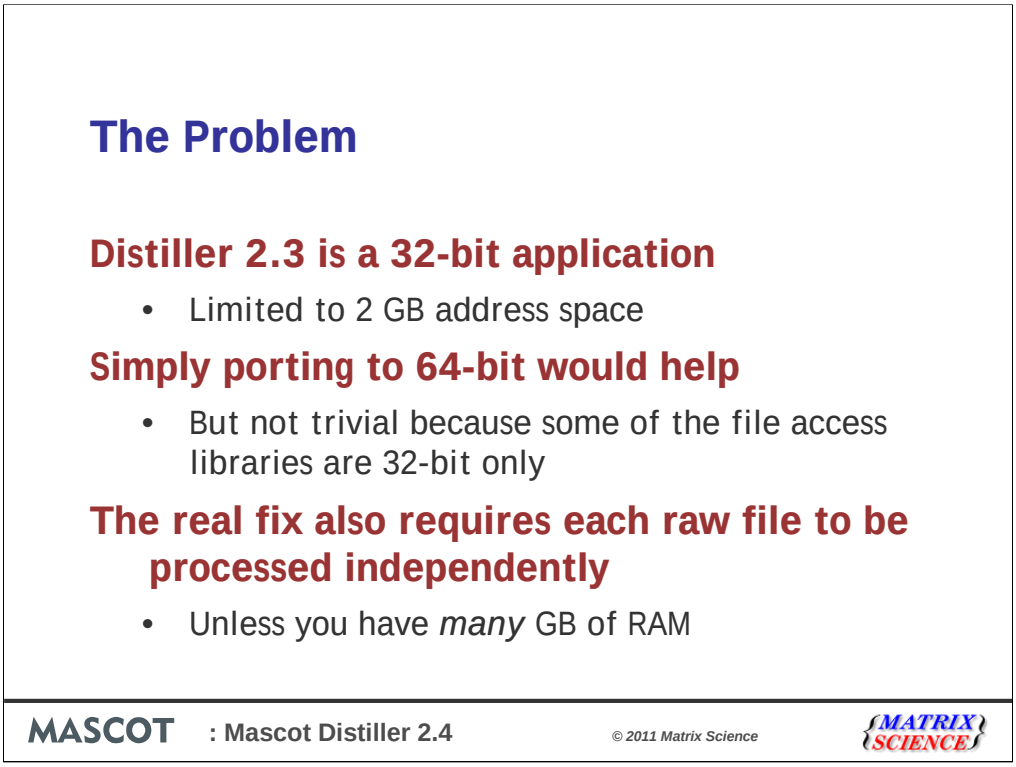
Mascot Distiller 2.4: Quantitation of Very Large Datasets

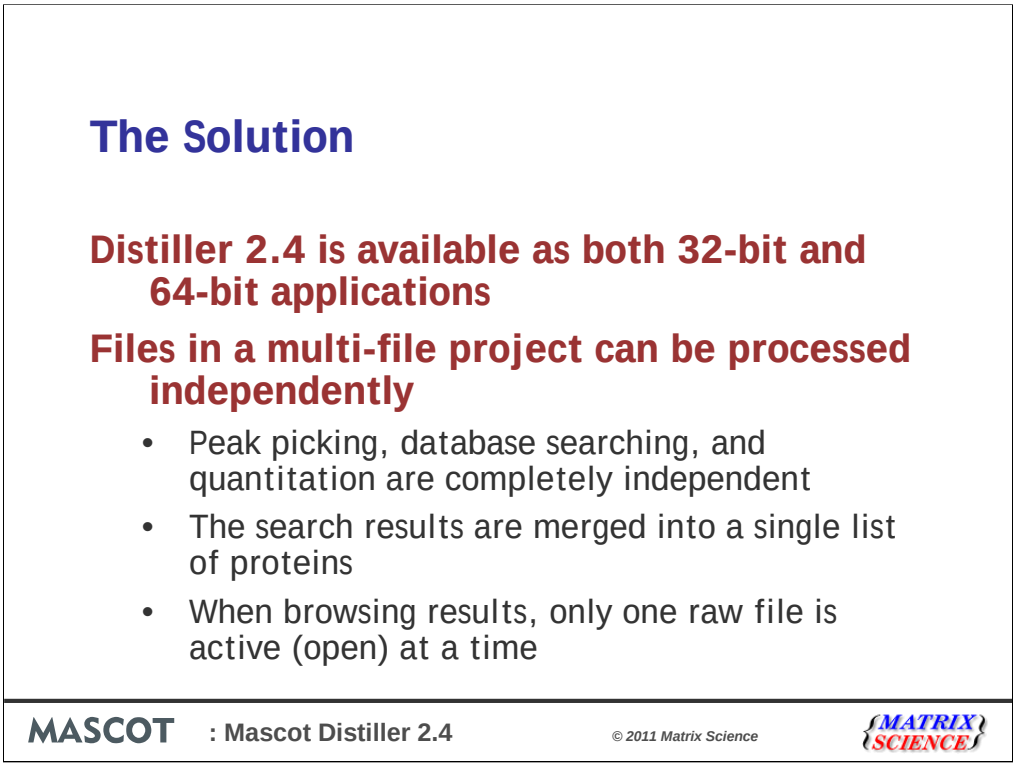
MASCOT

MATRIX SCIENCE



When we added support for quantitation to Mascot Distiller, we failed to anticipate how many people would want to process large collections of raw files as a single experiment. The original software architecture was designed to read large chunks of data into memory to make processing fast. Because it was a 32-bit application, which cannot use more than 2 GB of address space, this meant it would crash if you had a project where the total size of the raw files was more than 5 GB or so.

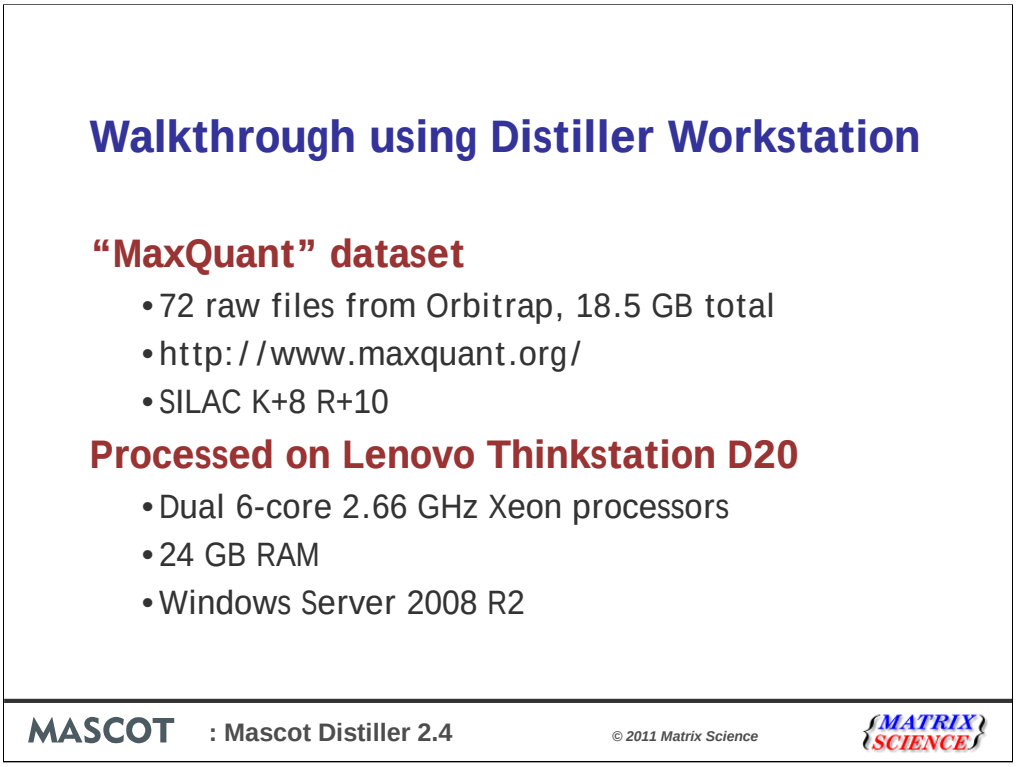
Making a 64-bit version of Distiller removes this limit but, for many people, this would mean moving to a new version of Windows and possibly new hardware. Also, unless you have a huge amount of RAM, a very large project would run out of RAM even though it hadn't run out of address space. The real fix also requires the processing to be serialised so as to keep as much of the data as possible on disk



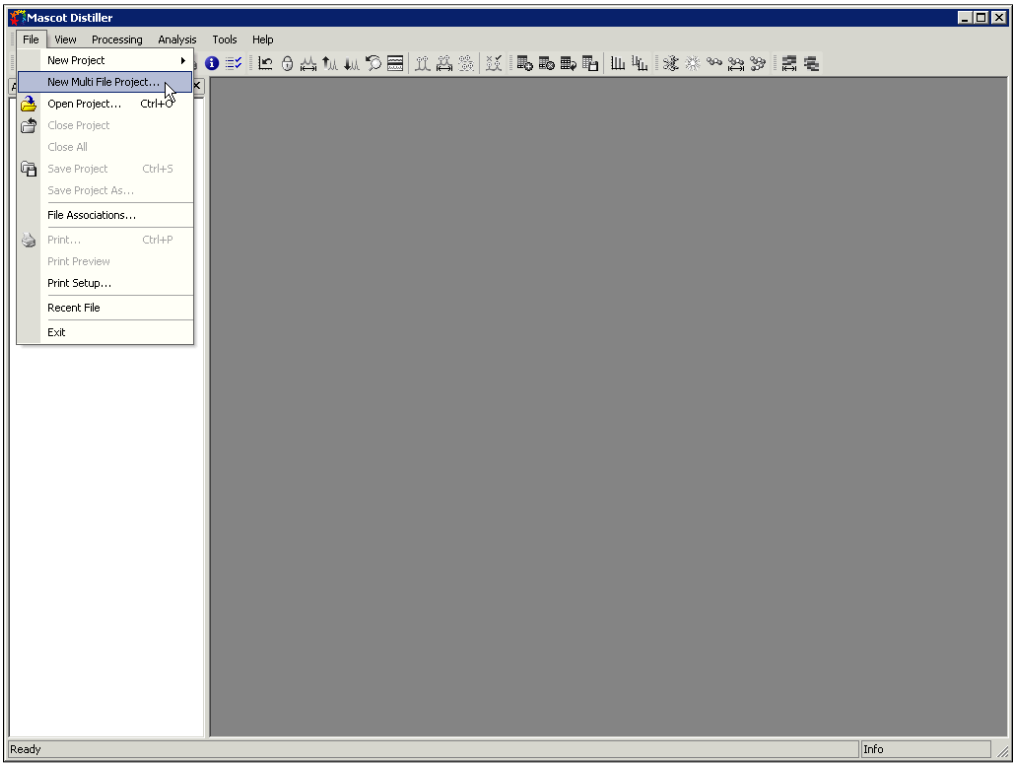
Distiller 2.4 is available in both 32-bit and 64-bit versions. This means that you aren't forced to move to a new PC immediately, although 64-bit is the future and is required for very large projects

