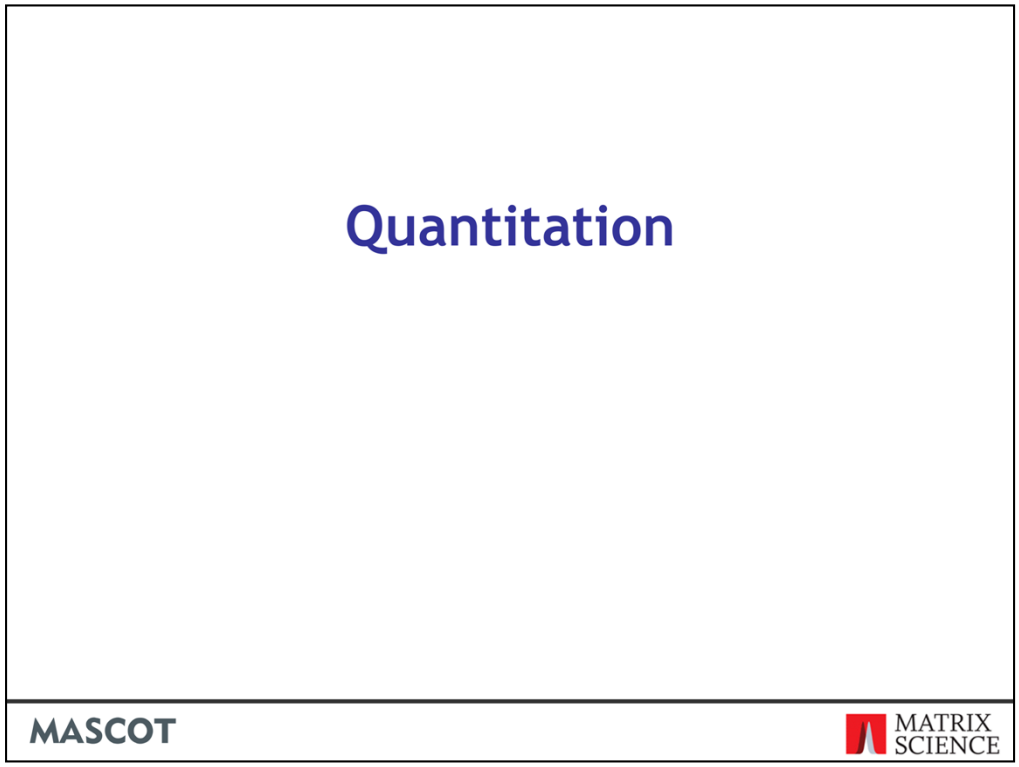


Quantitation was first introduced in Mascot 2.2. Our goal is to support all of the popular methodologies.

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

MASCOT : *Quantitation* MATRIX SCIENCE

To make this task manageable, we have classified the various approaches into a limited number of protocols. So far, we have identified 6 distinct protocols.

Reporter is quantitation based on the relative intensities of fragment peaks at fixed m/z values within an MS/MS spectrum. For example, iTRAQ or Tandem Mass Tags

Precursor is quantitation based on the relative intensities of extracted ion chromatograms (XICs) for precursors within a single data set. This is by far the most widely used approach, which can be used with any chemistry that creates a precursor mass shift. For example, 18O, AQUA, ICAT, ICPL, Metabolic, SILAC, etc., etc.

Multiplex is quantitation based on the relative intensities of sequence ion fragment peaks within an MS/MS spectrum. This is a novel approach, which can be used with any chemistry that labels one peptide terminus, creating a small mass shift, such as 18O or SILAC under certain conditions.

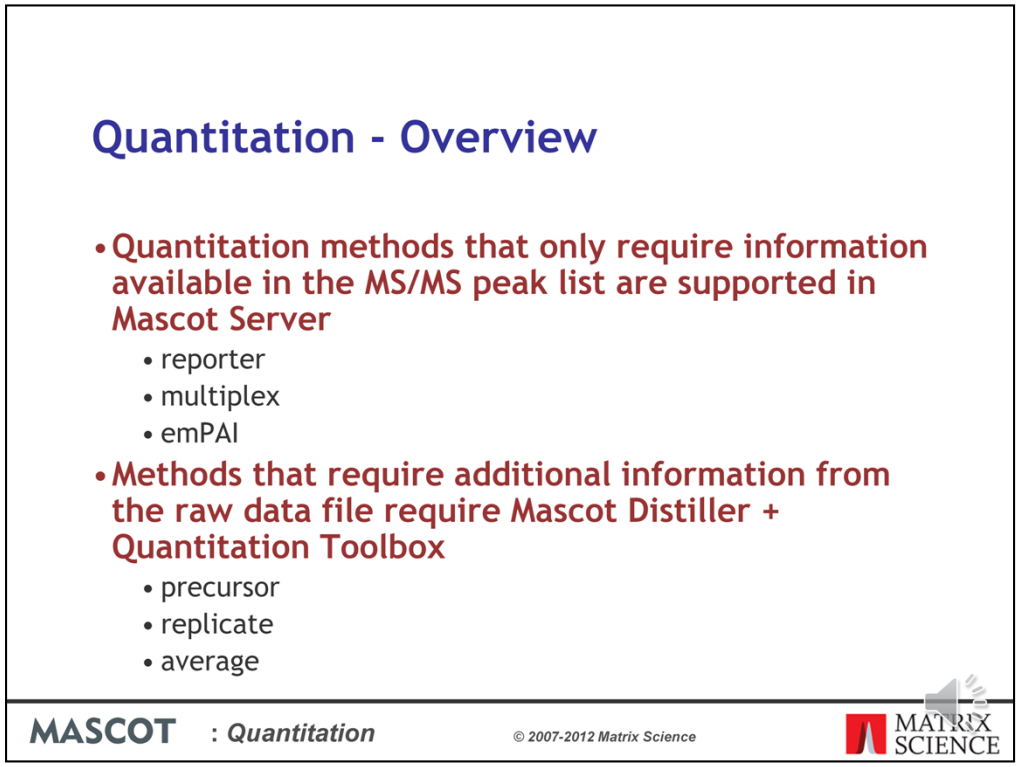
Replicate is label free quantitation based on the relative intensities of extracted ion chromatograms (XICs) for precursors in multiple data sets aligned using mass and elution time.

All these four methods are used to measure the relative abundance of a protein from sample to sample. For example, whether a particular protein is up or down regulated when an organism is stressed or diseased. The next two methods are used to estimate the relative abundances of different proteins within a single mixture.

emPAI is quantitation for the proteins in a mixture based on protein coverage by the peptide matches in a database search result.

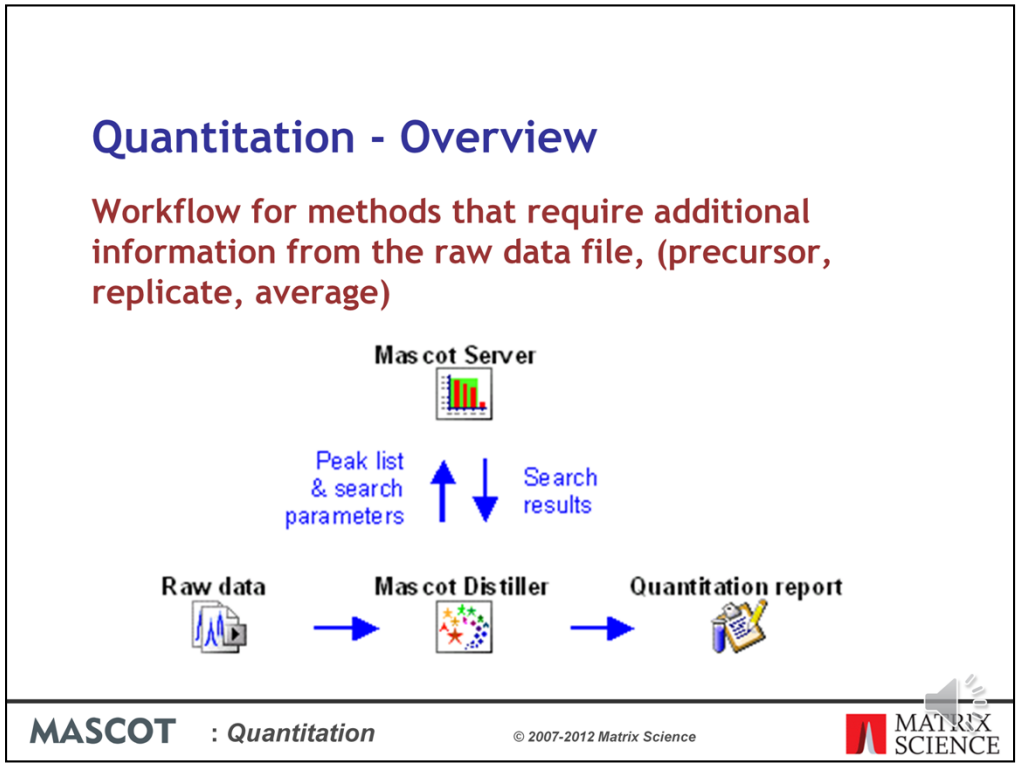
Average is quantitation for the proteins in a mixture based on the application of a rule to the intensities of extracted ion chromatograms (XICs) for the peptide matches in a database search result. For example, the average intensity for the three strongest peptide matches per protein

The rows with a blue background are the protocols that implemented in the search engine, and don't require any additional software.



The common factor for these protocols is that all of the information required for quantitation is contained in the peak list.

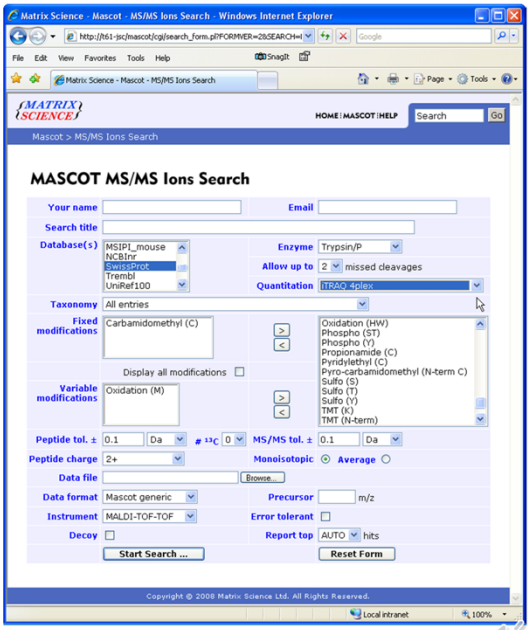
The other three methods require additional information from the raw data file, either because it is necessary to integrate the elution profile of each peptide or because information is required for multiple peaks in the survey scan. These methods require that the raw data files are processed using Mascot Distiller.



For methods that require additional information from the raw data file, the workflow looks like this. The raw data file is processed in 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.

Quantitation

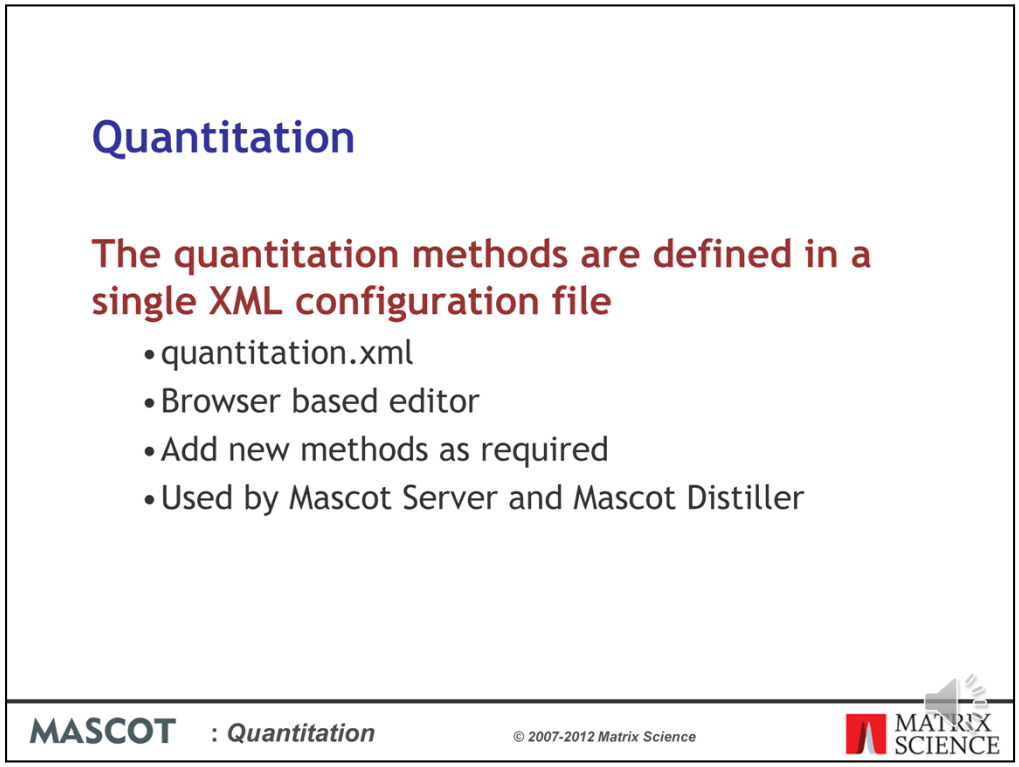
Named quantitation methods keep the search form uncluttered



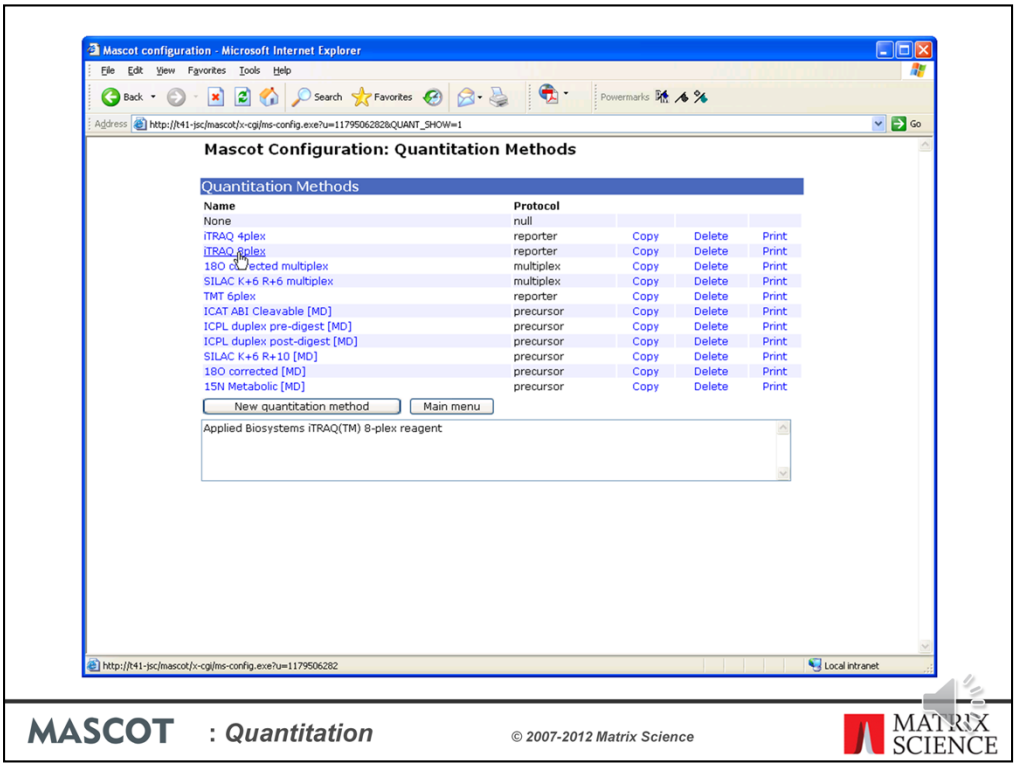
We wanted to keep the user interface simple. Quantitation adds a huge number of choices and parameters, but there is no point in exposing all of these in the search form.

