Mascot Distiller - the key to automation

ASMS 2003

MATRIX SCIENCE

Mascot Distiller - the key to automation

- What is Mascot Distiller and why do we need it
- Mascot Distiller - internals and functionality
- Mascot Distiller in an automated environment

ASMS 2003

MATRIX SCIENCE

We are pleased to announce that Mascot Distiller is now available. In this talk I'll describe the rationale for developing the product, and demonstrate some of its rich feature set.

I'll show that there are actually two parts to this piece of software - an engine that can easily be used from many programs and a standard Windows application.

Finally, I'll explain why Mascot distiller is a key to automation - and how to use it in an automated environment.

What is Mascot Distiller?

- **Mascot Distiller reads raw data from most mass spectrometers and produces peak lists suitable for Mascot**
- **It consists of two parts:**
  - **An 'engine' (A Windows COM library)**
  - **A Windows application**

ASMS 2003

MATRIX SCIENCE

Mascot Distiller does what sounds to be a very straight forward job. Essentially, it just reads raw data files from any mass spectrometer, and produces high quality peak lists.

Sounds simple doesn't it? I'll describe some of the difficulties shortly.

Mascot Distiller consists of two parts - an engine and a standard Windows application.

Now, this may sound un-interesting to you, but it is in fact critical. We want to use the engine from a number of products, and also enable other people to use the engine. We wanted to be sure that the same algorithms were used by all the different applications. This has been made possible by making it a COM library.

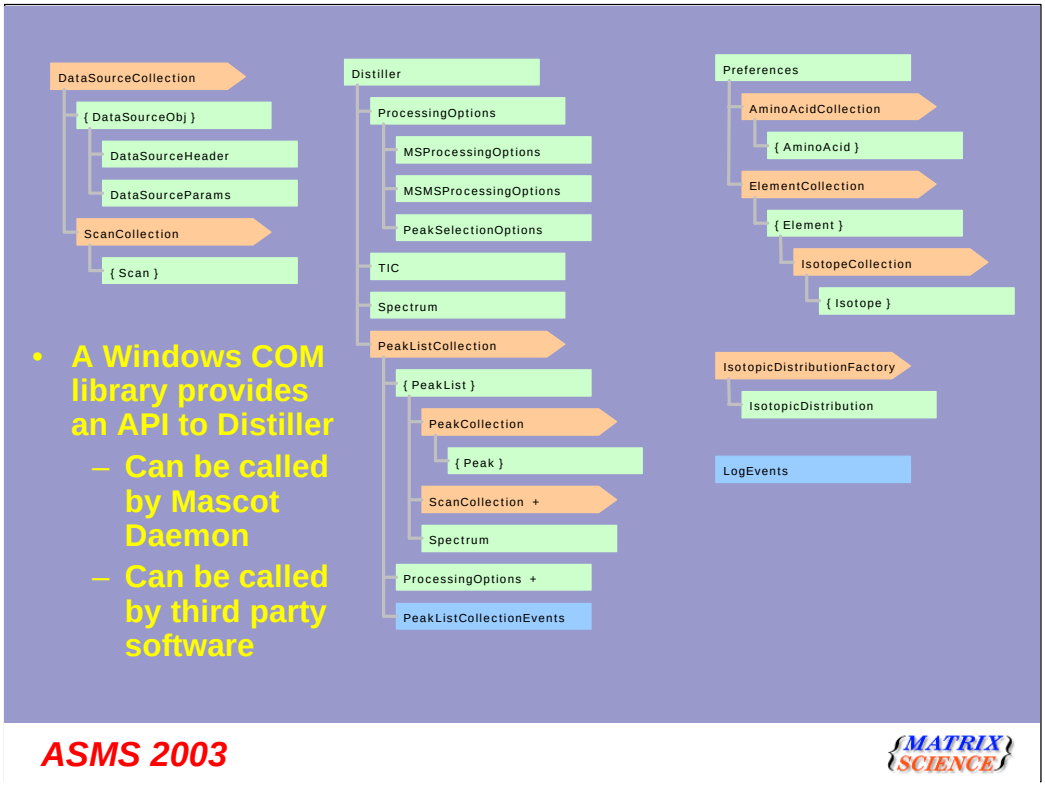
Before I go on to briefly describe Mascot Distiller, I'll just explain why we felt we needed to develop it.

It remains a fact that mass spectrometers are complex to operate and it is often hard to get good data from them. It also seems to be a universally accepted opinion that getting the data is only just the beginning - many more hours of work will follow to interpret the data.

Most labs seem to have instruments from more than one manufacturer - or at least have instruments with different data systems. We have observed that people struggle with generating peak lists suitable for searching - or even just finding their way around a complex data set. Key strokes, mouse actions, terminology, icons are all different between different data systems. So, we have seen a need for a single user interface for accessing date from multiple systems.

Secondly, the quality of the peak lists can best be described as variable. What we often notice is that in expert hands a reasonable peak list can be obtained - however, for the average user the results are often sub-optimal, and result in poorer Mascot results than would otherwise be obtained.

Finally, we some data systems it is not even possible to automate the output of peak lists - it has to be performed manually.



I'll briefly describe the Mascot Distiller 'engine' - where all the hard work takes place. However, you will be relieved to know that I'm not going to go into great detail here.

We've gone through many iterations to try and get the correct 'model', and are quietly confident that we have something that will be future proof and robust.

This engine can be called from any Windows program - you can write a small piece of code in VB, or you can write an even smaller VBScript. If you are brave, you can probably even use Microsoft's latest J# language.

A developers kit with full documentation is available now.

For this product, we are breaking our tradition of supporting lots of different Unix platforms and we will only support Windows. This is because for most instruments we require the data system to be present, and these are now mostly Windows based systems.

Distiller "Engine" used by

- Mascot Distiller application
- Mascot Daemon 2.0 and later
- Mascot Wizard
- Your application

ASMS 2003

MATRIX SCIENCE

The Windows application that I am about to show uses this COM engine.

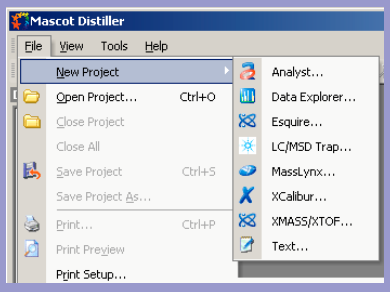
When we release Mascot 2.0 later this year, Mascot Daemon will use this engine for working directly with the raw data files.

For those of you that know Mascot Daemon, this simply means that we will add to the list of filters.

I'll also show you another small product 'Mascot Wizard' that we are due to release a little later this year.

Data files supported

- Analyst (QStar, QTrap)
- Data Explorer (Voyager)
- Data Analysis (Esquire)
- Data Analysis (LC/ MSD)
- MassLynx (QTof, M@ldi)
- XCalibur (LCQ, Deca)
- XTof/XMass (autoflex, ultraflex, omniflex)
- Text



ASMS 2003

MATRIX SCIENCE

I'll now describe and show you the Windows application.