When you create a multi-file project, the default is to process the files independently. If you want the old behaviour, to process all the data as if it was in one huge raw file, this is still available as an option. By default, peak picking, database searching, and quantitation are completely independent. The search results from the individual files are merged into a single list of proteins, because we need this to organise the quantitation results. This is the point at which you might run into problems with the 32-bit version if the data set is extremely large.

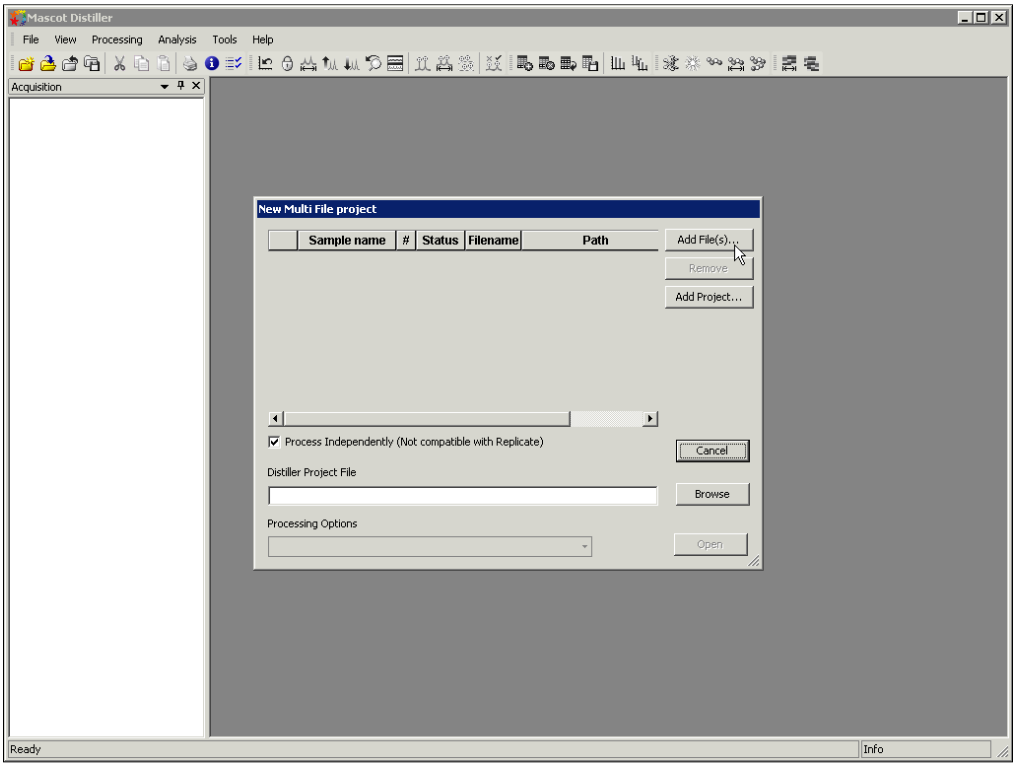
When browsing search or quantitation results, only one raw file is open at a time, so there is a slight delay when you move between files.



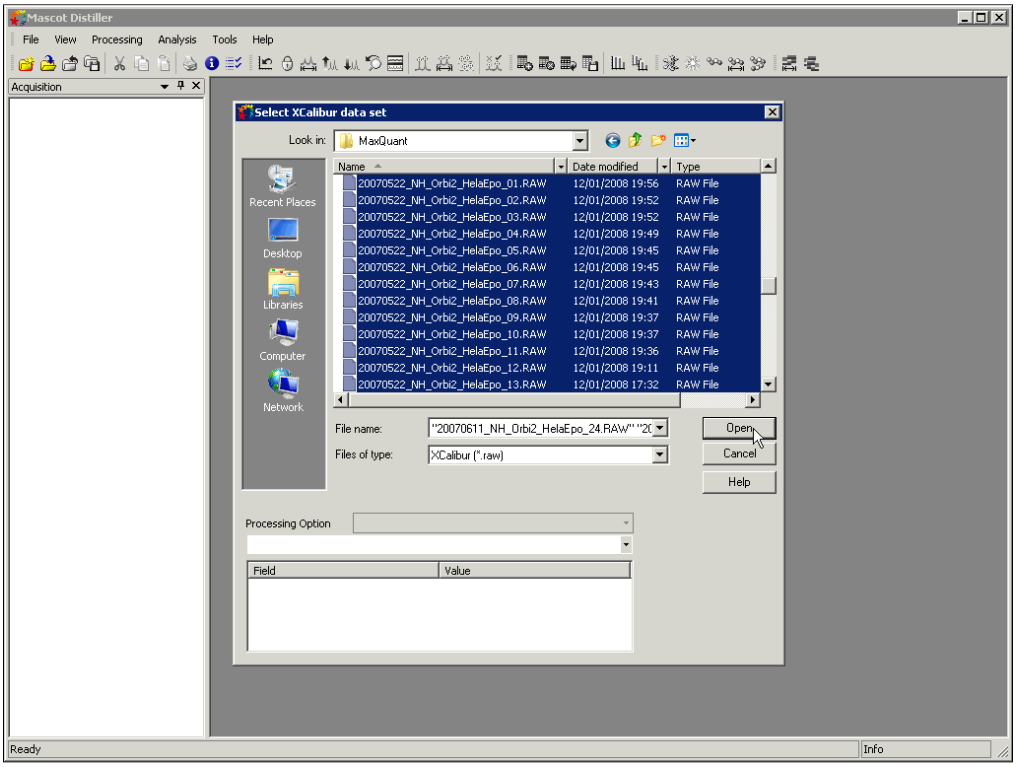
I'd like to show how it works by walking through some screen shots. The dataset is a set of public domain Orbitrap SILAC files. There is a link on the maxquant.org site to download the files from Proteome Commons. We used a reasonably high spec PC. First, lets look at processing everything in Distiller workstation



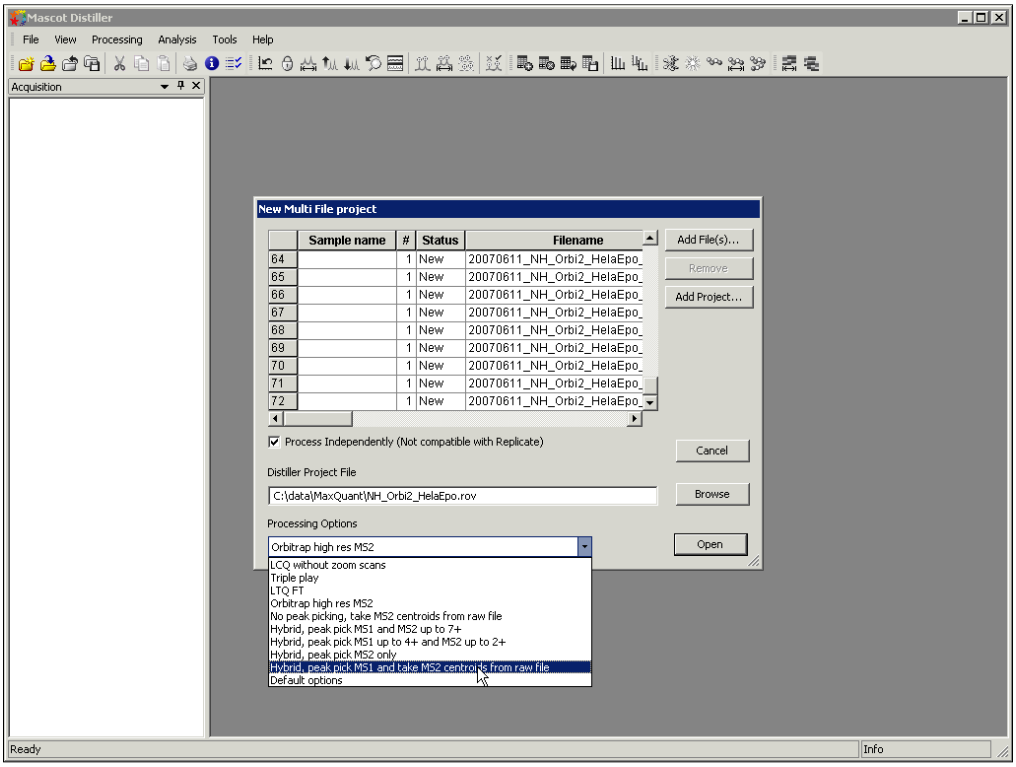
As in 2.3, we choose to create a new multifile project



