What's new in Mascot 2.3

MASCOT

MATRIX SCIENCE

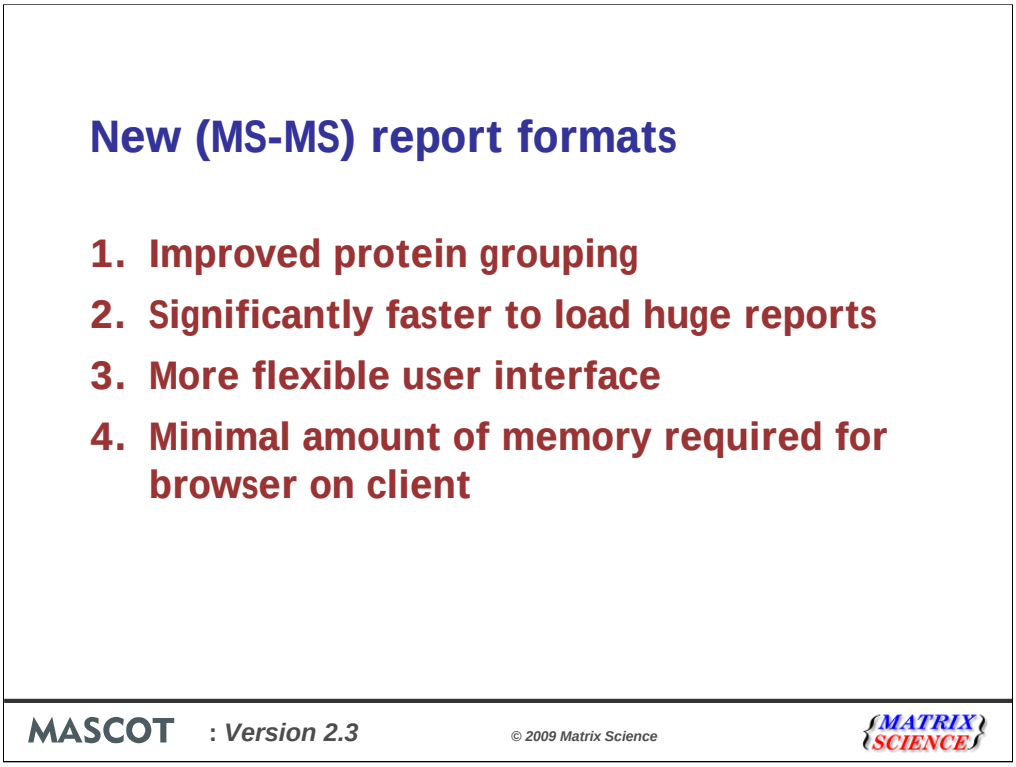
What's new in Mascot 2.3

- New report formats
- Searching multiple fasta files
- Support for HUPO PSI standard formats
- Support for Percolator
- Improvements to Mascot Daemon
- Mascot Distiller 2.3

MASCOT : Version 2.3 MATRIX SCIENCE

In this session, I'll be describing some of the changes in Mascot 2.3

I'll be describing some significant improvements to Mascot Server, Mascot Daemon and also to Mascot Distiller.



We have put a significant amount of effort into changing the way we report results from Mascot searches.

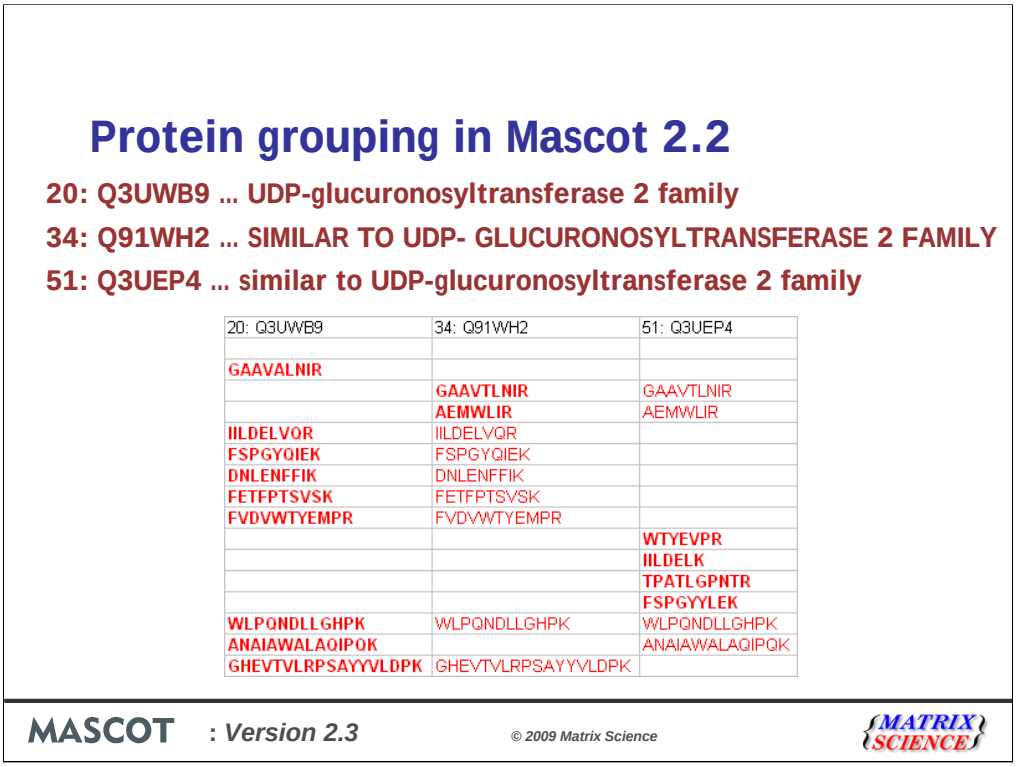
The main improvements that I want to highlight are:

•Improved protein grouping

•Significantly faster to load huge reports

•More flexible user interface

•Minimal amount of memory required for browser on client



I'd like to show an example of the protein grouping in Mascot 2.2 producing less than ideal results.

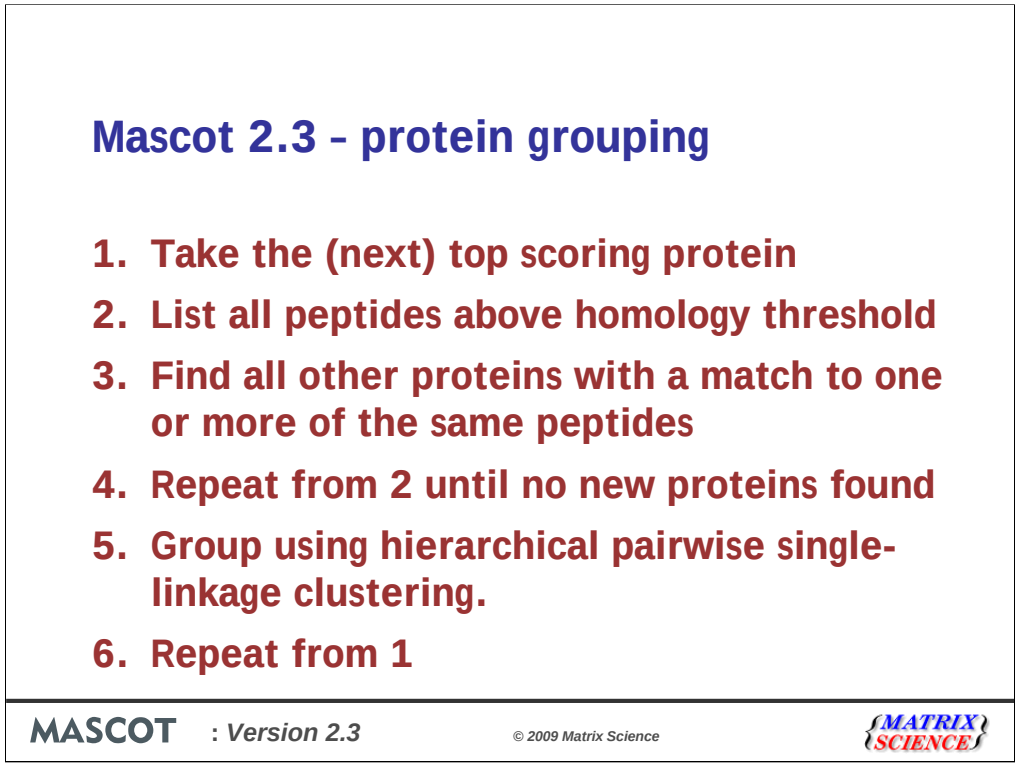
If you look at the description lines for these three proteins, I think that we would all agree that they should be grouped together. However, they are are reported as three separate proteins in Mascot 2.2

The first column shows the peptide matches for the hit 20.

Hit number 34 shares 7 peptide matches with hit number 20, but also matches these 2 additional peptides, which is why it is recorded as a separate match.

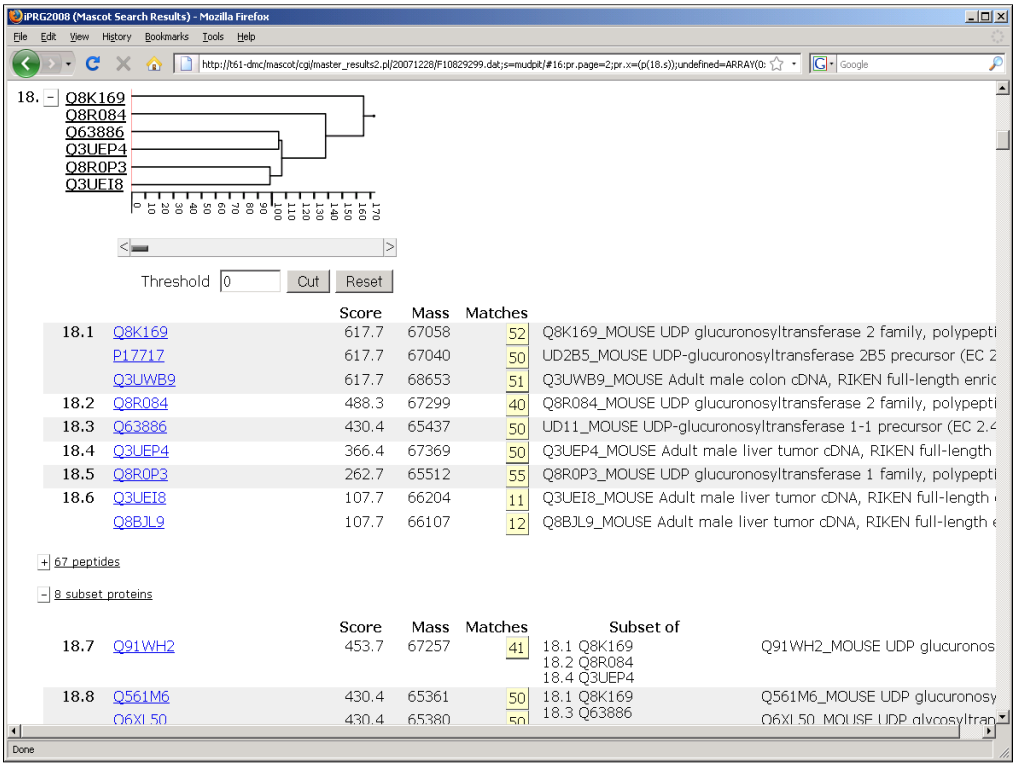
Hit number 51 has 4 new peptide matches, and the rest are either in hit 34, hit number 20 or both.

If we were to swap the order of hits 34 and 51, we can see that hit 34 just contains peptides in the other two – i.e. there are no new peptides.



In Mascot 2.3, these proteins are grouped together.

Here's an overview of how the new protein grouping works



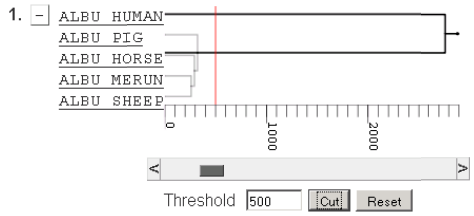
If we look at this search using the Mascot 2.3 reports, we can see that it has also grouped together a number of other glucuronosyltransferase proteins.

We can see that although they are all part of the same ‘family’, we have used hierarchical clustering to differentiate between protein groups. So, Q3UEP4 is in a separate group within the family.

As expected we can also see that Q91WH2 is a subset of Q3UEP4 and Q3UWB9

I’ll explain a little bit about the dendrogram now, but it’s easier to do this with a different example:

New reports – dendrogram



Threshold 500 Cut Reset

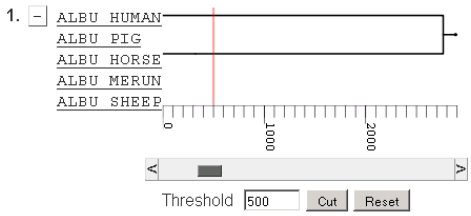
	Score	Mass	Matches	
1.1 ALBU_HUMAN	74522.0	71317	4597	Serum albumin precursor - Homo sapiens (Human)
1.2 ALBU_PIG	2640.0	71643	417	Serum albumin precursor - Sus scrofa (Pig)
1.3 ALBU_HORSE	2275.1	70550	375	Serum albumin precursor (Allergen Equ c 3) - Equus
ALBU_EQUAS	2275.1	70490	372	Serum albumin precursor - Equus asinus (Donkey)
1.4 ALBU_MERUN	721.8	70892	247	Serum albumin precursor - Meriones unguiculatus (Mo
1.5 ALBU_SHEEP	404.8	71139	176	Serum albumin precursor - Ovis aries (Sheep)

MASCOT : Version 2.3 © 2009 Matrix Science MATRIX SCIENCE

This is an example of a match to a human serum albumin protein, where we have searched all species.

