Exercises – Spectral Library Searching Using SpectraST

These exercises take you through the process of preparing a spectral library for searching, searching a small dataset against it, and validating the results using PeptideProphet.

Step 1. Preparing the Library for Searching

Point your browser to http://www.peptideatlas.org/NISTDownload.php, and read through the license agreement. Note that new and/or updated spectral libraries will be made available from time to time at this site, so be sure to check back occasionally for updates. To save time, the library files are already downloaded to your machine for you.

Log onto the Petunia interface, and select the **TOOLS** link. Then click on the tab **SpectraST Libraries**. Click **Add Files**, browse to class/SpectraST/lib, and select yeast_consensus.msp.(You can also **View** the file for your amusement – it is in a somewhat human readable text format. Hit the Stop button on your browser to stop loading.)

Now, click on **Create Indexed Library Files** to run SpectraST in the *Create* mode. This will convert the raw library file (in .msp format) to an indexed library file for fast searching. While it is running, you can monitor its progress in the **Output so far** box. (You can click **UPDATE THIS PAGE** periodically if your browser is not updating.)

The entire process should take anywhere from 8-15 minutes. When the command is finished running (**Command Status** panel turns orange), click the "**View results of previous commands**" link in the Command Status panel for the library statistics. How long did it take to run the command? How many spectra are there in the library yeast_consensus.msp? (35,135) How many of them are spectra of +2 peptide ions? (22,286) How many are tryptic peptides? (29,778) How many of them have an acid-cleavable ICAT labeled cysteine? (10,533)

Note that you only have to do this once whenever you download a new raw library file.

Step 2. Searching a test dataset

A yeast dataset is provided for you to test-run the program. This is a generic shotgun proteomics run of a fractionated yeast extract tryptic digest on a ThermoFinnigan LCQ.

Click on the **HOME** link at the top. Then in the **Analysis Pipeline** panel, select **SpectraST** in the drop-down box. Click on the **ANALYSIS PIPELINE** link at the top, and click on the **SpectraST Search** tab. To set up the search, first click **Add Files** in the first box (**1. Specify mzXML Files**), browse to the directory class/SpectraST/data, and select "LQ20060324_s_1X_26.mzXML". In the second box (**2. Specify Library File**), browse to the directory class/SpectrasST/lib, and select yeast_consensus.splib, which is the indexed library file that you created in Step 1. In the fourth box (**4. SpectraST**

Options), click the check-box "Read search options from file: spectrast.params." You are now set to go – scroll down to the bottom of the page and hit **Run SpectraST**. Again, use the **Command Status** box and the **Output so far** box to monitor the progress.

The search should take about 2-6 minutes. In the meantime, a word about the options. SpectraST has a seemingly overwhelming set of options for you to choose from, but the defaults will do just fine in most searches. In the search that you just started, all the options are specified in a parameter file ("spectrast.params"), a feature that saves you from remembering to set your favorite options every time. (To check out the "spectrast.params" file, click on **Tools** link at the top, then click on the **Browse Files** tab, navigate to /class/SpectraST/data, and **View** the file "spectrast.params". The options specified here are basically the defaults. This text file can be modified in any text editor.)

Some more commonly-used options that you may want to know about are: (1) expected cysteine modification, for which you can select CAM (carbamidomethyl), ICAT_cl (acid-cleavable ICAT) and ICAT_uc (uncleavable "old" ICAT) -- this tells SpectraST not to consider candidates with a different cysteine modification; (2) database file, for which you need to select a sequence database (in .fasta format) – this does not affect SpectraST in any way, but that sequence database will be included in the output for downstream processing (e.g., peptide-to-protein inference by ProteinProphet); (3) output format, for which you can choose pepXML, tab-delimited text or fixed-width text; the latter two formats are provided in case you wish to view the output manually or in a spreadsheet. More information about the options can be found in the documentation of the SpectraST source code.

When the search is done, be sure to check the **Output so far** panel to view the search statistics. How many spectra were searched? (2,678) On average, how much time did it take to search one spectrum? (~0.04 second) Note that in comparison, a typical SEQUEST search will take a few seconds per spectrum.