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 control.

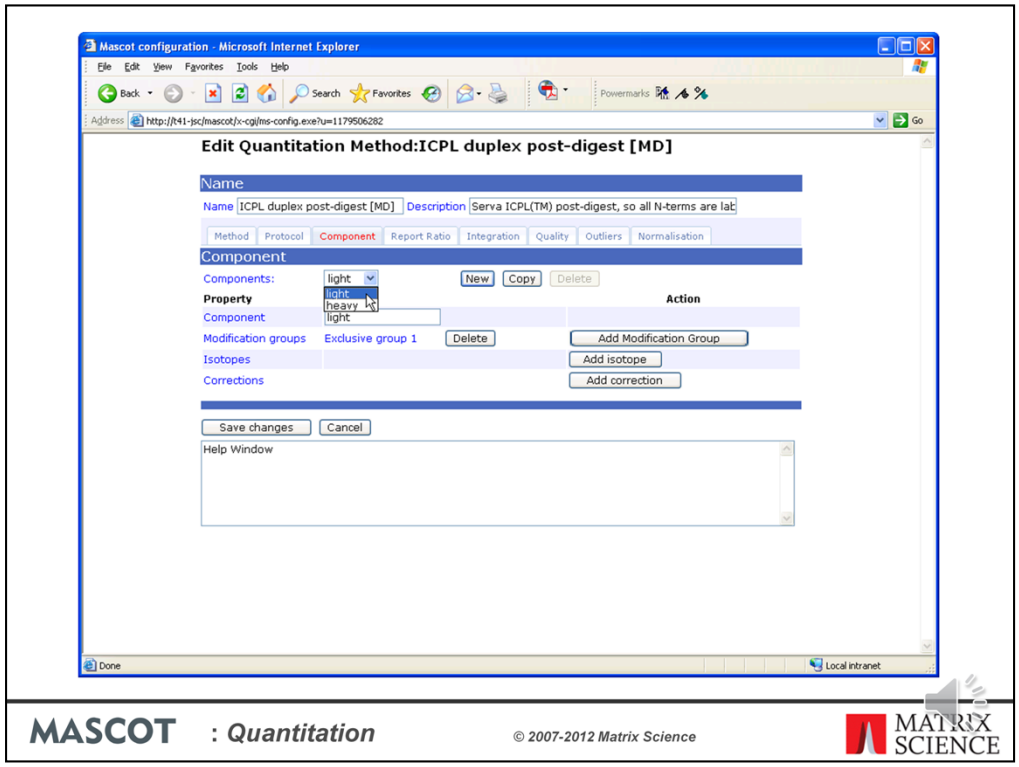
Methods that have [MD] at the end are the ones that require Mascot Distiller



The configuration file that encapsulates the choices and parameters for each quantitation method is called quantitation.xml. This is an XML file, and there is a browser based editor for modifying methods and creating new ones. quantitation.xml lives on the Mascot server and is read by both the search engine and Mascot Distiller

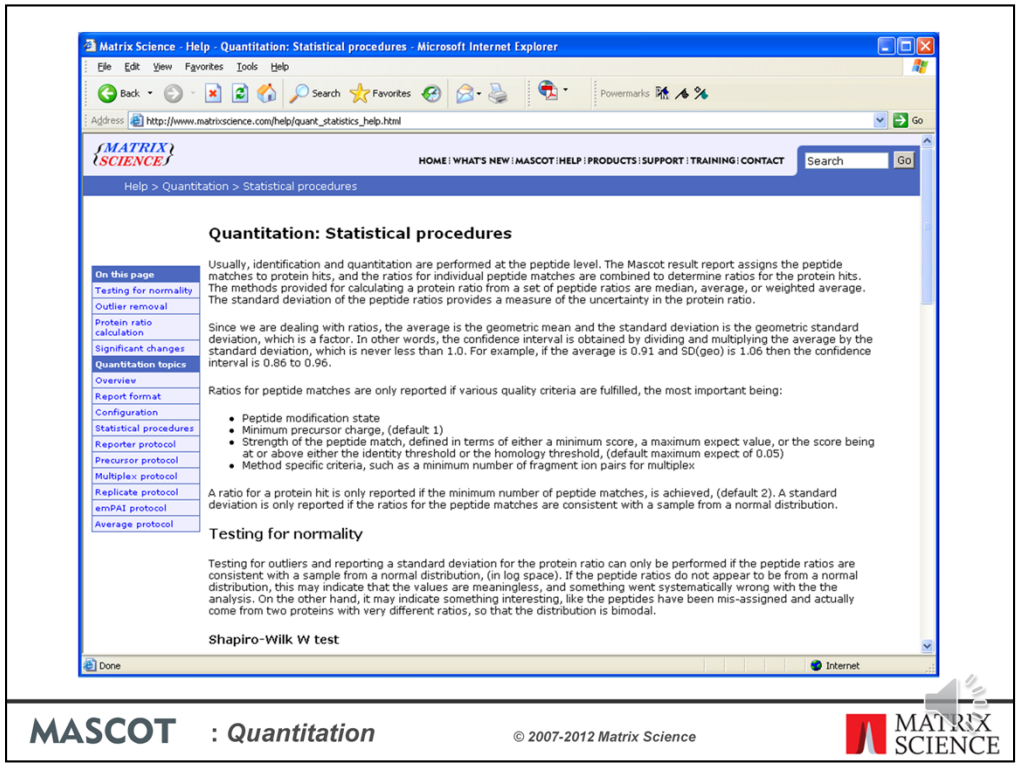


The browser-based Configuration Editor provides an interface to all the Mascot configuration files. In the case of quantitation, you can edit an existing method or make a copy of it as the basis for a new method



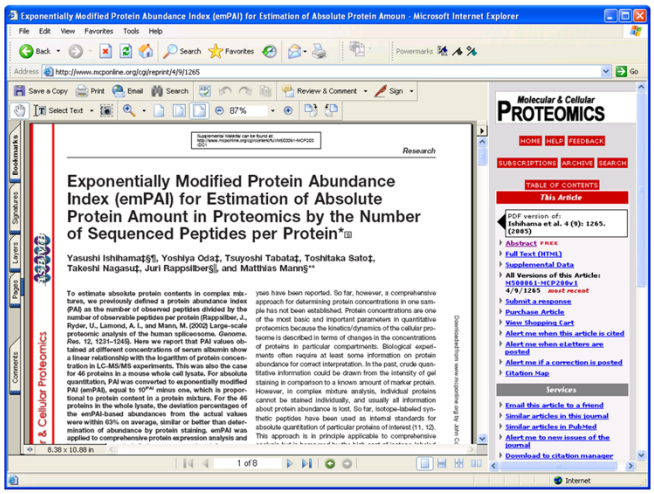
For each method, a tabbed dialog is used to navigate between property pages. In many cases, the property pages correspond to XML elements, but sometimes elements have been combined onto a single page or split across multiple pages so as to give a balanced layout.

Here, we can see a duplex ICPL method. The unlabelled and labelled components have been called heavy and light, but you are free to choose your own names so as to make the final report as clear as possible.



We have taken trouble to ensure that appropriate statistical procedures are correctly used. For example, we test that a set of peptide ratios is consistent with a normal distribution before rejecting outliers or reporting a standard deviation. Standard deviations are always geometric, because we are dealing with ratios that conform to a normal distribution in log space.

Quantitation - emPAI



emPAI quantitation offers approximate, label-free, relative quantitation of the proteins in a mixture based on protein coverage by the peptide matches in a database search result. This approach was developed by Ishihama and colleagues

Quantitation - emPAI

- Very simple

$$emPAI = 10^{\frac{N_{observed}}{N_{observable}}} - 1$$

- Very approximate
  - Many assumptions in $N_{observed}$ and $N_{observable}$
- 'Always on'

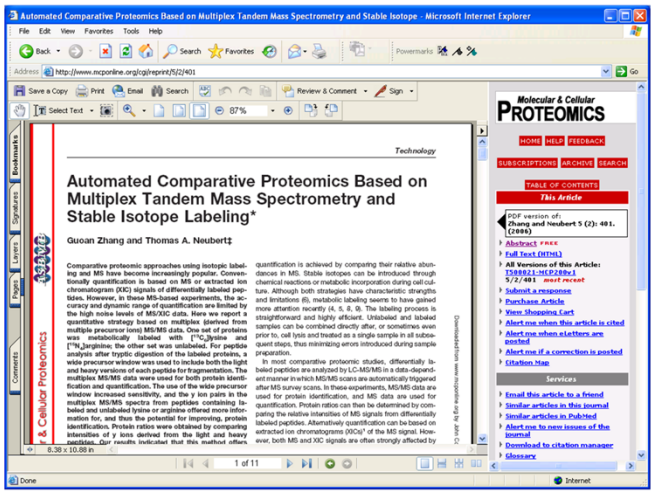
1. PPB1_HUMAN **Mass:** 58259 **Score:** 452 **Queries matched:** 17 **emPAI:** 1.04
Alkaline phosphatase, placental type precursor (EC 3.1.3.1) (PLAP-1) (Regan isozyme) - Homo sapiens
Check to include this hit in error tolerant search

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
27	462.6807	923.3468	923.5116	-0.1649	0	33	0.25	1	R.FPYVALSK.T
41	517.1760	1032.3375	1032.5604	-0.2229	0	71	6.4e-05	1	R.GSSIFGLAPGK.A
62	564.6804	1127.3463	1127.5764	-0.2301	0	9	1.2e+02	1	R.GFFLFVEGGR.I

MASCOT : *Quantitation* MATRIX SCIENCE

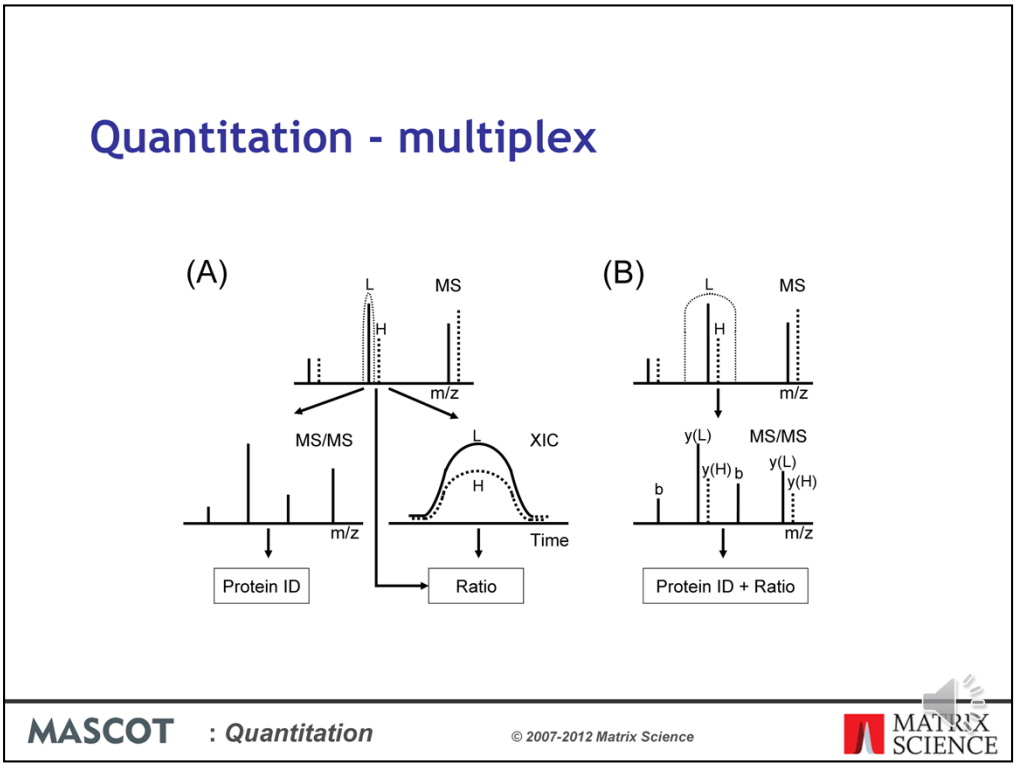
It is very simple. It is also very approximate, because there are many arbitrary assumptions in the way that the number of observed and observable peptides are calculated. Nevertheless, Ishihama's paper shows that it can be a useful guide to relative amounts. emPAI doesn't require a label or special data processing, so it is always reported in a standard Mascot results report, as long as the number of MS/MS spectra is at least 100

