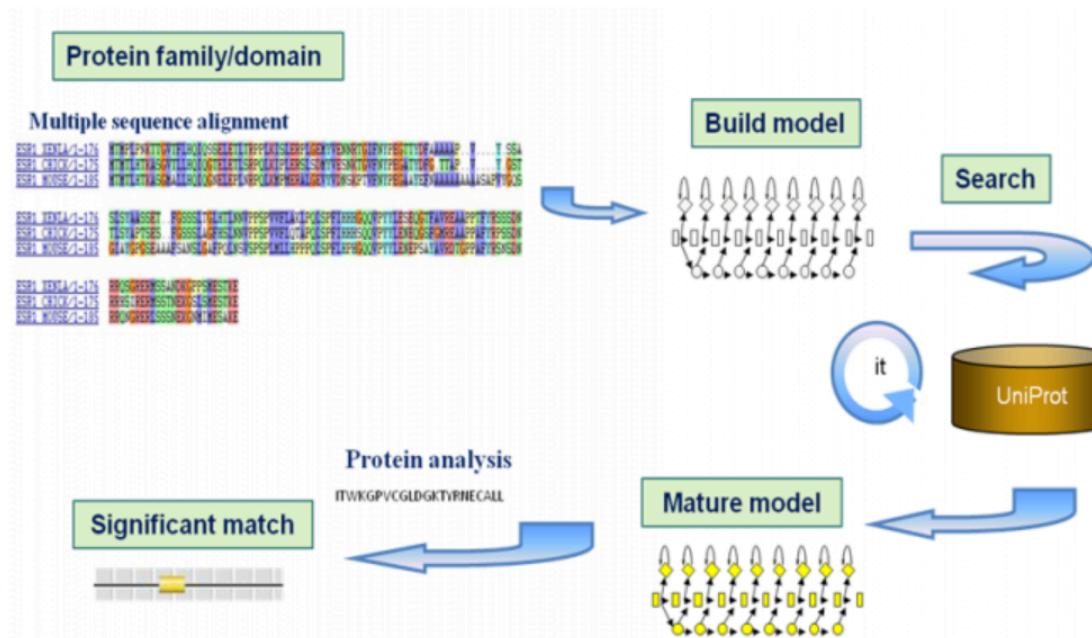
**Figure 9** 3D-structure of a plant-type ferredoxin with its 2Fe-2S cluster. The conserved cysteine (Cys) residues that help form the binding site are highlighted in red. The iron and sulphur atoms bound to the cysteines are displayed as spheres.

# Protein Databases

## InterPro

### 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) (Figure 10). When building the initial model, the level of amino acid 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.



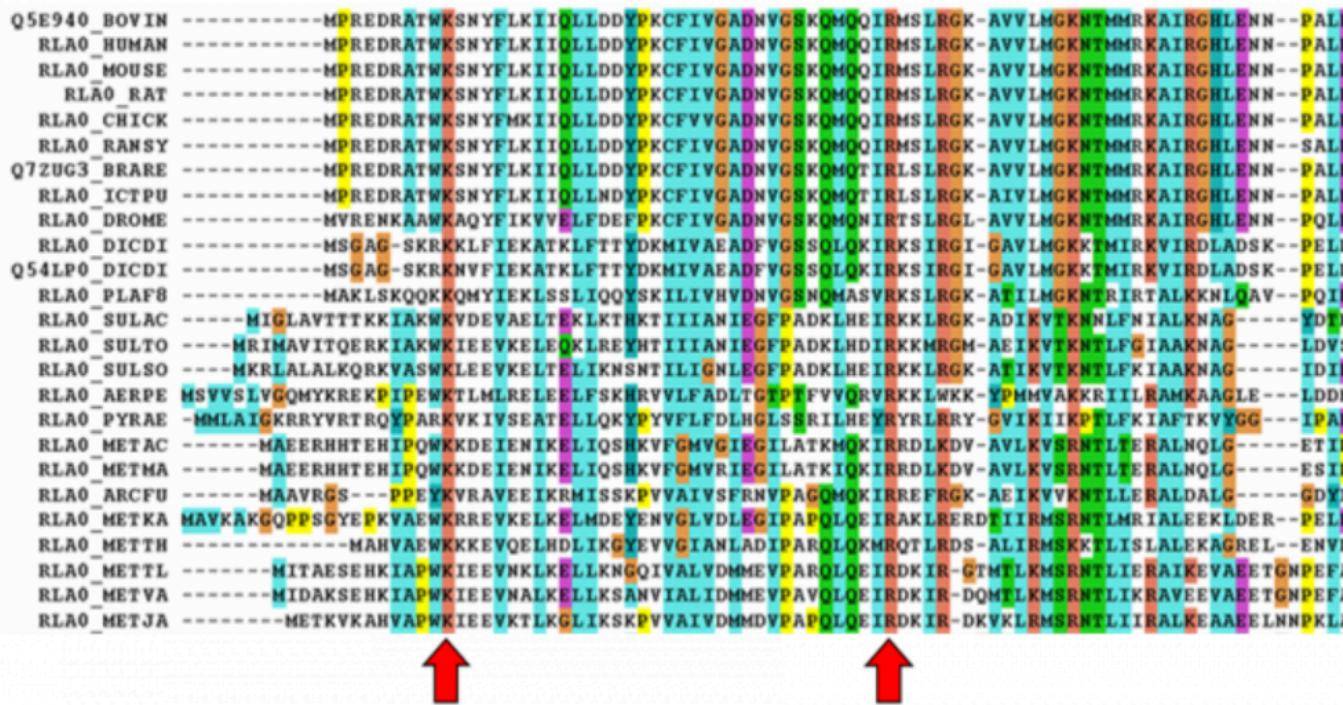
**Figure 10** The process of building a protein signature starts with a multiple sequence alignment, which is used to build a predictive model. By searching a protein database in an iterative way, more distantly related sequences can be identified. This information is used to create a final mature model.

