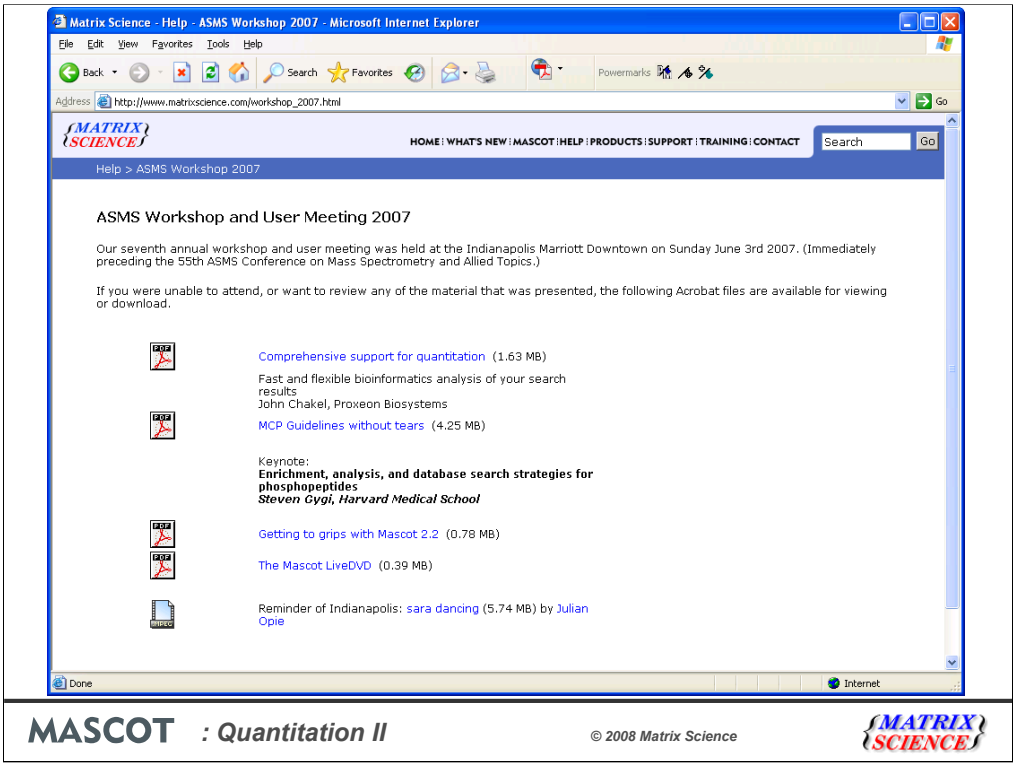
Quantitation II
Mascot Distiller strikes back

MASCOT

MATRIX SCIENCE



We talked about our strategy for supporting quantitation at last year's ASMS User Meeting. I want to give you an update, and concentrate on what's new, so I won't be repeating the material presented last year. You can find the slides and notes from last year's talk, along with all the other talks, on our web site. Search for Indianapolis

Quantitation - Overview

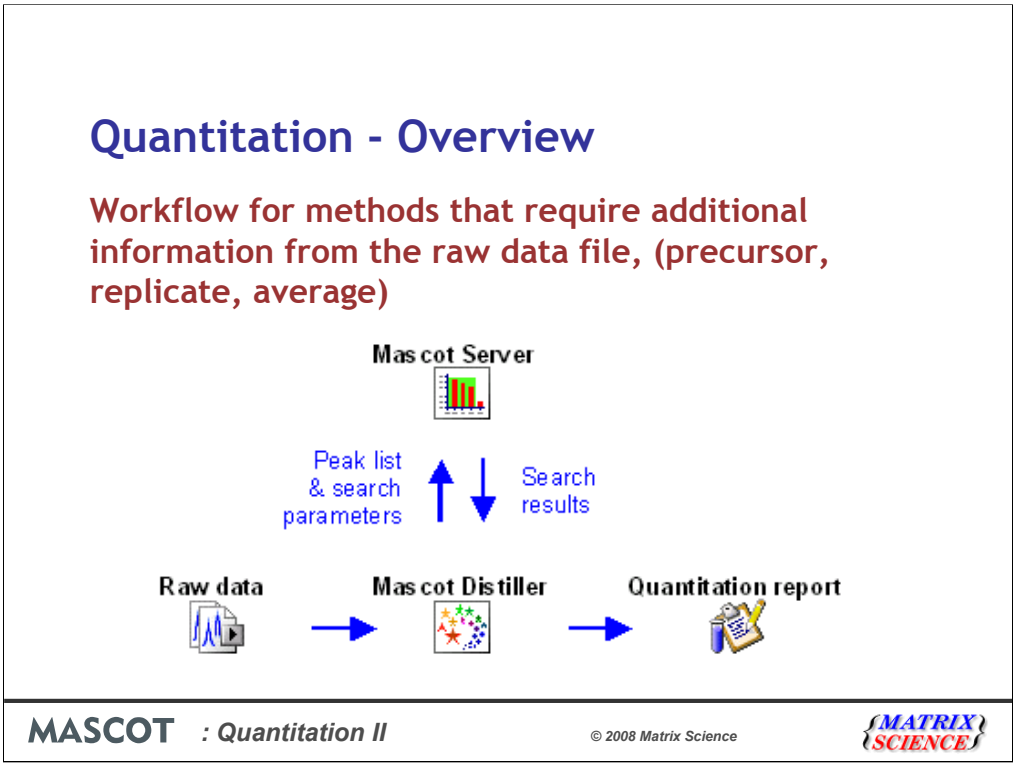
Protocol	Basis	Ratios	Examples
reporter	Specific reporter ion peaks within a single MS/MS spectrum	Inter-sample	iTRAQ, ExacTag, TMT
precursor	Extracted ion chromatograms for related precursors within a single dataset	Inter-sample	ICAT, SILAC, ^{18}O, ICPL, AQUA, Metabolic
multiplex (Neubert et. al.)	Pairs of sequence ion fragment peaks within a single MS/MS spectrum	Inter-sample	SILAC, ^{18}O
replicate	Extracted ion chromatograms for identical precursors across two or more datasets	Inter-sample	Label-free
empai (Ishihama et. al.)	Protein coverage from a database search result	Intra-sample	N/A
Average (Silva et. al.)	Extracted ion chromatograms for selected peptides per protein within a single dataset	Intra-sample	N/A

I'll just show three of the slides from last year, so as to provide the context for this talk

We described how we have classified the various quantitation strategies into six protocols. For details of each protocol, see the 2007 talk.

The rows with a blue background are the protocols that were fully implemented in Mascot 2.2, because they only require information from a standard peak list.

The rows with a white background require additional information from the raw data file. Information that is not present in a standard peak list

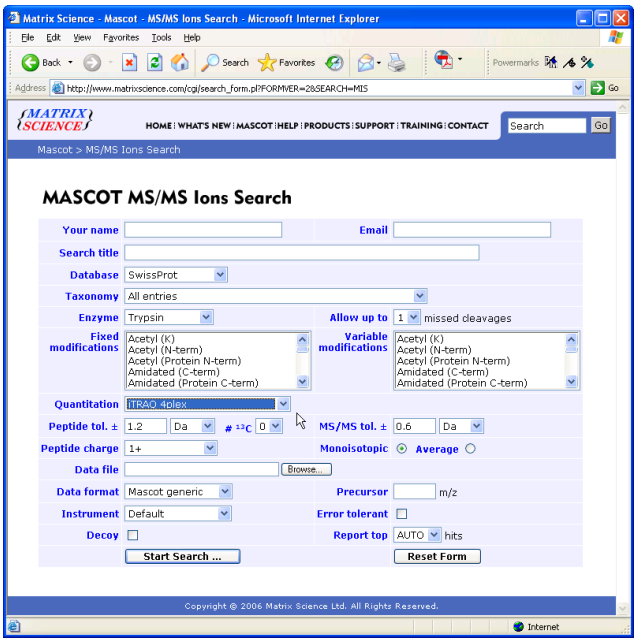


For these methods, the workflow looks like this. The raw data file is processed in Mascot Distiller and the search submitted to Mascot. When the search is complete, the results are returned to Distiller. The quantitation report can then be generated in Mascot Distiller, which has access to both the Mascot search results and the raw data.

Although the Mascot Server side of things was released in March 2007, the Quantitation Toolbox for Distiller is still under development. I wish I could say it was now released. Unfortunately, this is not the case. The plan is to release the toolbox to a beta test group on June 30th with a general release date of August 30th. So, we are very close, and I can show you how it works in some detail.