Quantitation - multiplex

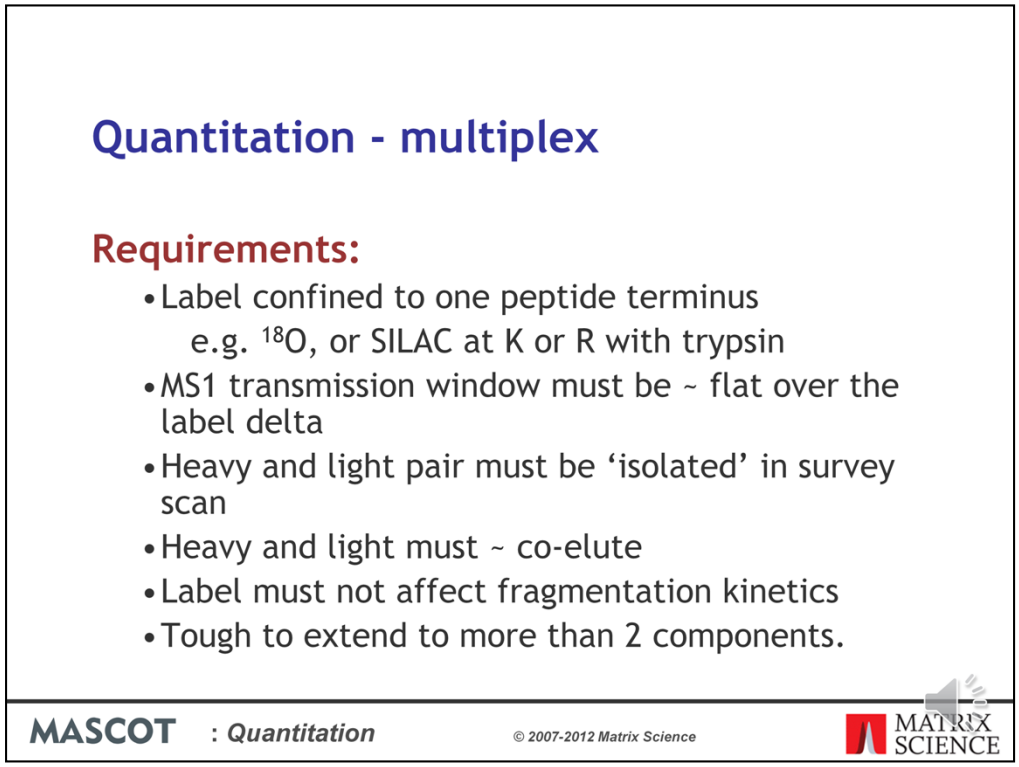


MASCOT : *Quantitation* *© 2007-2012 Matrix Science* MATRIX SCIENCE

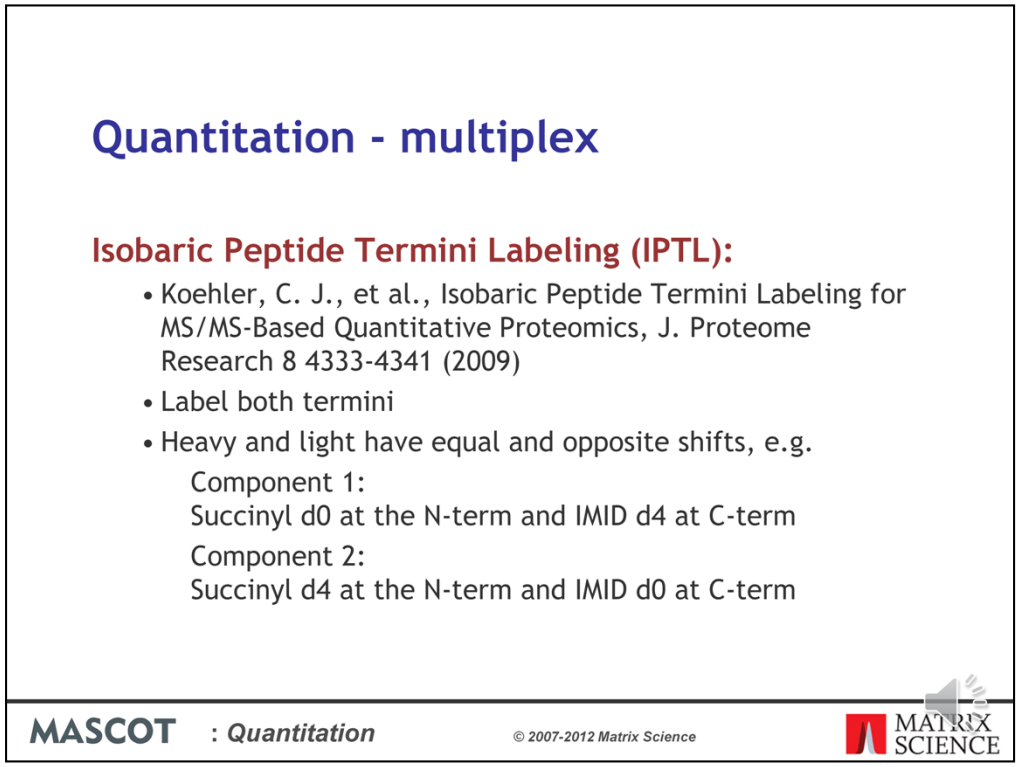
Multiplex is quantitation based on the relative intensities of sequence ion fragment peaks within an MS/MS spectrum. This approach, developed Zhang and Neubert, can be used with any chemistry that labels one peptide terminus and has a reasonably small mass shift.



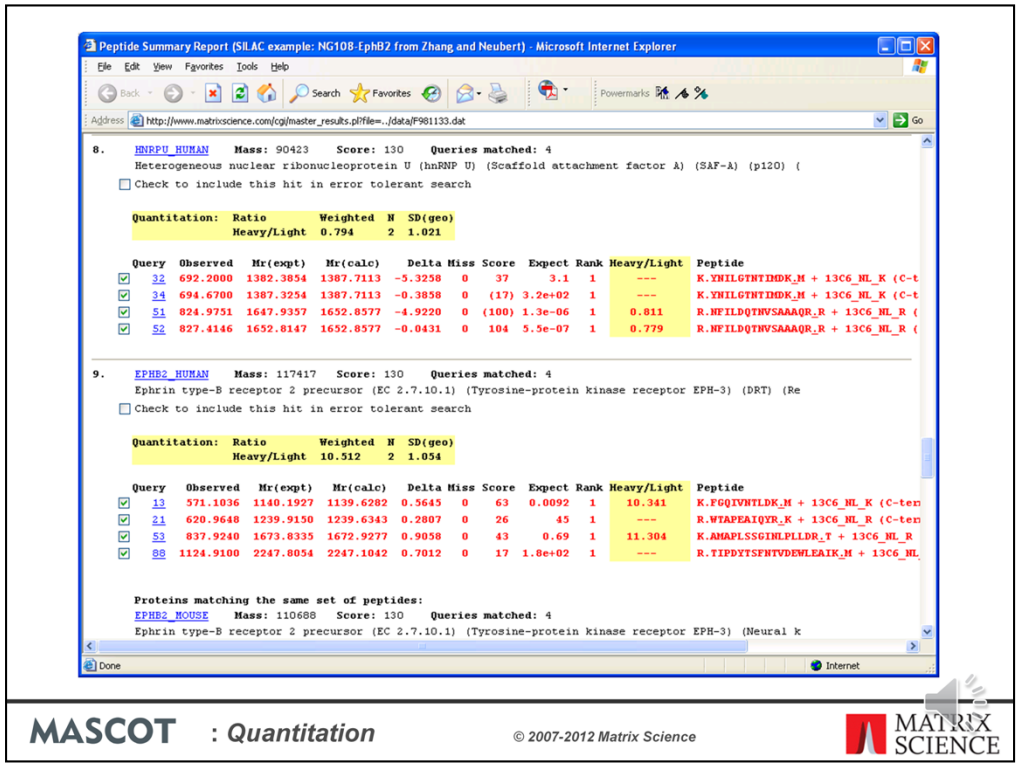
This diagram, copied from the MCP paper, illustrates how it works. On the left, we have conventional quantitation; the ‘precursor protocol’ in Mascot terms. This requires the precursor intensity for each component to be integrated across its elution profile. In the case of the multiplex protocol, the MS1 transmission window is set wide enough to allow both components through simultaneously, giving a mixed MS/MS spectrum. The relative amounts can be measured from the sequence ions that include the labelled terminus. If the label is on the carboxy terminus, we see the ratios in the y ions.



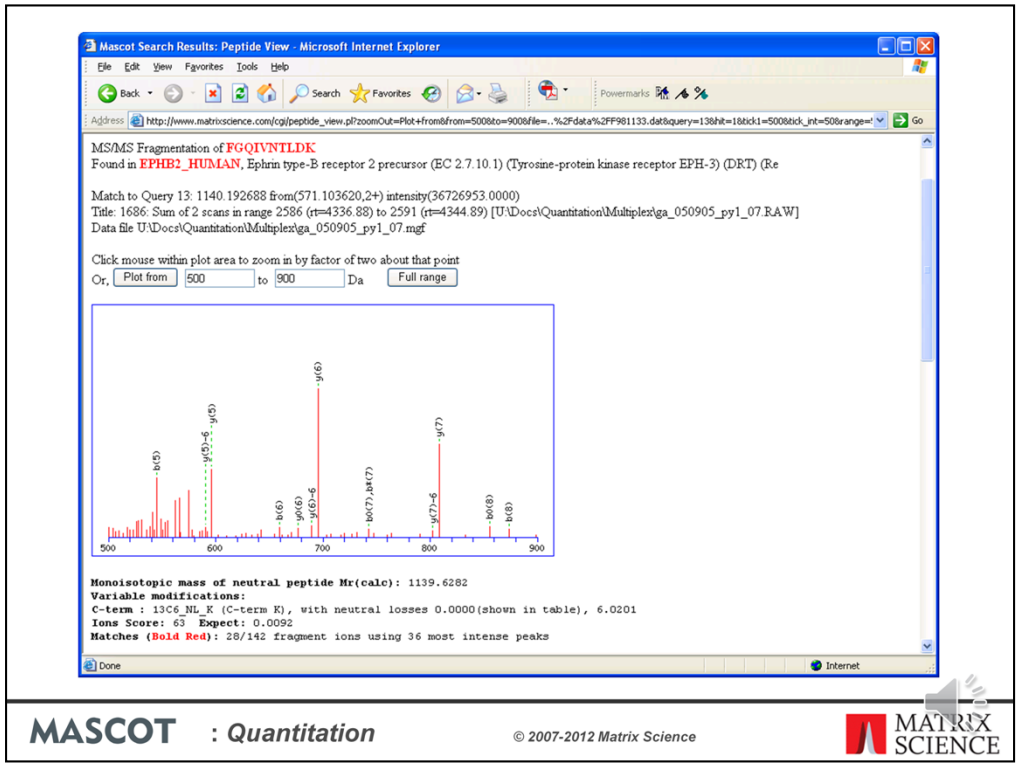
The multiplex method has the potential to give excellent precision, because each ratio is represented by multiple sequence ion pairs. On the other hand, the ratio will only be accurate if several constraints are met.



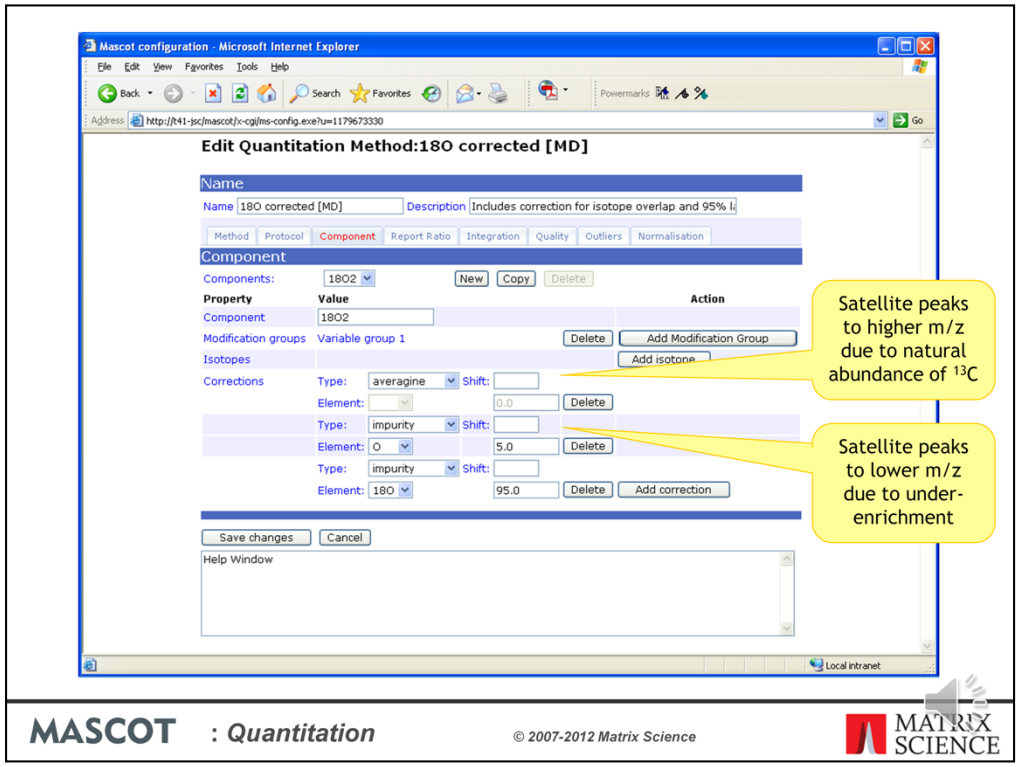
Isobaric Peptide Termini Labeling (IPTL) is a recent improvement to multiplex. This labels both termini and the difference between the two components is a mass increase at one terminus exactly balanced by a mass decrease at the other. Having isobaric precursors removes the requirement for the transmission window between MS1 and MS2 to be wide enough to accommodate the mass shift introduced by the label.



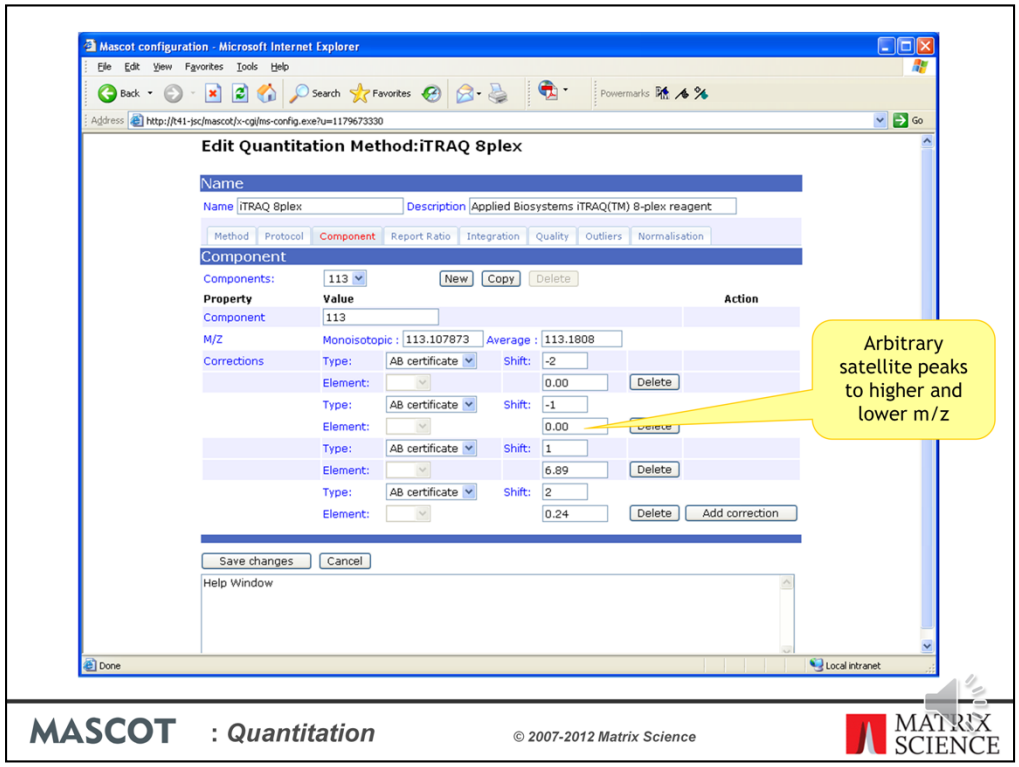
This is an example of Multiplex using a dataset courtesy of Zhang and Neubert. The instrument was an ion trap and the label is 13C(6) SILAC on K and R.