Step 3: Validating the Results by PeptideProphet

Select the tab **Analyze Peptides**, click **Add Files**, navigate to class/SpectraST/data, and select the file "LQ20060324_s_1X_26.xml". **Run XInteract** (which will run PeptideProphet) in the same way you did previously. Note that SpectraST searches can be handled in the same manner as SEQUEST and Mascot searches. **View** the "interact.shtml" file once it is finished running. This will bring up the PepXML Viewer that you should be familiar with by now.

View the actual spectra by clicking on any of the spectrum names (e.g. "LQ20060324_s_1X_26.0990.0990.2"). The matched library spectrum is on top, displayed as a "mirror image" of the query (unknown) spectrum at the bottom. Be sure to try a few with high probabilities, and a few with lousy probabilities to see how much the library spectrum resembles that of the query spectrum in each case. Ask yourself if you would believe that the unknown spectrum originates from the same peptide ion as that of

the library spectrum, and that if you would believe in the assignment to that particular *peptide sequence* indicated by the library entry.

The coloring scheme for the spectrum viewer is as follows. In the library (top) spectrum: *Red lines* = peaks assigned to common ions (default are y's and b's – you can customize this on the left panel); *Blue lines* = unassigned peaks; *Red peak labels* = indicates that this peak is present in the query spectrum as well; *Black peak labels* = indicates that this peak is missing in the query spectrum. In the query (bottom) spectrum: *Red lines* = peaks that match assigned peaks in the library spectrum; *Black lines* = peaks that do not match any assigned peaks in the library spectrum. In the ion table underneath the spectra: *Red colored boxes* = ions that are present in both spectra; *Pink-colored boxes* = ions that are present in the library spectrum only; *White colored boxes* = ions that are not present in either spectra.

Now, filter the hits for only those with probability >= 0.99. How many peptides are left? (661) How many unique peptides? (645)

Click on any of the probabilities to view the PeptideProphet analysis summary. Were the models fitted properly to the actual scores? What are the predicted sensitivity and error rate at the probability threshold of 0.99? (Sensitivity = 0.844, error = 0.001) Also, check if the positive and negative distributions make sense. What is different in these distributions (e.g. number of tryptic termini) compared to those you have seen before in SEQUEST or Mascot searches? Why do you think that is?

Step 4: Comparing SpectraST results to SEQUEST results

For comparison, SEQUEST search was performed on the same dataset for you. Click on the **Tools** link at the top, select the **Browse Files** tab, navigate to the directory "class/SpectraST/data", then **View** "interact-SEQ.shtml".

Again, filter the hits for only those with probability >= 0.99. How many peptides are left? (365) How many unique peptides? (359) How are these numbers different from what you obtained for SpectraST?

Click on any of the probabilities to view the PeptideProphet analysis summary. Were the models fitted properly to the actual scores? Comparing SpectraST and SEQUEST, which program is able to separate the positive and negative hits better? (Hint: Note the amount of overlap between the positive and negative distributions.) What are the predicted sensitivity and error rate at the probability threshold of 0.99, and how does that compare to what you have seen for SpectraST? (Sensitivity = 0.379, error = 0.005)

If you still have time and are curious, try to determine why some spectra are identified by SpectraST but missed by SEQUEST. To do so, view both PepXML files side-by-side, and sort by spectrum name to help you cross-check queries between the two lists. View some spectra for cases where SpectraST found a confident identification but SEQUEST did not. Watch out for factors like noisiness, presence of dominant peaks,

fraction of unassigned peaks, etc. Another possibility is that the SpectraST identification may be out of the precursor mass tolerance of SEQUEST (3 Da). For this search, SpectraST uses a precursor ***m/z*** tolerance of 3 Th, which is equivalent to 3 Da for +1 peptides, 6 Da for +2 peptides, and 9 Da for +3 peptides. (Hint: The mz_diff column will help you determine if this is indeed the case: if mz_diff is greater than 1.5 for +2 peptides or 1.0 for +3 peptides, then it lies outside of SEQUEST's precursor mass window.)