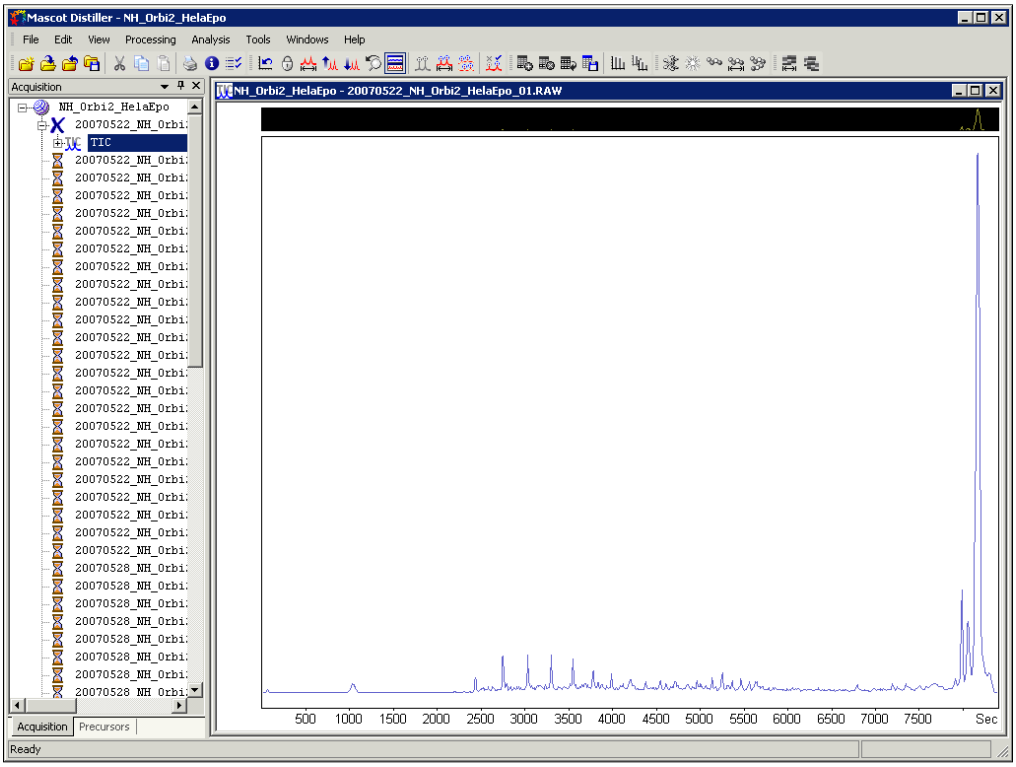
There is a change here. The dialog allows us to choose raw files or existing projects or a mixture. We'll look at using existing projects later.



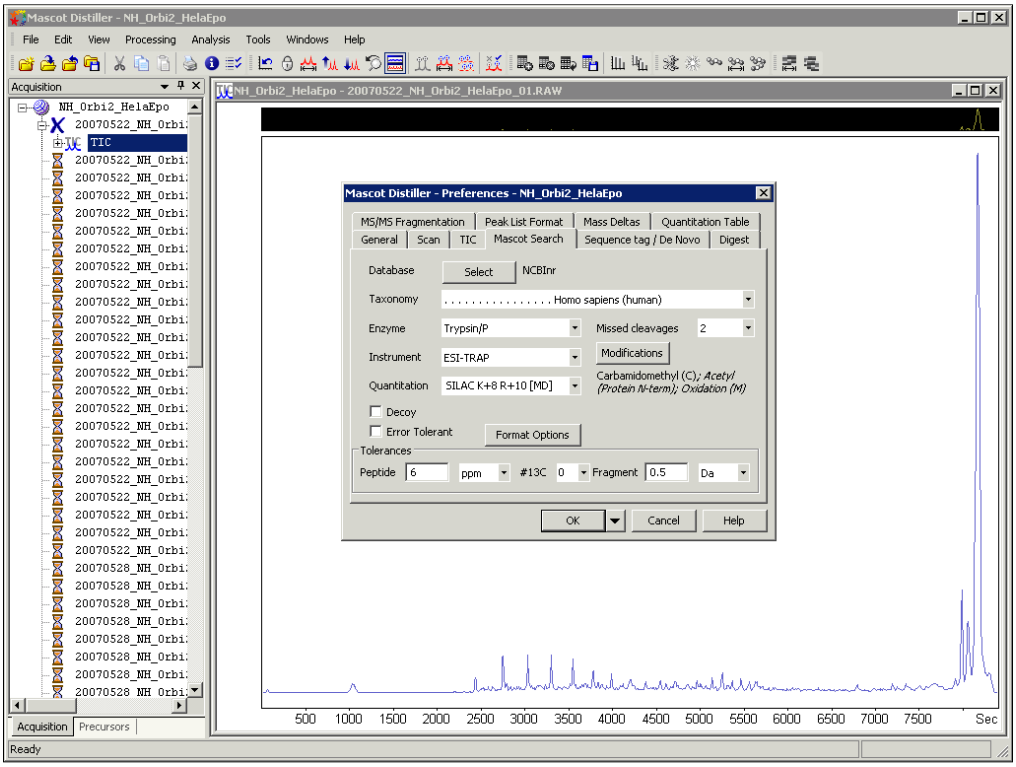
For now, we browse to the raw files and select all 72



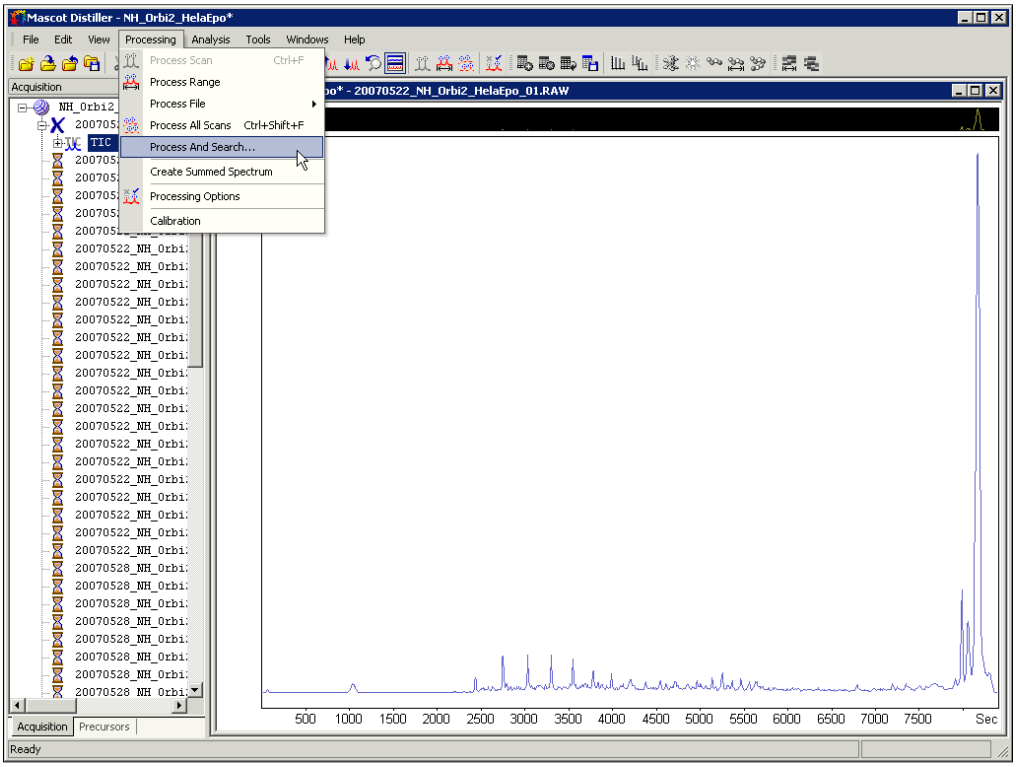
The selected files are listed. You can add further files or remove files if you change your mind. This is where you choose a name for the project and a chance to select the processing options for peak picking. When everything looks OK, choose Open



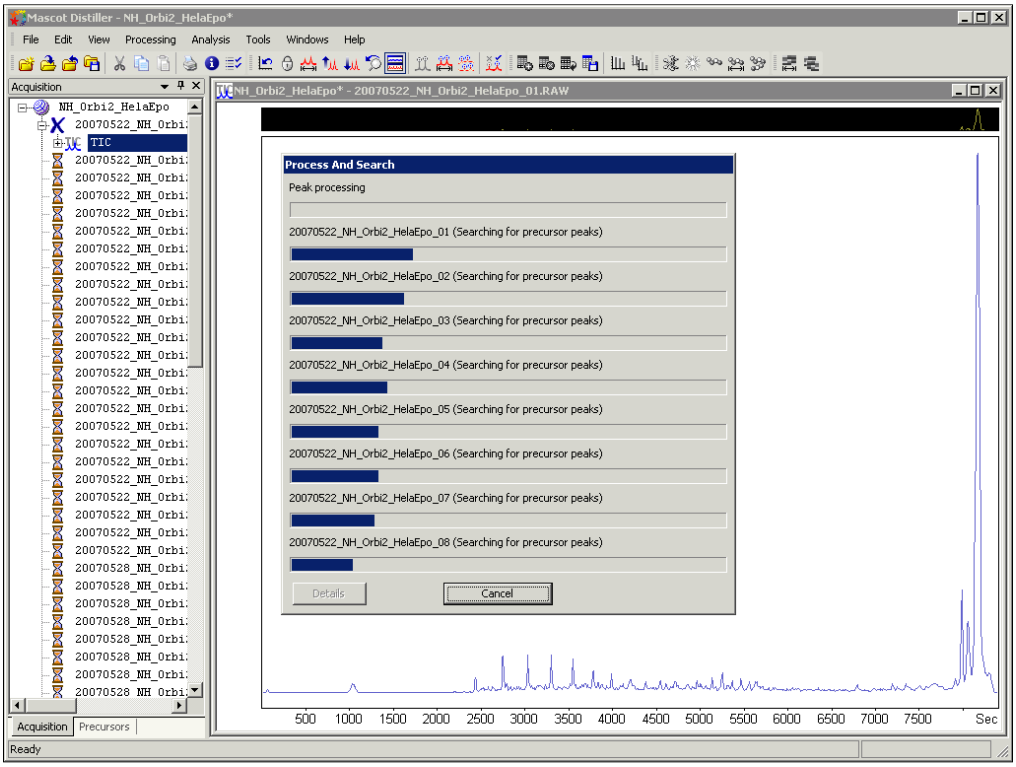
You'll notice that one file, the active one, shows on the acquisition tree as an Xcalibur icon with a TIC. The others, that are not currently in memory, show as hour glass icons. If you wanted to browse the raw data, and click on one of the files with an hourglass, there would be a few seconds delay while the first file was closed and the selected file was opened. In this walk through, we'll go directly to peak picking and database searching