# Protein Databases

## InterPro

### 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 residues that are conserved in distantly related proteins (Figure 11). It is not possible to identify such important residues with pairwise alignment techniques, such as BLAST. As a consequence, protein signatures built from multiple sequence alignments are usually better at detecting divergent homologues than pairwise comparison methods.



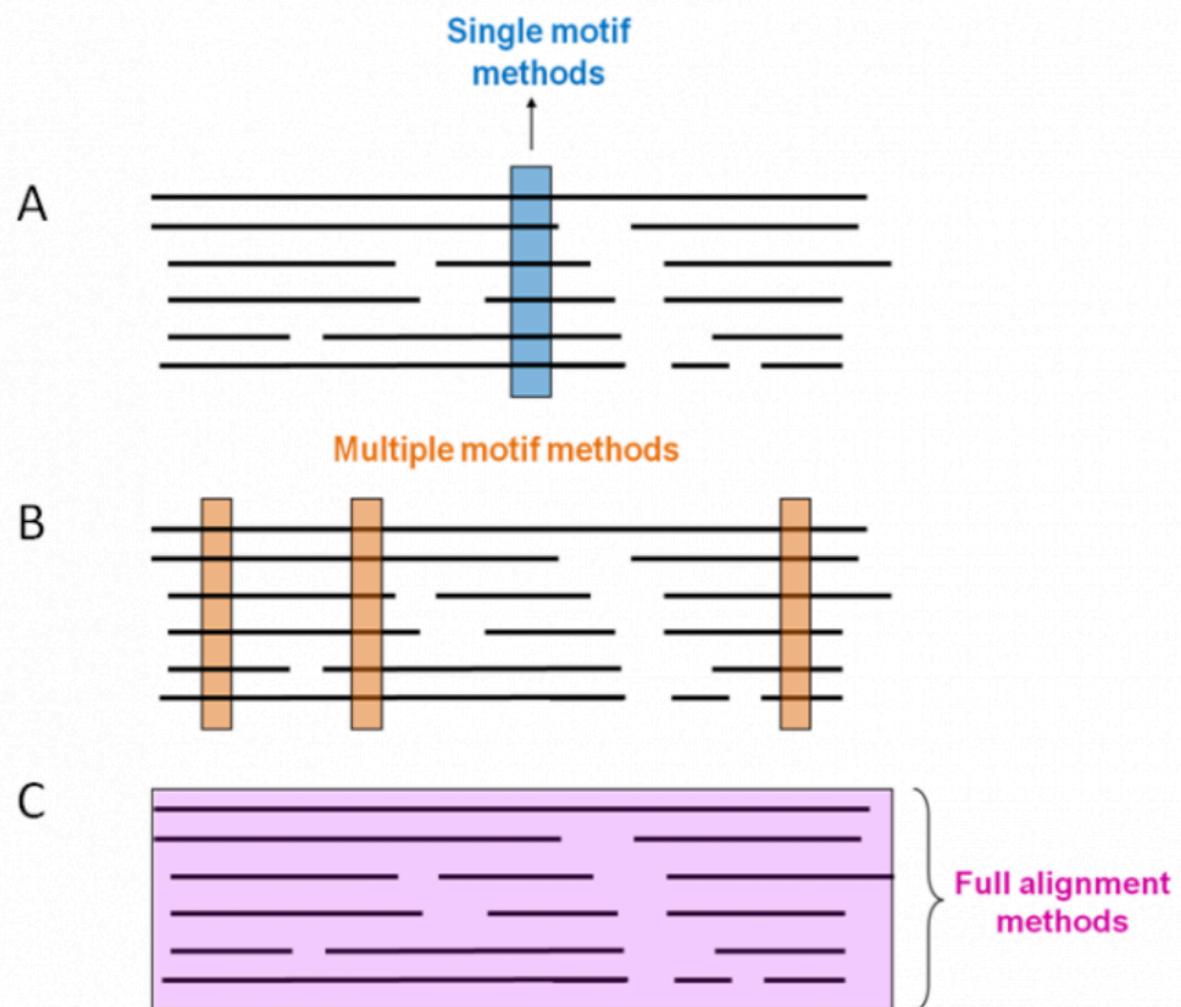
**Figure 11** Multiple sequence alignment for 60S acidic ribosomal protein P0 from different organisms (eukaryota and archaea). There are two amino acids indicated by red arrows, lysine (K) and arginine (R), that are conserved in all sequences. Multiple sequence alignment methods are important for identifying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

