Peptide to Spectrum Matching

Shotgun proteomics relies on the assignment of a large number of spectra to theoretical peptides derived from a sequence database. Various search engines have been developed for this task, each with its own advantages and drawbacks. We are going to search the mgf file obtained in the “Peak List Generation” chapter against the database obtained in the “Database Generation” chapter using OMSSA[**1**](#_ENREF_1) and X!Tandem[2](#_ENREF_2" \o "Craig, 2004 #46), two freely available proteomics search engines. The necessary spectrum and database files can be found in the resources folder.

Peptide 1

Peptide 2

Peptide 3

Peptide 4

Peptide 1

Peptide 2

Peptide 3

Peptide 4

Peptide 1

Peptide 2

Peptide 3

Peptide 4

Spectrum Collection



Protein Database



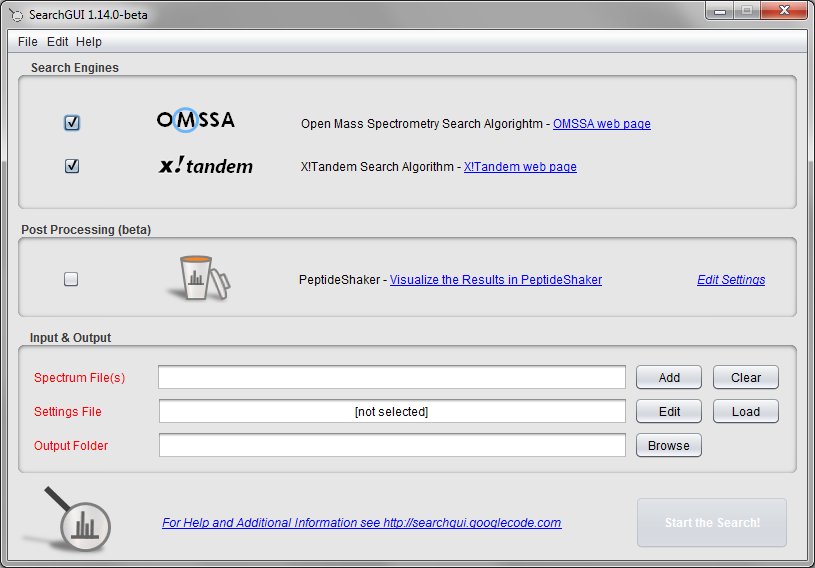
Results



Search Engine

Both OMSSA and X!Tandem can easily be used together *via* a simple user interface called SearchGUI.[3](#_ENREF_3) SearchGUI for Windows platforms is provided in the software folder together with OMSSA and X!Tandem. For Mac and Linux versions, please see the SearchGUI web page: [http://searchgui.googlecode.com](http://searchgui.googlecode.com/). Start SearchGUI by double clicking the file SearchGUI-X.Y.Z.jar (replace X.Y.Z with the current SearchGUI version number).

You will then see the following dialog:



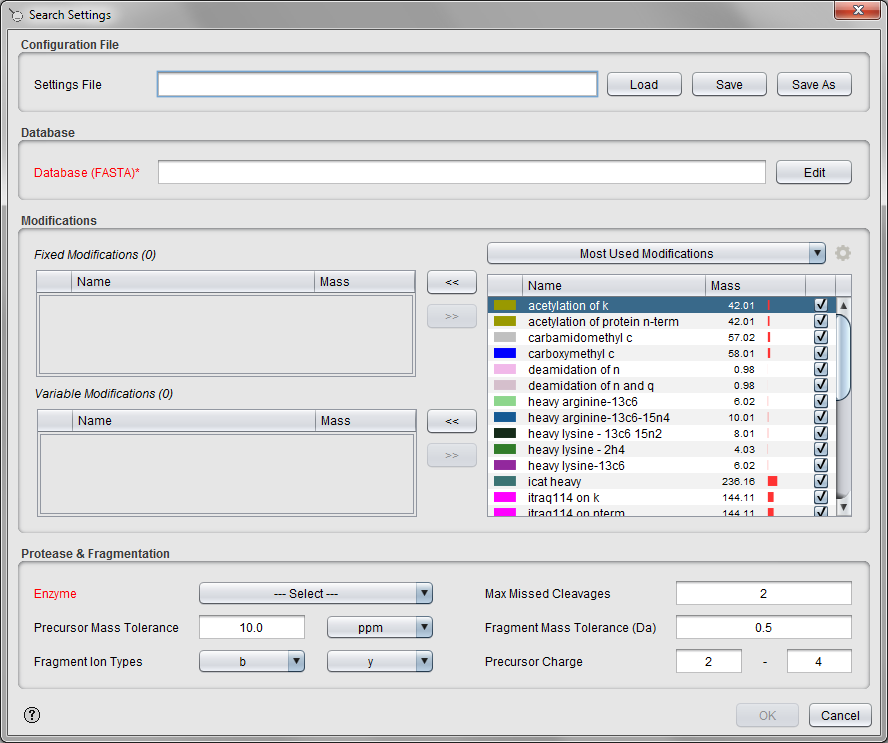
You will notice that OMSSA and X!Tandem is already selected. In fact, keen observers may already have noticed the search engines in the SearchGUI home folder. This means that when you have downloaded the SearchGUI zip file and unzipped it (which comprises the entire installation procedure), you have also already installed OMSSA and X!Tandem along with it!