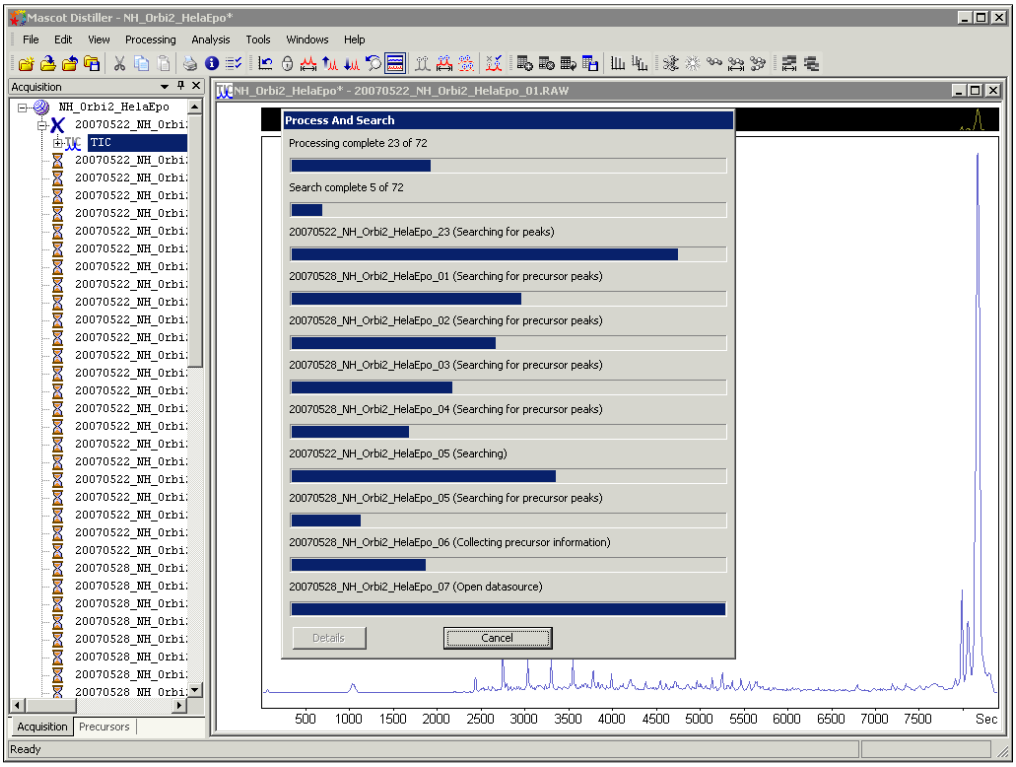
You specify the peak picking via the processing options file. For routine work, you will simply choose one that has been optimised for the type of data. The search conditions can be saved as part of the project preferences or entered into a search form at the point of submitting a search. These are the conditions we will use for this dataset



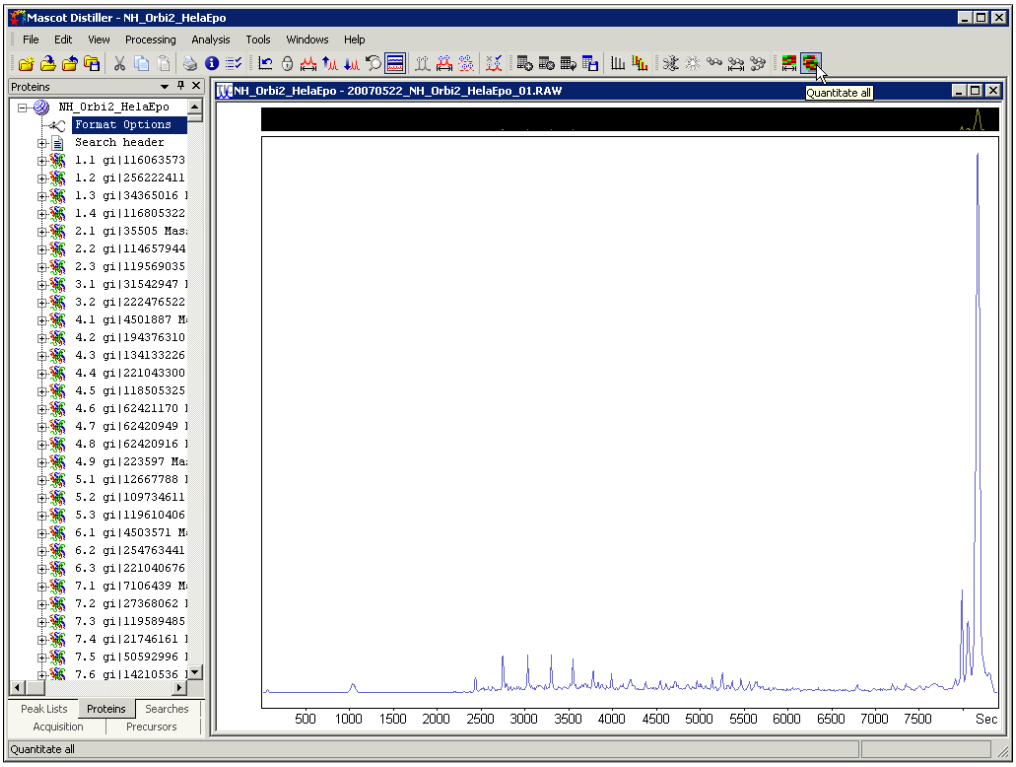
Peak picking and database searching can be performed for all files in the project by choosing ‘Process and Search’



This is where you'll see a substantial improvement in speed compared with Distiller 2.3. The code is threaded so that all the processor resources can be used. Initially, the progress dialog looks like this

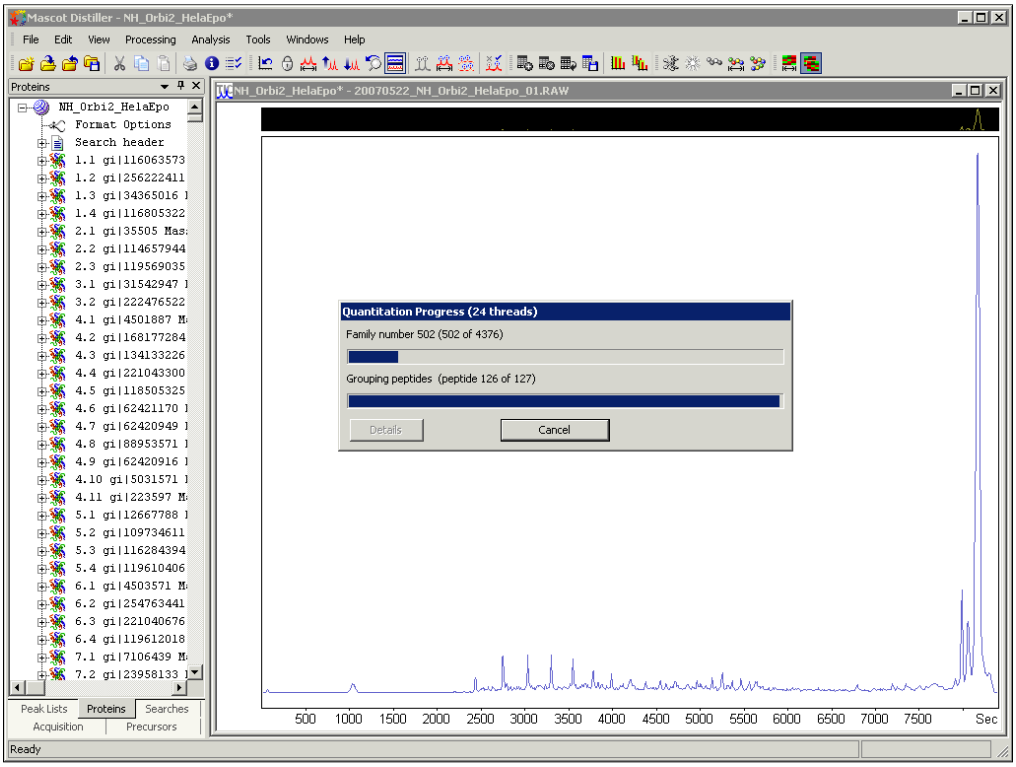


A short while later, we are about 1/3 the way through peak picking and 5 searches have been completed

