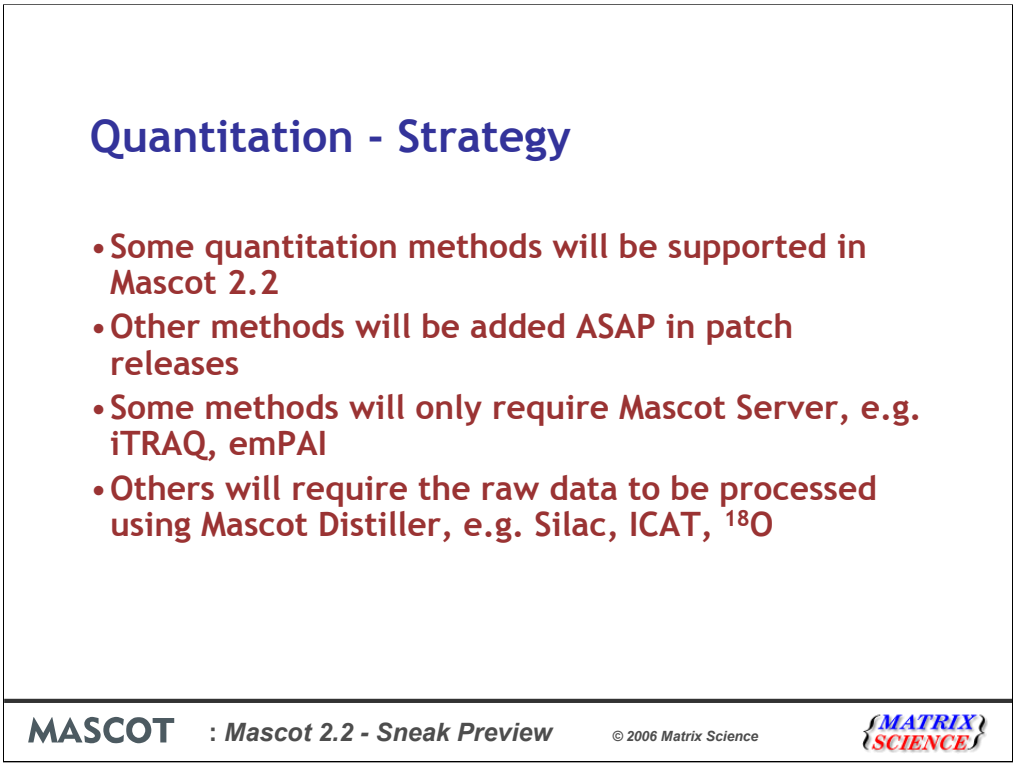
Sneak preview of new features in Mascot 2.2

MASCOT

MATRIX SCIENCE



The major new piece of work in Mascot 2.2 is quantitation. It seems like there is a strong need for more systematic and reliable quantitation software. Many people are working on developing chemistries and protocols, but the software to report the numbers seems to be a little neglected.

Since identification is an essential aspect of most MS based quantitation, it seems natural to try and integrate this functionality more closely into Mascot. So, quantitation will be one of the main components in the next release of Mascot. There is still more work to be done, but I wanted to give you a preview of what's in the pipeline.

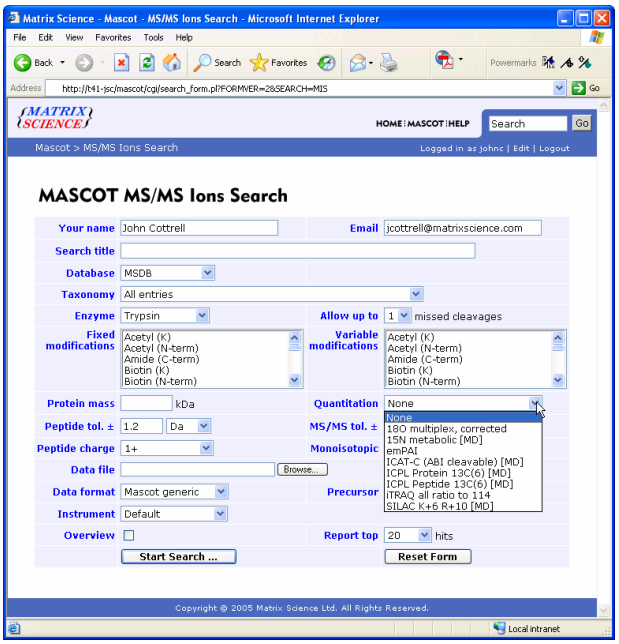
We intend to support as many of the popular methods as possible. Only some of these will be finished in time for the initial release of Mascot 2.2. Other will be added as soon as possible in patch releases.

