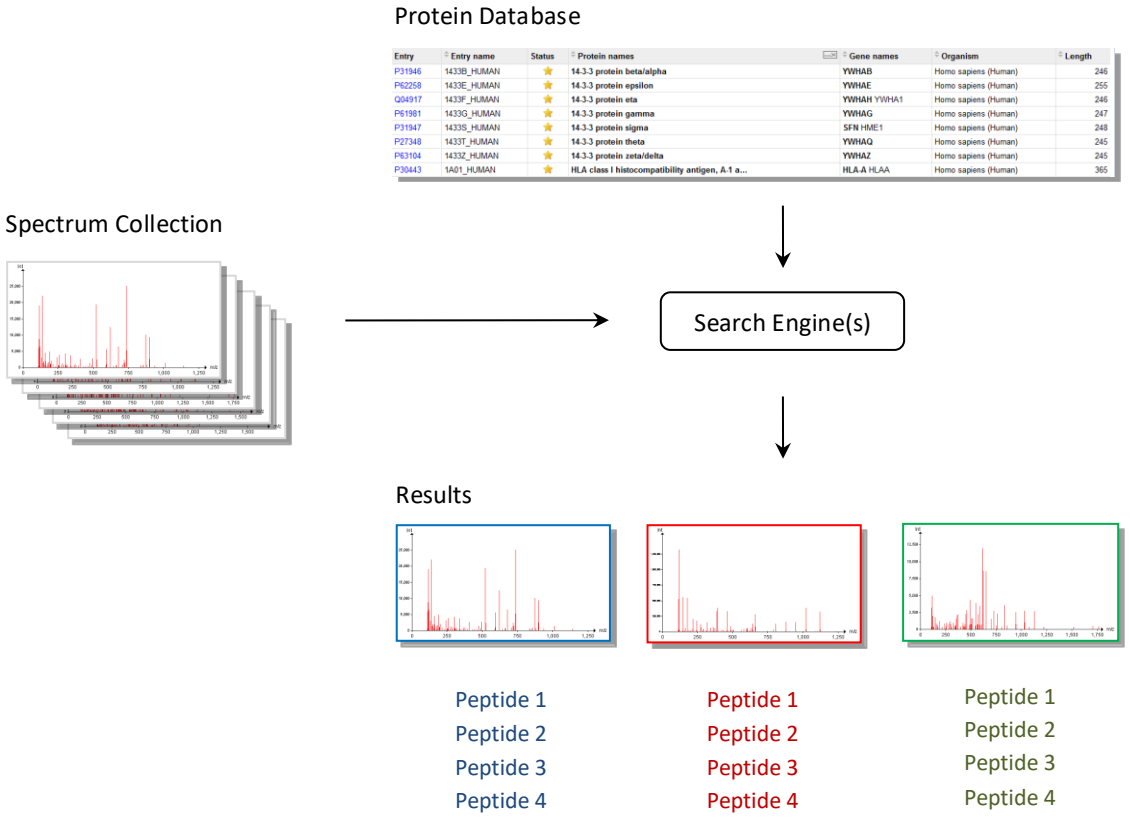
Peptide to Spectrum Matching

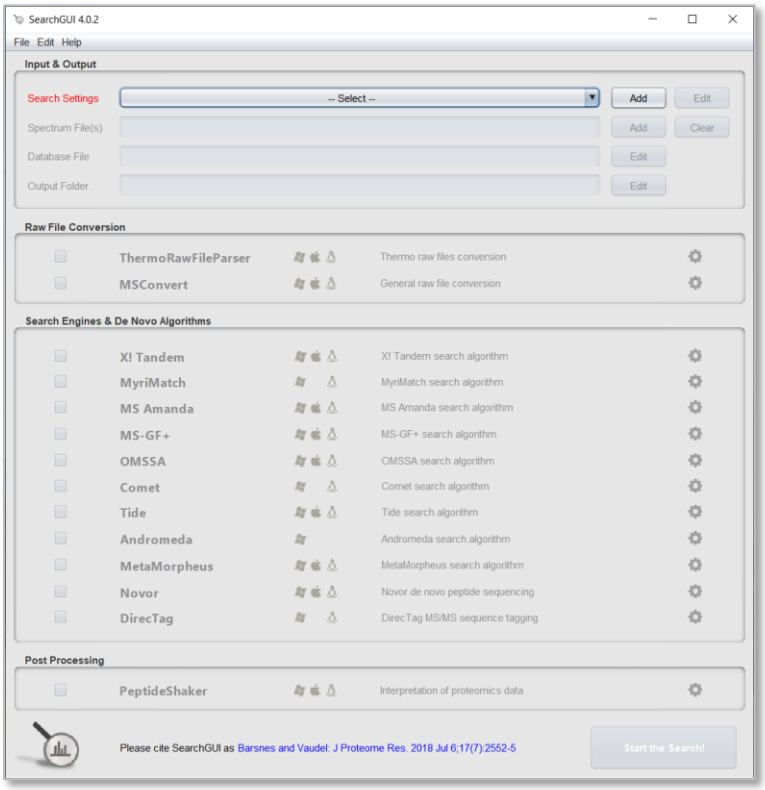
Shotgun proteomics relies on the assignment of a large number of spectra to theoretical peptides derived from a sequence database. Multiple search engines have been developed for this task, each with its own advantages and drawbacks. We are going to search the mgf file obtained in the "Peak List Generation" chapter against the database obtained in the "Database Generation" chapter using freely available proteomics search engines. The necessary spectrum and database files can be found in the resources folder.



Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

To easily run nine different search engines, we will rely on SearchGUI[1], a user-friendly graphical user interface including the following search engines: X! Tandem[2], MyriMatch[3], MS Amanda[4], MS-GF+[5], Comet[6], Tide[7], Andromeda[8], MetaMorpheus[9] and