# Protein Databases

## InterPro

### Signature types

- Different approaches can be used to generate signatures. These include:
  - patterns
  - profiles
  - 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), multiple conserved motifs, or the full alignment of the entire protein or a particular domain (Figure 12).



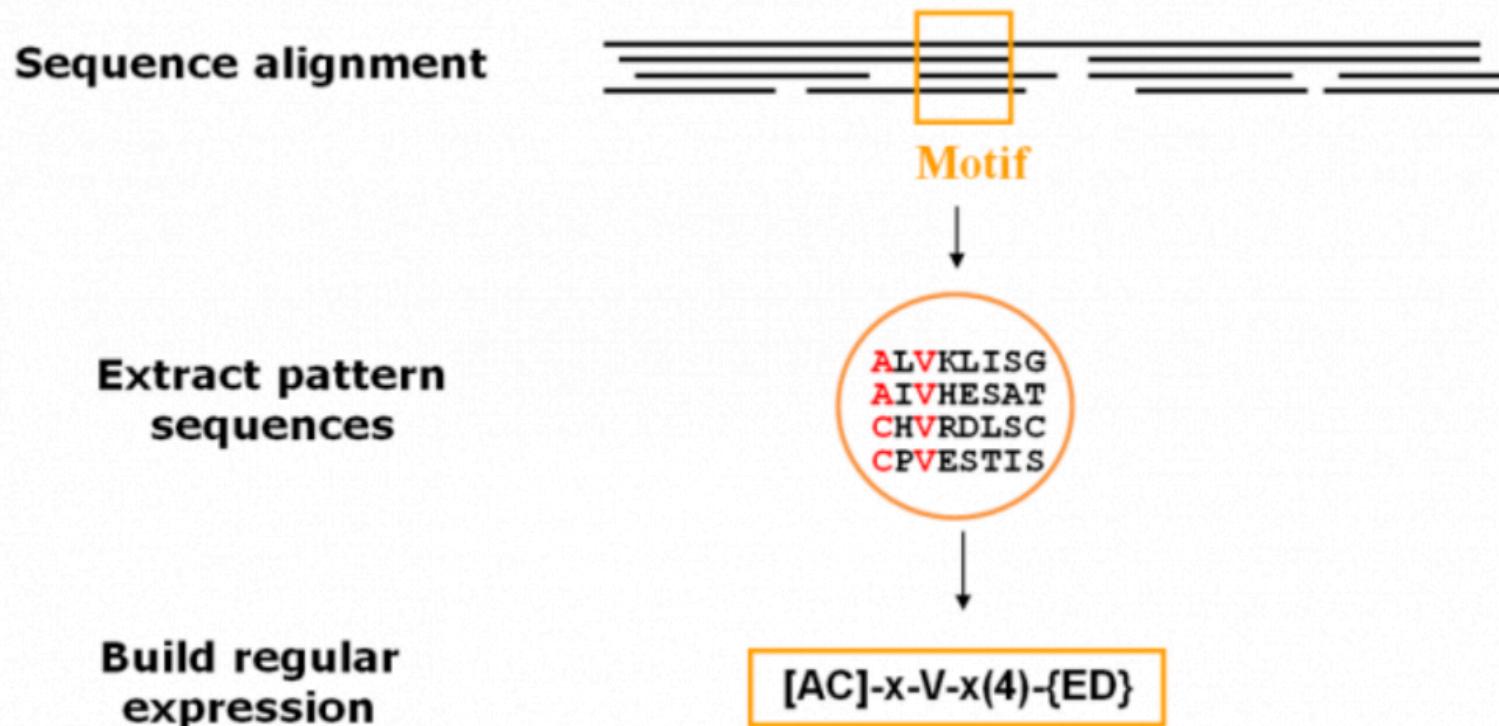
**Figure 12** Representation of different strategies for building signatures. A) single motif methods, B) multiple motif methods; and C) full alignment methods. Protein multiple sequence alignments are represented by black lines and the conserved regions used to build the signatures are indicated by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

# Protein Databases

## InterPro

### 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 of conservation within the sequence feature is then modelled as a regular expression, as is indicated in Figure 13. An example of a database that uses patterns is PROSITE.