Firstly, we can open data files from most of the instruments. In alphabetical order we have:

Analyst - both Analyst QS and Analyst 1.3 are supported, which means that we can read in files from the Qstar and the Qtrap instruments.

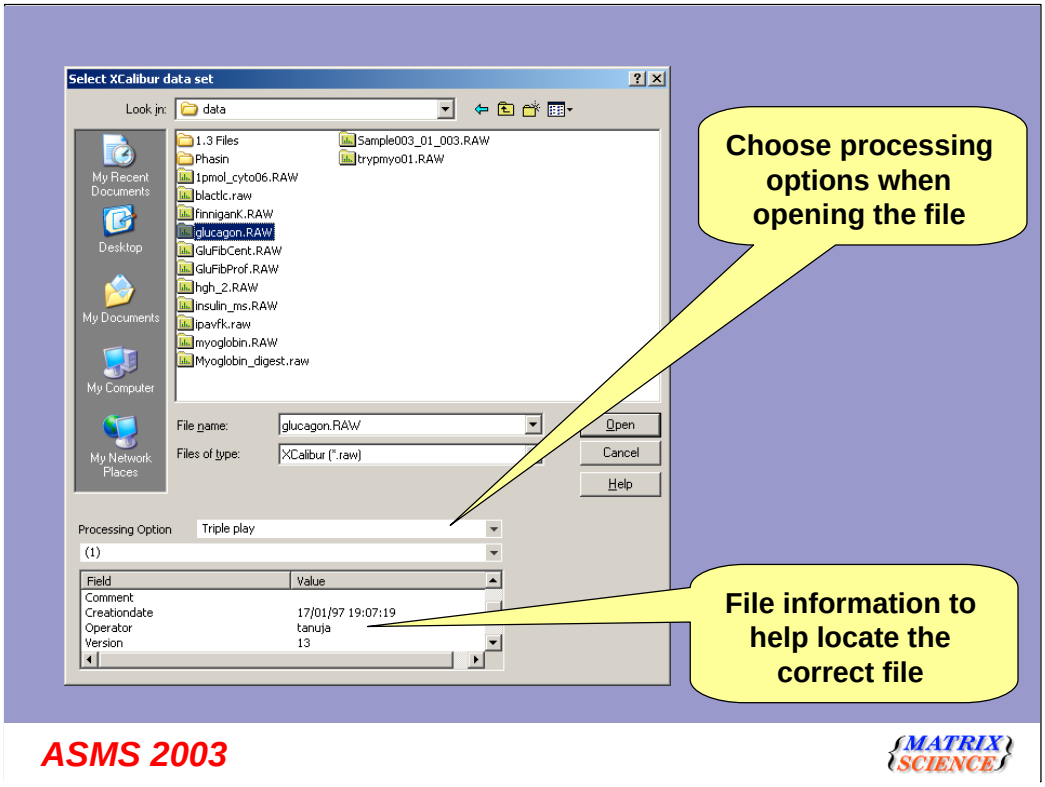
We have full support for the Voyager family of instruments - although we only support Data Explorer 3.5 and later - we don't support the old Windows 3.1 Grams software

The ion trap instruments from Bruker and Agilent are both supported.

All the Waters / Micromass instruments that use MassLynx are supported - e.g the Qtof and the M@ldi.

With the ions traps from Thermo Finnigan, we can open Xcalibur 1.2 and Xcalibur 1.3 files.

Note that for Analyst, Data Explorer, MassLynx and XCalibur, Mascot distiller needs to be installed on the computer with the relevant data system because it uses certain DLLs provided by the manufacturer.



There is a consistent file open dialog box. For each of the data file types, some key file information will be shown in the list at the bottom. As you may be able to see here, this is a very old LCQ data file - creation date 17th January 1997 at 7pm in the evening, and the operator is someone called Tanuja.

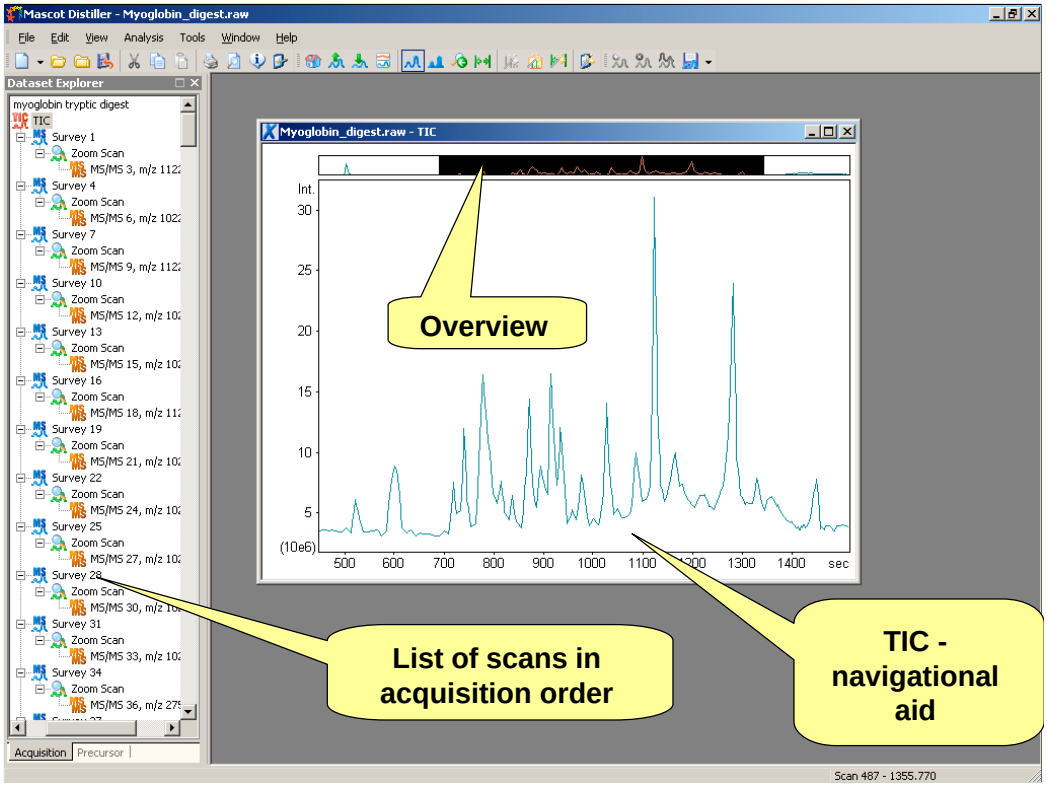
One of the reasons for the complexity in Mascot Distiller is the different modes that an instument can run in.

The simplest type of spectrum is a single MS spectrum from say a Maldi time of flight instrument.