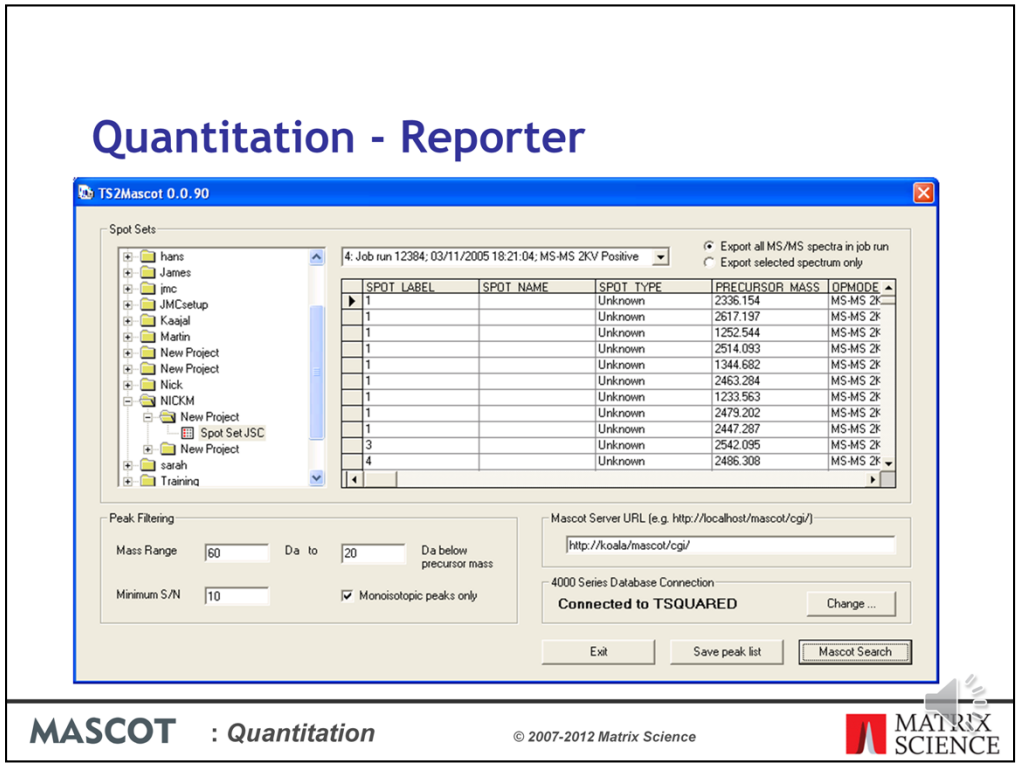
We can see that the heavy component has been strongly up-regulated in this peptide from human ephrin



One of the complications of any type of isotope labelling is isotope impurity. It is rarely possible to get 100% enrichment. In the Mascot quantitation schema, this is described by a correction element. An ‘impurity’ correction works "downwards". That is, in this 18O method, some of the intensity of peptides labelled with the 18O label will appear at lower mass values because the heavy water is only 95% enriched. A second type of isotope correction, ‘averagine’, works “upwards”. This describes how some of the intensity will be found at higher mass values because of the natural abundances of heavy isotopes. An averagine correction only matters when the mass delta is small, as in the case of 18O labelling.



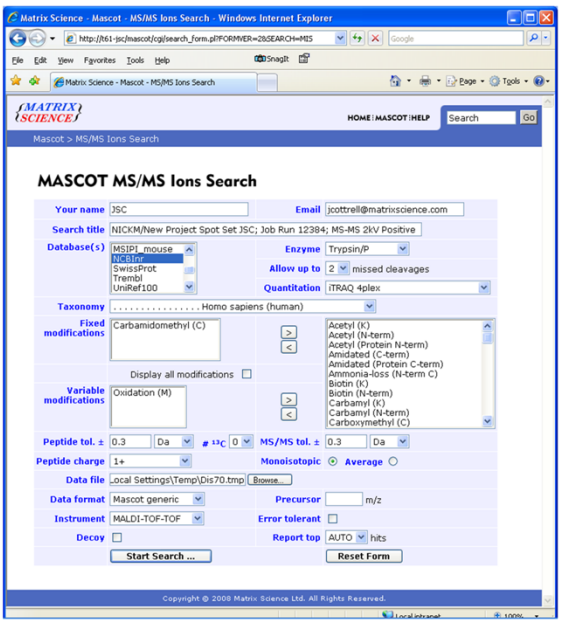
A third type of isotope correction is used in iTRAQ and TMT, where the correction factors are obtained experimentally, by analysing the isolated reagents. This combines both upward and downward corrections for labels which have complex, multi-isotope compositions



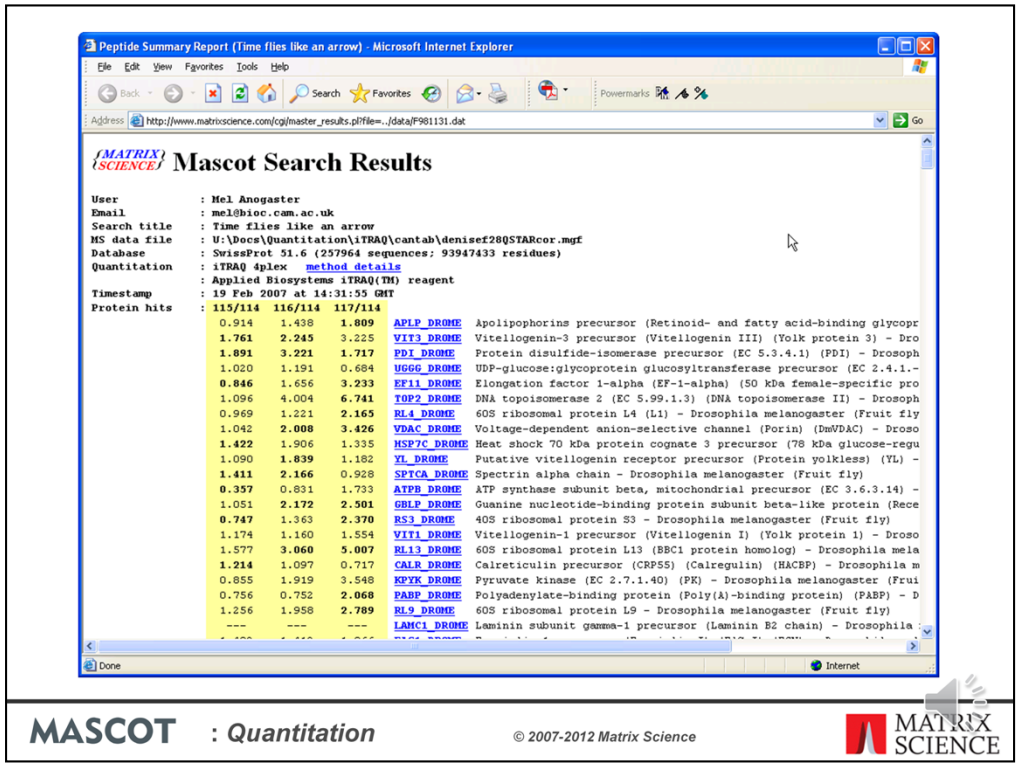
In data processing terms, the reporter protocol is one of the simplest. However, we did find that the peak list exported from the 4000 series data system or submitted to Mascot from GPS Explorer did not have the correct peak areas for the reporter ions. We had to write our own application to export an uncorrected peak list from the Oracle database. We've released this utility, called TS2Mascot, as freeware, and you can download it from our web site.

So, for iTRAQ, we could launch TS2Mascot and choose Mascot Search …

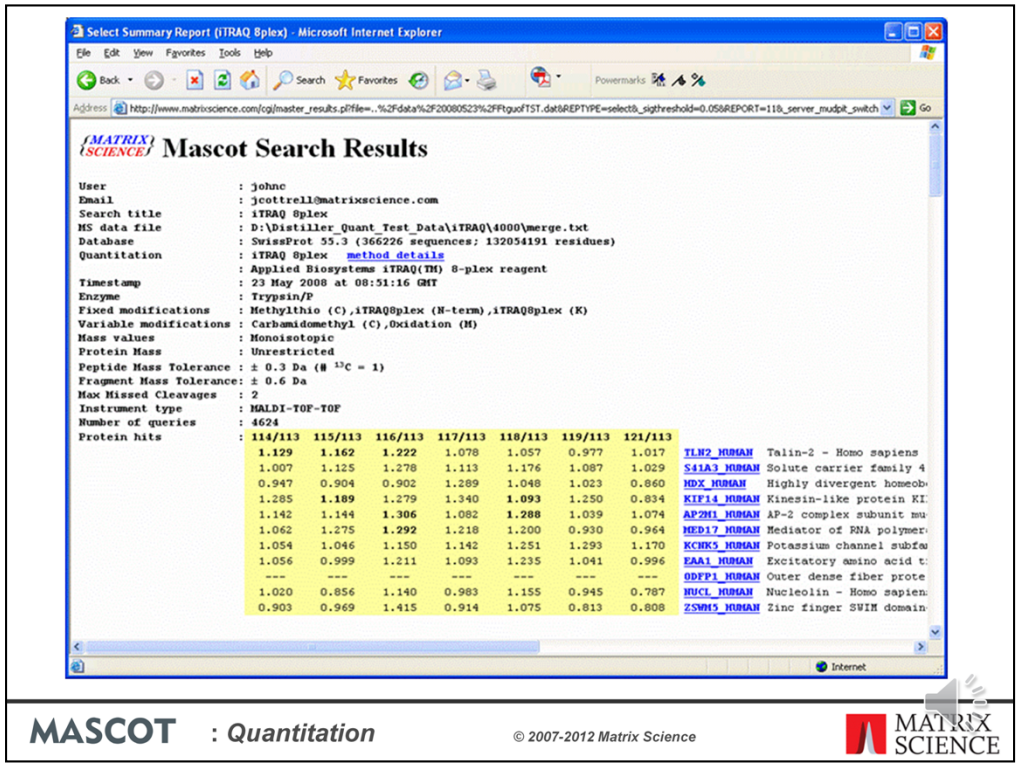
Quantitation - Reporter



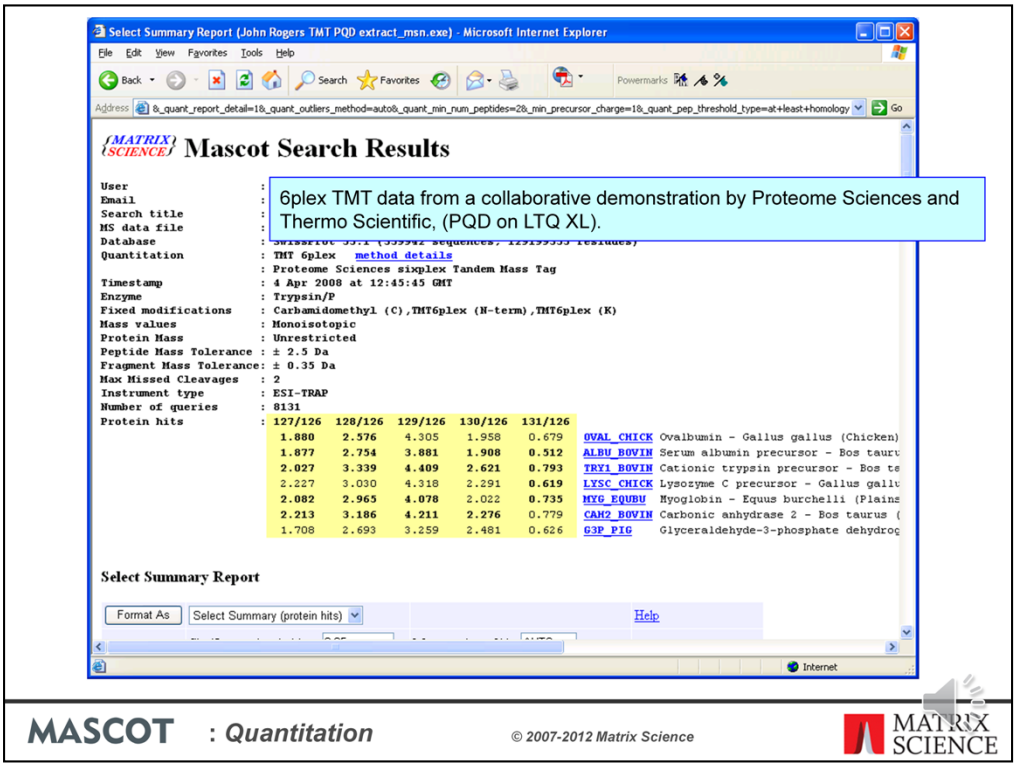
Which brings up the search form. We choose an appropriate quantitation method. We don't need to specify the iTRAQ modifications because these are pre-defined in the quantitation method. Submit the search…



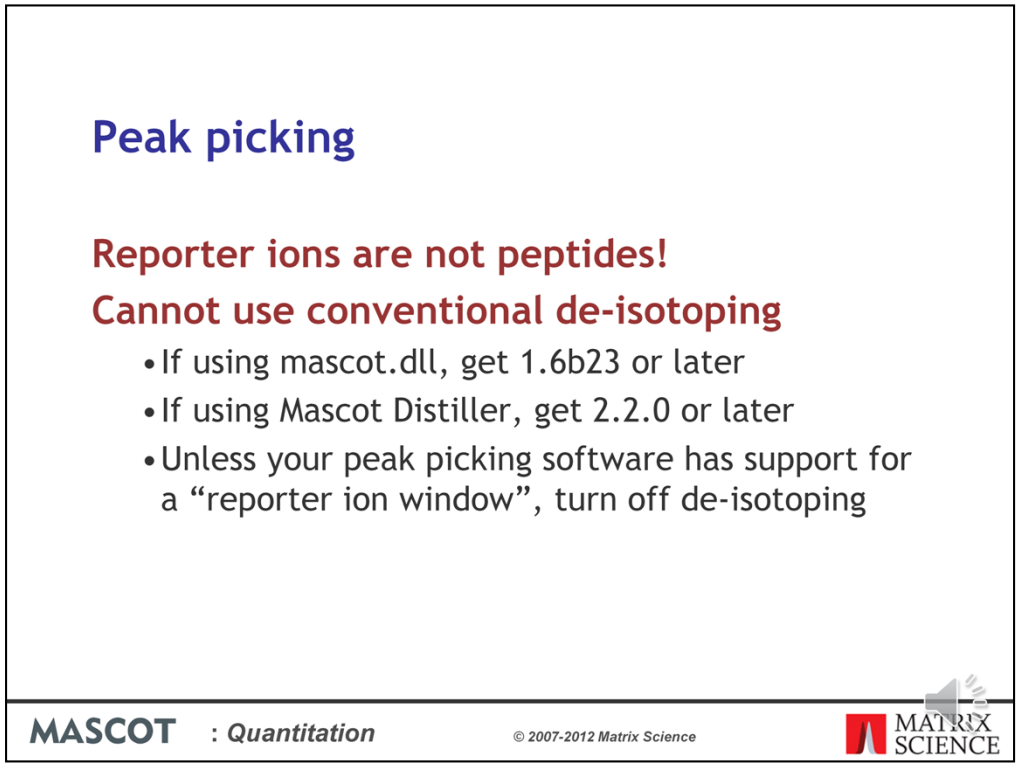
And back comes the report. At the top is a summary of the protein ratios. In this example, the method asks for ratios to 114, but you have total flexibility. You can edit the quantitation method to report two pairs, e.g. 115/114 and 117/116, or something more complex, like ratios to the sum of all four channels. Note that you can't do this if you are using our public web site, because this is a shared resource, so you don't have access to the configuration editor.



