Introduction to the IntAct Database

www.ebi.ac.uk/services

The IntAct database is a freely available, open source database system and analysis tool for protein interaction data providing a resource for the storage, presentation and analysis of protein interactions. All the interactions are modelled using literature curation, as well as through direct user submissions. The web interface provides textual and graphical representations of protein interaction networks, which users can explore in the context of Gene Ontology annotations of the interacting proteins. The aim of this tutorial is to familiarise users with the wealth of annotated protein interaction data available within the IntAct database, how to extract this information, and how to use IntAct to help in the functional annotation of proteins.

For this tutorial, we will access IntAct through the UniProt database, using the MPsrch sequence search tool.

ALL THE TOOLS AND DATABASES USED IN THIS TUTORIAL CAN BE ACCESSED VIA THE PAGE: **www.ebi.ac.uk/services**

The Starting Point – a sequence (peptide or protein)

Protein A:
```
MSEQSTSLGSRRVGPPLHKKALRVCFLRNGDRHFKGVNLVISRAHFKDFPALLQGVTESLKRHVLL
RSAIAHFRRTDGSHLTSLSCFRETDIVICCCKNEEIICVKYSINKDFQRMVDSCKRWGQHHLDSGT
LESMKSHDLPEAIQLYIETIEPVEHNTRTLIYRGQTRANRTKCTVKMVNKQTQSNDRGDTYMEAEV
LRQLQSHPNIIELMYTVEDERYMYTVLEHLDCNMQKVIQKRGILSEADARSVMRCTVSALAHMHQL
QVIHRDIKPENLLVCSSSGKWNFKMVKVANFDLATYYRGSKLYVRCGTPCYMAPEMIAMSGYDYQV
DSWSLGVTLFYMLCGKMPFASACKNSKEIYAAIMSGGPTYPKDMESVMSPEATQLIDGLLVSDPSY
RVPIAELDKFQFLAL
```

➢ **From the EBI web services page, follow the link to the *MPsrch* page under "Toolbox". Use MPsrch to find the proteins with the closest sequence identity to your search sequence.**
➢ **On the MPsrch page:**
   • **Paste sequence into MPsrch**
   • **Add e-mail address if you want to keep the results to look at later**
   • **Check "Database" is set at "UniProt"**
   • **Check "Program" is set at "MPsrch_pp"**
   • **Check "Results" is set at "Interactive"**
   • **Press "Run".**

> **On the MPsrch results page, click on the link to the protein with a 100% match, Q9VCL7_DROME.**

From both the "Keywords" section and the "InterPro" line in the "Cross-references" section, Q9VCL7_DROME is predicted to be a serine/threonine kinase.

> **On the UniProt entry page for Q9VCL7_DROME, scroll down to the "Interaction" line in the "Comments" section.**

The interaction data is supplied by IntAct, with links to the IntAct database. Each line in the table represents a protein that interacts with this entry sequence, namely Q9VCL7_DROME. Listed in the table are the UniProt accession number of the interacting protein, its gene name, the number of experiments that have been recorded, and the IntAct interaction accession number for the interaction. Links to IntAct interaction can also be found in the Cross-reference section, where the link provides a summary of all the interactions involving the entry protein.

> **Scroll down to the "IntAct" line in the "Cross-reference" section.**
> **Click on the hyperlinked accession number "Q9VCL7".**

The IntAct table shows the query protein (Q9VCL7_DROME) highlighted in red, and any proteins that it interacts with listed below. The information in the table, which includes the IntAct accession number, UniProt accession number, gene name, description and number of interactions, contain various hyperlinks to the relevant data.

? Which proteins does your sequence interact with?

> ➢ **Click on the hyperlinked UniProt accession number, "P40945", for the interacting protein in order to see details of this protein.**

? Has the interacting protein been characterised in Swiss-Prot?

If the unknown query protein interacts with an annotated protein of known function, the details of the interaction can tell you more about your protein.

? What is the function of this protein?
? What family of proteins does it belong to?

