Duncan Legge, Jennifer McDowall, Sarah Burge, Phil Jones (V5.0, Aug 2013)

# Getting Started with InterPro

This tutorial provides an introduction to InterPro, the new user interface and the database content. Please remember that this is a beta release of the new website. Exercises are provided to help you practice what you learn in this tutorial. Further information can be found in the InterPro user manual: http://www.ebi.ac.uk/interpro/user\_manual.html. This manual is due for a complete overhaul – it currently describes the existing public web-site, so watch this space!

### *You will learn about:*

* Querying InterPro via the web interface
* Using signatures to classify and annotate protein sequences
* Understanding signature relationship hierarchies
* Relating signatures to protein structure

*What is InterPro?*

InterPro is a curated resource for classifying and annotating protein families, domains and sites. InterPro combines a number of independent protein signature databases (referred to as member databases) into a single searchable resource. InterPro provides protein signatures for over 80% of proteins in the UniProt database. InterPro provides additional information through GO mapping, structural links, external database links and comprehensive abstracts and references.

### *What are protein signatures?*

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

* 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.

### *The InterPro home page*

## The development version of the InterPro home page 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...)
* Link to tools, such as the BioMart.

### *How do I search InterPro?*

InterPro can be searched a number of different ways:

* **Text search**
  + Via the search bar at the top of all InterPro web pages.
  + Search using: UniProtKB accessions; InterPro entry IDs; GO terms; plain text.
* **InterProScan**
  + Link to InterProScan on InterPro home page.
  + Search using protein or nucleotide sequence.
  + Incorporates all the search engines from the member databases.
* **BioMart**
  + Link to BioMart on InterPro home page.
  + Easy querying; retrieve results in HTML, plain text or Microsoft Excel spreadsheet.

Note: To analyse a specific protein, do a text search for its UniProtKB accession number – this will provide both the signature hits and structural information for the protein. By contrast, the InterProScan search only provides the signature hits. InterProScan is best used for novel sequences not found in UniProtKB. If you have an accession number from GenBank, Xref, EMBL or Ensembl, you can convert it to a UniProtKB accession number using PICR (<http://www.ebi.ac.uk/Tools/picr/>).

EXERCISE 1:

Searching InterPro using a text search

* Open the InterPro homepage (<http://www.ebi.ac.uk/interpro/>) in a web browser.
* Using the “Text” Search in the silver box, 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.

* Read below to understand how to interpret this protein match:

### *How do I interpret an InterPro protein view?*

In InterPro, the Protein View consists of the following features:

* **Top Section**
  + First of all we present the basic information about the protein: name, short name, identifier, links to UniProtKB and taxonomic information.
* **Protein family membership**
  + The protein families that a given protein belongs to are presented in a hierarchy. Clicking on the links takes you to the entry pages for each family.
* **Sequence features**
  + Each solid coloured bar represents a signature that matches the protein.
  + Domain signatures are presented at the top, 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. You will often find that were there are multiple matches to the same region of the protein, they are the same colour as these are matches to signatures describing the same type of feature.
  + If you hover your mouse over the coloured bar, a pop-up will display the signature accession and the sequence matched. Clicking on the link will take you to the member database page about this signature
  + The length of the protein is indicated by the white vertical bars, which are marked (in this case) every 20 amino acids.
  + The InterPro entry the signature belongs to is indicated and linked on the left-hand column.
* **Structural Features**
  + One match represents coverage of a PDB chain (PDB accession on the left).
  + The other two bars represent the structural classification databases CATH and SCOP, respectively.
  + CATH and SCOP break the PDB into its structural domains.
* **Structural Predictions**
  + Two matches are presented here, one to the homology databases ModBase and the other to the Swiss-Model database.
  + These homology models estimate the structure based on the closest homologue.
* Looking at the InterPro protein view for O15075, how many InterPro entries (not individual signatures) match the query protein sequence?
* How many domains is the protein divided up into?
* How many signatures are contained within IPR003533? (Hint: clicking on the link to IPR003533 will take you to the entry page for that domain).

