

Mascot Distiller

***a new tool for data
reduction***

*{MATRIX}
{SCIENCE}*

Mascot Distiller - a tool for Data Reduction

- *What is data reduction and why do we need it?*
- *Problems and issues with most peak detection and data reduction software*
- *Mascot Distiller - the solution?*

*{MATRIX}
{SCIENCE}*

For this final session, I would like to spend a little time describing a new product that we are developing - Mascot Distiller.

Mascot Distiller is used for 'data reduction' and in this session I want to describe:

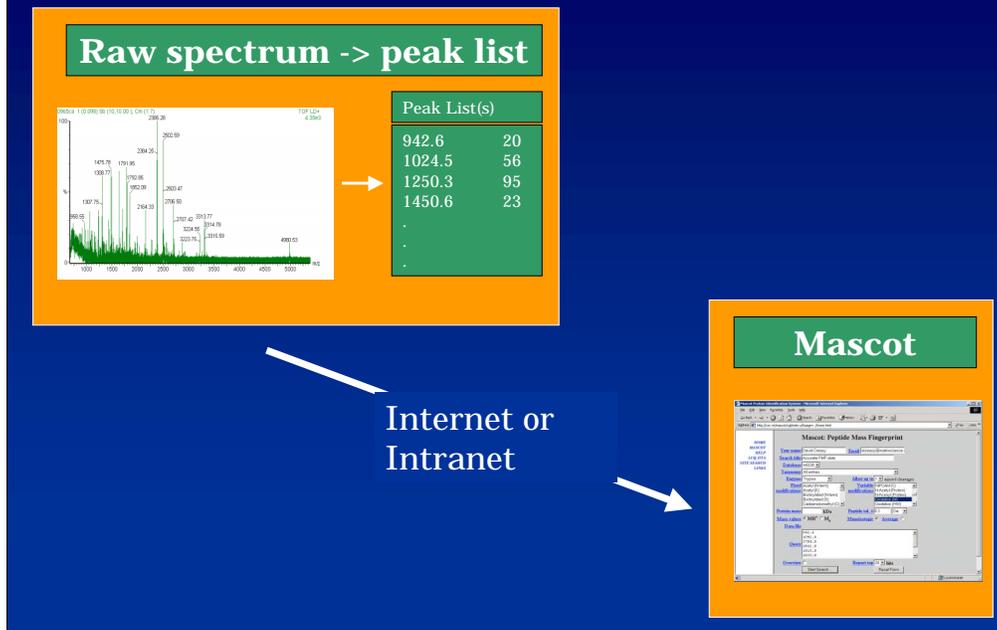
what we mean by data reduction,

what are some of the issues with peak detection and data reduction software.

During the session , I will show some examples from Mascot Distiller.

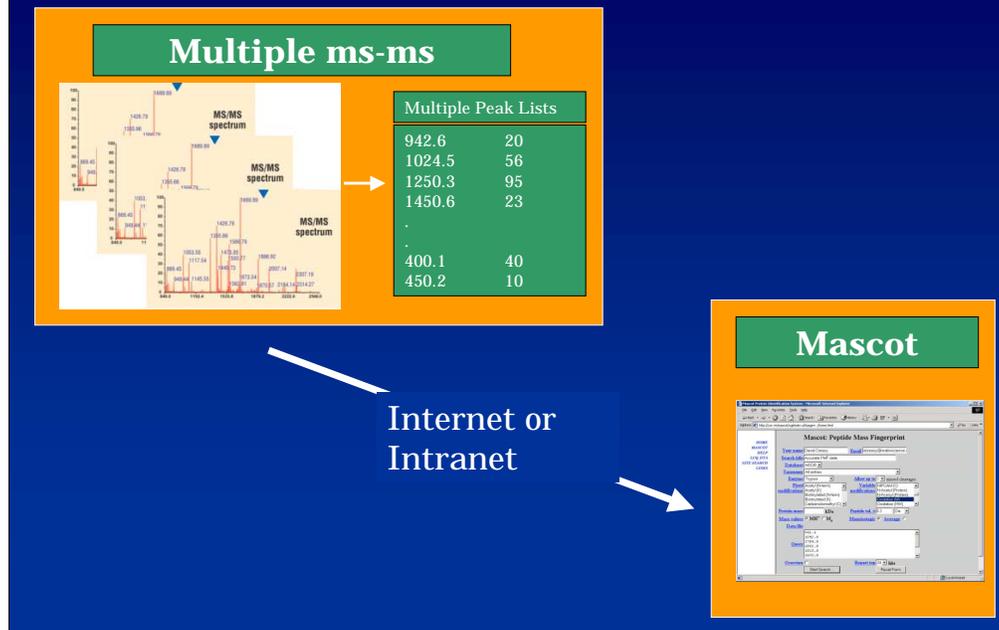
So, first of all, what do we mean by data reduction