> ➢ **Go back to the IntAct table page.**

Now we know more about the protein with which Q9VCL7_DROME interacts, we can take a look at the interaction details. IntAct stores an interaction in the context of the experimental method that originally described the interaction. If an interaction has been described several times using different methodologies, this will increase the "Number of Interactions" and also our confidence in the veracity of the interaction.

> ➢ **Click on "1" under the "Number of interactions" section to see details of the experimental methodology.**

The experimental summary page provides all the annotated experimental data extracted from the source literature or submission. There are many different types of annotation, including modifications of experimental methods, confidence mappings in the experiment and an author list, as shown in this entry. Direct links are provided to the source literature, as well as to external databases, such as for taxonomy.

? What type of experimental method was used?

Note that this type of high-throughput experiment can sometimes yield artificial results in cases where the two interacting proteins do not come into proximity with one another *in vivo*.

? How high a confidence do the authors attribute to the experimental data?

The interaction data can also be viewed graphically.

> ➢ **Go back to the original IntAct table.**
> ➢ **Check the box to select the ARF2_DROME protein, and then click "Graph" to see a graphical view.**

> ➢ **Centre the graph by checking the "centre the view" option on the left-hand panel (usually the default setting), and clicking on "ARF2_DROME".**

In the top right hand corner, the GO terms for the ARF2_DROME protein can be found, which help define the function of the interacting protein.

? Which processes does GO annotation suggest this protein is involved in?

The interactions can be viewed either in a graphically view, or in tabular form.

> ➢ **Click on the hyperlinked interaction network accession number at the top of the graph, namely "EBI-134336".**

? What other proteins does ARF2_DROME interact with?

The graph can be expanded out to see entire interaction networks.

> ➢ **Go back to the IntAct Graph page.**
> ➢ **Highlight the "Add a network" option on the left-hand panel, and then click on "Expand" Graph on the same panel.**

The position of both ARF2_DROME and our query protein, Q9VCL7_DROME, can now be seen in relation to the entire network of interactions.

? Are there anymore ARF family proteins in this network?
? What other Swiss-Prot characterised proteins are present?
(Hint: TrEMBL entries begin with an accession number, whereas Swiss-Prot entries begin with a condensed name).

All these proteins could be explored to add further information about the network that our query protein is involved in.

We know that our protein (Q9VCL7_DROME) is predicted to be a serine/threonine kinase, which act to phosphorylated their substrates. Also, a single Y2H experiment suggests that it may interact with an ARF family protein as part of a large interaction network.

? Are Arf family members regulated by phosphorylation?

To address this question, we need to carry out a search of the UniProt database.

➢ **To search for Arf family members in *Drosophila melanogaster*, return to www.ebi.ac.uk/services.**
➢ **Select UniProt, and then Power Search.**

<u>Drosophila melanogaster Search</u>:

➢ **Find all the UniProt entries with the species *Drosophila melanogaster*:**
   • **Select "Species/Organism (OS)" under the "Query line type"**
   • **Type "Drosophila melanogaster" as the query text**
   • **Select "Search and View".**

We now have a data set containing all *Drosophila melanogaster* proteins.

Arf Family Member Search:

➤ **Go back to the Power Search page.**
➤ **Find all the UniProt entries with the word "arf":**
  • **Select "Any Field" under the "Query line type"**
  • **Type "arf" as the query text**
  • **Select "Search and View".**

We now have two data sets, one with *Drosophila melanogaster* proteins, and the other with arf-related proteins.

Manipulate Data Sets Using Data Set Manager:

➤ **On the results page, select the link "Data Set Manager".**

The Data Set Manager provides a means of combining data using specific operators, thereby permitting very complex data sets to be built. The two pull-down menus are used to select the two data sets you wish to manipulate, denoted "set A" and "set B". The Venn diagrams are three standard operators: union, intersection and difference. The first Venn diagram (solid colour) is the union operator, which will combine the two data sets into one group. The second Venn diagram (centre coloured) is the intersection operator, which will select the proteins in common between the two data sets. The third Venn diagram (left coloured) is the difference operator, which will select the proteins found in set A that are not in set B.