Quantitation Overview

Simple user interface

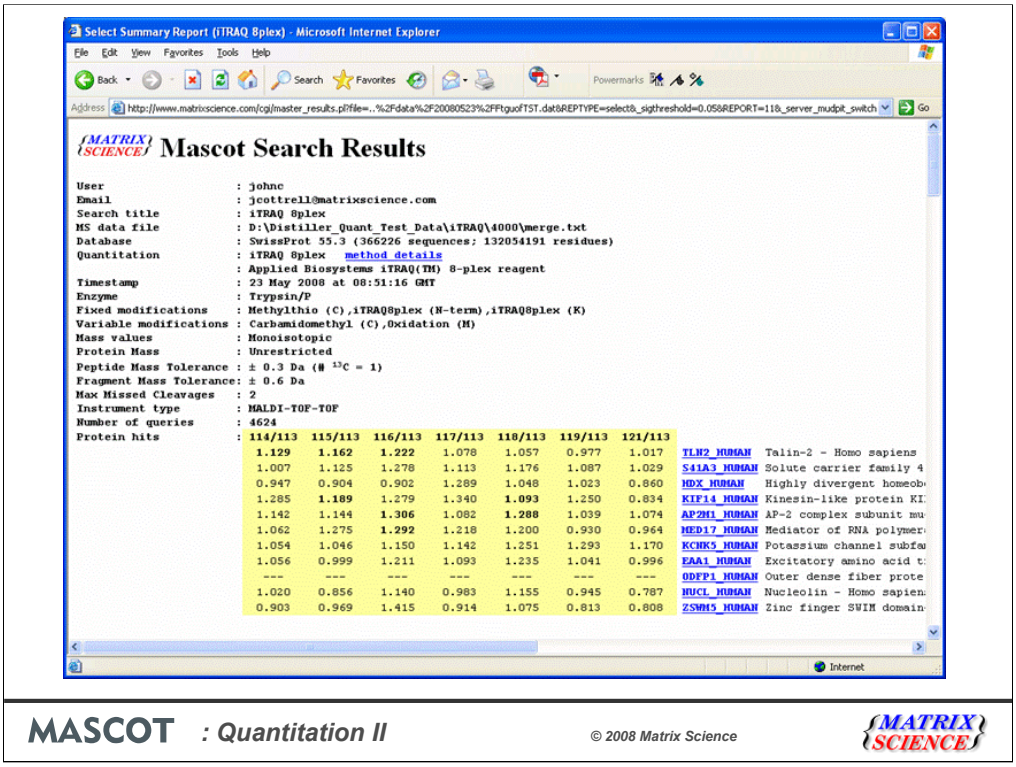


The final review slide is to illustrate that one of the main goals in our implementation has been to keep the user interface simple. Quantitation adds a huge number of choices and parameters, but there is no point in exposing all of these in the search form.

The approach we have chosen is encapsulate these choices and parameters into named quantitation methods. This means that the search form has just a single new control, which replaces the old ICAT checkbox.

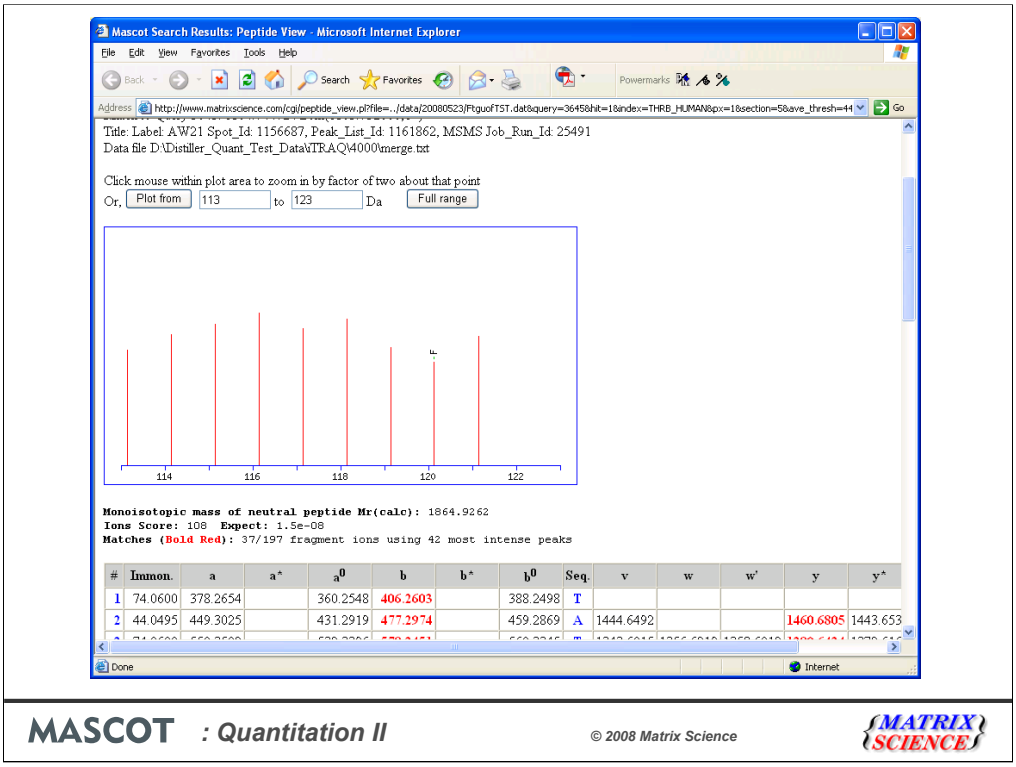
Methods that have [MD] at the end are the ones that require Mascot Distiller

I'll now talk in more detail about the reporter and precursor protocols

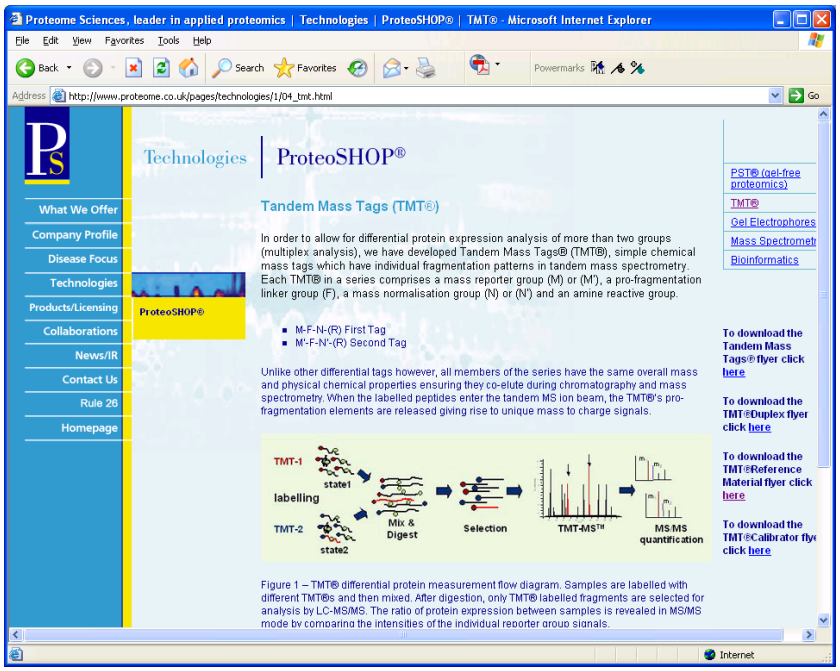


First, reporter. Although we have kept the user interface as simple as possible, this isn't at the expense of flexibility. Nothing is hard coded, so you can easily create new methods for new chemistries. For example, Applied Biosystems released their 8plex iTRAQ chemistry last year. The only change that was needed in Mascot was to add a new method with the new modification and reporter ion masses.

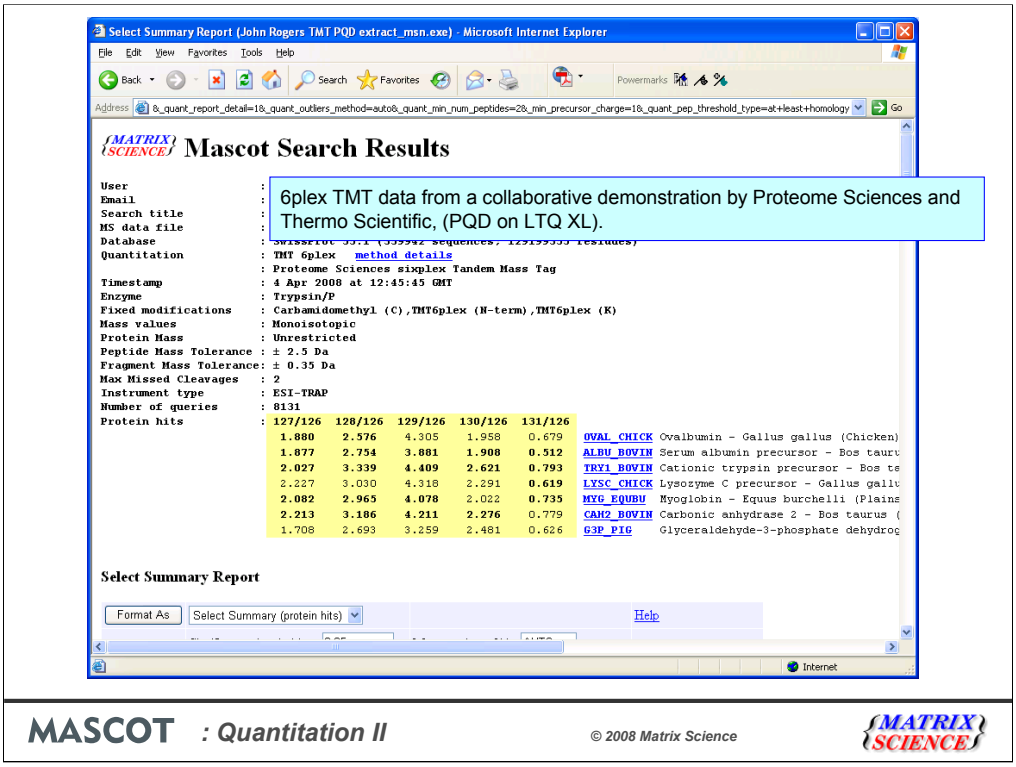
Also, remember that the ratios to be reported are defined in the method. Although this example shows all ratios relative to 113, you could just as easily configure the method to report four independent ratios, 114/113, 116/115, etc.