There can be more than one InterPro entry for a single domain, because InterPro represents domain hierarchies (i.e. structural domain superfamilies, domain families and domain subfamilies). Therefore, look at all the signatures together, and see how many regions the protein is broken into along the sequence. Also, several signatures can belong to the same InterPro entry if they represent the same biology.

**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.
* What is the name of this domain?
* Scroll down to the “Contributing signatures” section.

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

* Which signatures make up this entry and how many proteins are matched?
* Scroll down to the “GO term 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 TrEMBL proteins that do not otherwise have GO terms associated with them.

* What GO terms does this entry provide?
* 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 proteins with a specific function or those involved in a specific process.

* Go back to the IPR000719 entry page. 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.

* 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 on the "Overview" link on the left hand side.

InterPro links related signatures through Parent/Child relationships which indicate domain/family hierarchies. Parent entries are more general (i.e. are representative of more proteins) than Child entries, which are more specific entries.

* What “Child” entries are IPR000719 subdivided into?

Child entries subdivide IPR000719 into more closely related subgroups.

* What is the name of the “Parent” of IPR000719?

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 proteins more closely related by sequence.

**Exercise 4:**

### InterProScan

* Analyze the following sequence using InterProScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan/>) Note that InterProScan is very forgiving about file format – it won’t matter if there is some whitespace around the sequence.

>seq1

MELRVLLCWASLAAALEETLLNTKLETADLKWVTFPQVDGQWEELSGLDEEQHSVRTYEV

CDVQRAPGQAHWLRTGWVPRRGAVHVYATLRFTMLECLSLPRAGRSCKETFTVFYYESDA

DTATALTPAWMENPYIKVDTVAAEHLTRKRPGAEATGKVNVKTLRLGPLSKAGFYLAFQD

QGACMALLSLHLFYKKCAQLTVNLTRFPETVPRELVVPVAGSCVVDAVPAPGPSPSLYCR

EDGQWAEQPVTGCSCAPGFEAAEGNTKCRACAQGTFKPLSGEGSCQPCPANSHSNTIGSA

VCQCRVGYFRARTDPRGAPCTTPPSAPRSVVSRLNGSSLHLEWSAPLESGGREDLTYALR

CRECRPGGSCAPCGGDLTFDPGPRDLVEPWVVVRGLRPDFTYTFEVTALNGVSSLATGPV

PFEPVNVTTDREVPPAVSDIRVTRSSPSSLSLAWAVPRAPSGAVLDYEVKYHEKGAEGPS

SVRFLKTSENRAELRGLKRGASYLVQVRARSEAGYGPFGQEHHSQTQLDESEGWREQLAL

IAGTAVVGVVLVLVVIVVAVLCLRKQSNGREAEYSDKHGQYLIGHGTKVYIDPFTYEDPN

EAVREFAKEIDVSYVKIEEVIGAGEFGEVCRGRLKAPGKKESCVAIKTLKGGYTERQRRE

FLSEASIMGQFEHPNIIRLEGVVTNSMPVMILTEFMENGALDSFLRLNDGQFTVIQLVGM

LRGIASGMRYLAEMSYVHRDLAARNILVNSNLVCKVSDFGLSRFLEENSSDPTYTSSLGG

KIPIRWTAPEAIAFRKFTSASDAWSYGIVMWEVMSFGERPYWDMSNQDVINAIEQDYRLP

PPPDCPTSLHQLMLDCWQKDRNARPRFPQVVSALDKMIRNPASLKIVARENGGASHPLLD

QRQPHYSAFGSVGEWLRAIKMGRYEESFAAAGFGSFELVSQISAEDLLRIGVTLAGHQKK

ILASVQHMKSQAKPGTPGGTGGPAPQY

* Now, open a new tab and using InterProScan look at this second sequence that is from a patient with a cardio-vascular disease. What is the difference?
* Can you infer a possible reason for the patients disease?