Data reduction - Maldi data



This is the simplest example of data reduction. This is a typical maldi spectrum which contains about 50,000 data points. We need to reduce this to somewhere between about 30 and 150 peaks. Unfortunately, this is not as trivial as it seems, and we are still seeing people perform this process manually for some spectra.

Data reduction - MS-MS / LC MS-MS



More complex examples are for LC MS-MS data or nano spray ms-ms data. Here we have four issues:

- Firstly, as with the ms data, we need to get a peak list from each spectrum
- Secondly, we need to understand the structure of the file. There is no value in detecting peaks in a zoom scan for example and sending this data to Mascot. Clearly the data from a zoom scan needs to be used for accurate precursor mass and charge determination
- Thirdly, we need to average together spectra from the same peptide
- Fourthly, in any lc ms-ms run, there will be many 'junk' spectra. Ideally, we should discard these at this stage rather than submitting them to Mascot. You would be amazed to see how many ms-ms spectra with a single peak get submitted to the Mascot public web server

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**

{MATRIX}
{SCIENCE}

The issues with most data reduction software can be broken down as shown here

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

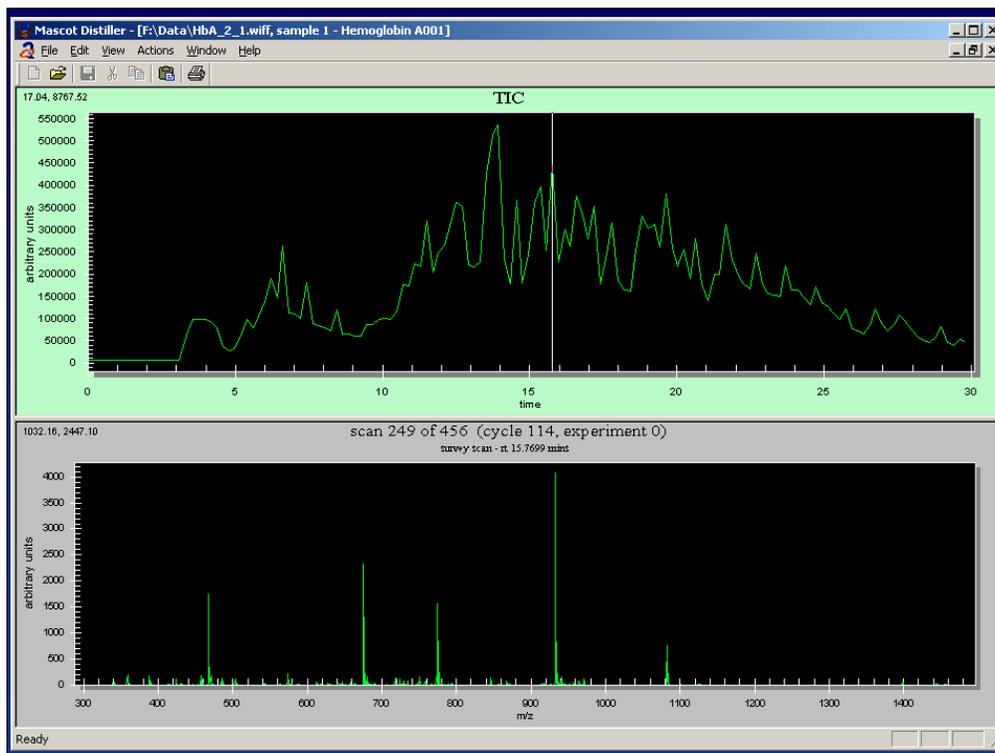
Peak detection issues

- **Low intensity peaks are often missed**
- **Software will fail if you need to set an intensity threshold**