➤ **Use the pull-down menus to select "arf" as set A, and "Drosophila melanogaster" as set B.**
➤ **Select the Venn diagram denoting the intersection operator (centre coloured), and then press "Combine Data Sets"**

The first data set in the table now contains the arf proteins that are found in *Drosophila melanogaster*.

? How many Drosophila melanogaster proteins are members of the Arf family?

➤ **Click on the link under "Query" for the combined data set.**

All the Arf member *Drosophila* proteins are now listed with links to their UniProt entries, with the Swiss-Prot entries listed first. The "Interaction" section of the Swiss-Prot entries can be checked for any known interactions with kinases.

➤ **Click on the UniProt entry "CEG1A_DROME", which is a well-characterised protein.**
➤ **Scroll down to the "IntAct" line in the "Cross-reference" section.**
➤ **Click on the hyperlinked accession name, "Q9NGC3".**

? Does this *Drosophila* arf family protein interact with any kinases?
? Now can you answer the question as to whether Arf family members may be regulated by phosphorylation?

➤ **Return to www.ebi.ac.uk/services.**

This is the end of the short tour of the IntAct database. Perhaps you might like to try it again with a more relevant sequence, such as one you are currently working with. Remember that all this information is at your disposal and much of the data can be downloaded and installed in-house.

Now try and repeat some or all of these searches on the following sequences:

Protein X:

```
MPYLLPGFFCDRVIRERDRRNGEGTVSQPLKFEGQDFVVLKQRCLAQKCLFEDRVFPAGTQALGS
HELSQKAKMKAITWKRPKEICENPRFIIGGANRTDICQGDLGDCWFLAAIACLTLNERLLFRVIP
HDQSFTENYAGIFHFQFWRYGDWVDVVIDDCLPTYNNQLVFTKSNHRNEFWSALLEKAYAKLHGS
YEALKGGNTTEAMEDFTGGVTEFFEIKDAPSDMYKIMRKAIERGSLMGCSIDTIVPVQYETRMAC
GLVKGHAYSVTGLEEALFKGEKVKLVRLRNPWGQVEWNGSWSDGWKDWSFVDKDEKARLQHQVTE
DGEFWMSYDDFVYHFTKLEICNLTADALESDKLQTWTVSVNEGRWVRGCSAGGCRNFPDTFWTNP
QYRLKLLEEDDDPDDSEVICSFLVALMQKNRRKDRKLGANLFTIGFAIYEVPKEMHGNKQHLQKD
FFLYNASKARSKTYINMREVSQRFRLPPSEYVIVPSTYEPHQEGEFILRVFSEKRNLSEEAENTI
SVDRPVPRPGHTDQESEEQQQFRNIFRQIAGDDMEICADELKNVLNTVVNKHKDLKTQGFTLESC
RSMIALMDTDGSGRLNLQEFHHLWKKIKAWQKIFKHYDTDHSGTINSYEMRNAVNDAGFHLNSQL
YDIITMRYADKHMNIDFDSFICCFVRLEGMFRAFHAFDKDGDGIIKLNVLEWLQLTMYA
```