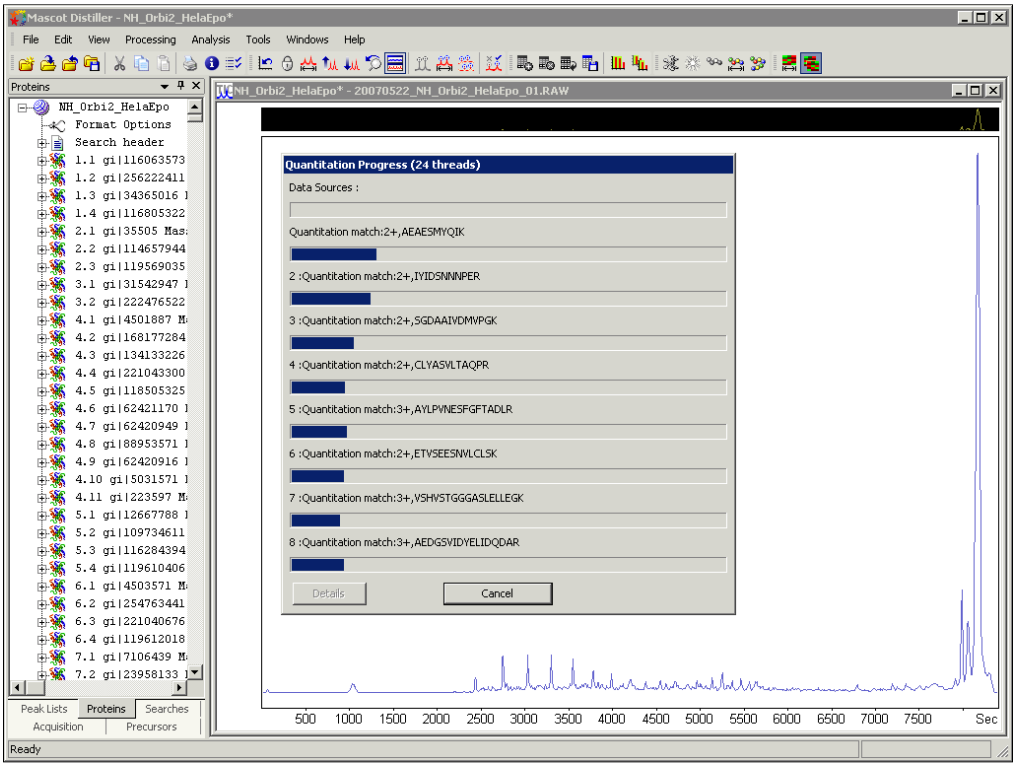


In this example, peak picking and searching all 72 files took 7 hours 45 minutes to complete. The search results have been merged into a single, minimal list of proteins. You'll notice that the proteins tab now uses the new family grouping, introduced in Mascot 2.3. This is particularly useful for quantitation because it ensures that proteins related by shared peptide matches are displayed together, making it easy to spot whether a particular peptide match should belong to one isoform rather than another. Family grouping is an option; you can choose the Select Summary-style list if you prefer it.

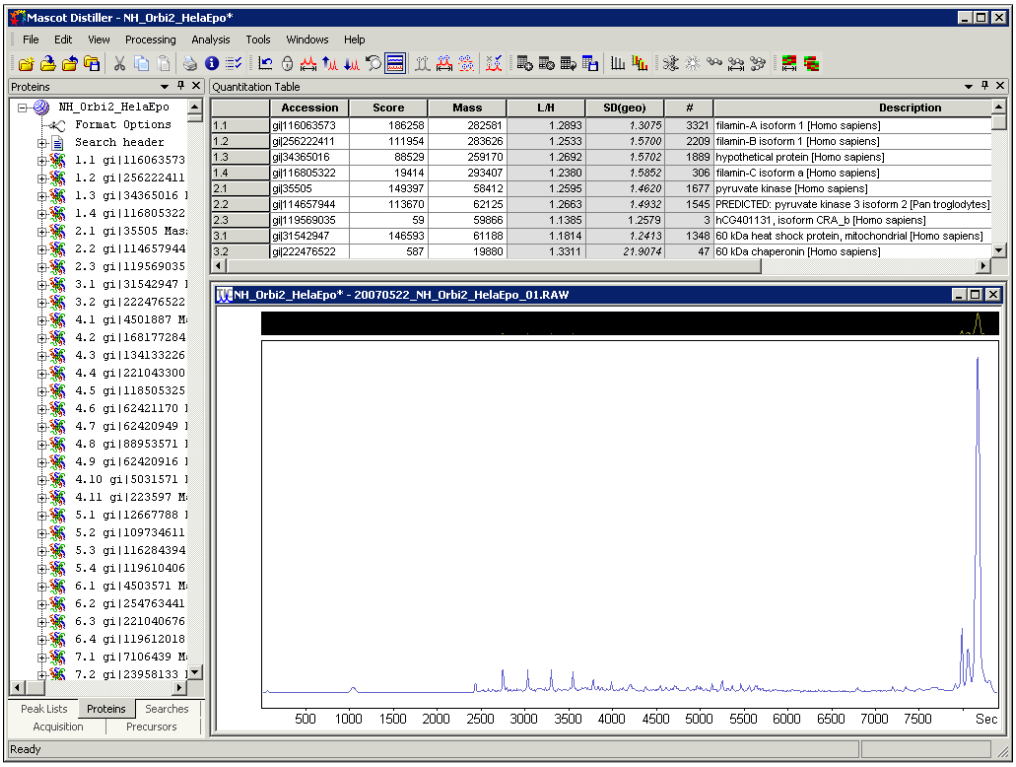
The next step is quantitation. You can process some or all of the proteins. We'll choose all.



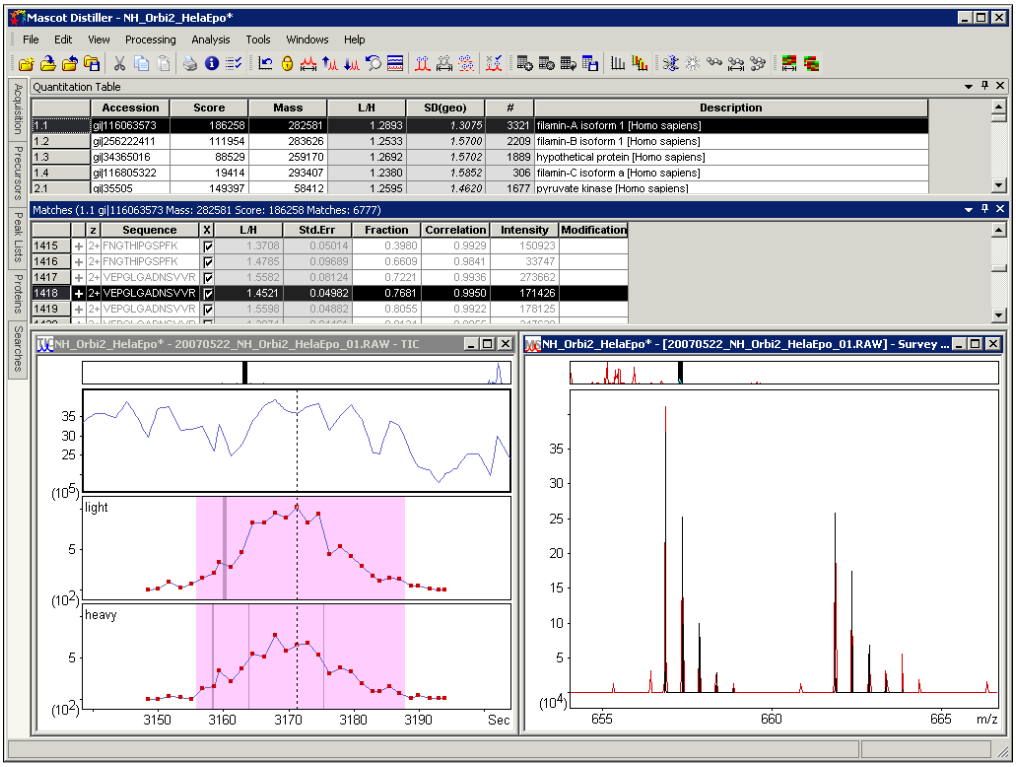
The first step is to collect together peptide matches from all the search results that correspond to the same sequence.



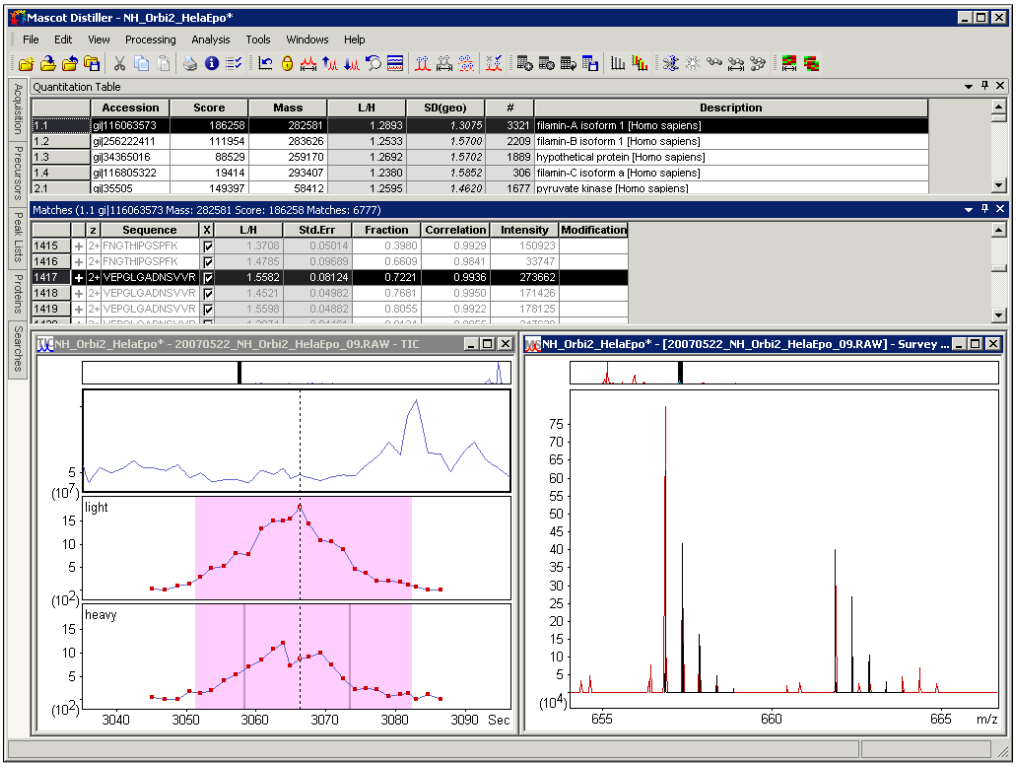
Once quantitation gets going, it also uses multiple threads. This system has dual 6-core processors, and each core is hyperthreaded, so 24 threads are used. Even so, it takes some 22 hours to quantitate all 4376 proteins



When complete, it looks much the same as before apart from the family grouping. There is a change in the way the quantitation table is displayed. You can now choose between having peptide rows indented in the main table or displaying peptides as a separate, linked table.