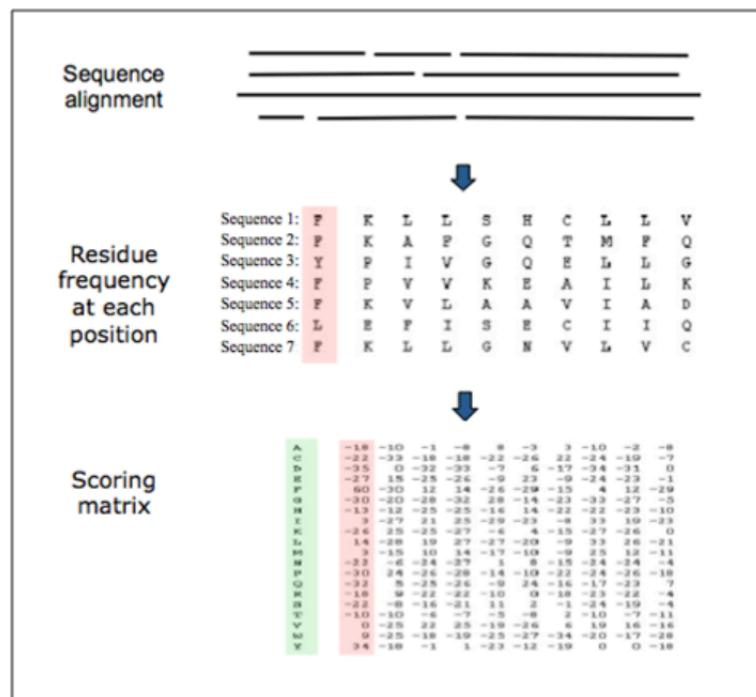
**Figure 13** When creating patterns, a conserved motif is used to build a regular expression. The pattern illustrated here is translated as: [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}.

# Protein Databases

## InterPro

### What are profiles

- Profiles 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) can be used to add evolutionary distance weighting these scores.
- Examples of databases that use profiles to classify proteins include CDD [2], HAMAP [3] and PROSITE (which produces profiles as well as patterns) [4]. The PRODOM [5] database also uses a related approach, using PSI-BLAST to create its profiles.



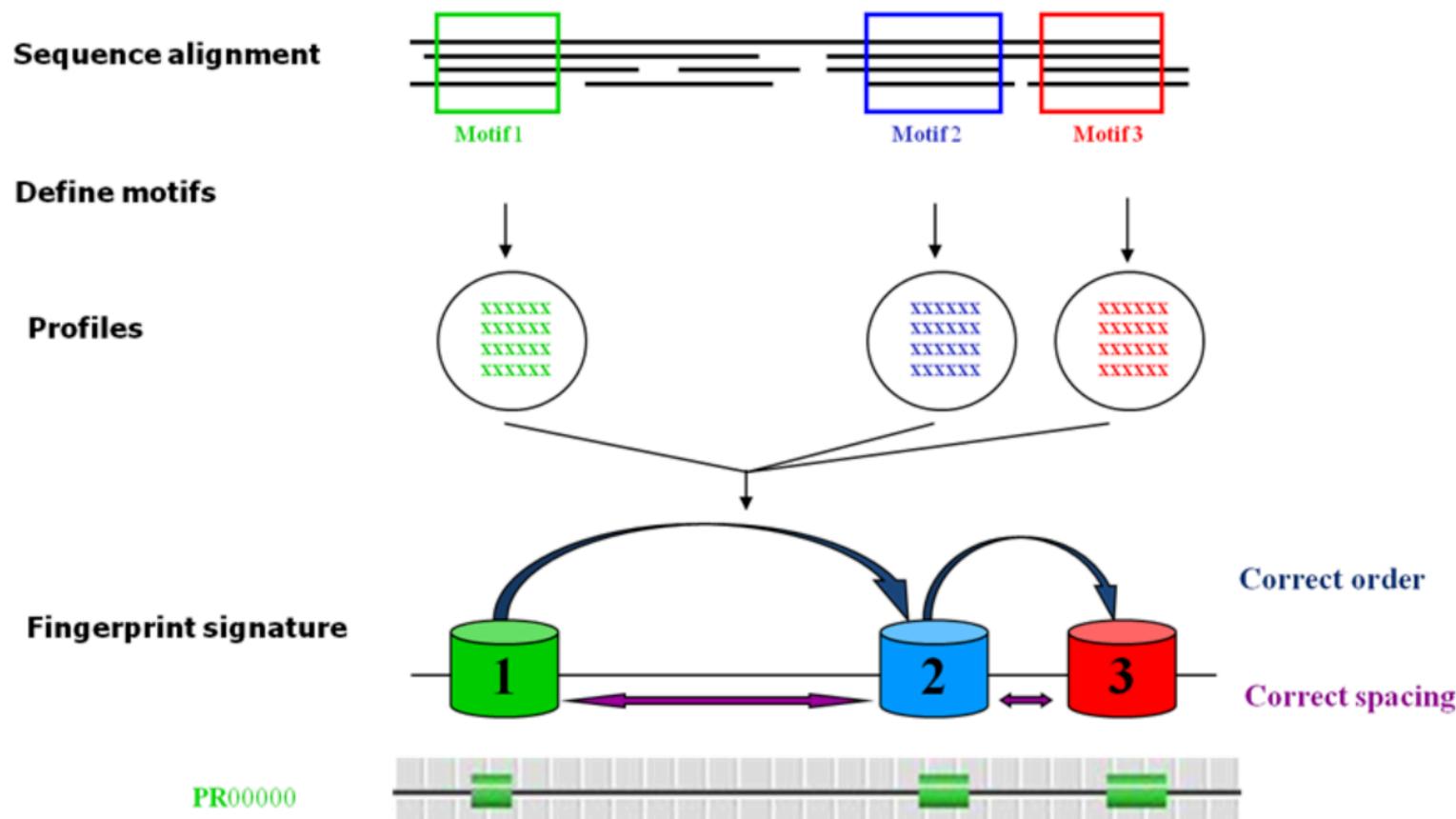
**Figure 14** Representation of a scoring matrix based on a multiple sequence alignment. Each of the 20 amino acids commonly found in proteins is given a score for each position in the sequence according to the frequency with which they occur in the original alignment. Other factors, such as evolutionary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