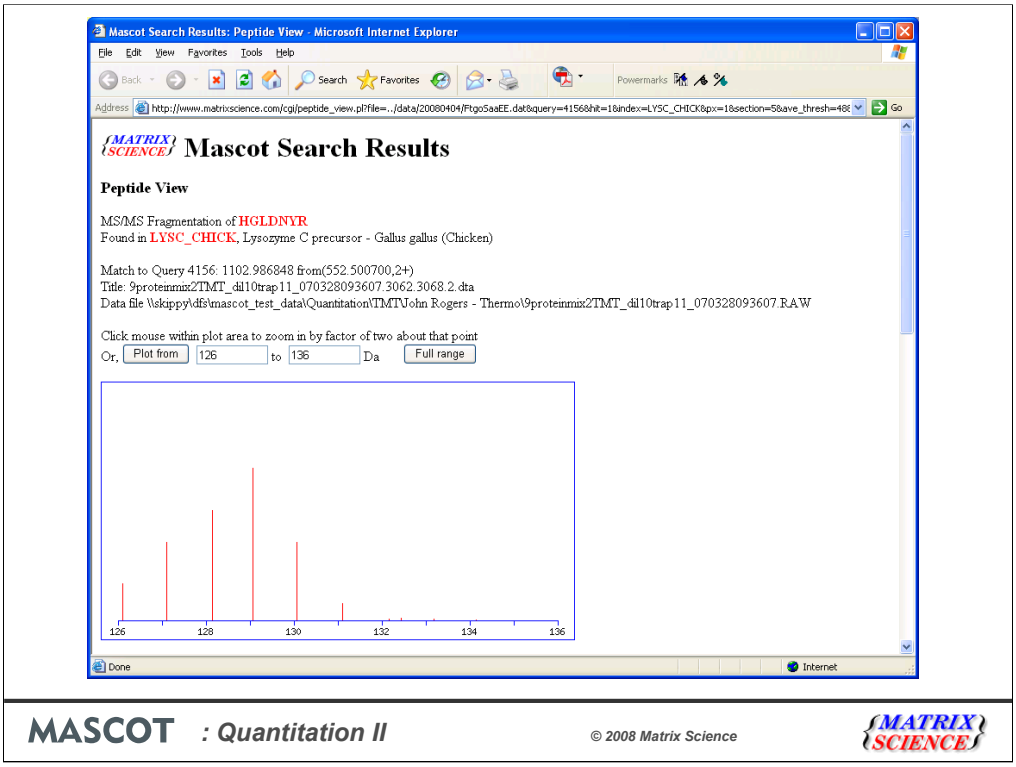
In peptide view, the initial display shows the reporter ion region. We can see why there is no reporter ion at 120 – this is the immonium ion of phenylalanine. Notice the high score. We sometimes get asked whether strong reporter ion peaks cause any problems with scoring. I think this shows that the answer is no.



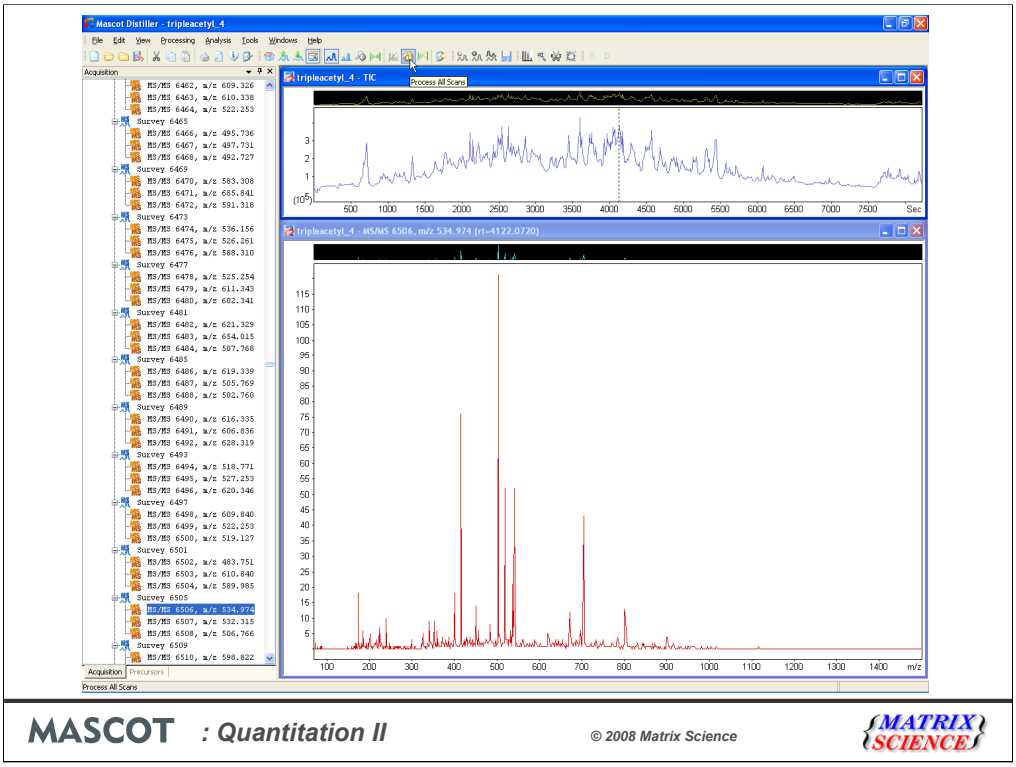
The Tandem Mass Tag chemistry from Proteome Sciences is now commercially available. Again, just a matter of putting the correct modifications and mass values into a method.



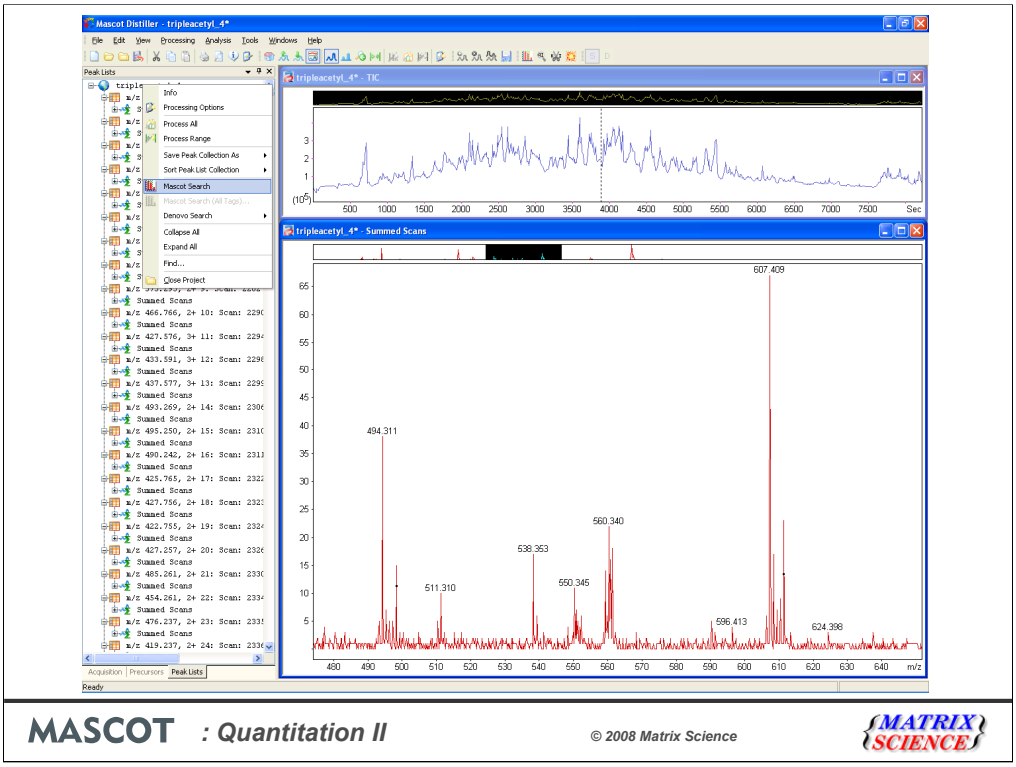
Here we have a report for the 6plex TMT tag, courtesy of Proteome Sciences and Thermo Scientific. This data is from an LTQ XL using Pulsed Q Dissociation