For some methods, like iTRAQ and emPAI, all of the information required for quantitation is contained in the peak list.

Most 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 will require that the raw data files are processed using Mascot Distiller, and the Mascot search results will be imported into Distiller to generate a quantitation report.

Quantitation

Simplicity for user



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 new control, which replaces the old ICAT checkbox.

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

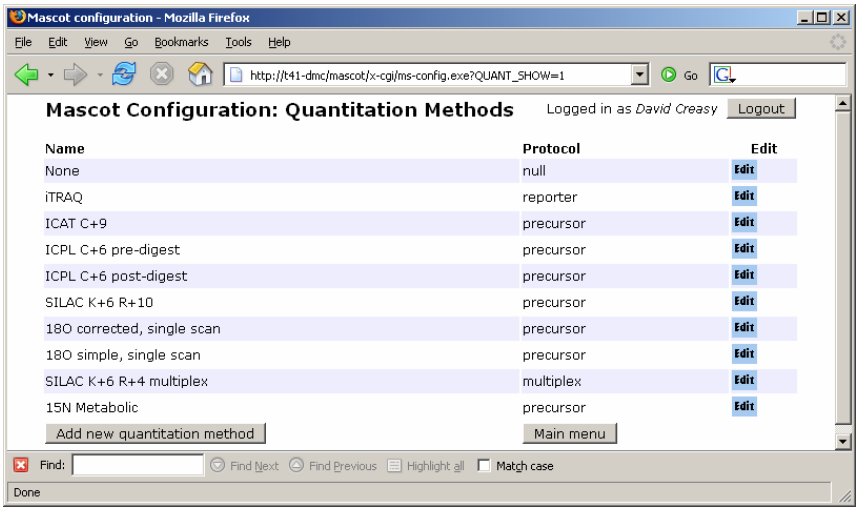
Quantitation

The quantitation methods are defined in a single XML configuration file

- quantitation.xml
- Browser based editor
- Add new methods as required
- Used by Mascot Server and Mascot Distiller

