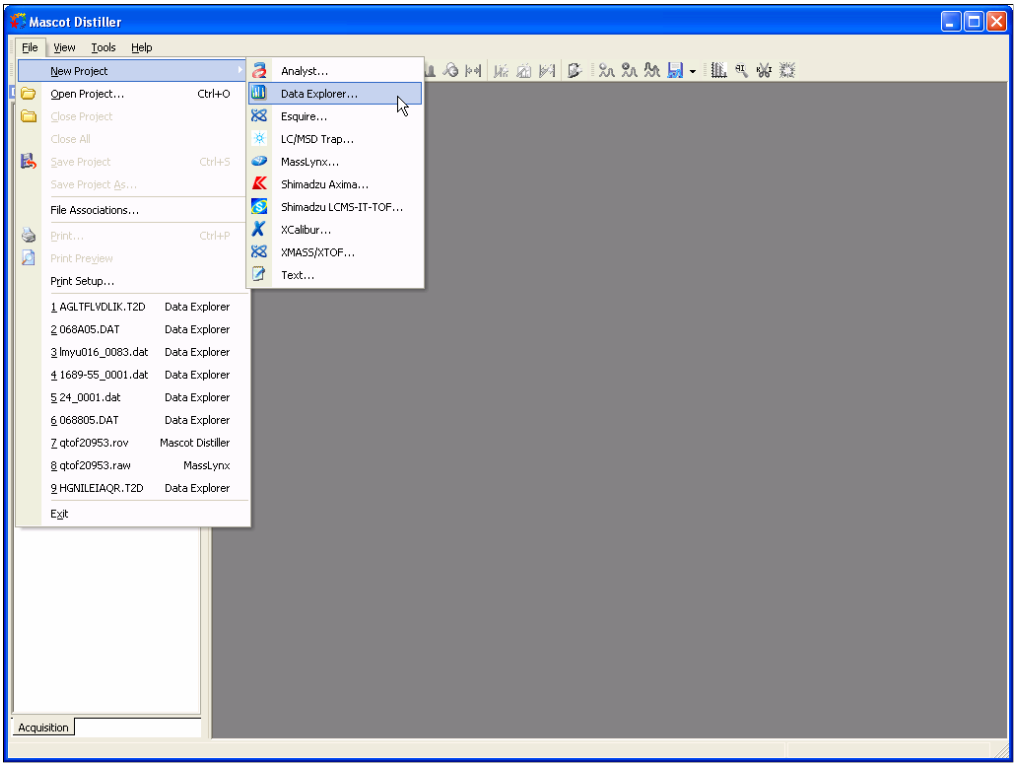
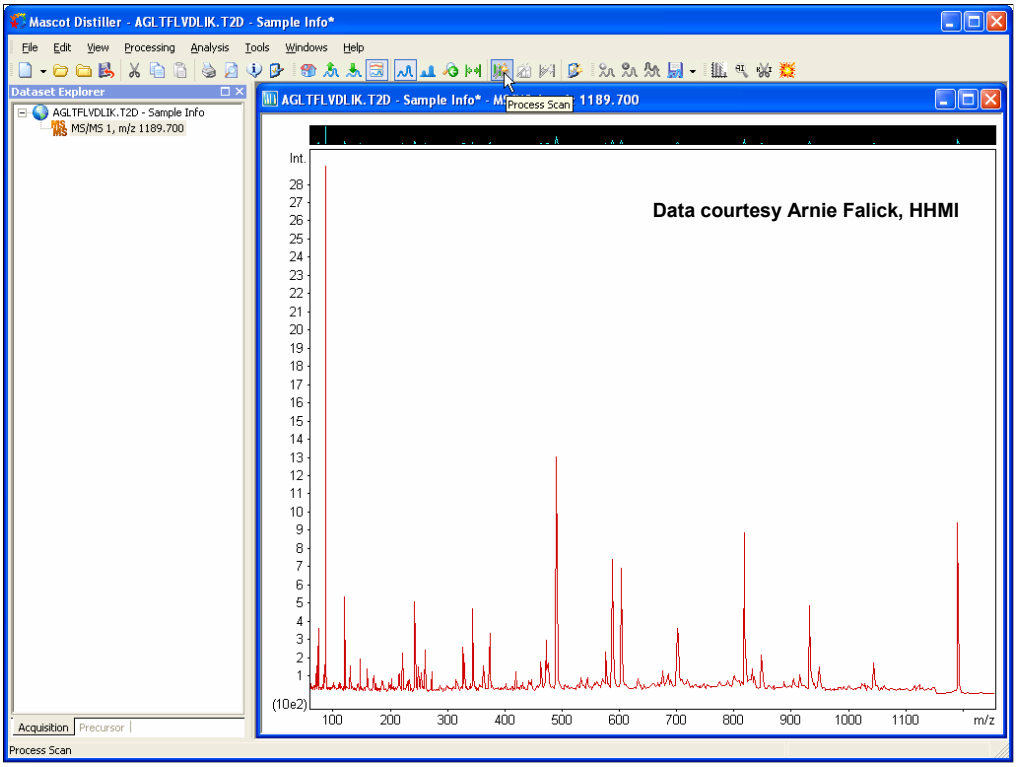
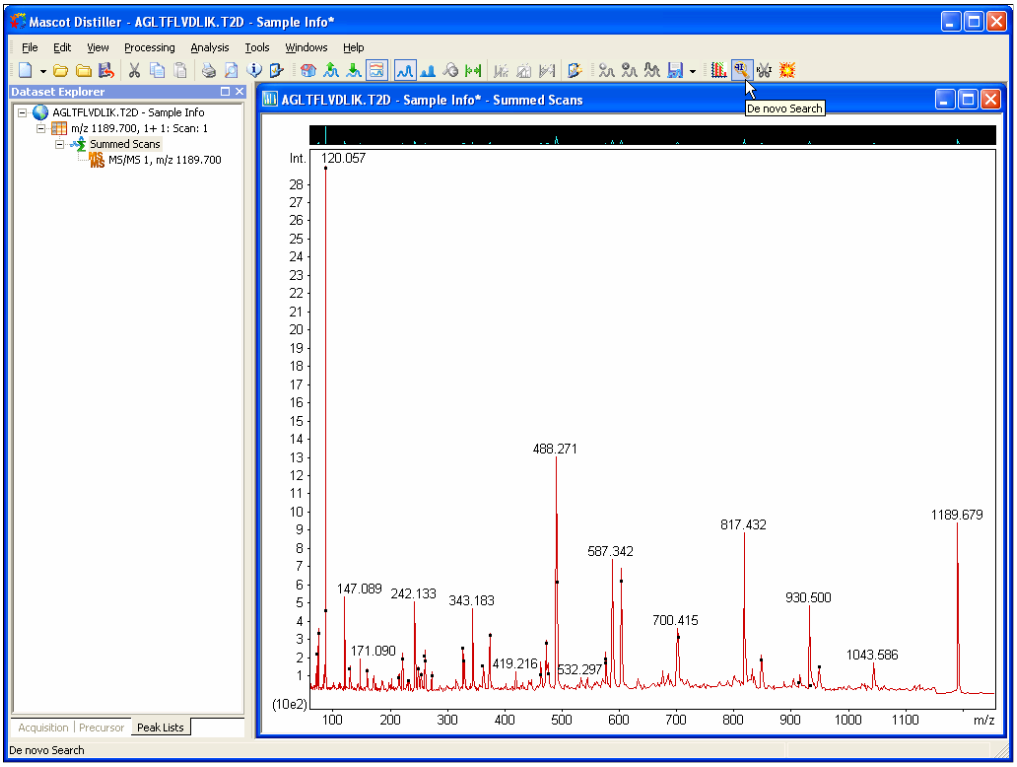
The forthcoming release of Mascot Distiller has many new features. One of them is de novo peptide sequencing. I'd like to show you how easy this is to use