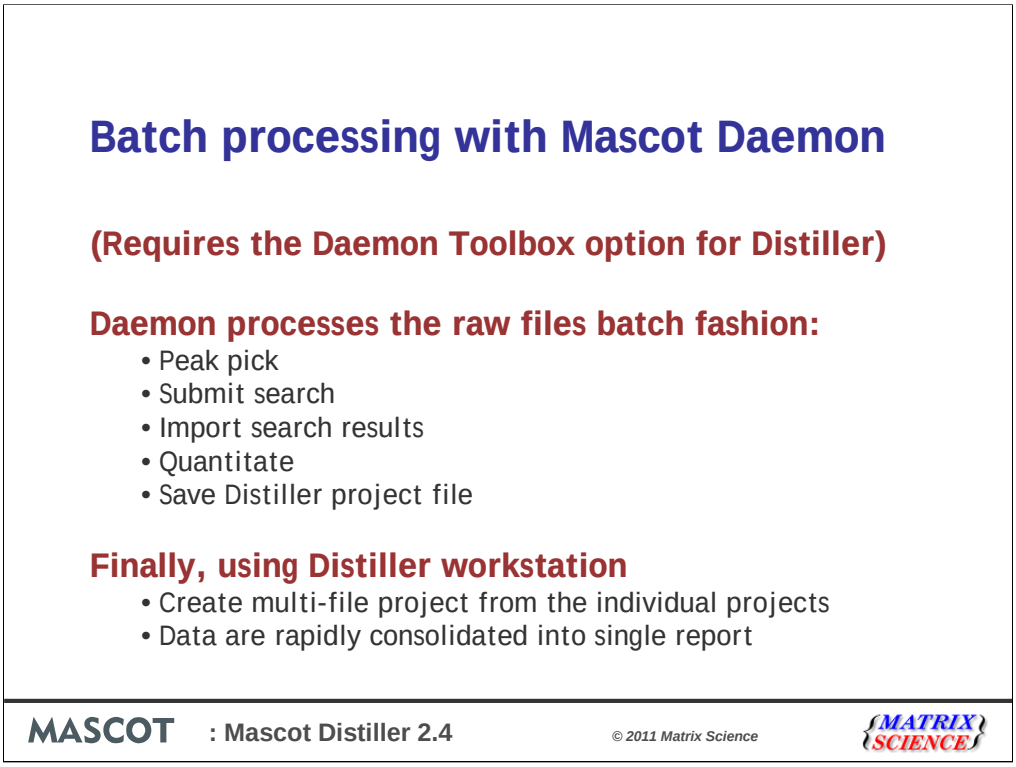


The linked table, as shown here, usually works better for very large tables. You'll notice that the peptide matches adjacent to the selected one are in grey. This is a visual clue that they come from a different raw file, and if you click on one of them, there will be a short delay while the files swap over. Peptide 1418 is from 01.raw

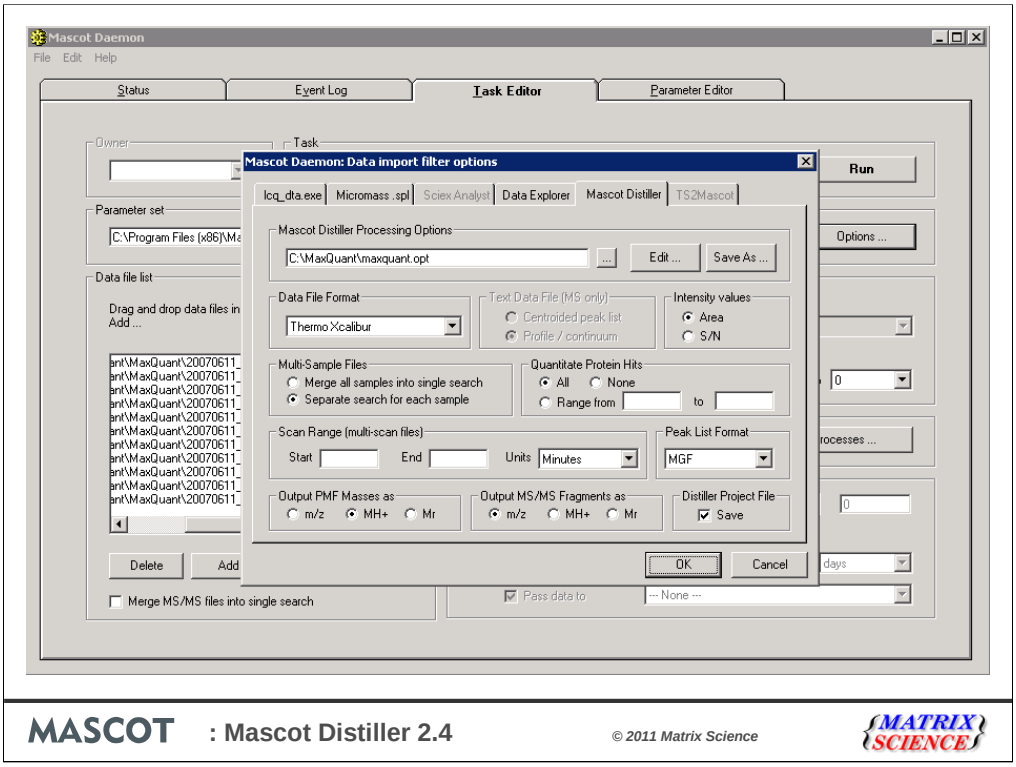


While 1417 is from 09.raw

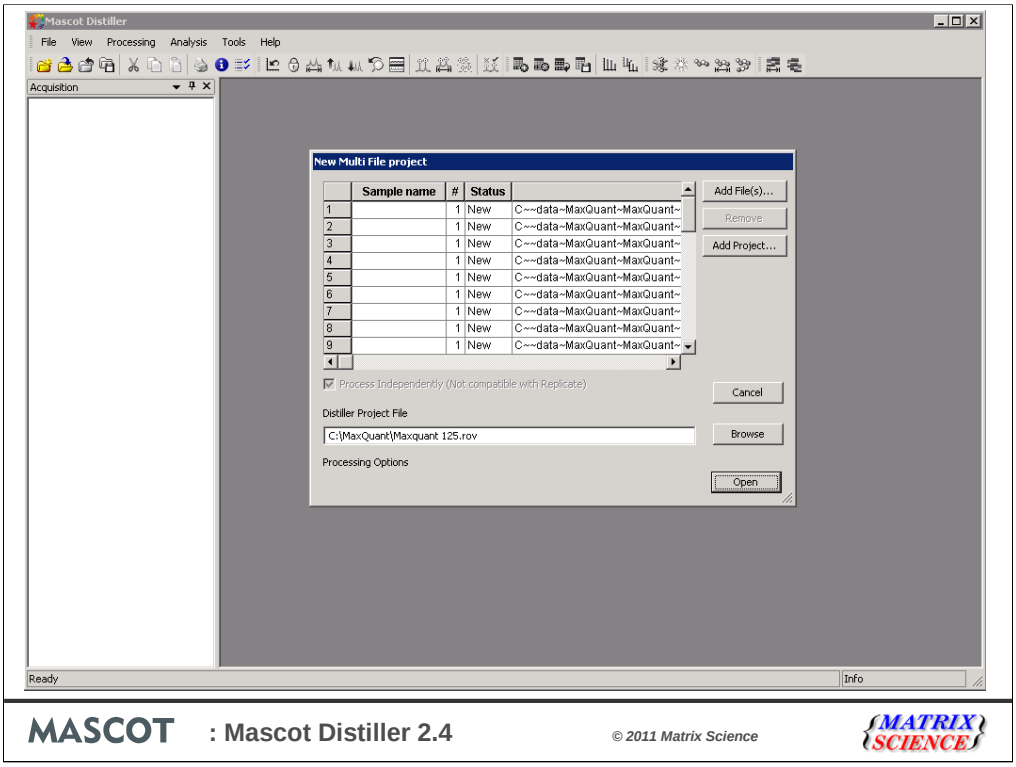
After you save the project, its reasonably fast to reopen. This particular one takes just under 3 minutes to open from disk. Remember that, if Distiller is not registered, it operates in viewer mode, so this is a powerful way of sharing search and quantitation results with colleagues



