EMBL-EBI Tutorial

# Predicting protein structure and function with InterPro


Phil Jones, Alex Mitchell, Amaia Sangrador (V4.0, Jun 2011)

# Contents

# Course Information

| | |
|---|---|
| Course description | This tutorial provides an introduction to InterPro, its web interface and content. You will learn how to search InterPro to obtain information about protein function, classification and sequence and structural features. |
| Course level | Suitable for graduate-level scientists and above who have not used InterPro before. Regular users may benefit from going through the tutorial to understand better the new InterPro website |
| Pre-requisites | Basic knowledge of biology and basic computer skills |
| Subject area | Proteins and Proteomes, Genes and Genomes, Sequence Analysis |
| Target audience | Scientists interested in Sequence Analysis |
| Resources required | Internet access (a current browser such as the latest Firefox or Internet Explorer) |
| Approximate time needed | 2 hours |

# Course learning objectives

- Querying InterPro via the web interface
- Using InterPro entries to classify and annotate protein sequences
- Understanding signature relationship hierarchies
- Relating signatures to protein structure
- “Power search” – The InterPro BioMart.

# An introduction to InterPro

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. It does this by combining protein signatures from a number of independent databases (referred to as member databases) into a single searchable resource. InterPro integrates the signatures, providing a name and abstract and, whenever possible, GO mapping, structural links and external database links. Together, the protein signatures combined within InterPro cover ~ 80% of proteins in the UniProt database.

### What are protein signatures?

Protein signatures are obtained by modelling the conservation of amino acids at specific positions within a group of related proteins (i.e., a protein family), or within the domains/sites shared by a group of proteins. The different member databases use different computational methods to produce protein signatures. These include:

- Regular expressions (PROSITE patterns)
- Fingerprints that use Position Specific Sequence Matrices (PRINTS)
- Sequence clustering via PSI-BLAST (PRODOM)
- Sequence matrices (PROSITE profiles, HAMAP)
- Hidden Markov Models (Pfam, TIGRFAMs, PIRSF, Superfamily, Gene3D, PANTHER, SMART)

These protein signatures are run against the UniProt database of protein sequences, and all significant matches are reported in InterPro, allowing users to check the protein matches corresponding to a signature. This information is also used for automatic annotation in TrEMBL (non-reviewed protein sequences of UniProtKB; see http://www.uniprot.org/).

### InterPro entry types

The signatures provided by the member databases are integrated by InterPro. Each InterPro entry is assigned a type, namely family F, domain D, repeat R, or site S (conserved site, active site, binding site or post translational modification site).

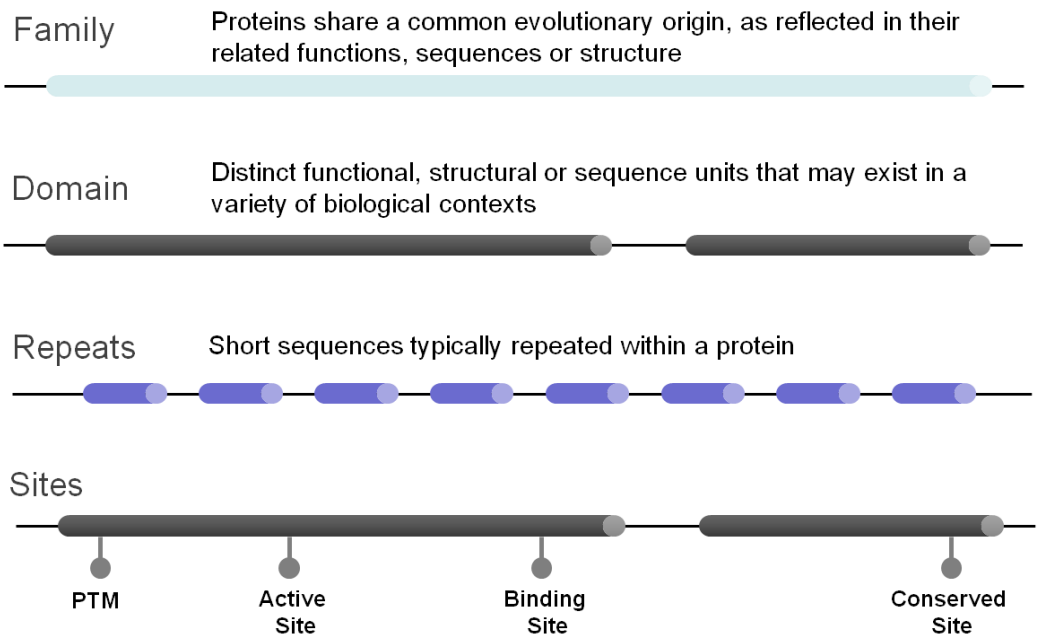


**Figure 1** InterPro entry types and definitions. More information can be found be found on the InterPro web site http://wwwdev.ebi.ac.uk/interpro/about.html#3

### InterPro hierarchies

InterPro entries are organised into hierarchical relationships, where possible. Entries in a hierarchy form "parent-child relationships". Entries at the top of these hierarchies (the "parents") describe more general functional families or domains, while entries at the bottom (the "children") are describe specific functional subfamilies or structural/functional subclasses of domains.