We can also support a single MS/MS spectrum

Multiple ms spectra, e.g. from the M@ldi are supported, as are multiple ms/ms spectra, e.g. from nano spray

The more complex experiments such as triple play need to be understood by the software if they are to take full advantage of the data. For example, if the software didn't know what a zoom scan was, it couldn't determine the charge state for the following ms-ms spectra.

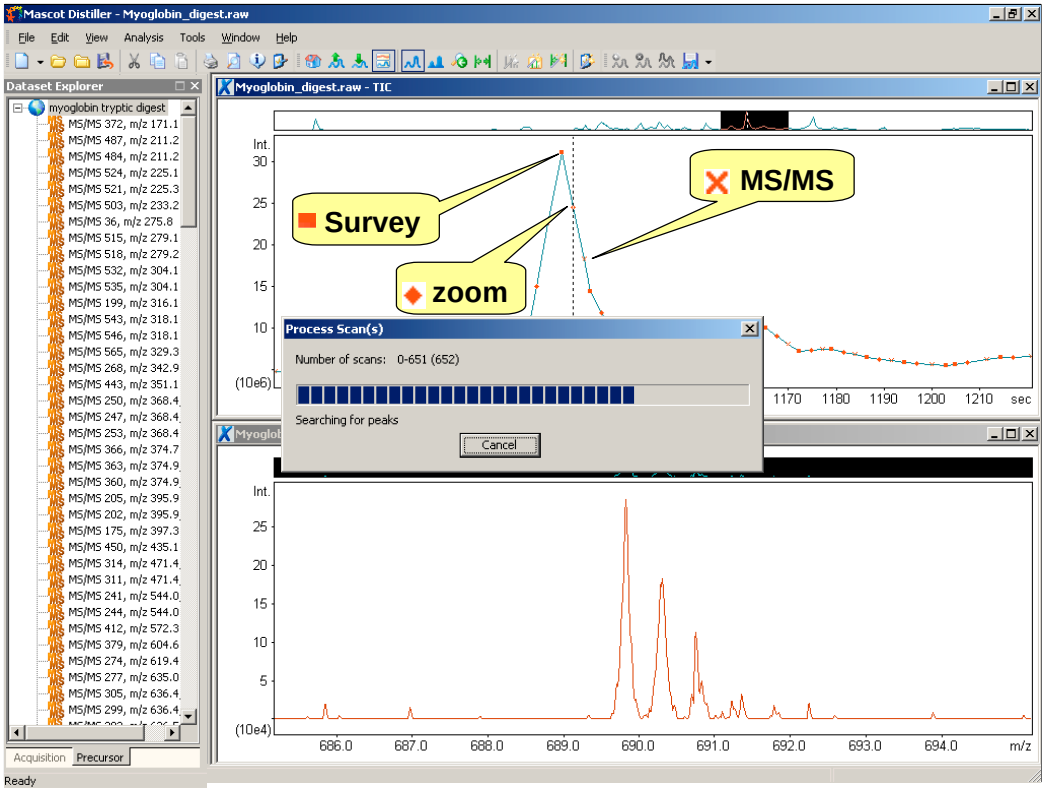


Since we have limited time for this talk, let's jump in and open one of the more complex data sets - a triple play lcq data set.

When you open the file for any multiple scan data set, a TIC window is displayed. This is chiefly a navigational aid. You can hopefully see at the top of the window a thin representation of the whole TIC - the black area shows the part of the TIC that is actually visible at the moment. You can obviously zoom into any part of the spectrum

On the left we have a list of the spectra that were accumulated in time order.

If we now double click on a point in the TIC...



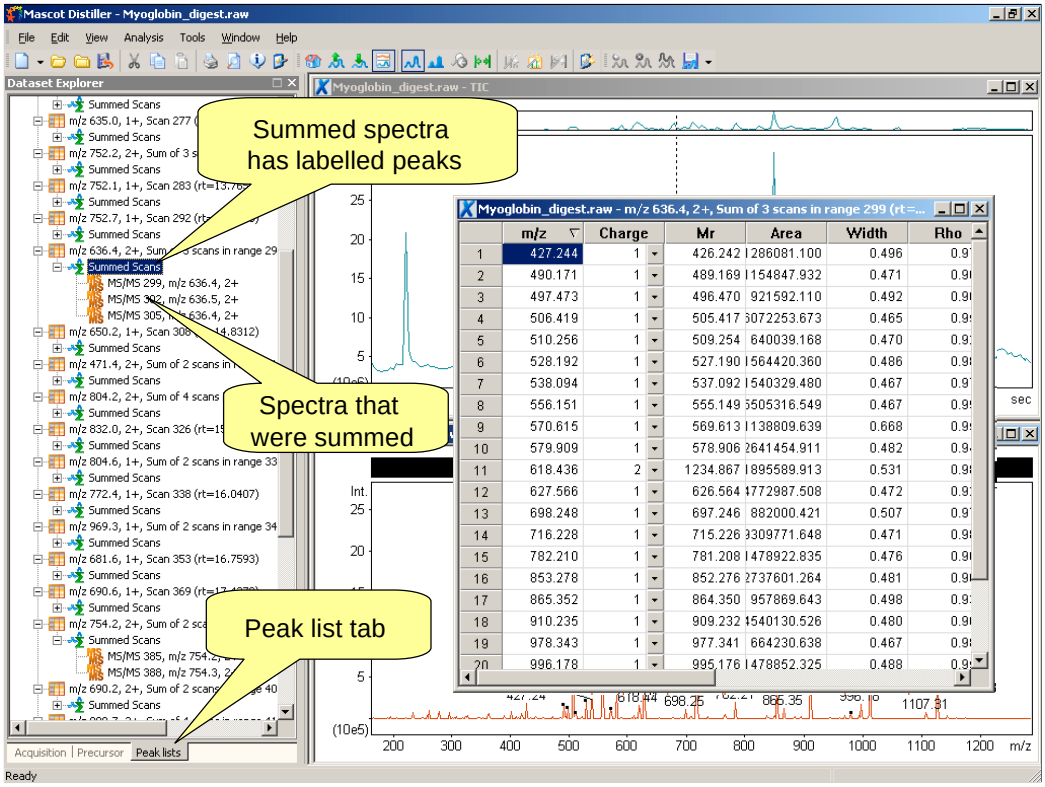
We can see the relevant spectrum for the point on the TIC.

The interface is simple, just click on a point in the TIC to display the relevant spectrum, and note that this is also highlit on the tree view on the left. Of course you can also navigate by clicking on the left.