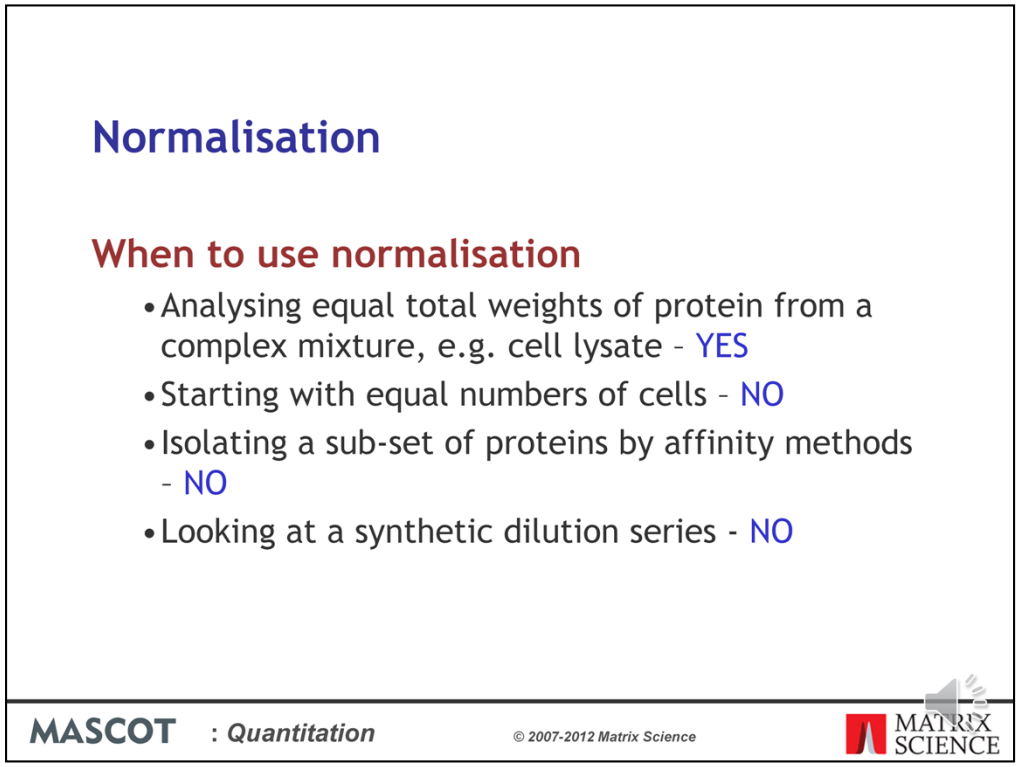
Its very easy to create a new method when a new chemistry becomes available. Here, for example, is the iTRAQ 8plex



And the TMT tags from Thermo



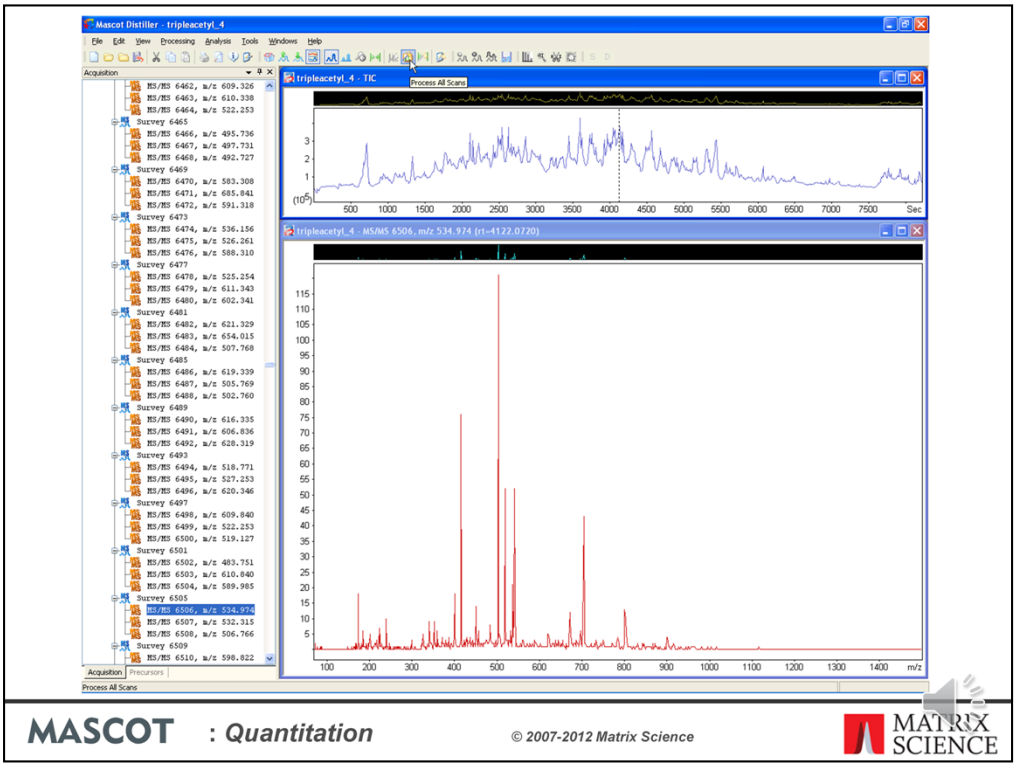
For the reporter protocol, i.e. iTRAQ or TMT, you have to be very careful with peak detection. Reporter ions do not have natural isotope distributions, so anything that assumes this will not be reliable. Whatever peak picking software you use, you'll probably need to experiment with the settings. For wiff files, the latest version of mascot.dll has a window around the reporter ion region where it does not perform de-isotoping. You can download this from the Analyst help page on our web site.



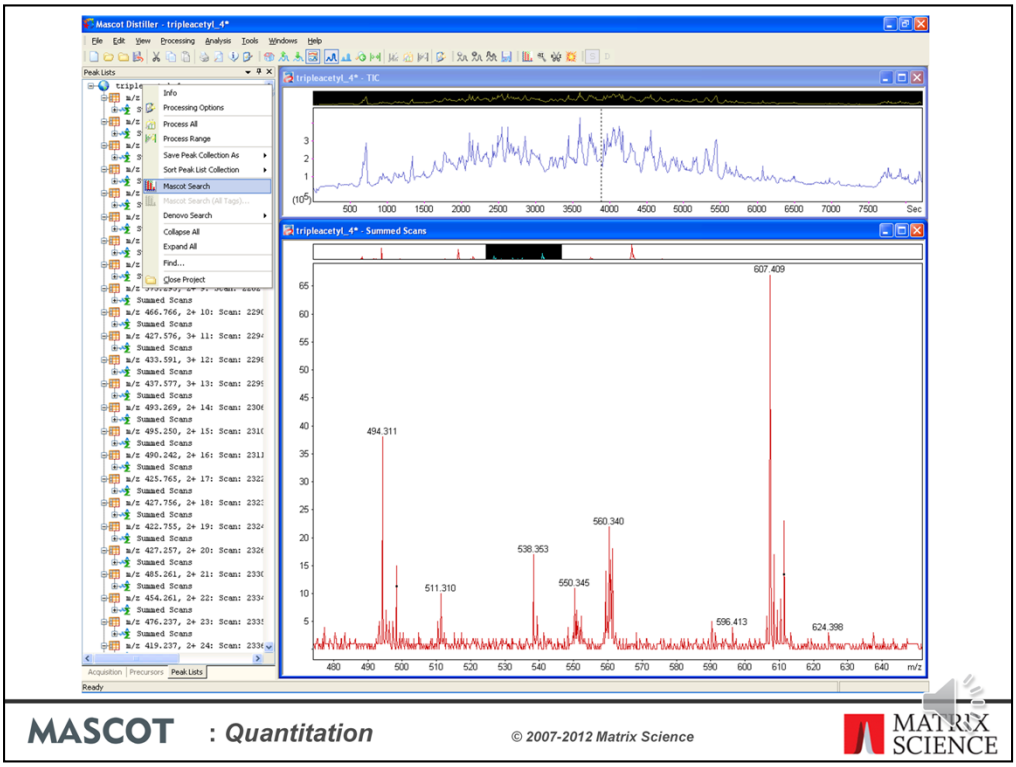
We've had a number of technical support questions about normalisation. What does it do and when should it be used?

First, you have to make an intelligent decision whether normalisation is appropriate. It depends entirely on the experiment. Normalisation forces the average or median ratio to be 1. If the average or median ratio is supposed to be 1, this is the smart thing to do, and eliminates systematic errors. In other cases, it is the wrong thing to do. For example, if you are analysing a standard, where the ratio is supposed to be (say) 3:1, you wouldn't want to force it to be 1:1

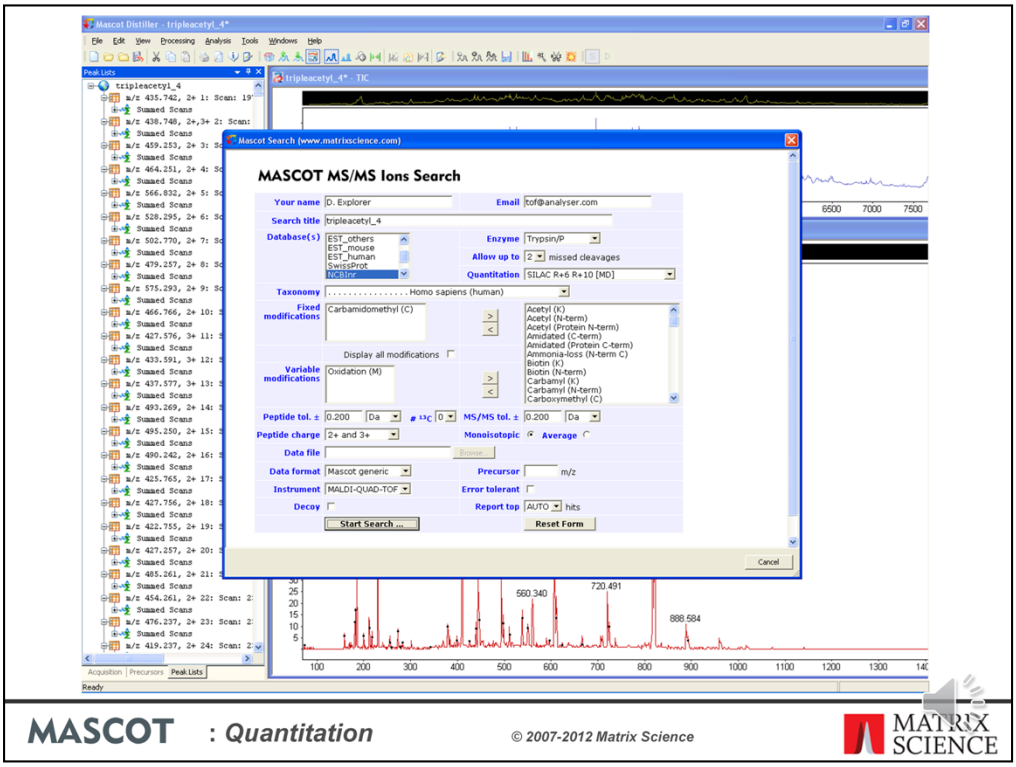
Whether to use sum, median or weighted, is best decided by running some standards (e.g. a cell lysate spiked with varying amounts of a known protein) and seeing which gives the best accuracy and precision.



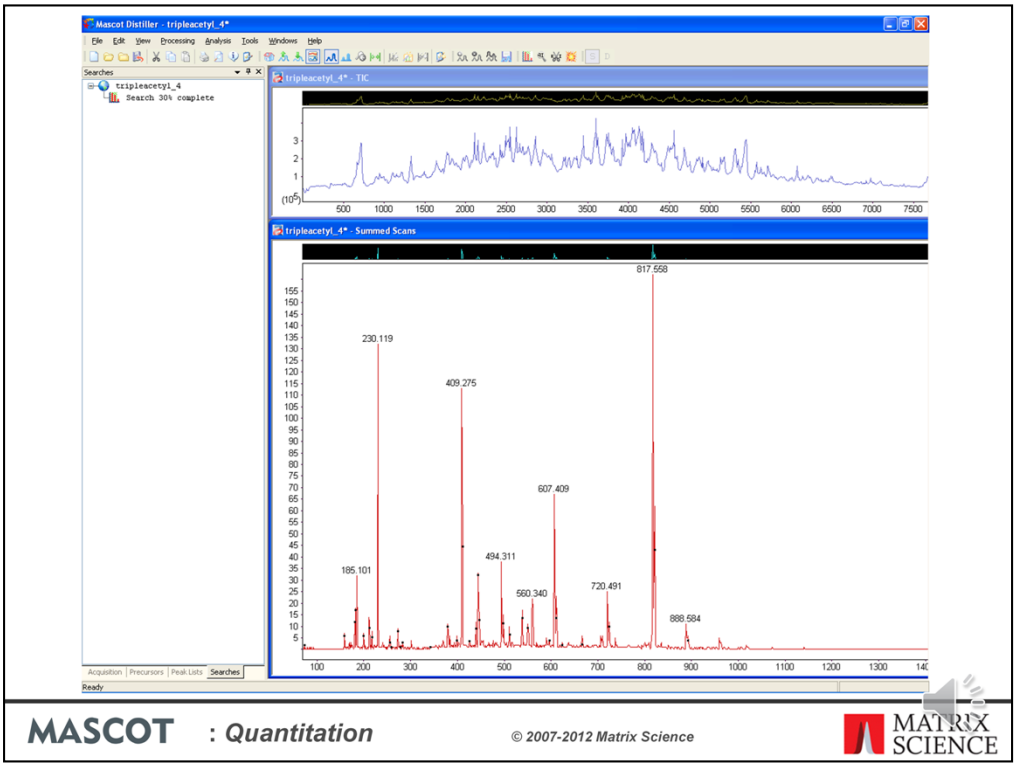
Lets turn to the precursor protocol. This is where we have to use the Mascot Distiller Quantitation Toolbox. I'm going to illustrate the workflow using one of the QStar sample data sets that can be downloaded from the MSQuant home page. This is a SILAC sample with three components: unlabelled, Arg labelled 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



