Analysis Tutorial

for the APEX Quantitative Proteomics Tool

03.20.2010

This document describes the mechanics of using the APEX Tool to generate APEX protein expression values. This guide focuses on the use of the tool rather than underlying concepts of the APEX technique. Prior to running through this tutorial, please read the ***APEX Process Overview*** in the manual **OR** the ***APEX Overview*** available within the tool's help pages in order to set the context for the steps described in this tutorial.

Launching the APEX Tool

To launch the APEX tool using a computer running Windows, double click on *apex.bat* within the main APEX folder. Linux users can launch by navigating a terminal window to the main apex folder and typing 'sh apex.sh' on the command line. Mac users can either use the apex.sh file as described for Linux users, or can double click on the APEX_Mac_App application bundle icon. If the APEX tool fails to launch, it is likely that the Java Runtime Environment (JRE) needs to be installed or updated. Follow the directions in the manual for getting the latest Java JRE release.

Interface Orientation

The APEX interface is made up of a set of tabbed panes. Each tabbed pane is responsible for controlling a specific processing task. Each of these tabbed panes, called *Process Panels*, includes an information button near the lower left corner which opens a help page that describes the parameters and process details for the corresponding process panel. Each help page has a link back to the help index page. Take a few minutes to view the various process panels and click on the help buttons in the lower left (question marks) to explore the help page system.

Input Data

The sample data are contained in the *sample_input_files* directory within the *data* directory in the main apex directory. Within this folder you will find three files that can be used to run through this tutorial.

yeast_top_protein_acc_list.txt – a list of high abundance proteins used to generate training data.

yeast_sequence_2004.fasta – a sequence file that covers the protein sequences under study.

yeast_MSMS_LCQDecaPlus_5inj-prot.xml – a ProteinProphet output protXML file.

Generating Results

The process of generating protein abundance values is divided into three basic steps, generating training data, computing O_i values, and APEX Score computation. Each of these basic tasks are described in the following sections.

Building Training Data

In this step we will construct an ARFF format file that contains training data. This data will be used in the next stage to train a classifier during O_i generation.

1.) Click on the ***Build Training Data*** tab on the APEX interface.

2.) Specify training proteins by choosing *Supply Protein Accession List File* option in the upper left area of the process panel. Select the Accession List File by pressing the *Select Accession File* button. The file chooser will default to the APEX *data* folder. Navigate the file selection dialog into the *sample_apex_input_files* folder and select the ***yeast_top_protein_acc_list.txt*** file. This file contains 50 high abundance protein gi accessions.

3.) Select the sequence fasta file, ***yeast_sequence_2004.fasta*** , using the sequence file selection area. The file selection dialog should start within the ***sample_apex_input_files*** folder. This file will supply the sequence for the 50 proteins corresponding to the protein accessions in the list file.

4.) Select the ProteinProphet XML (***yeast_MSMS_LCQDecaPlus_5inj-prot.xml***) file using the XML file selection button. This file will supply information on which tryptic peptides associated with the selected proteins have been observed in prior MS results.

5.) Specify an output file name. Note that the .arff extension will be appended automatically. The output file will be in ARFF format, Attribute Relation File Format, a comma delimited data file which captures peptide attributes related to the peptides cleaved from the top proteins. The file format Appendix in the manual has additional information on the ARFF format.

6.) Accept the default values, 0, for number of consecutive missed cleavages, the minimum peptide length, and minimum and maximum peptide mass filters. These filters exclude peptides cleaved from the top proteins prior to computation of the peptide physicochemical properties.

7.) Accept the default values for Peptide Properties specified in the area on the right. The selected peptide properties, often referred to as peptide attributes, are computed for each tryptic peptide.

8.) Hit the ***Construct ARFF Data File*** button at the bottom of the panel to start the process. A small progress dialog will be launched to report on the stage of process. Click 'OK' once the process is complete. The output training ARFF file will ready to view and use in the next process.

Note that the Progress Log Panel (last tab on the right) has been updated to reflect the ARFF training file construction process. The ARFF file contains many comments about the input parameters and the results. Open the resulting ARFF file using a text editor to examine the format of this file. The ARFF format is described in the manual in the appendix on file formats. In that manual section there is also a link to a web page that defines the ARFF format.

