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



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.

Why do we need Mascot Distiller?

- It provides a single user interface for processing data
- Quality of data reduction in instrument vendor software is 'variable'
- Not possible to automate the output of peak lists from some data systems

ASMS 2003



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.

DataSourceCollection	Distiller	Preferences
{ DataSourceObj }	ProcessingOptions	AminoAcidCollection
DataSourceHeader	MSProcessingOptions	{ AminoAcid }
DataSourceParams	MSMSProcessingOptions	ElementCollection
ScanCollection	PeakSelectionOptions	{ Element }
{ Scan }	ТІС	IsotopeCollection
	Spectrum	{ Isotope }
A Windows COM	PeakListCollection	IsotopicDistributionFactory
library provides	{ PeakList }	IsotopicDistribution
an API to Distiller	PeakCollection	
- Gan be called	{Peak}	LogEvents
Daemon	Spectrum	
- Can be called	ProcessingOptions +	
by third party	PeakListCollectionEvents	
software		
ASMS 2003		(MATRIX) SCIENCE)

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.

<section-header><section-header><section-header><section-header><list-item><list-item><list-item><list-item><list-item>

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.



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.

Select XCalibur data set	2 2	<u>×</u>
Look jr: data I	SampleO03_01_003.RAW SampleO03_01_003.RAW W S.RAW W W W W W W W W W V W V V	Choose processing options when opening the file File information to help locate the correct file
ASMS 2003		(MATRIX SCIENCE)

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.

Mascot Distiller - understands

- Single MS spectrum
- Single MS/MS spectrum
- Multiple MS spectra
- Multiple MS/MS spectra (e.g. nano spray)
- 'Triple play' (survey, enhanced/zoom, MS/MS)
- LC-MS/MS (complex arrangements of survey and MS/MS scans)

ASMS 2003



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

Mascot Distiller - Myoglobin_digest.raw								_ 8 ×
File Edit View Analysis Tools Window Help								
Dataset Explorer	.raw - TIC						[
E		\						
Bummed spectr	2			im	-l	A		
	a			1				
has labelled pea	ks			1				
🗄 😽 Summed Scans 🛛 🔰 25 -	Mun	India diamata			- ()	200 (-1		r II
□ m/z 752.7, 1+, Scan 292 (rtz	Myog	iobin_aigest.	raw - m/2 63t	5.4, Z+, SUM	or 3 scans in r	ange 299 (Pt=.	-비스	
🗄 📲 Summed Scans 20 .		m/z ⊽	Charge	Mr	Area	Width	Rho 🔺	
E m/z 636.4, 2+, Sum o scans in range 29	1	427.244	1 -	426.242	286081.100	0.496	0.91	
E> Summed Scans S MS/MS 200 m/z 626 4 21 15 1	2	490.171	1 -	489.169	154847.932	0.471	0.91	
M5/M5 392, m/z 636.5, 2+	3	497.473	1 -	496.470	921592.110	0.492	0.91	
M5/M5 305, 636.4, 2+ 10	4	506.419	1 -	505.417	072253.673	0.465	0.9	
e III m/z 650.2, 1+, Scan 308 4.8312)	5	510.256	1 -	509.254	640039169	0.470	0.91	
E Summed Scans		500.400	1	503.234	593555.100	0.406	0.0	m
m/2 4/1.4, 2+, 5um of 2 scans in	- 0	520.192		527.180	504420.300	0.400	0.8	
E miz 804.2. 2+. Sum of 4 scans		538.094	1 -	537.092	540329.480	0.467	0.9	
Spectra that	8	556.151	1 -	555.149	505316.549	0.467	0.9!	Sec
m/z 832.0, 2+, Scan 326 (rt=15	9	570.615	1 -	569.613	138809.639	0.668	0.9	
🗄 📲 Summed Scans 🛛 🔍 WEI'E SUITIITIEU	10	579.909	1 -	578.906	641454.911	0.482	0.9	<u>م مع</u>
m/z 804.6, 1+, Sum of 2 scans in range 33	11	618.436	2 -	1234.867	895589.913	0.531	0.9	
±-∞∑ Summed Scans	12	627 566	1 -	626 564	772987 508	0.472	0.9	
The summed Scans 25	12	609.749	1 -	607 246	892000 421	0.607	0.01	
m/z 969.3, 1+, 5um of 2 scans in range 34	13	746.220	1	745.000	002000.421	0.307	0.0	
😥 📲 Summed Scans	14	710.228	1 -	/15.226 :	1309771.048	0.471	0.9	L 11
m/z 681.6, 1+, Scan 353 (rt=16.7593)	15	782.210	1 💌	781.208	478922.835	0.476	0.91	
±-∞∑ Summed Scans T	16	853.278	1 -	852.276	737601.264	0.481	0.9	L 11
10/2 090.0, 14, Scans	17	865.352	1 -	864.350	957869.643	0.498	0.9:	L 11
B m/z 754.2, 2+, Sum of 2 sca Dook list tob	18	910.235	1 👻	909.232	540130.526	0.480	0.9	
E Summed Scans	19	978.343	1 -	977.341	664230.638	0.467	0.9	L 11
M5/M5 385, m/z 754.2,	20	996.178	1 -	995 176	478852.325	0.488	0.9:	L 11
MS/MS 388, m/z 754.3, 2 5	•						•	
e 40		21.24	618.44 6	98.25 104.4	' 88 <mark>6</mark> .35	990. 0 11f	17 81	·
	Khunnel	when the		سلالاستعال	الساله_	U	<u> </u>	[]
	300 40	n són	ein	700 80	n gin	1000 11	nn 1200) m/z
Acquisition Precursor Peak lists			000		0 000	.000 11		/ 11/2
Ready								