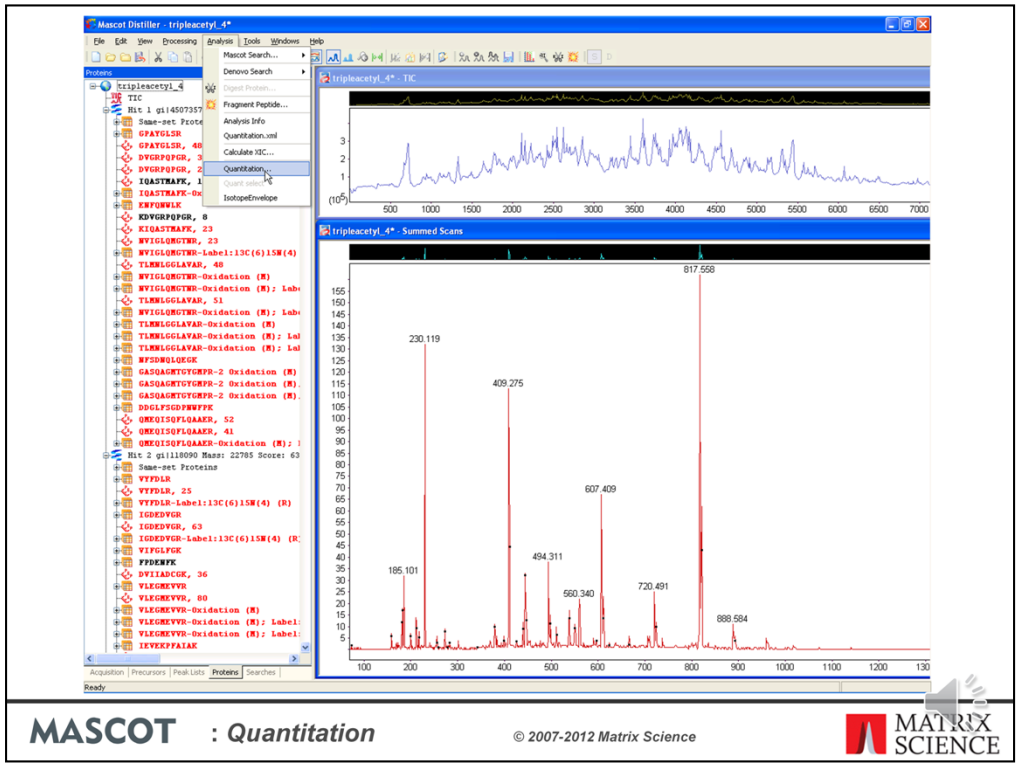
The peak lists are submitted to a Mascot Server to be searched



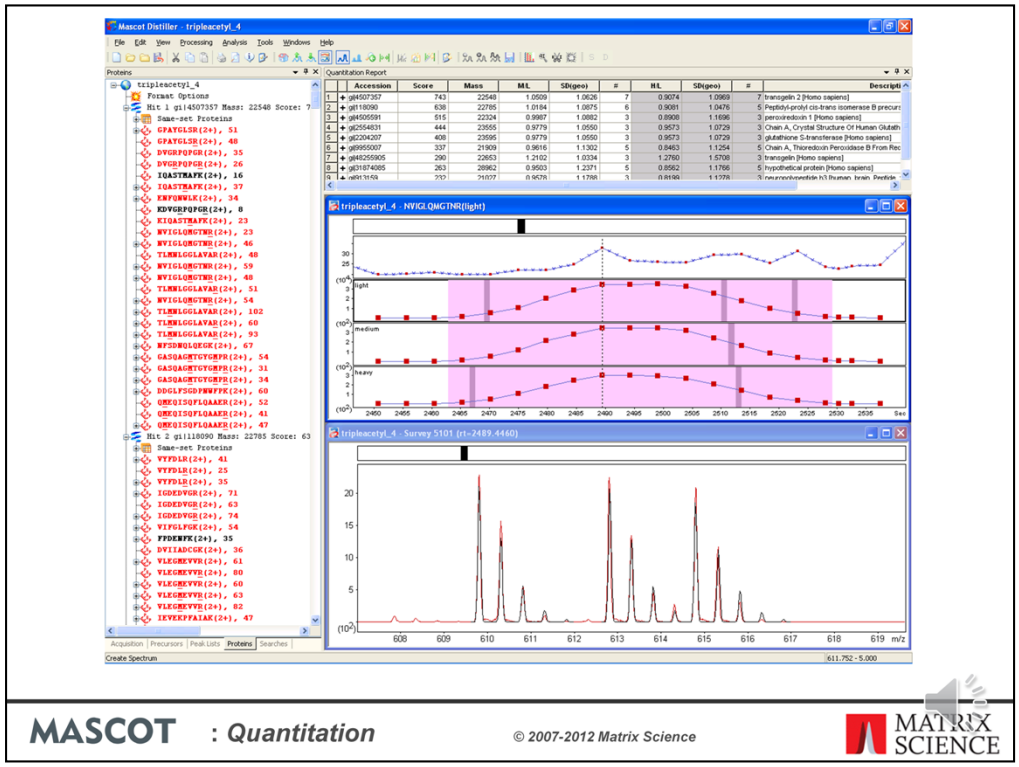
There’s an opportunity to tweak the search conditions, but most of the important settings, including the modifications, are embedded in the selected quantitation method



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



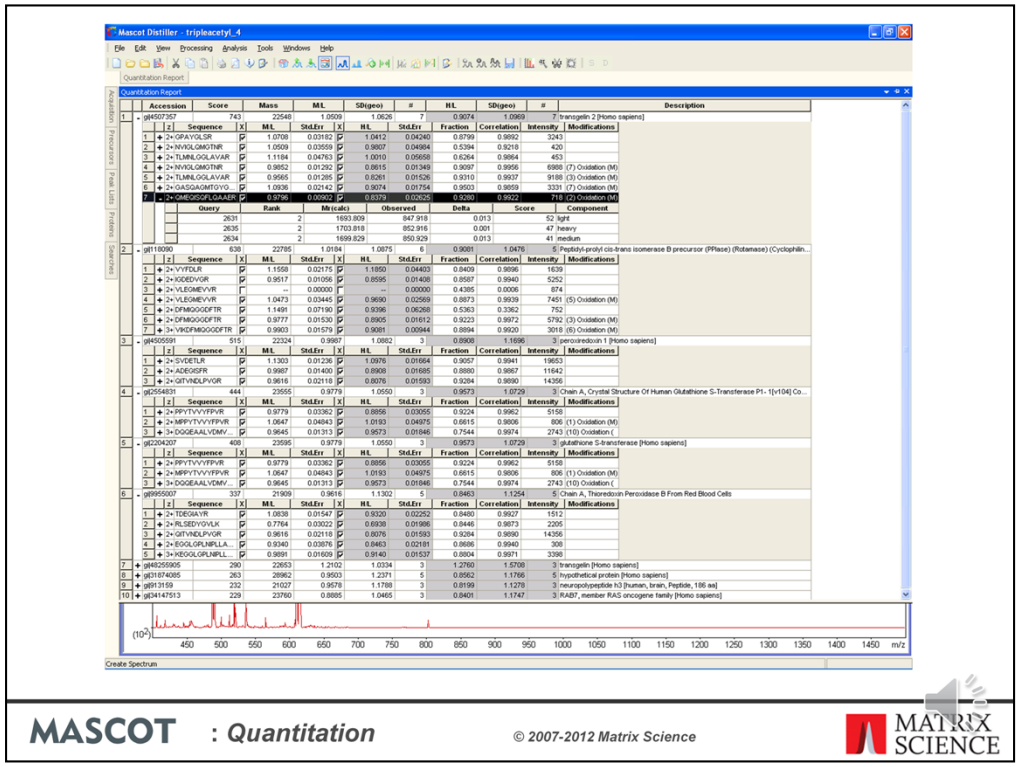
The quantitation results are displayed in a grid control, spreadsheet-style. More about this on the next slide

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

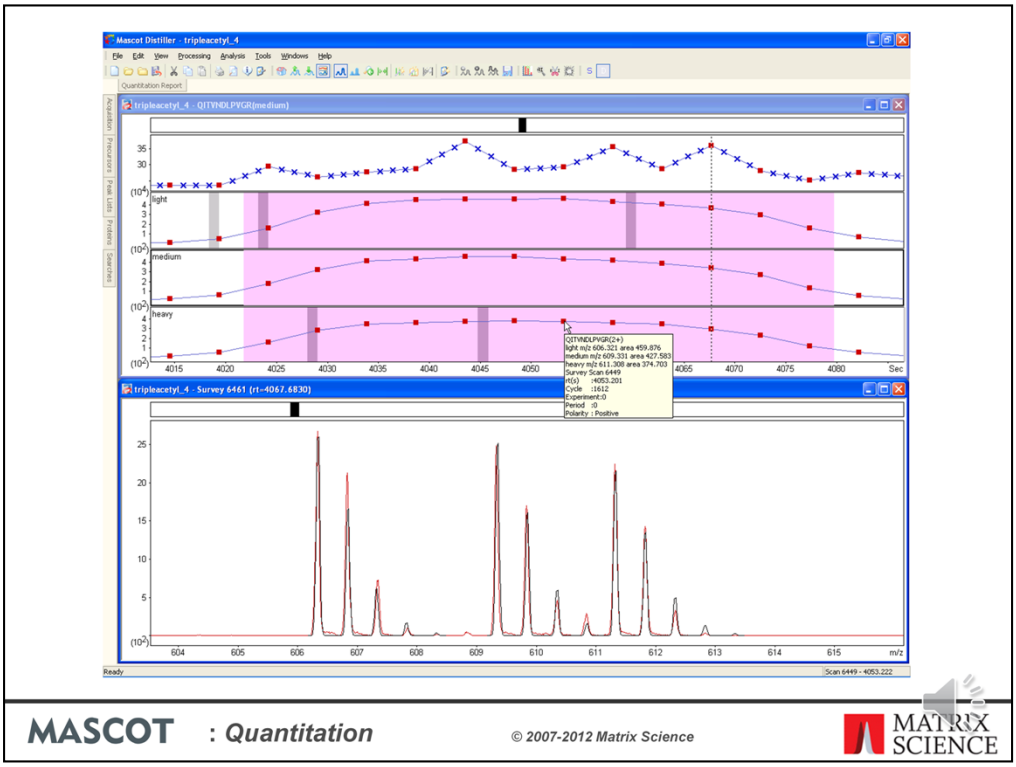
However big your screen, its always difficult to find enough room to display everything. To try and make best use of limited space, the grid and tree can be unpinned, so that they fly out when required and disappear when you move the mouse away.



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

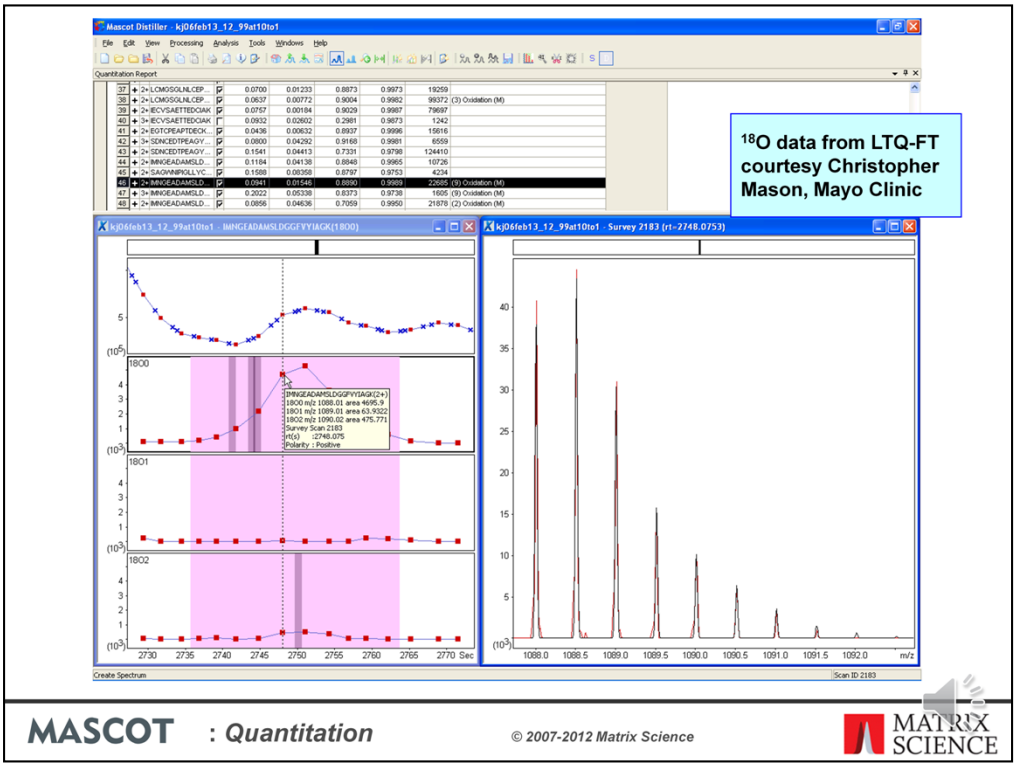


You can spend many happy hours devising different ways to arrange the Windows.

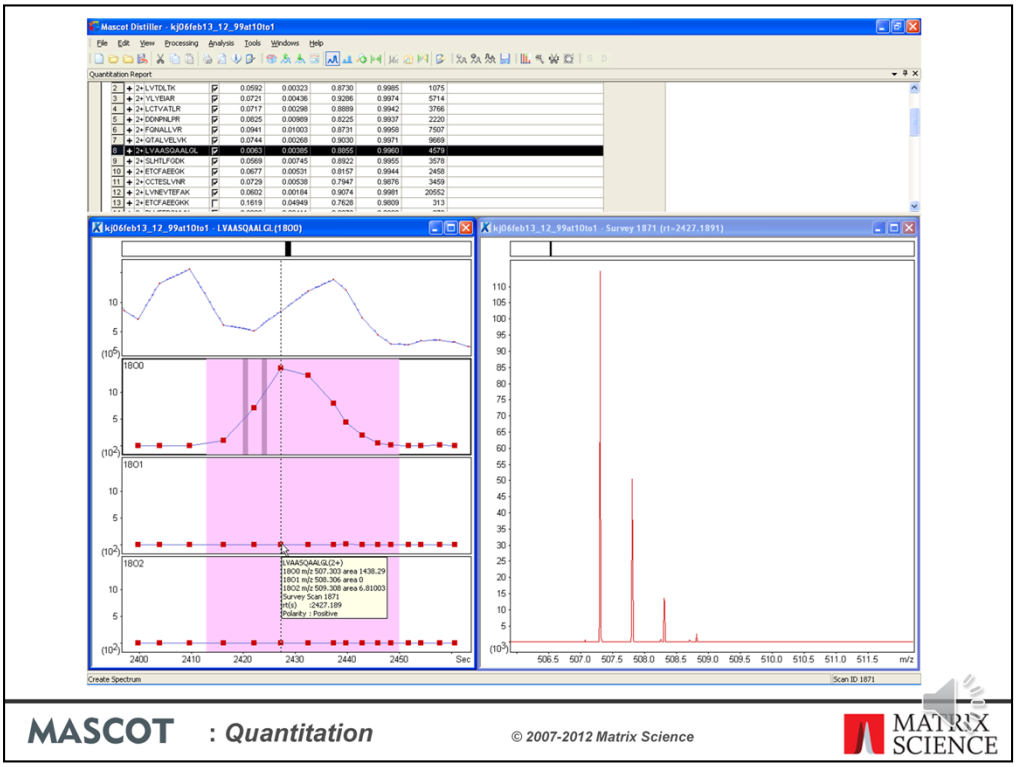
With the tree and quant grid unpinned, we can see the chromatogram and scan windows more clearly. A tooltip for each scan provides mass and charge information.