**Figure 2** Two examples of InterPro hierarchies. Families and domains are placed into separate hierarchies

### The InterPro home page

You can find the InterPro site at http://www.ebi.ac.uk/interpro/ or by clicking on the InterPro link on the EBI home page. However, for this tutorial we will use the newly developed beta version of the InterPro web site (see Fig. 3 below), which offers increased functionality and will shortly be replacing the old website. The beta site can be accessed by pointing your web browser at http://wwwdev.ebi.ac.uk/interpro/. The home page provides:

- search tools.
- documentation (user manual, release notes, etc).
- link to tools, such as BioMart.

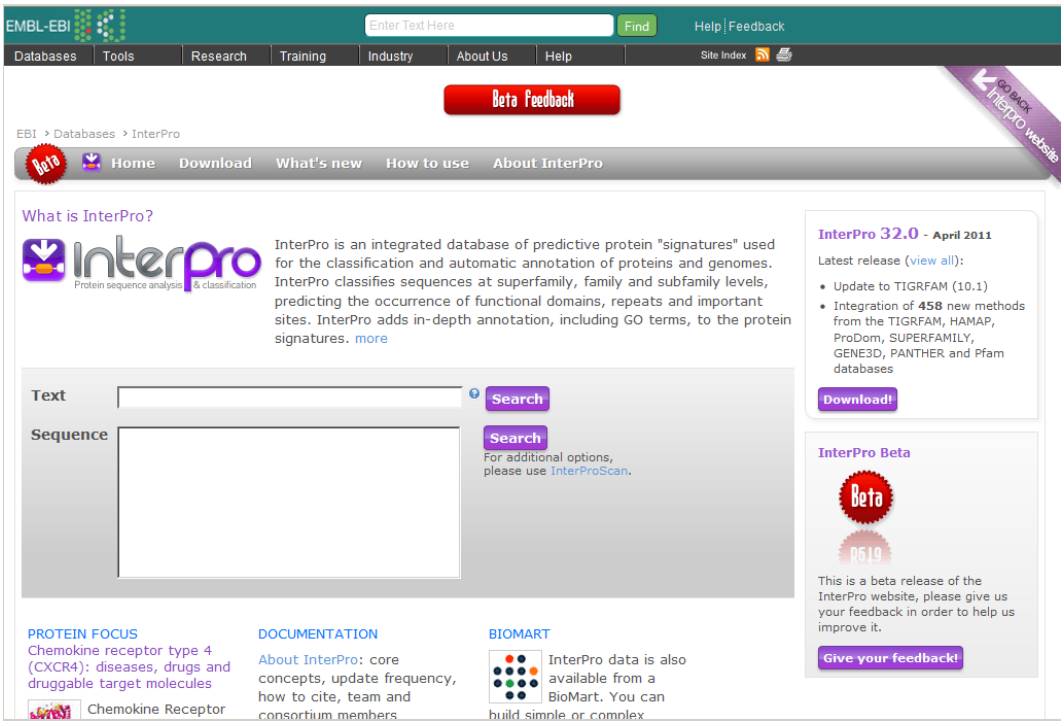


**Figure 3** InterPro beta web site home page..

### Searching InterPro

InterPro can be searched a number of different ways:

- *Text search*
    - Via the text box in the main page or at the search bar at the top right of all other InterPro web pages
    - Search using: UniProtKB accessions; InterPro entry IDs; GO terms; plain text
- *InterProScan*
    - Search using protein sequence

- Use the sequence search box on the InterPro home page or follow the link to InterProScan for more search options

- *BioMart*
  - More flexible and powerful querying; retrieve results in HTML, plain text or Microsoft Excel spreadsheet
  - To access the InterPro BioMart, follow the link on the InterPro home page

# I. Searching InterPro using a text search

## Learning objectives

In this section, you will learn how to:

- query InterPro using the text search option.
- interpret the information in the InterPro protein view.
- interpret the information of in the InterPro entry page.

The text search bar can be found in the text box in main page and at the top right of all InterPro web pages. It can be used to search with plain text, a UniProtKB protein identifier, an InterPro identifier, a Gene Ontology (GO) term or a protein structure code.