The dendrogram is a pictorial representation of the hierarchical clustering. We calculate a distance tree based on the scores of significant peptides found in one protein but not the other. It's immediately obvious that the pig, horse and sheep albumins are similar. It's also obvious that there is big difference between them and the human protein. We could decide that we aren't interested in distinguishing between the different animal proteins, and cut the dendrogram at, say 500

New reports – dendrogram



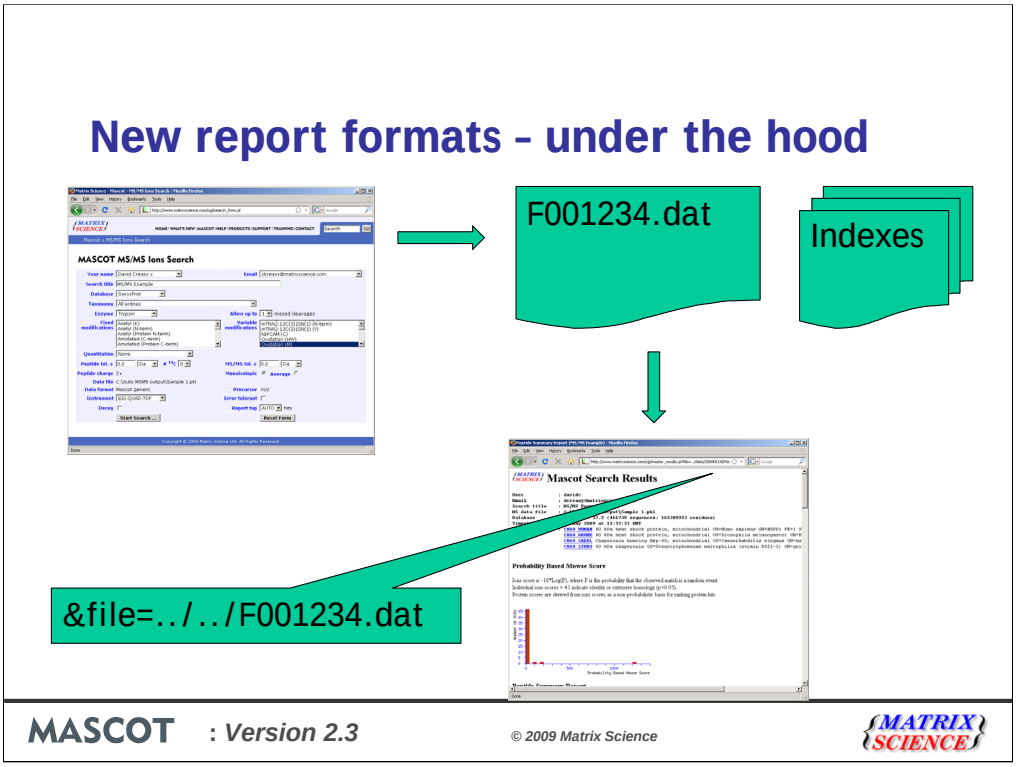
Threshold 500 Cut Reset

		Score	Mass	Matches	
☑ 1.1	ALBU_HUMAN	74522.0	71317	4597	Serum albumin precursor - Homo sapiens (Human)
☑ 1.2	ALBU_PIG	2640.0	71643	417	Serum albumin precursor - Sus scrofa (Pig)
	ALBU_HORSE	2275.1	70550	375	Serum albumin precursor (Allergen Equ c 3) - Equus
	ALBU_MERUN	721.8	70892	247	Serum albumin precursor - Meriones unguiculatus (Mc
	ALBU_SHEEP	404.8	71139	176	Serum albumin precursor - Ovis aries (Sheep)
	ALBU_EQUAS	2275.1	70490	372	Serum albumin precursor - Equus asinus (Donkey)

MASCOT : Version 2.3 © 2009 Matrix Science MATRIX SCIENCE

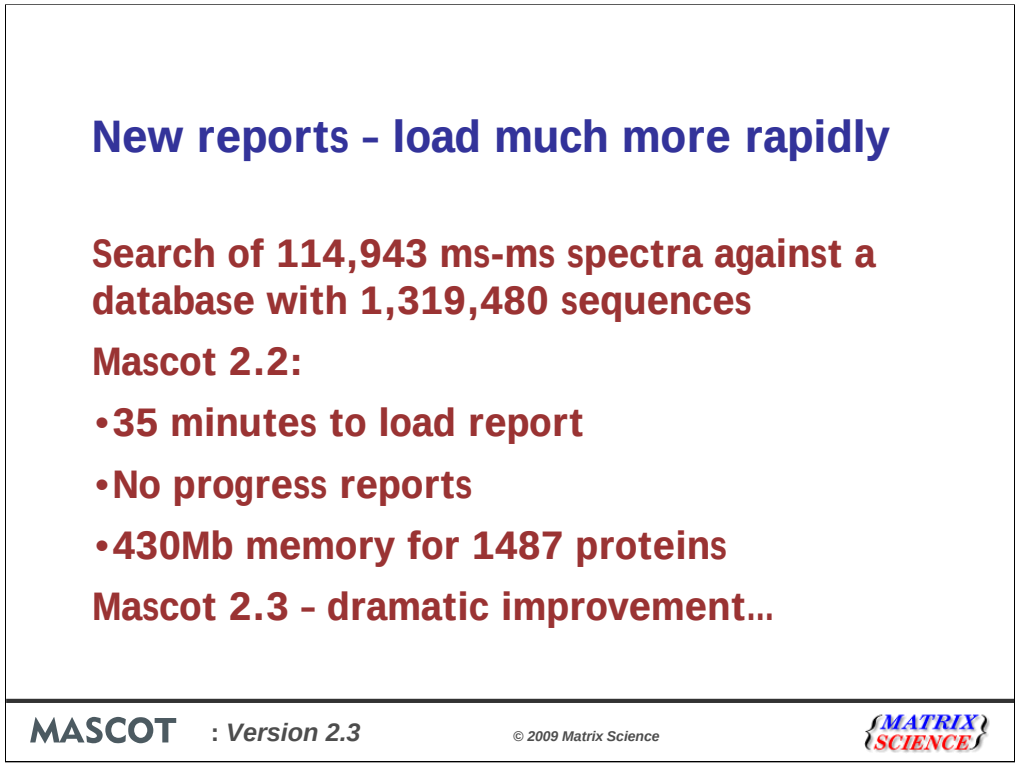
And we now just have two family members. You can see here that the horse, donkey and sheep albumin proteins are all now considered as ‘samesets’ of the pig protein.

We aren’t sure how useful this feature is, so we’d appreciate any feedback.



I will now describe some of the other changes to the reports, but first a bit of background information… when a search is submitted, it runs a cgi program called nph-mascot.exe. This program saves a file with all the search results onto the Mascot server. It automatically then loads another program which displays the results in a friendly manner.

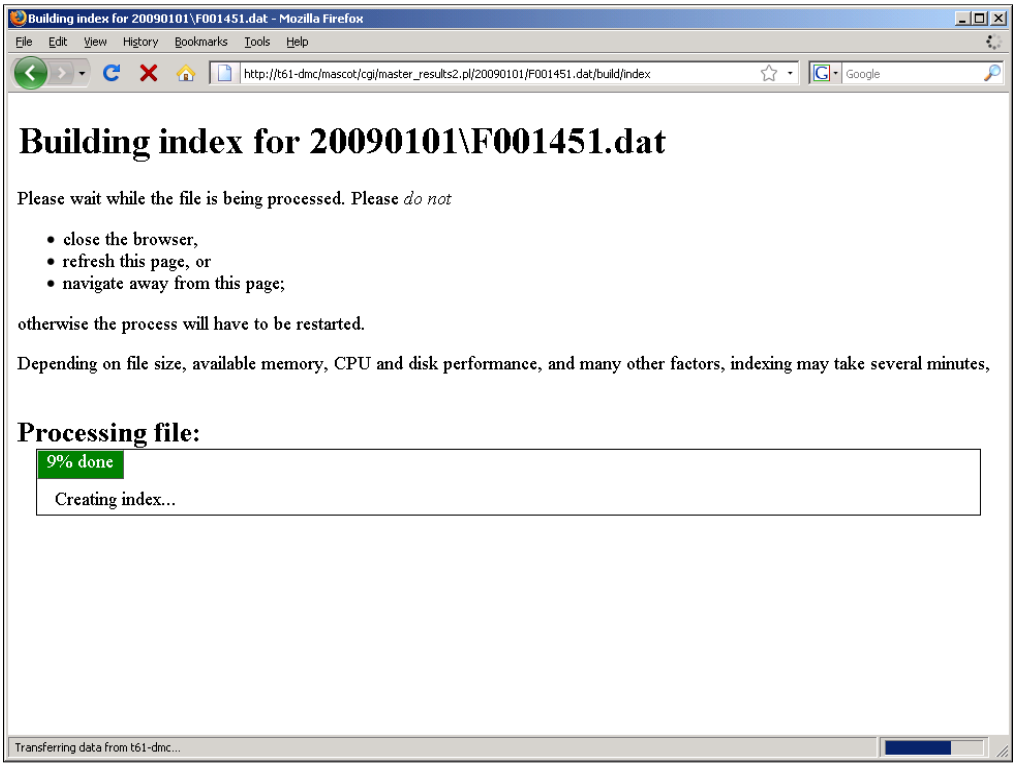
It's the changes to this second piece of software that I am describing. Because of this architecture, it means that old searches can be viewed using the new protein grouping and with the other enhancements that I'm about to describe. To help speed things up, we've also added some indexes.



The first thing that you will notice is that the reports load much more rapidly. As an example, I'm going to show a search of 114 thousand spectra against a sequence database with 1.3 million entries. The resulting Fxxxx.dat file is just over a GB in size.

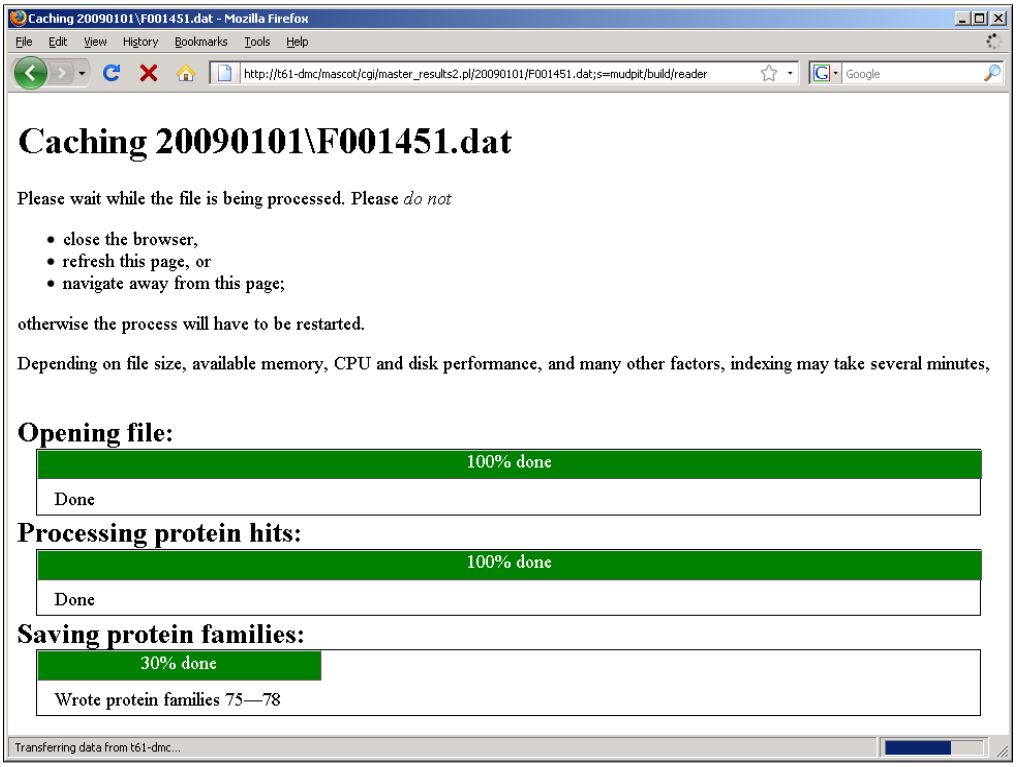
Mascot 2.2 takes 35 minutes to load the report which contains 1487 proteins. One of the problems is that there is no indication as to how much longer you need to wait for the report to be displayed.