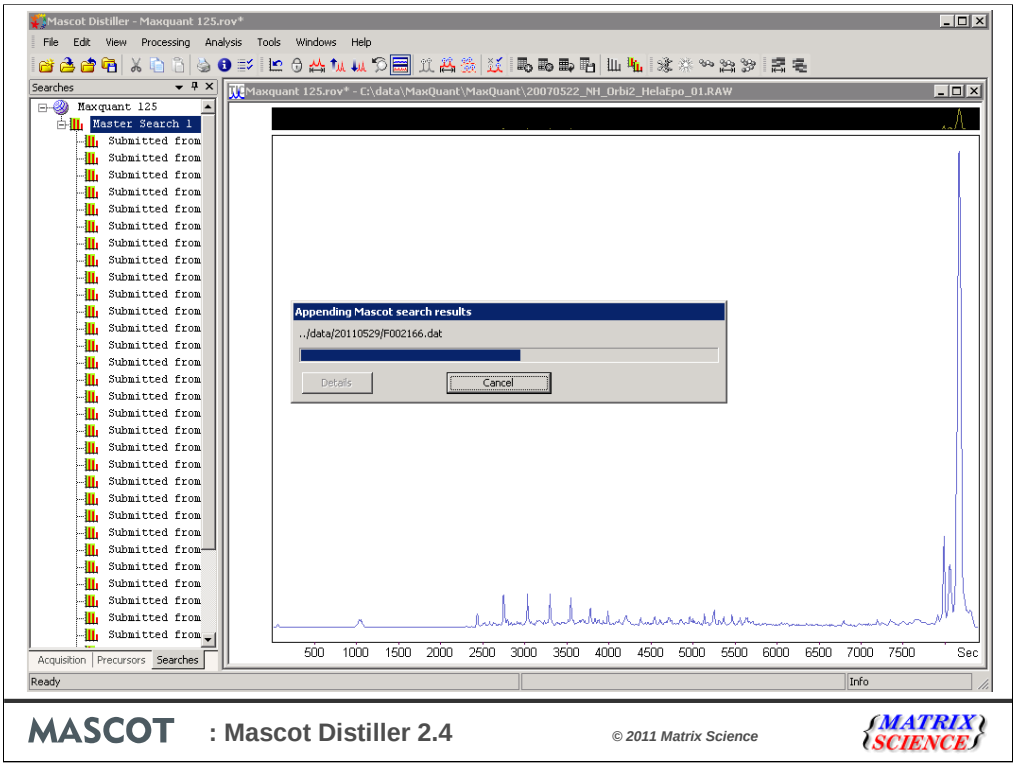
The other workflow is to use Mascot Daemon to batch process the individual files. This requires Distiller to include the Daemon Toolbox option, so that Distiller can be called by Daemon. Daemon automates all of the processing steps, from peak picking to quantitation, and saves each project to disk. You can then open the set of projects in Distiller Workstation to create a multi-file project where all the data processing for the individual raw files has already been completed



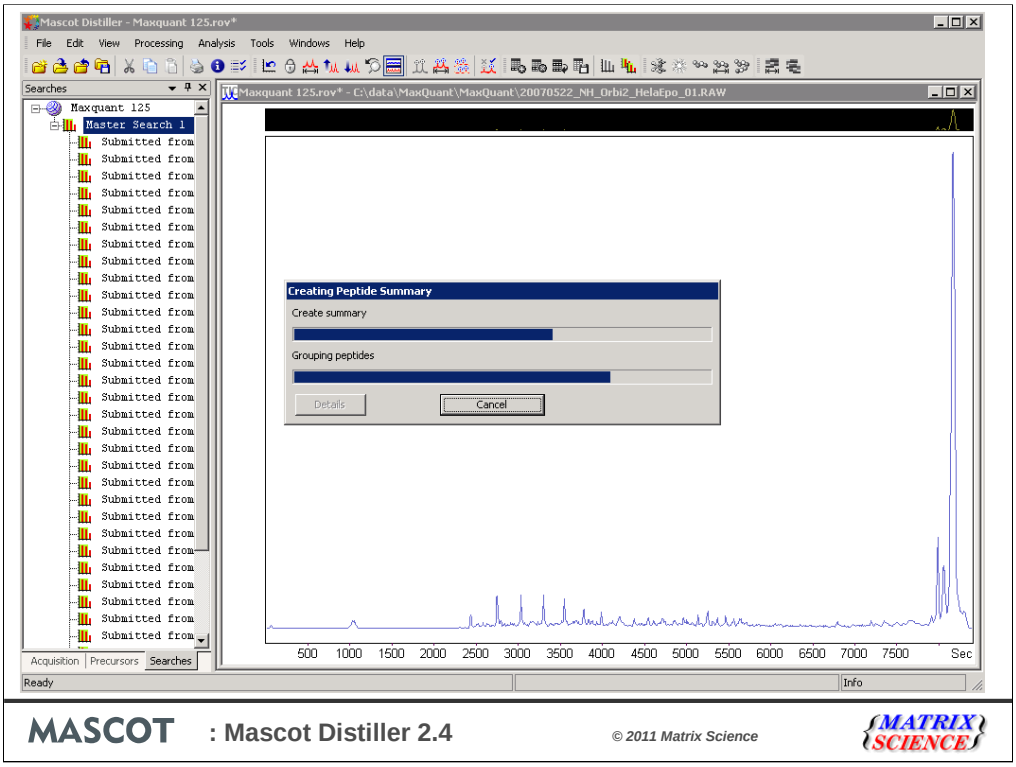
Daemon is great for routine work. You don't have to remember any settings. Just clone a previous task and change the list of files. Make sure the box is checked to save the project and choose to quantitate all hits, because there is no guarantee that hit 10 in an individual file will be hit 10 in the merged search results



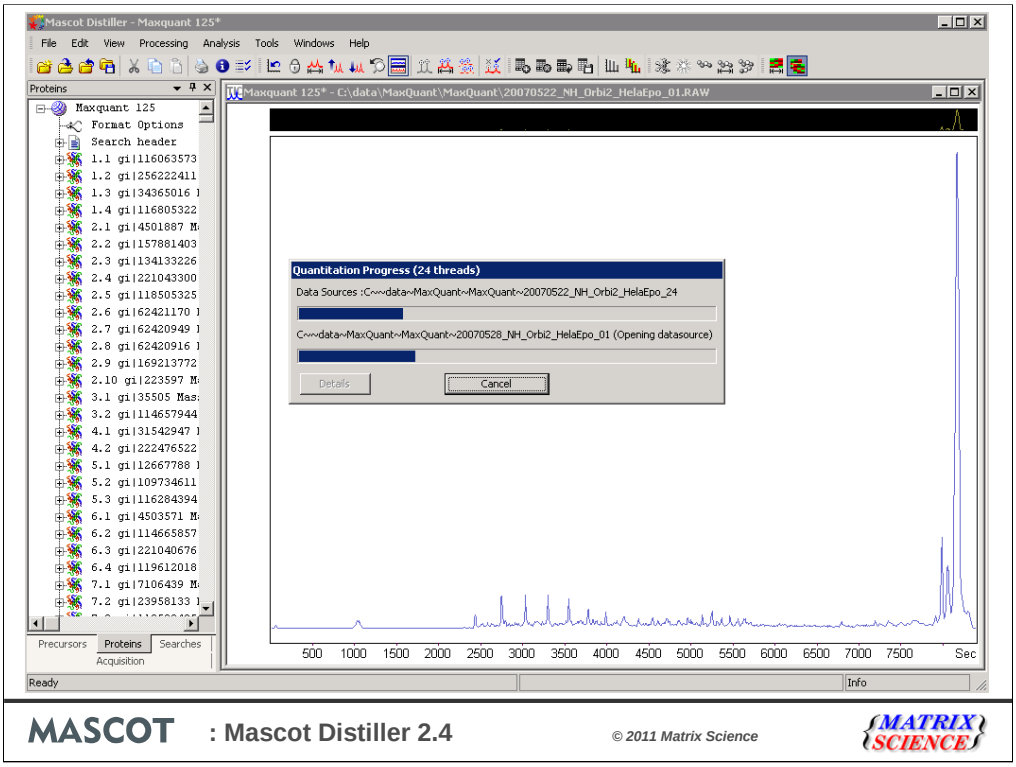
Once the Daemon tasks are complete, we can select the project files for the multi-file project



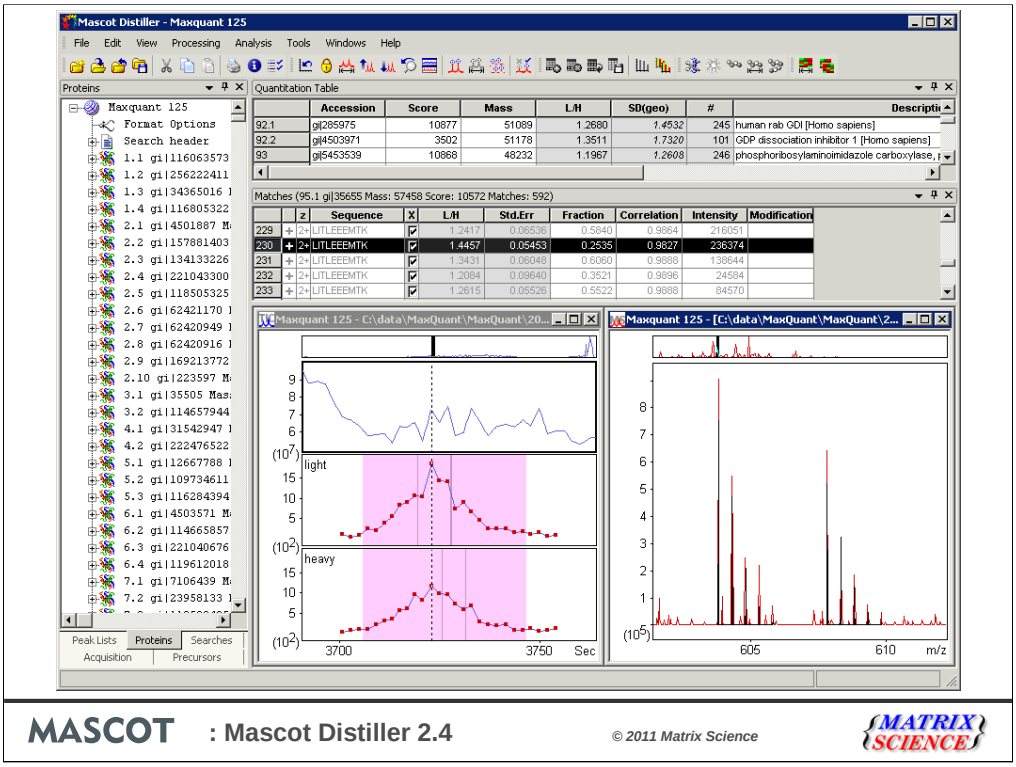
You still choose Process and Search. This checks that the peak picking and search conditions are identical across all the projects, then extracts the search results. If any of the projects used different peak picking or search settings, they would be re-processed and/or re-searched during this procedure. As would any raw files that had been included.