OMSSA[10]. SearchGUI for Windows is provided in the software folder. For Mac and Linux users, please see the SearchGUI web page: http://compomics.github.io/projects/searchgui.html.

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:



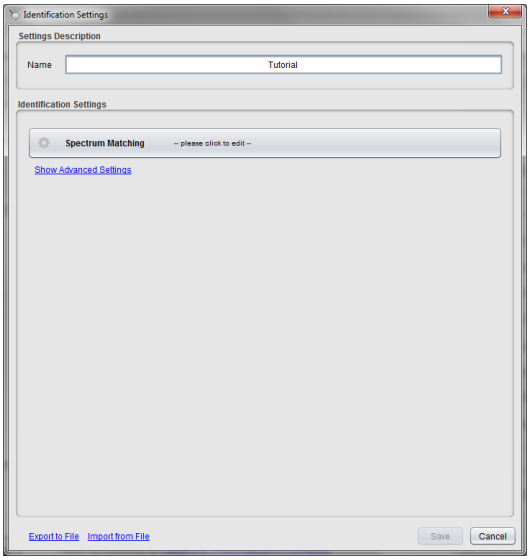
You will notice that all the nine search engines are already available (plus two *de novo* sequencing algorithms, Novor and DirecTag). In fact, keen observers may already have noticed the search engines in the SearchGUI resources folder. This means that when you have downloaded the SearchGUI zip file and unzipped it (which comprises the entire installation procedure), you have also already installed all the nine search engines along with it!

Is this legal? Can the SearchGUI developers do this? They did not make these search engines? [1.3a]

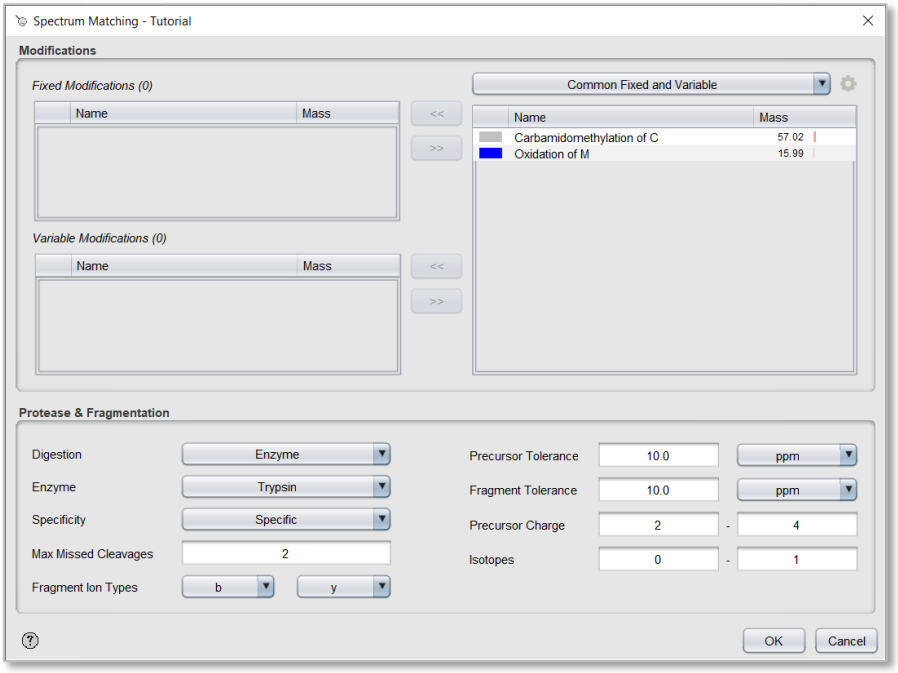
Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