In Mascot 2.3 …



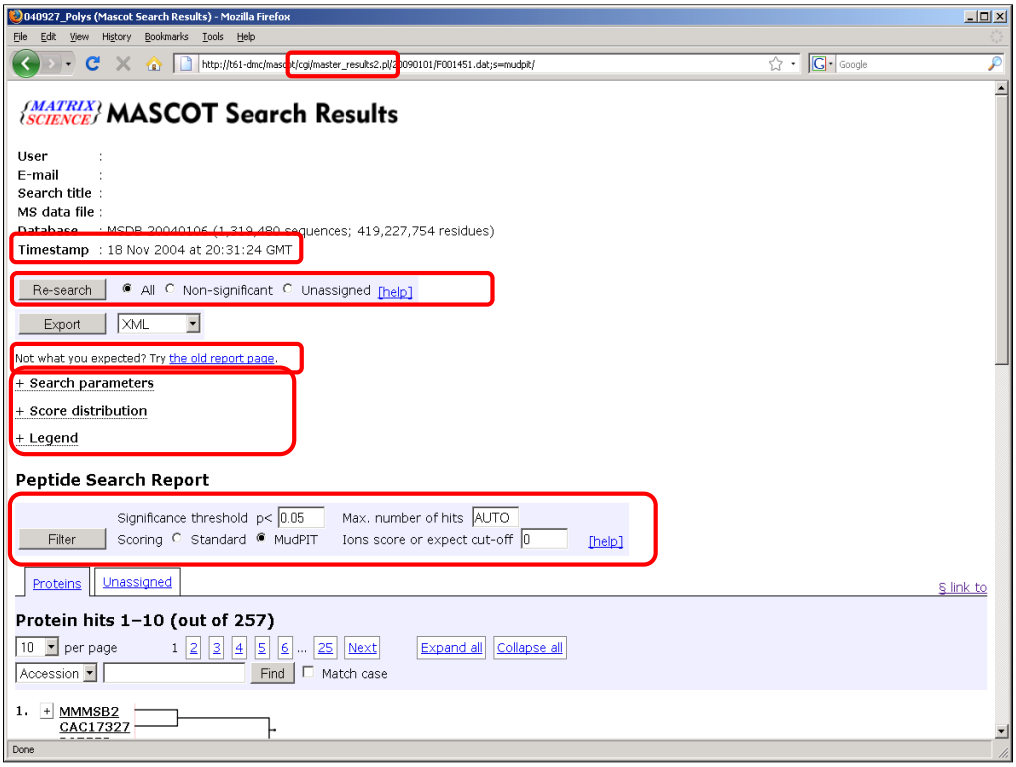
you instantly get some feedback. The first time that you load a report, various indexes get created. You may not see this page if you are running searches from Mascot Daemon, because the first index gets created as part of running the search. This part for our 1Gb file takes less than 3 minutes. I'm running this on a 2GHz laptop with a fairly slow disk, so you'd expect any recent server to be faster.

Apart from our own reports, any 3rd party programs that use Mascot Parser can also take advantage of these indexes without any change to their code.



Once the main index has been created, the file and creates an initial list of candidate proteins from the peptide hits and then groups the proteins.

It then saves the results to cache files. These 3 steps takes a further 3 and a half minutes on my laptop.



And about 5 seconds later the report loads. If you decide to re-load the report later this will take just 15 seconds because it doesn't need to re-create the cache files. So, I hope you'll agree that this is a rather dramatic improvement – 35 minutes to 15 seconds. Furthermore, it only uses 50Mb memory in the browser compared with 430Mb

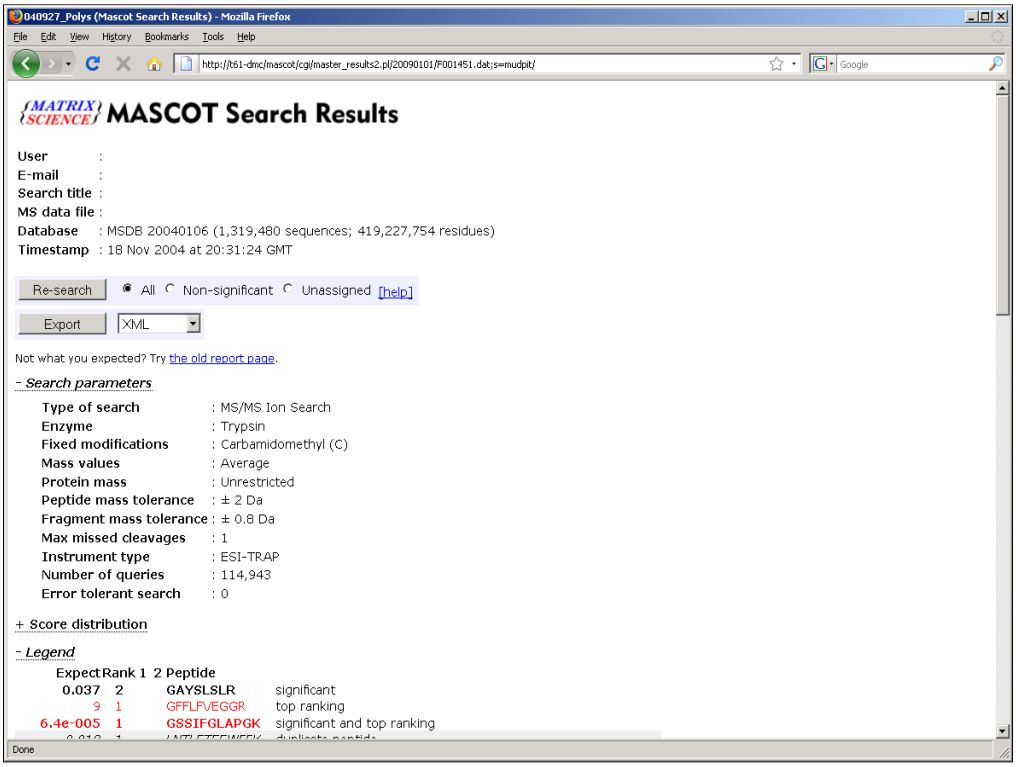
As you can see, this is an old search from 2004. As I mentioned earlier, the new reports can be used with old searches – there should be no need to repeat the search unless, for example, you want to search against a new database.

The next thing to point out is that it is now much easier to repeat a search, either for all ms-ms spectra, or just ones below the significance threshold.

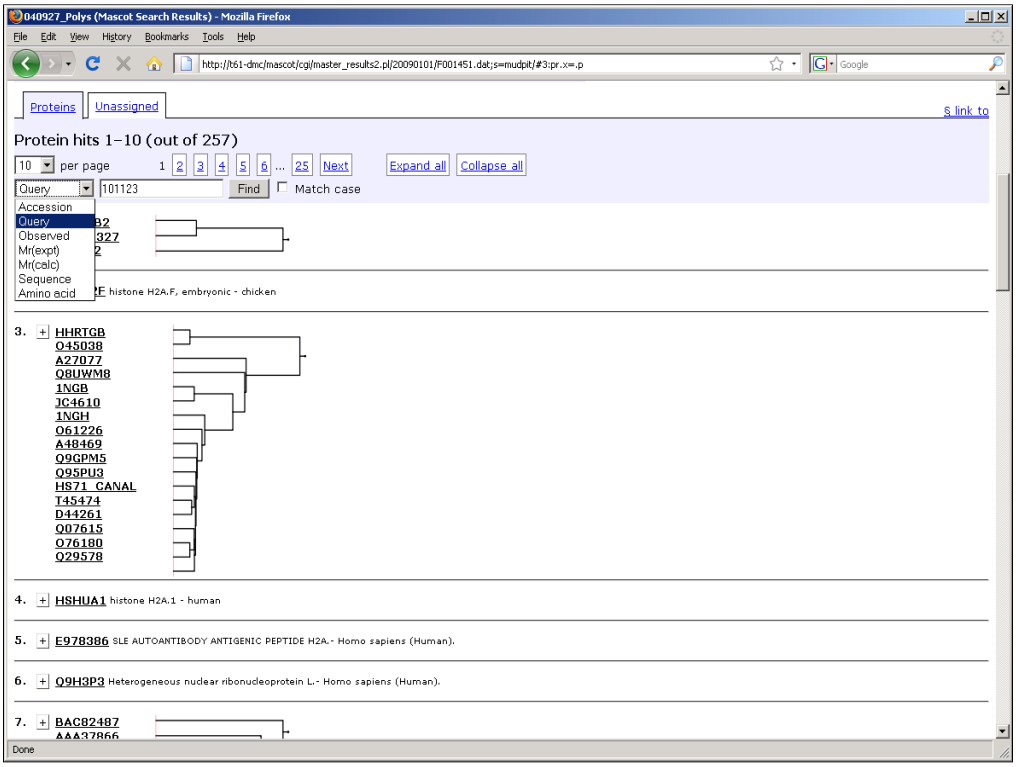
Some people will find that the new reports take a bit of getting used to. There will always be a link to the old reports, or for anyone that really doesn't like the new report format, there's a configuration option available to load the old report by default. Also, those with 20 20 vision will be able to see that the new report is called master_results2.pl – this means that any third party software that automatically loads the reports and then parses the html may still work because the original master_results.pl is almost unchanged. I do of course need to point out that this is a bad thing to do, and you really should be using the Export function for this type of task.

The filtering options are much the same as in previous versions of Mascot. If you change any of these values, new cache files need to be created, so there will be a delay. You'll notice that there is no longer a 'require bold red' option. This was useful in earlier versions of Mascot where the grouping was less sophisticated. There is also no show subsets option because we can now show or hide these for a particular protein 'on demand'.

In fact, many parts of the new reports can be expanded/contracted on demand. We are using 'AJAX' so this is generally very fast. If I click here, for example,



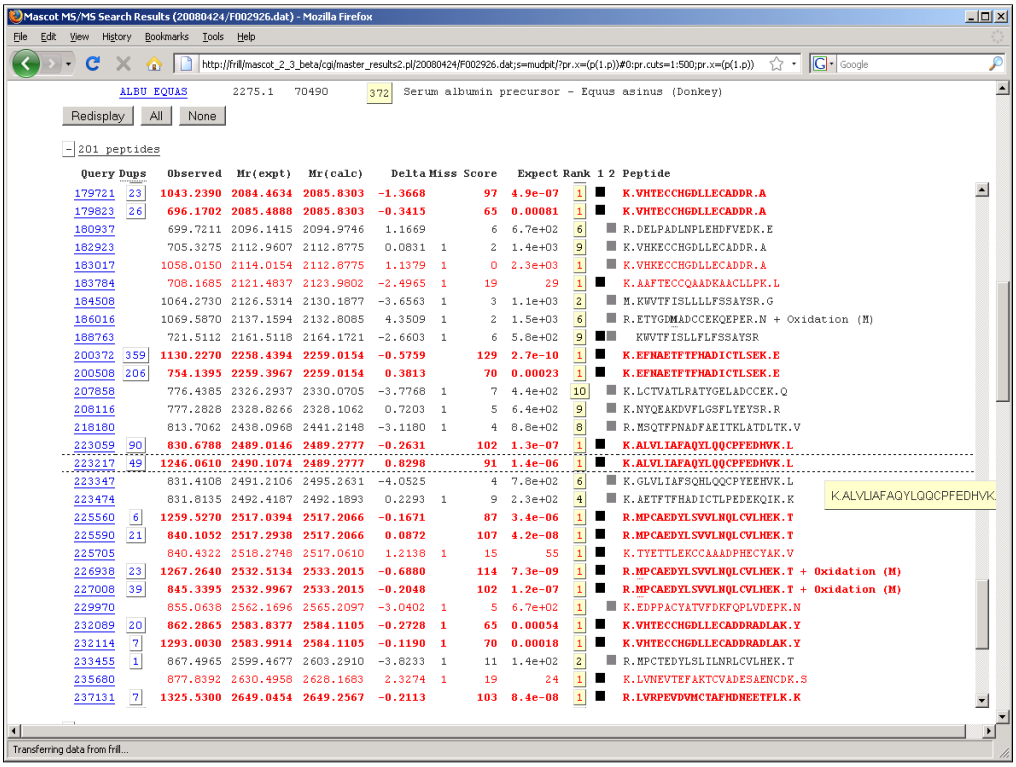
…we can see the search parameters.



The next thing that you will notice is that by default, we only show 10 hits per page, although this can easily be changed from the drop down list here. The significant reduction in time to load the report is due to this being a paged report where not all hits are shown at once.

There's also an option to search for a particular query, accession or even precursor match – very important with a paged report.

