InterPro: Functional and structural analysis of protein sequences

Alex Mitchell [1]

- Proteins
- Beginner
- 1 hour

InterPro is a classification database that provides predictive information about protein sequences. This course will show you what kind of information is available in InterPro, how to access it, and how to search the database.

Learning objectives:

- Be able to search the InterPro database in a number of different ways
- Be able to interpret the information in an InterPro entry page
- Be able to review the graphical representation of the signature matches on an InterPro protein page

What is InterPro?

InterPro [2] is a bioinformatics resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites (Figure 1). 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.

Figure 1. The InterPro homepage can be found here [3] and is easily accessible by clicking on the InterPro link in the Proteins panel on the EBI services [4] page.
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 [2] consortium (Figure 2). 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.

To learn more about the different methods that can be used to classify proteins and the different types of models used by InterPro's member databases, see our Introduction to Protein Classification [5] tutorial.

Figure 2. InterPro member databases grouped by signature construction method and focus of interest.

Why do we need InterPro?

Protein signature [6] 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 [2] because it:

- Reduces redundancy and simplifies protein sequence analysis by integrating signatures from different member databases that represent the same protein family [7], domain [8] or site.
• Unites the member databases, capitalising on their individual strengths to produce a powerful classification tool.

• Provides a single convenient searchable location, allowing simultaneous querying of all member databases.

• Adds information (including descriptive abstracts and Gene Ontology [9] terms) to the signatures, which may be used to annotate the proteins they match.

What is InterPro used for?

InterPro is used to provide annotation for UniProtKB

• The sequences stored in the universal protein database UniProtKB are analysed regularly using InterPro. In this way, InterPro helps to provide annotation for uncharacterised sequences in the UniProtKB database (Figure 3).

InterPro can be used to analyse any protein sequence

• Users can also choose to analyse their own sequences for predictions about their function and/or the presence of certain domains and sequence features.
Figure 3. Flow of data between InterPro and UniProt databases.

How can InterPro help with your research?

When to use InterPro

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

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

You can also use InterPro for a variety of other purposes, such as predicting GO terms for a set of sequences, or identifying all sequences in the UniProtKB database that are predicted to belong to a given family or to possess a particular domain.

When not to use InterPro

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 (intron / exon predictions, identification of promoter regions, etc)

What is an InterPro entry?

At its heart, each individual InterPro entry consists of one or more member database signatures that are characteristic of the same protein family, domain or sequence feature. The entry is given a name and a unique InterPro identifier. It is also annotated with additional information, which can be found in different sections on the entry page.

For example, Figure 4 below shows the InterPro entry for the type 2 malate dehydrogenase protein family. The entry has a name (malate dehydrogenase, type 2) and identifier (IPR010945). The contributing signatures are shown on the right-hand side, with links to the signatures on the individual member database web sites. A descriptive abstract explains what these proteins are and what their function is. It also provides links to publications in PubMed for more detailed information. A set of GO terms is also provided, which describe the characteristics of the proteins matched by the entry.
**InterPro entry types**

**InterPro entry types**

InterPro [2] entries are classified into one of five categories, depending on the biological entity they represent: homologous [14] superfamily [15], protein family [7], domain [8], repeat or site.

The entry type is indicated by a specific icon (Figure 5), which can be found on the top left hand side of an InterPro entry page [16].

![Icons denoting the different type of entries](image)

**Figure 5.** Icons denoting the different type of entries (homologous superfamily, family, domain, repeat or site) that can be found [16] in the InterPro database.

**InterPro entry types: Family and Homologous Superfamily**

An InterPro [2] protein family [7] 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 (Figure 6). This subfamily consists of nuclear receptors that regulate the metabolism of several important lipids, including oxysterols.

**Figure 6.** Example of a protein family hierarchy. The steroid hormone receptor family can be subdivided into a number of smaller, closely related subfamilies.