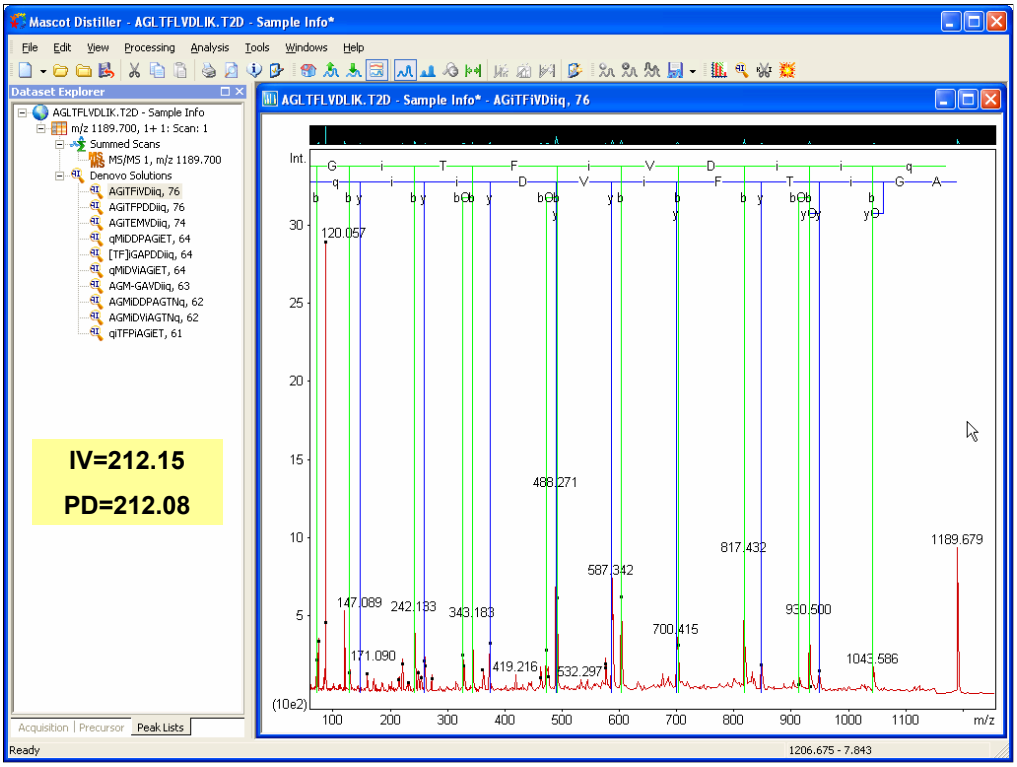
We open a raw data file. For this example, it is an Applied Biosystems 4700 file, courtesy of Arnie Falick at HHMI



Press the tool button to process the scan into a peak list



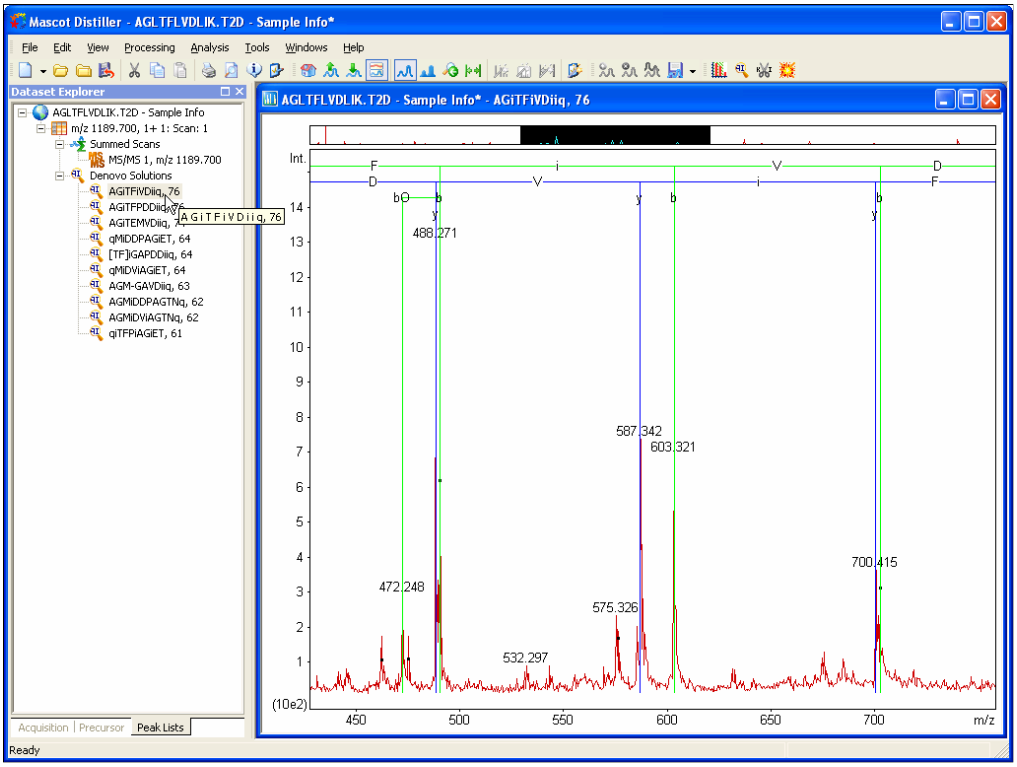
Then press the de novo tool button



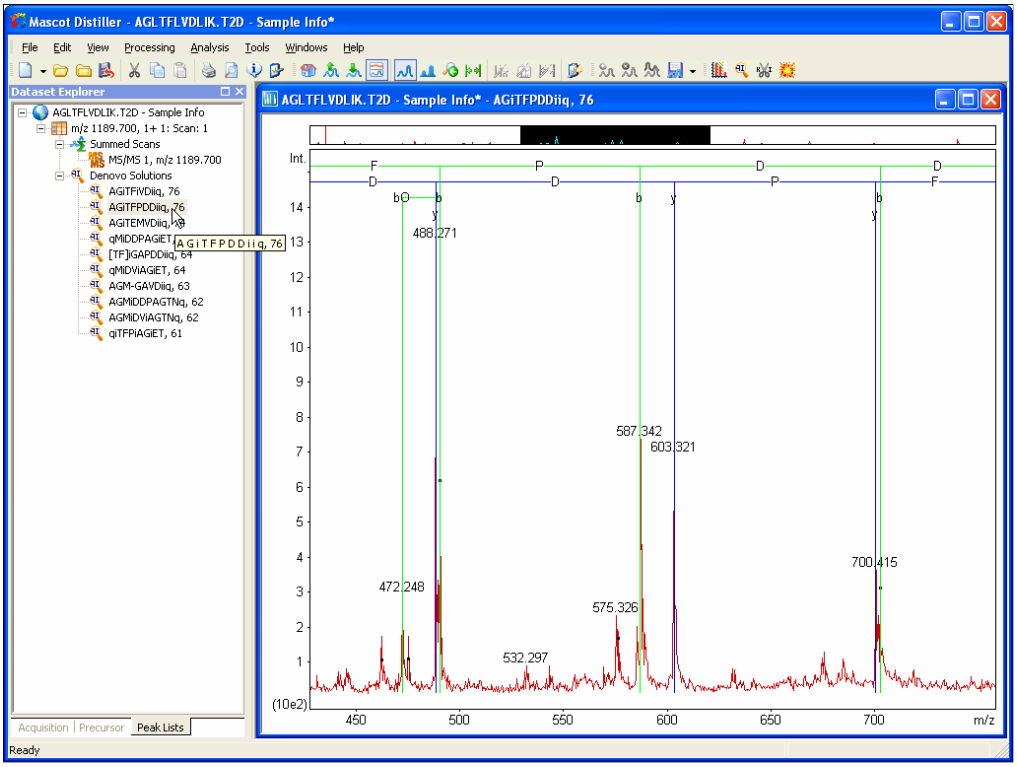
Which shows us the top 10 best solutions. The sharp eyed amongst you will notice that the sequence is in the file name. I'm not going to try and pretend this is an unknown.

You'll also notice that the top two de novo solutions have the same score, 76, and very similar sequences. The only difference is that one has IV in the middle and the other has PD, both of which are the same mass to one decimal place

By the way, the lower case i is used for I or L, the lower case q is used for Q or K



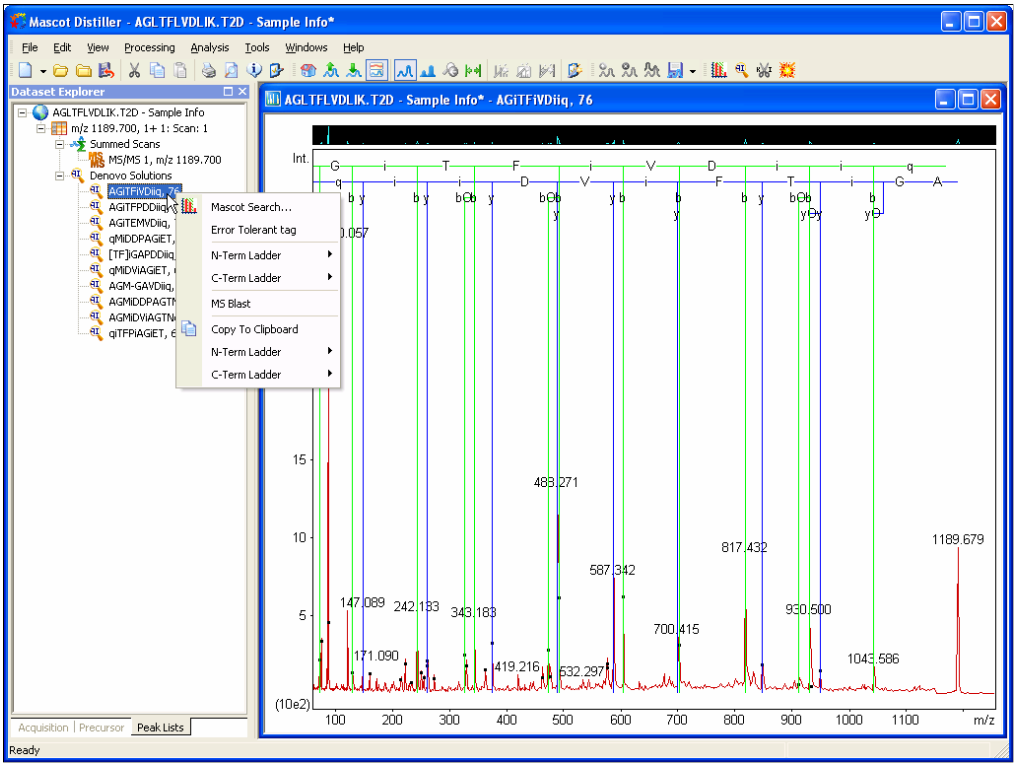
If we zoom in a little, here is the mid-region of the “correct” match