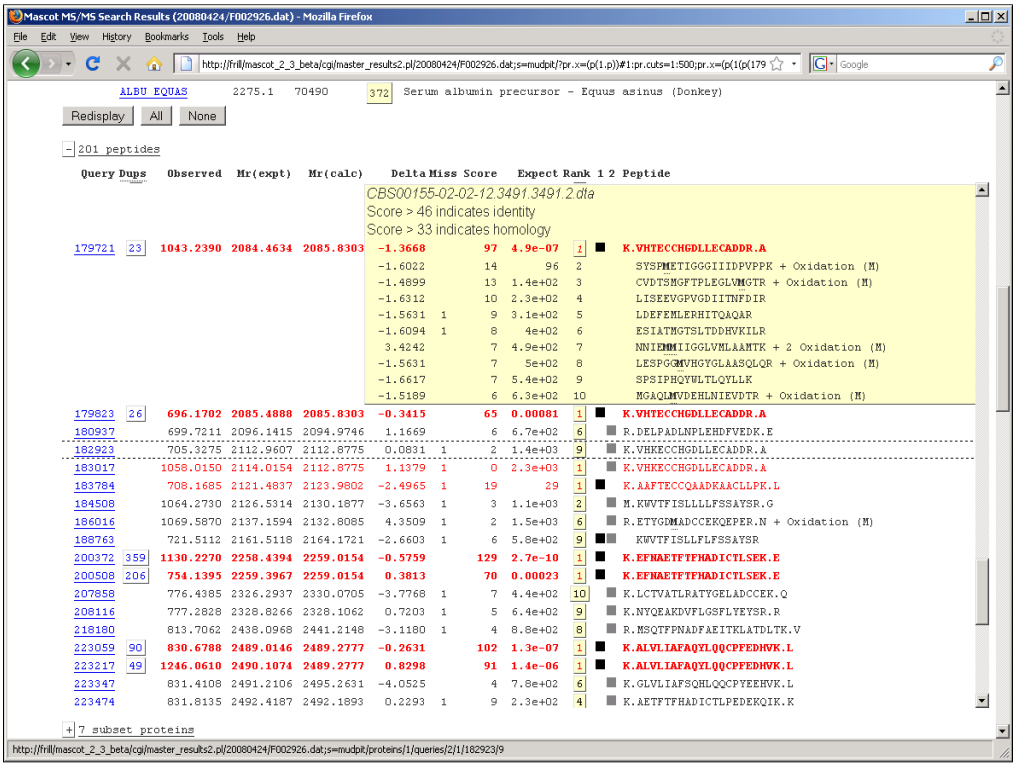
Anybody who is really wide awake will notice that I said that there were 1487 protein matches in Mascot 2.2, but here you can see that there are only 257 protein families. This is because the protein grouping is so much more efficient in Mascot 2.3



Scrolling down for this match, we can see a summary of the 201 peptide matches to the albumin proteins. You'll see on the left there are a lot of cases where more than one spectrum matched to the same peptide.

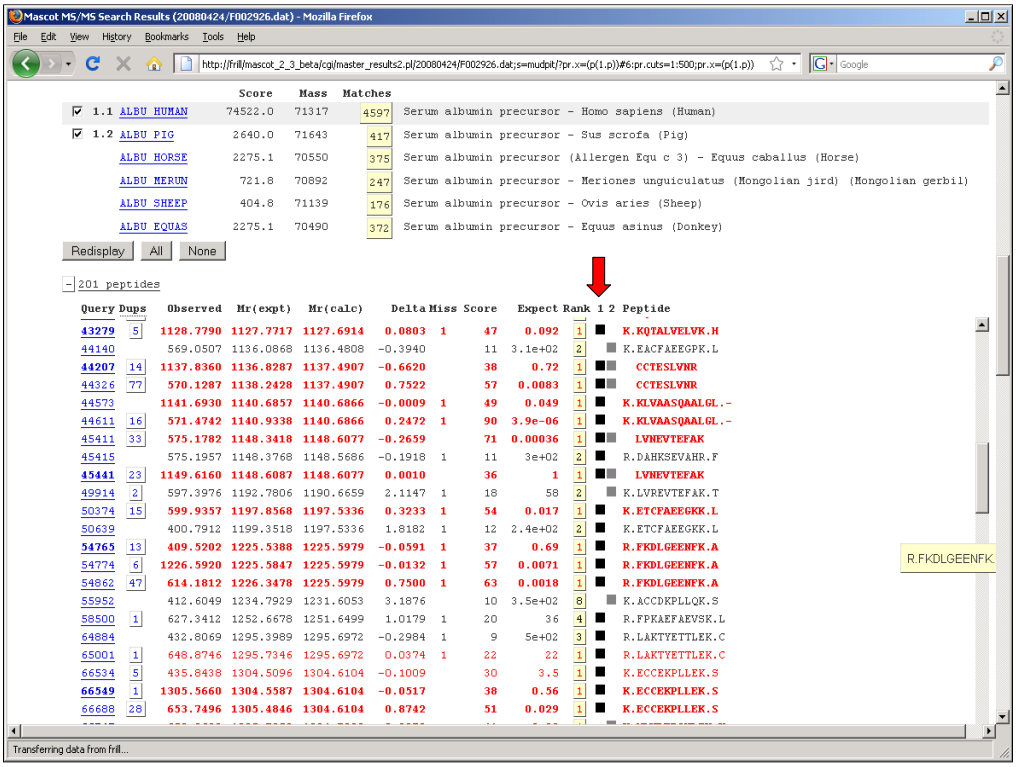
We found that many people believed that bold red meant a significant match, so we've now given in and that's what it means in Mascot 2.3.

Much of this should be familiar. You can still click on the query number to get the peptide view. However, there is no yellow popup window when you hover the mouse over the query number. To get the same information, click on the rank number



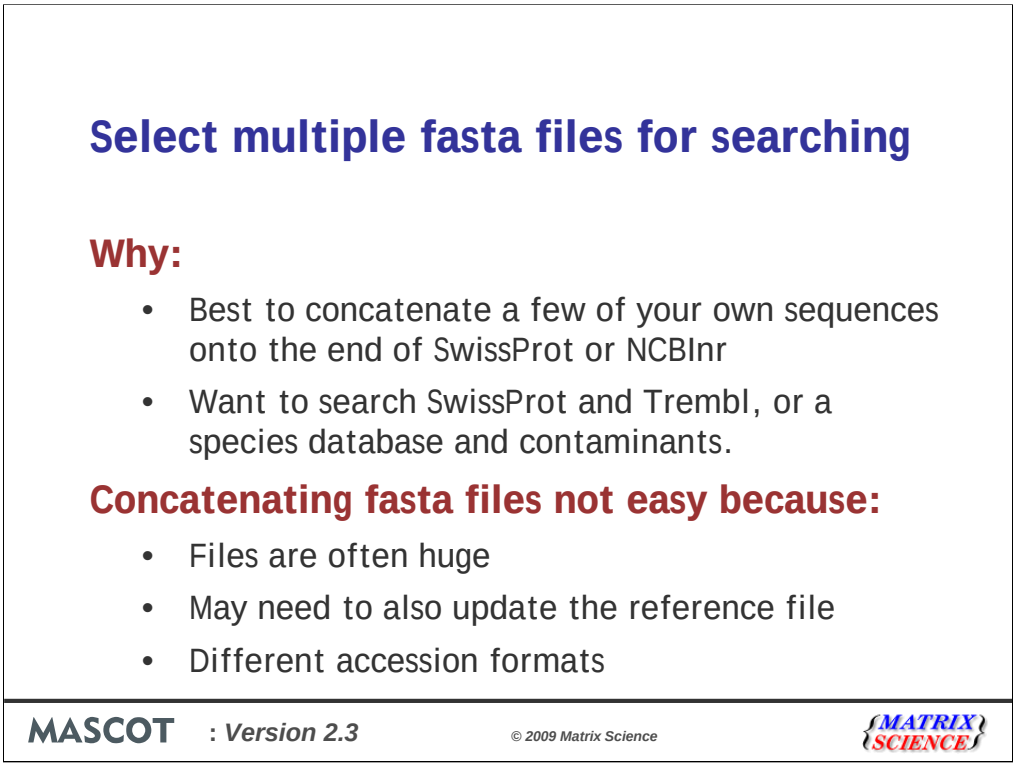
You will see that this information is now included ‘in-line’ which means that you can display more than one at a time – something that was impossible with the popup windows.

Again, we are using AJAX (shorthand for asynchronous JavaScript and XML) to just load these parts on demand.



The little square boxes need some explanation. The number at the top of the column signifies the protein number – so in this case 1 is the human albumin and 2 is the pig, horse, donkey albumin. For query 50374, you can see that this peptide was just found in the human protein. A grey box shows that a match was found in some, but not all of the proteins in the family member.

As you can see, this new report is a replacement for both the peptide and select summary. There is no longer a requirement for separate reports for small and large searches.



In Mascot 2.3, you can select multiple databases for searching in a single search.

It's not advisable to search databases with a single, or just a few entries. We recommend that you add your own sequences to the end of a reasonable size database such as SwissProt or NCBInr so that the statistics are more reasonable. Also, a common requirement is to search a species specific database and some common contaminants.

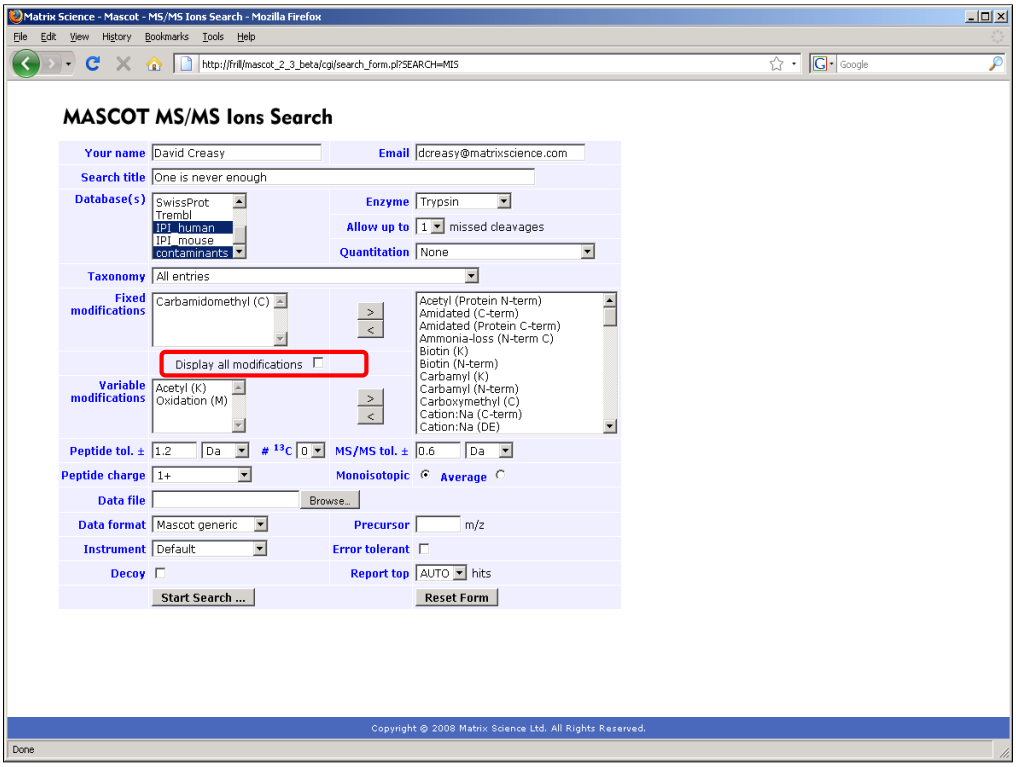
These points illustrate why the new feature will make it easier to maintain the databases.

Dealing with multi GB files is never easy.

It's also difficult to add a sequence to a database with a reference file, such as SwissProt, because the reference file needs to be updated too.

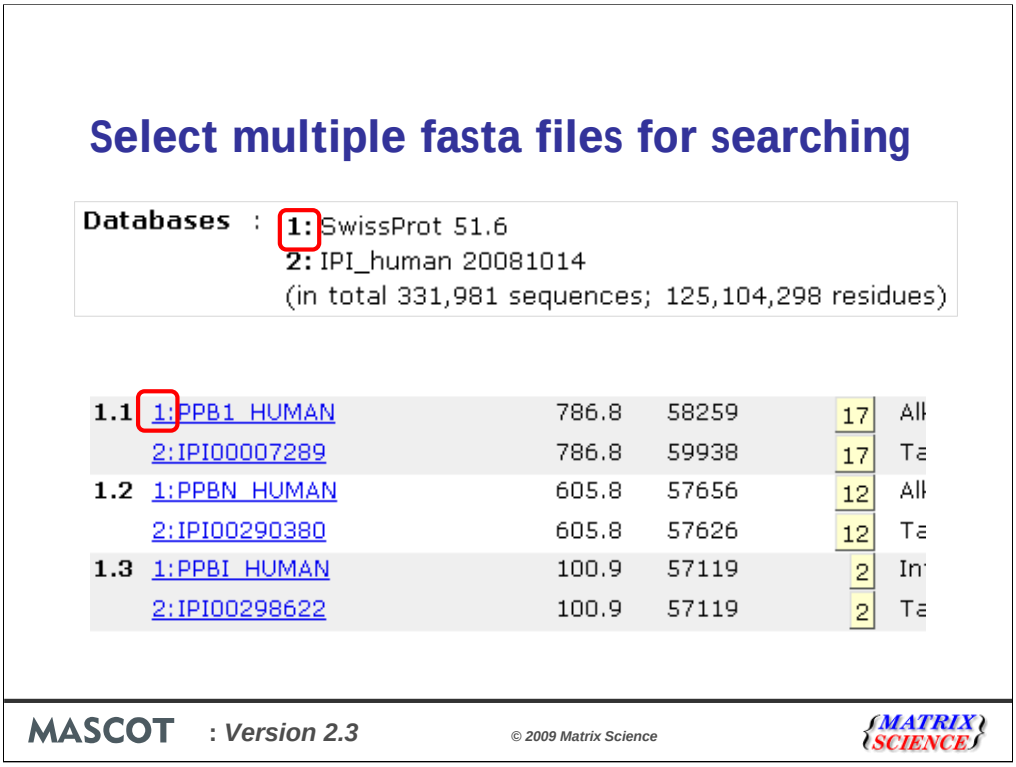
You need to make sure that the accession string format is similar in both databases and that there are no duplicate accessions.