A Homologous Superfamily is a large diverse family, usually with a shared tertiary structure. They are exclusively composed of CATH-Gene3D and/or SUPERFAMILY methods, both of which utilise a collection of underlying profile hidden Markov models (HMMs) to represent diverse structural families, rather than one single model. Unlike Family entries, Homologous Superfamily entries do not exist in hierarchies.

**InterPro entry types: Domain**

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 (Figure 7). 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.
Figure 7. Graphical representation of the domain architecture of beta-adrenergic receptor kinases.

InterPro entry types: Sites and Repeats

Sites
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 [2] 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 [22]
- conserved sites, which are found in specific types of proteins, but whose function is unknown.

For example, the stretch of residues involved in the catalytic function of the S1B subfamily of serine peptidases is specific to this type of peptidases and constitutes the active site [23] of these proteins.

Repeats
Repeats are typically short amino acid [10] 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.

Relationships between InterPro entries
InterPro [2] entries often share relationships with other entries in the database. For example, an entry might represent a functionally-specific subfamily of a larger family, or a subclass of a broad class of domains. These relationships are stored in InterPro as hierarchies, which are displayed as trees in the relationship section of an entry page (Figure 8).

Both family and domain [8] entries are able to form hierarchical relationships, although family and domain hierarchies are kept separate in the database and do not overlap (for example, a subclass of a particular domain can not be a subtype of a protein family [7]). Repeats and sites are not arranged into hierarchical relationships.

![Figure 8. Examples of hierarchical relationships between InterPro entries.](image)

**Side menu links**

More information is available from the side menu of an InterPro [2] entry page, including links to the proteins matched by that entry and the species in which they are found.

The different types of information found in this menu are shown in Figure 9.

A more detailed description of some of the information available from the side menu is given in the following sections, where we explore searching the database.
Figure 9. Overview of the information available from the side menu of an InterPro entry.

How to search InterPro?

You can search InterPro [2] (Figure 10) using either an amino acid [10] sequence or via a text search (e.g., with a word or short phrase, a UniProt [24] or InterPro identifier [12]). The following sections will take you through both of these options.
Sequence searching

To search InterPro [2] with a protein sequence, copy and paste an amino acid [10] sequence into the large box on the home page and click on the search button located below the box (Figure 11). InterPro will then attempt to assign the sequence to a protein family [7] and identify any domains, repeats and sites.

Once the analysis is complete, a results page will be returned showing the InterPro matches to your query sequence in graphical form (Figure 12). Next we will look at interpreting these results to understand what they mean, using a guided example.
Guided example: searching InterPro with an amino acid sequence

Getting started
In this guided example, we will look at searching InterPro [2] with an amino acid [10] sequence.

First, we need to locate the sequence search box on the InterPro homepage.

Navigate to the InterPro [3] homepage and find the sequence search box.

Now we need a sequence to search against the database.
Copy and paste the following sequence into the sequence search box, and press ‘Search’

>MY_SEQ
MAAGGSGAESAPPTPSMSSLPLAALNVRVRHRLSLFLNVRTQVAADWTGLAEEM
NFEYLEIRRLETHPDPTSSLDDLQGGRPGASVGRLELLAKLGRDDVLVELGPS
IEEDCRKYLIKQQEAAEKPQLQVDSVDSSIPWMSGITIRDDPLGQMPFHDFAFI
CYCPSSIDQFQVQEMIRQLQTONYRLKLCVSDRDVLPGTCVWSIASIELIEKCRRM
VVVYSSDDYLQSKEDQFQTKFALSISPGAHQKRLIPVKSMKKEFPSILRFITV
CDYTNPCTKSWFWTRLARALSLP

We will now go through the search results in the slideshow below.

Sequence search results: overview page

The search result page should look something like the one shown in Figure 13 below. This is an interactive page that contains a large amount of information.

Section A shows the family to which InterPro [2] predicts the sequence belongs. This is displayed as a hierarchy, where appropriate. Clicking the link will take you to the InterPro entry page for the family, where detailed information about its function may be found.