Although we use the term 'zoom' scan, this is obviously the same as the MaxRes scan for Bruker / Agilent data and Enhanced resolution scan for Sciex QTrap data

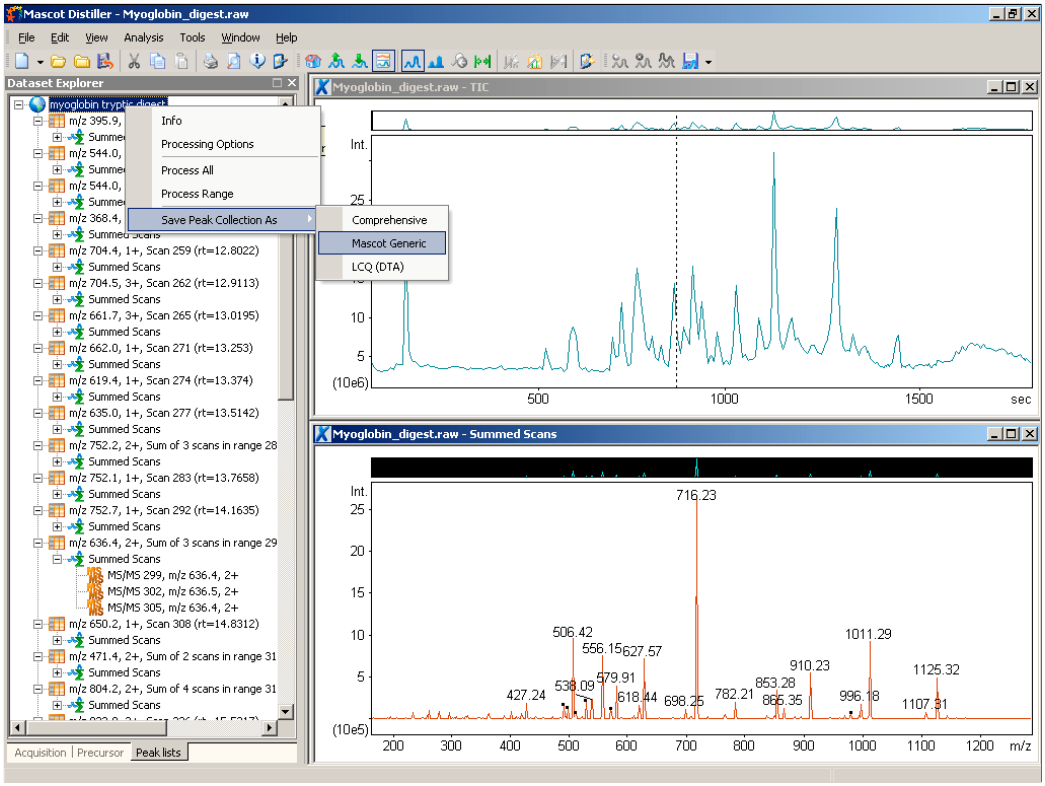
Note the two tabs at the bottom here - acquisition and precursor - we can view in acquisition order or in precursor mass order.

If we process all the scans to detect peaks etc then we end up with



A third tab at the bottom - the peak list tab. Spectra that have the same precursor mass and are within a specified retention time are merged together and peak detection is performed on the merged spectrum. It's also easy to view the spectra that were merged into this.

And finally, we can view, and even edit the peak list in a nice easy to use table.



So, the last stage in the process is to save the peak list in Mascot generic format for searching and to submit the data to be searched.

Issues with data reduction s/w

- Peak detection issues
- Incorrect precursor charge or mass determination for ms-ms
- Poor grouping of similar spectra
- Inability to recognise and remove 'junk' spectra

ASMS 2003

MATRIX SCIENCE

Most conventional mass spec software fails in one or more of the areas of data reduction.

So, either peak detection is poor

Or the Precursor mass or charge is determined incorrectly

Or the grouping of spectra is incorrect.

One other major problem is that junk spectra are often not removed.

I will go into some detail on each of the issues, and describe how Mascot Distiller attempts to resolve them

Problems with conventional peak detection

- Failure to pick low intensity peaks
- Picking peaks that are just noise
- Selection of the wrong peak(s) or all peaks in an isotopic cluster
- Need to continually 'tweak' parameters

ASMS 2003

MATRIX SCIENCE

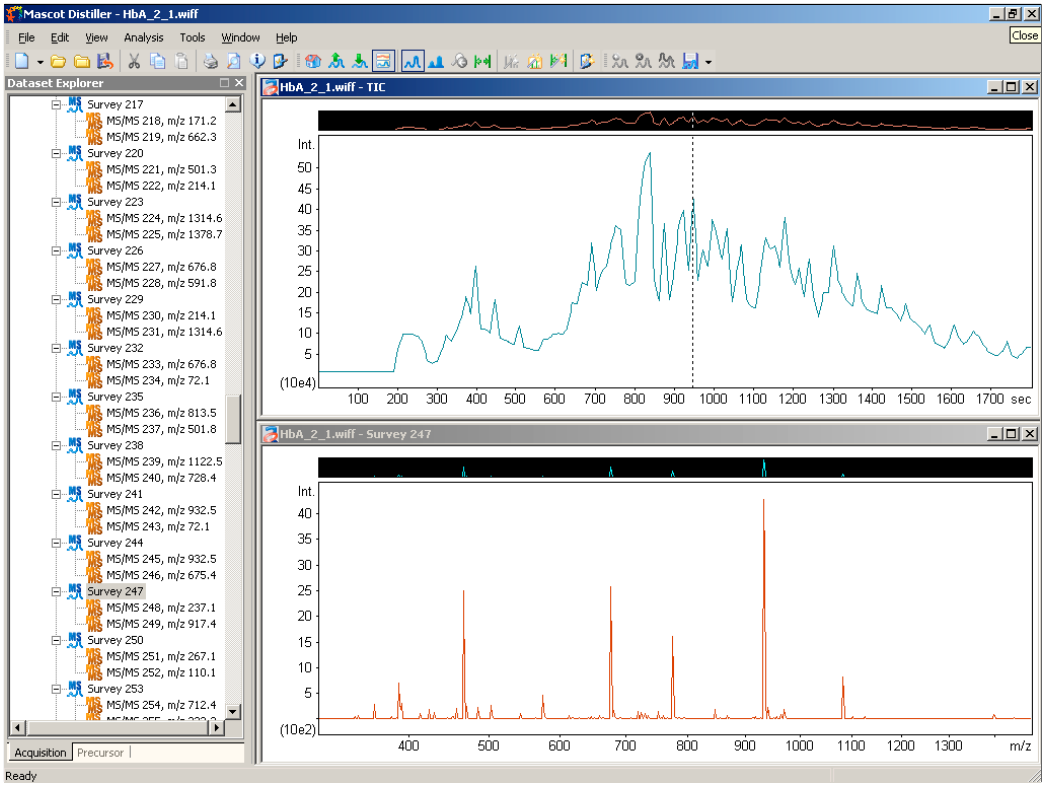
Conventional peak detection software routinely fails to pick out low intensity peaks. There is often just a simple threshold to reject peaks.

If the threshold is set too low, then a noise peak is labelled.

A very common problem is that the wrong peak or all the peaks are picked in an isotopic cluster.