# Protein Databases

## InterPro

### What are fingerprints

- While single motif 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.



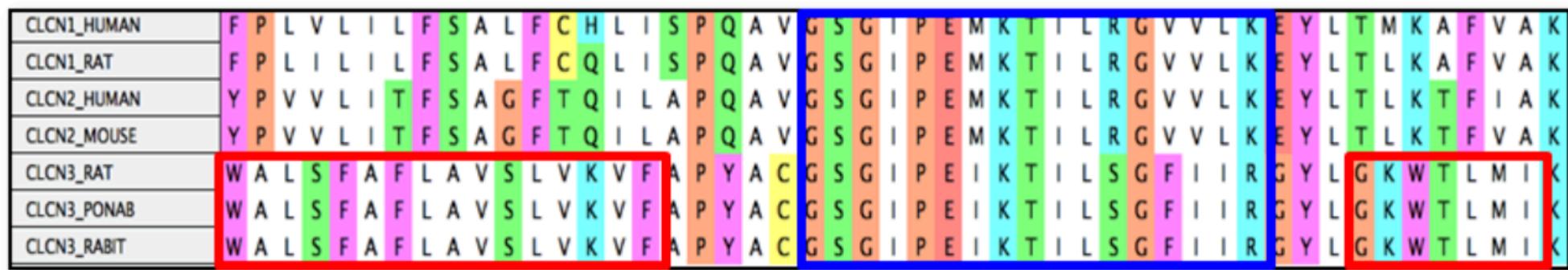
**Figure 15** Representation of the steps involved in creating a fingerprint signature.

# Protein Databases

## InterPro

### What are fingerprints

- Fingerprints are used by the PRINTS database. They are very good at modeling the often small differences between closely related proteins, as illustrated in the example in Figure 16. This means fingerprints can distinguish individual subfamilies within protein families. This allows functional characterisation of sequences at a high level of specificity (identifying individual cellular pathways in which a protein might be involved, the ligand it might bind, the exact reaction it may catalyse, and so on).