Finally, if you update the large public databases such as NCBInr, then you need to concatenate additional sequences to the end of the file after every update.



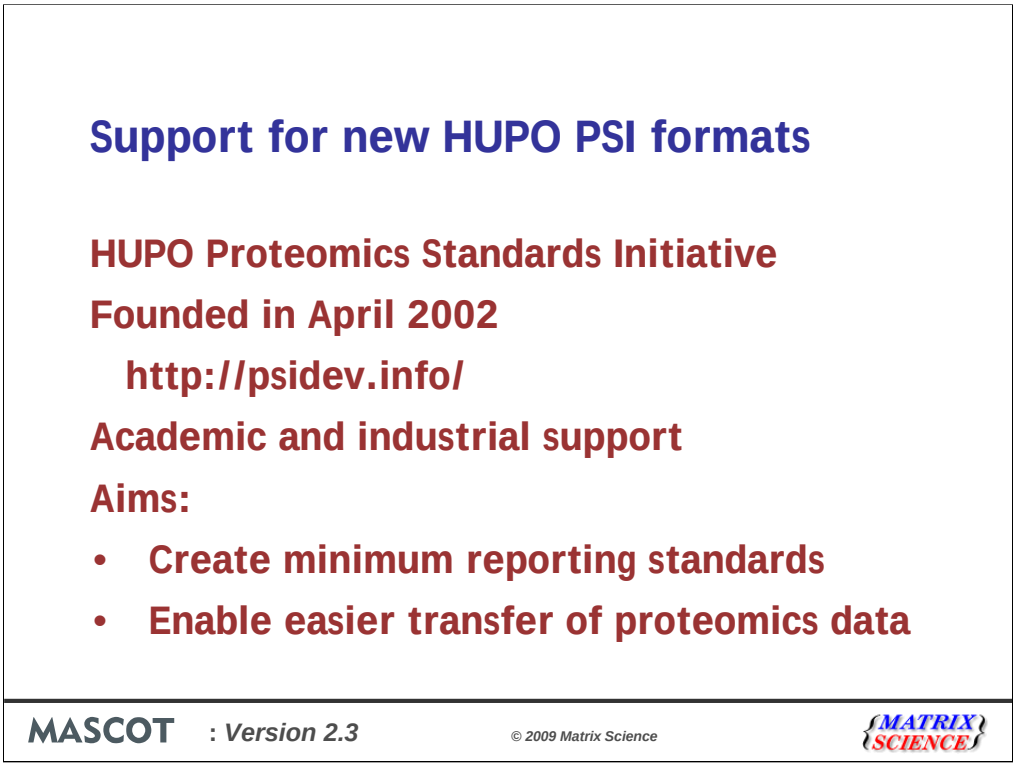
You can now simply select more than one database at a time.

We've also changed the way that you select modifications. It's now very easy to see which modifications have been selected, and there's a checkbox to show all the modifications.



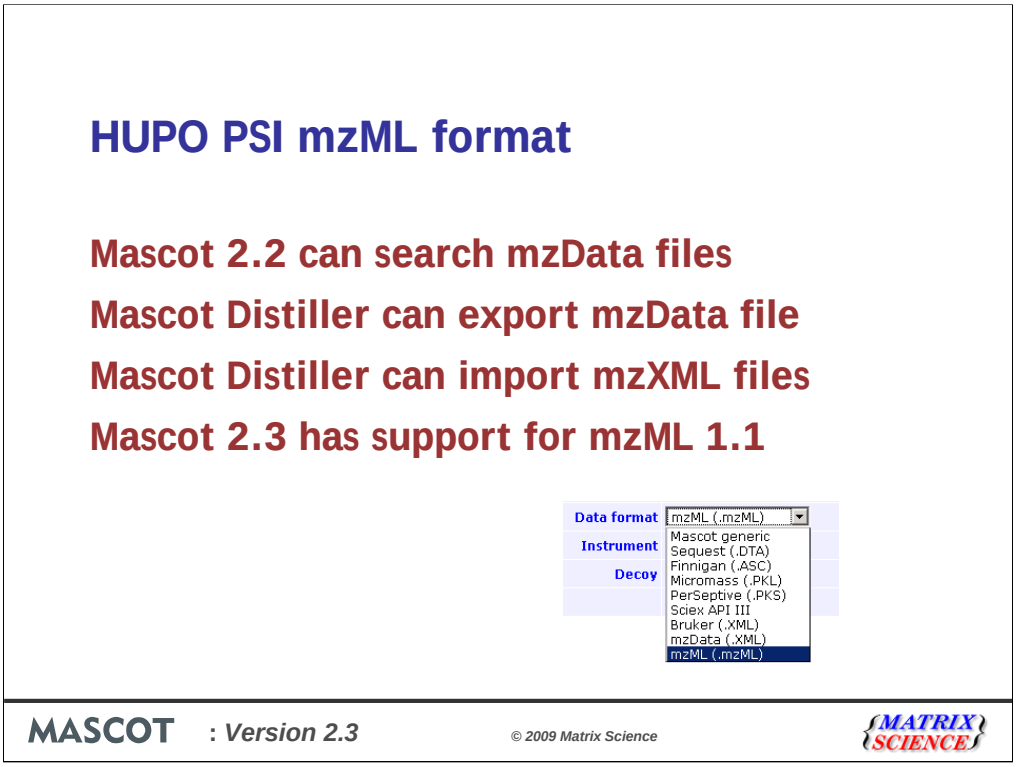
At the top of the report, there is a list of the databases that were searched and the total number of sequences and residues.

In this case, the search was just against SwissProt and IPI_human. The databases are numbered sequentially, and these numbers are used to refer to the database in the body of the report.



The HUPO Proteomics standard initiative was founded in April 2002 and has had consistent support from a number of academic groups and mass spec instrument and software vendors. Matrix Science has been involved throughout this period.

The main aims are to create standards which recommend the minimum information about a proteomics experiment that should be reported. From these, XML schema have been developed which should allow data to be more easily moved between different proteomics applications and repositories.

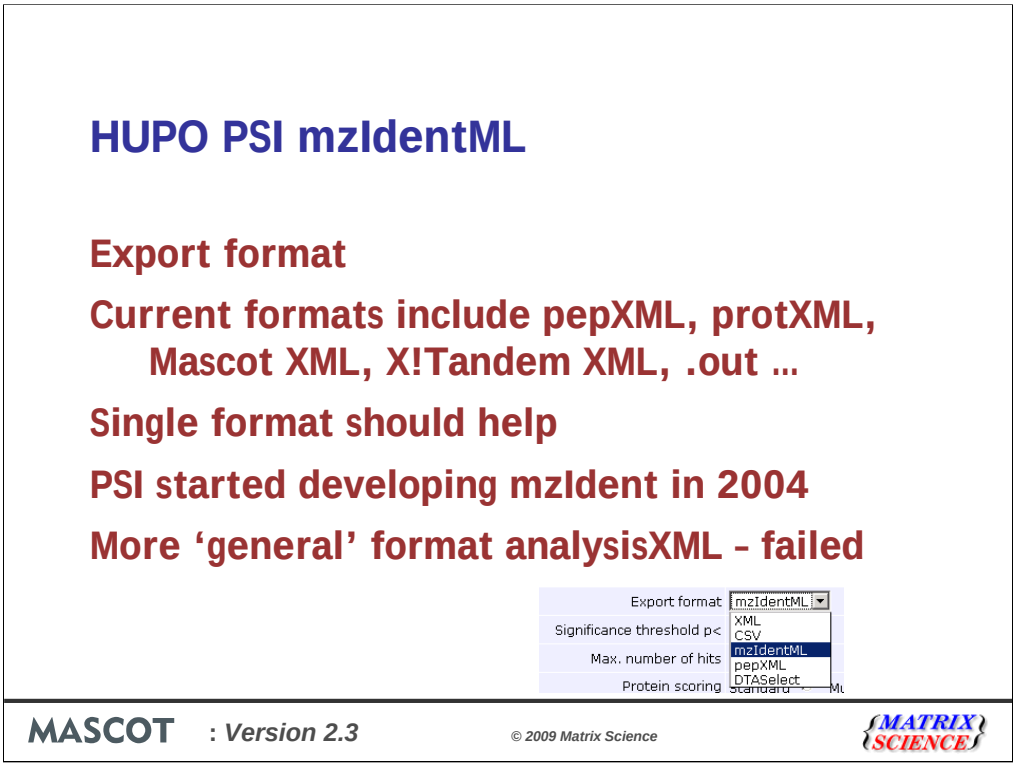


The first mass spectrometry based schema to be developed by the PSI was the mzData format. This has been used by a number of vendors and Mascot has included support for this format for many years.

However, the ISB has had a ‘competing’ standard called mzXML. A couple of years ago, it was agreed to merge the mzXML and mzData standard, and this has now been achieved. Almost. Version 1.0 had a few shortcomings, and it was agreed to fix these in version 1.1. The review period for 1.1 is almost over, and we’ve written the parsers for mzML 1.1 assuming that there will be no further changes. See the PSI web site for tools that can output mzML files

Unfortunately, there’s not much to show for quite a lot of work, just an option on the list of file formats. However, we’ve tested with a number of files and we do pass a lot of the additional information provided in the mzML file through to the Mascot results file – perfect for developers building their own pipelines.

We’ll be adding support for mzML to Mascot Distiller later in the year.



The other PSI format that is relevant for Mascot is the mzIdentML format.

This is an export format and there is currently no real standard. All the search engines output in different formats.

A single format would certainly help for repositories such as Pride and for submission to journals.

The PSI started developing mzIdent in early 2004 when it was claimed that it would be "functional by the year's end".

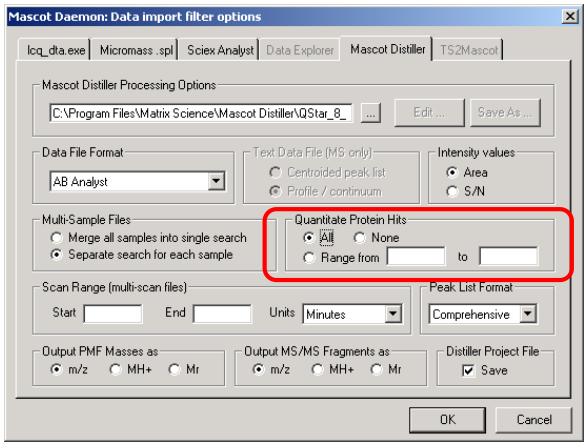
However, a few people thought that this was an opportunity to create something more general and useable by other proteomics processing such as 2D gels, and even chromatography. For this reason, the name was changed to analysisXML

However, after a couple of years, it became clear that this wasn't going to succeed, so there was a change back to reporting for just mass spec protein identification. A 60 day review period has just ended, and minor changes are currently being implemented. One of the requests for change was to the name analysisXML which was considered to be too general, so it is now called mzIdentML. The 'ML' incidentally stands for modelling language.

Again, there's not much to say or show, but it is in Mascot 2.3 and a good number of the example instance documents on the PSI web site are from Mascot.