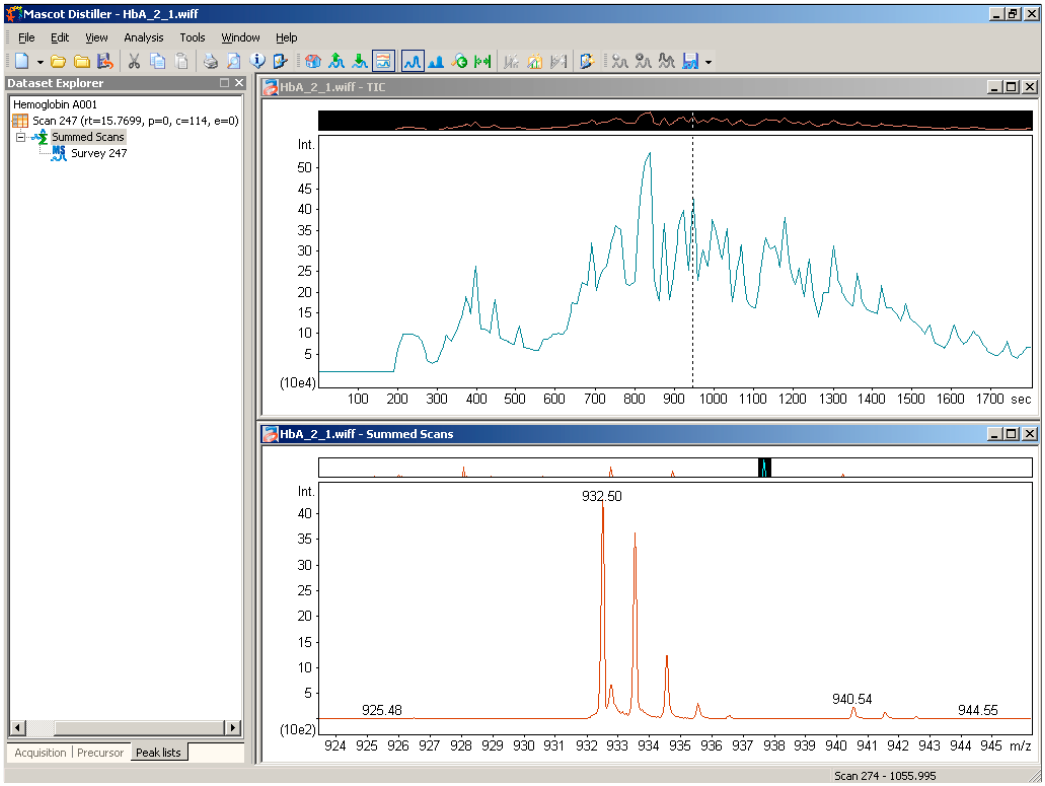
The other problem is that there is a continual need to tweak parameters.

Mascot Distiller detects peaks by attempting to fit an ideal isotopic distribution to the experimental data. This ideal distribution is predicted from the elemental composition expected for a peptide of average amino acid composition at that point on the mass scale. The profile of the ideal distribution is then adjusted by varying the mass, resolution, intensity, and charge state, so as to maximise the correlation with the experimental data. Once the best fit has been obtained, the peak is added to the peak list and the corresponding signal subtracted from the spectrum. The process is repeated until all the significant peaks have been fitted.



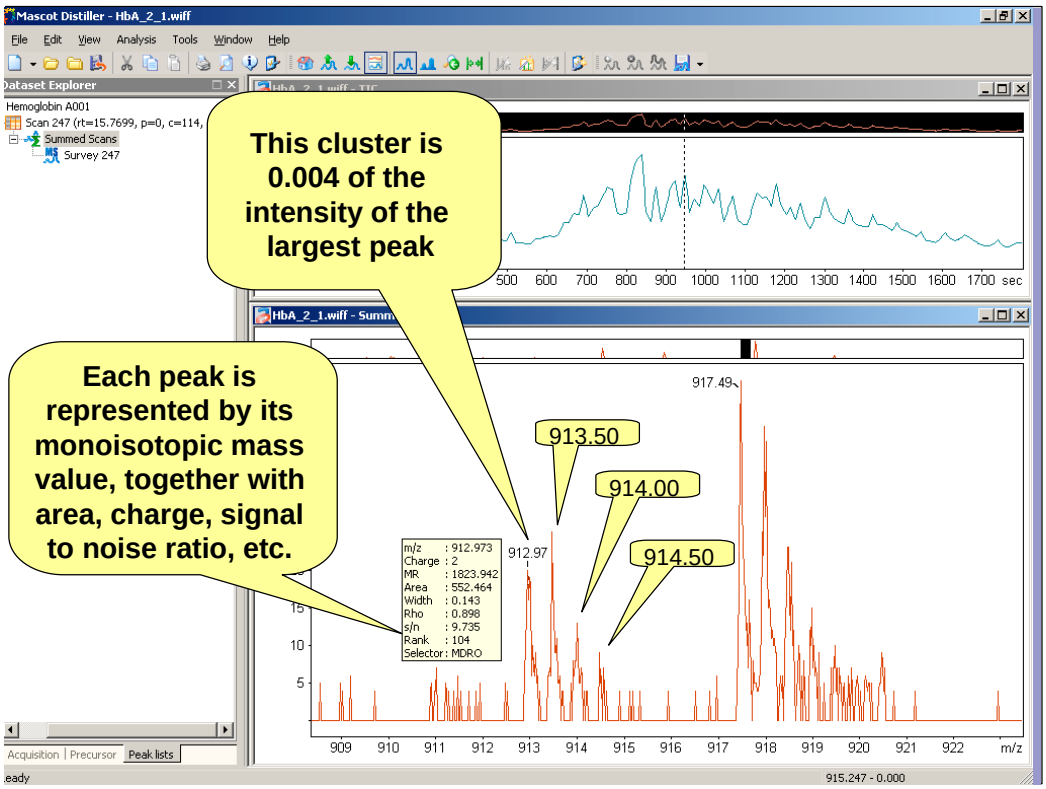
Here we can see some data from a Sciex QStar. This is very nice data, and from this survey scan we can clearly see some large peaks that any respectable software will detect.