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



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

- 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



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 iteratively searches for the best correlation between the expected isotope distribution at a given mass and the experimental data
- · Similar methods have been described by
 - Peter Berndt et. al., Electrophoresis (1999) 20 3521-3526
 - Robin Gras et. Al., Electrophoresis (1999) 20 2535-3550

ASMS 2003



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 $96\,$

Import the DTA file into Mascot Distiller	Ambiguous Data Data Type C MS MS Precursor m/z 789.11 precursor charge 3+ • Scan state Profile Centroided CK Cancel	
ASMS 2003		(MATRIX) (SCIENCE)

We can import the file into Mascot



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

🕷 Mascot Distiller - C:\Distiller test data\Text\centroid\60_original.dta 💶 🖻 🗙									
Elle Edit View Analysis Tools Window Help									
Dataset Explorer 🛛 🗙 🚺 🚰 🗠 Distiller test data 💶 🖉 web central/ 60. orginial dia = MS/MS.0. mp 789.1.3 ±									
C:Distiller test data/Text/centroid									
🖻 🌇 MS/MS 0, mz 789.1, 3+									
Scan 0 (rt=0) Int.	ç	915.56							
34	-								
33	C:\D	vistiller test da	ta\Text\cen	troid\60_orig	inal.dta - Scar	0 (rt=0)		_	
32 - 31 -		Mass V	Charge	Mr	Area	Width	Rho	S/N	
30 -	20	486.220	1 -	485.218	1657.029	0.510	0.901	10.255	_
29	21	490.419	1 -	489.417	2992.102	0.872	0.947	18.517	
28 27	22	501.310	1 -	500.308	21265.482	0.471	0.998	131.604	
26	23	525.913	1 -	524.911	1946.467	1.070	0.966	12.046	
25	24	530.220	1 -	529.218	2150.558	0.522	0.907	13.309	
	25	536.369	1 -	535.367	1883.755	0.487	0.808	11.658	
	26	540.023	1 -	539.021	1955.308	1.070	0.974	12.101	
	27	551.387	1 🔻	550.385	7784.462	0.473	0.990	48.175	
	28	569.270	1 -	568.268	1926.081	0.778	0.902	11.920	
	29	587.234	1 -	586.232	1944.920	0.531	0.886	12.036	
	30	597.419	1 -	596.417	4903.695	0.778	0.967	30.347	
Peak detect and	31	611.843	1 -	610.841	2143.298	1.070	0.881	13.264	
r oan dotoot and	32	614.391	1 -	613.389	16420.183	0.476	0.995	101.618	
save a list of the de-	33	619.803	1 -	618.801	2131.910	1.070	0.855	13.194	
icotonod nooko with	34	622.414	1 -	621.411	1637.582	0.565	0.785	10.134	_
isotoped peaks with	35	625.950	1 -	624.948	6374.862	1.070	0.978	39.452	_
signal / noise > 10	36	634.037	1 -	633.034	3554.869	1.070	0.895	22.000	_
	37	641.269	1 •	640.267	2632.373	0.778	0.760	16.291	_
	38	655.320	1 •	654.318	2251.173	1.070	0.906	13.932	
5	 	001.097	•	000.095	2482.728	1.070	0.951	10.300	
4 900 3 897.5 1 1 (10e2) 830 90	39 5 0 0 910	918.62 _{923.}	78 942.4 	950.57 965.6 15 950 960 950 960	968.58	982 <u>81</u>	101 	3.48	
Acquisition Precursor Peaklists 1									

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.



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.

Masco • Free tool • Will be a	<i>t Wizard</i> for peptide mass /ailable for downlo	fingerpint only bad from our web site				
Mascot Wizard		Mascel Search Results - Microsoft Internet Deployer				
File Options Help	lascot Wizard	Bie Edit Yew Favorites Iods 1990 😽 😪 Bauch 👷 Favorites 🜒 Media 🕢 😥 - 😓 🗔 📃				
i i i i i i i i i i i i i i i i i i i	e Options Help	Address (ii) Hose/Innen matricsdenze.com/cg/mater_results.pht/ie=loba/20000000/PTmmx/C5.dx.				
		(MATRIX) Mascot Search Results				
Drag and dr	Processing	User : Borid Creary Borid Creary Scatch 1112 : beta gulariostates.com States : beta gulariostates.com States : biology 2000 : 100514 sequence: 37020403 residence States : biology 2000 : 100514 sequence: 37020403 residence Tambrade : biology 2000 : 1110510 : 0000 Tambrade : biology 2000 : 1110510 : 0000 : 0000 Tambrade : biology 2000 : 0000 : 0000 : 0000 : 0000 : 0000 : 0000 Tambrade : biology 2000 : 0				
	beta galactosidase.dat - Sample Info	Probability Based Mowse Score				
Read	y Searching	Score is -10*Log(F), where P is the probability that the observed match is a random event. Protein scores greater than 73 are significant (p<0.05).				
	Cancel					
		Done Done				
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



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