ProteinProphet – Tutorial and exercises
Alexey Nesvizhskii

1. Yeast Orbitrap data

We will use the same Yeast LTQ-Orbitrap dataset to evaluate the performance of ProteinProphet.

All MS/MS spectra were searched using X-Tandem (with k-score plug-in) against a Yeast database appended with an equal number of decoy sequences and common contaminants. In this dataset all decoy sequences have names that start with REV0_ or REV1_. All decoy proteins are incorrect identifications.

The search results were analyzed using PeptideProphet. You do not need to run ProteinProphet, we have done it for you already.

Using the File Browser in Petunia, navigate down to the class/ProteinProphet/xtandem-k/semitryptic/ folder, and open the interact.prot.shtml file by clicking on the View link. This file is the main ProteinProphet output file.

Do the following:

i) It is always good to check first that PeptideProphet worked fine and that computed peptide probabilities are likely to be accurate. If there was a problem running PeptideProphet and peptide probabilities are not accurate, then ProteinProphet results are not going to be accurate as well. To do this, first follow **any** peptide link. A new window should open showing more information about that particular peptide identification. Click on the peptide probability link, and it should bring you to the PeptideProphet output page. Look at the discriminant score distributions learned by the model. Do they look Ok? Look at the model output below and check that other parameters learned by the model are reasonable (e.g., the distribution of NTT parameter among correct and incorrect identifications).

ii) Familiarize yourself with ProteinProphet output.
Find entry #**344a**, YEL034W.
- What is the probability assigned to this protein?
- How many different peptide sequences are identified that correspond to this protein?
- What is the number of "unique peptides", and how is it defined in ProteinProphet?
- Are there any peptides identified multiple times?
- Are there any *'shared peptides'*, i.e. peptides present not only in YEL034W but also in some other protein(s)?

Look at the peptide 2_NGFVVIK.
- How many siblings does it have?
- Compute the NSP value for that peptide by summing the probabilities of its sibling peptides. Does this computed NSP value agree with the NSP number shown for that peptide?
- Does adjustment for peptide grouping information (NSP) increase or decrease the probability that this particular peptide identification is correct?

iii) *Extra questions for those who like math and statistics*: Find entry #**474**. This one is a single-hit protein identification (and an incorrect one). The initial peptide probability computed by PeptideProphet was somewhat high, **0.9761**. However, ProteinProphet penalized this peptide identification (and, therefore, reduced the probability of the protein identification) for being a single hit. As a result, the adjusted peptide probability is only **0.5006**. Please repeat the calculations yourself. Follow the NSP link and find the values of p(NSP|+) and p(NSP|-) for the corresponding bin (NSP value 0, NSP bin 0). Plug the numbers in the expression (5) of the Anal. Chem. (2003) paper and see how the initial probability **0.98** gets reduced to **~0.50**.

iv) Check the accuracy of protein probabilities computed by ProteinProphet. This is a small dataset, and we do not really know for sure what identifications are correct. So this is just a simple estimate.

1. Consider all protein identifications in the probability range between 0.65 and 0.75. Count the total number of proteins in that range (N), and the number of decoys among them (N_d, names start with REV0_ or REV1_).
2. Estimate the number of correct proteins in the 0.65-0.75 range by assuming that the number of incorrect Yeast protein identifications in that range is equal to the number of decoy protein identifications, $N_c = N - N_d - N_d = N - 2*N_d$
3. Calculate the ratio of the estimated number of correct identifications to the total number of identifications N_c/N in this probability range. Is this ratio close to the expected value of ~ 0.7?

v) As discussed in the lecture, adjustment of peptide probabilities to account for peptide grouping information (NSP) makes peptide probabilities (and, therefore, protein probabilities) more accurate. Check if this is the case. Open the file *interact-nonsp.prot.shtml*, it is in the same directory as interact.prot.shtml. This file has protein probabilities computed without adjustment for NSP. Again, look at the same probability range (0.65-0.75). Are computed probabilities accurate? (no need to count proteins, the answer should be obvious).

vi) What are the ProteinProphet predicted sensitivity and false discovery rate (FDR) for this dataset when filtered using minimum protein probability threshold of 0.7? To find that, follow the Sensitivity/Error Info link at the top of the file (note FDR=err). Compare the predicted FDR with that estimated based on decoy counts, $FDR = 2N_d / N$, where N is the total number of proteins with probability above 0.7, and N_d is the number of decoys among them.

vii) Think about different sources of false positives. What are we NOT taking into account when performing target-decoy based FDR estimates, or when using ProteinProphet computed probabilities? How does it affect the error rate estimates?