Generation of O_i Data Files

In this stage the APEX Tool will utilize a sequence FASTA file and the ARFF training data generated in the last process to generate an O_i file. O_i values reflect the predicted number of observed or detected peptides that are derived from a particular protein *i*. These values will be used in the next process to generate APEX abundance values. During O_i generation, protein sequences undergo an in silico trypsin digestion to produce peptides. Each of these peptides has a set of peptide attributes or properties computed which characterize the peptide. The training data that was produced in the previous processing task is used to build or *train* a classifier that is used to estimate a probability of being observed for each peptide. The O_i values in the output file are the summation of the individual peptide observation probabilities for the peptides derived from protein *i*.

1.) Select the ***Generate Oi Data File*** processing panel tab to display the parameter selection options for this task.

2.) Select the input sequence fasta file, ***yeast_sequence_2004.fasta***, using the file selection button in the upper left area of the process panel. This file contains all known proteins for the species/strain under study in this example. Each protein in this file will have an associated O_i value computed in this process. Only proteins having O_i values can be quantified so this sequence file should cover all known possible proteins that one might want to quantify for a particular study.

3.) Select the training data ARFF file that was created in the previous step by using button under the training data ARFF file selection area on the left side of the panel.

4.) Specify an output file location and name using the output file section of the dialog. The extension .oi will be appended to the file name. The output file will contain an O_i value for each protein and will list the parameters and file used to generate the file.

5.) Accept the default, 0, consecutive missed cleavages in the digestion parameter area in the upper right area of the process panel.

6.) The ARFF training file contains a set of protein properties. The peptide properties selection area allows one to select to use all ARFF properties or to select to use a subset of ARFF properties. Keep the option default option labeled *Use All ARFF Properties*.

7.) Mouse over the '?' in the Classification Parameters area (just to the left of the button) to display a tool tip that displays the current parameters. Note that the options and values displayed in this tool tip relate to the Weka data mining software. The meaning of the option tags is not informative until you become familiar with the parameters associated with these tags. The *Select/View Classification Parameters* button can be used to view and select classifier parameters. Accept the default parameter selections for this run.

8.) Hit the 'Generate O_i Data File' button at the bottom of the page to launch the process. This process takes approximately two minutes on this data set of more than 6000 proteins. Construction of the classifier should happen in roughly 1 minute. The remainder of the time is spent computing peptide properties and applying the classifier to the peptides.

As a test we have tested multi-species fasta files containing nearly 100,000 proteins. In this case the classifier is built one time and proteins are processed in batches of 1000 proteins at an overall rate of approximately 3000 proteins per minute.

Dismiss the progress dialog and open the output O_i file to examine the format of this file. This file will supply O_i values for APEX score computation.

APEX Score Computation

Once an O_i data file has been created, it can be used for multiple analyses that involve the proteins represented in the O_i file. This means that the two preceding steps are run very infrequently. The main reasons to generate a new Oi file would be to either cover the proteins within a different organism (to support APEX abundance computations), or to support a different MS technology in which a protein's O_i value, the predicted peptide detection count, might vary due to the nature of the MS technology.

1.) Select the ***APEX Computation*** tab to display the process panel for this task.

2.) Select the XML file, ***yeast_MSMS_LCQDecaPlus_5inj-prot.xml***, using the file selection button in the upper left. This file supplies protein identification probabilities and peptide counts for each protein.

3.) Select the O_i file created from the previous task section using the O_i file selection panel.

4.) Specify an output file location and name using the output file section of the dialog. The extension .apex will be appended to the file name.

5.) Enter a value for the 'C' parameter. For this sample enter a value of 2500000 as an estimate of 2.5E10+6 proteins per cell.

6.) Hit the 'Compute APEX Scores' button to launch the process. A progress dialog will be displayed and will very soon be followed by the protein selection dialog.

7.) The Protein Selection Dialog provides a list of proteins. Users can use protein identification probability values (p_i) or false positive error rate (Est. FPR) to select a set of proteins. Click in this table to select a set of proteins. As you click in this table, you are selecting the protein that you click and all of the proteins above it in the list. The information at the top of the dialog will update on each click to indicate the number of proteins that will be quantified. APEX scores will be computed for the selected set. Scroll down the list and select a set of about 680 proteins. This is a set that has a false positive error rates less than 0.10. Note that usually the false positive error rate or p_i is used as the protein set selection criteria.

The apex result will be presented in the table in the process panel. The output file is also created and can be opened for viewing in a spreadsheet application. The output file contains all input parameters and file names.

You have just completed a run-through of the basic tasks of the APEX tool. The manual contains additional information on these processes and additional utilities and analysis tasks. The manual has a more complete description of the APEX technique. Further information can be found in the original APEX technique paper that is referenced in the references section at the end of the manual.