Mascot Daemon

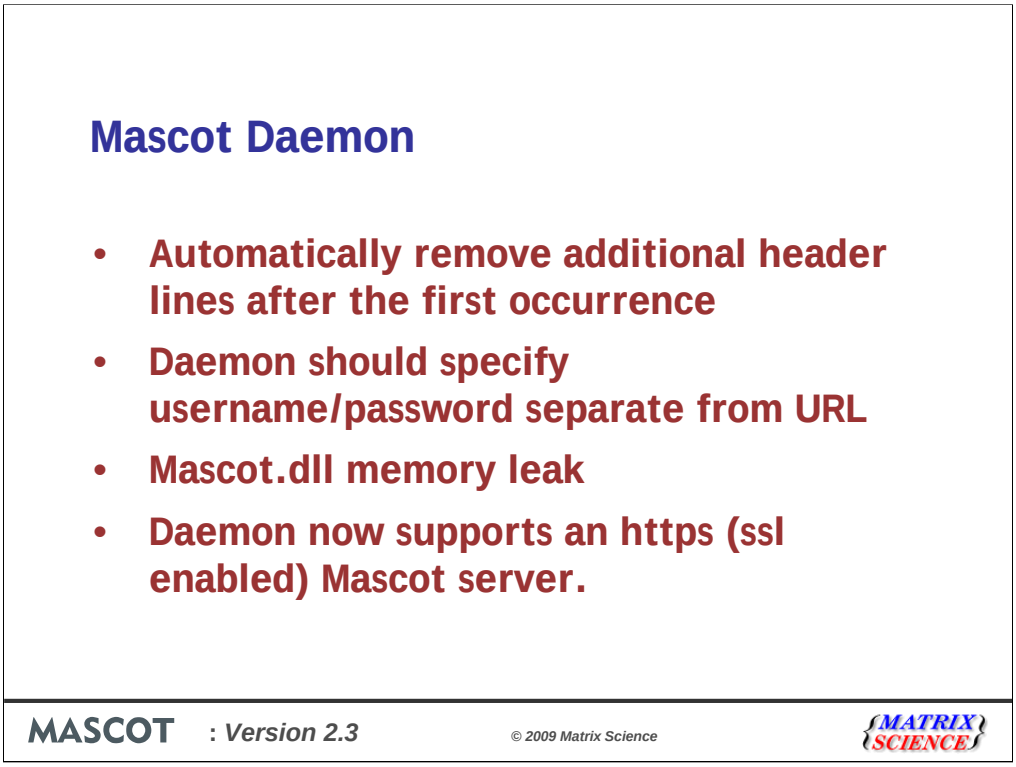
Add support for Distiller quantitation



MASCOT : Version 2.3 MATRIX SCIENCE

One of the major changes for Daemon 2.3 is to enable the automation of quantitation using Mascot Distiller. Those people who currently use Mascot Daemon with the Mascot Distiller import filter will know that you can automate the peak processing using the Distiller libraries and the search the data automatically. Daemon saves the Distiller project file which can then be opened in Mascot Distiller. If you are performing a quantitation experiment, you then need to open the project file in Distiller and press the quantitate button – fine for one or two files, but tedious otherwise.

So, it now possible to chose to perform the quantitation automatically from here to automate the whole process



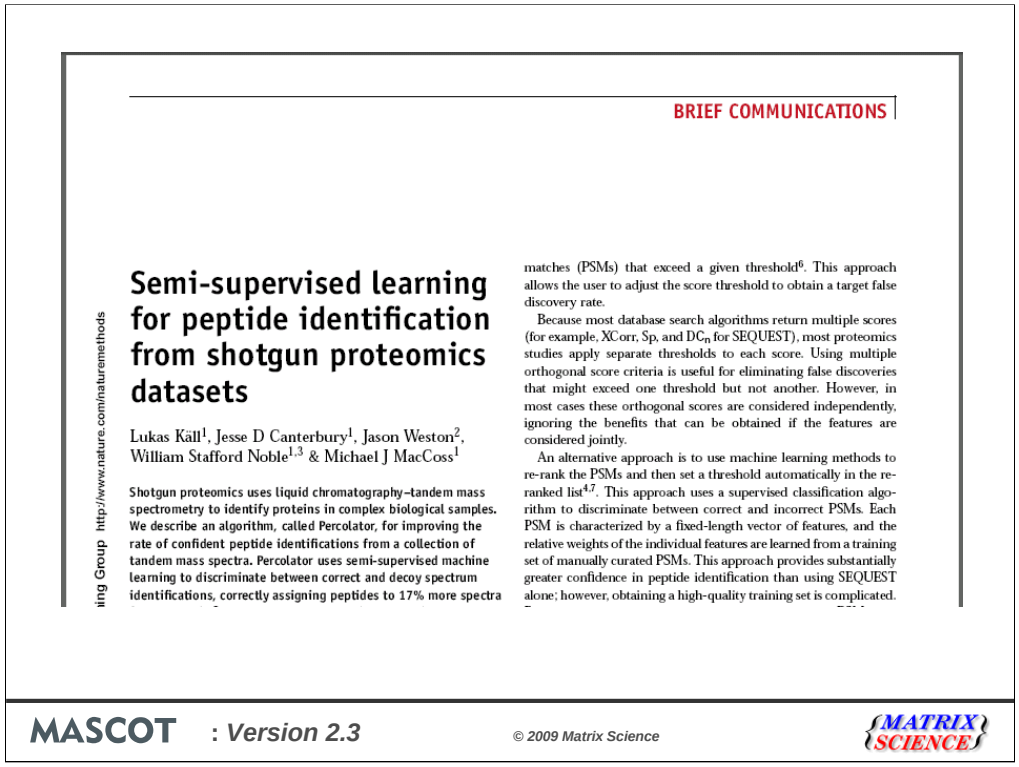
There's a few other important changes in Daemon.

The first occurs when you chose to produce mgf files from an instrument data system and then want Daemon to merge all the files together. If the data system outputs header lines, then when these are merged, the header lines appear in the middle of the mgf file which causes an error. Daemon now strips the header lines after the first file.

If you are required to enter a username and password to access your Mascot server, you used to have to enter this on the url, which is then easily visible to anyone else. There's now a separate place to enter this.

A number of people have had problems with a memory leak in Mascot.dll for Analyst files. Unfortunately, it appears that the problem was not in Mascot.dll, but in Analyst itself. It seemed that there was unlikely to be a fix for this in the near future, so we've played some clever tricks to work around the problem in Mascot Daemon

Finally, Mascot Daemon now supports Mascot running on https.



BRIEF COMMUNICATIONS

Semi-supervised learning for peptide identification from shotgun proteomics datasets

Lukas Käll[1], Jesse D Canterbury[1], Jason Weston[2], William Stafford Noble[1,3] & Michael J MacCoss[1]

Shotgun proteomics uses liquid chromatography–tandem mass spectrometry to identify proteins in complex biological samples. We describe an algorithm, called Percolator, for improving the rate of confident peptide identifications from a collection of tandem mass spectra. Percolator uses semi-supervised machine learning to discriminate between correct and decoy spectrum identifications, correctly assigning peptides to 17% more spectra

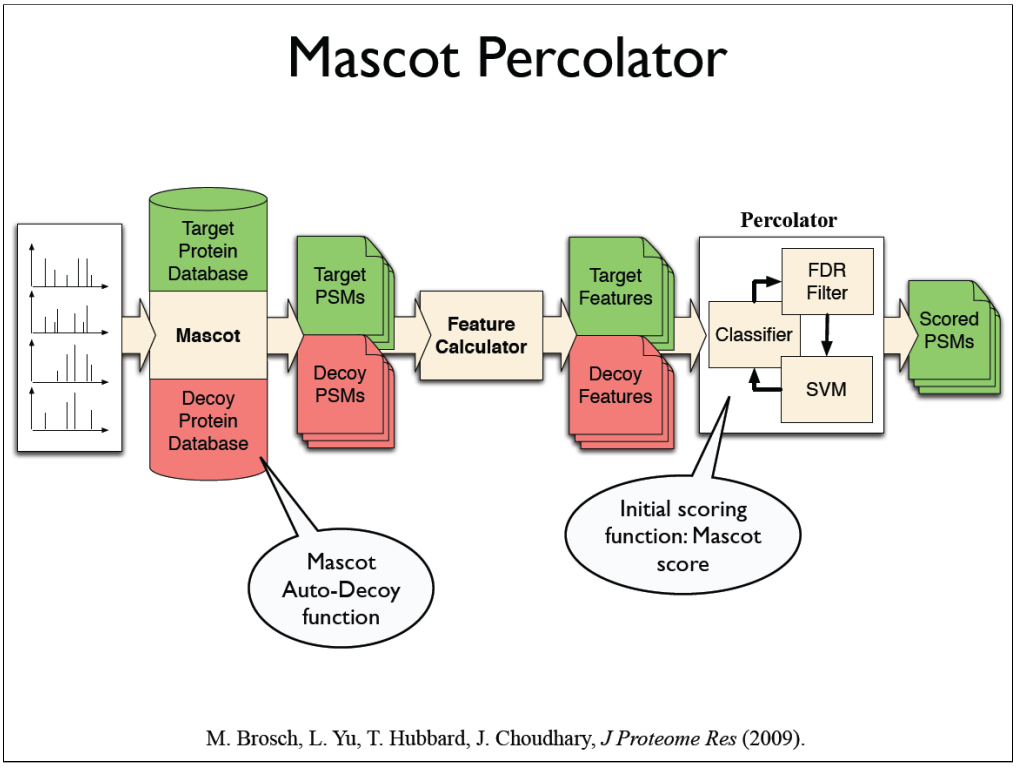
matches (PSMs) that exceed a given threshold[6]. This approach allows the user to adjust the score threshold to obtain a target false discovery rate.

Because most database search algorithms return multiple scores (for example, XCorr, Sp, and DC_n for SEQUEST), most proteomics studies apply separate thresholds to each score. Using multiple orthogonal score criteria is useful for eliminating false discoveries that might exceed one threshold but not another. However, in most cases these orthogonal scores are considered independently, ignoring the benefits that can be obtained if the features are considered jointly.

An alternative approach is to use machine learning methods to re-rank the PSMs and then set a threshold automatically in the re-ranked list[4,7]. This approach uses a supervised classification algorithm to discriminate between correct and incorrect PSMs. Each PSM is characterized by a fixed-length vector of features, and the relative weights of the individual features are learned from a training set of manually curated PSMs. This approach provides substantially greater confidence in peptide identification than using SEQUEST alone; however, obtaining a high-quality training set is complicated.

MASCOT : Version 2.3 MATRIX SCIENCE

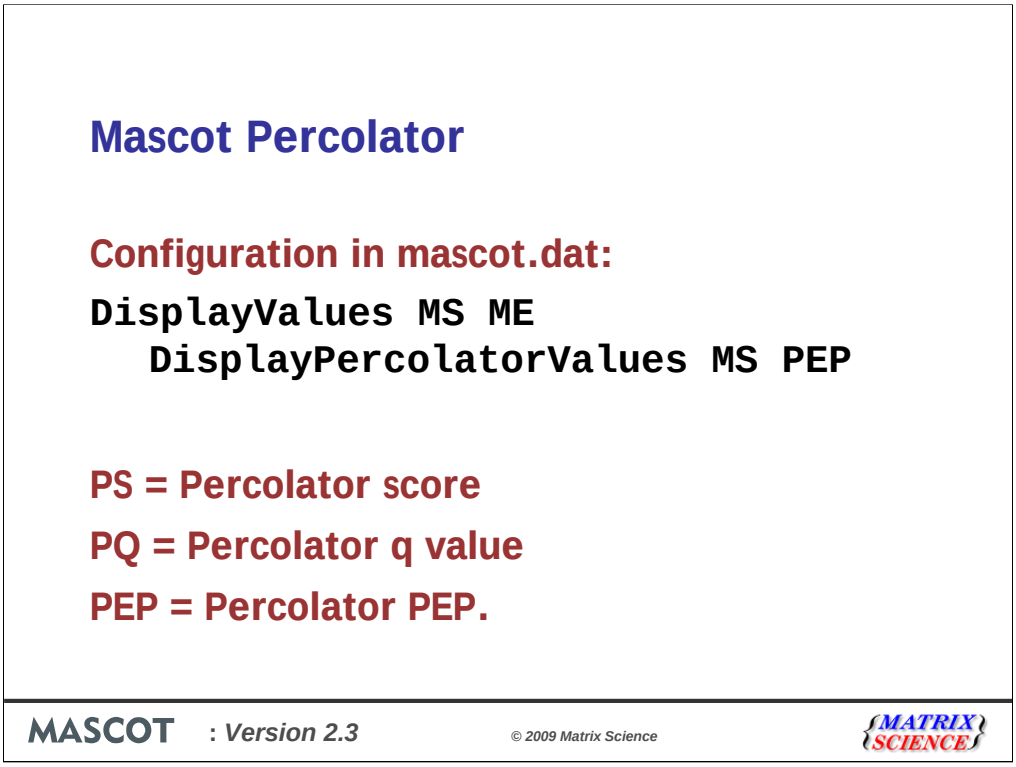
We've heard earlier from Markus about Percolator which was developed by Lukas Kall and in Mike MacCoss.



Markus described how Mascot Percolator provides a neat interface between Percolator and Mascot version 2.2. This provides alternative and improved scoring in cases where you perform a decoy search and have sufficient data for the results to be reliable.

To use Mascot Percolator with Mascot 2.2, you'll need to download the appropriate files and run Mascot percolator. You then have a choice of how to view the results. As Markus has shown, you can run a script to modify the Mascot results file and then view the scores in the standard Mascot results, or you can combine the results using spreadsheets.