Section B summarises the domain [8] and repeats that InterPro predicts the protein to contain. The sequence is represented as a grey bar with its length in amino acids displayed along the bottom. Domains and repeats are indicated as coloured bars. Mousing over the bars reveals the type of domain or repeat that they represent, along with their position on the sequence and a link to the relevant InterPro entry page.

Section C holds detailed signature match information, showing the raw match position of all the different signatures in InterPro to the sequence, including (where available) signatures representing families, domains, repeats and sites, and unintegrated signatures that are not associated with InterPro entries. The information displayed in this section can be controlled using the interactive menu on the left hand side of the screen (Section D).

Section E shows the Gene Ontology [9] (GO) terms predicted for the protein. These terms are assigned based on the matches to the InterPro entries shown above.
Figure 13. InterPro sequence search results overview page.

Sequence search results: domain information (1)

Looking at the results overview page, we will first examine the domains that InterPro predicts the query protein to contain.

On the results page, place your mouse over the left hand domain in the Domains and Repeats section (Figure 14).

Figure 14. Mousing over domains reveals their name, position and the InterPro entries with which they are associated.
Mousing over the left hand domain, we see its name (death domain) and InterPro accession number (IPR000488), its position on the sequence (amino acid residues 5-125), and that it belongs to a wider class of death-like domains.

Click on the link to visit the entry page for the death domain (IPR000488)

Examining the entry page for the death domain (Figure 15), we can see that our domain is a protein-interaction module that is involved in the association of receptors so that they can signal downstream events, including programmed cell death (apoptosis). The other proteins in UniProtKB that InterPro predicts to contain such a domain can be accessed by clicking the 'Proteins matched' link on the left hand side menu. The species in which these proteins are found, the pathways in which they may be involved, and any solved protein structures can also be accessed from the left hand menu, via the 'Species', 'Pathways & interactions' and 'Structures' links, respectively.

![Figure 15. InterPro entry page for the death domain.](image)

**Sequence search results: domain information (2)**

Navigating back to the search results page, we can examine the second domain that InterPro predicts the query protein to contain.

On the results page, place your mouse over the right hand domain in the Domains and Repeats section
Figure 16. Mousing over the second domain reveals its name, amino acid residue position and a link to the relevant InterPro page.

Mousing over the domain, we can see that it is a Toll/interleukin-1 receptor homology (TIR) domain. Following the link to the InterPro entry page, we can find out more about this domain, including its involvement in signalling (Figure 17). As with the death domain entry page, more information on the proteins predicted to contain a TIR domain (such as the species in which they are found and the pathways in which they are involved) can be found using the left hand side menu.

Figure 17. InterPro entry page for the TIR domain.

Sequence search results: family information

Returning to the results overview page, next we look at the family membership section (Figure 18). We can see that InterPro [2] predicts that the protein belongs to the myeloid differentiation primary response protein MyD88 family.
Figure 18. Predicted protein family membership for the query sequence.

Click on the link to visit the entry page for myeloid differentiation primary response protein MyD88.

Clicking on the link will take us to the InterPro entry page for this family (Figure 19). This page explains what MyD88 is, what its functions are, and the GO terms that can be applied to members of the family. Clicking on the left hand menu allows us to examine the different UniProt [24] proteins that are predicted to belong to the family (Proteins matched), their different domain [8] architectures (Domain organisations) and the species in which they are found (Species). Clicking on any of these menu items opens a new page with the appropriate information and links that allow you to download specific datasets (e.g., all the proteins in a family, or all the proteins with a particular domain architecture, or all the protein family members from a particular kingdom, class or species).

Figure 19. InterPro entry page for the myeloid differentiation primary response protein MyD88 family of proteins.

Sequence search results: exploring other proteins in the family

It is possible to find all of the UniProtKB [11] proteins that match a protein family [7], using the ‘Proteins matched’ link on the left hand side of the family page (Figure 20).
Figure 20. The protein families side menu page, where links to more information on proteins belonging the same family can be found.

Click the 'Proteins matched' link to see other proteins in UniProtKB with the same family membership.

