Introduction to Bioinformatics

EMBO Practical Course on Computational analysis of protein-protein interactions: From sequences to networks

Mon 28th September – Sat 3rd October 2015 TGAC, Norwich, UK

Monday 28th September 2015

Aidan Budd
EMBL Heidelberg

Lots of different bioinformatics resources

- 1500+ listed in NAR database collection: www.oxfordjournals.org/our_journals/nar/database/cap/
- most researcher use many tools, often meet new ones
  - e.g. you'll meet many during this course
- thus, ability to effectively and efficiently learn new tools, and critically interpret their results, is very useful

Learning/critically assessing new tools

Things that can help us learn new tools:

- recognising common features of bioinformatics tools e.g.
  - unique record identifiers
  - cross-references to other resources
  - ontologies
- not trying to understand/learn all features of a tool
  - focus instead on those features most relevant to your questions

Learning/critically assessing new tools

Things that can help us learn new tools:

- understanding link between the tool, and its reported results, to experimental data
  - accuracy of tool's results depends on the quality of this data
- effective searching for relevant information/help
  - search function in your web browser
  - Internet searching e.g. "uniprot phosphorylation" often more effective than using tool's own search engine
  - writing to mailing lists/developers/maintainers with questions
- domain-specific knowledge

Using UniProt to illustrate these ideas

why UniProt?

extremely widely used, linked to, referenced to, protein bioinformatics resource

we'll explore it by considering the question:

Can I trust the information/results I get from this tool?

(a question we hear quite often from biologists)

UniProt

A UniProt record (http://www.uniprot.org/) describes the protein(s) associated with the a single gene in a single taxonomic group (usually "species")

e.g. http://www.uniprot.org/uniprot/P07550 is the record for:
human Beta-2 adrenergic receptor
the unique identifier of this record in the UniProt database is P07550

Records in SwissProt section of UniProt are manually annotated and reviewed

Top page of UniProt online manual gives the many different types of information that can be found in records http://www.uniprot.org/manual/

Let's try judging the accuracy of different information in a UniProt record

UniProt

Information in a record depends on experimental observations of either:

A. protein/gene/entity described in that record - sometimes referred to as "**direct assay**"

B. other protein(s)/gene(s)/entiti(es) somehow "similar" to the protein(s)/ gene(s)/entiti(es) described in the record - often referred to as a "**prediction**"

Information reported by many (most? all?) bioinformatics tools depends on either "**direct assay**" or "**prediction**"

direct assay examples

- atom positions in a 3D structural model viewed in e.g. Chimera
  - e.g. X-ray crystalography used to build a model describing locations of atoms of the molecule
- amino acid sequence described in a UniProt record
  - e.g. mass spectrometry used to measure the sequence of the protein

UniProt

Information in a record depends on experimental observations of either:

A. protein/gene/entity described in that record - sometimes referred to as "**direct assay**"

B. other protein(s)/gene(s)/entiti(es) somehow "similar" to the protein(s)/ gene(s)/entiti(es) described in the record - often referred to as a "**prediction**"

Information reported by many (most? all?) bioinformatics tools depends on either "**direct assay**" or "**prediction**"

prediction examples

- protein families reported by Pfam in a protein sequence
  - query sequence examined for similarity to sequences known to contain a given family
- intrinsically-unstructured regions reported by IUPred in a protein sequence
  - query sequence examined for similarity to sequences known to be IUP

UniProt

Information in a record depends on experimental observations of either:

A. protein/gene/entity described in that record - sometimes referred to as "**direct assay**"

B. other protein(s)/gene(s)/entiti(es) somehow "similar" to the protein(s)/ gene(s)/entiti(es) described in the record - often referred to as a "**prediction**"

Information reported by many (most? all?) bioinformatics tools depends on either "**direct assay**" or "**prediction**"

of course it's more complicated than that…

- human genomic sequence described e.g. in Ensembl - it depends on direct assay of genomic DNA to measure the sequence AND on clustering of many such sequences to assemble a model of the genomic sequence

UniProt

Information in a record depends on experimental observations of either:

A. protein/gene/entity described in that record - sometimes referred to as "**direct assay**"

B. other protein(s)/gene(s)/entiti(es) somehow "similar" to the protein(s)/ gene(s)/entiti(es) described in the record - often referred to as a "**prediction**"

Information reported by many (most? all?) bioinformatics tools depends on either "**direct assay**" or "**prediction**"

How we address the "trustworthiness" of results/output differs depending on whether it is based on "**direct assay**" or "**prediction**"

So it's important we can identify which of these a given result/output is

"Direct Assay" information

E.g.s of "**direct assay**" information in http://www.uniprot.org/uniprot/P07550

Gene involved in "activation of transmembrane receptor protein tyrosine kinase activity" (Function, GO - Biological Process; links to PubMed article)

Link to description of binary interaction with SRC (Interaction, Binary interactions; links to IntAct record)

Location of transmembrane helices (Structure, Show more details, Helix, link to PDB record)

Note the links to descriptions (e.g. articles) of direct assays in other resources

Database cross-references/x-refs make exploring reported data about the protein much easier than if they weren't there

"Direct Assay" information

E.g.s of "**direct assay**" information in http://www.uniprot.org/uniprot/P07550

Gene involved in "activation of transmembrane receptor protein tyrosine kinase activity" (Function, GO - Biological Process; links to PubMed article)

Link to description of binary interaction with SRC (Interaction, Binary interactions; links to IntAct record)

Location of transmembrane helices (Structure, Show more details, Helix, link to PDB record)

Note the use of structured descriptions of information (ontologies/controlled vocabularies) to find other entities assigned the same features

Ontologies can help organising and integrating "messy" biological data, particularly for automated analysis

"Direct Assay" information

E.g.s of "**direct assay**" information in http://www.uniprot.org/uniprot/P07550

Gene involved in "activation of transmembrane receptor protein tyrosine kinase activity" (Function, GO - Biological Process; links to PubMed article)

Link to description of binary interaction with SRC (Interaction, Binary interactions; links to IntAct record)

Location of transmembrane helices (Structure, Show more details, Helix, link to PDB record)

Note how we use the browser's "search" facility to navigate the page

Appropriate use of browser "search" can make it much easier to find the specific information we need from long webpages

Domain-specific knowledge of terms/entities relevant to our topic of interest make this easier i.e. knowing that IntAct could be a source of relevant information about interactions makes these links easier to find

"Direct Assay" information

E.g.s of "**direct assay**" information in http://www.uniprot.org/uniprot/P07550

Gene involved in "activation of transmembrane receptor protein tyrosine kinase activity" (Ontologies, GO, Biological Process)

Link to description of "binary interaction" with SRC in IntAct

Location of trans-membrane helices (Sequence annotation, Regions)

Note how we use the browser's "search" facility to navigate the page

Appropriate use of browser "search" can make it much easier to find the specific information we need from long webpages

Domain-specific knowledge of terms/entities relevant to our topic of interest make this easier i.e. knowing that IntAct could be a source of relevant information about interactions makes these links easier to find

"Predicted" information

E.g.s of "**predicted**" information in http://www.uniprot.org/uniprot/P07550

Several phosphorylation sites

Two glycosylation sites

(I assume this is via sequence similarity to reported sites in other proteins)

Lack of a link to the data on which these statements are based makes it very difficult to trace how/why these assertions are made

We notice again:

Database cross-references/x-refs make exploring reported data about the protein much easier than if they weren't there

"Predicted" information

E.g.s of "**predicted**" information in http://www.uniprot.org/uniprot/P07550

Several phosphorylation sites

Two glycosylation sites

(I assume this is via sequence similarity to reported sites in other proteins)

Almost all information in this record is either by direct assay, or it's not clear whether direct assay or prediction

Contrast with the UniProt record for the gorilla version of the gene, which has an **identical sequence to the human protein**

http://www.uniprot.org/uniprot/G3QRR6

(Any suggestions on how we could check that their sequences are identical?)