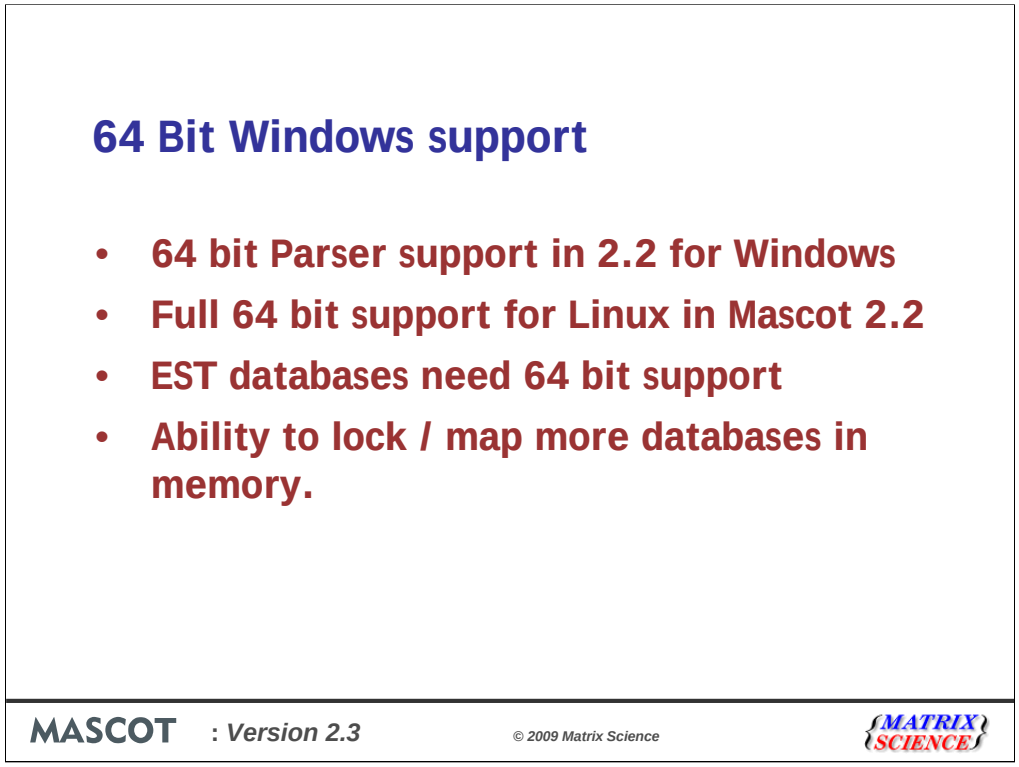
In Mascot 2.3, we will simply be providing an easier to use interface for this. We will be shipping the percolator binaries and providing the infrastructure to run percolator automatically.



In mascot.dat, you'll be able to select what values to display in the standard Mascot reports.

In this case, the Mascot score and Percolator PEP (probability that individual match with this Percolator score is random match) will be displayed. If percolator 'fails' in any way, for example if you didn't search against a decoy database, then the DisplayValues will be used instead.

It will be possible to display the Percolator score, the percolator q value or the percolator PEP value.

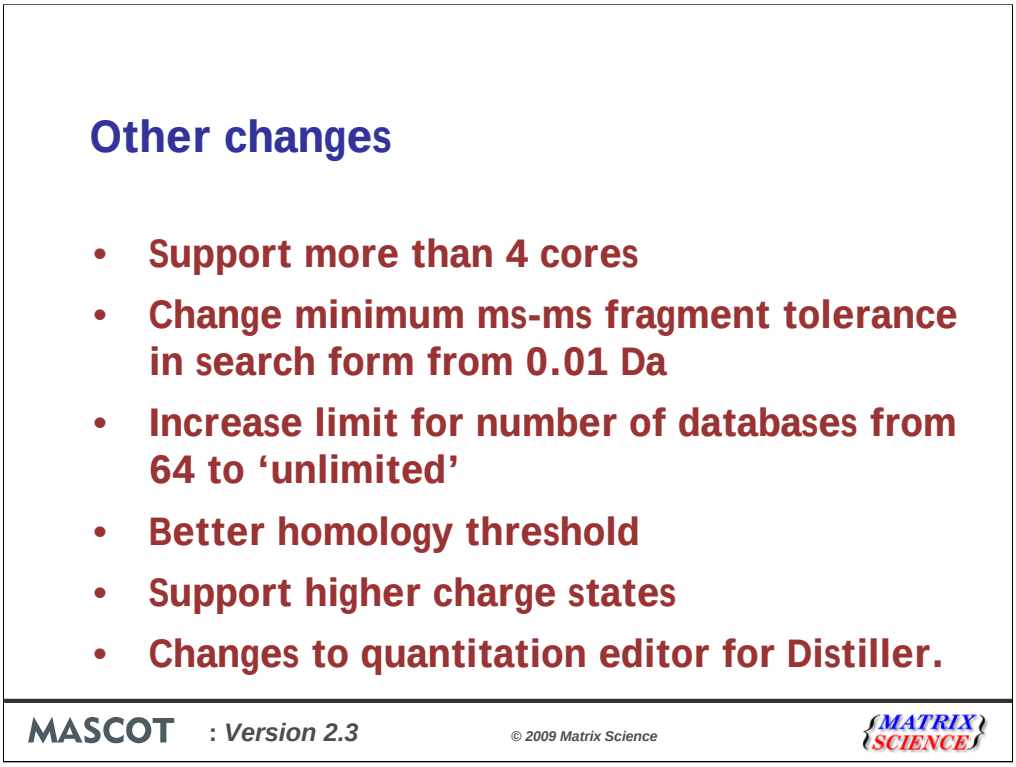


We are also now including true support for 64 bit Windows.

In version 2.2 we added support for 64 bit Mascot Parser, which means that you can open very large results files if you install on a 64 bit platform.

Mascot 2.2 also provided full support for 64 bit Linux.

The most important issues resolved by adding 64 bit support are the ability to use the huge EST databases and the ability to lock more databases in memory.



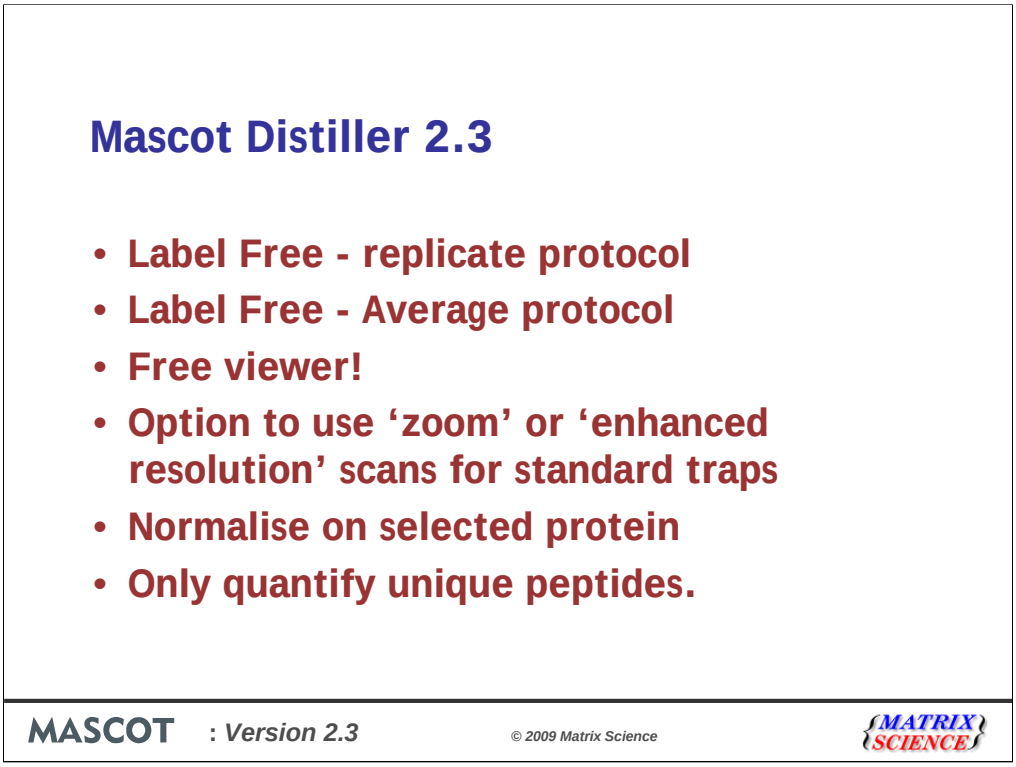
There have been a couple of patch releases to Mascot 2.2 to resolve issues with the more recent Intel processors. However, there is a limit of 4 cores per cpu in Mascot 2.2. Mascot 2.3 will provide support for processors with more than 4 cores, but you'll need one license for each 4 cores. A system with an 8 core processor, will therefore require a 2 cpu license.

A number of people have thought that Mascot can't work with a fragment tolerance less than 0.01 Da. This is simply because the search form doesn't allow you to enter a value less than this unless you use millimass units. This is changed in Mascot 2.3

We've increased the limit for the number of databases from 64 to a configurable 'unlimited' maximum value.

In some cases it wasn't possible to calculate a homology threshold, particularly with more accurate precursor masses, so we've been able to improve this.

We also now support higher charge states for the precursor which helps with top down experiments. The limit has been increased from 8+ (or 8-) to an unlimited value.



The new version of Distiller supports two protocols for label free quantitation.

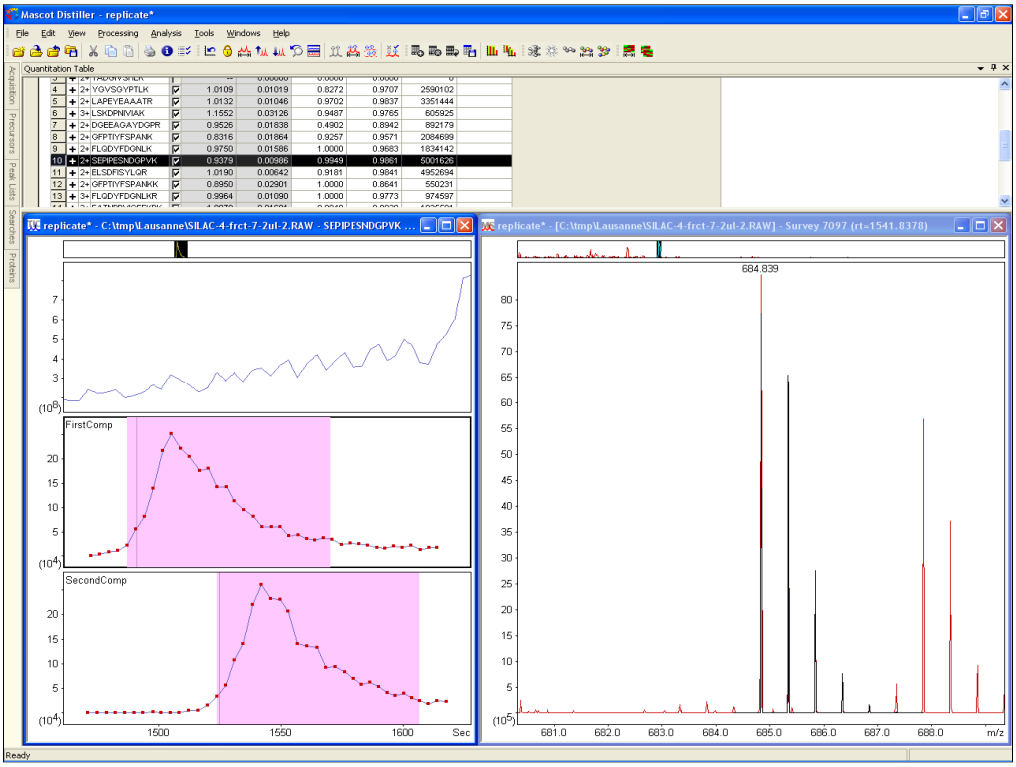
The first, is what we call the replicate protocol, and is used to determine the relative abundance of proteins between two or more samples. There will typically be a different raw file for each sample.

The Average protocol, is label-free, absolute quantitation for the proteins in a mixture in a single sample.

If you want a colleague who doesn't have a Mascot Distiller license to be able to view results, just send them the raw data, the project file and then get them to download and install Distiller. Without a license, it runs as a viewer. If you have a 30 day evaluation license, then Distiller becomes a free viwer when the license expires.

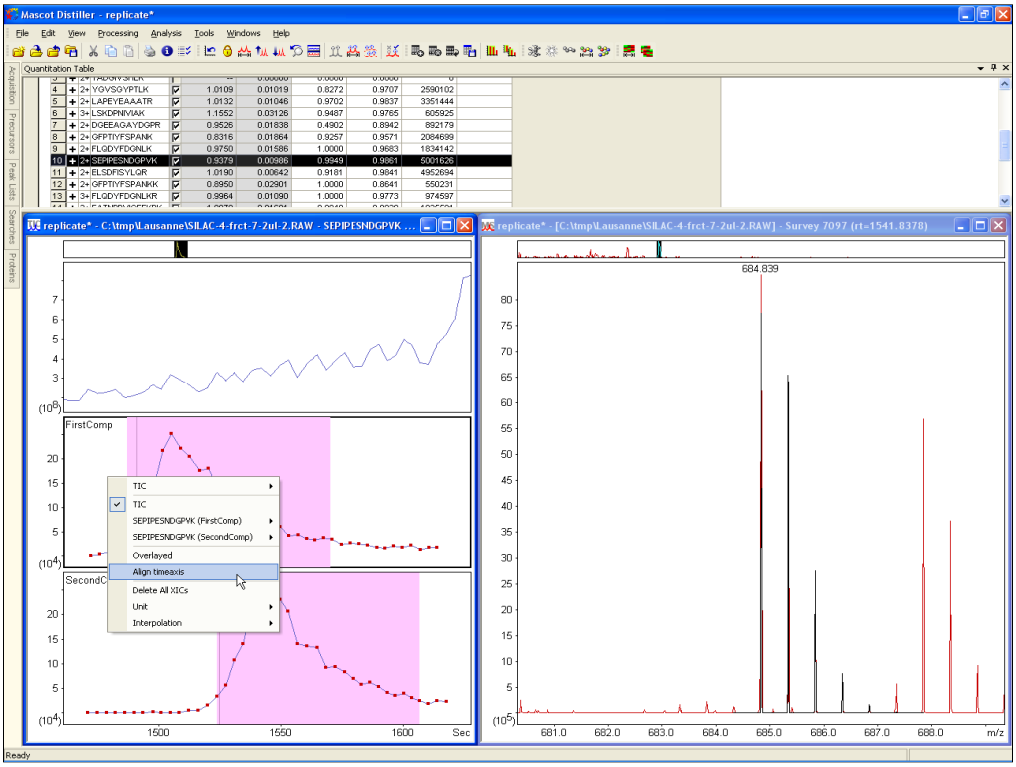
With a standard ion trap, the survey scans are often not high enough resolution for accurate quantitation. In Distiller 2.3, we've added the option to use zoom scans or enhanced resolution scans as they are called on some instruments.