>seq2

MELRVLLCWASLAAALEETLLNTKLETADLKWVTFPQVDGQWEELSGLDEEQHSVRTYEV

CDVQRAPGQAHWLRTGWVPRRGAVHVYATLRFTMLECLSLPRAGRSCKETFTVFYYESDA

DTATALTPAWMENPYIKVDTVAAEHLTRKRPGAEATGKVNVKTLRLGPLSKAGFYLAFQD

QGACMALLSLHLFYKKCAQLTVNLTRFPETVPRELVVPVAGSCVVDAVPAPGPSPSLYCR

EDGQWAEQPVTGCSCAPGFEAAEGNTKCRACAQGTFKPLSGEGSCQPCPANSHSNTIGSA

VCQCRVGYFRARTDPRGAPCTTPPSAPRSVVSRLNGSSLHLEWSAPLESGGREDLTYALR

CRECRPGGSCAPCGGDLTFDPGPRDLVEPWVVVRGLRPDFTYTFEVTALNGVSSLATGPV

PFEPVNVTTDREVPPAVSDIRVTRSSPSSLSLAWAVPRAPSGAVLDYEVKYHEKGAEGPS

SVRFLKTSENRAELRGLKRGASYLVQVRARSEAGYGPFGQEHHSQTQLDESEGWREQLAL

IAGTAVVGVVLVLVVIVVAVLCLRKQSNGREAEYSDKHGQYLIGHGTKVYIDPFTYEDPN

EAVREFAKEIDVSYVKIEEVIGAGEFGEVCRGRLKAPGKKESCVAISTLKGGYTERQRRE

FLSEASIMGQFEHPNIIRLEGVVTNSMPVMILTEFMENGALDSFLRLNDGQFTVIQLVGM

LRGIASGMRYLAEMSYVHRDLAARNILVNSNLVCKVSDFGLSRFLEENSSDPTYTSSLGG

KIPIRWTAPEAIAFRKFTSASDAWSYGIVMWEVMSFGERPYWDMSNQDVINAIEQDYRLP

PPPDCPTSLHQLMLDCWQKDRNARPRFPQVVSALDKMIRNPASLKIVARENGGASHPLLD

QRQPHYSAFGSVGEWLRAIKMGRYEESFAAAGFGSFELVSQISAEDLLRIGVTLAGHQKK

ILASVQHMKSQAKPGTPGGTGGPAPQY

# Answers

### EXERCISE 1: Searching InterPro using a text search

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

7

* How many domains is the protein divided up into?

2 – Doublecortin & Kinase

* How many signatures are contained within IPR003533? (Hint: clicking on the link to IPR003533 will take you to the entry page for that domain).

5

**Exercise 2:**

* What is the name of this domain?

IPR000719Protein kinase, catalytic domain

* Which signatures make up this entry and how many proteins are matched?

Pfam - PF00069 (Pkinase)

PROSITE profile - PS50011 (PROTEIN\_KINASE\_DOM)

* What GO terms does this entry provide?

Biological Process:

GO:0006468 protein phosphorylation

Molecular Function:

GO:0004672 protein kinase activity

GO:0005524 ATP binding

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

Very wide – as it appears in many different species.

**Exercise 3:**

* What “Child” entries are IPR000719 subdivided into?

Protein kinase-like domain, Apicomplexa (IPR027916)

Serine-threonine/tyrosine-protein kinase catalytic domain (IPR001245)

Serine/threonine- / dual-specificity protein kinase, catalytic domain (IPR002290)

* What is the name of the “Parent” of IPR000719?

Protein kinase-like domain (IPR011009)

**Exercise 4:**

* Now, open a new tab and using InterProScan look at this second sequence that is from a patient with a cardio-vascular disease. What signature is missing?

Tyrosine-protein kinase, active site.

* Can you infer a possible reason for the patients disease?

The kinase function is ineffective in this variant (Seq2) as the ATP binding site is not present (IPR017441).



### *Further reading*

*Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation..* Burge S, *et al*. Database (Oxford) 2012

<http://dx.doi.org/10.1093/database/bar068>

*InterPro in 2011: new developments in the family and domain prediction database..* Hunter S *et al.* Nucleic Acids Res [2012 (40) ] page:d306-12

<http://dx.doi.org/10.1093/nar/gkr948>