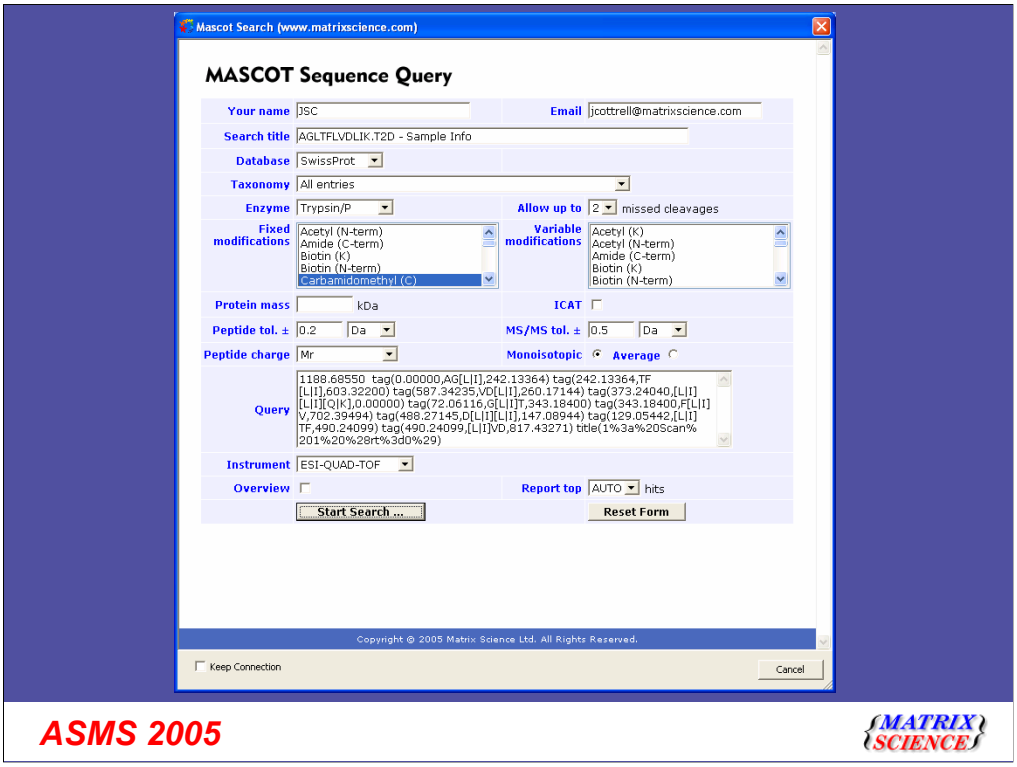
And here is the mid region of the second match. The b and y ion matches for these two peaks simply swapped over.

This is a fact of life with de novo. The search space of all possible peptide sequences is so large that, most times, we are left with some ambiguity. Its hard to blame the algorithm. These two solutions are equally valid.

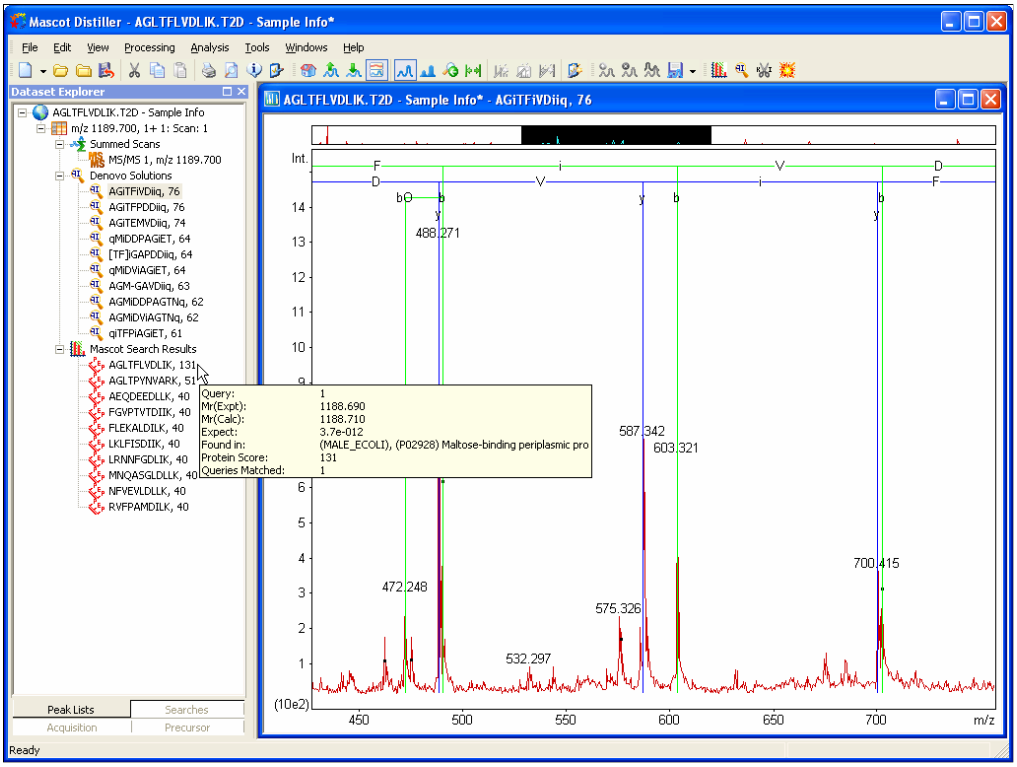
If the data quality is very high, maybe we get something close to a complete peptide sequence. With average data, all we can hope for is to get a partial sequence. The next step might be to take this partial sequence and search it as a sequence tag or using a some kind of sequence homology software.



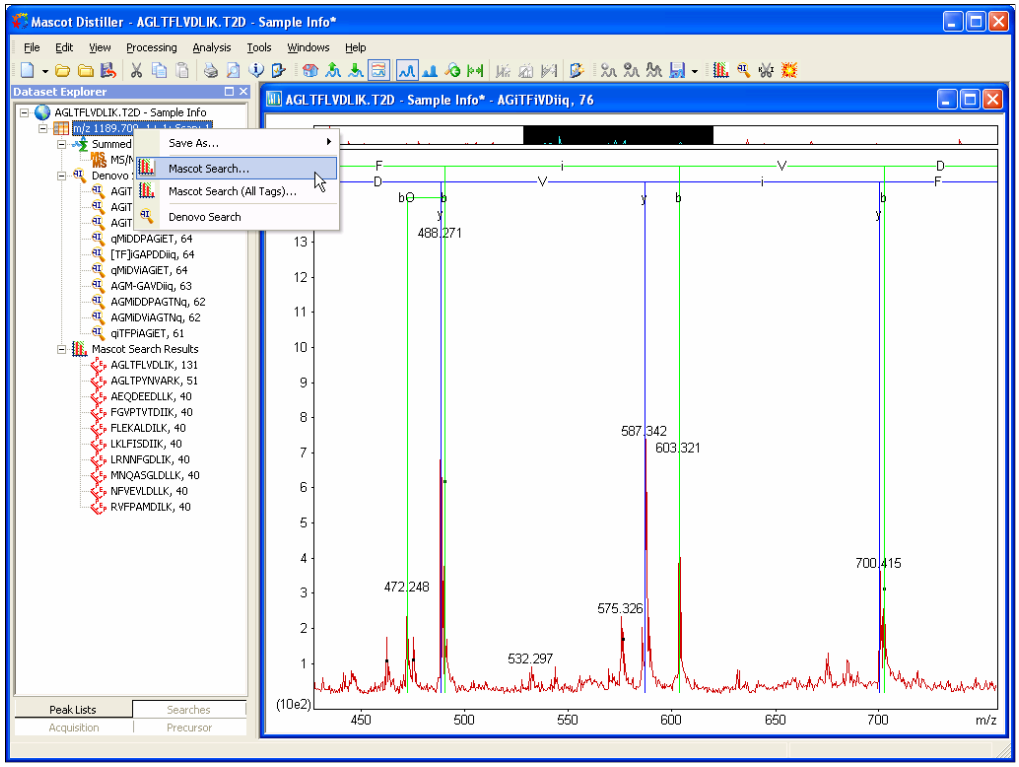
By right clicking on a solution, we get a context menu. If we choose Mascot search …