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.

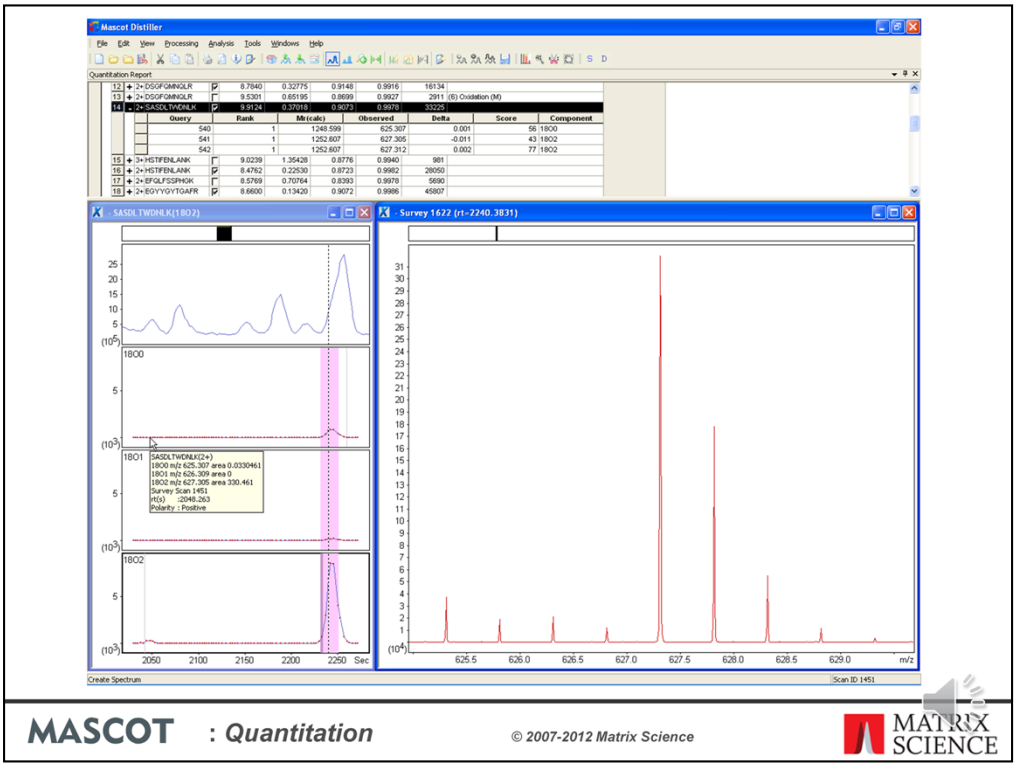
Whenever isotope distributions overlap, as happens here for the medium and heavy, this is accounted for in all calculations, as is incomplete enrichment. Obviously, in an individual scan, the fit is unlikely to be perfect. However, I hope you'll agree that we can immediately judge that the ratio isn't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



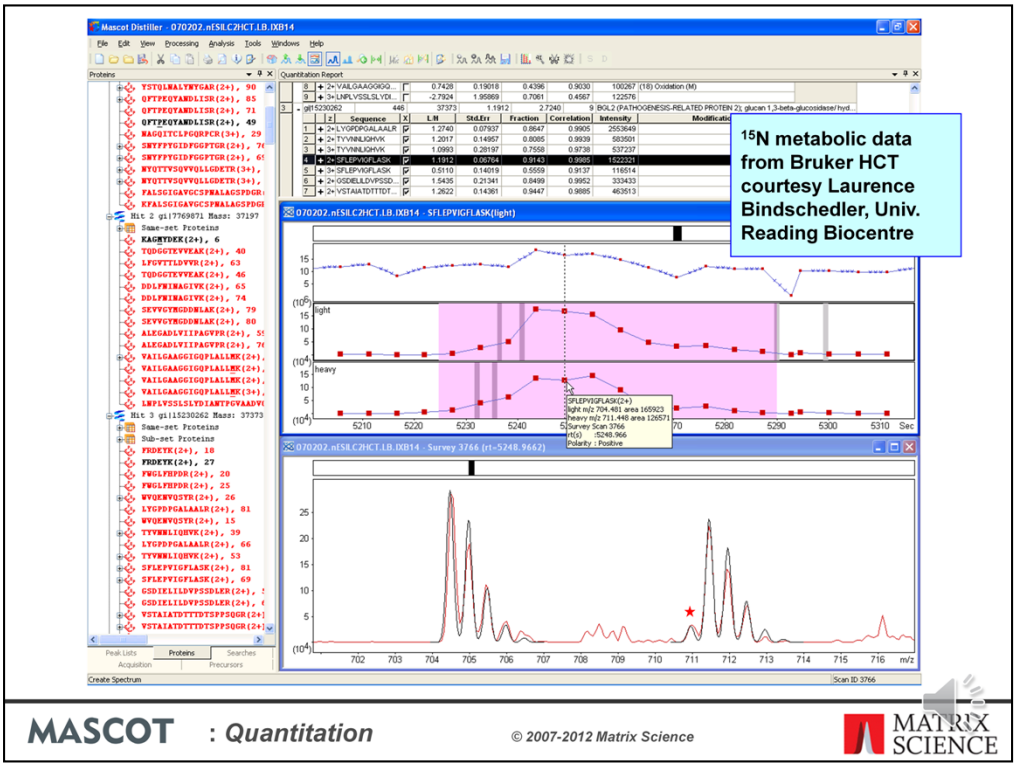
This is O-18 data from an LTQ-FT, courtesy Christopher Mason, Mayo Clinic. The sample is part of a dilution series and this one is particularly tricky because the ratio is 10:1, light to heavy. Obviously, with only a 4 Da separation, the isotope distribution for the heavy is smoothly overlapped by the 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.



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

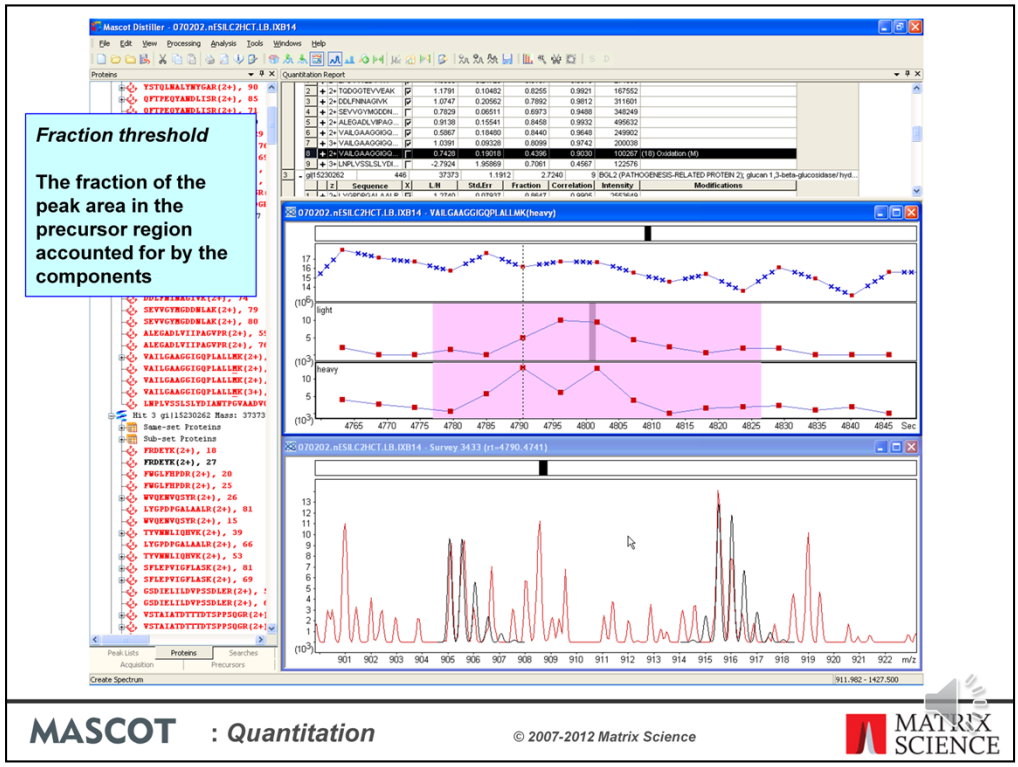


Sometimes, XIC peak detection is a challenge. The starting point is set of scans for which we have Mascot matches. You can hardly see, but in this case, there is a grey band way out here for a match obtained some 3 minutes before the main peak came through. The XIC peak detection algorithm has decided 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.

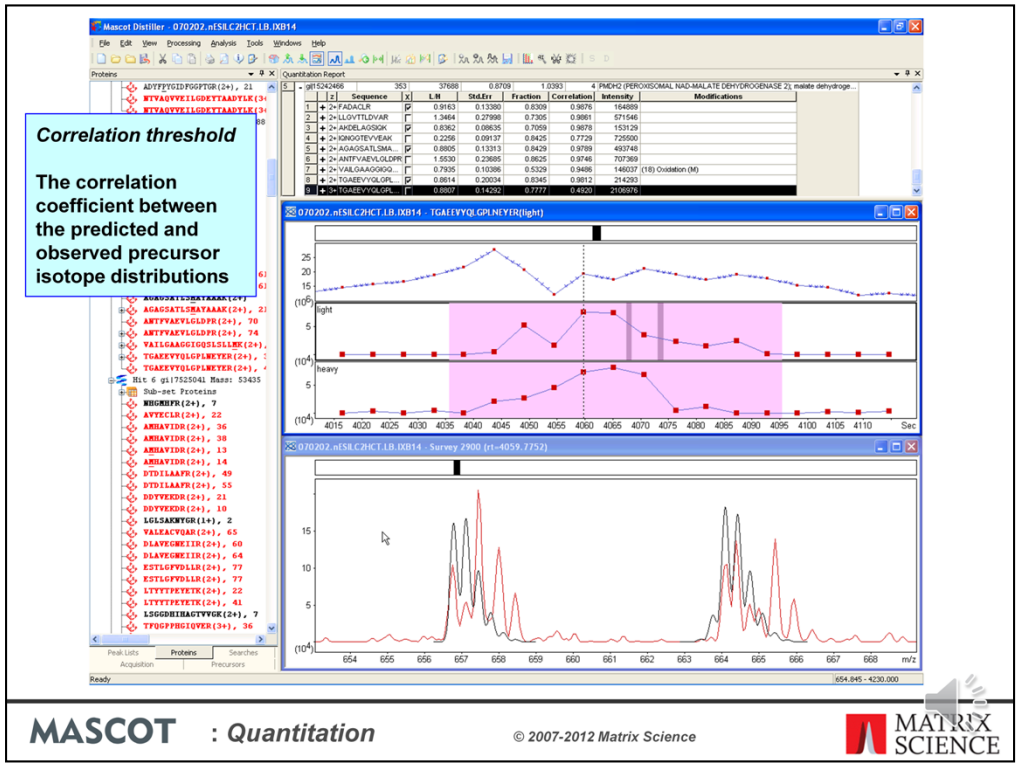


Now some metabolic data, courtesy of Rainer Cramer's group at the Reading Biocentre. In the heavy component, all the nitrogens in the proteins have been replaced by N-15. The calculated distributions (black) fit reasonably well to the observed peaks (red). Notice the peak marked with a red star. This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

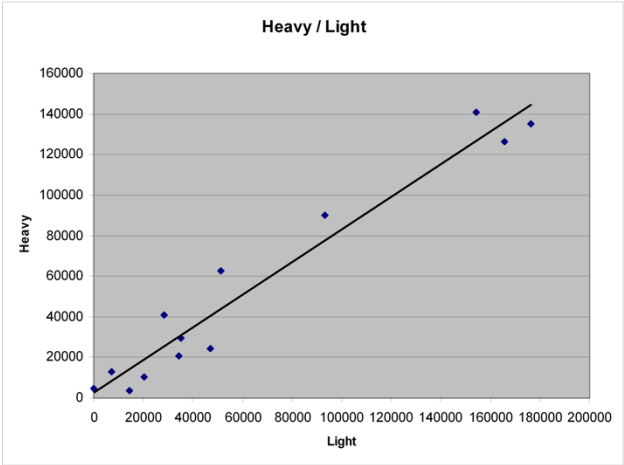


Fraction is the fraction of the peak area in the precursor region accounted for by the components. Here is a spectrum where there are a lot of interfering peaks. When we add up the areas, the expected precursors only account for 44% of the area, so the ratio is rejected. This threshold, like the others,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