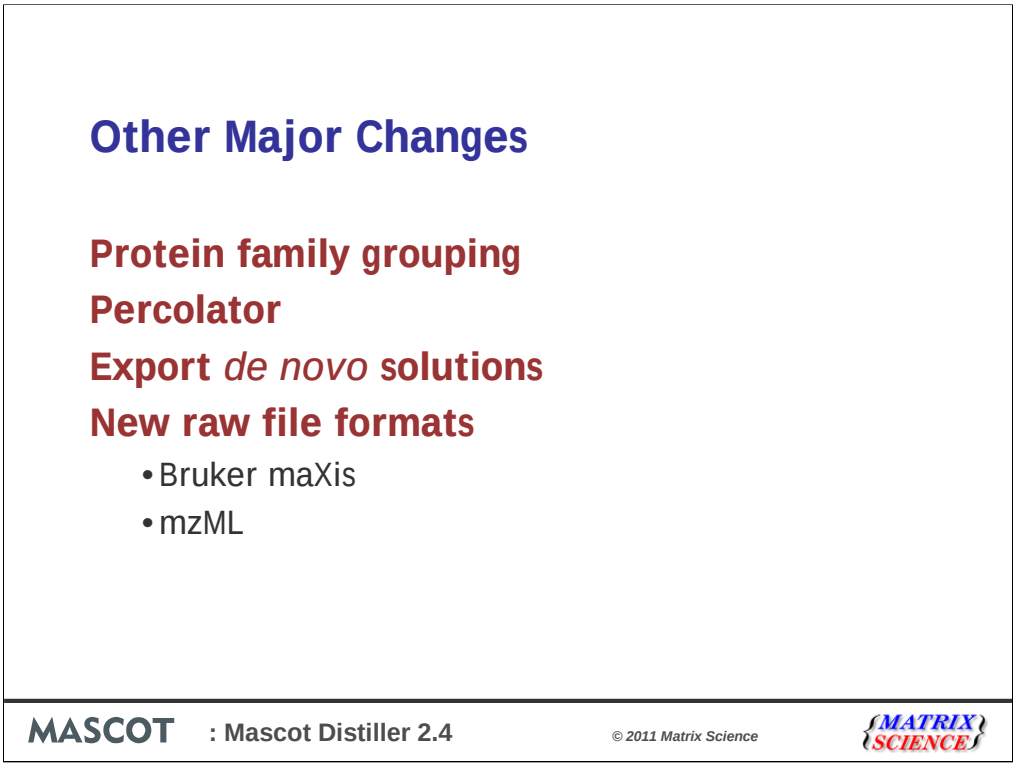
Creating the combined peptide summary is the most time consuming step



Quantitation is just a case of extracting the existing data from the individual project file



The final result is exactly the same as if we started from raw files, but the time taken for this particular example was under 2 hours. You can easily remove projects or add new ones, as long as they were processed using identical settings. This is ideal for experimenting with replicates. You can look at the results for the individual replicates and the combined results without having to start from scratch every time.



Besides fixing the memory and speed problems for multi-file projects, there are a number of other new features. The 2.4 release will bring Distiller back into line with a couple of features that were new in Mascot Server 2.3: protein family grouping and re-scoring using Percolator. (Although, you cannot use Percolator for multi-file projects.) By popular request, you can now export de novo solutions as an XML file. Some of the data format libraries have been updated, including the Bruker libraries, so we can finally open maXis data.

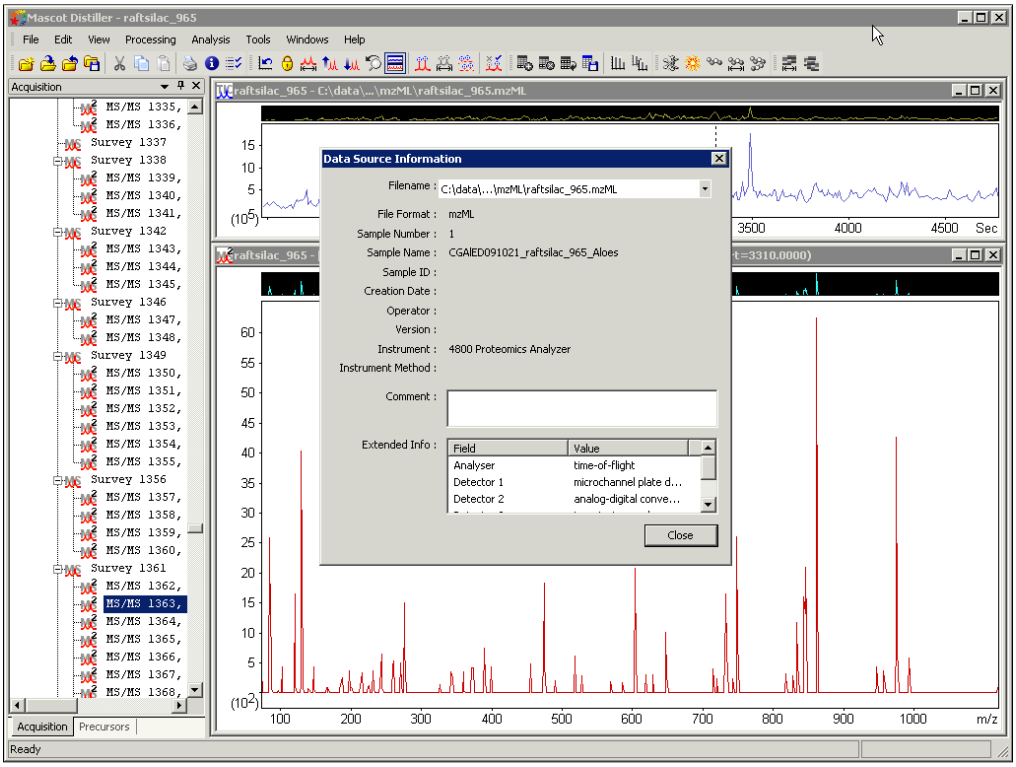
The ability to open mzML files is particularly important for AB Sciex TOF-TOF data. Previously, there was no easy way to work with this data in Distiller because it is stored in tables in an Oracle database.

mzML

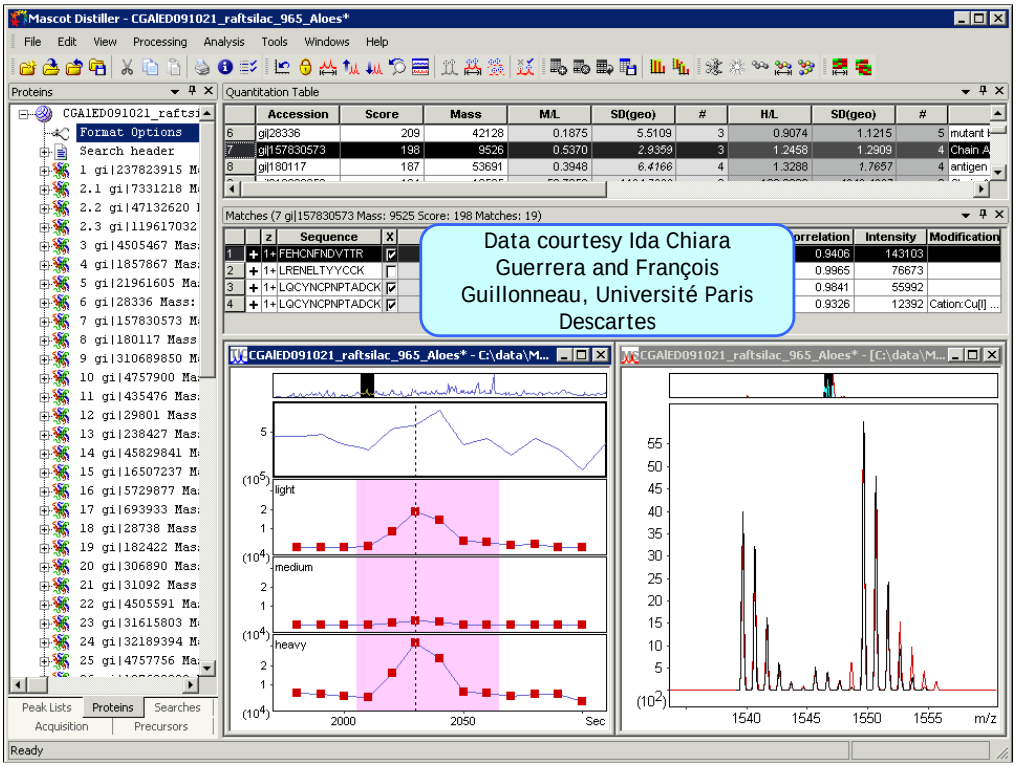
AB SCIEX MS Data Converter

- Converts TOF-TOF and Wiff to MGF and mzML
- Command line utility
- Still in beta test

Recently, AB Sciex developed a utility that can export a spot set as an XML file, with conventional parent child relationships between the MS and MS/MS scans. This is still in beta test, so I can't provide a download link at this time.



Here is a small mzML file of 4800 TOF-TOF data, courtesy of Ida Chiara Guerrera and François Guillonneau, Université Paris Descartes. It looks just like conventional LC-MS/MS data when opened in Distiller



Which opens up the possibility of quantitation using precursor protocol methods such as the SILAC experiment shown here

Let me anticipate the first question: When can I get it? Currently, we are in beta test. All being well, beginning of July