Protein Y:

```
QLEEEVKDLADKKESVAHWEAQITEIIQWVSDEKDARGYLQALASKMTEELEALRNSSLGTRAT
DMPWKMRRFAKLDMSARLELQSALDAEIRAKQAIQEELNKVKASNIITECKLKDSEKKNLELLS
EIEQLIKDTEELRSEKGIEHQDSQHSFLAFLNTPTDALDQFETVDSTPLSVHTPTLRKKGCPGS
TGFPPKRKTHQFFVKSFTTPTKCHQCTSLMVGLIRQGCSCEVCGFSCHITCVNKAPTTCPVPPE
QTKGPLGIDPQKGIGTAYEGHVRIPKPAGVKKGWQRALAIVCDFKLFLYDIAEGKASQPSVVIS
QVIDMRDEEFSVSSVLASDVIHASRKDIPCIFRVTASQLSASNNKCSILMLADTENEKNKWVGV
LSELHKILKKNKFRDRSVYVPKEAYDSTLPLIKTTQAAAIIDHERIALGNEEGLFVVHVTKDEI
IRVGDNKKIHQIELIPNDQLVAVISGRNRHVRLFPMSALDGRETDFYKLSETKGCQTVTSGKVR
```

```
HGALTCLCVAMKRQVLCYELFQSKTRHRKFKEIQVPYNVQWMAIFSEQLCVGFQSGFLRYPLNG
EGNPYSMLHSNDHTLSFIAHQPMDAICAVEISSKEYLLCFNSIGIYTDCQGRRSRQQELMWPAN
PSSCCYNAPYLSVYSENAVDIFDVNSMEWIQTLPLKKVRPLNNEGSLNLLGLETIRLIYFKNKM
AEGDELVVPETSDNSRKQMVRNINNKRRYSFRVPEEERMQQRREMLRDPEMRNKLISNPTNFNH
IAHMGPGDGIQILKDLPMPGFPYPSPHHHSGLISSPINFEHIYHMTVNSAEKFLSPDSINPEYS
PSLRSVPGTPSFMTLRNPRPQESRTVFSGSVSIPSITKSRPEPGRSMSASSGLSARSSAQNGSA
LKREFSGGSYSAKRQPMPSPSEGSLSSGGMDQGSDAPARDFDKEDSDSPRHSTASNSSNLSSPP
SPVSPRKTKSLSLESTDRGSWDP
```