In order to perform the search, we need to provide the spectra, the database and experiment dependent search settings.

We will start with the identification settings. Click the 'Add' button after the 'Search Settings' text field and provide a name for your identification settings as shown below.



We are now going to provide the settings used to map our spectra to the peptides. Click the 'Spectrum Matching' option:



Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

Start by specifying the modifications to consider. As fixed modifications choose Carbamidomethylation of C, and as variable modifications choose Oxidation of M. *Are these all the modifications you would expect for a standard shotgun experiment? How do you decide which modifications should be variable and which should be fixed?* *[1.3c]*

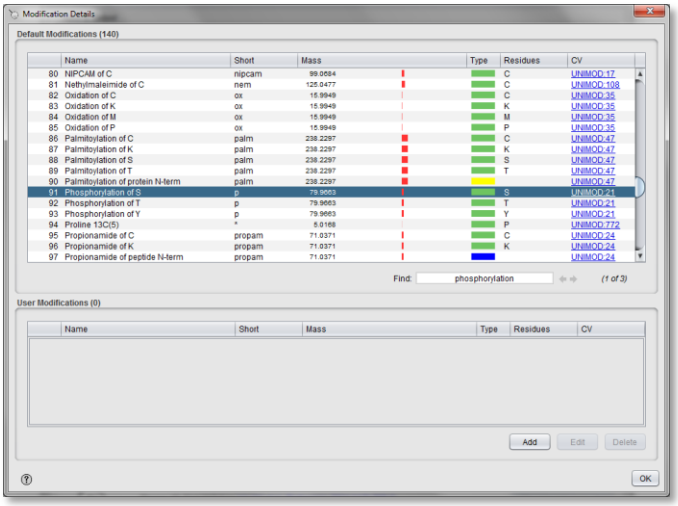
Next, we will choose the enzyme, makes sure that digestion is set to "Enzyme", enzyme set to "Trypsin" and specificity set to "Specific". Keep the maximum number of allowed missed cleavages at 2. *What does "Specific" mean and how does it affect the processing? What is a missed cleavage? Why 2 and not 0? Or 1?* *[1.3d]*

Keep the precursor ion mass tolerance at 10 ppm and change 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 precursor charge and isotope ranges are fine as they are. *Why?* *[1.3f]*

Note that only the most commonly used modifications are initially listed in the dialog. You can see the other modifications by selecting a different modification category above the modifications table. To see all the modifications at once, select the cog wheel icon next to the drop-down menu above the modifications table. In this new dialog you can also create your own custom modifications.

