Introduction to Bioinformatics

EMBL-Australia Masterclass on Protein Sequence Analysis

Sydney, Australia
Monday 21st October 2013

Aidan Budd
EMBL Heidelberg

Introduction to Bioinformatics

Becoming better, in general, at using bioinformatics resources, is often something trainees on courses like this are looking for

Or, another way of casting it, is wanting to learn to think more like an experienced bioinformatician

In this hour, I've chosen to focus on one specific aspect of this, one that I hear again and again as being important for trainees i.e.

becoming better able and more confident judge the likely accuracy/utility/"trustworthiness"/reliability of a tool/result

Would this be interesting for some of you? [I'm hoping the answer is yes!]

If very few of you are looking for that, we can instead do some small-group work to focus on your specific needs/interests

Beginning with UniProt...

Each UniProt (http://www.uniprot.org/) record is associated with the protein (s - i.e. includes information about different splice forms) from a single gene in a single organism

E.g. http://www.uniprot.org/uniprot/P07550 describes the record for the human Beta-2 adrenergic receptor

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/

Getting better/more comfortable at judging the likely accuracy/utility/re;liability/"trustworthiness" of a tool/result is one of our aims in this session

So let's begin with this extremely commonly-used tool - i.e. trying to judge the likely accuracy of different kinds of information in a UniProt record

Beginning with UniProt...

Information in a protein's UniProt record is based on experimental results/observations obtained from working with either:

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

B.other protein(s)/gene(s) that are somehow identified as similar to the protein described in that record - often referred to as a "prediction"

The information/results obtained from/reported by many (most? all?) bioinformatics tools can be similarly classified

The way we address how "trustworthy" such information is differs considerably depending on whether it is based on "direct assay" or "prediction"

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

"Direct Assay" information

Some examples of "direct assay" information from 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)

Notice that here we can link through to the published descriptions of the relevant direct assay results

"Predicted" information

Some examples of "predicted" information from
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 everything else in the 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

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.

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 fraudexperiments were carried out improperly i.e. the results are just wrong i.e. if repeated "correctly' you'd get different resultsexperiments were correctly carried out but inappropriately interpreted in terms of function by experimenters and/or annotators e.g. transient over-expression to draw conclusions about localisation i.e. the conclusion drawn from the correctly-conducted experiment was wrong
- 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
- 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) be vary waryyou 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?

- repeat the experiment yourself in the lab
- carry out a yourself in the lab different analysis to try and find additional evidence to support the conclusion
- ask (several) expert(s) for their opinions - check papers that cite a given result, any that contradict it
- community annotation to highlight potentially wrong things

Possible reasons why "predictions" might be wrong

How do we 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

Aidan Budd, EMBL Heidelberg

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)

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

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!
- 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