Here is the reporter ion region. This sample was a test mix with ratios 1:2:3:4:2:0.5

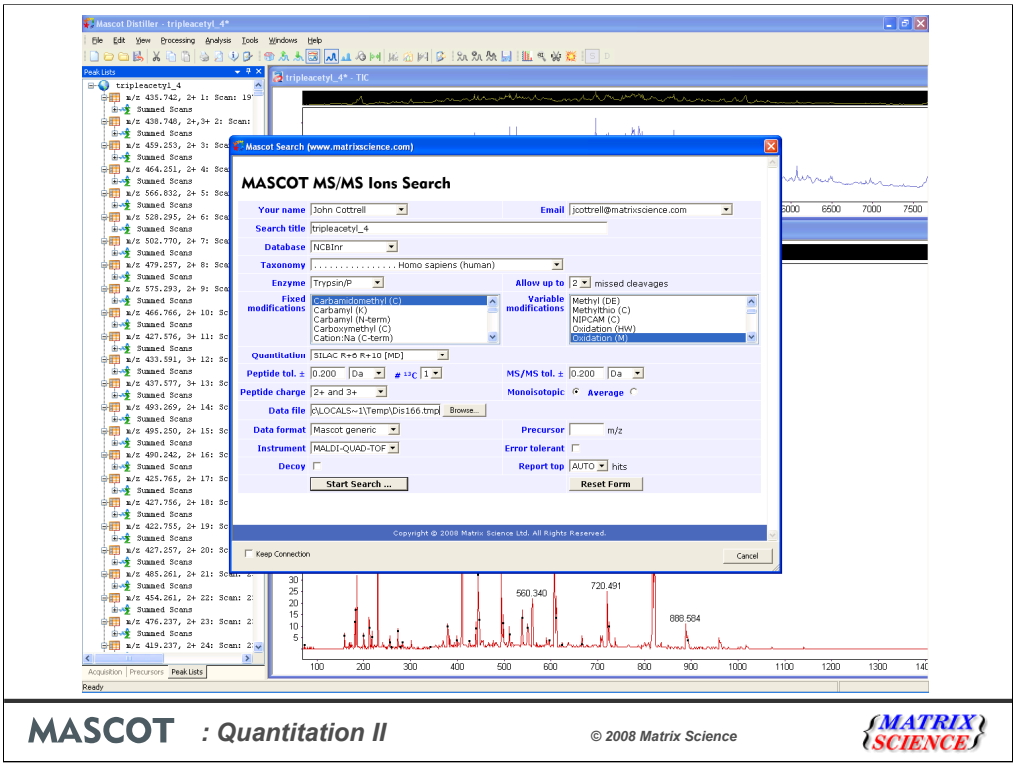


Now lets turn to the precursor protocol. This is where we have to use the Mascot Distiller Quantitation Toolbox. I'm going to illustrate the workflow using one of the QStar sample data sets that can be downloaded from the MSQuant home page. This is a SILAC sample with three components: unlabelled, Arg labelled with 13C(6), and Arg labelled with 13C(6)15N(4). First, we open the Wiff file in Mascot Distiller, and process it into peak lists

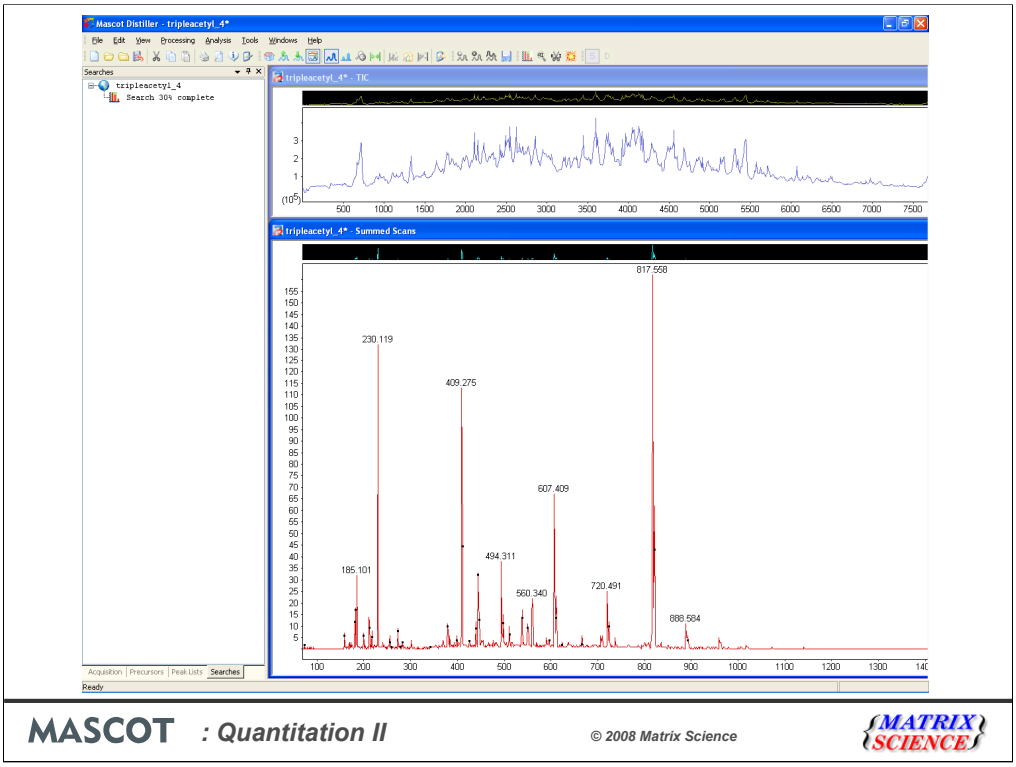


MASCOT : Quantitation II © 2008 Matrix Science MATRIX SCIENCE

The peak lists are submitted to a Mascot Server to be searched

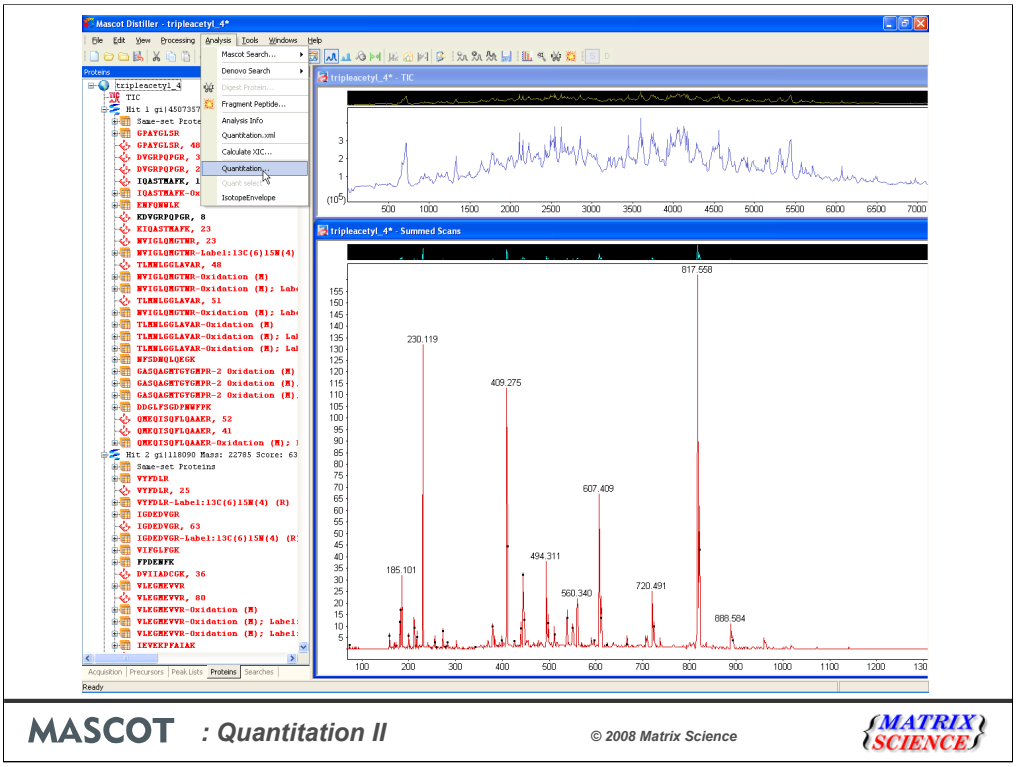


There's an opportunity to tweak the search conditions, but most of the important settings, including the modifications, are embedded in the selected quantitation method