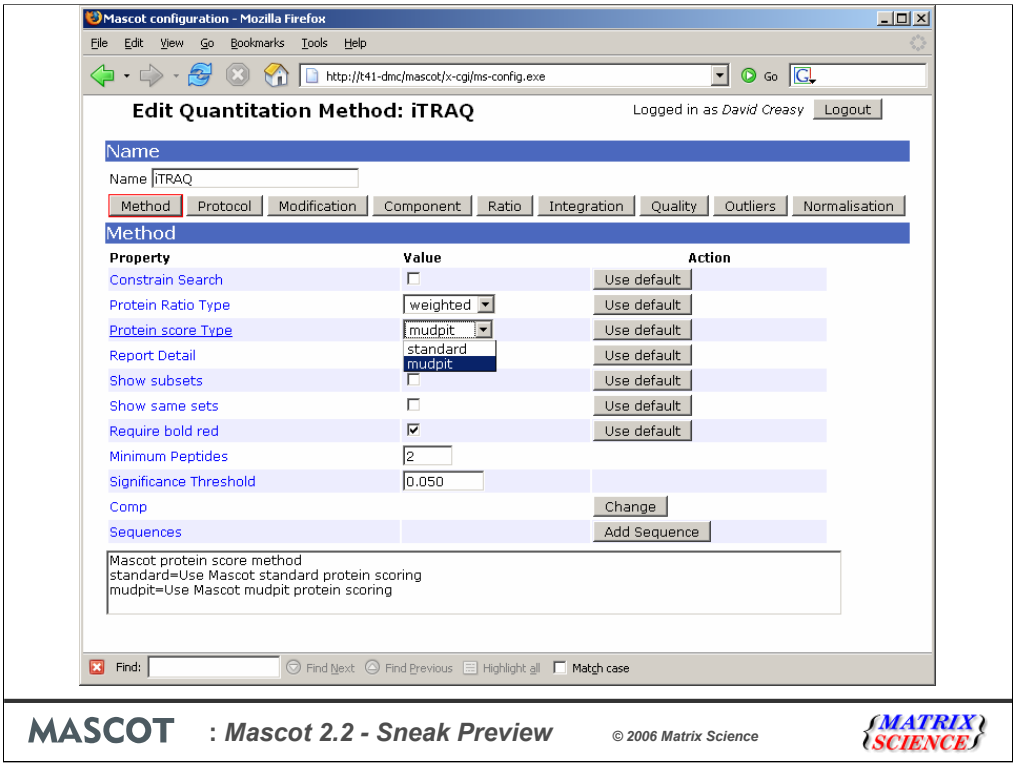
The configuration file that encapsulates the choices and parameters for each quantitation method is called quantitation.xml. This is an XML file, and there is a browser based editor for modifying methods and creating new ones. quantitation.xml lives on the Mascot server and is read by both the search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

Mascot 2.2 - Quantitation editor



MASCOT : *Mascot 2.2 - Sneak Preview* *© 2006 Matrix Science* MATRIX SCIENCE

When you first open the quantitation editor, there will be a list of all the currently defined modifications. You can also of course add a new one from here. I'll show an example with ITraq. Clicking on the Edit button brings up this page:

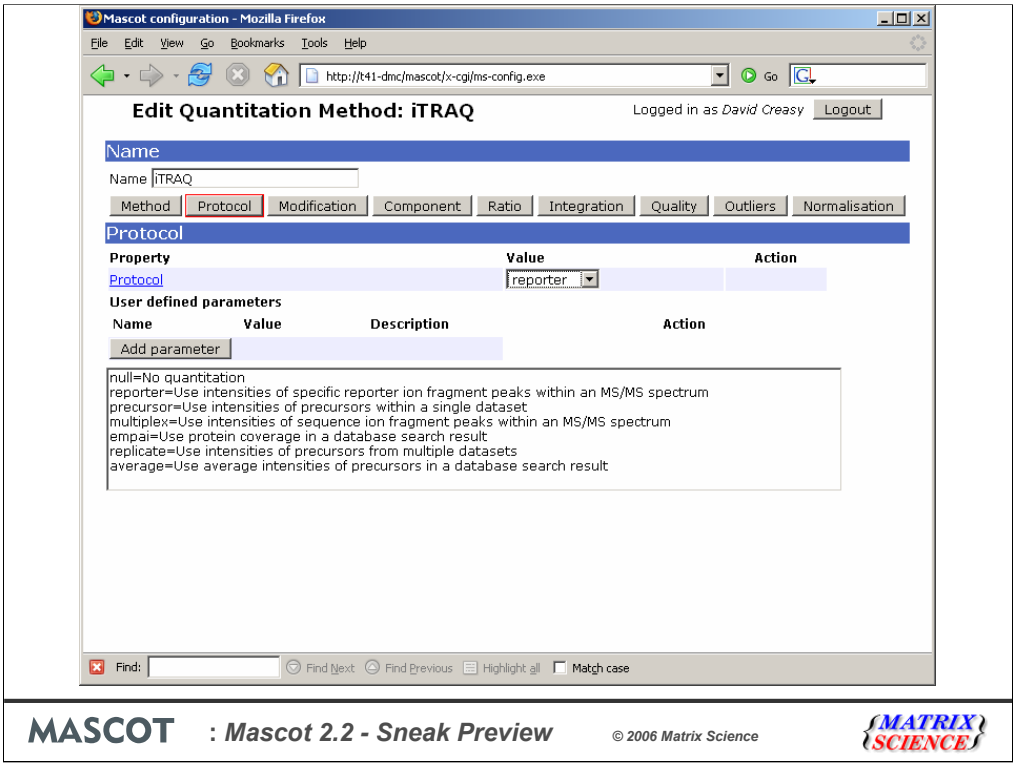


You'll see that there are a lot of different parameters, split into groups. So, there are settings for the method, protocol, modifications, etc. Some techniques won't have entries for some of these settings.

As with the database maintenance and security administration tool, holding the mouse over any of these items gives further help. So, we can see, for example that the help for constrain search says that "If any modification group specifies exclusive mode, then apply this constraint during the search so that only matches that can be used for quantitation will be returned." This isn't a sensible option for iTraq where the modifications would normally be specified as Fixed

The protein ration type is the method of calculating the ratio for the protein from the individual peptide ratios. You can specify whether to use the median, the average or a wighted average.

For the protein scoring, either standard or MudPIT scoring can be chosen.

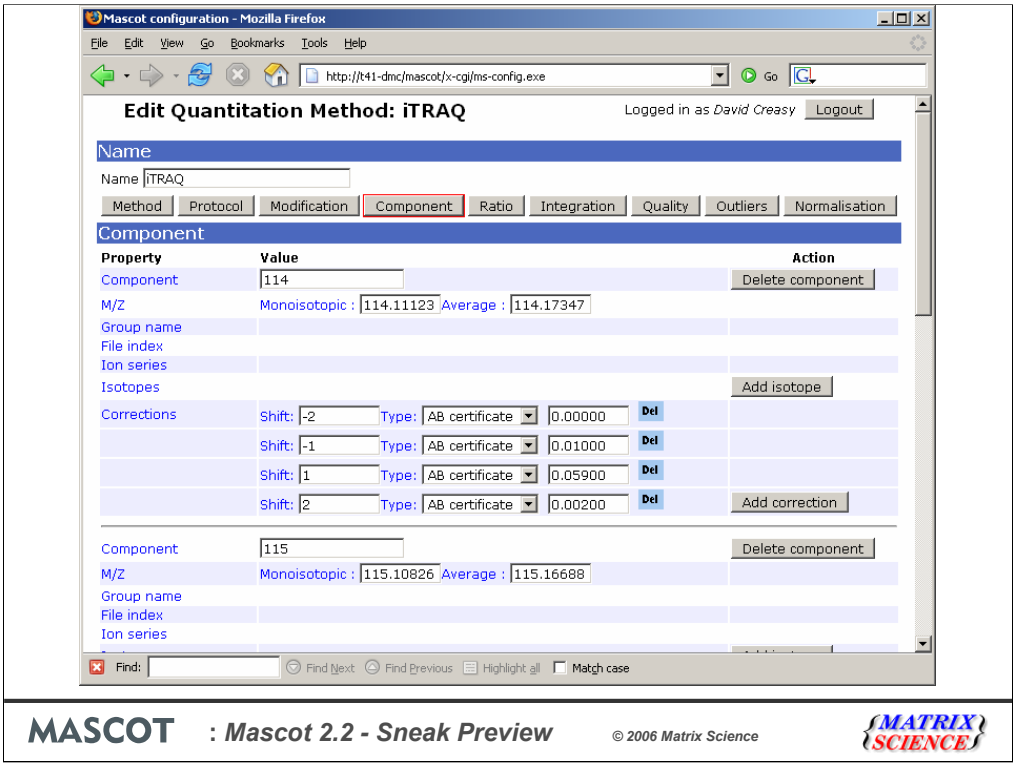


The protocol selection is perhaps the most important. It's a simple drop down list with the following options:

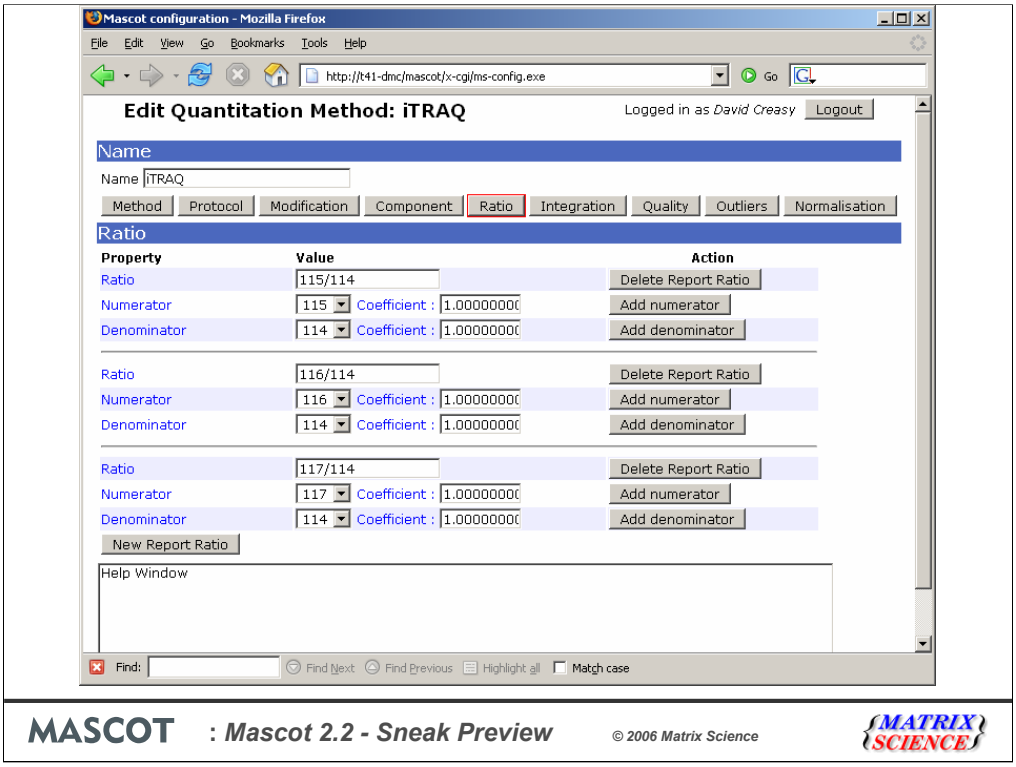
Quantitation - Protocols

Protocol	Description	Examples
null	No quantitation	
reporter	Specific reporter ion peaks within a single MS/MS spectrum	iTRAQ
precursor	Extracted ion chromatograms for related precursors within a single dataset	ICAT, SILAC, ^{18}O, ICPL, AQUA, Metabolic
Multiplex (Neubert et. al.)	Pairs of sequence ion fragment peaks within a single MS/MS spectrum	SILAC, ^{18}O
Empai (Ishihama et. al.)	Protein coverage in a database search result	Label-free
replicate	Extracted ion chromatograms for identical precursors across multiple datasets	Label-free
average	Extracted ion chromatograms for selected peptides per protein within a single dataset	Label-free