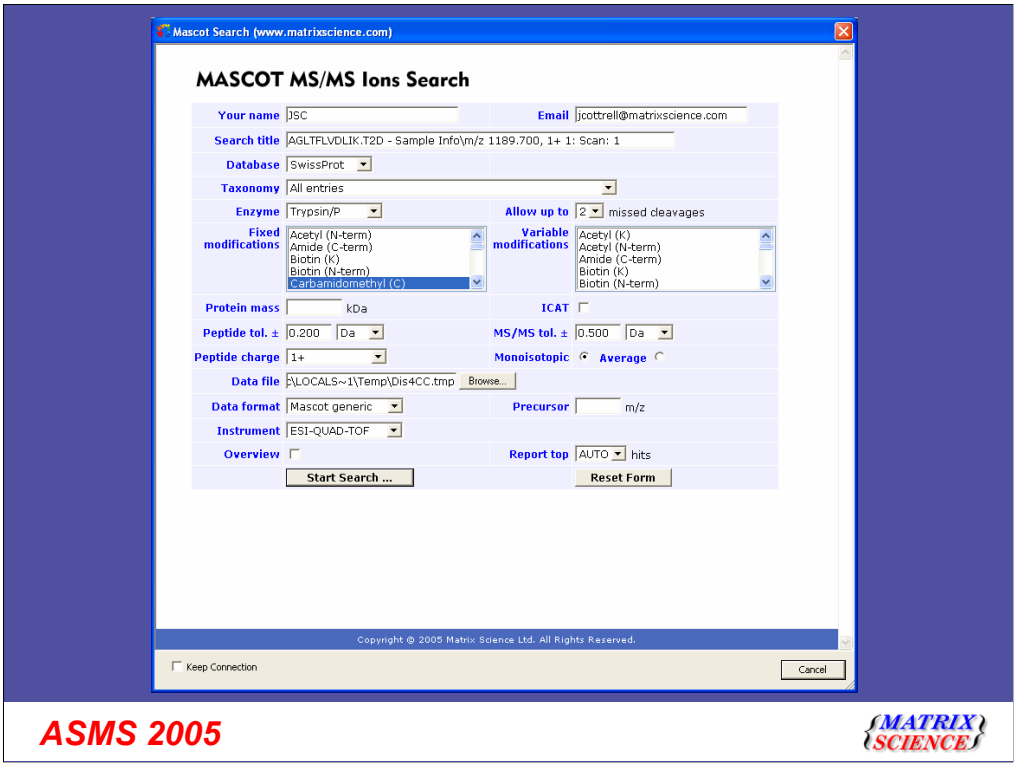
Distiller enters a list of all possible 3 residue sequence tags into the form. The idea is that some of these will be correct, but others may be incorrect. A Mascot tag search scores tags on a probabilistic basis, it does not require all of them to be correct. When we press submit and run this search, Distiller will import the results automatically.



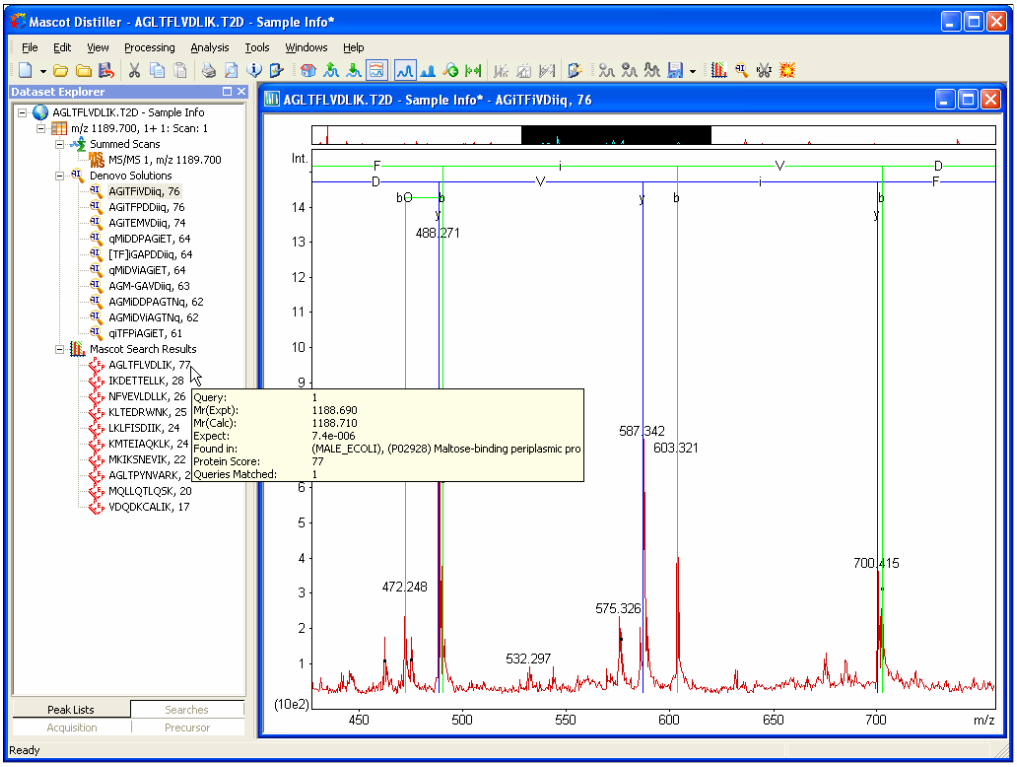
The results are displayed on the explorer tree as shown here. Only the first sequence is actually found in the database. The score is high because, in this particular case, the sequence was entirely correct so all of the tags were correct



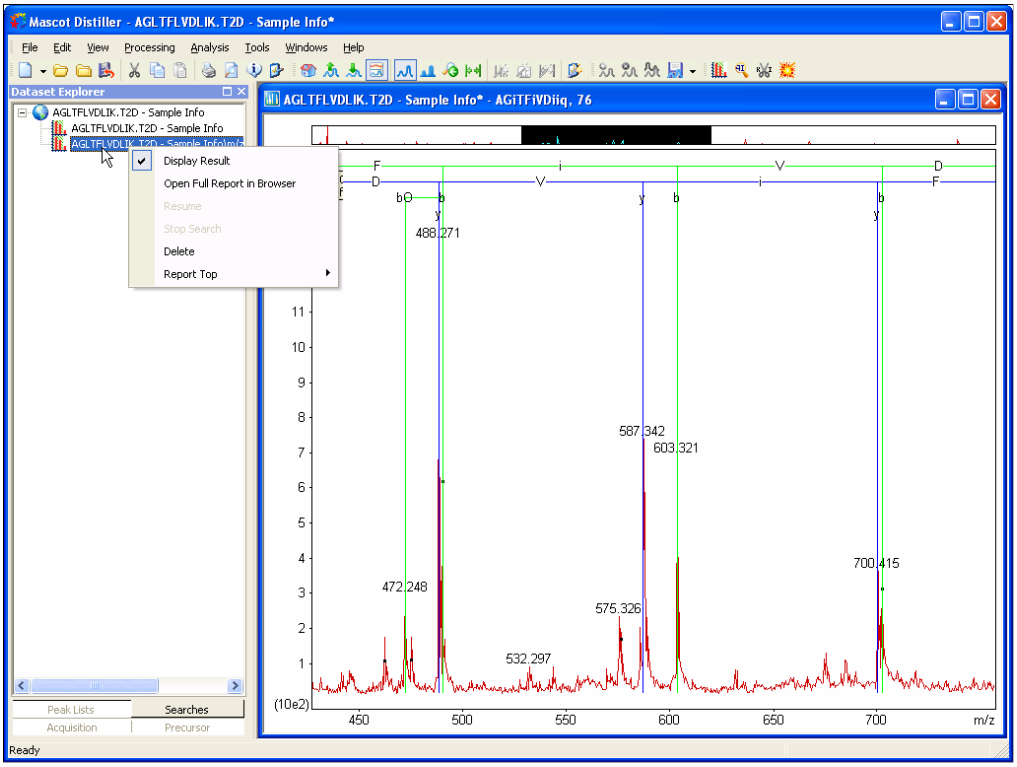
Alternatively, if we right click the peak list node and choose Mascot search off this context menu …



