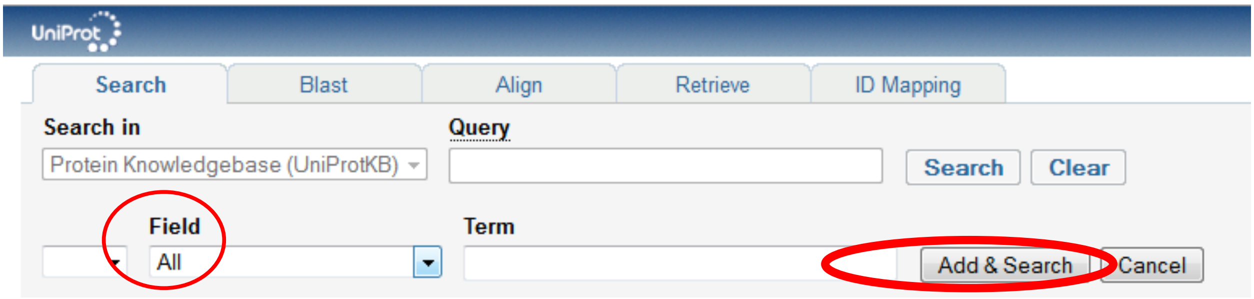
Sandra Orchard (V2 03/10/12)

UniProt – the protein sequence database

www.uniprot.org

UniProt (Universal Protein Resource) is the world's most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function produced by the UniProt Consortium, comprised of the European Bioinformatics Institute, the Swiss Institute of Bioinformatics and the Protein Information Resource, Washington. The UniProt consortium aims to support biological research by maintaining a high quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community. The **UniProt Knowledgebase** (UniProtKB) is the central access point for extensive curated protein information, including function, classification, and cross-reference. The database is divided into two section UniProtKB/SwissProt which is manually curated and UniProtKB/TrEMBL which is automatically maintained. During this course you will concentrate on UniProtKB/SwissProt and learn how to access the entries in the database, extract the maximum amount of information from them.

All data stored in UniProt can be downloaded from the Download Centre at http://www.ebi.uniprot.org/database/download.shtml



Exercise 1 – Exploring UniProtKB

Use the Advanced Search to search on the keyword 'Malaria' using the 'Field' button

You will now be searching all the entries in UniProtKB where the protein has been identified as playing a role in the human disease, malaria.

From your results list, select Q8I5D2, the 101 kDa malaria antigen protein from Plasmodium falciparum (isolate 3D7).

Q 1. Where in the host cell is this protein found

Move up to the section headed ‘Names and Origin’. You will see under the sub0heading ‘Organism’ that Plasmodium falciparum (isolate 3D7) is described as a ‘Reference Proteome’. Click on the hyperlink to access all proteins within that reference protoeome.

Q 2. How many proteins are there in this reference proteome. How many have been manually reviewed in UniProtKB/Swiss-Prot and how many are unreviewed in TrEMBL?

Retaining the current query, use the Advanced search to find out how many of the proteins in this reference proteome have been crystallized (Hint – under ‘Field’ select ‘Cross-reference’, then PDB from the additional drop-down list which then appears, the wildcard the search using ‘*’ in the query box.

Q 3. How many proteins in the Plasmodium falciparum (isolate 3D7) have been crystallized?

Open the entry for **Q9TY95** (SERA_PLAF7).

Q 4. How many times has this protein been crystallized

Click the PDBe hyperlink for crystal 3CH2 to access the details of this crystal.

Q 5. Does this protein most likely crystallize as a monomer or a dimer?

Return to the entry and scroll to Binary Interation. SERA has been shown to interact with PfSUB1 (Q8I0V0). Click on the hyperlink to look at this protein in IntAct.

Q 6. Is this a Swiss-Prot or TrEMBL entry?

Q 7. What reaction is catalysed by this enzyme? (hint – click on the EC number)

The human protein ‘Low affinity immunoglobulin gamma Fc region receptor II-b’ P31994 (FCG2B_HUMAN) has been associated with malaria. Search and open this entry in UniProt.