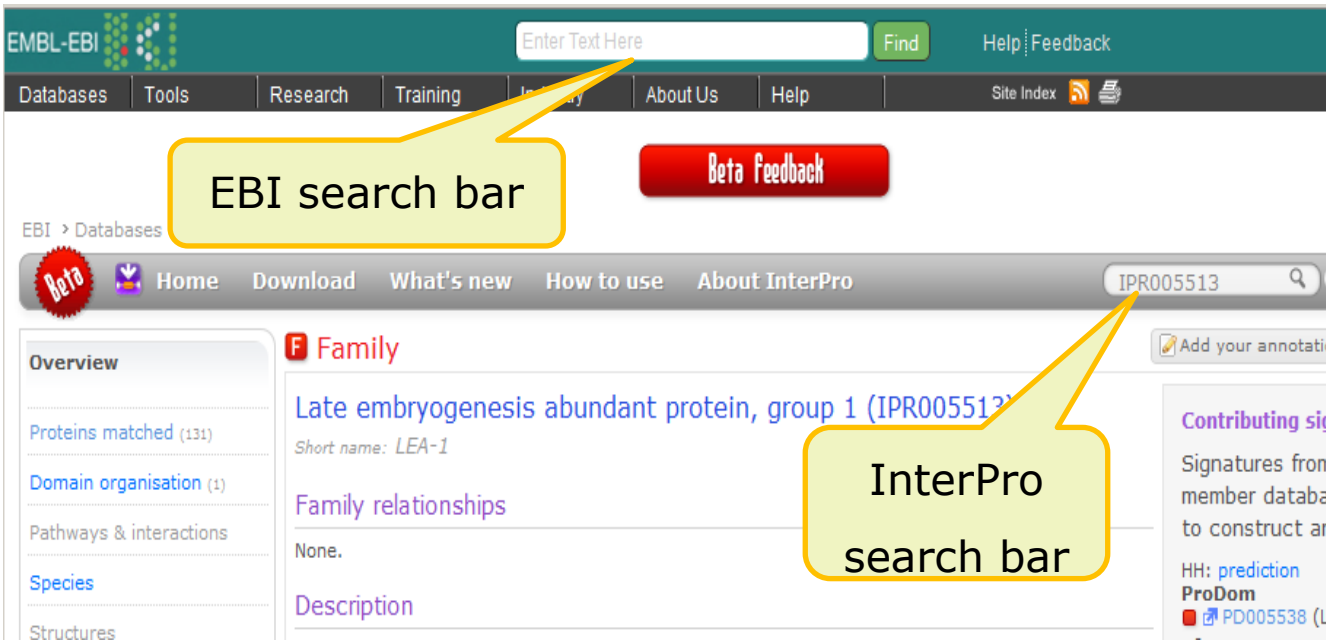


**Figure 4** The InterPro search bar can be found at the top right of InterPro web pages.

To examine the precalculated analysis results that InterPro has stored for a protein in UniProtKB, perform a text search using its UniProtKB accession number or identifier (e.g.O15075). This will bring you to the protein page view, which provides both the signature hits and structural information for the protein. If you have an accession number from GenBank, Xref, EMBL or Ensembl, you can convert it to a UniProtKB accession number using the EBI's PICR service (http://www.ebi.ac.uk/Tools/picr/).

Searching with an InterPro identifier (e.g.IPR020405 or 20405) will provide the entry page with all the information corresponding to that entry. You can also use a member database signature accession, which will return the InterPro entry that signature is integrated into.

A simple text search query (with a word, GO term, etc) will give you all the InterPro entries associated with that term.

## What information can be found in the InterPro entry page?

The InterPro entry page consists of the following features (see Fig.5):

A. Entry type and name
B. Contributing signatures
C. Entry relationships, representing existing parent and children entries.
D. Description of the InterPro entry with links to references
E. GO (Gene Ontology) terms associated to that entry. GO terms are divided in three categories: biological process, molecular function and cellular component.