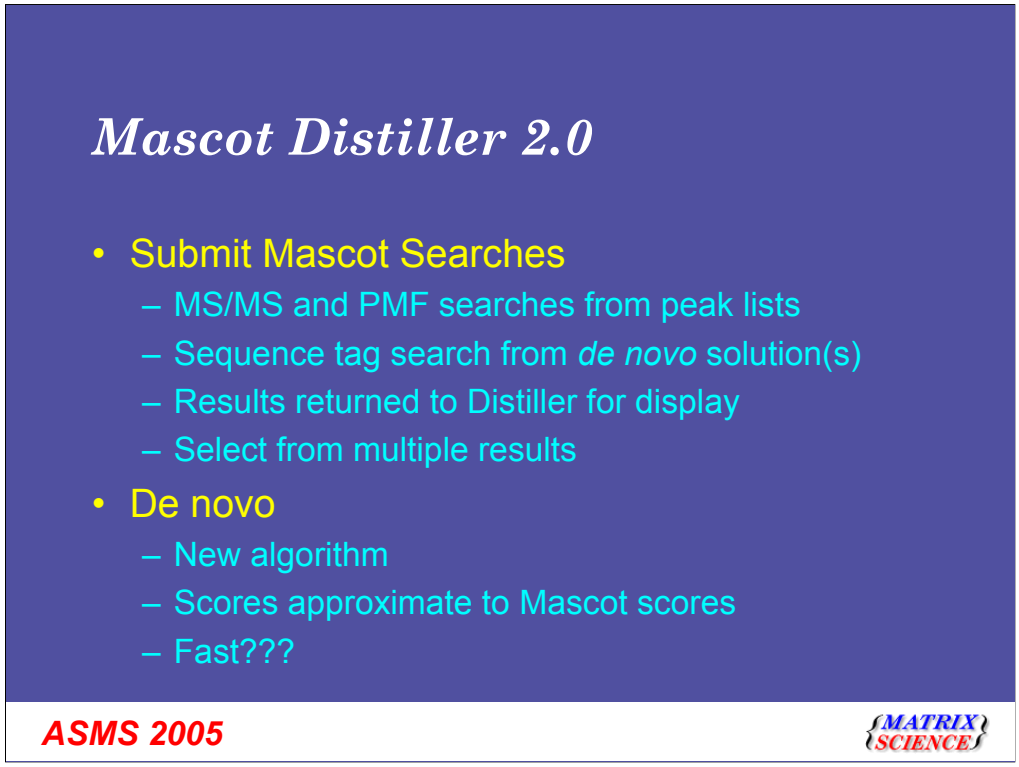
We get a search of the uninterpreted peak list.



The results from this new search are displayed on the explorer tree. Please not that the top score here is very close to the score for the de novo solution. We are actually using the Mascot scoring algorithm on the de novo solutions.

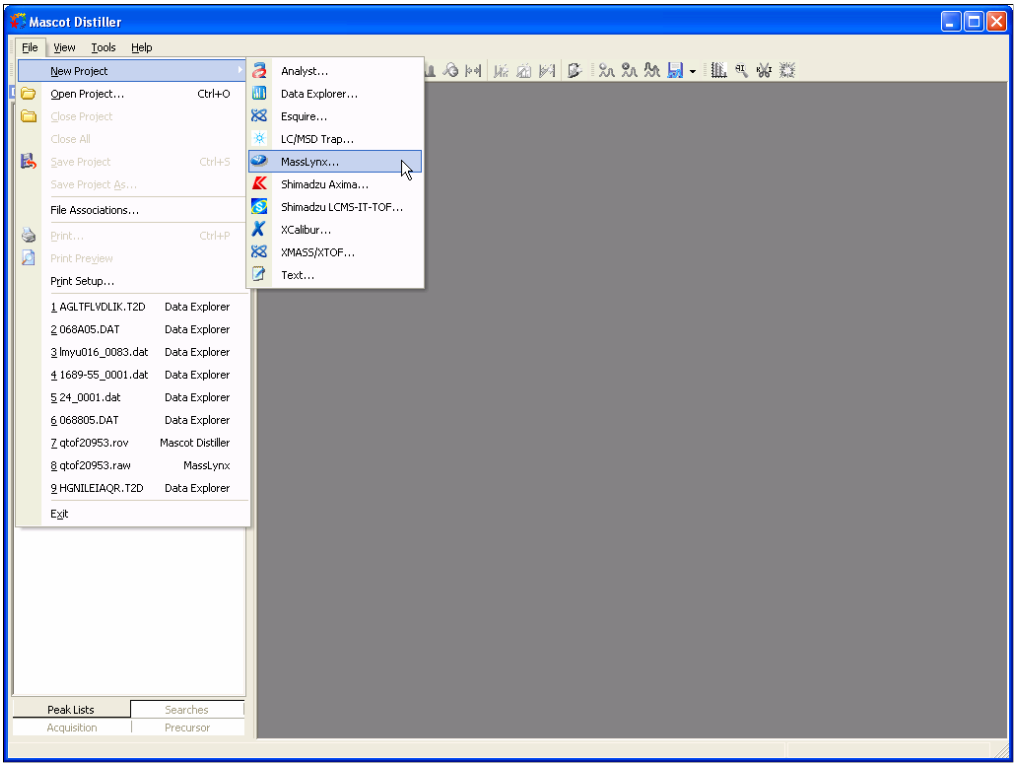


When a search completes, the results are displayed on the tree. If you want to go back to an earlier set of results, you can do this easily. However, only one set of results can be displayed at any one time. Otherwise, it would be too confusing. You can also use the context menu to display the standard Mascot report in a web browser.



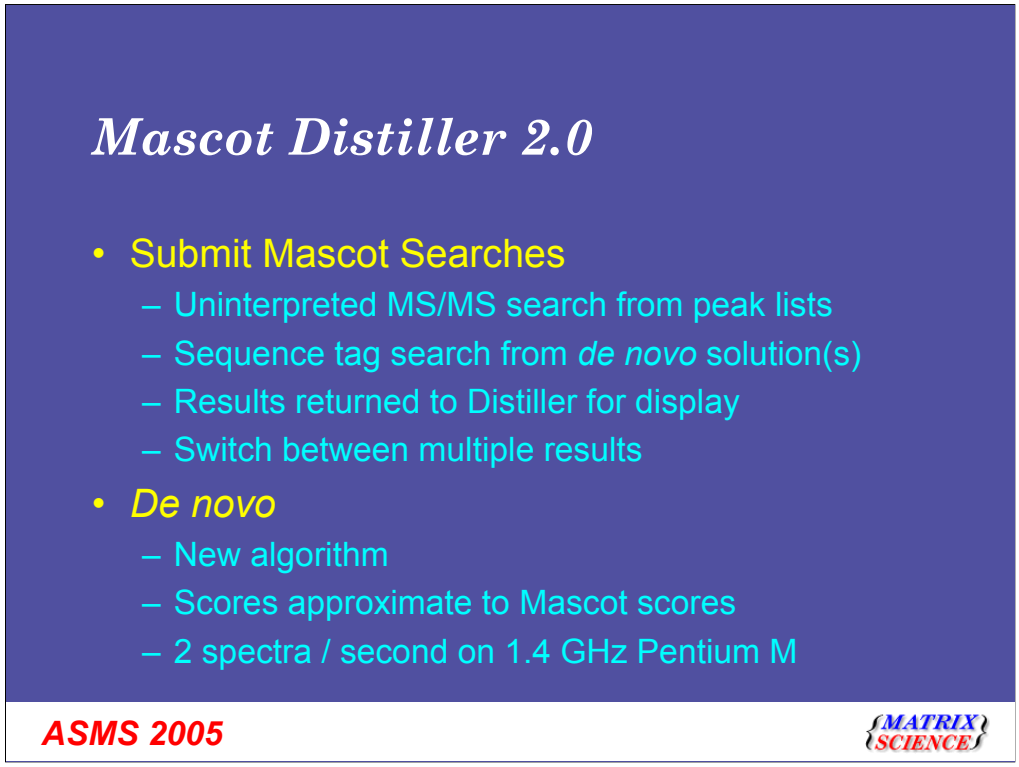
Let me summarise some of the things we just saw.

The title of this talk claimed that we had “Fast de novo sequencing”. What does this mean?

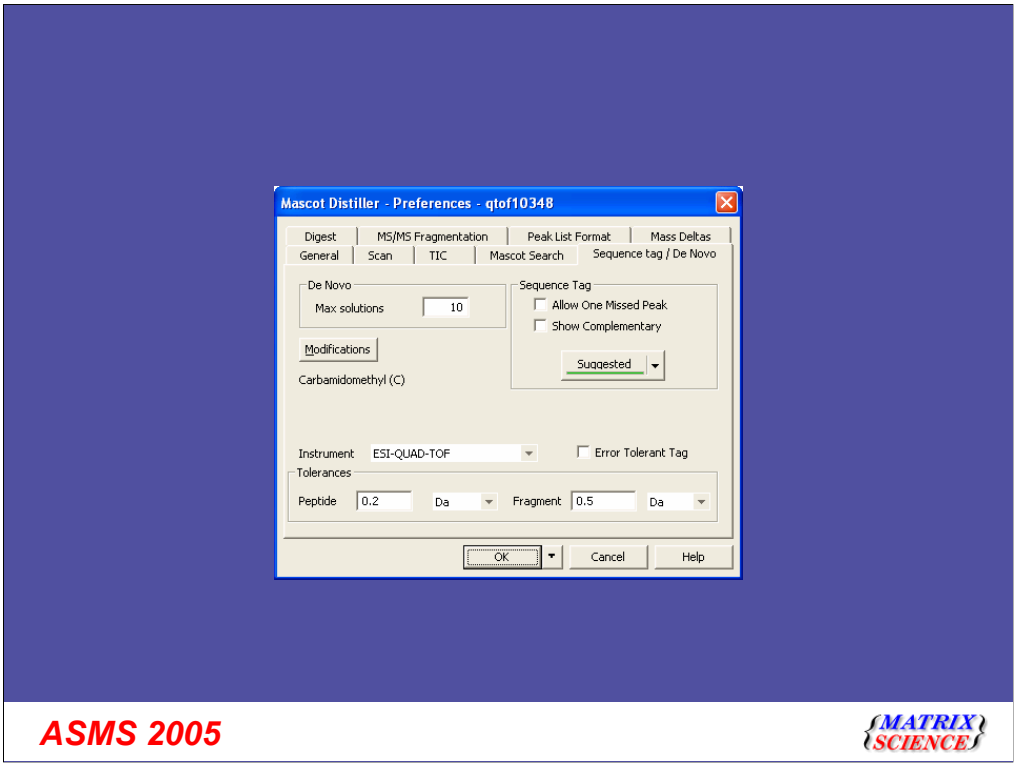


Lets open an LC-MS/MS dataset

And process some some of the scans so as to create 30 odd summed spectra. This movie shows all these spectra being de novo sequenced in real time