Q 8. Which natural variant has been associated with a susceptibility to malaria? Hint – look in the Sequence annotation section

Alternative Products in UniProtKB

Return to the entry and go to the Alternative Products section.

FCG2B is a protein for which multiple (3) isoforms have been identified. All of these are mapped within UniProtKB, and given stable identifiers.

Press the button “Align” to see how all the isoforms differ.

Return to the Alternative Products section.

Q 9. Which isoform has a shortened extra-cellular domain?

Adding information using the InterPro Database

InterPro is a classification resource for protein families, domains and sites, combining a number of member databases that use different methodologies and biological information on well-characterised proteins to derive protein signatures. Signatures describing the same protein family, domain, repeat or site are grouped into unique InterPro entries. Each combined InterPro entry has a unique accession number, an abstract describing the features of proteins associated with the entry and literature references and has links to the relevant member database(s). The InterPro graphical view shows the position of the signatures on the protein, mousing over the signature brings up a pop-box, giving the accession, name and position.

InterPro graphically represents the location of a protein domain and information pertaining to the origin of that domain and the proteins that contain it. Families are also defined and may contain several InterPro domains which are often, but not always, in the same order. InterPro entries can be linked to one another through PARENT/CHILD relationships which indicate superfamily/family/subfamily relationships, as well as domain hierarchies, where sequences can be subdivided into more specific sub-sets. InterPro entries that belong to a UniProtKB entry can be found in Cross-references -> Family and domain databases

InterPro and InterProScan are accessible for interactive use over the EBI web server (www/ebi.ac.uk/interpro), they are distributed as stand-alone copies by anonymous ftp.

Exercise 2 – Exploring InterPro

Open the InterPro homepage (InterPro) in a web browser (http://www.ebi.ac.uk/interpro/).

Using the "Text" Search box mid-way down the page, type in the UniProtKB accession '**Q8I0V0**'. Click on the purple "Search" button.

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

Q10. Looking at the InterPro protein view for Q8I0V0, how many InterPro entries (not individual signatures) match the query protein sequence?

Q11. How many sites does the protein contain?

Q12. How many member database signatures contribute to InterPro entry IPR000209? Hint - you can see the contributing member database signatures to InterPro entry IPR003533 in the view or alternatively click on the link to IPR003533, which will take you to the entry page for that domain.

Now search on Q9TY95. Go to the "Overview" box in the top left hand corner and select "Structures"

Under the "Structures" 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.

Q13. Comparing the sequence feature to the structural domain, what region is covered by the PDB structure (ie which domain)?

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).

At the bottom of the page, you can view the solved structures either by mouse over or opening up in a separate window

Exploring InterPro entries

General annotation

Return to the overview of Q9TY95, look at the match to the entry IPR013128.

Click on the hyperlink to IPR013128

Q14. What is the name of this Family?

Relationships

Q15. What “Child” entries is IPR013128 subdivided into?

InterPro links related signatures through Parent/Child relationships which indicate domain/family hierarchies. Child entries subdivide IPR013128 into more closely related subgroups.

Taxonomy

Click on the "Species" link on the left hand side of the InterPro entry page for IPR013128. InterPro divides all the protein hits in an entry by their taxonomy. Scroll down to ‘Taxa’

Q16. How wide a taxonomic coverage do proteins of the Peptidase C1A, papain family have?

GO (Gene Ontology) terms

Return to the Overview for Q9TY95. Scroll down to the “GO terms annotation” section.

Q17. What GO terms are predicted for this entry?

InterPro provides its own mappings to GO terms based on the curated UniProtKB/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.

A tour of the IntAct Portal

http://www.ebi.ac.uk/intact

IntAct is a molecular interaction database which provides the user with all the experimental detail described in the originating paper. Entries are fully IMEx- and MIMIx-compliant and provide extra levels of detail beyond these minimum requirements. IntAct makes extensive use of a number of controlled vocabularies, primarily PSI-MI to describe the technical details of the experiment, binding sites, protein tags and mutations and Gene Ontology to describe the subcellular location an interaction may be shown to occur in or the function of an enzyme in an enzyme/substrate assay. Interacting molecules are mapped to stable identifiers from public databases such as UniProtKB for proteins, ChEBI for small molecules, Ensembl for genes and the DDBJ/EMBL/GenBank nucleotide databases for nucleic acids. Features within a molecule, such as the binding site of a protein, are mapped to the sequence/structure given in the relevant database and remapped should a revised version of the sequence be released. Binding sites are cross-referenced to the InterPro database, whenever possible.

