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 shortcomings. We are going to search the mgf file obtained in the “Peak List Generation” chapter against the database obtained in “Database Generation” chapter using OMSSA[**1**](#_ENREF_1) and X!Tandem[2](#_ENREF_2), two freely available 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



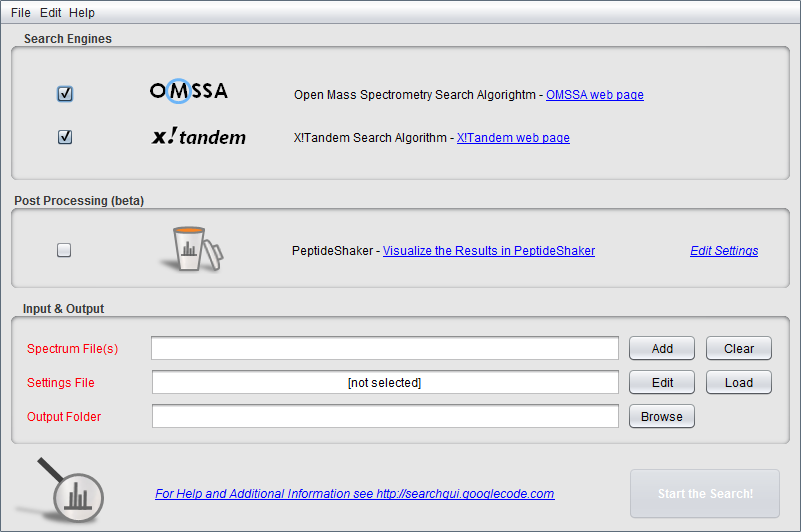
Results



Search Engine

Every search engine has its own specificity and it is recommended to study them on their respective web pages: [http://pubchem.ncbi.nlm.nih.gov/omssa](http://pubchem.ncbi.nlm.nih.gov/omssa/) and [http://www.thegpm.org/tandem](http://www.thegpm.org/tandem/). However, it is possible to use them together *via* a simple 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:



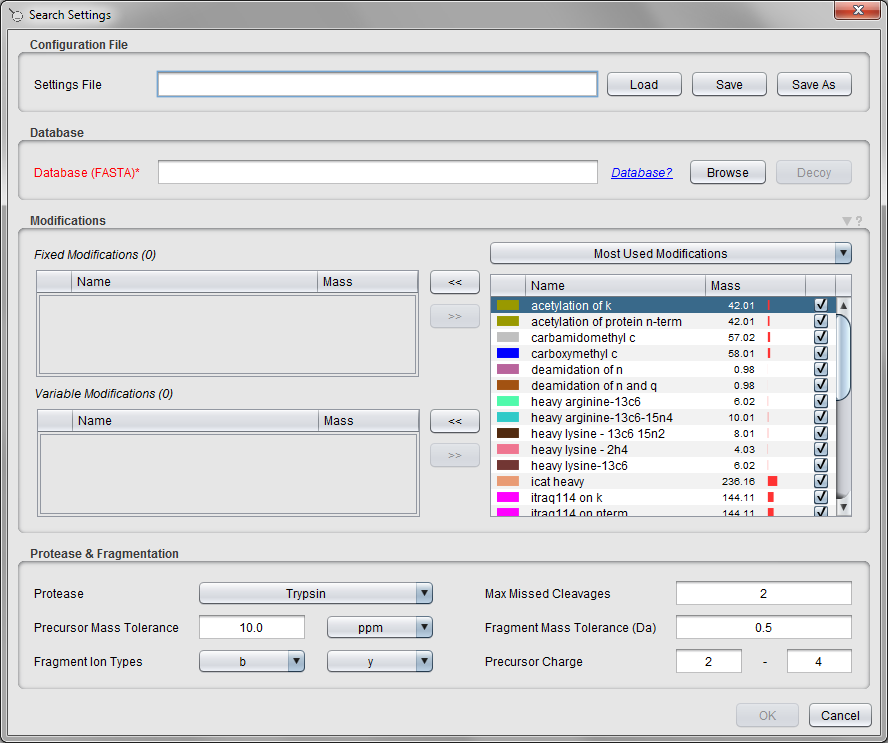
You will notice that OMSSA and X!Tandem is already selected. In fact, keen observers may already have noticed the search engines in the SearchGUI home folder. This means that when you have downloaded the SearchGUI zip file and unzipped it (which comprises the entire installation procedure), you have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

*Is this legal? Can the SearchGUI authors do this? They did not make OMSSA or X!Tandem?*