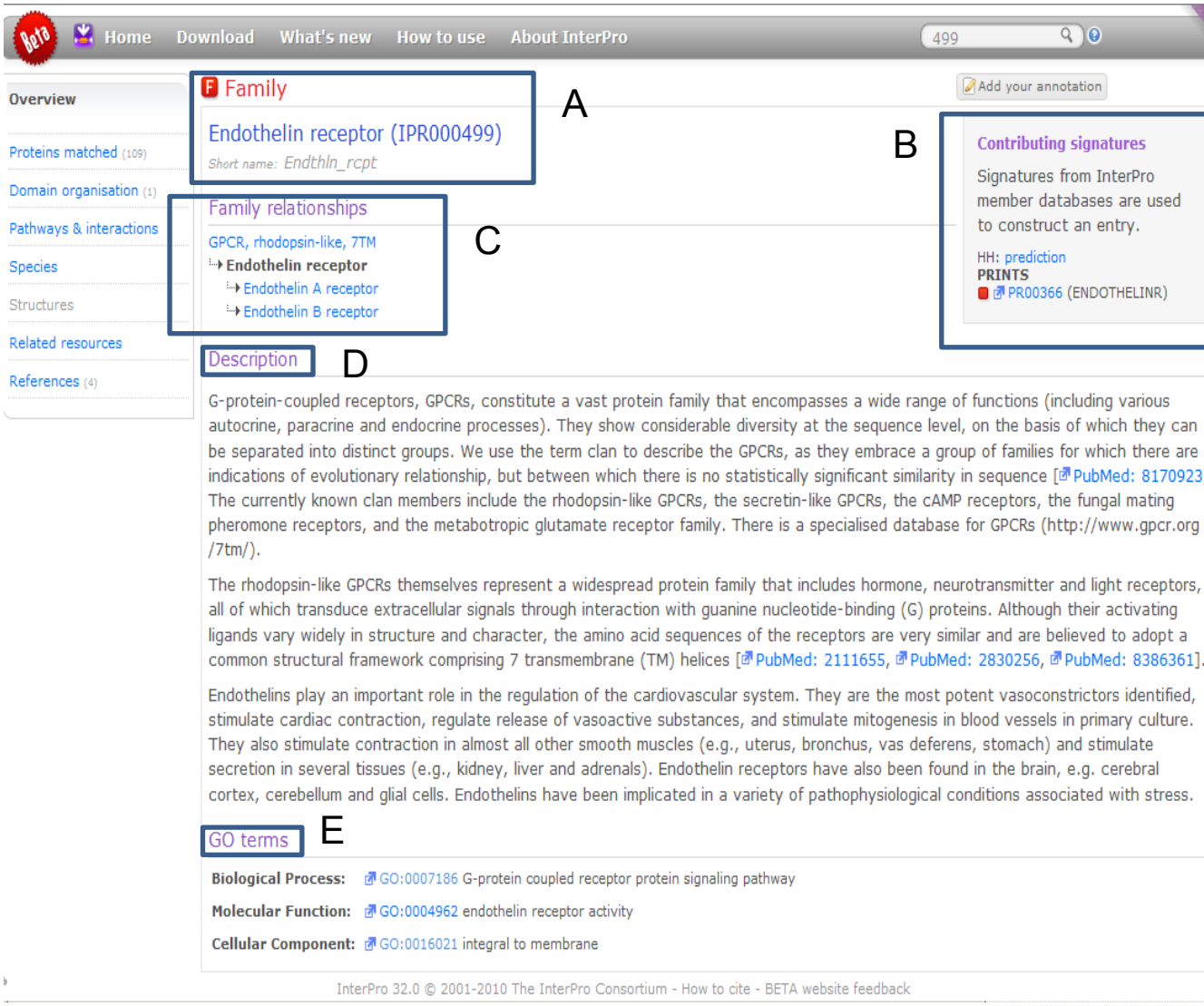


**Figure 5** The InterPro entry page

Following the links in the side menu (on the left hand side of the page) information can be found about proteins matched by that entry, their domain organisation, pathways & interactions in which they are involved, and their taxonomic coverage (i.e., the species in which the proteins are found).

## How do I interpret an InterPro protein view?

In InterPro, the Protein View consists of the following features (see Fig.6):

**Top Section**

- First of all we present the basic information about the protein: UniProt name, short name and accession (with a link to UniProtKB) and taxonomic information.

**Protein family membership**

- The protein families that a given protein belongs to are presented in a hierarchy, where appropriate. Clicking on the links takes you to the entry pages for each level of the hierarchy matched.

**Sequence features**

- The length of the protein is indicated by the white vertical bars, which are marked (in this case) every 20 amino acids.
- Each solid coloured bar represents a signature that matches the protein. The InterPro entry the signature belongs to is indicated and linked in the left hand column. Note: If you want to see the graphical representation of the matches for all of the individual member database signatures you can click on “Detailed results” in the left hand side menu.
- Domain signatures matches are presented, followed by sites. Unintegrated signatures (i.e. those that have not yet been curated) are presented at the bottom in gray. The type of each entry is represented by the icon on the left hand side beside the InterPro accession number.
- The bar colours group the entry matches – bars of the same colour are matches to entries in the same hierarchy.
- If you hover your mouse over the coloured bar, a pop-up will display the InterPro entry accession and name, and the region of the protein that the signature matches. Clicking on the linked accession will take you to the InterPro entry page (or the member database’s in the case of unintegrated signatures).

**Structural Features**

- Representative matches from PDB, SCOP and CATH are given when available. PDB contains information about experimentally-determined structures and provides structures that can cover part or the whole protein, while CATH and SCOP break proteins into structural domains.

**Structural Predictions**

- Two matches may be presented here, one to ModBase and the other to Swiss-Model. These homology databases predict protein structure based on the closest homologue.

**GO term prediction**

- GO (Gene Ontology) terms associated with the protein can be found at the bottom of the page. GO terms are divided in three categories: biological process, molecular function and cellular component. More information about GO can be found at http://geneontology.org/.

