ISB/SPC Software Proteomics Course
ProteinProphet – Alexey Nesvizhskii

## 1. Continue with the same *HaloICAT* dataset used in the tutorial

iv) Check the accuracy of protein probabilities computed by ProteinProphet. Count the number of correct protein identifications (names start with Chr_ORF or pNRC200_ORF) having probability between 0.4 and 0.6. Count the total number of protein identifications in that probability range. Is the ratio of the number of correct identifications to the total number of identifications close to the expected value of 0.5?

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.xml* using View link; this file is located in the directory *class/ProteinProphet/haloICAT/Prerun*. This file has protein probabilities computed without adjustment for NSP. Again, look at the intermediate probability range (0.4-0.6). Are computed probabilities accurate? (no need to count proteins, the answer should be obvious).

vi) What are the expected sensitivity and false positive identification error rates for this dataset when filtered using minimum protein probability threshold of 0.7? (Hint: follow the "Sensitivity/Error Info" link at the top of the file).

vii) What is the probability of the highest scoring incorrect protein? How many incorrect proteins have probability above 0.9? What is the model estimate for the number of incorrect proteins above that threshold? Is it an accurate estimate?

viii) *Earn bonus points by doing this exercise*! Find entry #109. This one is a single-hit protein identification (and an incorrect one). The initial peptide probability computed by PeptideProphet was somewhat high, 0.83. 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 0.52. 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 0.83 gets reduced to ~0.52.

## 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 *class/ProteinProphet/RAFT/IPI/* folder, and open the file *interact-prot.xml* using View link.

(Please note that this dataset was searched using SEQUEST)

Go through the following examples:

i) Indistinguishable proteins.
Find entry #10, IPI00026185 IPI00218782
This is a typical example of multiple (in this case 2) proteins that cannot be distinguished on the basis of identified peptides. In this case, the two proteins are two isoforms of the F-actin capping protein beta subunit, SW:P47756-1 and SW:P47756-2. You can follow the Ensembl or Swiss-Prot links to learn more about these proteins. In this case, one can only say a product of a particular gene (CAPZP) is present in the sample, but it is not possible to specify what isoform.

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) Differentiable proteins
Find entry #158, IPI00023138
This is another interesting example. 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 (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. One can conclude that both proteins are present in the sample.

iii) Subset proteins
Find entry #19, 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). Investigate the situation. What is the name of the other protein that also contains these peptides? What probability did ProteinProphet assign to it? Explain (hint: Occams's razor: the other protein is a subset, i.e., none of the identified peptides in the dataset corresponding to this protein are unique to it). What can be concluded about the presence of that other protein in the sample?

iv) Subsumable proteins: Special case of "protein groups"
Find entry # 208.
This is a typical example of a 'protein group'. Which protein in this group can explain the presence of all peptides observed in the dataset that correspond to proteins from this group (and, therefore, the most likely candidate)? Can we be certain what protein(s) are present in the sample?