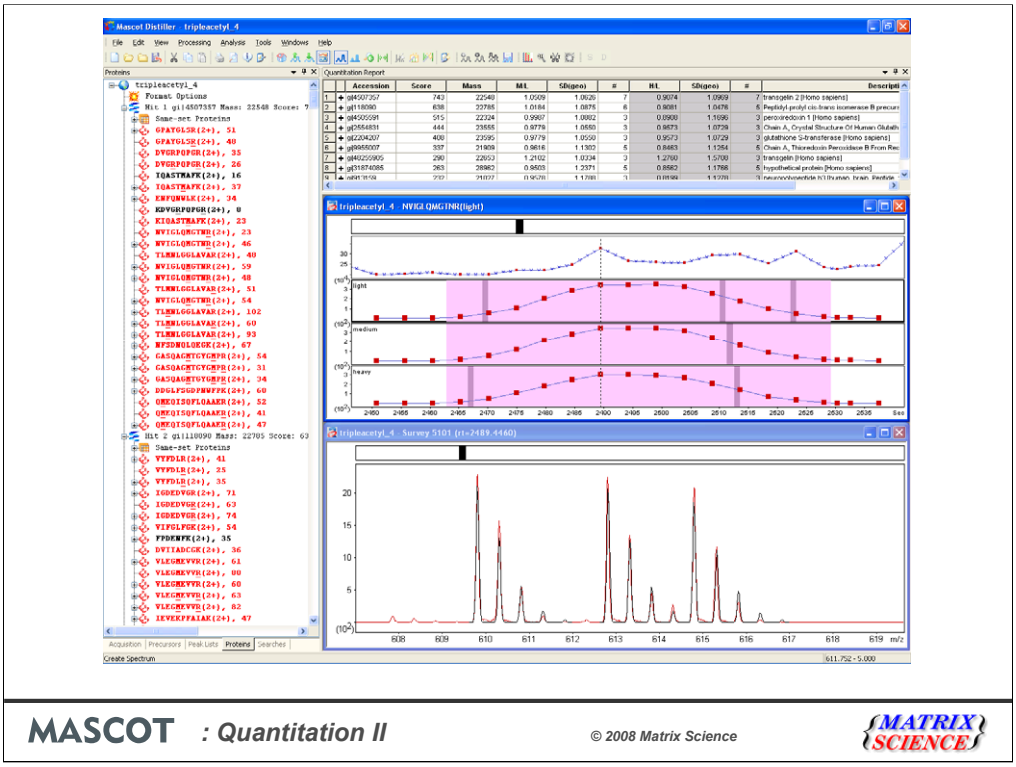


MASCOT : Quantitation II © 2008 Matrix Science MATRIX SCIENCE

We get progress reports while the search is running



When the search is complete, the results are imported into Distiller. All the required information is now available, so Distiller is in a position to generate a quantitation report



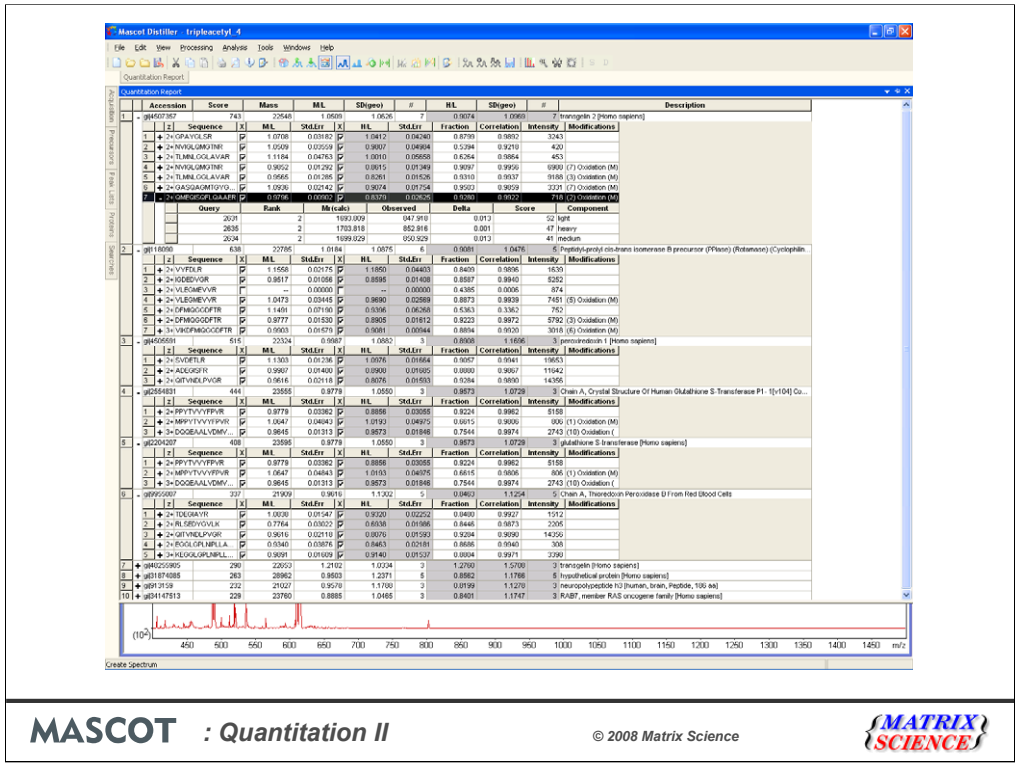
The quantitation results are displayed in a grid control, spreadsheet-style. More about this on the next slide

Below the grid is a chromatogram window. This is displaying the TIC plus extracted ion chromatograms for the three components, light, medium, and heavy.

When you click on an XIC, the scan window at the bottom shows the precursor region of the survey scan. The observed spectrum is in red. This is overlaid with black traces showing the isotope distributions calculated for the ratio being reported. You can make a visual judgement about the quality of the fit. The overlay can be turned on and off using a context menu.

To the left is an explorer tree showing the search results. If you are familiar with the explorer trees in Distiller, you might notice that this tree is a new one, which shows the peptide matches grouped into proteins, like the standard Mascot Peptide Summary report

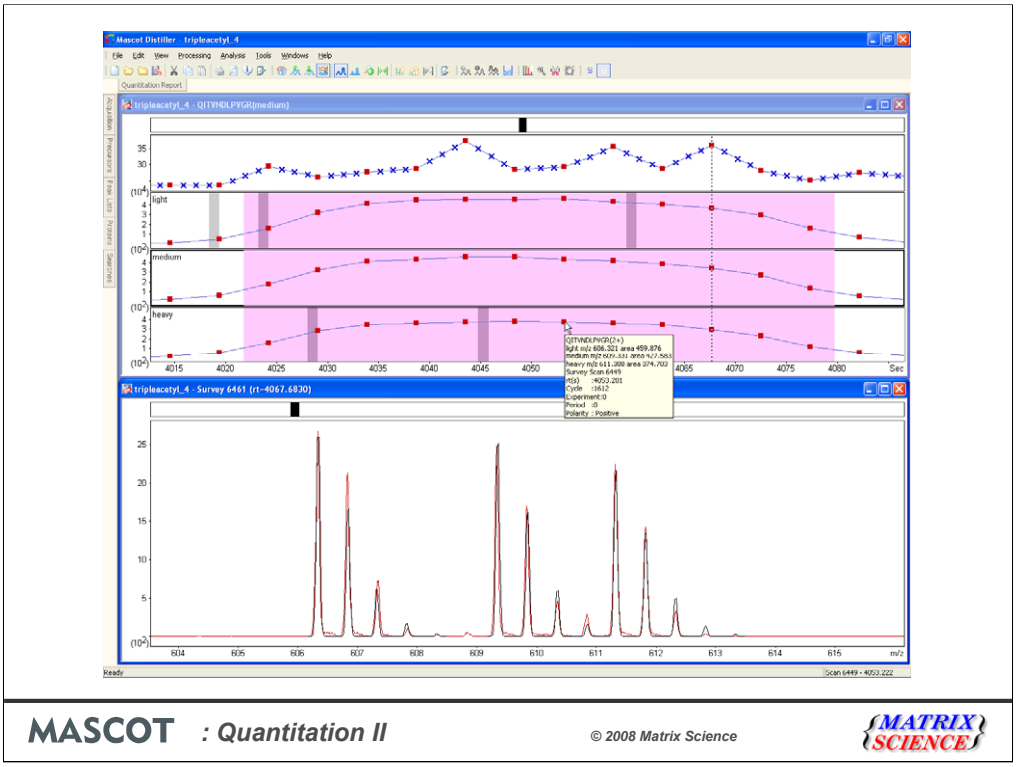
However big your screen, its always difficult to find enough room to display everything. To try and make best use of limited space, the grid and tree can be unpinned, so that they fly out when required and disappear when you move the mouse away.