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 created in the “Peak List Generation” chapter (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 Chapter 11. *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?*

*Selecting the correct database is a crucial step in proteomics. First, it needs to be as comprehensive as possible: you cannot find a protein which is not in the database. Moreover, if a protein is missing, the search engines might attach spectra derived from this protein to another resembling protein – hence making a false identification. It is thus crucial that you leave enough room for the search engine to “distribute” mistakes. However, using a too large database will lower your probability to find your proteins.*

*Generally, it is recommended to use the reference database of your species of interest completed with the sequences of expected contaminants: keratin, proteases used for protein digestion, etc. See the “Database Generation” chapter for more details.*

*Finally, bear in mind that the content of sequence databases evolves with time. It is hence important to constantly use the same database for a given project and document its version in every communication.*

Most proteomics databases 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 chapter “Peptides and Proteins Validation”. For now simply select the human database created in chapter “Database Generation” (also available in the resources folder), and select 'Yes' when SearchGUI offers the option to add decoy sequences. Note that the selected database has changed to use the target/decoy version.

**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?*

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

*There are two types of modifications: modifications induced by the experimental workflow and natural modifications of the sample. Among the modifications occurring when conducting the experiment, some are produced voluntarily like carbamidomethylation of cysteine here and some are experimental artefacts like oxidation of methionine here. These have hence to be selected in order to identify the proteins. The biological modifications on the other hand are selected in order to target biological functions. However these are typically low abundant: we have very little chance to identify a phosphorylated protein without enrichment*[*4*](#_ENREF_4) *– we actually here selected phosphorylation for illustrative purpose only.*

*Selecting variable modifications has a similar effect than using a large database: it increases the number of possible results, hence reducing our chances to identify our proteins. It is hence advised to reduce the number of variable modification. This can be done by selecting fixed modifications: for these, every targeted residue will be a priori considered as modified. Non-modified peptides will hence not be identified: it is to be used only when all peptides are expected to be modified. Here, carbamidomethylation is a high yield chemical process which will target all residues.*

*In case of doubt, it is very easily to control the level of modifications by doing a pre-search with the modification of interest as variable.*[*5*](#_ENREF_5) *Here, searching with oxidation of methionine and carbamidomethylation of cysteine as variable modification returned >98% of cysteine residues modified. The modification can thus reasonably considered as fixed. Note that such quality control steps are for crucial importance when working with chemically labelled samples.*[*6*](#_ENREF_6)

Then we’ll need to choose the enzyme - leave it set at Trypsin, and keep the number of allowed missed cleavages at 2*. What is a missed cleavage? Why 2 and not 0 or 1?*

*Missed cleavages are parts of the peptide sequence where one would expect the protease to cleave. Missed cleavages can occur due to incomplete digestion. Due to the impossibility for the protease to access some cleavage site or protease quality,*[*7*](#_ENREF_7) *some missed cleavages will always remain,*[*8*](#_ENREF_8) *in our experience up to two with trypsin.*[*5*](#_ENREF_5)

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

*With the first low resolution mass spectrometers, searches were conducted with a fixed tolerance in m/z – using the unit Dalton. With the advent of high resolution mass spectrometry, search engines adapted the tolerance to the m/z actually measured – one would allow a higher tolerance when measuring the mass of an elephant than the mass of a mouse – hence introducing ppm tolerance defined as:*

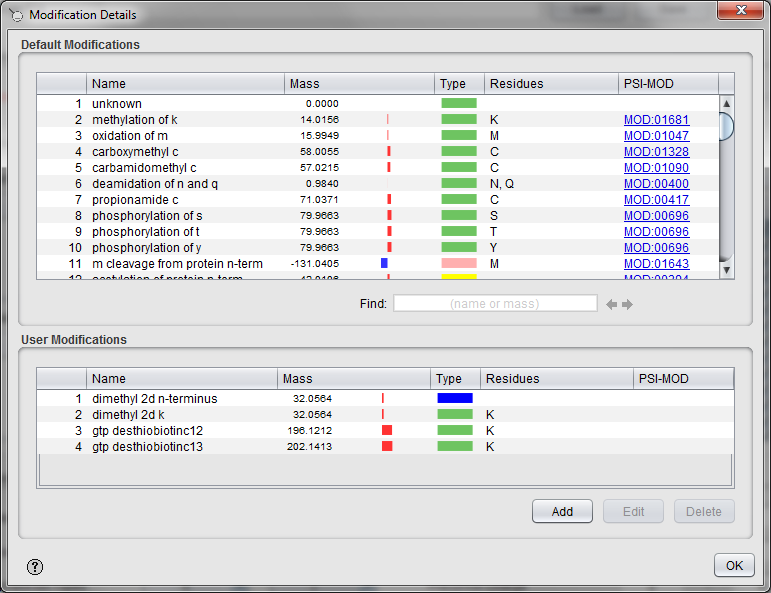
*The mass tolerances depend on the resolution of the mass spectrometer. Here, the data was recorded in the orbitrap where a 10 ppm tolerance gives the best results on our setup. OMSSA and X!Tandem do not allow us to set the fragment ion tolerance in ppm so we use the value of 0.02 Da.*