That was 33 scans in approximately 16 seconds … approx 2 scans per second on an average laptop



The conditions for the de novo interpretation are shown here, in the preferences dialog. You can have up to two variable modifications, although none were used here. The ion series to be considered are specified by choosing an instrument type, just like in a Mascot search.

One thing that is missing in this version is the ability to specify enzyme constraints. This will be added and can make a large difference to the success rate

So, it is fast, but is it any good?

Comparison of [redacted] and [redacted]

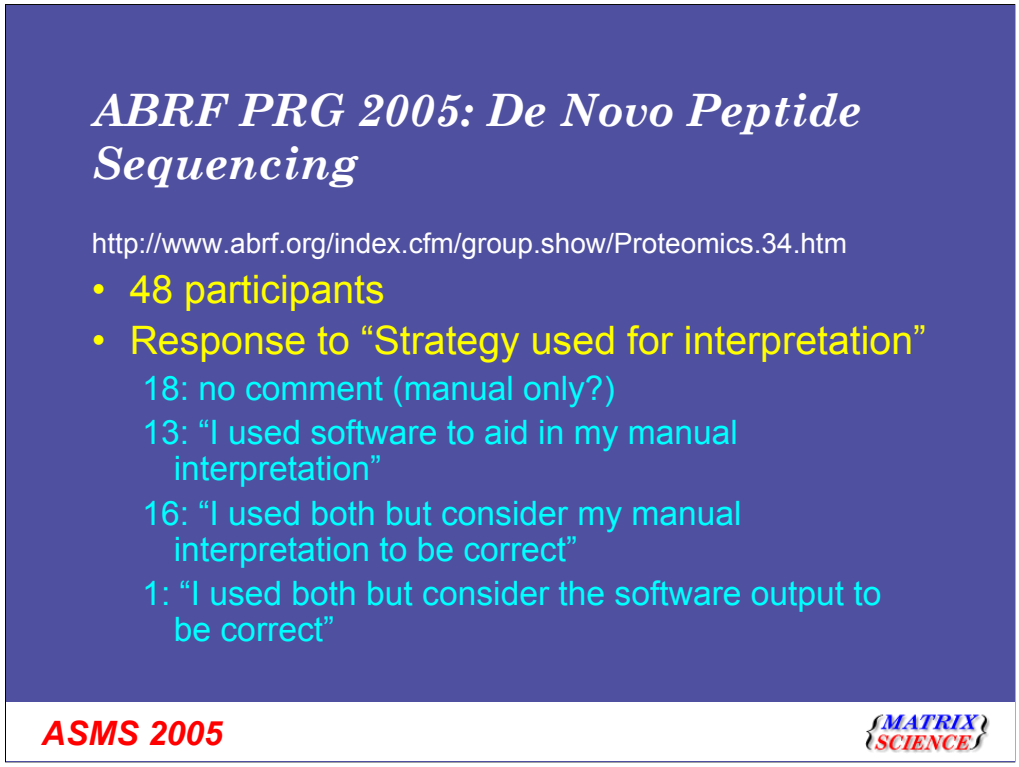
m/z	z	Correct Sequence	[redacted] *(de novo)*	Comments	[redacted] *(de novo)*
MALDI MS/MS					
BSA					
927.4	1	YLYEIAR	YLYEIAR	correct	[276.14]EY[184.08]R
1439.7	1	RHPEYAVSVLLR	GVLMVDVPPADNGR	Wrong (?)	No results
1479.8	1	LGEYGFQNALIVR	LWYGFQNALIVR	correct	No results
1639.8	1	KVPQVSTPTLVEVSR	RAPKVPQVSTPTLVEVSR	correct	No results
ESI MS/MS					
Cyt- c					
482.7	2	EDLIAYLK	EDLIAYLK	correct	[357.15]LAYLK
584.8	2	TGPNLHGLFGR	TGPNLHGLFGR	correct	TGPNLHGLFGR
589.3	1	GDVEK	VDVEK	V = Ac-G	VDVEK
634.4	1	IFVQK	IFVQK	correct	IFVQK
678.3	1	YIPGTK	YIPGTK	correct	YIPGTK
728.8	2	TGQAPGFSYTDANK	TGQAPGFSYTDANK	correct	[199.10]SAPGF[250.09]TWNK
779.4	1	MIFAGIK	MIFAGIK	correct	[244.12]FAGLK
792.9	2	KTGQAPGFSYTDAMK	KTGAGAPGFSYTDAMK	almost	[229.15]QGAPGAYQNHANK
817.3	2	IFVQKCAQCHTVEK	QFVTHMACCHTVEK	partial	[257.08][218.08][GP][260.08][HM]TVEK
Apo-Myoglobin					
662.3	1	ASEDLK	ASEDLK	correct	[244.07]SALK
689.9	2	HGTVVLTALGGILK	HGTVVLTALGGILK	correct	HGTVVLTALG[170.1]LK
748.4	1	ALELFR	ALELFR	correct	[184.12]ELFR
803.9	2	VEADIAGHGQEVLIR	LDADIAGHGQEVLIR	almost	no results
908.4	2	GLSDGEWQQVLNVWGK	GLSDGEWQQVLNVWGK	correct	[170.11]SG[244.07]WQQVLNVWGK
943.2	2	YLEFISDAIIHVLHSK	YLEFISDAIIHVLHSK	correct	[276.1]EFLSD[184.12]LHVLHSK

Red = Correct

That’s a very difficult question to answer. If you look on the web, or in ASMS extended abstracts, you’ll find many comparisons like this one, showing how “our” software knocks the spots off “their” software. I don’t feel comfortable about presenting this kind of comparison. There is too much temptation to repeat the experiment until you get the result you want to present. Also, you know how to get the best from your own software, but may not be so expert with someone else’s.

Our preferred way to convince you that Mascot Distiller does a useful job is to make it available on a free 30 day evaluation, just like we do with the current release. Then, you can try it on your own data and make your own decision.

The other way to get a genuine comparison is a properly conducted blind study involving a reasonable number of independent groups.



One such study was the PRG 2005 exercise organised recently by ABRF. Unfortunately, we weren't in time to catch this. Hopefully, there will be something similar organised again in the near future.

You can get a spreadsheet of the results from the ABRF web site.Some of the comments make interesting reading.

ABRF PRG 2005: De Novo Peptide Sequencing

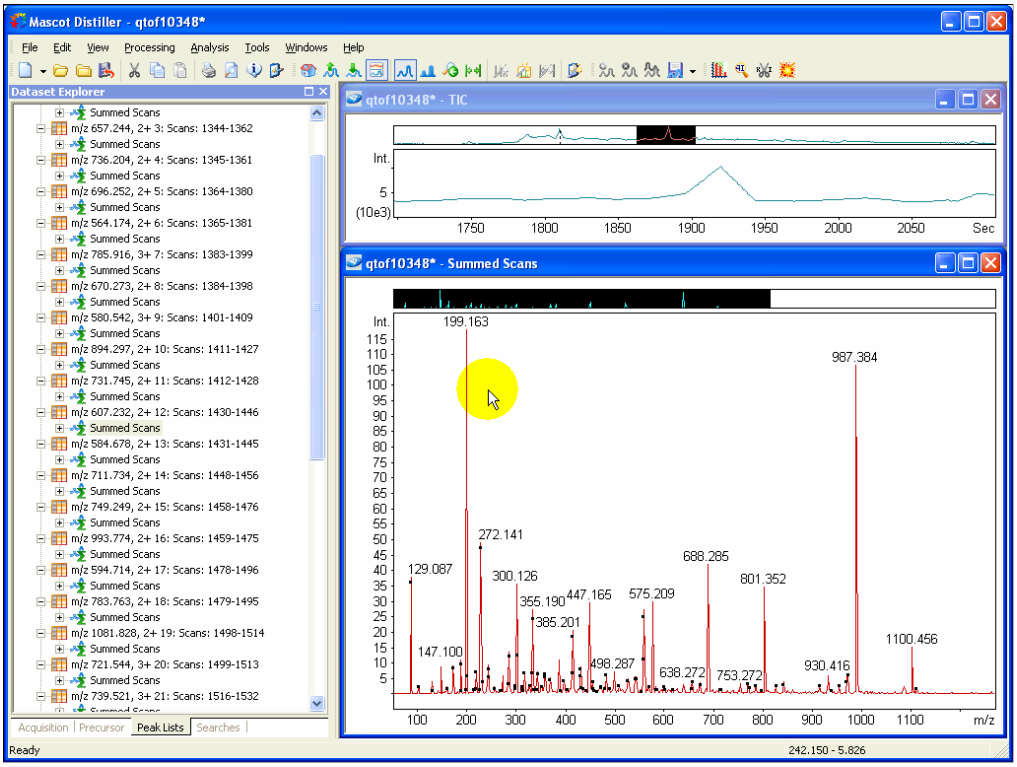
- Responses to “If you manually interpreted the spectra and also used software to assist in your interpretation, do you consider the software currently available to be adequate?”
  - 18: No comment (manual only)
  - 9: Yes
  - 21: No