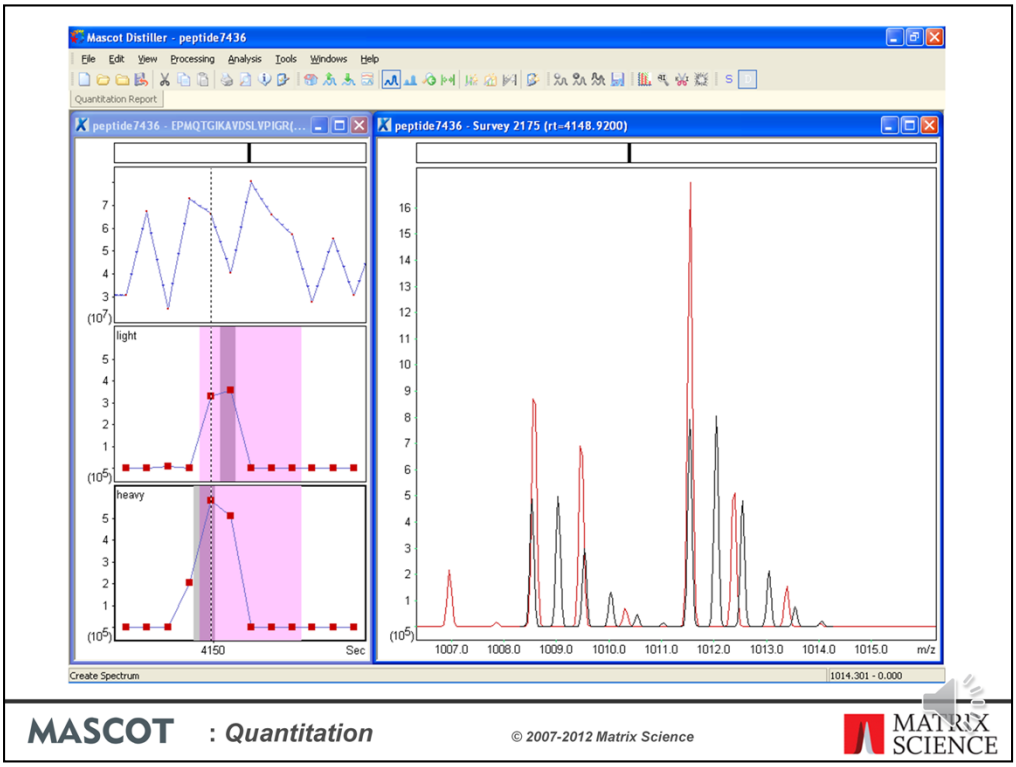
Each point represents the heavy and light intensities in a scan from the XIC peak. The straight line is a least squares fit

- 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

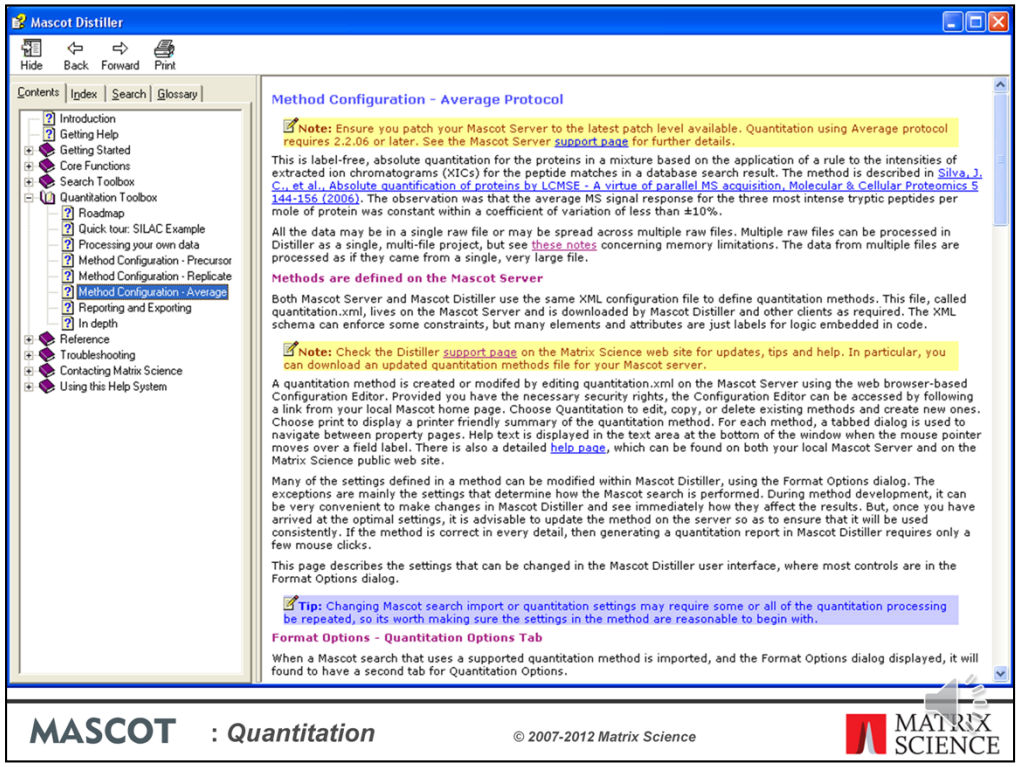
MASCOT : *Quantitation*

MATRIX SCIENCE

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.

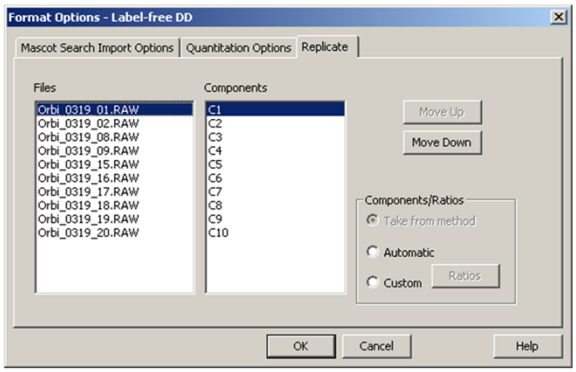


Of course, as always, garbage in means garbage out. Here is a case where the raw data are centroids, not profiles. Although we have good strong MS/MS, when you look at the survey scans, there are no proper isotope distributions. The monoisotopic peaks are approximately in the right place, but the spacings 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.



In the interests of time, we'll skip over the Average protocol, but use this opportunity to point out that Distiller comes with comprehensive help. This page gives an overview of how Average works and details of all the configuration settings. There are similar pages for precursor and replicate.

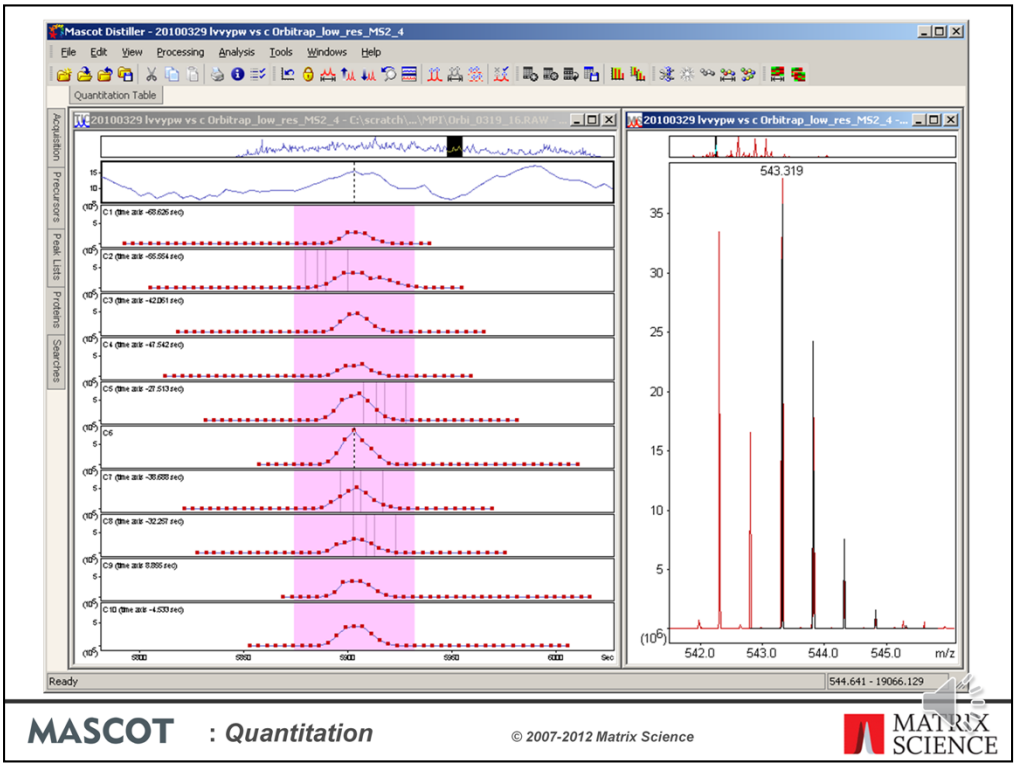
Replicate



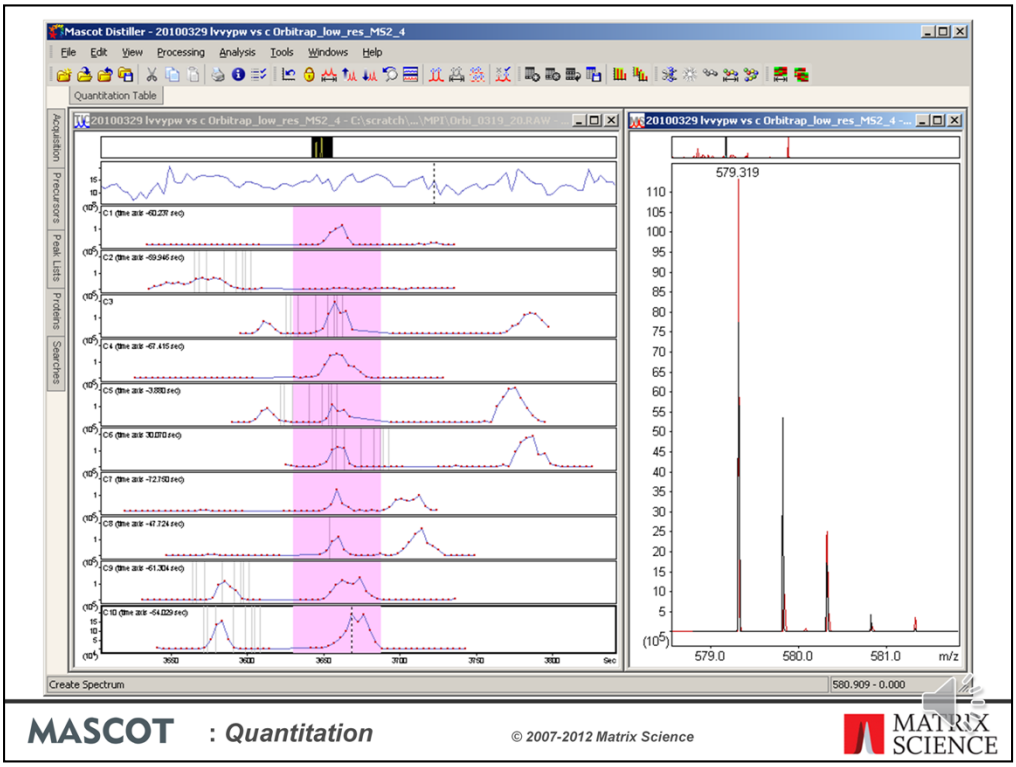
MASCOT : *Quantitation* MATRIX SCIENCE

Replicate, you may remember, is ‘label free’. Our implementation is identification driven, not feature driven. Distiller starts from the MS/MS data, imports the peptide matches from a Mascot search, then looks in the MS data for the precursors. As with precursor, you don’t need MS/MS in every file. You could have one file containing MS/MS and 9 files containing just MS and all identifications would be based on the first file.

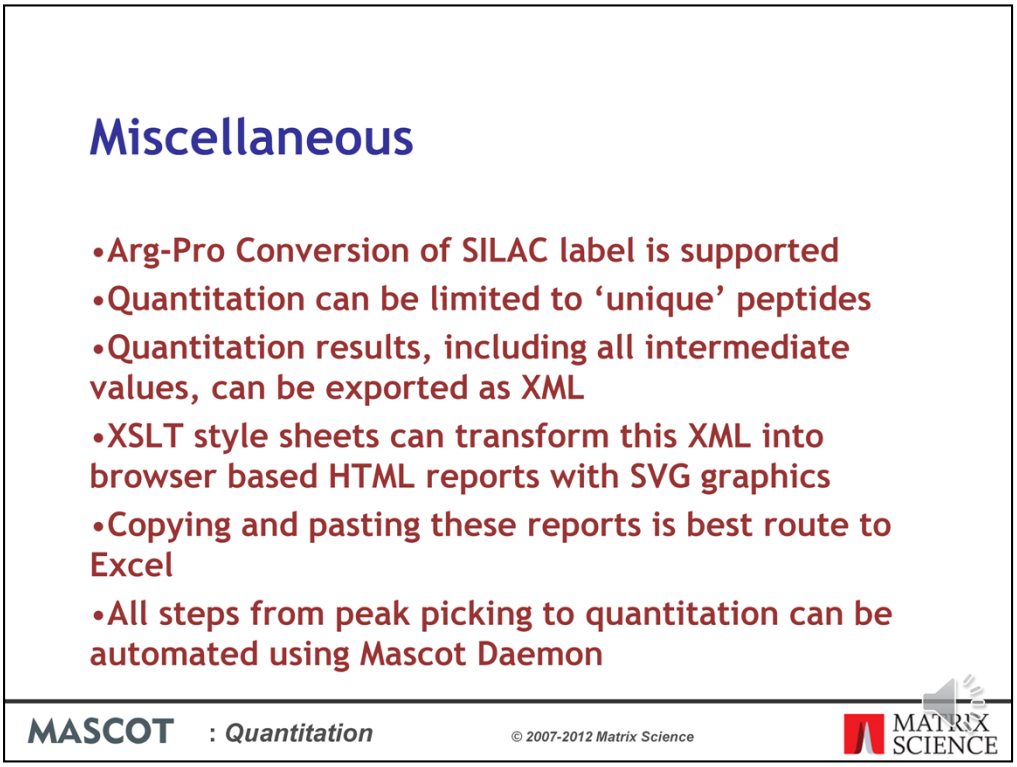
A replicate project is always a multi-file project, with one file for each component. You can define ratios or have them auto-generated, with one file selected as reference and all the others reported relative to it.



The tricky part of label free is time alignment of the chromatograms. Distiller does its best, and here is an example where things are working, even though the precursor region is heavily overlapped



But, here it fails. Distiller can manage small misalignments OK, but not severe misalignments or multiple XIC peaks. If alignment fails for a particular peptide, then the ratio is meaningless.



Some final points

Selected Literature

- Ross, P. L., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents*, Molecular & Cellular Proteomics 3 1154-1169 (2004) – iTRAQ
- Zhang, G. A. and Neubert, T. A., *Automated comparative proteomics based on multiplex tandem mass spectrometry and stable isotope labeling*, Molecular & Cellular Proteomics 5 401-411 (2006) - Multiplex
- Beynon, R. J. and Pratt, J. M., *Metabolic labeling of proteins for proteomics*, Molecular & Cellular Proteomics 4 857-872 (2005) - Metabolic
- Ong, S. E. and Mann, M., *Mass spectrometry-based proteomics turns quantitative*, Nature Chemical Biology 1 252-262 (2005) – General review
- Lill, J., *Proteomic tools for quantitation by mass spectrometry*, Mass Spectrometry Reviews 22 182-194 (2003) – General review
- Julka, S. and Regnier, F., *Quantification in proteomics through stable isotope coding: A review*, Journal of Proteome Research 3 350-363 (2004) – General review
- Bantscheff, M., et al., Quantitative mass spectrometry in proteomics: a critical review, Analytical and Bioanalytical Chemistry 389 1017-1031 (2007) – General review