{MATRIX}
{SCIENCE}

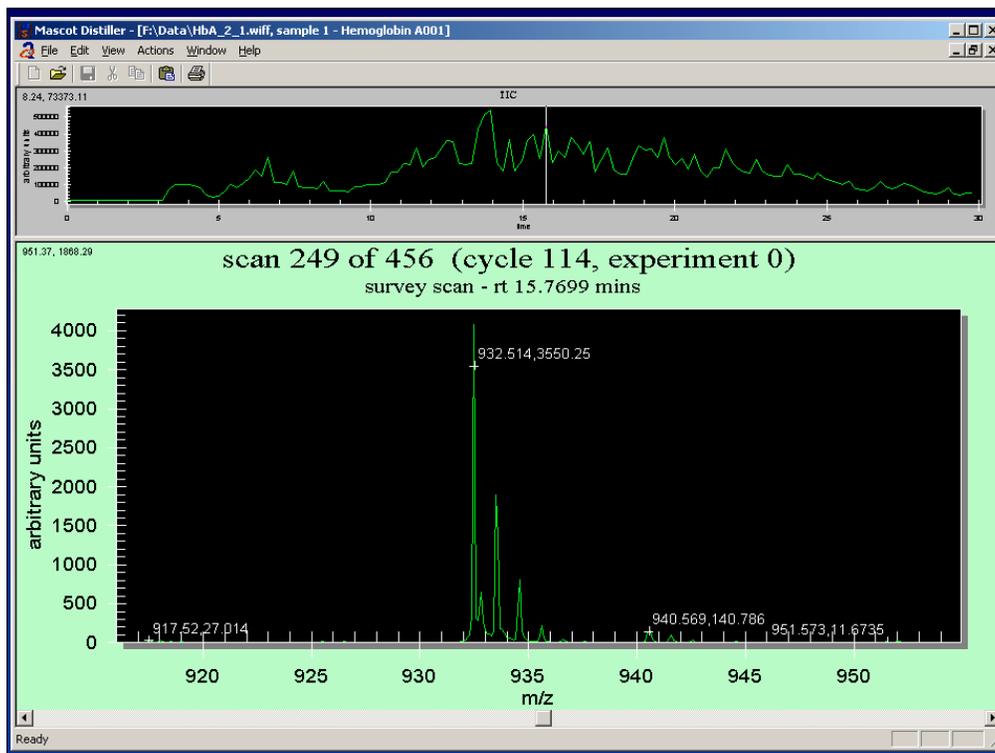
Firstly then, I'm going to describe some peak detection issues.

One key issue that we see is that high quality, but low intensity peaks are often missed. With many peak detection routines, you are required to set a threshold for peak detection. However, low intensity peaks are often real, and disregarding these will give a poorer match. I'll just show you an example:



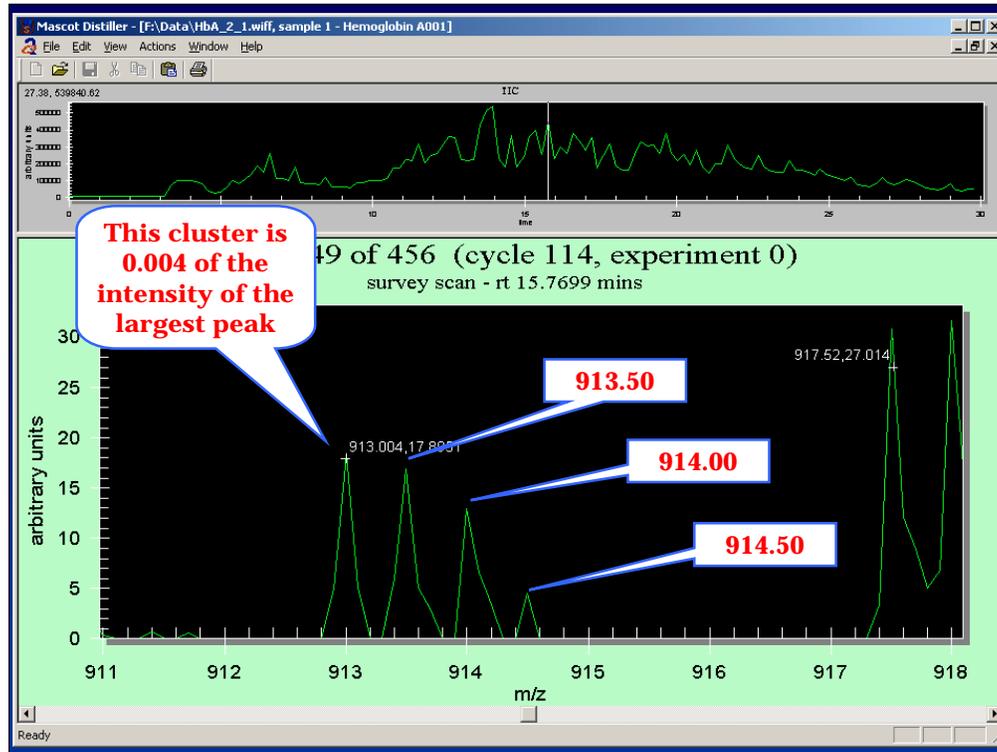
Here we can see some data from a Sciex QStar. This is very nice data, and from this survey scan we can clearly see some large peaks that any respectable software will detect.