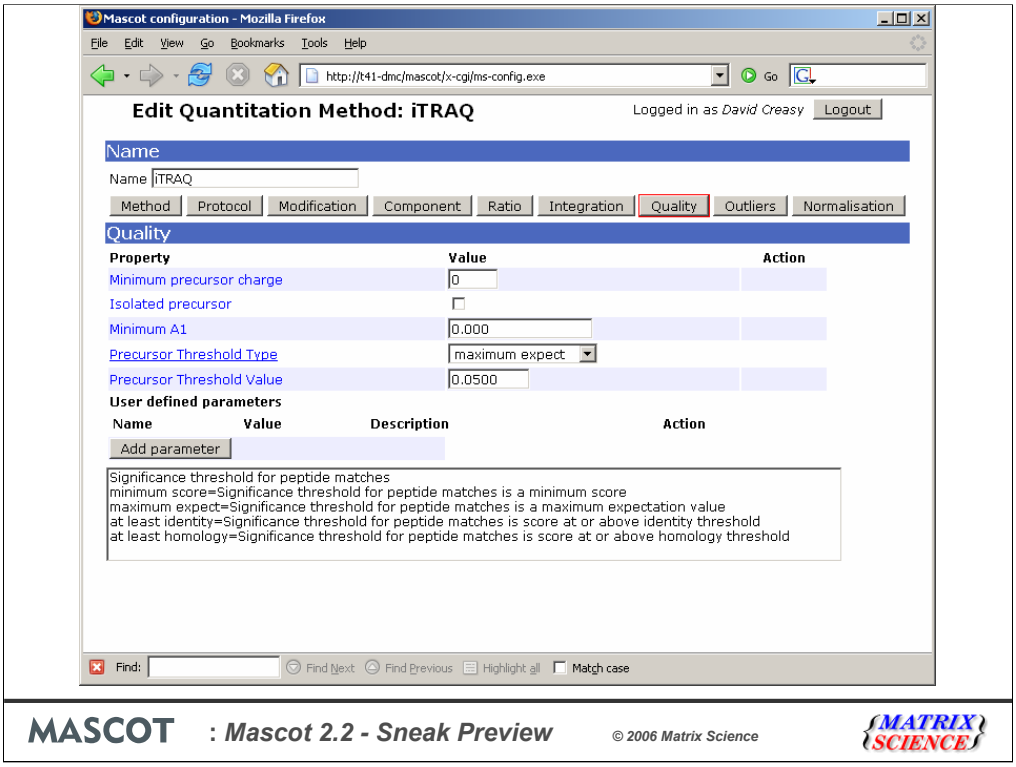
At this time, we have identified 6 protocols plus null, which means no quantitation.



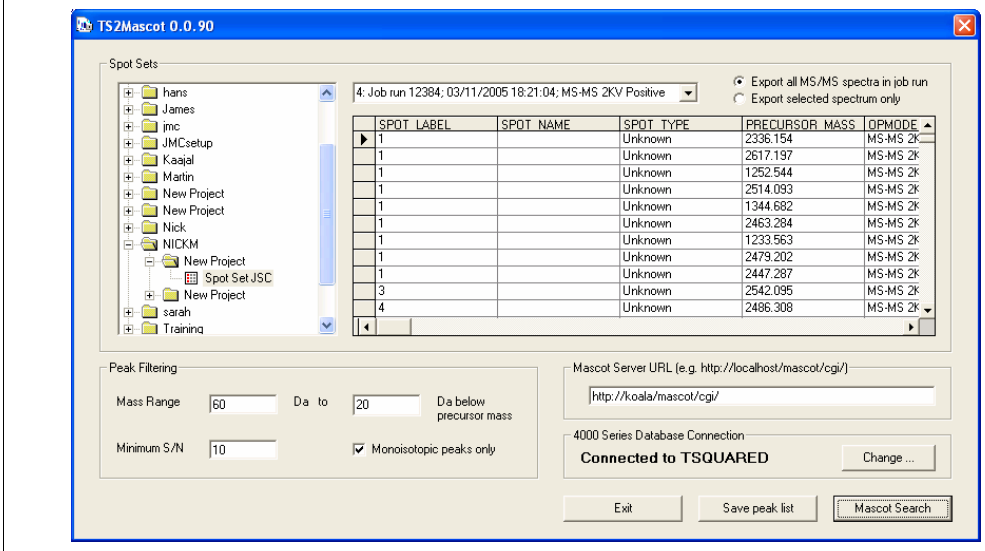
The contents of the component tab depend on the protocol. So, with ITraq for example, since we are using reporter ions, we have to specify the m/z value for each of the four reporter ions. Also, we have the option to enter some correction information that is supplied with the reagents.