Here, the grid has been unpinned and is being displayed over the top of the other windows. Some protein hits have been expanded to show individual peptides. One peptide has been expanded to show the corresponding Mascot matches. There's a lot going on here. If you want to study the numbers, this is the place to look. If you don't, then you can hide many of the columns to make the display simpler.

Because this is a three component experiment, two ratios are reported: M/L and H/L. These labels come from the method, and could equally well use sample IDs or whatever you prefer.

The checkboxes allow you to over-ride the decisions made by the software. If a checkbox is cleared, the peptide ratio is rejected and does not contribute to the protein ratio

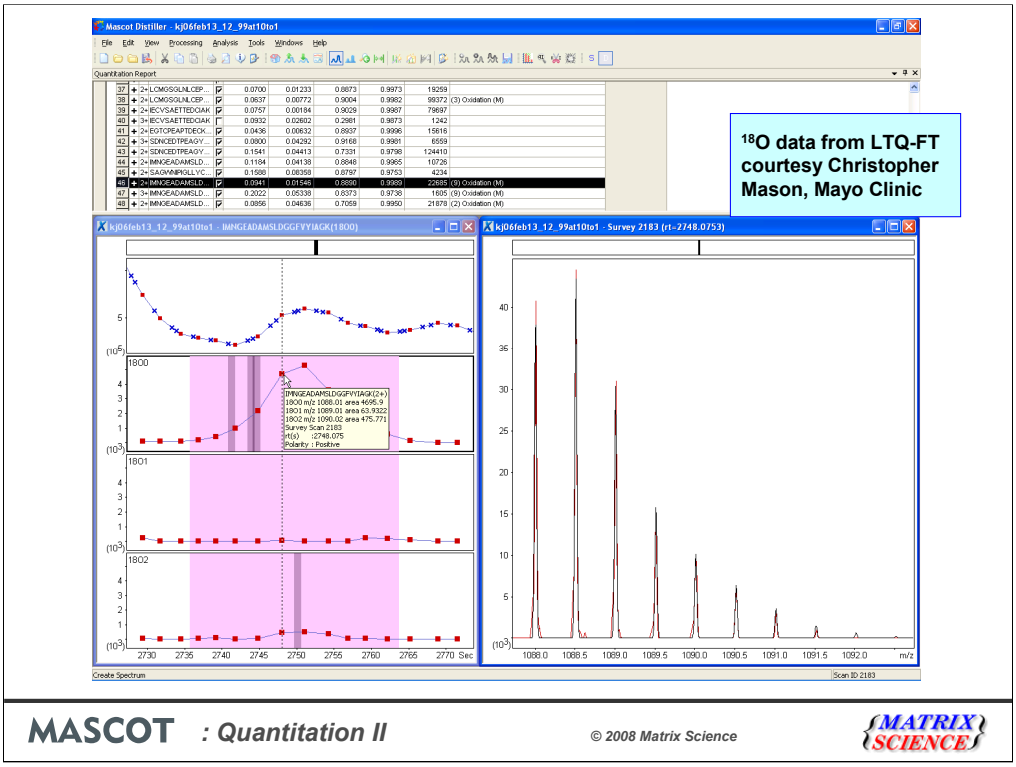


You can spend many happy hours devising different ways to arrange the Windows.

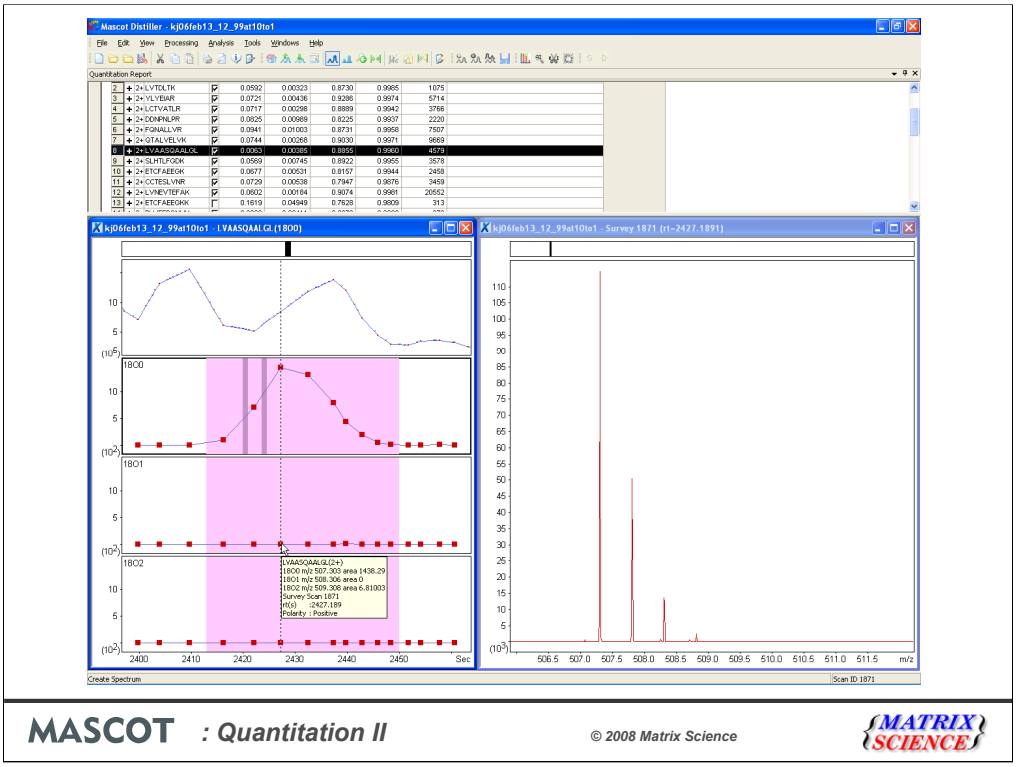
With the tree and quant grid unpinned, we can see the chromatogram and scan windows more clearly. A tooltip for each scan provides mass and charge information.

The light purple region is the XIC peak. These are the scans that have been integrated to determine the ratios. The grey bands, which look dark purple when within the XIC peak, are the scans for which we have Mascot matches. In this particular case, there is no match for the medium. The precursor is inferred from its mass and co-elution.

Whenever isotope distributions overlap, as happens here for the medium and heavy, this is accounted for in all calculations, as is incomplete enrichment. Obviously, in an individual scan, the fit is unlikely to be perfect. However, I hope you'll agree that we can immediately judge that the ratio isn't too bad and we can see there are no serious interferences or other problems in this particular scan

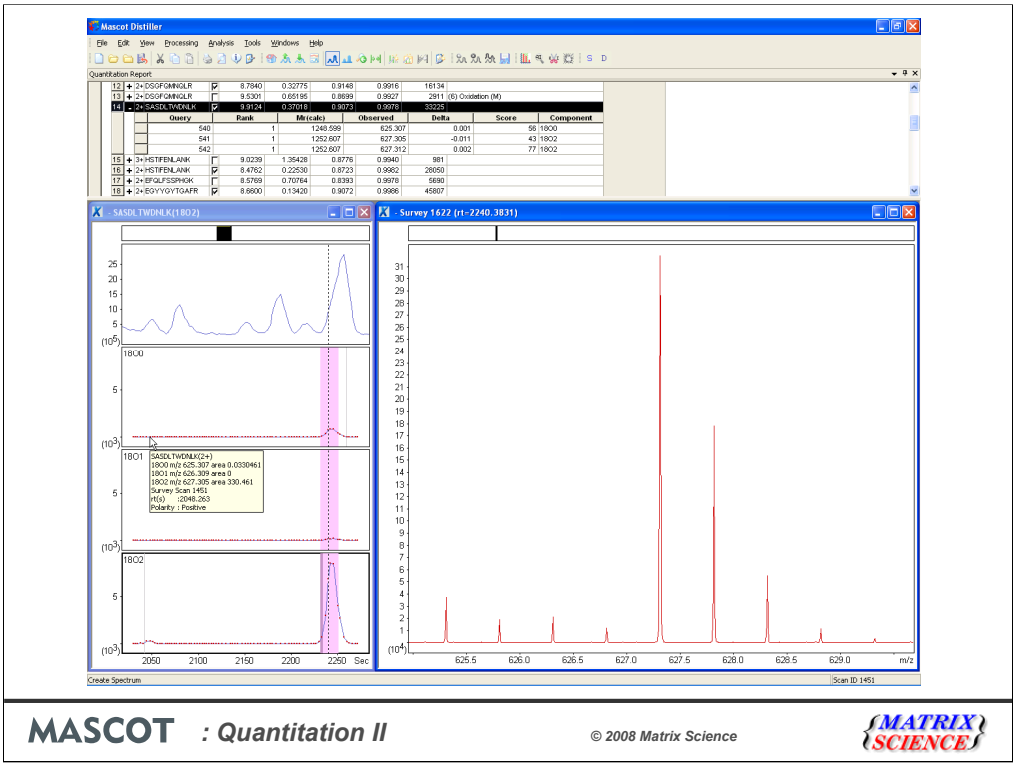


This is O-18 data from an LTQ-FT, courtesy Christopher Mason, Mayo Clinic. The sample is part of a dilution series and this one is particularly tricky because the ratio is 10:1, light to heavy. Obviously, with only a 4 Da separation, the isotope distribution for the heavy is smoothly overlapped by the tail of the light distribution, particularly for larger peptides. At the protein level, the ratios come out around 0.08, which we think isn't too bad.