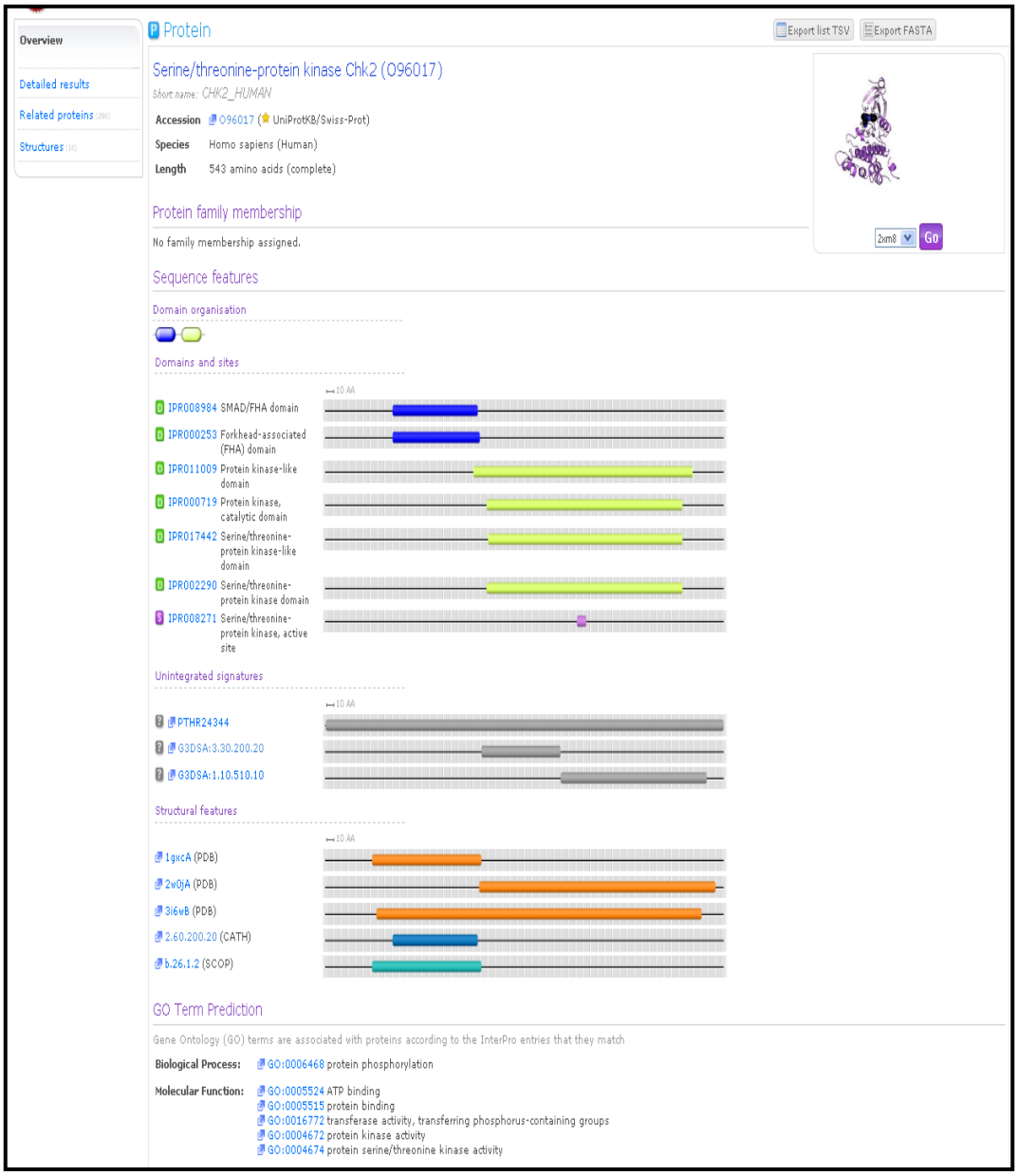


**Figure 6** The InterPro protein view

## Summary

- The InterPro text search can be queried with UniProtKB accessions, InterPro entry IDs, GO terms, structural identifiers and plain text
- To return precalculated InterPro analysis results for a protein in UniProtKB, perform a text search using its UniProtKB accession number or identifier
- InterPro entry pages provide information related to that entry (including the contributing signatures) and links to the proteins matched by that entry

- The InterPro protein view provides information about protein family membership, sequence features, structural features, and structural predictions for a particular protein.

## Exercises

### Exercise 1 – Searching InterPro using a UniProt identifier

- Open the InterPro beta website homepage (http://wwwdev.ebi.ac.uk/interpro/) in a web browser.
- Using the "Text" Search box mid-way down the page, type in the UniProtKB accession 'O15075' (*without* the quotes. That's a letter O at the start and a zero in the middle). Click on the purple "Search" button

You should now have a page describing the signature matches for this protein. (the protein view):

**Question 1:** Looking at the InterPro protein view for O15075, how many InterPro entries (not individual signatures) match the query protein sequence?

**Question 2:** How many domains is the protein divided up into?

**Question 3:** How many member database signatures contribute to InterPro entry IPR003533?

Hint: clicking on the link to IPR003533 will take you to the entry page for that domain.

### Exercise 2 – Exploring InterPro entries: General annotation

Still on the protein page for O15075, look at the match to the entry IPR000719. Notice that there are other domain entries that cover approximately the same sequence position.

- Click on the hyperlink to IPR000719.

**Question 1:** What is the name of this domain?

- Now look at the "Contributing signatures" section.