If we just zoom in on this peak now



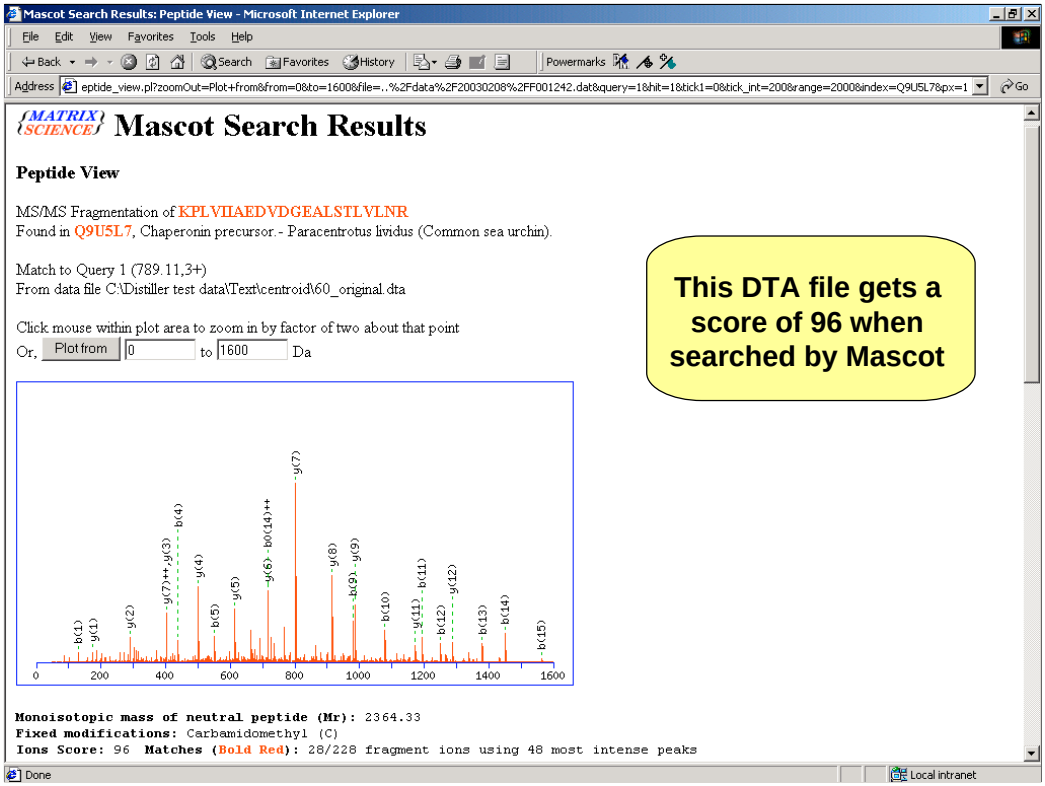
… we can see a nice looking isotopic cluster that any peak detection software would detect - although some of the poorer algorithms will unfortunately detect and label the carbon 13 and carbon 14 peaks as well. I would just like you to note that the intensity of this peak is about 4000.

Lets now zoom into the low intensity area down over here



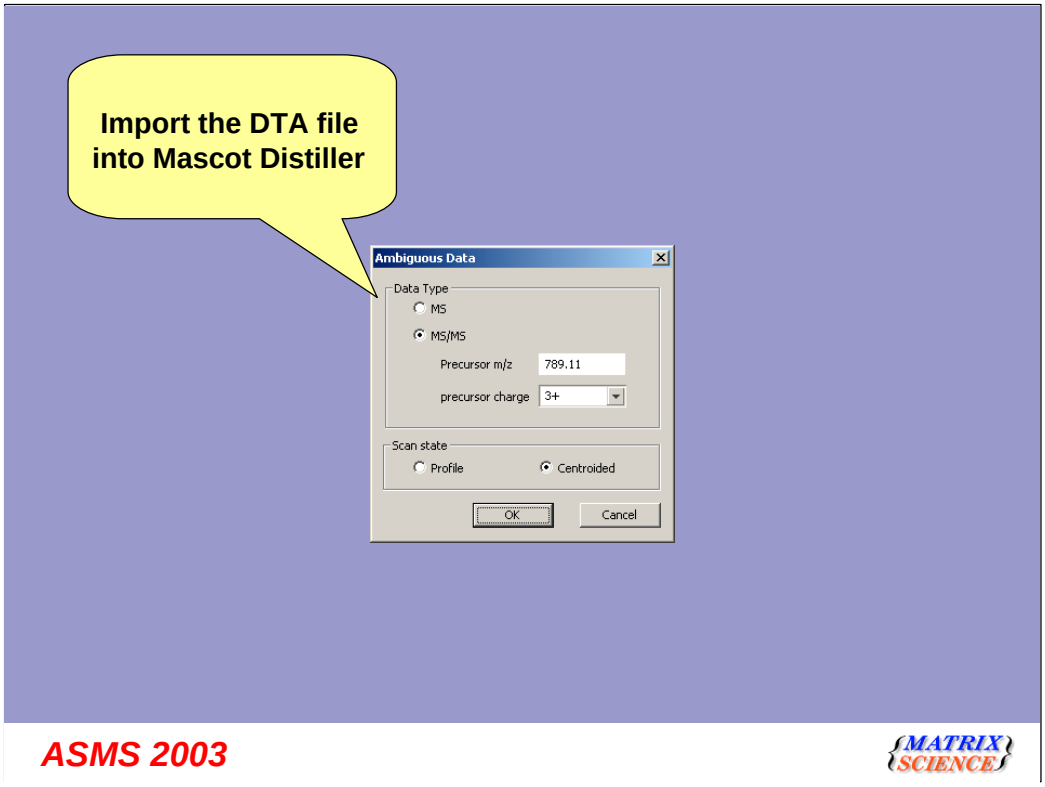
So, if we look carefully at this cluster, we can see that the peaks are exactly 0.5 Da apart. Also, note the shape of the cluster - it is almost an ideal shape - the individual peaks are not an ideal shape, but we shouldn't expect this at this low intensity. However, I am sure that you will agree with me that this is almost certainly a real peak from a peptide

However, look at the intensity level - this is about 20 - i.e. about 0.4% of the intensity of the most intense peak. If your software requires that you set an intensity threshold, then you are going to miss this peak cluster unless you set the threshold to a very low value.