Protein Z:

```
MLTDSGGGGTSFEEDLDSVAPRSAPAGASEPPPPGGVGLGIRTVRLFGEAGPASGVGSSGGGGSGS
GTGGGDAALDFKLAAAVLRTGGGGGASGSDEDEVSEVESFILDQEDLDNPVLKTTSEIFLSSTAEG
ADLRTVDPETQARLEALLEAAGIGKLSTADGKAFADPEVLRRLTSSVSCALDEAAAALTRMKAENS
HNAGQVDTRSLAEACSDGDVNAVRKLLDEGRSVNEHTEEGESLLCLACSAGYYELAQVLLAMHANV
EDRGNKGDITPLMAASSGGYLDIVKLLLLHDADVNSQSATGNTALTYACAGGFVDIVKVLLNEGAN
IEDHNENGHTPLMEAASAGHVEVARVLLDHGAGINTHSNEFKESALTLACYKGHLDMVRFLLEAGA
DQEHKTDEMHTALMEACMDGHVEVARLLLDSGAQVNMPADSFESPLTLAACGGHVELAALLIERGA
NLEEVNDEGYTPLMEAAREGHEEMVALLLAQGANINAQTEETQETALTLACCGGFSEVADFLIKAG
ADIELGCSTPLMEASQEGHLELVKYLLASGANVHATTATGDTALTYACENGHTDVADVLLQAGADL
EHESEGGRTPLMKAARAGHLCTVQFLISKGANVNRATANNDHTVVSLACAGGHLAVVELLLAHGAD
PTHRLKDGSTMLIEAAKGGHTNVVSYLLDYPNNVLSVPTTDVSQLPPPSQDQSQVPRVPTHTLAMV
VPPPQEPDRTSQENSPALLGVQKGTSKQKSSSLQVADQDLLPSFHPYQPLECIVEETEGKLNELGQR
ISAIEKAQLKSLELIQGEPLNKDKIEELKKNREEQVQKKKKILKELQKVERQLQMKTQQQFTKEYL
ETKGQKDTVSLHQQCSHRGVFPEGEGDGSLPEDHFSELPQVDTILFKDNDVDDEQQSPPSAEQIDF
VPVQPLSSPQCNFSSDLGSNGTNSLELQKVSGNQQIVGQPQIAITGHDQGLLVQEPDGLMVATPAQ
TLTDTLDDLIAAVSTRVPTGSNSSSQTTECLTPESCSQTTSNVASQSMPPVYPSVDIDAHTESNHD
TALTLACAGGHEELVSVLIARDAKIEHRDKKGFTPLILAATAGHVGVVEILLDKGGDIEAQSERTK
DTPLSLACSGGRQEVVDLLLLARGANKEHRNVSDYTPLSLAASGGYVNIIKILLNAGAEINSRTGSK
LGISPLMLAAMNGHVPAVKLLLLDMGSDINAQIETNRRNTALTLACFQGRAEVVSLLLDRKANVEHRA
KTGLTPLMEAASGGYAEVGRVLLDKGADVNAPPVPSSRDTALTIAADKGHYKFCELLIHRGAHIDV
RNKKGNTPLWLASNGGHFDVVQLLVQAGADVDAADNRKITPLMSAFRKGHVKVVQYLVKEVNQFPS
DIECMRYIATITDKELLKKCHQCVETIVKAKDQQAAEANKNASILLKELDLEKSREESRKQALAAK
REKRKEKRKKKKEEQKRKQEEDEENKPKENSELPEDEDEEENDEDVEQEVPIEPPSATTTTTIGIS
ATSATFTNVFGKKRANVVTTPSTNRKNKKNKTKETPPTAHLILPEQHMSLAQQKADKNKINGEPRG
GGAGGNSDSDNLDSTDCNSESSSGGKSQELNFVMDVNSSKYPSLLLHSQEEKTSTATSKTQTRLEG
EVTPNSLSTSYKTVSLPLSSPNIKLNLTSPKRGQKREEGWKEVVRRSKKLSVPASVVSRIMGRGGC
NITAIQDVTGAHIDVDKQKDKNGERMITIRGGTESTRYAVQLINALIQDPAKELEDLIPKNHIRTP
ASTKSIHANFSSGVGTTAASSKNAFPLGAPTLVTSQATTLSTFQPANKLNKNVPTNVRSSFPVSLP
LAYPHPHFALLAAQTMQQIRHPRLPMAQFGGTFSPSPNTWGPFPVRPVNPGNTNSSPKHNNTSRLP
NQNGTVLPSESAGLATASCPITVSSVVAASQQLCVTNTRTPSSVRKQLFACVPKTSPPATVISSVT
STCSSLPSVSSAPITSGQAPTTFLPASTSQAQLSSQKMESFSAVPPTKEKVSTQDQPMANLCTPSS
TANSCSSSASNTPGAPETHPSSSPTPTSSNTQEEAQPSSVSDLSPMSMPFASNSEPAPLTLTSPRM
VAADNQDTSNLPQLAVPAPRVSHRMQPRGSFYSMVPNATIHQDPQSIFVTNPVTLTPPQGPPAAVQ
LSSAVNIMNGSQMHINPANKSLPPTFGPATLFNHFSSLFDSSQVPANQGWGDGPLSSRVATDASFT
VQSAFLGNSVLGHLENMHPDNSKAPGFRPPSQRVSTSPVGLPSIDPSGSSPSSSSAPLASFSGIPG
TRVFLQGPAPVGTPSFNRQHFSPHPWTSASNSSTSAPPTLGQPKGVSASQDRKIPPPIGTERLARI
RQGGSVAQAPAGTSFVAPVGHSGIWSFGVNAVSEGLSGWSQSVMGNHPMHQQLSDPSTFSQHQPME
RDDSGMVAPSNIFHQPMASGFVDFSKGLPISMYGGTIIPSHPQLADVPGGPLFNGLHNPDPAWNPM
IKVIQNSTECTDAQQASLLPSVPALKGEIPSPQLTRPKKRIGRPMVASPNQRHQDHLRPKVPAGVQ
ELTHCPDTPLLPPSDSRGHNSSNSPSLQAGGAEGAGDRGRDTR
```