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

*The data was acquired with higher-energy collisional dissociation (HCD) fragmentation*[*9*](#_ENREF_9) *which principally generates b and y ions.*

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. Click the small triangle above the modification table and select the 'Edit Modifications' option. (The modification details are also available in the main SearchGUI frame, Edit menu > Modifications.)

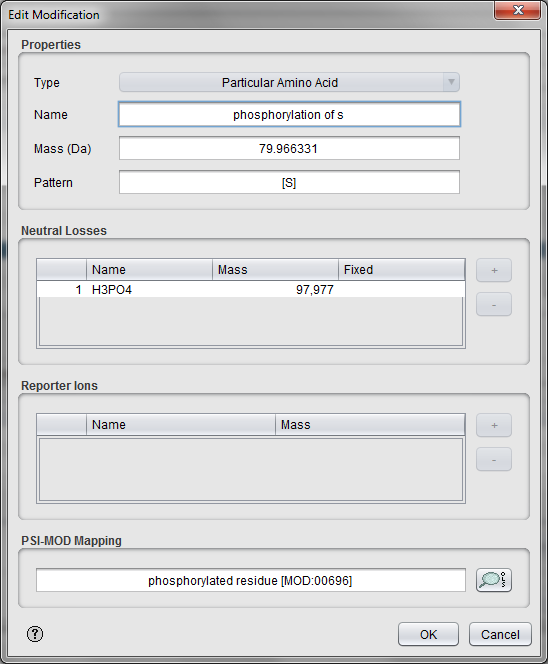
In the 'Find' field type ‘phosphorylation’. You will see that many modifications are available:



*What is the difference between the different phosphorylation possibilities? How does the selection affect your search results?*

*These modifications are all the OMSSA compatible modifications. Some of them will be better suited for your setup than others. Note that X!Tandem might not account for the difference between these OMSSA modifications. For more information on the handling of modifications by search engines, please contact the developers of these.*

Double clicking on a modification brings up the modification details:



*What is a neutral loss? What is a reporter loss?*

*Before a peptide or a fragment ion is recorded, it can lose a moiety named neutral loss. Most encountered neutral losses are water (H2O) and ammonia (NH3) losses. Some modifications like phosphorylation can also generate neutral losses and these can be set in this dialog. Note that this information is not accounted for by OMSSA and X!Tandem.*

*Some modifications can also lose charged moieties, named reporter ions or diagnostic ions. This is for instance used for reporter ion based quantification.*[*10*](#_ENREF_10)*,* [*11*](#_ENREF_11)

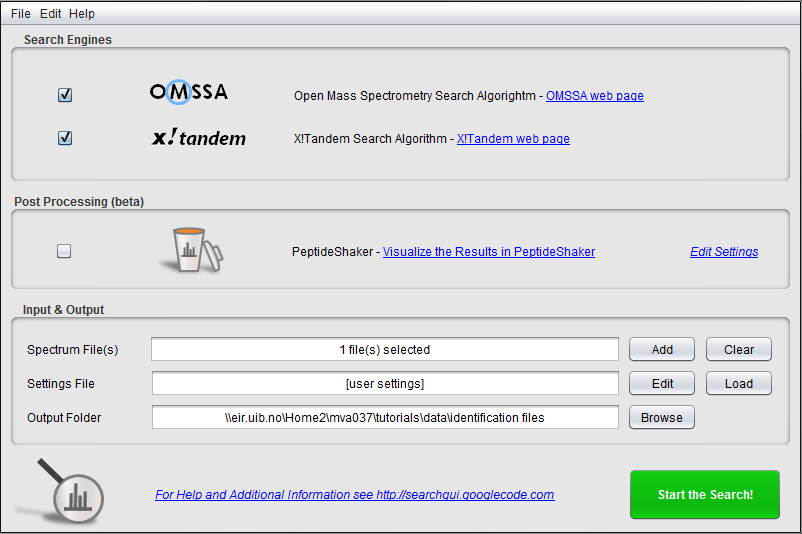
Close the modification details dialogs and go back to the Search Settings dialog. All the search settings are now filled in. Go to the top of the dialog, click the 'Save As' button and save the settings for future reuse. The next time you want to use the exact same search settings you can simply select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