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.

Peak detection problems

- **Picking peaks that are just noise**
- **Selection of the wrong peak(s) in an isotopic cluster**

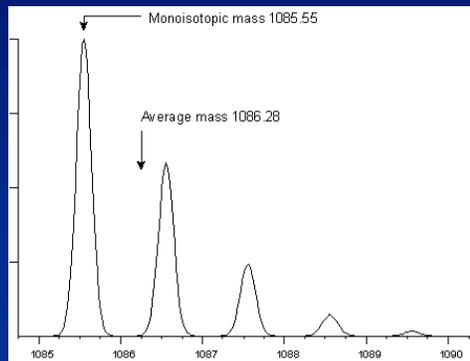
{MATRIX}
{SCIENCE}

Another classic problem with peak detection is that either the wrong peak or too many peaks are selected from an isotopic cluster.

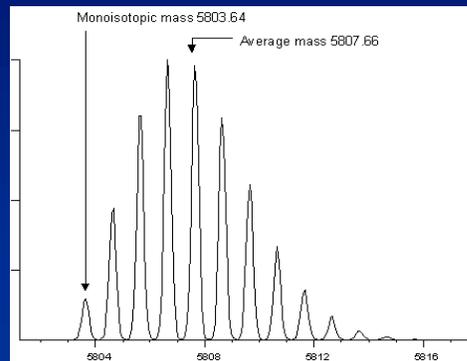
As I am sure that many of you are aware, the cluster shape changes depending on mass and resolution.