"Predicted" information

Contrast with the UniProt record for the gorilla version of the gene, which has an identical sequence to the human protein

http://www.uniprot.org/uniprot/G3QRR6

By following the links to predicted features of this protein, we can in most cases determine which (sets of, regions of) proteins this one is similar to, and how that similarity is being assessed

the GO terms "Inferred from electronic annotation. Source: Ensembl"

the "keywords"

Possible reasons why "direct assay" annotation might be wrong

We judge how well we trust this information by considering **reasons why it might be inaccurate**

Considering examples of this from http://www.uniprot.org/uniprot/P07550 e.g.

Gene involved in "activation of transmembrane receptor protein tyrosine kinase activity" (Ontologies, GO, Biological Process)

Link to description of "binary interaction" with SRC in IntAct

Write down, without discussing, a list of reasons why this information might be incorrect. For example:

"a human (or automatic) annotator assigned this information to the wrong protein"

I'll tell you when to stop this, and to then compare your list with your neighbour, and thus build a consensus list you both agree with

Then we'll discuss them all together

Possible reasons why "direct assay" annotation might be wrong

- experiments were fraudulent
- experiments were carried out improperly i.e. the experiments were repeated "correctly" you'd get a different, more accurate result
- experiments were correctly carried out but inappropriately functionally interpreted by experimenters and/or annotators e.g. transient over-expression results used to draw conclusions about localisation. Put differently: the conclusion drawn from the correctly-conducted experiment is inappropriate
- experimental results mistakenly assigned to the wrong entity
- the assay has low accuracy - "the assay may be direct, but the assay sucked e.g. has lots of false positives and/or false negatives"

How could you go about checking for potential wrong "direct assay" information?

What could/should you do to avoid serious consequences of such "wrong" data?

Again, think about this by yourself, then compare notes with your neighbour, and then we'll discuss it all together

For example: **look for multiple, somewhat independent, experiments which reach the same conclusion**

How could you go about checking for potential wrong "direct assay" information?

before you spend ages trying to determine if something is correct, remember **you can't check everything!**

Effort invested in checking should depend on how crucial its accuracy is for your work!

- look for multiple, somewhat independent, experiments from which you reach the same conclusion e.g.:
  - experiments done in different labs
  - different experimental methods used
  - different source of reagents
- learn which methods are commonly used inappropriately e.g. transient over-expression to indicate localisation
- you can sometimes spot fraud in figures e.g. by noticing duplicated bands, obvious photoshopping of gels etc.
- carefully read primary paper/evidence critically and carefully to spot potential problems

How could you go about checking for potential wrong "direct assay" information?

before you spend ages trying to determine if something is correct, remember **you can't check everything!**

Effort invested in checking should depend on how crucial its accuracy is for your work!

- repeat the experiment yourself in the lab
- carry out a yourself in the lab a different analysis/experimental approach to look for additional evidence to support the conclusion
- ask (several) expert(s) for their opinions - check papers that cite a given result, looking for any that contradict it
- community annotation to highlight potentially wrong information e.g. https://pubpeer.com/

Possible reasons why "predictions" might be wrong

How do we (ideally) predict function/structure?

1. Collect a set of "true positives" (TPs) i.e. features that you believe have the property you want to predict (e.g. residues that you believe are phosphorylated in some cellular contexts)

2. Collect a set of "true negatives" (TNs) i.e. features you believe do not have property you want to predict (e.g. residues you believe are not phosphorylated in similar cellular contexts - can be very tricky to find)

3. Choose a way of scoring/classifying unknown features (e.g. amino acid sequences) according to how similar they are to features in these two sets.

4. If a query feature is much more similar to TPs than TNs by this similarity measure, then you predict the query feature is likely to be also a positive

Possible reasons why "predictions" might be wrong

To judge how much to trust this information, we need to think about the reasons why it might be inaccurate

Considering examples of this from http://www.uniprot.org/uniprot/P07550 e.g.