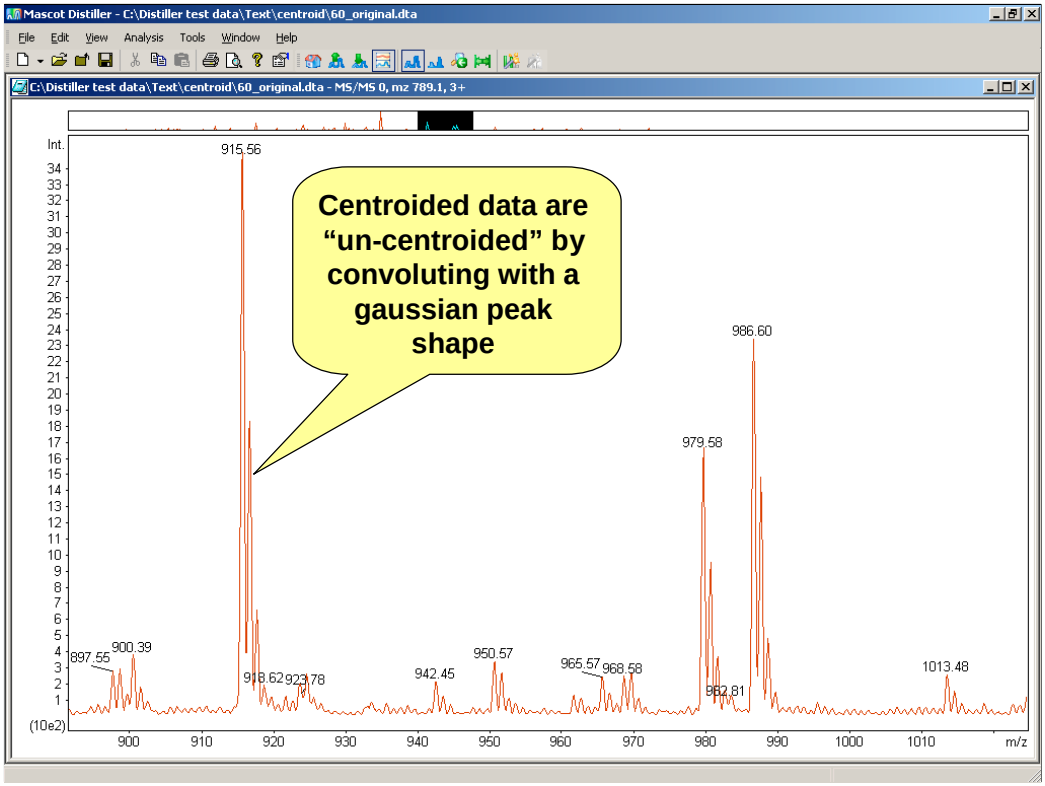


I'll just show a quick example of how this peak detection can improve a Mascot score.

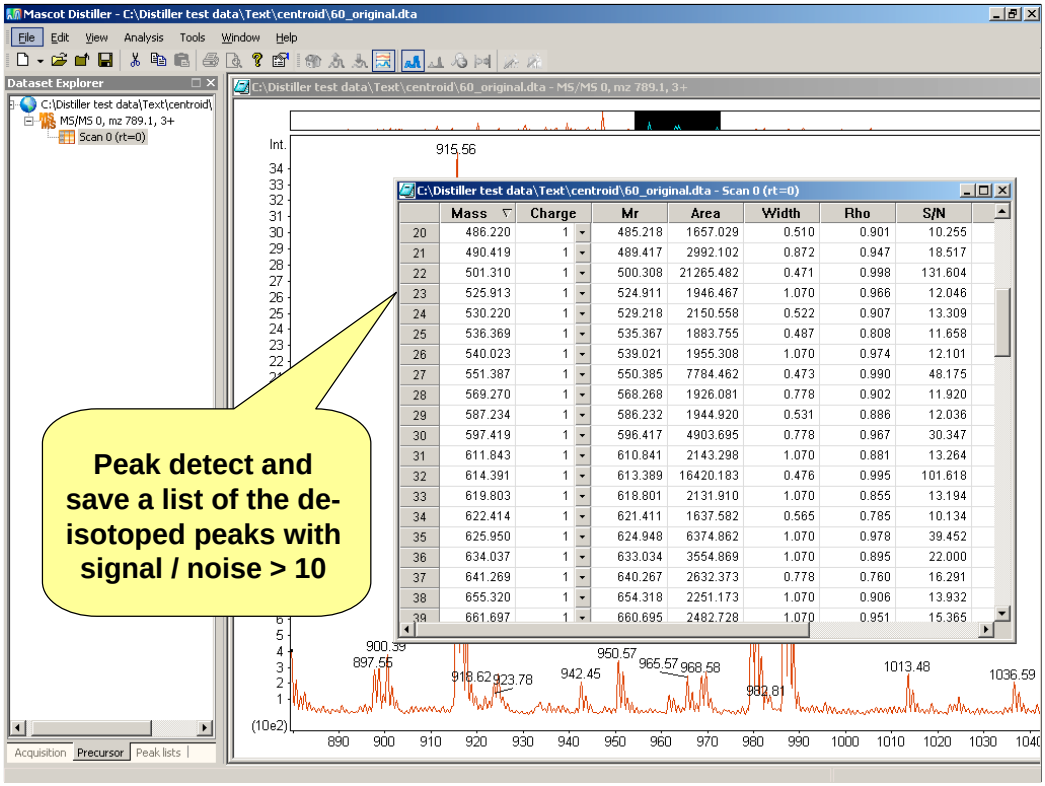
This is a DTA file from centroided LCQ,data and as you can it gets a very respectable Mascot score of 96



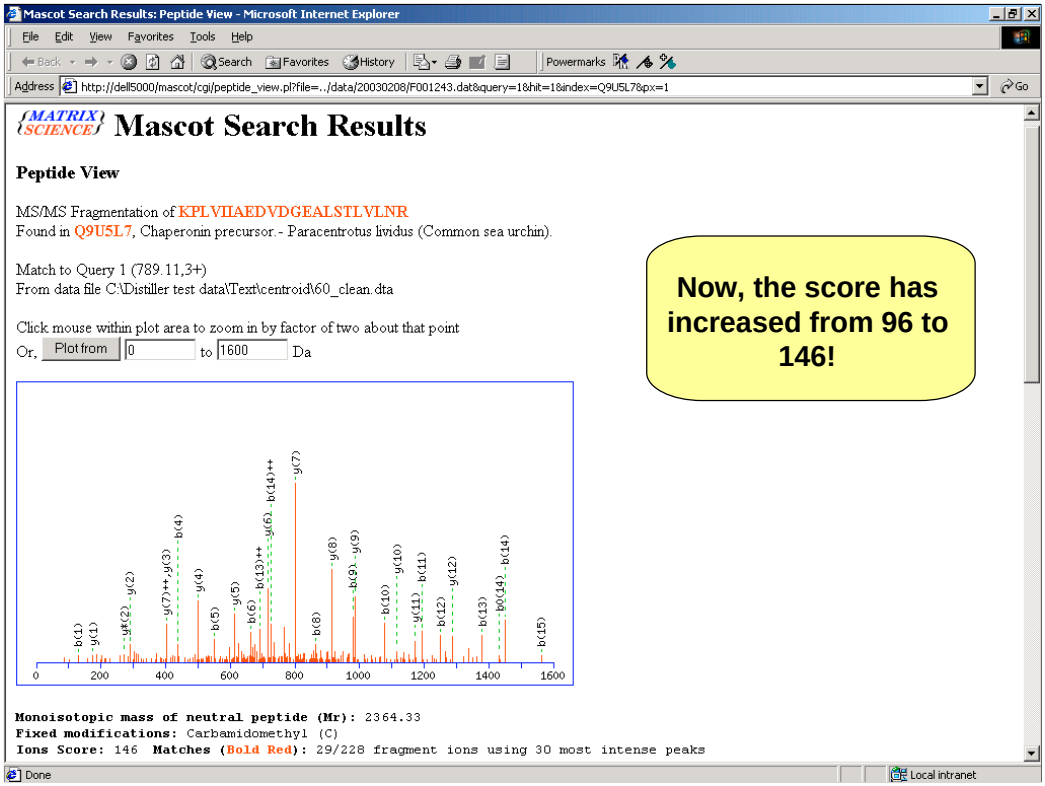
We can import the file into Mascot



Each centroid is convoluted with a gaussian profile to reconstruct an approximation to the original profile data.