Selecting first peak in envelope

Peptide: HLKTEAEMK



Insulin

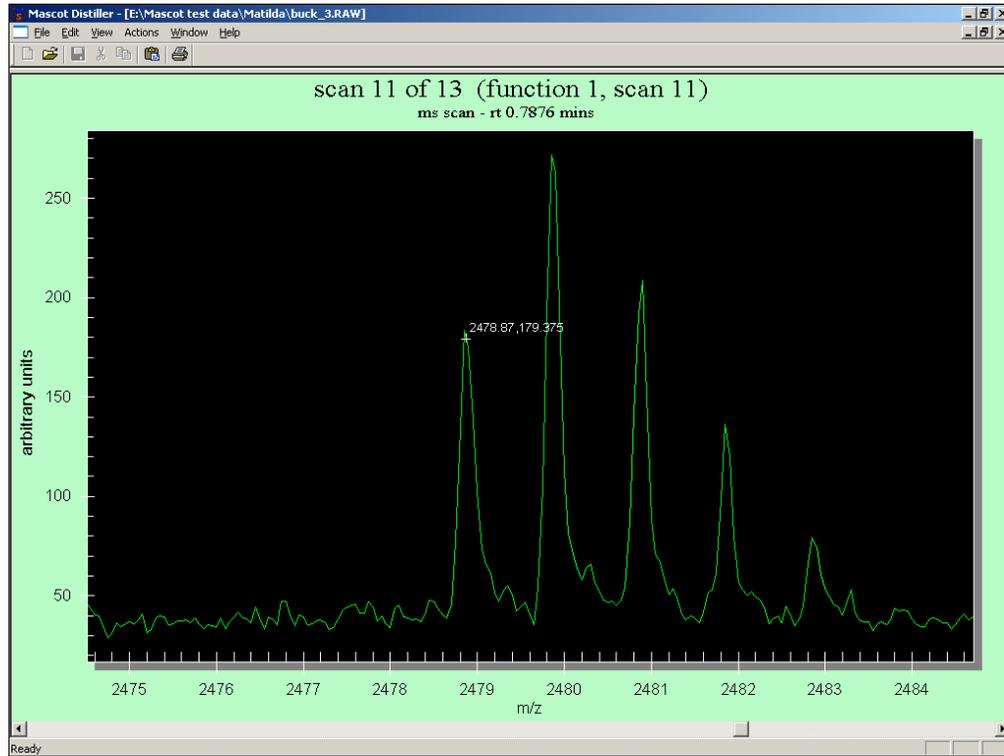


{MATRIX}
{SCIENCE}

For a peptide of mass 1000 at a resolution of about 5000 we expect to see a peak cluster like this. The first peak is of course the monoisotopic peak, and at this mass and resolution will be the largest peak.

At mass 5000, the shape is very different, and it is going to be harder to pick out the first peak which is quite small.

Somewhere between these two masses...



We have a cluster shaped like this, and you can see that Mascot Distiller has correctly identified the first peak in the cluster.

Peak detection

- **Uses similar method to that described by:**

Peter Berndt, Uwe Hobohm and Hanno Langen in
Electrophoresis 1999, 20, 3521-3526 section 2.3 'Peak
detection in mass spectrometric data'

and

Robin Gras et. al in Electrophoresis 1999, 20, 2535-3550
"Improving protein identification from ... and optimized
peak detection"

- **Mascot Distiller iteratively calculates the ideal isotopic peak shape at a given mass, searching for the best correlation with the real data**

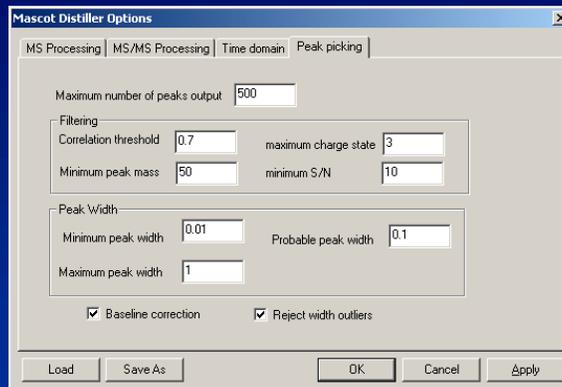
{MATRIX}
{SCIENCE}

The method that we use to implement the peak detection is not new, but is based upon the work of some of other experts in this field. The technique has been described by Peter Berndt and colleagues at Hoffman La Roche, and also by Robin Gras, Ron Appel and others at the SIB in Geneva.

The technique is to calculate an isotopic distribution at a given mass and try and correlate this with a small section in the spectrum. If the correlation is good, then the calculated distribution is removed and the process is continued. This is an iterative process, and is computationally very intensive.

So what we have done is essentially nothing new, but we have implemented it in a robust and efficient manner, making the finished product available to many people.

Peak detection parameters



The screenshot shows the 'Mascot Distiller Options' dialog box with the 'Peak picking' tab selected. The parameters are as follows:

Category	Parameter	Value
General	Maximum number of peaks output	500
Filtering	Correlation threshold	0.7
	Maximum charge state	3
	Minimum peak mass	50
	Minimum S/N	10
Peak Width	Minimum peak width	0.01
	Probable peak width	0.1
	Maximum peak width	1
Options	Baseline correction	<input checked="" type="checkbox"/>
Options	Reject width outliers	<input checked="" type="checkbox"/>

Buttons at the bottom: Load, Save As, OK, Cancel, Apply.

{MATRIX}
{SCIENCE}

We have observed people performing peak detection using standard packages and listened to peoples feedback. The process that many people follow is to 'tweak' parameters using trial and error until they get the desired result. In most cases the users don't really understand what many or all of the parameters mean. To compound the problem, different instrument manufacturers use different techniques and different terms, making it almost impossible for a user of several different instruments to achieve optimal results.

Ideally, there would be no parameters at all, but this is simply not achievable. Our goal is that there should be a fixed set of parameters per instrument, that doesn't require any tweaking.