*Is this legal? Can the SearchGUI developers do this? They did not make OMSSA or X!Tandem? [1.3a]*

In order to perform the search, we need to provide the spectra, the database and experiment dependent search settings. Load the mgf file qExactive01819.mgf created in the “Peak List Generation” chapter (also available in the resources folder).

**Tip:**  
*Note that you can load multiple mgf files and even entire folders.*

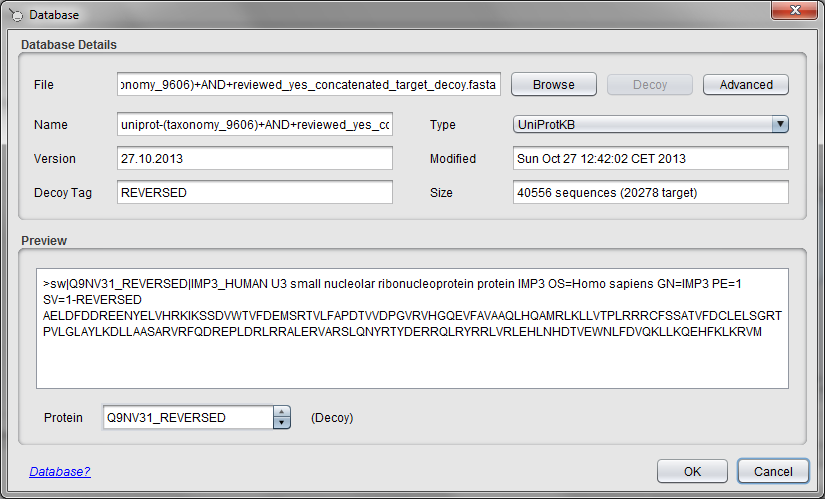
We are now going to set the search settings in the Search Settings dialog. Click the 'Edit' button after the 'Settings File' text field.



Note that at the top of this tab, you can load or save the settings you will enter - this makes it easy to keep track of your search settings, and to reuse them *verbatim* later on. These can also be loaded directly in the main SearchGUI display. For now, we will define the settings manually.

First we need to specify the database to search against. We will use the database generated in the "Database Generation" chapter. *How does the database used affect the results? Will we always find the same proteins? How does the size of the database affect the significance/score of the proteins we find? [1.3b]*

Most proteomics database searches are performed as so-called target/decoy searches, and to perform such a search you first have to add the decoy protein sequences to your database file. More details on target/decoy searches will follow in the chapter called “Peptides and Proteins Validation”. For now simply select the human database created in the “Database Generation” chapter (also available in the resources folder), and select 'Yes' when SearchGUI offers the option to add decoy sequences. After the decoys have been added you will see a dialog with database details. Click “OK“ to close this dialog.



**Tip:**  
*Decoys can also be added manually by clicking the 'Decoy' button.*

The next step is to specify the modifications to consider. As fixed modifications choose carbamidomethyl c, and as variable modifications choose phosphorylation of s, phosphorylation of t, phosphorylation of y and oxidation of m. *Are these all the modifications you would expect for a standard shotgun experiment? How do you define which modifications are variable and which are fixed? [1.3c]*

**Tip:**  
*'CTRL + Mouse Click' allows you to select multiple entries.*

Next we will choose the enzyme, set it to Trypsin, and keep the number of allowed missed cleavages at 2*. What is a missed cleavage? Why 2 and not 0? Or 1? [1.3d]*