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

Note that both search engines are selected at the top. This means that one can run a search using 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. Finally select an output folder, you should see the following screen:

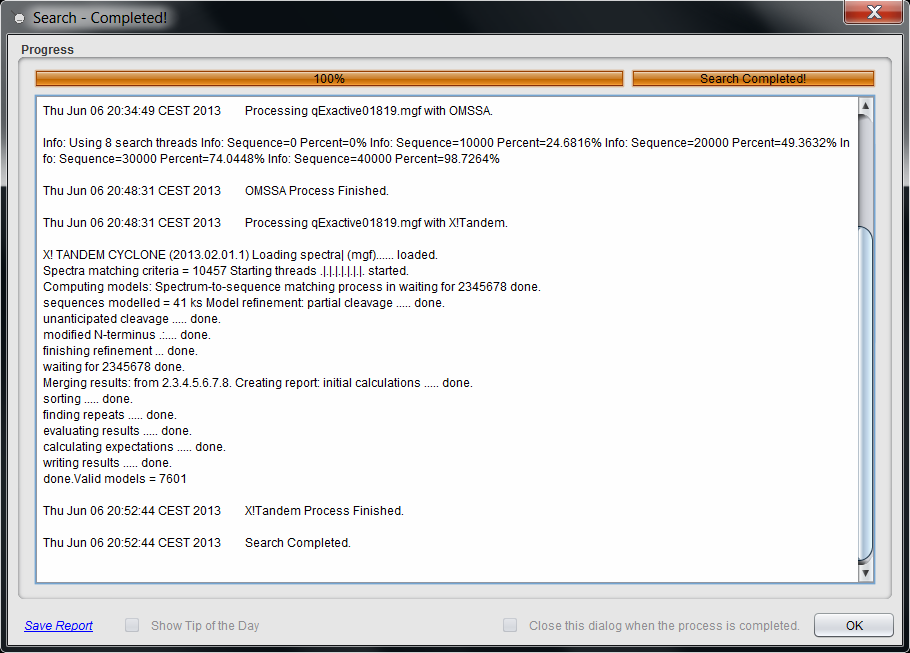
**Tip:**  
*Using an empty folder for the search output simplifies the post processing!*



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 searches. *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?*

*The search time usually scales with the number of spectra and their complexity. A similar effect goes for the database size. Notably, when using large databases, OMSSA will get stuck at ~98% progress during hours or days apparently doing nothing. Just be patient! There is a limitation in file size which can be processed by OMSSA. If this limit is reached, SearchGUI will propose to split the spectrum file. The splitting preferences can be modified in the additional settings. Also, bear in mind that the larger these files, the more challenging their post-processing. As a result, standard desktop computers are often simply unable to process large datasets.*

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. Note that the search takes almost half an hour on a standard laptop: for the sake of 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 their important size. We will see how to interpret these matches in the next chapter.

If you happen to 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).

4. Eyrich, B., Sickmann, A. & Zahedi, R.P. Catch me if you can: mass spectrometry-based phosphoproteomics and quantification strategies. *Proteomics* **11**, 554-570 (2011).

5. Vaudel, M., Burkhart, J.M., Sickmann, A., Martens, L. & Zahedi, R.P. Peptide identification quality control. *Proteomics* **11**, 2105-2114 (2011).

6. Burkhart, J.M., Vaudel, M., Zahedi, R.P., Martens, L. & Sickmann, A. iTRAQ protein quantification: a quality-controlled workflow. *Proteomics* **11**, 1125-1134 (2011).

7. Burkhart, J.M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A. & Zahedi, R.P. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *Journal of proteomics* **75**, 1454-1462 (2012).

8. Fannes, T. et al. Predicting tryptic cleavage from proteomics data using decision tree ensembles. *Journal of proteome research* **12**, 2253-2259 (2013).

9. Olsen, J.V. et al. Higher-energy C-trap dissociation for peptide modification analysis. *Nature methods* **4**, 709-712 (2007).

10. Ross, P.L. et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics : MCP* **3**, 1154-1169 (2004).

11. Thompson, A. et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical chemistry* **75**, 1895-1904 (2003).