Peak detection problems

- **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**

{MATRIX}
{SCIENCE}

To summarise the peak detection issues, the problems with many packages are

Precursor Mass/Charge

- Precursor mass is often incorrect due to picking wrong peak in isotopic cluster
- Recent support call: “Mascot ICAT doesn’t work” turns out to be a 4.5 Da error in precursor mass! Value in survey scan was correct...
- Charge determination often wrong - has to be determined from survey or zoom scan

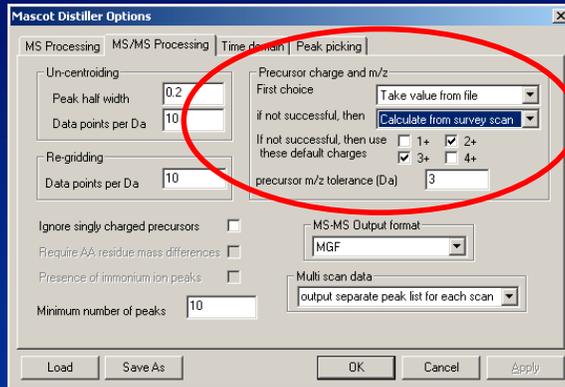
{MATRIX}
{SCIENCE}

The next big issue where things can go wrong with data reduction is failure to get the correct precursor mass and or charge for ms-ms data.

A very recent example of this was when I had a support call from a customer who said that Mascot was not working properly with ICAT data. On further investigation it turned out that the precursor mass had an error of 4.5 Da. The person concerned checked the survey scan and it soon became clear that there was a problem in the instrument data system. The problem with these errors is that it is impossible to get a match at all if there is an error, and the error is often hard to track down. More typically we frequently see errors of 1.0 or 2.0 Daltons. Of course, it is easy in Mascot to cope with this - simply set the precursor tolerance to be 2.0 or 2.5 Daltons.

Incorrect charge determination is more of a problem, and generally results in a total failure to obtain a match.

Precursor charge and mass



- Options are:
 - take value from file
 - calculate from survey/zoom scan
 - use one or more defaults
- Charge *and* mass are calculated

{MATRIX}
{SCIENCE}

There are three ways to determine the precursor mass and charge. Naturally, Mascot Distiller gives you all three choices, and most importantly allows you to decide on one method and if that fails try the next.

The choices are:

- take the value from the file. The mass value will normally be correct, although, as we have seen, there can be severe problems, and some values may be 1 or 2 Daltons out due to problems choosing the incorrect peak from an isotopic envelope. Furthermore, at least one system only saves the value as an integer.

- the second choice is to calculate the values from the survey or zoom scan. Mascot distiller uses its peak detection routines to find a peak nearest to the one in the file, and also determines the charge. If the survey or zoom scan is poor quality, then this will obviously fail. Obviously if there is no survey or zoom scan in the file, then this will not succeed.

- the final fail safe is to try a number of different charge states.

Issues with time domain processing

<input checked="" type="checkbox"/>	58	543.68	1085.35	1085.55	-0.20	1	34	1	HLKTEAEMK
<input checked="" type="checkbox"/>	82	636.18	1270.35	1270.66	-0.30	0	(52)	1	LFTGHPETLEK
<input checked="" type="checkbox"/>	83	636.21	1270.40	1270.66	-0.25	0	(43)	1	LFTGHPETLEK
<input checked="" type="checkbox"/>	84	636.23	1270.44	1270.66	-0.21	0	59	1	LFTGHPETLEK
<input checked="" type="checkbox"/>	89	680.90	1359.79	1359.75	0.04	1	41	1	ALELFRNDIAAK
<input checked="" type="checkbox"/>	90	689.82	1377.63	1377.83	-0.20	0	80	1	HGTVVLTALGGILK
<input checked="" type="checkbox"/>	91	689.89	1377.76	1377.83	-0.07	0	(18)	2	HGTVVLTALGGILK
<input checked="" type="checkbox"/>	93	751.64	1501.27	1501.66	-0.39	0	(56)	1	HPGDFGADAQGAMTK
<input checked="" type="checkbox"/>	94	751.71	1501.41	1501.66	-0.25	0	63	1	HPGDFGADAQGAMTK
<input checked="" type="checkbox"/>	95	752.14	1502.27	1501.66	0.61	0	(56)	1	HPGDFGADAQGAMTK
<input checked="" type="checkbox"/>	96	752.17	1502.32	1501.66	0.66	0	(39)	1	HPGDFGADAQGAMTK
<input checked="" type="checkbox"/>	97	752.25	1502.49	1501.66	0.83	0	(44)	1	HPGDFGADAQGAMTK
<input checked="" type="checkbox"/>	98	753.77	1505.52	1505.93	-0.41	1	78	1	HGTVVLTALGGILK
<input checked="" type="checkbox"/>	99	753.88	1505.74	1505.93	-0.19	1	(62)	1	HGTVVLTALGGILK
<input checked="" type="checkbox"/>	103	803.78	1605.54	1605.85	-0.31	0	(80)	1	VEADIAGHGQEVLR
<input checked="" type="checkbox"/>	104	803.80	1605.58	1605.85	-0.27	0	(68)	1	VEADIAGHGQEVLR
<input checked="" type="checkbox"/>	105	803.81	1605.60	1605.85	-0.25	0	86	1	VEADIAGHGQEVLR

