Using PeptideProphet to validate Mascot search results
Andrew Keller

In order to do these exercises, open a new browser which by default goes to the page for launching the ISB pipeline. Login as username: guest/password: guest and then click on the 'Analysis Pipeline' Tab Bar at the top of the page. From there, click on the 'Analyze Peptides' tab.

Introduction: control HaloICAT dataset

1. Run XInteract and PeptideProphet on the semi-tryptic constrained search results of the HaloICAT dataset (search conditions requiring at least one peptide terminus to be consistent with cleavage by trypsin), present in **C:\inetpub\wwwroot\ISB\data\class\PeptideProphet\haloICAT**. This dataset includes four Mascot pepXML files. Add all 4 XML files present in the **C:\inetpub\wwwroot\ISB\data\class\PeptideProphet\haloICAT** directory. Make sure the location of the output file is set for **C:\inetpub\wwwroot\ISB\data\class\PeptideProphet\haloICAT**, then push the 'Run XInteract' button at the bottom of the form. Update the page until the command finished message appears. A link to the results (**http://localhost/ISB/data/class/PeptideProphet/haloICAT/interact.shtml**) will appear when the analysis is complete.

Open the results link and click on any probability link. What is the total number of correct results predicted by the model? Do the learned discriminant score distributions among correct (pos) and incorrect (neg) results look reasonable, given the total distributions for the dataset?

Now scroll down the page. What distributions of discriminant score, numbers of tolerable tryptic termini (NTT), numbers of missed enzymatic cleavages (NMC), and mass difference did the model learn for the correct, and for the incorrect, search results? What fraction of the correct and incorrect results did the model learn to be 'icat compatible'

2. Sort the results by probability score.

First format the viewer to show all results sorted by probability. Under the pepXML Viewer Display options, select 'all' rows per page and sort by 'probability' descending. Click the 'UPDATE' link.

Estimate the accuracy of the computed probabilities using a sliding window of 10 spectra (the correct proteins are colored red). Does the fraction of correct proteins in the window correspond with its average computed probability? In the entire 0 to1 range

3. Filter the dataset using standard Mascot score thresholds. For example:

Ionscore > Identityscore
or: Ionscore > Homologyscore
or: Ionscore ≥ 30

Go into the Filtering Options section of the pepXML Viewer and check the appropriate box or enter the desired minimum ionscore threshold, then click the 'UPDATE' button.

How many correct and incorrect peptide assignments result? Hint: The number of incorrect peptide assignments can be determined by counting the blue corresponding proteins, while the number of correct peptide assignments can be determined by subtracting that number from the total number of entries passing the filter (indicated just below the green menu box).

The total number of correct search results in this dataset is 349. Compute the sensitivity (fraction of correct results in dataset that pass filter) and false positive error rate (fraction of results passing filter that are incorrect) resulting from the use of your chosen conventional threshold filter(s). How does this sensitivity compare with that predicted by the model for the same error rate (click on any probability to view the graph of model predicted sensitivity and error rate versus minimum probability threshold)?

Effect of ICAT information on the model

4. Run XInteract and PeptideProphet on the semi-tryptic constrained search results of the HaloICAT dataset present in **C:\inetpub\wwwroot\ISB\data\class\PeptideProphet\haloICAT**, but this timeanalyzed without using ICAT information and renamed 'interact-noicat.xml'. As before, use all 4 XML files present in the **C:\inetpub\wwwroot\ISB\data\class\PeptideProphet\haloICAT** directory. Make sure the location of the output file is set for **C:\inetpub\wwwroot\ISB\data\class\PeptideProphet\haloICAT**. Type 'interact-noicat.xml' in the text box following 'Write output to file', exclude ICAT information by selecting the checkbox before 'Do not use icat cysteine information…' in the PeptideProphet Options section of the form, and then push the 'Run XInteract' button at the bottom of the form. A link to the results (**http://localhost/ISB/data/class/PeptideProphet/haloICAT/interact-noicat.shtml**) will appear when the analysis is complete.

Sort the results by probability and assess the model accuracy using a sliding window of 10 spectra. Compare the predicted number of correct peptide assignments as a function of predicted error rate for the models learned in step 1 and step 5 (click on any probability to access this information). Which analysis yields more correct peptide assignments at any fixed error rate?

5. Pick out two search results in this dataset that were assigned probabilities close to 0.5: one result with an assigned peptide containing a cysteine, and one result with an assigned peptide without cysteine. Note: you can color cysteines red by entering 'C' in the 'hilight peptide text (regex)' text box in the Display Options section of the pepXML viewer and then pushing the 'UPDATE' button. Write down the spectrum names and computed probabilities. Now look for those spectra in the analysis you performed previously that employed icat information (interact.xml). What probability was computed for the result with assigned peptide containing cysteine in that analysis that used icat information? For the result with assigned peptide not containing cysteine? What might explain these observations?

6. Filter the results for those assigned to peptides containing the amino acid cysteine, by typing 'C' in the 'required peptide text (regex allowed)' text box in the Filtering Options section of the pepXML viewer and pressing the 'UPDATE' button. How does this filtering affect the accuracy of the remaining probabilities? Do you see a similar effect when you filter the results of step 1 in the same way?

Mystery dataset: More comparisons of the model with standard score thresholds

7. Run XInteract *without* PeptideProphet on search results of the Mystery dataset present in **C:\inetpub\wwwroot\ISB\data\class\PeptideProphet\Mystery.** Add the single XML file present in the **C:\inetpub\wwwroot\ISB\data\class\PeptideProphet\Mystery** directory. Make sure the location of the output file is set for **C:\inetpub\wwwroot\ISB\data\class\PeptideProphet\Mystery,** and rename the output by typing 'interact-noprob.xml' in the text box following 'Write output to file'. Turn off PeptideProphet by unchecking the box following 'RUN PeptideProphet' in the form, then push the 'Run XInteract' button at the bottom of the page. A link to the results (**http://locahost/ISB/data/class/PeptideProphet/Mystery/interact-noprob.shtml**) will appear when the analysis is complete.

Next, filter the dataset using conventional Mascot score thresholds:

Ionscore > Identityscore

or: Ionscore > Homologyscore

or: Ionscore ≥ 30

How many 'correct' identifications pass the filter?

8. Now run XInteract *with* PeptideProphet on search results of the Mystery dataset present in **C:\inetpub\wwwroot\ISB\data\class\PeptideProphet\Mystery,** including in the output all peptides with probability 0 or greater by setting the 'PeptideProphet probability' to 0 in 'Output File and Filter Options" section of the form. Look at the results in your browser at **http://locahost/ISB/data/class/PeptideProphet/Mystery/interact.shtml.**

How many search results of 1+, 2+, and 3+ precursor ions are predicted to be correct? What may explain these observations? Click on any probability link to view the models learned by PeptideProphet. Next, use Pep3D to view the Mystery RAW data by clicking on the 'Pep 3D' bar of the pepXML Viewer. Using the default settings, click the 'Generate Pep3D' button to overlay CID attempts and search results. What explanations for the results of PeptideProphet remain viable?