Exercise 3 – Exploring IntAct

1. *Using the Quick Search*

In this search panel you may type anything that might relate to interactions, whether it is properties of their interactor (gene name, *Accession Numbers, GO term…) or more specific to the interaction such as publication ID, authors, experimental detection method, …*

In this exercise you will perform a very simple search and look at your results in the IntAct viewer. You may perform the exercise below or alternatively, you may try a protein you are interested in through your own work. However, you may find there is little or no data for your protein in the database, particularly if it originates from a non-model organism which are much less well studied.

Open the IntAct homepage in a web browser (http://www.ebi.ac.uk/intact).

Type 'SERA' into the IntAct Quick search and hit Search

Q18 How many binary interactions does this find?

Searching on a non-specific gene name will bring you up a mixed set of results. Refine your search using either **Q9TY95** or SERA_PLAF7.

Q19 How many binary interactions does the refined search find?

Graphically view the data withinIntAct

Open the 'List' tab and select all proteins (use the check box at the top of the table) and use 'Search interactions' to perform a second round of search.

Go to the detailed view of the interaction of SERA (Q9TY95) with PfSUB1 (EBI-1568875)

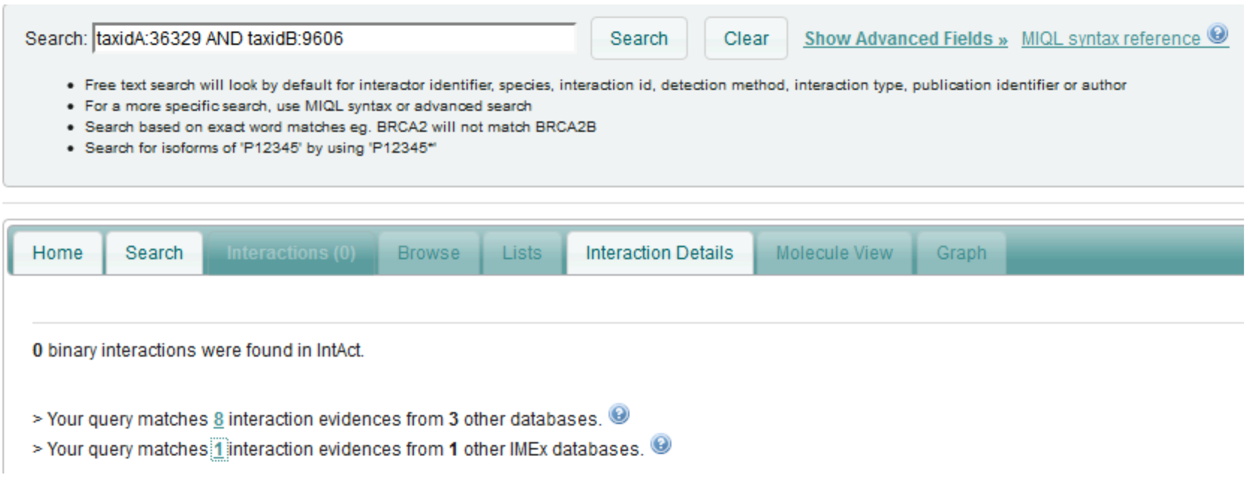
Q20 Which of these proteins acts as the enzyme in this reaction?

Q 21 What molecule inhibits this interaction?

2. *Refining your search using the Advanced Search*

If you want to construct more complex queries we recommend you take a look at the Molecular Interaction Query Language, accessible from the quick search panel. This will allow you to write more complex queries, for example:

To discover if any interactions are known between Plasmodium falciparum (isolate 3D7) and human proteins, Clear any previous searches, then type 'taxidA:36329 AND taxidB:9606'.



