ISB/SPC Software Proteomics Course
ProteinProphet Tutorial – Alexey Nesvizhskii

***HaloICAT* dataset**

HaloICAT is a dataset of MS/MS spectra generated from a complex *Halobacterium* sample. Here we use this dataset to test the accuracy of protein probabilities computed by ProteinProphet. All MS/MS spectra were searched using Mascot against a human protein database appended with a much smaller *Halobacterium* database. The search results were analyzed using PeptideProphet.

i) You do not need to run ProteinProphet, we have done it for you already. Using the File Browser in Petunia (under Tools), navigate down to the *class/ProteinProphet/haloICAT/Prerun/* folder, and open the *interact-prot.xml* file using View link. This file is the main ProteinProphet output file.

Alternatively (time permitting) you can rerun ProteinProphet yourselves starting from PeptideProphet, XPRESS, and ASAPRatio output files that are located in
class/ProteinProphet/haloICAT/Run/

ii) PeptideProphet quality check

It is always good to check first that PeptideProphet worked fine and that the computed peptide probabilities are correct. If there was a problem running PeptideProphet and peptide probabilities are not accurate, then ProteinProphet results are not going to be accurate as well (this does not happen often, but it is always good to do a simple quality control). 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? Follow the 'More Analysis Info' link and check that other parameters learned by the model are reasonable (e.g., the distribution of 'ICAT consistent' peptides and NTT parameter among correct and incorrect identifications).

iii) Familiarize yourself with ProteinProphet output

Note that in this dataset all correct identifications (*Halobacterium* proteins) have names that start with Chr_ORF or pNRC200_ORF; all human proteins are incorrect identifications.
Find entry #8, Chr_ORF0924.

- What is the probability assigned to this protein? _____ (1.00)
- How many unique peptides are identified in the dataset that correspond to this protein? _____ (4)
- Why are there 5 peptides listed? _____ (There are 4 unique sequences, but one of the peptide is identified in both 2+ and 3+ charge state. ProteinProphet treats peptides identified in different charge states, in modified/unmodified form, or in light/heavy labeled form as independent evidence for the presence of the protein in the sample)
- Are there any *'shared peptides'*, i.e. peptides present not only in Chr_ORF0924 but also in some other protein(s)? _____ (No)

Look at the peptide IDCAHAHNLNVIDSAR.

- How many siblings does it have? (4)
- 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?__(3.77, yes)
- Does adjustment for peptide grouping information (NSP) increase or decrease the probability that this particular peptide identification is correct? _____(increase)