It is quite common to see Mascot results as shown here. As you can see, there are many identical peptides. Better signal to noise can be obtained by averaging these spectra together.

The principles for doing this are quite simple, but somewhat surprisingly, many software packages don't provide nearly enough control over which scans will be grouped together

Options for grouping scans

The screenshot shows the 'MDRO Options' dialog box with the 'Time domain' tab selected. The dialog contains several input fields for configuring scan grouping parameters:

Parameter	Value
Minimum precursor mass (Mr)	500
Maximum precursor mass (Mr)	5000
Precursor m/z tolerance for grouping	1
Max number of intermediate scans	1
Minimum number of scans in a group	2
Start time	0
End time	0

At the bottom of the dialog, there are buttons for 'Load', 'Save As', 'OK', 'Cancel', and 'Apply'.

{MATRIX}
{SCIENCE}

Time Domain parameters control how scans are grouped together

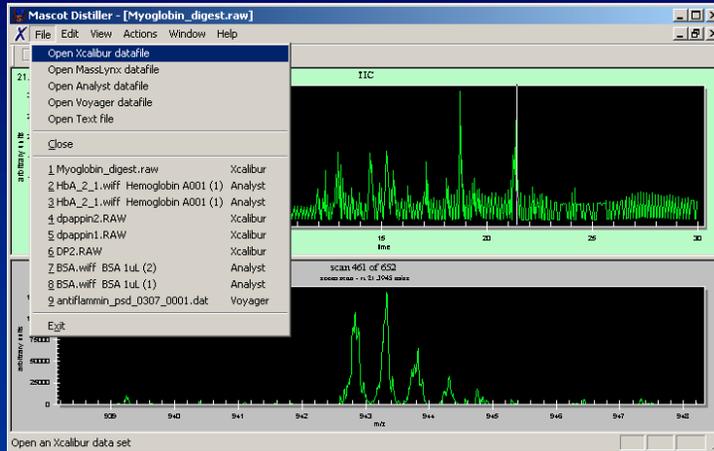
Mascot Distiller

- **Reads RAW data from all the manufacturers**
- **“Understands”:**
 - single maldi spectrum
 - single ms-ms spectrum
 - multiple ms-ms spectra (as in nano spray)
 - triple play (survey, zoom, ms-ms, survey, zoom...)
 - lc ms-ms (survey, ms-ms, survey, ms-ms)
- **Produces high quality peak lists**

{MATRIX}
{SCIENCE}

The features of Mascot Distiller are ...

Current GUI



MATRIX
SCIENCE

The GUI is very basic, and is only intended as a debugging tool.

Mascot Distiller

- **Important part is a COM library for Windows**
 - Can be called by Mascot Daemon
 - Can be used by third party software
- **Should be available this fall**

{MATRIX}
{SCIENCE}

More importantly, Mascot Distiller is a function library that can be called from Mascot Daemon or from your own code

Acknowledgements

Mascot Distiller is being developed in collaboration with:

- **BioVisioN in Hannover, Germany**
- **GSK in Philadelphia (Roland Annan's Group)**

{MATRIX}
{SCIENCE}

The development has been a three way partnership with Biovision and GSK

MASCOT[®]
Take the guesswork out of protein identification...



{MATRIX}
{SCIENCE}
www.matrixscience.com