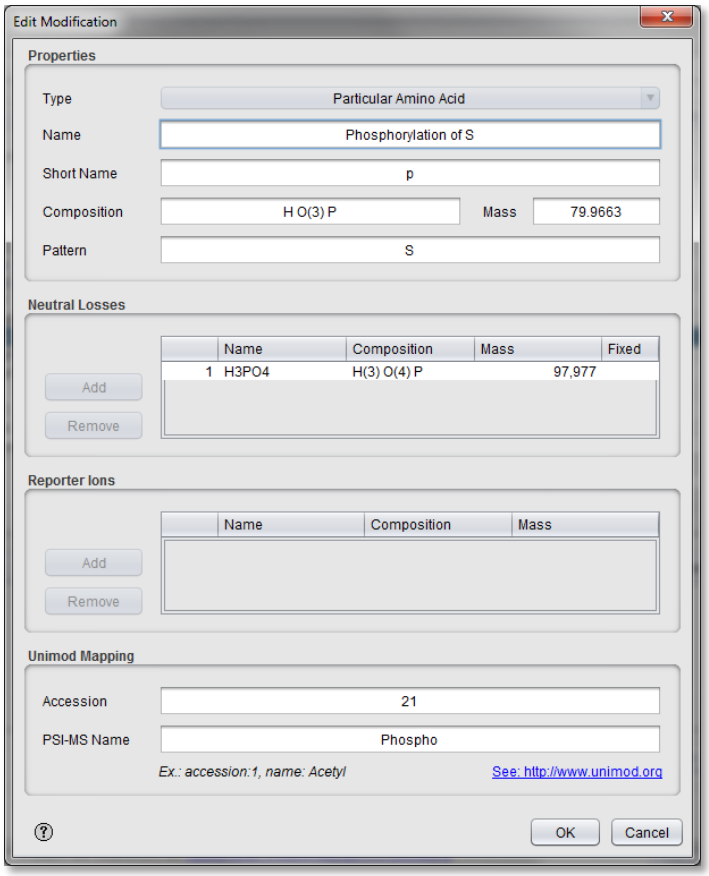
You can use the 'Find' feature to locate a given modification. For example, type 'phosphorylation' in this field and you will see that there are three modifications related to phosphorylation:



Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

It is of course crucial to select the correct set of 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:

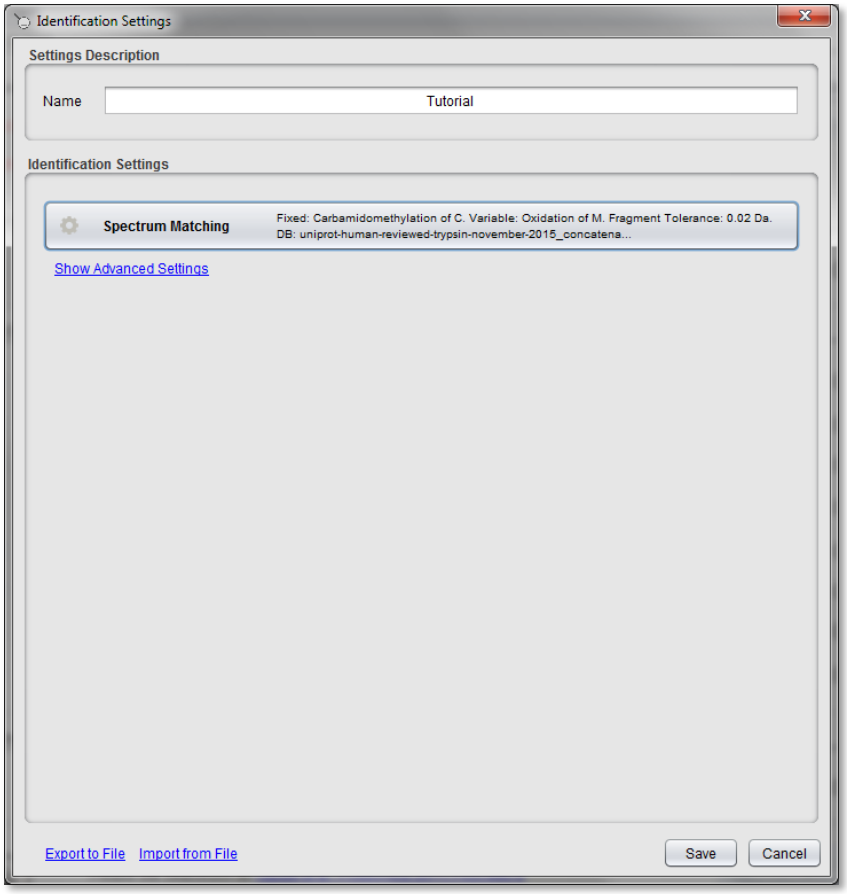


What is a neutral loss? What is a reporter ion? *[1.3h]*

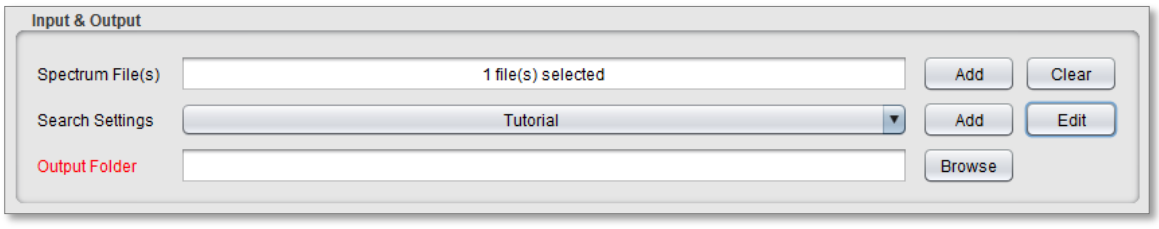
Close the modification details dialogs and go back to the Search Settings dialog. All the search settings are now filled in. Click the 'OK' button to close the Search Settings dialog.

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

Now we have all the identification settings needed.



Click the 'Save' button to save the Identification Settings for later reuse. The next time you want to use the exact same settings you can simply select them in the main SearchGUI dialog.



Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

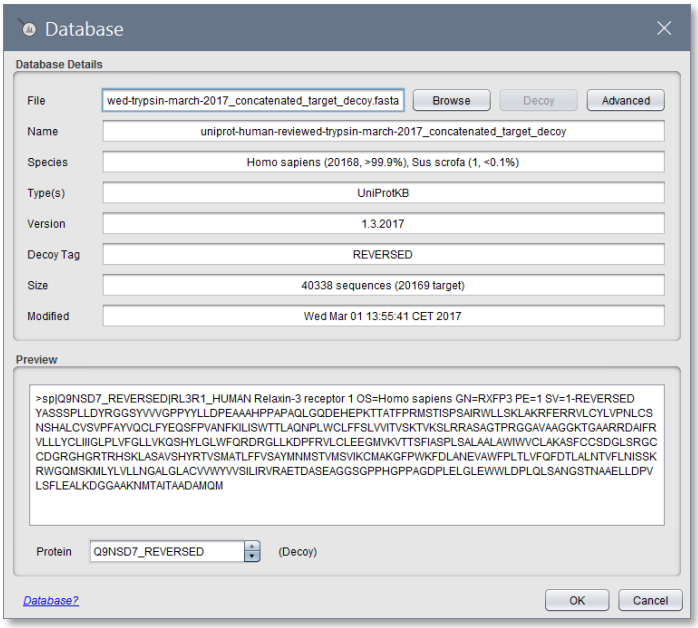
Next, load the mgf file qExactive01819.mgf created in the “Peak List Generation” chapter (also available in the resources folder). If getting a question about adding missing precursor charges choose “No”.

Tip:

Note that you can load multiple mgf files and even entire folders of mgf files.

We then have to specify the database to search against. We will use the database generated in the "Database Generation" chapter (containing all reviewed human protein sequences plus trypsin). *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.



Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

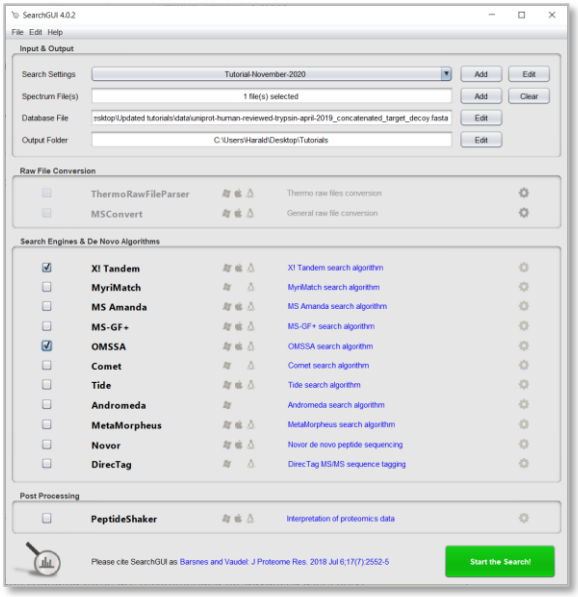
Tip:

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

Finally, select the 'Output Folder'. This is where your search results will be stored.

In the main SearchGUI dialog you will note that none of the search engines are yet selected. By selecting the wanted search engines, you can run a search with the selected search settings and get individual result files for each search engine. To save time we will now only use X! Tandem and OMSSA, so make sure that these are the only two search engines selected.