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



Amino acids uniquely conserved in chloride channel protein 3 subfamily members

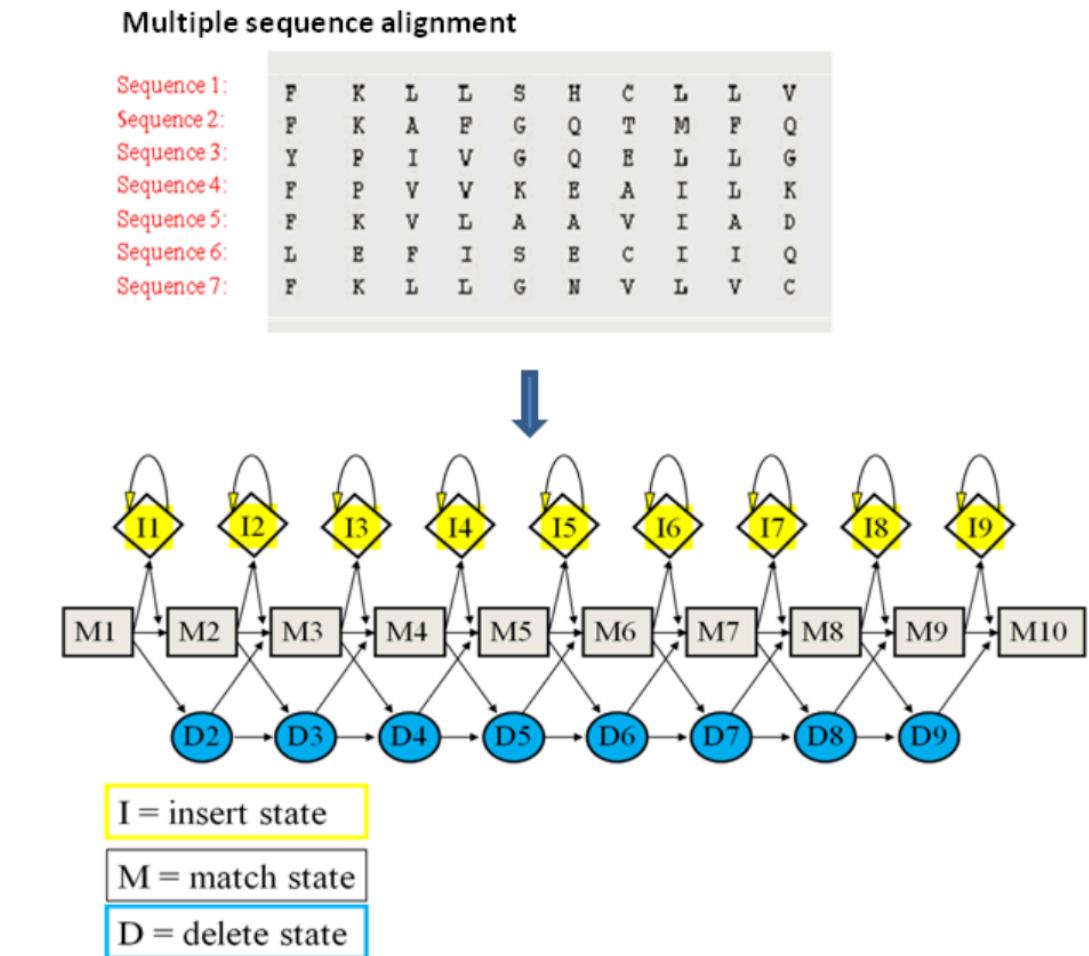
**Figure 16** Multiple sequence alignment showing amino acid conservation across chloride channel protein family members. By using multiple short conserved motifs, fingerprints are able to distinguish closely related subfamilies from each other.

# Protein Databases

## InterPro

### 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.
- HMMs have wide utility, as is clear from the numerous databases that use this method for protein classification, including Pfam, SMART, TIGRFAM, PIRSF, PANTHER, SFLD, Superfamily and Gene3D.