Consider entry #**387**, protein YNL014W.
>HEF3 Translational elongation factor EF-3; paralog of YEF3 and member of the ABC superfamily.
This protein is identified with a very high probability, **0.9991**. However, this identification is likely to be a false positive. Investigate this case. On what peptide is this identification based? Find an alternative explanation that makes this identification questionable. Hint: this is a high mass accuracy data (LTQ-Orbitrap), which can help.

viii) What is the advantage of generating data on high mass accuracy MS instruments with respect to the source of false positive protein identifications discussed in the previous question? How could you modify the database search parameters to lessen this problem (although not eliminate completely).

2. Human Raft Dataset searches against the Human IPI database

This is a subset of a much larger dataset from a human raft protein profiling experiment. This dataset demonstrates the complexity often encountered in proteomics experiments on higher eukaryote organisms. The purpose of this exercise is to get familiar with the difficulty of inferring what proteins are present in the sample given the list of identified peptides. To view the results, use Petunia to navigate down to the **C:\inetpub\wwwroot\ISB\data\class\ProteinProphet\RAFT\IPI** directory, and open the file **interact2.prot.shtml** using View link (or run yourself, data are in **C:\inetpub\wwwroot\ISB\data\class\ProteinProphet\RAFT\IPI\data***).*

Go through the following examples:

i) Indistinguishable proteins.
Find entry #**77**, IPI00026185 IPI00218782
This is a typical example of multiple proteins that cannot be distinguished on the basis of identified peptides. In this case, the two proteins are different isoforms of the F-actin capping protein beta subunit, SW:P47756-1 and SW:P47756-2. You can follow the Ensembl links (and from the Ensembl page Description sections, Swissprot links) to learn more about these proteins. What can you conclude about the presence of the isoforms in the sample?

For the most curious: spend some time playing with this example. Check the sequences of these two proteins. Are they significantly different? To do that, go to Swiss-Prot (the easiest way since this alternative splicing even is annotated in Swiss-Prot), or cut and paste protein sequences and align them using a sequence alignment program (e.g., utility bl2seq that can be found at http://www.ncbi.nlm.nih.gov/blast). In what situation would it be possible to determine which of the two proteins (or both) is actually present in the sample? (hint: are there tryptic peptides in these proteins that, if identified, would discriminate between the two isoforms?)

ii) Subset proteins
Find entry #**178a**, IPI00027500
This protein (Rho A, SW:P06749) is a member of a family of Rho proteins. Two of its peptides, IGAFGYMECSAK and modified form, are unique to this protein (marked with an asterisk). The other peptides are shared, i.e., they are also present in another protein from the same protein family. What is the name of the other protein that also contains these peptides? What probability did ProteinProphet assign to it? What can be concluded about the presence of that other protein in the sample?

iii) Differentiable proteins
Find entry #**167**
This is another interesting example. There are several members from the same protein family that are grouped together. For example, consider entry #**167e,** IPI00023138. This protein (Ras-related C3 botulinum toxin substrate 3) is identified by one unique peptide, HHCPHTPILLVGTK, and several

shared peptides. Some of the other peptides are shared between this protein and a different isoform (e.g. entry #**167b**: IPI00010270, Ras-related ... substrate 2). However, the other isoform is identified by several peptides that are unique to it, including HHCPSTPIILVGTK (note a two amino acid difference compared to the peptide that is unique to the first isoform). Thus, even though these proteins share a set of peptides, each of them has at least one unique peptide. What does ProteinProphet conclude about the presence of these proteins in the sample?

iv) Special case: a protein group containing proteins with no distinct peptides

Find entry # **165**, Protein Group **15**

This is an example of a special case where, strictly speaking, the parsimony rule (Occam's razor) cannot be applied. Four protein entries comprise this group with an assigned probability of 1. Is there definitive evidence that any particular group member is present in the sample? Which protein in this group can explain the presence of all peptides observed in the dataset that correspond to proteins from this group ('subsumes' all the others), and is therefore the most likely candidate? Can we be certain what protein(s) are present in the sample?