This section lists the signatures in an entry, the database they come from, and the number of proteins they match.

**Question 2:** Which signatures make up this entry?

- Scroll down to the “GO terms annotation” section.

InterPro provides its own mappings to GO terms based on the curated UniProt/Swiss-Prot proteins matching an entry. These are useful for the annotation of TrEMBL proteins that do not otherwise have GO terms associated with them.

**Question 3:** What GO terms are provided for this entry?

- Choose one GO term and copy/paste the GO ID into the text search box at the top of the InterPro entry page.

This will produce a list of all the InterPro entries associated with this GO term. This is of value if you are interested in searching for InterPro entries that match proteins with a specific function or those involved in a specific process.

- Go back to the IPR000719 entry page.
- Look at the left hand side menu.

**Question 4 :** How many proteins are matched by entry IPR000719?

- . Click on the "Structures" link on the left hand side menu.

InterPro provides a list of all the PDB entries associated with an entry. There are also structural links to SCOP and CATH at the bottom of the page, which provide structural classifications of the proteins that match this entry.

- Scroll to the bottom of the page and follow the “SCOP d.144.1.7” link to the SCOP database to find out the structural classification of this domain.

**Question 5:** What type of structure does the protein kinase-like fold consists of ?

Hint: look at the information under “Fold” in the “Linage” section.

- Click on the “Superfamily” link on the SCOP page, namely Protein kinase-like (PK-like)

**Question 6:** Which families of protein- kinase-like domains does SCOP list? (Note: this is not an exhaustive list of families, as only those with structural information in PDB are included).

- Click on the browser back button twice till you are in the InterPro page again, then click on the "Species" link on the left hand side of the InterPro entry page for IPR000719. You may explore the taxonomic spread by expanding the table.

InterPro divides all the protein hits in an entry by their taxonomy.

**Question 7:** How wide a taxonomic coverage do proteins containing a protein kinase domain have?

## Exercise 3– Exploring InterPro entries: Relationships

- Return to the overview of entry IPR00719 by clicking back button.

InterPro links related signatures through Parent/Child relationships which indicate domain/family hierarchies.

**Question 1:** What “Child” entries is IPR000719 subdivided into?

Child entries subdivide IPR000719 into more closely related subgroups.

**Question 2:** What is the name of the “Parent” of IPR000719?

In this case, the parent entry represents domains with a structural fold homologous to that of the protein kinase domain (even if they have no enzyme activity), whereas IPR000719 represents a more specific form of the domain that has catalytic protein kinase activity.

## Exercise 4– Exploring InterPro entries: Structure

- Return to the InterPro graphical view page of our protein, O15075.

Hint: you can do this quickly by typing "O15075" into the Search box

- Scroll down to the “Structural features” just below the view of unintegrated signatures.

Under the “Structural Features” heading you will find the PDB structure. Its length indicates the region of the protein for which the structure is known. You will also see bars representing a CATH database match and a SCOP database match,

both of which are structural classification databases that break down the PDB structures for the protein into their constituent domains.

**Question 1:** What region is covered by the PDB structure (ie which domain)?

Hint: Compare it to entry IPR003533.

Not all of the protein has been structurally characterised, shown by the fact that only a small region of this protein is covered by the PDB match. To help address this problem, there are homology models from both ModBase and Swiss-Model found under the "Structural Predictions" section. These are models based on aligning our protein with its closest homologue whose structure has been determined. (Note: these are predictive models that provide a 'best guess' at the remaining structure).

**Question 2:** Why does IPR003533 have two domain hits compared to the single domain the PDB structure?

Note the structural view at the top right of the page and click on the GO purple bottom. It will bring you to the PDB page, where you can find the AstexViewer for molecular structures (n the "Tools" menu). Open it (it will pop up in a new window) and take a look at the structure of the doublecortin domain.

# II. Querying sequences using InterProScan

## Learning objectives

In this section, you will learn:

- how to perform sequence-based queries using InterProScan.
- how to use the InterPro matches provided by InterProScan to compare sequences.

If you are using an unknown protein sequence to query InterPro, the simplest way is to copy and paste the amino acid sequence into the large box on the home page and click on the search button immediately to the right. This will run InterProScan on your sequence with the default parameters selected. For more

advanced search options use the InterProScan link, which allows query sequences to be either entered directly or uploaded from a file in different formats (GCG, FASTA, EMBL, GenBank, PIR, etc).

InterProScan incorporates all the analysis algorithms and result post-processing steps from the member databases. InterProScan outputs the resulting matches for a sequence in a graphical format. The matches can also be viewed as a table, which lists the signature match positions.

In addition to the online version of InterPro, a stand-alone version can be downloaded from the ftp server (ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/) and installed locally. Unlike the online version of InterProScan, the standalone version can accept multiple sequences as input.

## Summary

- InterPro Scan can be used with sequence queries as a predictive tool for protein sequence classification and comparison.