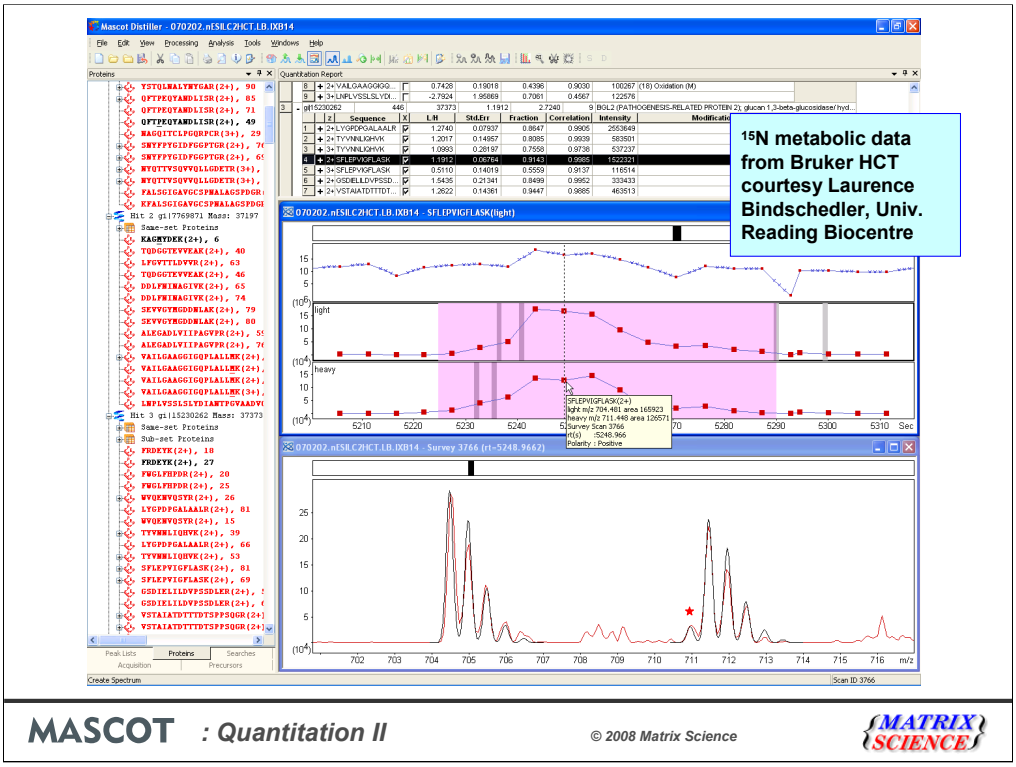


MASCOT : Quantitation II © 2008 Matrix Science MATRIX SCIENCE

This is the C-terminal peptide of BSA, which should be unlabelled, so reassuring to see the ratio is indeed close to zero

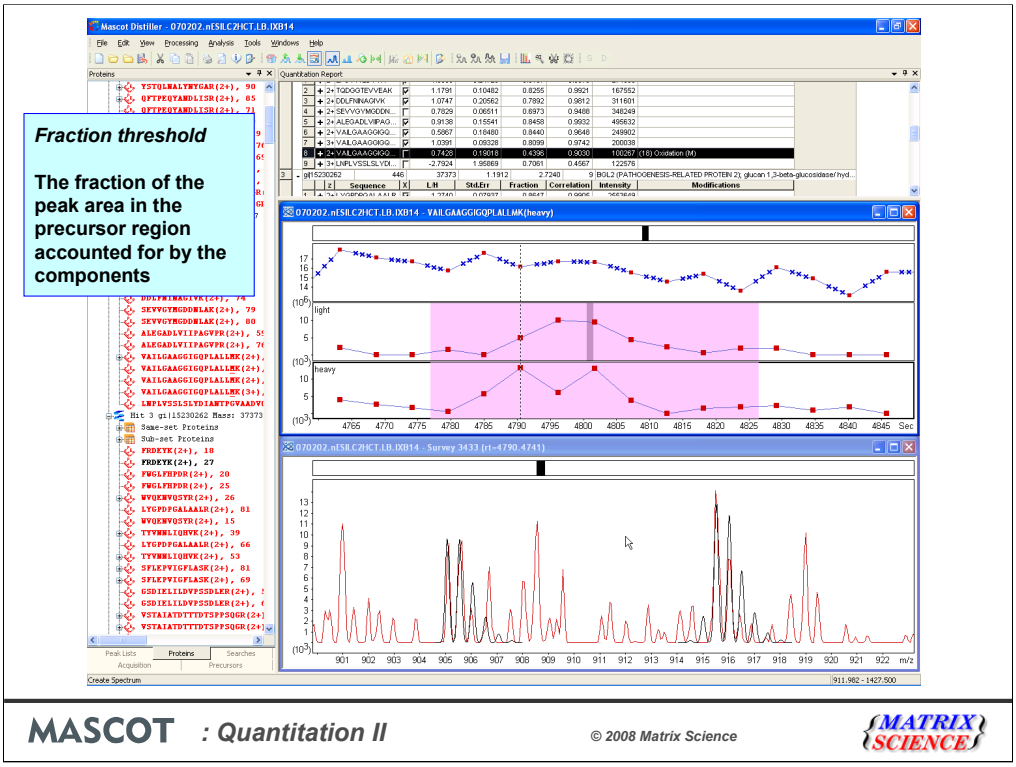


Sometimes, XIC peak detection is a challenge. The starting point is set of scans for which we have Mascot matches. You can hardly see, but in this case, there is a grey band way out here for a match obtained some 3 minutes before the main peak came through. The XIC peak detection algorithm has decided that the bulk of the signal is within the much narrower region to the right, as shown by the purple highlight. This, by the way, is still O-18, but the sample is now 1:10 light to heavy.

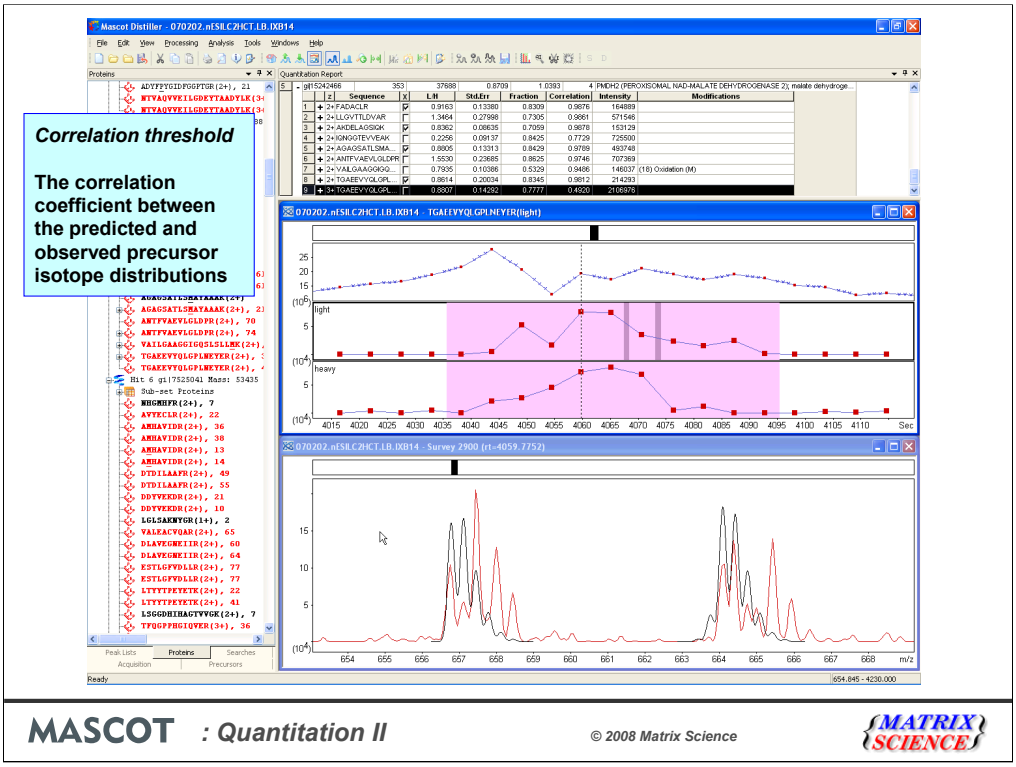


Now some metabolic data, courtesy of Rainer Cramer's group at the Reading Biocentre. In the heavy component, all the nitrogens in the proteins have been replaced by N-15. The calculated distributions (black) fit reasonably well to the observed peaks (red). Notice the peak marked with a red star. This is due to the 1% isotope impurity.