We've also added the option to normalise on a selected protein

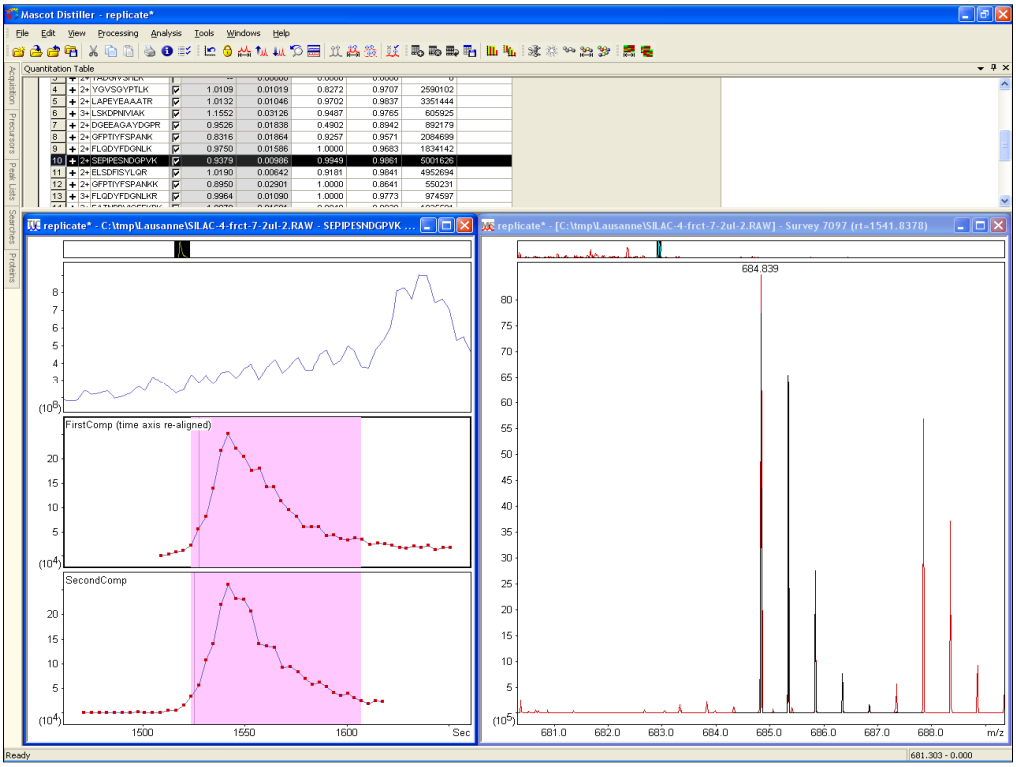


This is an example of label free using what we the replicate protocol. For clarity, there are just two raw files here, but it is of course possible to have 10 or 20 files. This method relies on getting the Mascot peptide matches first, and as with labelled methods, there is no need to get a match from both the data sets.

In this particular case, there is a shift of about a minute. The alignment is surprisingly good provided chromatograms don't get too far apart. This shows an XIC peak on a true time axis

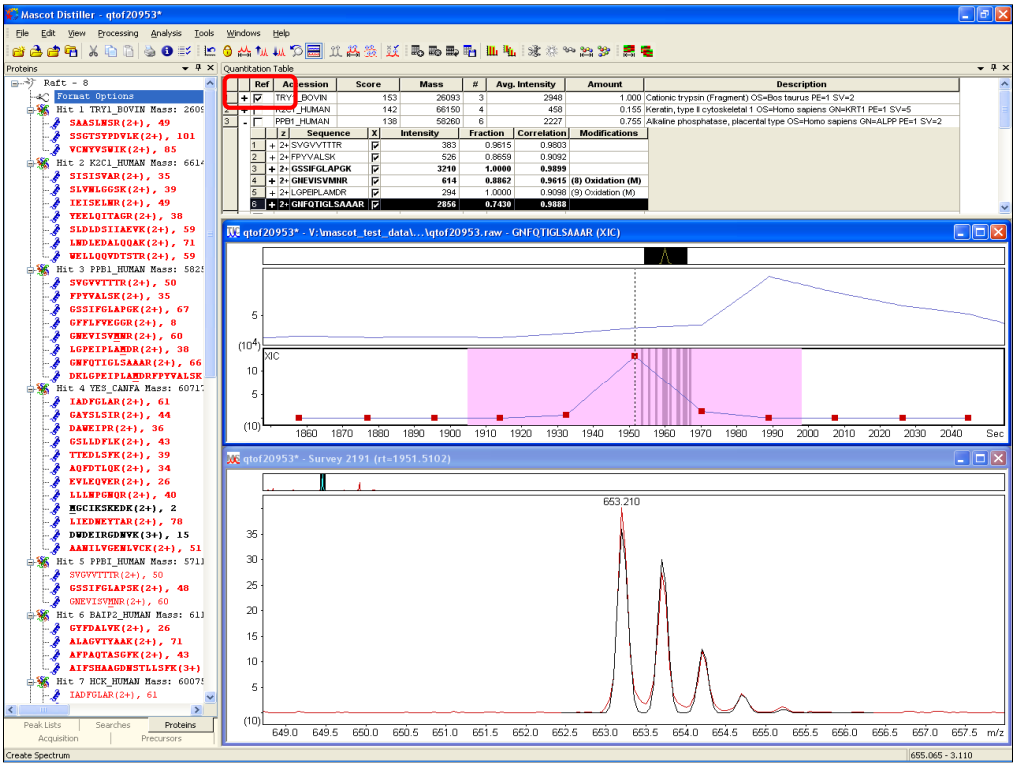


And there is a context menu to display aligned XICs



We can see here that the alignment is not bad at all.

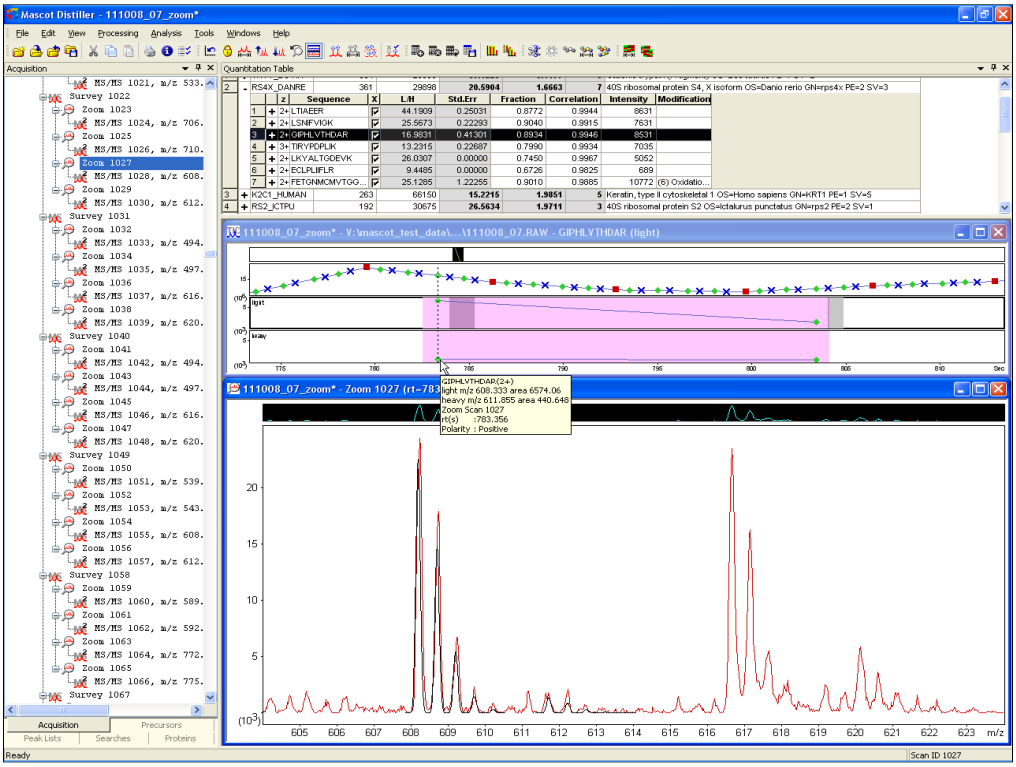
It will of course fail if XIC peaks are miles apart and there are no Mascot matches to tie them together. If there are matches to the peptides in both files, it is of course more likely to succeed.



The other label free protocol that we now support is average.

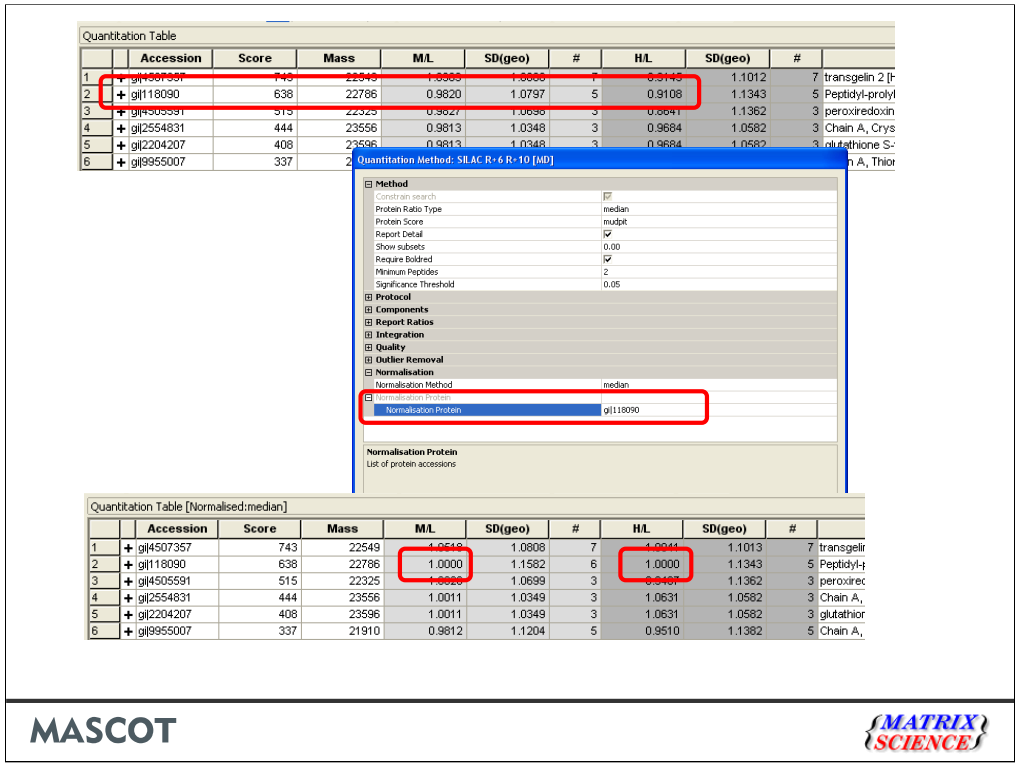
The Average protocol, is label-free, absolute quantitation for the proteins in a mixture based on the application of a rule to the intensities of extracted ion chromatograms (XICs) for the peptide matches in a database search result. The method was first described by Silva, J. C., and a group from Waters. Their observation was that the average MS signal response for the three most intense tryptic peptides per mole of protein was constant within a coefficient of variation of less than ±10%.

Amounts are shown relative to reference protein, which can be selected using checkbox. Rule here is that amount corresponds to summed intensities of 3 most intense peptide matches per protein. Alternative to spectral counting, but should be more accurate because it is using the survey scans.



With a standard ion trap, the survey scans are often not high enough resolution for accurate quantitation. In Distiller 2.3, we've added the option to use zoom scans or enhanced resolution scans as they are called on some instruments.

This is an example of standard LTQ data off zoom scans using SILAC labelling.



MASCOT

MATRIX SCIENCE

It’s now possible to normalise on selected protein(s) or peptide(s), spiked into sample. Before (no normalisation) and after (normalise on gi|118090)

Mascot Server 2.3

- New protein grouping
- New faster reports
- Support for HUPO PSI standards
- Percolator integration

Mascot Distiller 2.3

- Two label free protocols
- Quantitation from zoom scans
- Free viewer
- Mascot Daemon integration