ASMS 2005

MATRIX SCIENCE

Sounds like people feel there is room for improvement

I showed earlier how a de novo solution can be searched as a set of tags. Mascot Distiller 2.0 also supports manual tag interpretation. I think you can see this best if I show another short movie



-Maximise the window

-Choose a likely looking peak, such as 987.384

-Right click to start a tag

-Click on any arrow to extend the tag

-In general, I just go for the biggest peak

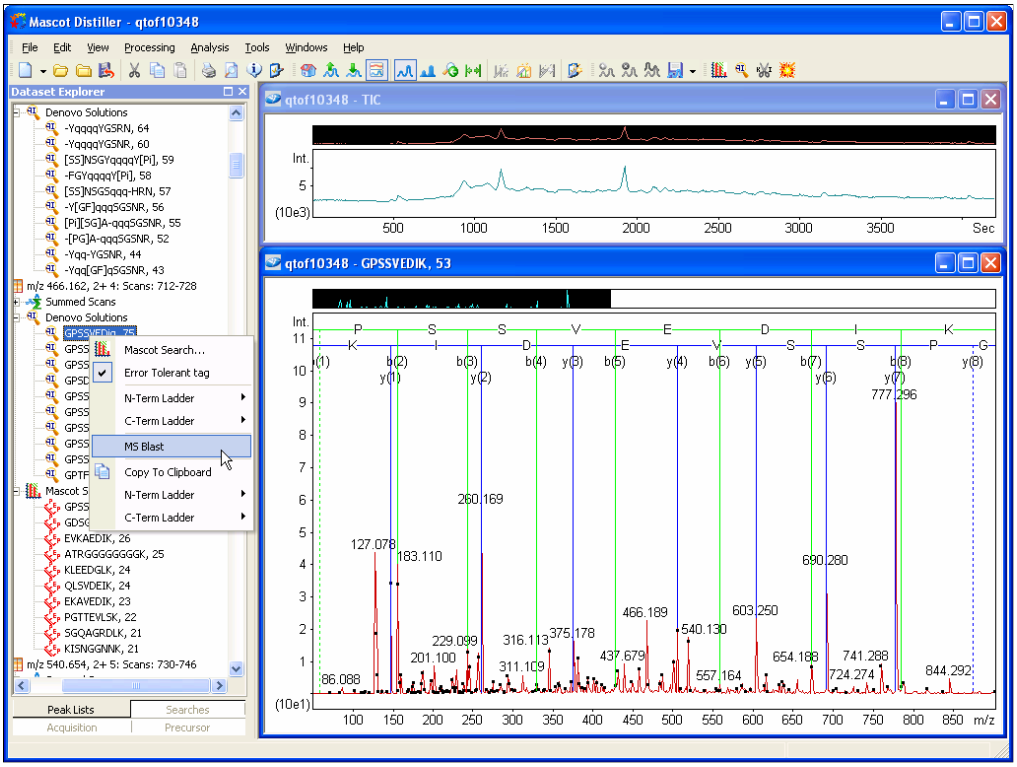
-Stop when it starts to look tricky

-Here's the tag

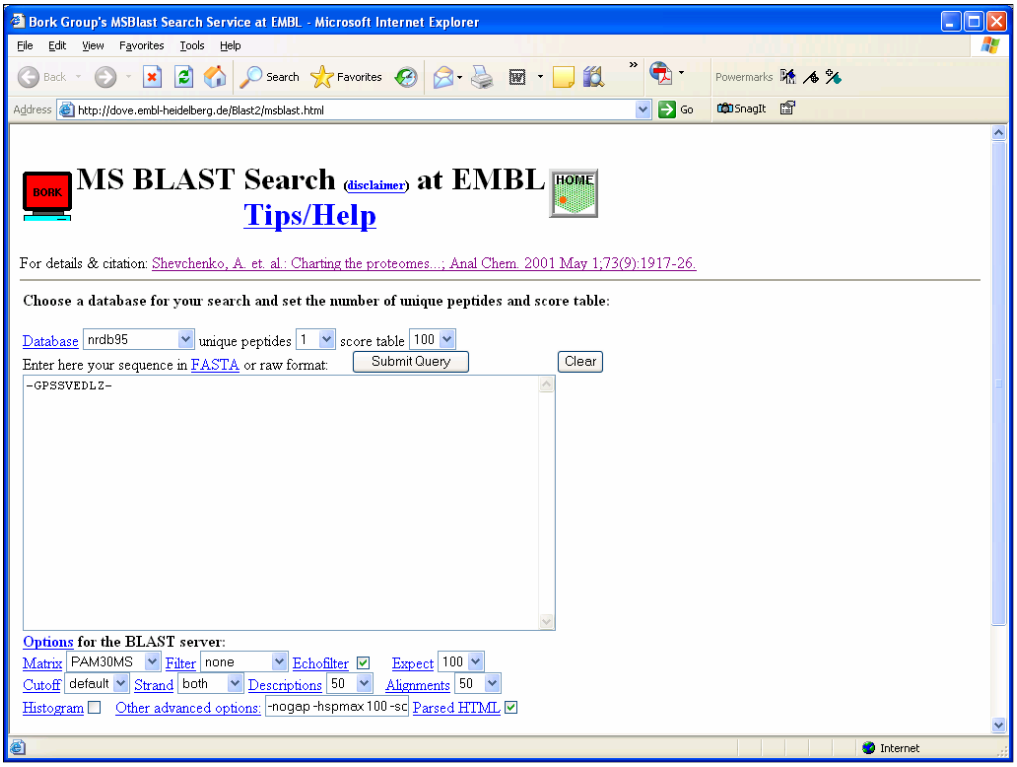
-Do a Mascot search of the peak list to see what the answer should have been. Here's one I prepared earlier.

-Whoops! Got it wrong, should have been GE, not W. I'll stick to the day job.

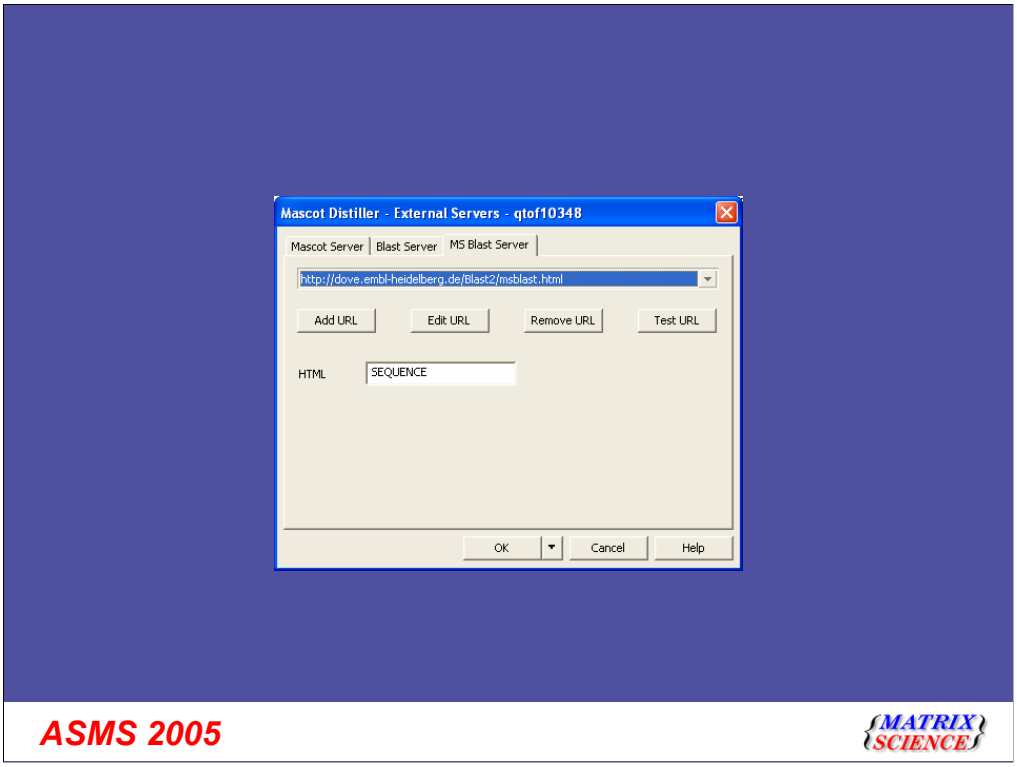
So, we can generate tags automatically or manually. We can then search them using Mascot or Blast or MS-Blast