Following the 'Proteins matched' link takes us to a page like the one shown in Figure 21. This is a paginated list of UniProtKB proteins, showing their name, accession number, species and domain architectures (if known). Those sequences with a gold star are from Swiss-Prot, the manually annotated section of UniProtKB, whilst those with a silver star are from the automatically annotated section TrEMBL. Proteins for which structural information is available are shown with a '3D' icon, next to their accession number. A FASTA file containing the sequences of all of the proteins can be downloaded by clicking the 'Export FASTA' button at the top right of the page.
Figure 21. The proteins matched page, displaying UniProtKB proteins belonging to the same protein family as the query sequence.

Searching InterPro with a batch of amino acid sequences

Often, people want to search InterPro with a batch of amino acid sequences - for example, to analyse whole genomes or proteomes. There are two ways in which it is possible to do this. The first way involves downloading and installing the InterProScan [28] analysis tool. This can be obtained from here [29].

The software comes with full installation instructions. It also has a lot of configurable features, including the ability to search nucleic acid sequences against the database.

The second way of searching InterPro with a batch of sequences involves the use of the EBI's web services [30].
These allow up to 30 sequences to be analysed per request. Instructions on how to use the SOAP [31]-based web service are available here [32].

Instructions on using the REST [33]-based web service can be found here [34].

**Text search**

InterPro can also be searched using text (Figure 22), this can take the form of:

- an InterPro identifier, which will open the corresponding entry page
- a word or a short phrase, in which case a list of all InterPro entries with the word or phrase in their title or annotation will be returned
- a Gene Ontology term identifier, which will return a list of all InterPro entries with that GO term
- a structural identifier, which returns a list of InterPro entries that match the protein to which the structural identifier relates
- a UniProtKB protein identifier, which returns a results page similar to a sequence search, but enhanced with UniProtKB information, as is explained in the next section.

![Text search example](https://www.ebi.ac.uk/training/online)

**Figure 22.** To search InterPro using a word, phrase or identifier type it in the text search box as indicated.

**Searching with a protein identifier**

Searching InterPro with a UniProtKB accession (e.g., O88551) or identifier (such as CLD1_MOUSE) returns a page similar to the sequence search results page, but containing extra information, drawn from UniProtKB. Family membership, domain, site and repeat information, and GO terms are displayed as normal, but at the top of the page an additional section is displayed (Figure 23). This contains information on the name of the protein in UniProtKB, its accession number and identifier, the species in which it is found, its length in amino acids, and whether it is a complete sequence or a fragment. Following the accession number link will take you to the UniProtKB entry for this protein. Additional information can also be found in the left hand menu: 'Similar proteins' lists all UniProtKB proteins with the same domain architecture, and 'Structures' lists any structural information in UniProt for the protein accession searched (in each case, these links will only be active where such information is present in UniProtKB).
Figure 23. Protein page showing InterPro predictions for UniProtKB protein O88551.

Summary

- InterPro is a resource for classifying proteins into families and predicting domains, repeats and functional sites.

- InterPro integrates protein signatures from 14 member databases that are part of the InterPro consortium: CATH-Gene3D, CDD, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE Patterns, PROSITE Profiles, SFLD, SMART, SUPERFAMILY and TIGRFAMs. Protein signatures are predictive models based on similarities among proteins that have the same structure or function.

- You can search InterPro with an amino acid sequence using the web site.

- The InterPro web site can also be searched with a word or phrase, an entry ID, UniProtKB accessions and identifiers, GO terms and structural identifiers.

- Batches of sequences can be searched against InterPro by downloading the InterProScan tool, or using the EBI's web services.

- The InterPro protein view provides a graphical representation of the signatures that match a particular protein, with information about protein family membership, sequence features, structural features, and structural predictions for that protein.

Quiz: Interpro Quiz

Questions: 5
Attempts allowed: Unlimited
Available: Always
Pass rate: 75 %
Backwards navigation: Allowed
Exercises

This exercise allows you to apply the knowledge you have gained on this course by providing an example of how InterPro [2] can be used and asking you to solve a given task. Start by clicking on the exercise title provided.