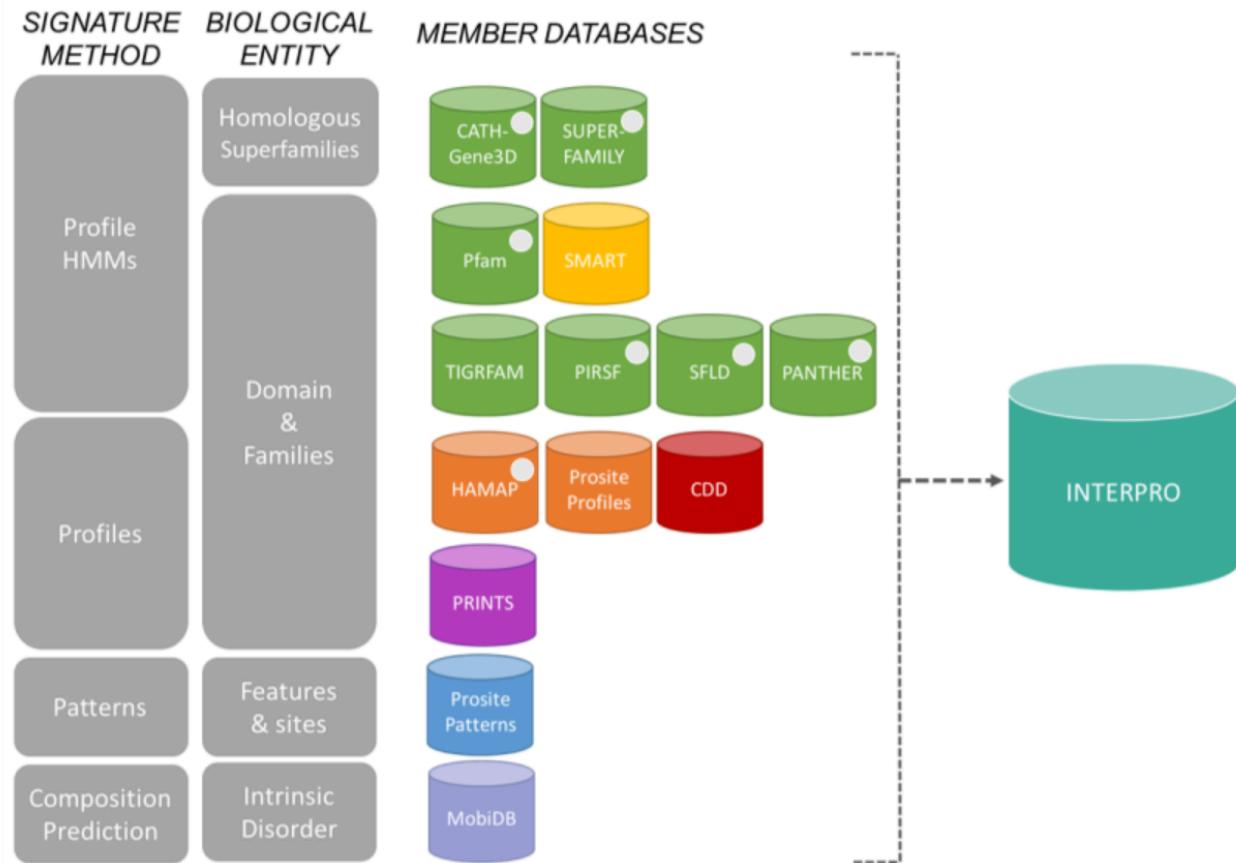
**Figure 14** Representation of a Hidden Markov model based on a multiple sequence alignment. Amino acids are given a score at each position in the sequence alignment according to the frequency with which they occur. Transition probabilities (i.e., the likelihood that one particular amino acid follows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

# Protein Databases

## InterPro

### Protein classification resources at the EBI: InterPro

- InterPro is the main resource for protein classification at the EBI.
- In InterPro, patterns, profiles, 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 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.

# Protein Databases

## InterPro

### What is an InterPro entry?

**Entry name:** Malate dehydrogenase, type 2 | IPR010945

**Entry type:** F (InterPro entry)

**Navigation menu:** Overview, Proteins 15k, Domain Architectures 34, Taxonomy 11k, Proteomes 2k, Structures 30, Genome3D 145

**Relationships to other entries:**

- L-lactate/malate dehydrogenase (IPR001557)
- Malate dehydrogenase, type 2 (IPR010945)
- Lactate dehydrogenase, protist (IPR011272)
- Malate dehydrogenase, NAD-dependent, cytosolic (IPR011274)

**Description:**

Malate dehydrogenases catalyse the interconversion of malate and oxaloacetate using dinucleotide cofactors [1]. The enzymes in this entry are found in archaea, bacteria and eukaryotes and fall into two distinct groups. The first group are cytoplasmic, NAD-dependent enzymes which participate in the citric acid cycle (1.1.1.37 [2]). The second group are found in plant chloroplasts, use NADP as cofactor, and participate in the C4 cycle (1.1.1.82 [3]). Structural studies indicate that these enzymes are homodimers with very similar overall topology, though the chloroplast enzymes also have N- and C-terminal extensions, and all contain the classical Rossman fold for NAD(P)H binding [2, 3, 4, 5]. Substrate specificity is determined by a mobile loop at the active site which uses charge balancing to discriminate between the correct substrates (malate and oxaloacetate) and other potential oxo/hydroxyacid substrates the enzyme may encounter within the cell [6].

**GO terms:**

Biological Process	Molecular Function	Cellular Component
Malate metabolic process (GO:0006108) [7]	Malate dehydrogenase activity (GO:0016615) [8]	None
Oxidation-reduction process (GO:0055114) [9]		