For this particular ratio, there is evidence for some interference between the light and heavy precursors, and also something happening off to the right, but in general, this is a clean spectrum and can be expected to give a good ratio. We expect that most people will want the software to make this judgement, most of the time. In the quantitation grid, the figures in the columns headed standard error, fraction, and correlation are measures of the quality of the spectrum in the precursor region. The software compares these numbers with thresholds to decide whether to accept or reject a ratio. If you want, you can over-ride these decisions using the checkboxes. I'll describe briefly how each of these criteria work

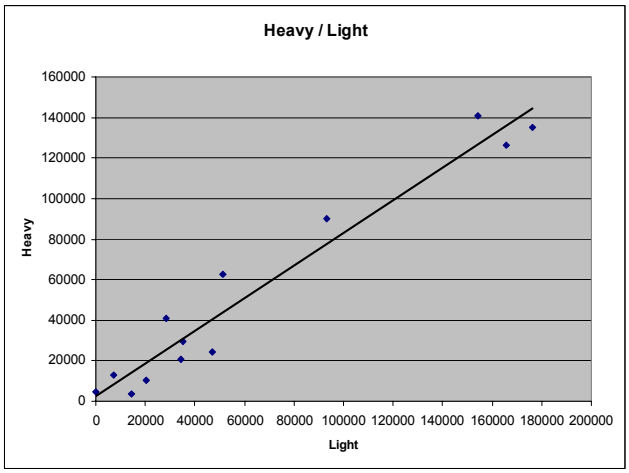


Fraction is the fraction of the peak area in the precursor region accounted for by the components. Here is a spectrum where there are a lot of interfering peaks. When we add up the areas, the expected precursors only account for 44% of the area, so the ratio is rejected. This threshold, like the others, is set as part of the quantitation method.



What about a case where the interfering peaks come right on top of the precursor peaks? A test on the matched fraction won't help. This is why we have a second test on the shape of the distribution. This uses the correlation coefficient between the predicted and observed precursor isotope distributions. Here is a case where a ratio fails the test, because the correlation coefficient is only 0.49

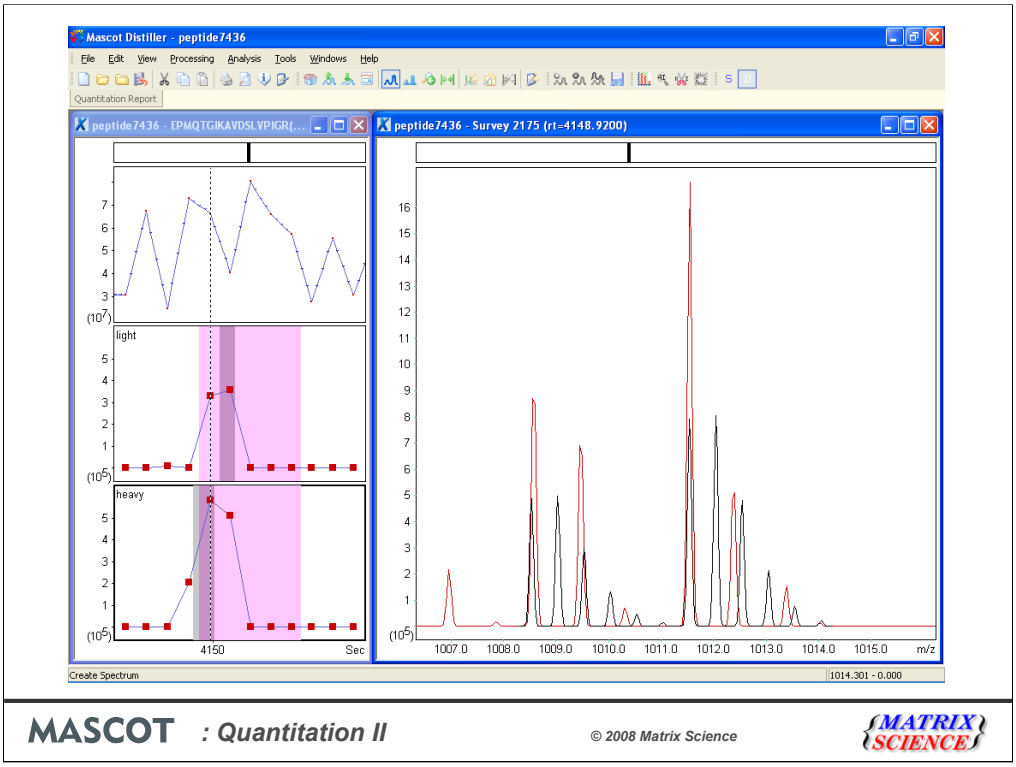
Quantitation - Precursor



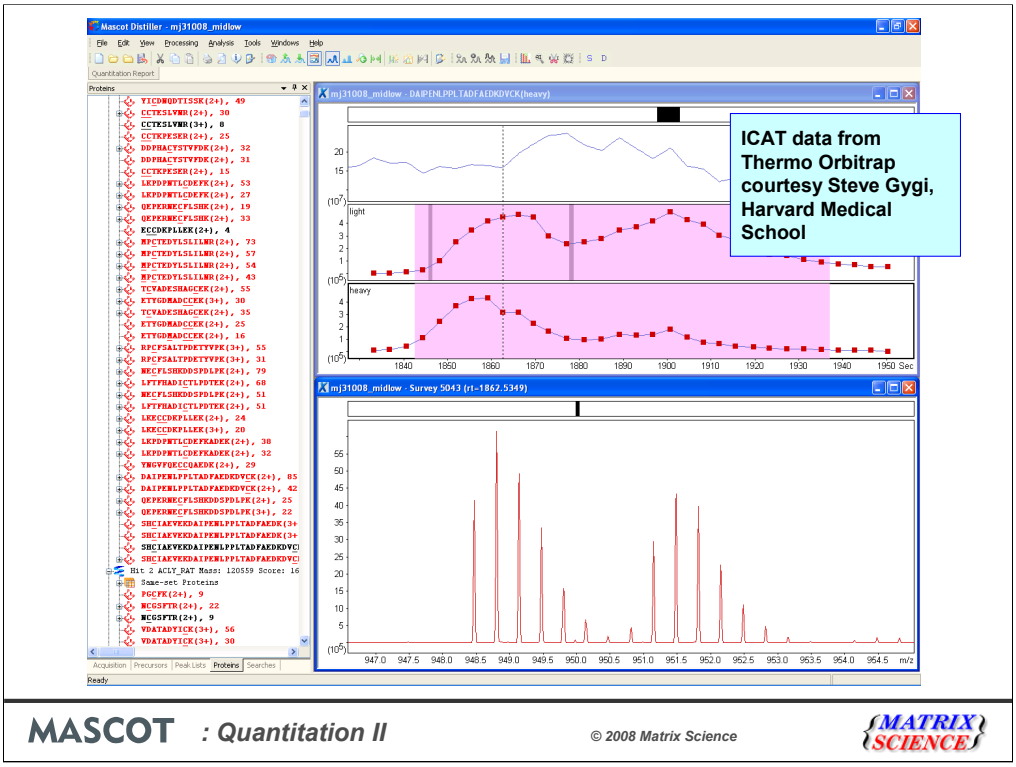
Each point represents the heavy and light intensities in a scan from the XIC peak. The straight line is a least squares fit

- The gradient is the best estimate of the ratio.
- Any background is corrected
- The standard error for the fit is a good measure of the reliability of the ratio

Another column reports the estimated standard error for the calculated ratio. Each ratio comes from making a least squares fit to the component intensities from the scans in the XIC peak. Here, for example, each point represents the heavy and light intensities in one scan. The gradient of the fitted line is the best estimate of the ratio. The standard error for the fit is a good measure of the reliability of the ratio, and can simply be tested against another threshold.



Of course, as always, garbage in means garbage out. Here is a case where the raw data are centroids, not profiles. Although we have good strong MS/MS, when you look at the survey scans, there are no proper isotope distributions. The monoisotopic peaks are approximately in the right place, but the spacings to the isotope peaks are almost random. The charge state is 2+, and the calculated overlay shows what the distributions should look like. Presumably, there was something seriously wrong with the original peak picking. We simply cannot expect to get decent quantitative information out of data like this.



So, just like with protein ID, the better the data, the better the quantitation results. Precursor resolution is clearly a good thing, because it makes peak detection easy and delivers high signal to noise ratios. This is ICAT data from an Orbitrap courtesy of Steve Gygi. Because this is deuterium ICAT, the components do not precisely co-elute, and the software has to try and align the chromatograms. This makes data quality especially critical.