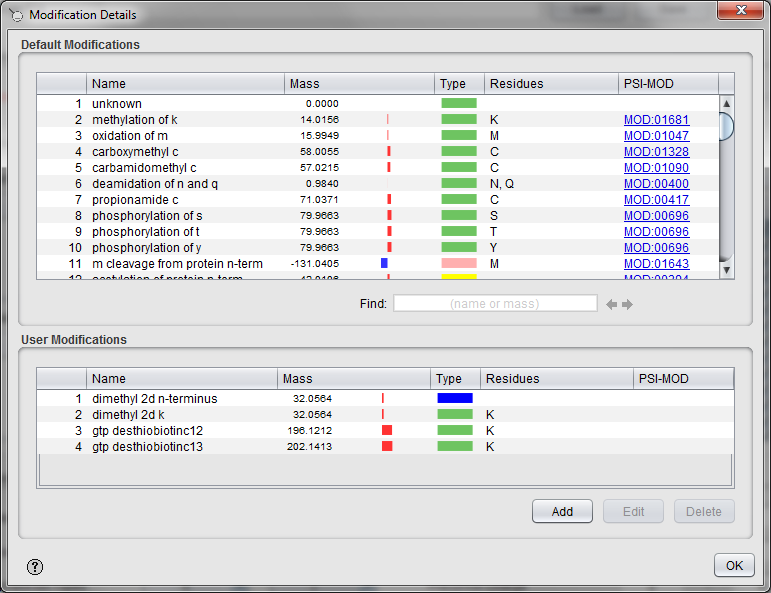
Keep the precursor ion mass tolerance at 10 ppm and set the fragment ion mass tolerance to 0.02 Da. *How do we choose these values? What is the difference between using a mass tolerance in ppm or Dalton? [1.3e]*

The fragment ion types and the charge bounds are fine as they are. *Why? [1.3f]*

Note that only the most commonly used modifications are listed in this dialog. There are more modifications available in SearchGUI, and you can also set up your own modifications. To see all the modifications select "All Modifications" in the drop down menu above the modifications table.

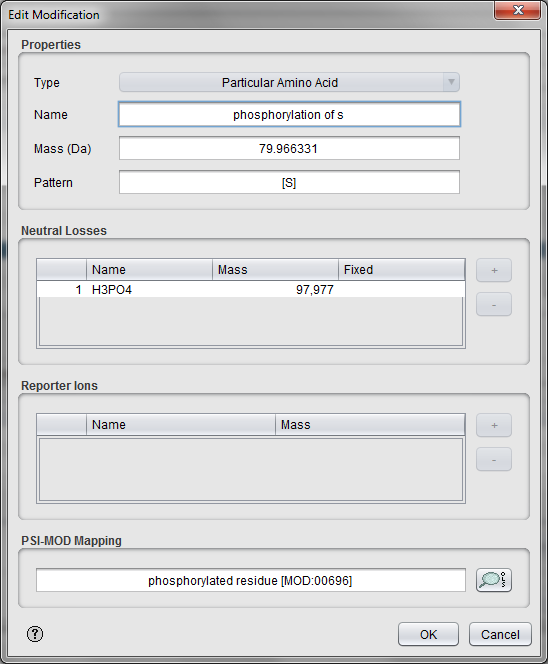
To see the modification details or to add your own modifications click the settings ions next to the drop down menu. (The modification details are also available in the main SearchGUI frame, Edit menu > Modifications.)

You can use the 'Find' feature to locate a given modification. For example, type ‘phosphorylation’ in this field and you will see that there are 14 modifications related to phosphorylation:



It is of course crucial to select the correct modifications. *What is the difference between the different phosphorylation possibilities? How does the selection affect your search results? [1.3g]*

Double clicking on a modification (or right click > "Edit") brings up the modification details:



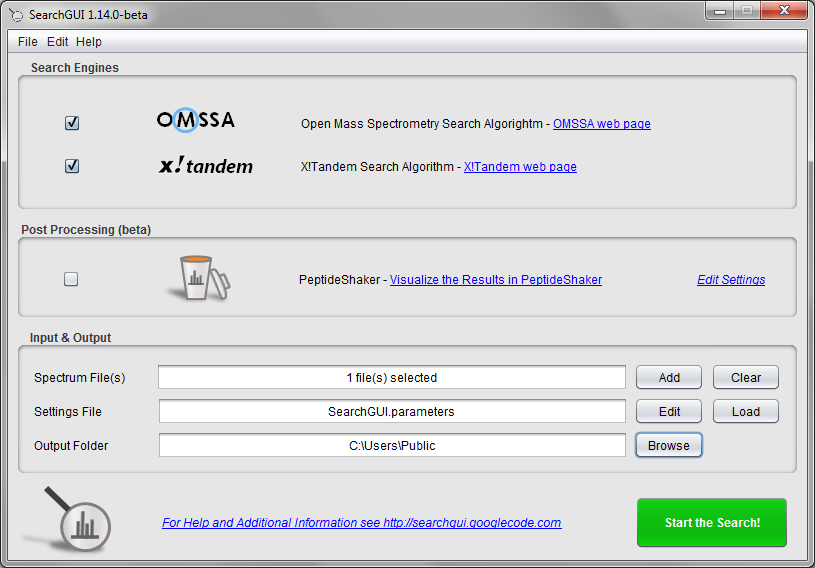
*What is a neutral loss? What is a reporter ion? [1.3h]*