If you need help to complete this section you can look in the 'Need some help?' and 'Want to know how we did it?' sections.

Analyse a protein sequence

Scenario

You have isolated a protein from squirrel. You want to know more about this protein, perhaps the presence of certain domains or features can give you a clue about its role.

Exercise

Find information about the following protein using InterPro.

>squirrel_seq

MALPARLVPLCCLALLALPAQSCGPRGPVGRRRYVRKQLVPLYKQ
FVPSSPERTLGASGPAEGRVARSGERFRDLVPYNPDIIIFKDEENSG
ADRLMTERCKERVNAIAVMNMWPGVRLRTEGWDGEDGHQAQDSLH
YEGRALDITTSRDRNKBLLARLVEAGFDVYYESRNHHVHSVKA
GTVGGGCNRETEAAQLWGDARGLRELHRAWLAAADAGAVTVPTPVL
FLRDRLQRARSFVATEPVRPRKLLTPWHLVFAARGPAPAPGTFAP
VFARRLAGDSVLPAGGALRPARVARVAREEAAGVFAPLTAHGTLL
VNDVLASCYAVLESQWAHRAFAPLRLHALLPGAVQPPTGMHW
YSRFLYRLAEELLG
Need some help?

1. Copy and paste the FASTA sequence into the ‘analyse your protein sequence’ box on the InterPro home page and press ‘Search’.
2. Once you are in the protein overview page, explore the information on the graphical results view. What family does this protein belong to? What domains does it have? What processes is it involved in?
3. Investigate further the entries that match this protein by clicking in the InterPro identifiers and reading the descriptions for those entries.

Want to know how we did it?

When you search InterPro with the FASTA sequence provided, you will get a graphical representation of results, similar to that shown in Figure 24.
Figure 24. InterPro sequence search results overview page.

From the results in InterPro you can infer:

- That the protein is predicted to contain an N-terminal signalling domain, which is part of a wider class of domains that also includes the zinc-binding domains of DD-peptidase.
- That the protein is also predicted to contain a C-terminal Hint domain, which is involved in autoprocessing of the protein. The Hint domain itself can be broken down into two smaller subdomains.
- That the protein queried is predicted to be a member of the hedgehog protein family, and therefore involved in a range of biological processes, including development and cell-cell signalling. How? From the GO terms and description from entry IPR001657 (see Figure 25 below).
Figure 25. The InterPro entry page for IPR001657 (Hedgehog protein).

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.

Get help and support on InterPro

Documentation

You can find documentation about InterPro [2] (the user manual, release notes, etc) linked from our web page [3] or you can download the information from our FTP server [35].

Contact

For all support enquiries, please contact the InterPro interhelp [at] ebi.ac.uk (help desk)
Learn more

Recommended courses

Online course: Introduction to Protein Classification at the EBI [5]

The EBI offers hands-on courses covering a range of subjects, including training on InterPro [2]. An up-to-date list of training courses is available here [36].

Contributors

Alex Mitchell [1]
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.

Amaia Sangrador [37]
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.

Mindi Sehra [38]
EMBL-EBI
EMBL-EBI ALUMNI Scientific Training Officer - eLearning Content Developer

Mindi Sehra was the Scientific Training Officer (eLearning) for the Outreach and Training Team at EMBL-EBI. Mindi was responsible for expanding and consolidating the EBI’s range of online training materials and monitoring and maintaining the portal, which includes investigating ways to exploit electronic technologies. Mindi completed a Genetics Degree at Sheffield University before moving into genome analysis at the Wellcome Trust Sanger Institute. She completed a MSc in Medical Genetics and Immunology at Brunel University, her thesis on the Swine Leukocyte Antigen secured her a position in the Human and Vertebrate Annotation [39] and Analysis group as a computer biologist. She then joined the UniProt [24] team at the EBI as a protein curator [40] working on automatic and manual annotation [41].

References


Source URL: https://www.ebi.ac.uk/training/online/course/interpro-functional-and-structural-analysis-protein-sequence