## Exercises

## Exercise 5 – Analysing and comparing sequences using InterProScan

- To select a sequence for analysis, use the following url **http://www.ebi.ac.uk/~amaia/** and select seq1

>seq1
MELRVLLCWASLAAALEETLLNTKLETADLKWVTFPQVDGQWEELSGLDEEQHSVRTYEVCDQRAPGQ
AHWLRTGWVPRRGAVHVYATLRFTMLECLSLPRAGRSCKETFTVFYYESDADTATALTPAWMENPYIK
VDTVAAEHLTRKRPGAEATGKVNVKTLRLGPLSKAGFYLAFQDQGACMALLSLHLFYKKCAQLTVNLTR
FPETVPRELVVPVAGSCVVDAVPAPGPSPSLYCREDGQWAEQPVTGCSCAPGFEAAEGNTKCRACAQ
GTFKPLSGEGSCQPCPANSHSNTIGSAVCQCRVGYFRARTDPRGAPCTTPPSAPRSVVSRLNGSSLHL
EWSAPLESGGREDLTYALRCRECRPGGSCAPCGGDLTFDPGPRDLVEPWVVVRGLRPDFTYTFEVTAL
NGVSSLATGPVPFEPVNVTTDREVPPAVSDIRVTRSSPSSLSLAWAVPRAPSGAVLDYEVKYHEKGAEG
PSSVRFLKTSENRAELRGLKRGASYLVQVRARSEAGYGPFGQEHHSQTQLDESEGWREQLALIAGTAV
VGVVLVLVVIVVAVLCLRKQSNGREAEYSDKHGQYLIGHGTKVYIDPFTYEDPNEAVREFAKEIDVSYVKI
EEVIGAGEFGEVCRGRLKAPGKKESCVAIKTLKGGYTERQRREFLSEASIMGQFEHPNIIRLEGVVTNSM
PVMILTEFMENGALDSFLRLNDGQFTVIQLVGMLRGIASGMRYLAEMSYVHRDLAARNILVNSNLVCKVS
DFGLSRFLEENSSDPTYTSSLGGKIPIRWTAPEAIAFRKFTSASDAWSYGIVMWEVMSFGERPYWDMSN
QDVINAIEQDYRLPPPPDCPTSLHQLMLDCWQKDRNARPRFPQVVSALDKMIRNPASLKIVARENGGAS
HPLLDQRQPHYSAFGSVGEWLRAIKMGRYEESFAAAGFGSFELVSQISAEDLLRIGVTLAGHQKKILASV
QHMKSQAKPGTPGGTGGPAPQY

- Open the InterProScan web page (http://www.ebi.ac.uk/Tools/pfa/iprscan/), paste your sequence into the text box and press submit. Note that InterProScan is very forgiving about file format – it won't matter if there is some whitespace around the sequence.

**Question 1:** What functional information can you infer from the domains and sites associated with this protein? [Hint: check out the second and third InterPro hits]

- Now look at this second sequence that is from a patient with a cardio-vascular disease (to select a sequence for analysis, use the following url **http://www.ebi.ac.uk/~amaia/** and select seq2).

>seq2
MELRVLLCWASLAAALEETLLNTKLETADLKWVTFPQVDGQWEELSGLDEEQHSVRTYEVCDVQRAPGQAHW
LRTGWVPRRGAVHVYATLRFTMLECLSLPRAGRSCKETFTVFYYESDADTATALTPAWMENPYIKVDTVAAEHL
TRKRPGAEATGKVNVKTLRLGPLSKAGFYLAFQDQGACMALLSLHLFYKKCAQLTVNLTRFPETVPRELVVPVA
GSCVVDAVPAPGPSPSLYCREDGQWAEQPVTGCSCAPGFEAAEGNTKCRACAQGTFKPLSGEGSCQPCPAN
SHSNTIGSAVCQCRVGYFRARTDPRGAPCTTPPSAPRSVVSRLNGSSLHLEWSAPLESGGREDLTYALRCREC
RPGGSCAPCGGDLTFDPGPRDLVEPWVVVVRGLRPDFTYTFEVTALNGVSSLATGPVPFEPVNVTTDREVPPAV
SDIRVTRSSPSSLSLAWAVPRAPSGAVLDYEVKYHEKGAEGPSSVRFLKTSENRAELRGLKRGASYLVQVRAR
SEAGYGPFGQEHHSQTQLDESEGWREQLALIAGTAVVGVVLVLVVIVVAVLCLRKQSNGREAEYSDKHGQYLI
GHGTKVYIDPFTYEDPNEAVREFAKEIDVSYVKIEEVIGAGEFGEVCRGRLKAPGKKESCVAISTLKGGYTERQR
REFLSEASIMGQFEHPNIIRLEGVVTNSMPVMILTEFMENGALDSFLRLNDGQFTVIQLVGMLRGIASGMRYLAE
MSYVHRDLAARNILVNSNLVCKVSDFGLSRFLEENSSDPTYTSSLGGKIPIRWTAPEAIAFRKFTSASDAWSYGI
VMWEVMSFGERPYWDMSNQDVINAIEQDYRLPPPPDCPTSLHQLMLDCWQKDRNARPRFPQVVSALDKMIRN
PASLKIVARENGGASHPLLDQRQPHYSAFGSVGEWLRAIKMGRYEESFAAAGFGSFELVSQISAEDLLRIGVTLA
GHQKKILASVQHMKSQAKPGTPGGTGGPAPQY

**Question 2:** What is the difference between the two sequences? [Hint: Its about ATP]

**Question 3:** Can you infer a possible reason for the patient's disease? [Hint: think of the functionality of the protein]

This is the end of the tour of the InterPro database, available at the EBI. Perhaps you might like to try it again with a sequence relevant to your research.