Several phosphorylation sites

Two glycosylation sites

Again:

Make, for yourself, a list of reasons why this might be the case - for example **there are only small differences between TPs and TNs on average using a given similarity measure**

Then compare your list with those of your neighbour - and build a consensus list

Then we'll discuss them together

Possible reasons why "predictions" might be wrong

1. training sets contain wrongly-assigned features (e.g. some of the sites listed as "phosphorylated" aren't, some that are listed as "nonphosphorylated" are) could be due to fraud, badly-carried out experiments, mis-interpreted experiments (by experimenters and/or curators):

    In particular it can be difficult/impossible to identify genuine TNs

2. information used for prediction does not contain all that is needed to distinguish P from N:

    e.g. a particular activity is only relevant in a particular sub-cellular localisation; without that data, you're going to have a bad predictor e.g. RGD motifs only interact with integrin extracellularly

3. there are only small differences between TPs and TNs on average using a given similarity measure (related to 2 - both are due to problems with the similarity measure)

How could you go about checking for potential wrong predictions?

What could/should you do to avoid serious consequences of such "wrong" data?

Again, think about this by yourself, then compare notes with your neighbour, and then we'll discuss it all together

For example: **check whether several non-identical analyses giving the same/similar answers - this increases our confidence in the prediction**

How could you go about checking for potential wrong predictions?

again...

before you spend ages trying to determine if something is correct, remember **you can't check everything!**

Effort invested in checking should depend on how crucial its accuracy is for your work!

Text

- check whether several non-identical analyses giving the same/similar answers - this increases our confidence in the prediction
- critically examine the training sets, and similarity measures, used to build the tool (typically by reading the publication describing the tool) to identify possible inaccuracies - if two results give contradictory predictions, use this reading to try and decide which is more likely to be correct

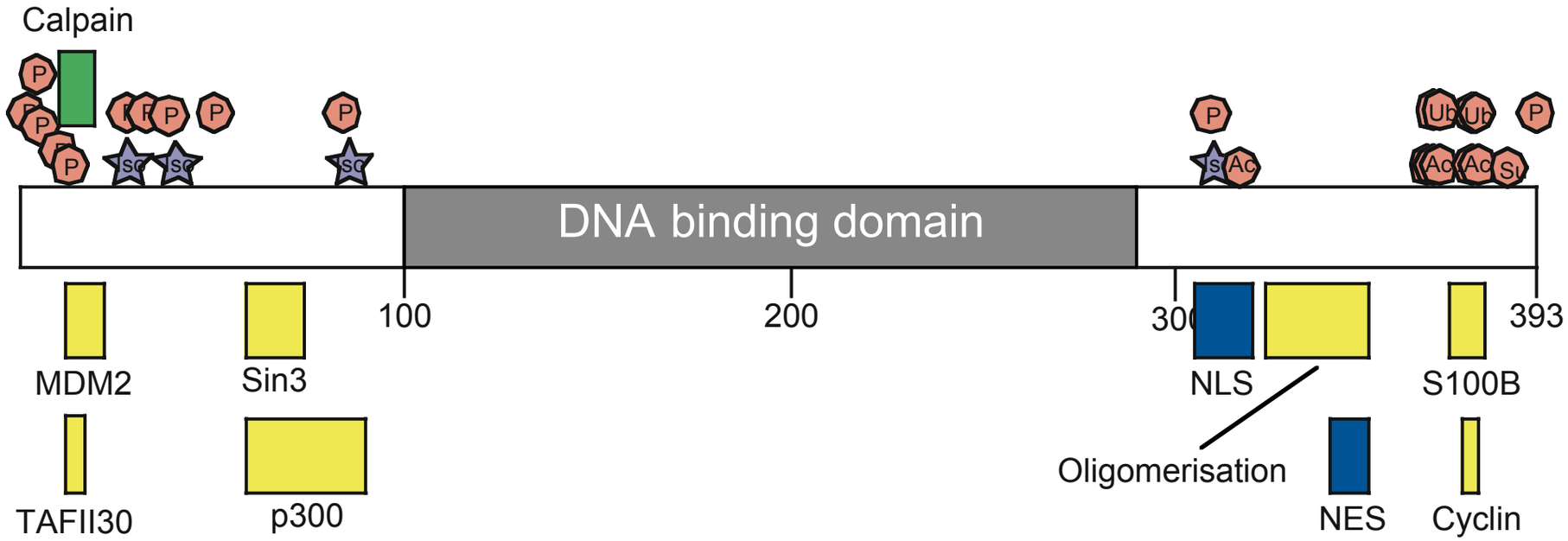
Protein Modules

Aidan Budd, EMBL Heidelberg

Protein Modules

Protein modules:

Protein regions associated with specific functions/activities of the protein that, if isolated from the rest of the protein, retain similar activity/function



Functional modules mapped on P53

Morten Mattingsdal & Paal Puntervoll, Bergen

Aidan Budd, EMBL Heidelberg

Protein Modules

Many modules are found in many different proteins, where they are associated with similar activities

e.g. PFAM shows some of the proteins predicted to contain a kinase domain

Protein Modules

Many modules are found in many different proteins, where they are associated with similar activities

e.g. ELM instances of LIG_SH3_3 SH3-domain binding motif

16 Instances for LIG_SH3_3

(click table headers for sorting; Notes column: =Number of Switches, =Number of Interactions)

Protein Name	Gene Name	Start	End	Subsequence	Logic	#Ev.	Organism	Notes
P73_HUMAN	TP73	407	413	HLQPPSYGPVLSPMNKVHGG	TP	1	Homo sapiens (Human)	
WASF1_MOUSE	Wasf1	319	325	AGRTPVFVSPTPPPPPPLP	TP	1	Mus musculus (House mouse)	
LAS17_YEAST	LAS17	216	222	TASAPTTPAPALPPASPEVR	TP	2	Saccharomyces cerevisiae (Baker"s yeast)	
SOS1_MOUSE	Sos1	1135	1141	GTDEVPVPPPVPPRRRPESA	TP	1	Mus musculus (House mouse)	PDB 2GBQ
CY24A_TURTR	CYBA	154	160	IKQPPSNPPPRPPAEARKKP	TP	1	Tursiops truncatus (Bottlenosed dolphin)	
CY24A_TURTR	CYBA	149	155	QVGGTIKQPPSNPPPRPPAE	TP	1	Tursiops truncatus (Bottlenosed dolphin)	
CY24A_TURTR	CYBA	153	159	TIKQPPSNPPPRPPAEARKK	TP	1	Tursiops truncatus (Bottlenosed dolphin)	
ORF3_HEVBU	ORF3	102	108	RPSAPPLPHVVDLPQLGPRR	TP	3	Hepatitis E virus (strain Burma)	
TIP_SHV24		135	141	DPGMPKPTLPPRPANLGASQ	TP	1	Herpesvirus saimiri (strain 484-77)	
NS1_I72A2	NS	161	167	AIVGEISPLPSFPGHTIEDV	TP	4	Influenza A virus (A/Udorn/307/1972(H3N2))	
CD2_HUMAN	CD2	294	300	HRSQAPSHRPPPPGHRVQHQ	TP	2	Homo sapiens (Human)	1
DAG1_HUMAN	DAG1	888	894	KGSRPKNMTPYRSPPPYVPP	TP	2	Homo sapiens (Human)	2
TIP_SHV24		132	138	ESWDPGMPKPTLPPRPANLG	TP	1	Herpesvirus saimiri (strain 484-77)	1
ORF3_HEVBU	ORF3	93	99	HSAPLGVTRPSAPPLPHVVD	TP	3	Hepatitis E virus (strain Burma)	2
BIN1_HUMAN	BIN1	308	314	SQLRKGPPVPPPPKHTPSKE	TP	6	Homo sapiens (Human)	1
TAU_HUMAN	MAPT	213	219	GSRSRTPSLPTPPTREPKKV	TP	2	Homo sapiens (Human)	1 1

Protein Modules

Looking for a source of hypotheses for the function of a protein?

e.g. because you want to choose some appropriate bench science experiments to carry out on this protein?

Using bioinformatics tools to predict the modular architecture of your protein can be a great source of such hypotheses