Leave the PeptideShaker[11] post-processing option unchecked for now (we will get back to this option later). You should now see the following:



Tips:

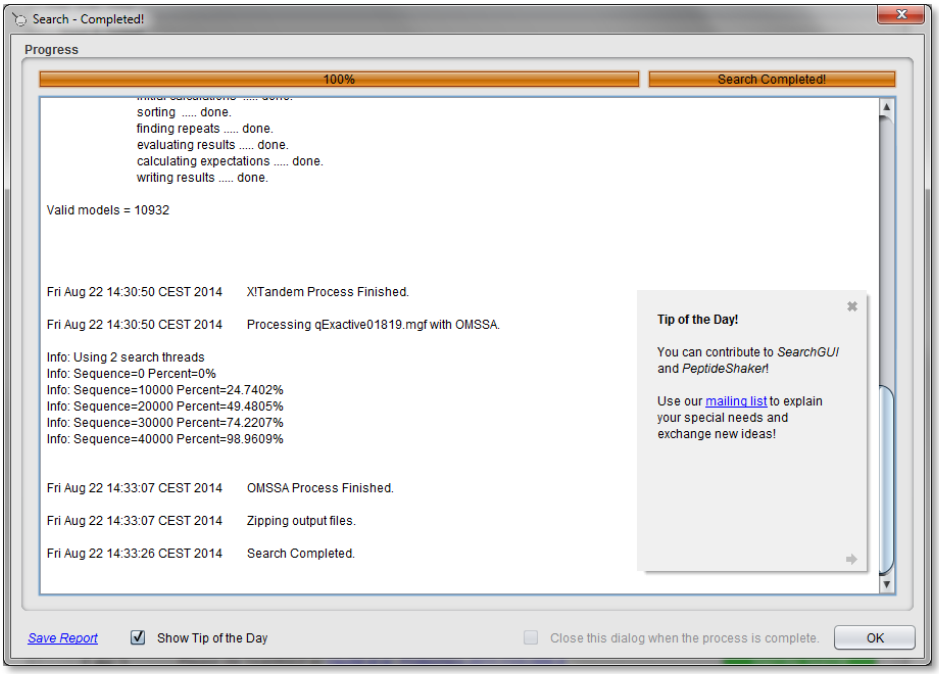
Advanced search engine settings are available by clicking the settings icon to the right of each search engine. Note: changing these are recommended for advanced users only!

Always use an empty folder for the search output as this simplifies the post-processing!

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

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 a zip file with the results called searchgui_out.zip. The search should take maximum a minute or two on a standard laptop. The OMSSA output file is called qExactive01819.omx, while the X! Tandem output file is called qExactive01819.t.xml. These files contain the 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 this information in the next chapter.

If you encounter any issues with SearchGUI, please consult the troubleshooting section at: http://compomics.github.io/projects/searchgui.html.

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)

References

1. Barsnes, H. & Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J Proteome Res* **17**, 2552-2555 (2018).
2. Craig, R. & Beavis, R.C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20**, 1466-1467 (2004).
3. Tabb, D.L., Fernando, C.G. & Chambers, M.C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **6**, 654-661 (2007).
4. Dorfer, V. et al. MS Amanda, a Universal Identification Algorithm Optimized for High Accuracy Tandem Mass Spectra. *J Proteome Res* (2014).
5. Kim, S. MS-GF+: https://bix-lab.ucsd.edu/pages/viewpage.action?pageId=13533355.
6. Eng, J.K., Jahan, T.A. & Hoopmann, M.R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22-24 (2013).
7. Diament, B.J. & Noble, W.S. Faster SEQUEST searching for peptide identification from tandem mass spectra. *J Proteome Res* **10**, 3871-3879 (2011).
8. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805 (2011).
9. Wenger, C.D. & Coon, J.J. A proteomics search algorithm specifically designed for high-resolution tandem mass spectra. *J Proteome Res* **12**, 1377-1386 (2013).
10. Geer, L.Y. et al. Open mass spectrometry search algorithm. *J Proteome Res* **3**, 958-964 (2004).
11. Vaudel, M. et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotech* **33**, 22-24 (2015).

Harald Barsnes (harald.barsnes@uib.no) and Marc Vaudel (marc.vaudel@uib.no)