And for the ratios, we are simply choosing to report the ratio of the 115, 116 and 117 peaks to the 114 peak. Obviously, it's easy for people to select other ratios if desired.



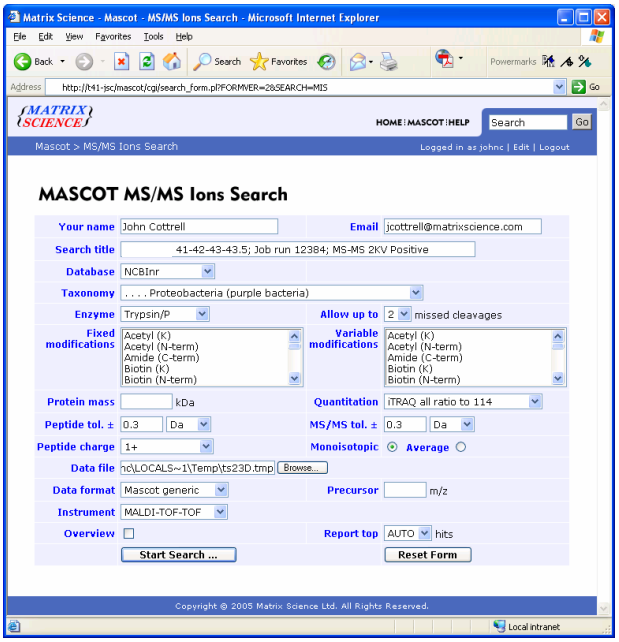
The quality page allows you to select which peptides will be used for calculating ratios - so it would be normal to disregard anything with a 1 in 20 chance of it being a random match.



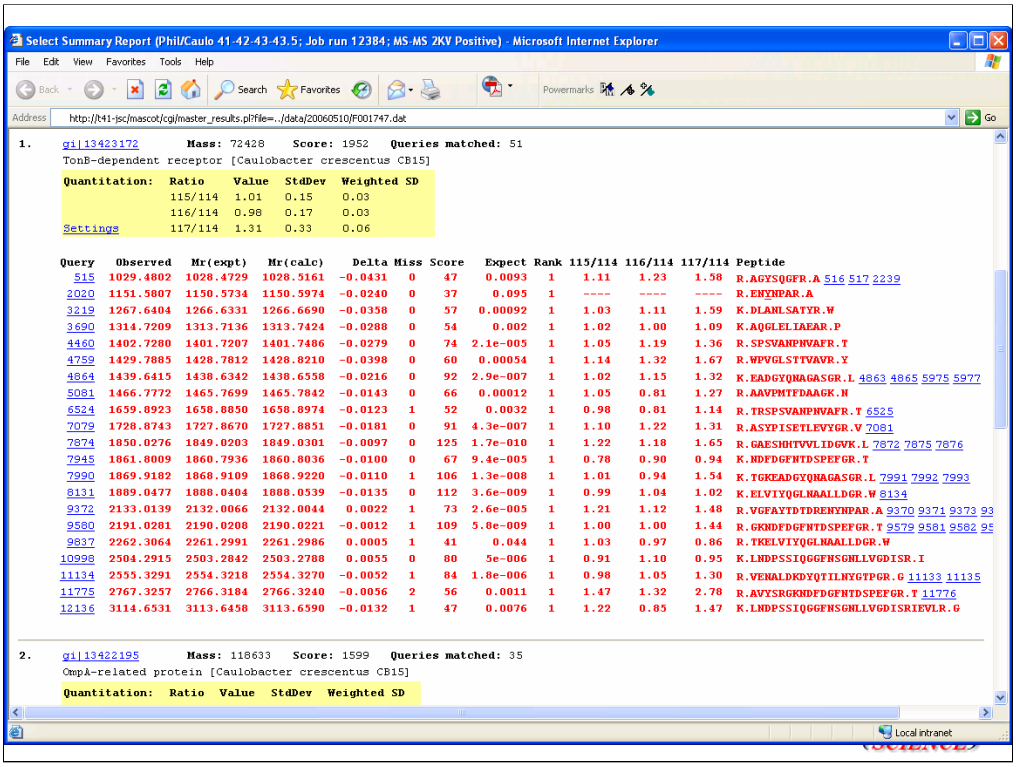
Even the simpler methods can become complicated. With iTRAQ, for example, we found that the peak list exported from the 4000 series data system or submitted to Mascot from GPS Explorer did not have the correct peak areas for the reporter ions. The numbers are different from those used within GPS Explorer for quantitation. We have had to write our own application to export a suitable peak list from the Oracle database.