Search: taxidA:36329 AND taxidB:9606

- Free text search will look by default for interactor identifier, species, interaction id, detection method, interaction type, publication identifier or author
- For a more specific search, use MIQL syntax or advanced search
- Search based on exact word matches eg. BRCA2 will not match BRCA2B
- Search for isoforms of 'P12345' by using 'P12345*'

0 binary interactions were found in IntAct.

> Your query matches 8 interaction evidences from **3** other databases.

> Your query matches 1 interaction evidences from **1** other IMEx databases.

We currently (Jan 2013) do not have any such data in IntAct but you can access some via our IMEx partners. Click on 'Your query matches 1 interaction evidences from 1 other IMEx databases' to access this information,

However, it is much easier to use the IntAct 'Show Advanced Fields' option. Clicking on this to the right of the Quick Search box on the IntAct homepage will open up the Advanced Search, allowing you to specify one or more fields you wish to search in, and building the query for you as you progress.

Clear the previous search, then use Advanced Fields to search for 'organism' = human AND 'Interaction detection method' = two hybrid

Q Q22 How many interactions do see for 'organism' = human AND 'Interaction detection method' = two hybrid?

3. Using the Ontology Search

Open the Search Tab. This panel is specialised to give you an easy access to ontology search. So far you can search on 4 ontologies:

Gene Ontology
InterPro
PSI-MI
ChEBI

Whenever you start typing a query in this search panel, the system will search as you type and propose a list of matching controlled vocabulary terms. You can then select one of them and select matching interactions.

Type: innate immune response in the Ontology Search box.

You will be presented with a few choices, please note that each term is followed by the count of matching interactions in the IntAct database.

Select the parent term 'innate immune response' (GO:0045087) using the keyboard cursor keys, complete the search and you will be taken to the interaction tab. This now gives you ALL the interactions for proteins in IntAct which GO have annotated as being involved in the process of innate immunity. Add the term 'AND species:human' to limit this to interactions in which one of the interactors is of human origin.

A look at pathways in Reactome

We will now look at Reactome - the user interfaces and the database content. Further information can be found in the online Reactome user guide at http://www.reactome.org/userguide/Usersguide.html.

Experimental infection of humans with *Plasmodium falciparum* primes Toll-like receptor (TLR)-mediated proinflammatory responses.

Exercise 4 – Exploring Reactome

Search on TLR3_HUMAN in the Quick Search on the Home Page

Q23. How many pathways does this protein appear in?

Click on the Pathway: Toll Receptor Cascades and expand the pathway down via ‘Trafficking and processing of endosomal TLR’ to ‘Full-length TLR3/7/8/9 binds to UNC93B1’

Q24 Is this reaction conserved in mouse, or in Plasmodium falciparum (Hint – expand ‘Details’ triangle)

Return to the home page and use the Species Comparison option to compare human with Plasmodium falciparum

Q25 How well conserved is the Eukaryotic Translation Initiation pathway between the 2 species, and how well is the innate immune system

Now use the Sample dataset given in the Analyse Expression Data to get a feel for how you can analyse your data using Reactome.

Further Info

Sequence Searching

For more sequence searching tools visit http://www.ebi.ac.uk/Tools/similarity.html.

There are two main programs that implement BLAST searches; WU-BLAST 2.0 and NCBI BLAST2. They are distinctly different software packages, although they have a common lineage for some portions of their code, so the two packages do their work differently and obtain different results and offer different features. You can also check for vector contamination with Blast2 EVEC.

Fasta can be very specific when identifying long regions of low similarity especially for highly diverged sequences. You can also conduct sequence similarity searching against complete proteome or genome databases using the Fasta programs.

MPsrch – Smith and Waterman algorithm, capable of identifying hits in cases where Blast and Fasta fail and also reports fewer false-positive hits.