Right click a de novo solution and choose MS-Blast

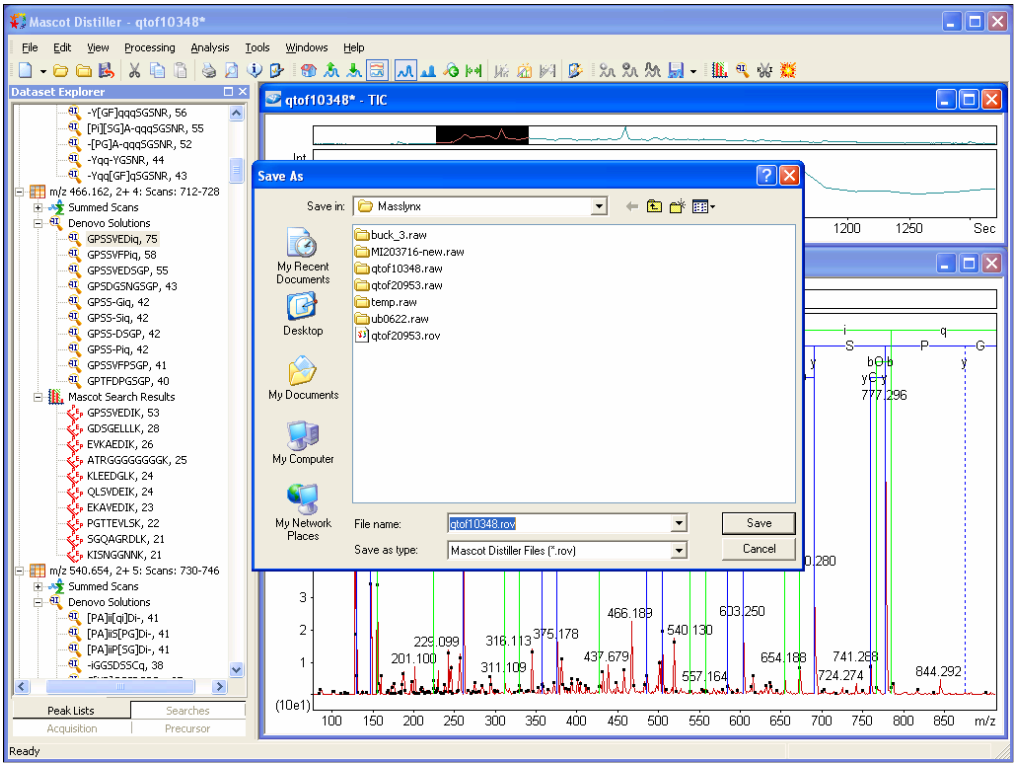


And the sequence is translated into MS-Blast syntax and pasted into the search form.



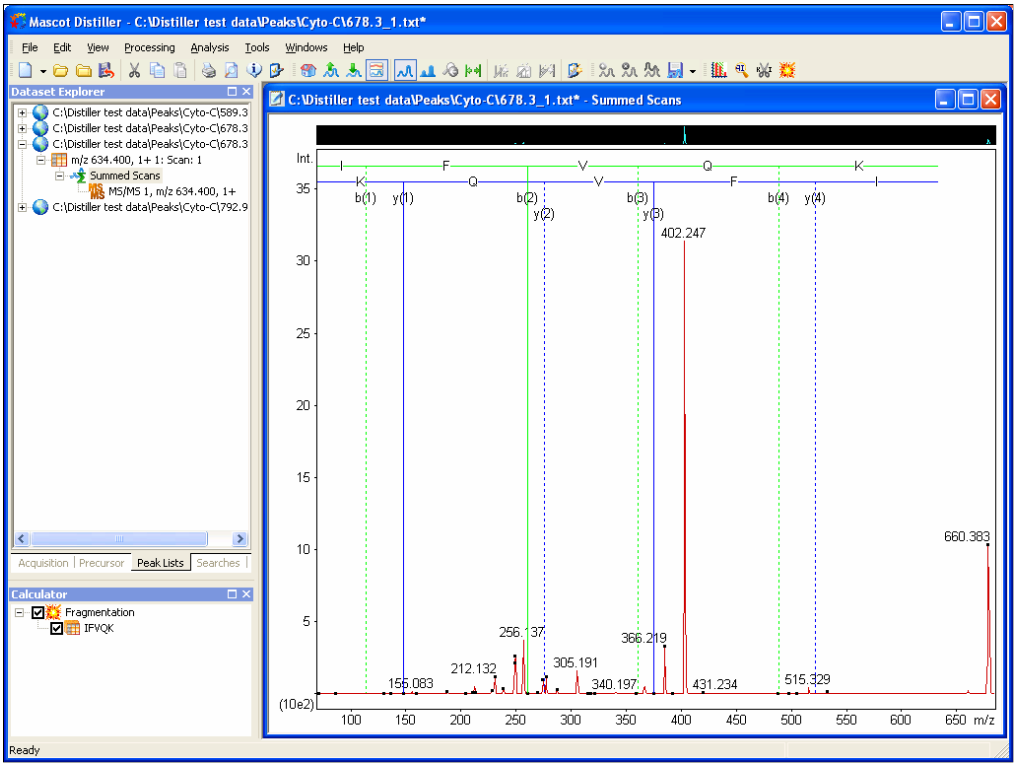
The URL's and any required parameters are configured in this "External Servers" dialog

Another new feature of 2.0 is that we can finally save and open projects. This allows the entire workspace to be saved. In fact, we now have an option in Mascot Daemon to save a project file when batch processing raw files through Distiller. This means you can examine Daemon search results in Distiller by simply clicking a hyperlink

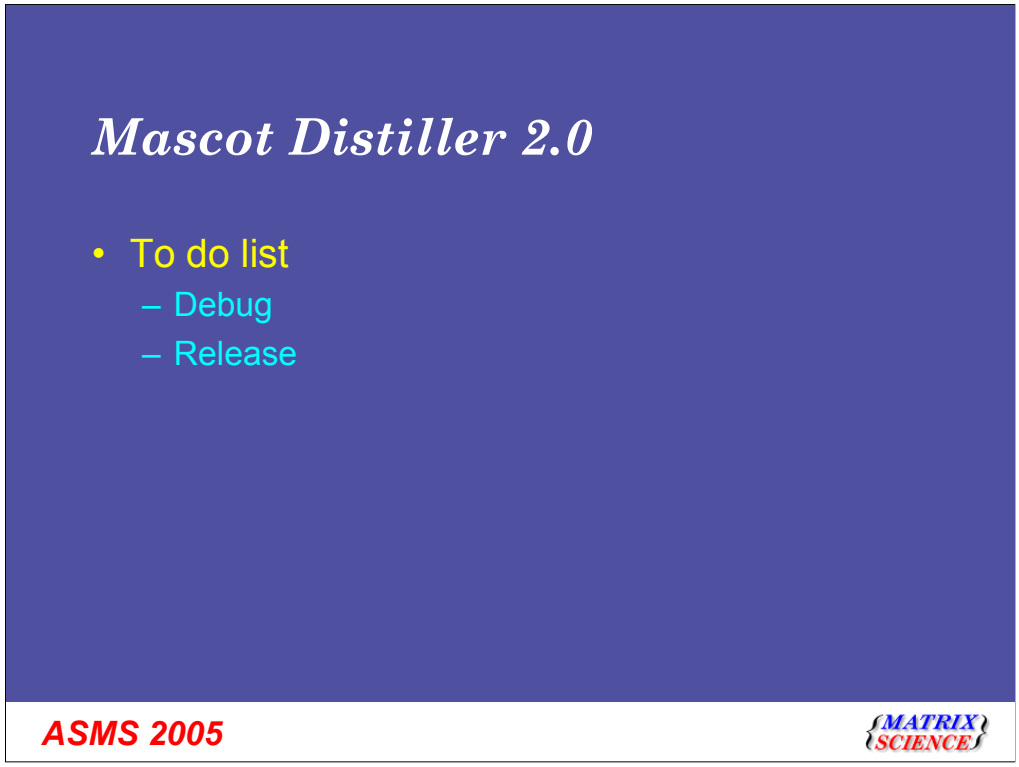


Proof statement

A project file is actually a zip archive containing a number of XML files



Mascot Distiller 2.0 also allows you to enter a protein or peptide sequence and display the fragments against the spectrum.



What's left to do? Just finish it off and get it released!