Close the modification details dialogs and go back to the Search Settings dialog. All the search settings are now filled in. Go to the top of the dialog, click the 'Save As' button and save the settings for future reuse. The next time you want to use the exact same search settings you can simply select this file in the main SearchGUI frame. Click 'OK' to close the Search Settings dialog and go back to the main dialog.

**Tip:**  
*A well-organized library of search parameter files can save a lot of time!*

Note that both search engines are selected at the top. This means that one can easily run a search with the provided search settings on both OMSSA and X!Tandem, and get result files for both of them at the same time.

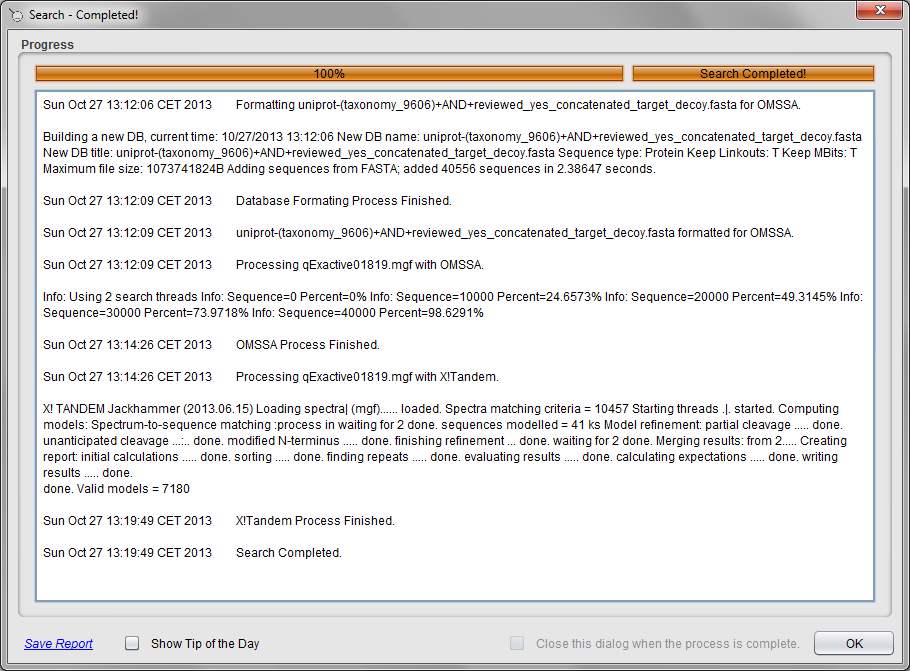
Leave the PeptideShaker post-processing option unchecked for now (we will get back to this option later). Finally, select an output folder. You should now see the following screen:



**Tip:**  
*Using an empty folder for the search output simplifies the post-processing!*

Pressing the ‘Start the Search!’ button will launch the search. A progress bar and scrolling text will keep you informed on the progress of the search. *How does the size of the spectrum file affect the search time? What about the database size? The search parameters? Can all searches be performed on a standard desktop computer? [1.3i]*

A screenshot of the dialog after completion is shown below:



After completion, the output folder will contain several files, where the two most important are the output files for the search engines. The search takes around eight minutes on a standard laptop: to save time, you can cancel the process and use the files provided with the tutorial: the OMSSA output file is called qExactive01819.omx, while the X!Tandem output file is called qExactive01819.t.xml. These files contain so-called Peptide to Spectrum Matches (PSMs) inferred by the search engines. Note that these files can become quite big. We will learn how to interpret these matches in the next chapter.

If you encounter any issues with SearchGUI, please consult the troubleshooting section at: <http://searchgui.googlecode.com>.

References

1. Geer, L.Y. et al. Open mass spectrometry search algorithm. *J Proteome Res* **3**, 958-964 (2004).

2. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466-1467 (2004).

3. Vaudel, M., Barsnes, H., Berven, F.S., Sickmann, A. & Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **11**, 996-999 (2011).