We can then perform peak detection using Distiller's algorithm



This de-isotopes the spectrum. If the same spectrum is now searched again in Mascot, we see a significant improvement in the score.

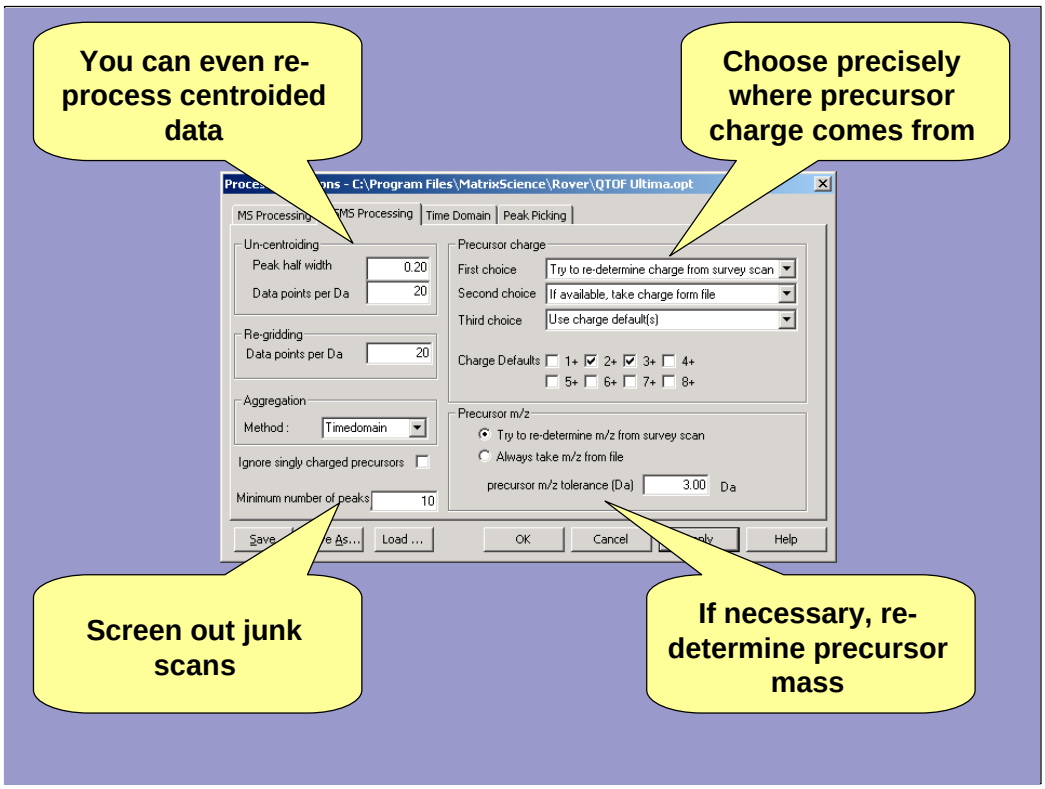
Issues with data reduction s/w

- Peak detection issues
- *Incorrect precursor charge or mass determination for ms-ms*
- *Poor grouping of similar spectra*
- Inability to recognise and remove 'junk' spectra

ASMS 2003

MATRIX SCIENCE

Moving on to the last three items here, the incorrect precursor mass or charge determination, the poor grouping of similar spectra and the inability to remove junk spectra



Mascot Distiller allows precise control over where the precursor charge state comes from. As you can see, there are three choices for getting the charge - it only tries the next stage if the first one fails. In this case, the first choice is to try and determine the charge state from the survey or zoom scan. If that fails, then it tries to take the value from the file. If the instrument software fails to determine the charge then it won't be in the file. In that case the multiple default values are use.

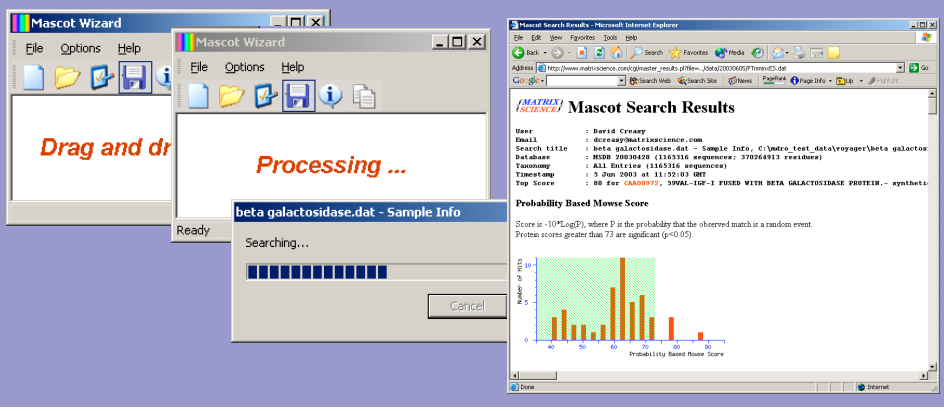
Having found the precursor charge, it is of course possible to re-determine the precursor mass.

Junk scans are rejected by insisting on a minimum of say 10 peaks. As we have seen earlier, Mascot Distiller detects peaks and not noise, so this is a reasonable option.

Finally, it is possible to re-process centroided data. LCQ data is typically saved as centroided data.

Mascot Wizard

- Free tool for peptide mass fingerpint only
- Will be available for download from our web site



ASMS 2003

MATRIX SCIENCE

As mentioned earlier, we will be releasing a new product within the next few months - Mascot Wizard.

This will be a free tool for peptide mass fingerprint searches

It will be available for download from our web site

For many people who have automated their protein discovery pipelines, the first stage of data processing has become a stumbling block. Some people have written their own data processing software. Mascot distiller allows a programmer to get raw spectrum data, peak lists and other useful file information in a consistent manner for all the major instruments. There is no other product like this.

Secondly, Mascot Daemon will use Mascot distiller engine.

Mascot Distiller - summary

- **Distiller consists of two parts:**
  - **An 'engine' (A Windows COM library)**
  - **A powerful Windows application**
- **Support for all the major instruments**
- **Common user interface for all instruments**
- **Mascot 2.0 - Daemon will use the engine**
- **Mascot Wizard available shortly**

ASMS 2003

MATRIX SCIENCE

Distiller consists of two parts:
- An 'engine' (A Windows COM library)
- A powerful Windows application

Support for all the major instruments

Common user interface for all instruments

Mascot 2.0 - Daemon will use the engine

Mascot Wizard available shortly