# Course summary

InterPro is a diagnostic resource for protein families, domains and functional sites, which integrates the following protein signature databases: PROSITE, PRINTS, ProDom, Pfam, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D and PANTHER. The core concept of InterPro is that similarities and differences between proteins that have the same function or structure can be modelled; and the resultant predictive models provide a powerful tool for the prediction of protein structure and function and hence classification.

# Further reading

Hunter S, et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acid Res* , **37**, D211-5.

Mulder NJ, Apweiler R (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods in Molecular Biology*, **396**, 59-70.

Mulder NJ, et al. (2007) New developments in the InterPro database. *Nucleic Acids Research* ,**35**, D224-8.

# Where to find out more

You can find links to documentation about InterPro (user manual, release notes) in our web main page, as well as download information from our ftp server (ftp://ftp.ebi.ac.uk/pub/databases/interpro/).

# Course exercise answers

## Exercise 1- Searching InterPro using a UniProt identifier.

**Question 1:** Looking at the InterPro protein view for O15075, how many InterPro entries (not individual signatures) match the query protein sequence?

**Answer 1:** Eight (seven domain/sites entries plus the family membership entry)

**Question 2:** How many domains is the protein divided up into?

**Answer 2:** Three (one of them is found repeated in the protein)

**Question 3:** How many signatures are contained within IPR003533?

**Answer 3**: Five

## Exercise 2- Exploring InterPro entries: General annotation

**Question 1:** What is the name of this domain?

**Answer 1**: Protein kinase, catalytic domain

**Question 2:** Which signatures make up this entry?

**Answer 2:** One signature makes up this entry (PS50011)

**Question 3:** What GO terms does this entry provide?

**Answer 3**: GO:0006468 (protein phosphorylation), GO:0004672 (protein kinase activity), GO:0005524 (ATP binding)

**Question 4:** How many proteins are matched by IPR000719?

**Answer 4:** 97768 proteins are matched

**Question 5:** What type of structure does the protein kinase-like fold consists of ?

**A** **Answer 5:** Consists of two alpha+beta domains, and the C-terminal domain is mostly alpha helical.

**Q** **Question 6:** Which families of protein- kinase-like domains does SCOP list?

**A** **Answer 6:** There are seven families listed belonging to the superfamily protein- kinase-like: (1) protein kinase, catalytic unit, (2) actin-fragmin kinase catalytic domain, (3) MHCK/EF2-kinase, (4) phosphoinositide 3-kinase catalytic domain, (5) choline kinase, (6) APH phosphotransferase and (7) RIO-like kinases.

**Q** **Question 7:** How wide a taxonomic coverage do proteins containing a protein kinase domain have?

**A** **Answer 7:** It is widely spread, being found in eukaryota, bacteria, archaea and viruses.

## Exercise 3- Exploring InterPro entries: Relationships

**Q** **Question 1:** What "Child" entries are IPR000719 subdivided into?

**A** **Answer 1:** Two: Serine-threonine/tyrosine protein kinase and serine-threonine protein kinase like.

**Q** **Question 2:** What is the name of the "Parent" of IPR000719?

**A** **Answer 2:** Protein kinase-like domain

## Exercise 4- Exploring InterPro Entries: Structure

**Q** **Question 1:** What region is covered by the PDB structure (ie which domain)?

**A** **Answer 1:** The first doublecortin domain of the two represented by InterPro entry IPR003533.

**Q** **Question 2:** Why does IPR003533 have two domain hits compared to the single domain the PDB structure?

**A** **Answer 2:** IPR003533 predicts the presence of two doublecortin domains, but only the area corresponding to the first one has been structurally characterised and therefore appears in the PBD structure.

## Exercise 5- Analysing and comparing sequences using InterProScan

**Q** **Question 1:** What functional information can you infer from the domains and sites associated with this protein?

**A** **Answer 1:** It is a tyrosine protein kinase with ephrin receptor activity.

**Q** **Question 2:** What is the difference between the two sequences?

**A** **Answer 2:** The second sequence is missing an ATP binding site.

**Q** **Question 3:** Can you infer a possible reason for the patient's disease?

**A** **Answer 3:** The fact that the second sequence is lacking the ATP binding site will probably result in a non-functional or malfunctioning protein. Ephrin receptors seem to have a role in the early development of the circulatory system.