**References:**

- Malate dehydrogenase: a model for structure, evolution, and catalysis. Goward CR, Nicholls DJ. *Protein Sci.* 3, 1883-8, (1994). [View article](#) PMID: 7849603
- Determinants of protein thermostability observed in the 1.9-A crystal structure of malate dehydrogenase from the thermophilic bacterium *Thermus flavus*. Kelly CA, Nishiyama M, Ohishi Y, Beppu T, Birktoft JJ. *Biochemistry* 32, 3913-22, (1993). [View article](#) PMID: 8471603
- Structural basis for cold adaptation. Sequence, biochemical properties, and crystal structure of malate dehydrogenase from a psychrophile *Aquaspirillum arcticum*. Kim SY, Hwang KY, Kim SH, Sung HC, Han YS, Cho Y, J. *Biol. Chem.* 274, 11761-7, (1999). [View article](#) PMID: 10206992
- Structural basis for light activation of a chloroplast enzyme: the structure of sorghum NADP-malate dehydrogenase in its oxidized form. Johansson K, Ramaswamy S, Saarinen M, Lounsbury M, Issakidis-Bourget E, Mignac-Maslow M, Ekstrand H. *Biochemistry* 38, 4319-26, (1999). [View article](#) PMID: 10194350
- Chloroplast NADP-malate dehydrogenase: structural basis of light-dependent regulation of activity by thiol oxidation and reduction. Carr PD, Verger R, Ashton AR, Ollis DL. *Structure* 7, 461-75, (1999). [View article](#) PMID: 10196131
- Structural basis of substrate specificity in malate dehydrogenases: crystal structure of a ternary complex of porcine cytoplasmic malate dehydrogenase, alpha-ketomalonate and tetrahydNAD. Chapman AD, Cortes A, Daftow TR, Clarke AR, Brady RL. *J. Mol. Biol.* 285, 703-12, (1999). [View article](#) PMID: 10075524

**Further reading:**

Crystal structure of NAD-dependent malate dehydrogenase complexed with NADP(H). Tomita T, Fushinobu S, Kuzuyama T, Nishiyama M. *Biochem. Biophys. Res. Commun.* 334, 613-8, (2005). [View article](#) PMID: 16009341

**Cross References:**

**EC:** 1.1.1.37

**SCOP:** d.162.1.1, c.2.1.5

**CATH:** 3.40.50.720, 3.90.110.10

**External references:**

# Protein Databases

## InterPro

InterPro entry types



Homologous  
Superfamily



Family



Domain



Repeat



Site

# Protein Databases

## InterPro

### InterPro entry types



Homologous  
Superfamily



Family



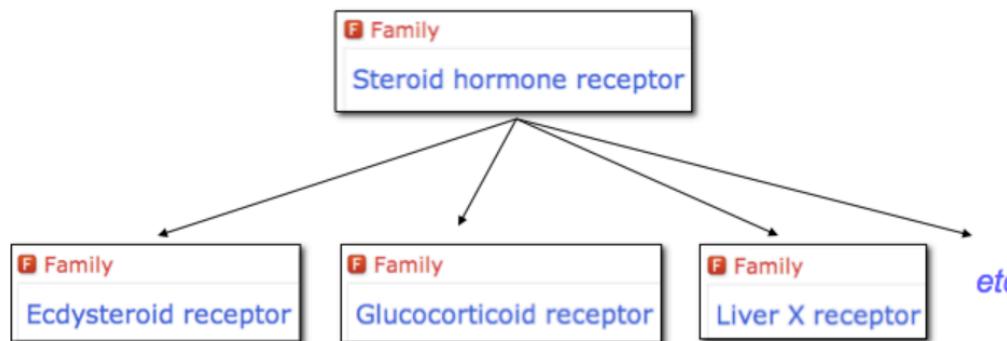
Domain



Repeat



Site



- An InterPro protein family 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 subdivided into smaller, more closely related groups. For example, steroid hormone receptors constitute a family of nuclear receptors responsible for signal transduction mediated by steroid hormones, and can be subclassified into different groups, including the liver X receptor subfamily (see Figure). This subfamily consists of nuclear receptors that regulate the metabolism of several important lipids, including oxysterols.

# Protein Databases

## InterPro

InterPro entry types



Homologous  
Superfamily

Family

Domain

Repeat

Site



- 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, the pleckstrin homology (PH) domain is a small modular domain that occurs in a large variety of proteins and is involved in phospholipid binding. One group of proteins containing a PH domain are the beta-adrenergic receptor kinases (see Figure). Four domains have been identified in these proteins: an RSG (regulator of G protein signalling) domain, a protein kinase (PK) domain, an AGCK domain, involved in regulation by phosphorylation, and a C-terminal PH domain.

# Protein Databases

## InterPro

InterPro entry types



Homologous  
Superfamily



Family



Domain



Repeat



Site

Sites and repeats

- Sites are groups of amino acids that confer certain characteristics upon a protein, and may be important for its overall function. Sites are usually quite small (often only a few amino acids long). The types of site covered by InterPro are:
  - active sites, which contain amino acids involved in catalytic activity
  - binding sites, containing amino acids that are directly involved in binding molecules or ions
  - post-translational modification (PTM) sites, which contain residues known to be chemically modified (phosphorylated, palmitoylated, acetylated, etc) after the process of protein translation
  - conserved sites, which are found in specific types of proteins, but whose function is unknown.
- Repeats are typically short amino acid sequences that are repeated within a protein, and may confer binding or structural properties upon it.
  - For example, pentapeptide repeats are sequence motifs of five amino acids found in multiple tandem copies. They were first identified in cyanobacterial proteins, where they can be found in many copies. Their function is currently unknown.