Further Reading

- The UniProt Consortium "Reorganizing the protein space at the Universal Protein Resource (UniProt)." Nucleic Acids Res. (2012) 40:71-75
- Leinonen R, Nardone F, Zhu W, Apweiler R "UniSave: the UniProtKB sequence/annotation version database." Bioinformatics (2006) 22:1284-1285
- Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH "UniRef: comprehensive and non-redundant UniProt reference clusters." Bioinformatics (2007) 23:1282-8
- Mulder NJ et al "New developments in the InterPro database." Nucleic Acids Res. (2007) 35:D224-8
- Kerrien S et al "The IntAct molecular interaction database in 2012." Nucleic Acids Res. (2012)40:D841-846

Exercise 1 – UniProt

Q Q 1. Where in the host cell is this protein found?

A *At the merozoite surface and within the parasitophorous vacuole (click on hyperlinks to find further definitions of locations)*

Q Q 2. How many proteins are there in this reference proteome. How many have been manually reviewed in UniProtKB/Swiss-Prot and how many are unreviewed in TrEMBL?

?

A *Reference proteome =5354. UniProtKB/Swiss-Prot=144, UniProtKB/TrEMBL = 5210*

Q Q 3. How many proteins in the Plasmodium falciparum (isolate 3D7) have been crystallized?

A *94*

Q Q 4. How many times has this protein been crystallized?

A Twice

Q Q 5. Does this protein most likely crystallize as a monomer or a dimer?

A *Monomer*

Q Q 6. Is this a Swiss-Prot or TrEMBL entry?

A *TrEMBL*

Q Q 7. What reaction is catalysed by this enzyme?

A *Hydrolysis of proteins with broad specificity for peptide bonds, and a preference for a large uncharged residue in P1. Hydrolyzes peptide amides*

Q Q 8. Which natural variant has been associated with a susceptibility to malaria?

A *Position 232 I → T*

Q Q 9. Which isoform has a shortened extra-cellular domain?

A *Isoform IIB3 (identifier: P31994-3) Positions 39-45: Missing.*

Exercise 2 - InterPro.

Q **Q 10.** Looking at the InterPro protein view for Q8I0V0, how many InterPro entries (not individual signatures) match the query protein sequence?

A *Six (IPR015500, IPR017314, IPR000209, IPR023827, IPR022398, IPR023828)*

Q *Q 11.* How many sites does the protein contain?

A *Three*

Q Q 12. How many member database signatures contribute to InterPro entry IPR000209?

A *Three*

Q Q 13. Comparing the sequence feature to the structural domain, what region is covered by the PDB structure (i.e. which domain)?

A *Peptidase C1A, papain C-terminal domain (IPR000668).*

Q Q14. What is the name of this Family?

A *Peptidase C1A, papain (IPR013128)*

Q Q15. What "Child" entries is IPR013128 subdivided into?

A *Peptidase C1A, cathepsin B (IPR015643), Peptidase C1A, cathepsin K (IPR015644), Peptidase C1A, placentally-expressed cathepsin (IPR015645)*

Q Q16. How wide a taxonomic coverage do proteins of the Peptidase C1A, papain family have?

A *They are widely spread, being found in eukaryota, bacteria, archaea and viruses.*

Q Q17. What GO terms are predicted for this entry?

A *Biological Process - GO:0006508 proteolysis; Molecular Function - GO:0008234 cysteine-type peptidase activity*

IntAct

Q Q18 How many binary interactions does this find?

A *68*

Q Q19 How many binary interactions does the refined search find?

A *29*

Q Q20 Which of these proteins acts as the enzyme in this reaction?

A *PfSUB1*

Q Q21 Which molecule acts as an inhibitor of this reaction?

A MRT12113, a compound originally purified natural products from temperate plants.

Q Q22 How many interactions do see for ‘organism’ = human AND ‘Interaction detection method’ = two hybrid?

A 27,792

Reactome

Q Q23. How many pathways does this protein appear in?

A 13

Q Q24. Is this reaction conserved in mouse, or in Plasmodium falciparum (Hint – expand ‘Details’ triangle)?

A 13

Q Q25 How well conserved is the Eukaryotic Translation Initiation pathway between the 2 species, and how well is the Innate Immune System

A Eukaryotic Translation Initiation pathway 88% (99/112 proteins), Innate Immune System 4% (24/558 proteins)