So, for iTRAQ, you might choose Mascot Search …

Quantitation



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



And back comes the report. This is how the reports will look for methods that can be handled entirely within the search engine. For methods like Silac and ICAT, the quantitation report will be generated by Mascot Distiller

Quantitation - Definition example

```
<method name="ICPL Protein 13C(6) [MD]" constrain_search="true" protein_ratio_type="average"
report_detail="true" description="Label pre-digest, so peptide N-terms are not labelled">
        <modification_group mode="exclusive" name="light">
                <mod_file>ICPL_light (K)</mod_file>
                <mod_file>N-ICPL_light (Protein)</mod_file>
        </modification_group>
        <modification_group mode="exclusive" name="heavy">
                <mod_file>ICPL_heavy (K)</mod_file>
                <mod_file>N-ICPL_heavy (Protein)</mod_file>
        </modification_group>
        <component name="light">
                <group_name>light</group_name>
        </component>
        <component name="heavy">
                <group_name>heavy</group_name>
        </component>
        <report_ratio name="L/H">
                <numerator_component name="light"/>
                <denominator_component name="heavy"/>
        </report_ratio>
        <quality isolated_precursor="false"/>
        <integration method="trapezium" source="survey"/>
        <protocol>
                <precursor allow_mass_time_match="false"/>
        </protocol>
</method>
```


Here is an example of a quantitation method. This is for ICPL labelling of a protein prior to performing the digest.

Note that we are introducing a new way of treating modifications. Exclusive means that peptides must conform to one of these modification groups. You cannot have a peptide that has a mixture of the heavy and light labels.

It will also be possible to specify user defined modifications and isotope substitutions for metabolic labelling

The mechanics of quantitation, which is to say, where the numbers come from, is defined by the protocol. In this case, we use a protocol called precursor.

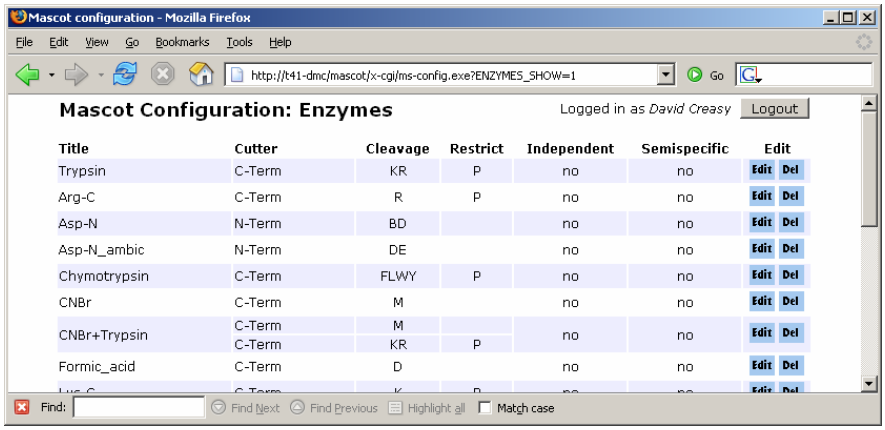
Configuration file editors

Configuration files on the Mascot server:

- enzymes
- fragmentation_rules
- quantitation
- taxonomy
- unimod (replaces mod_file and masses file)

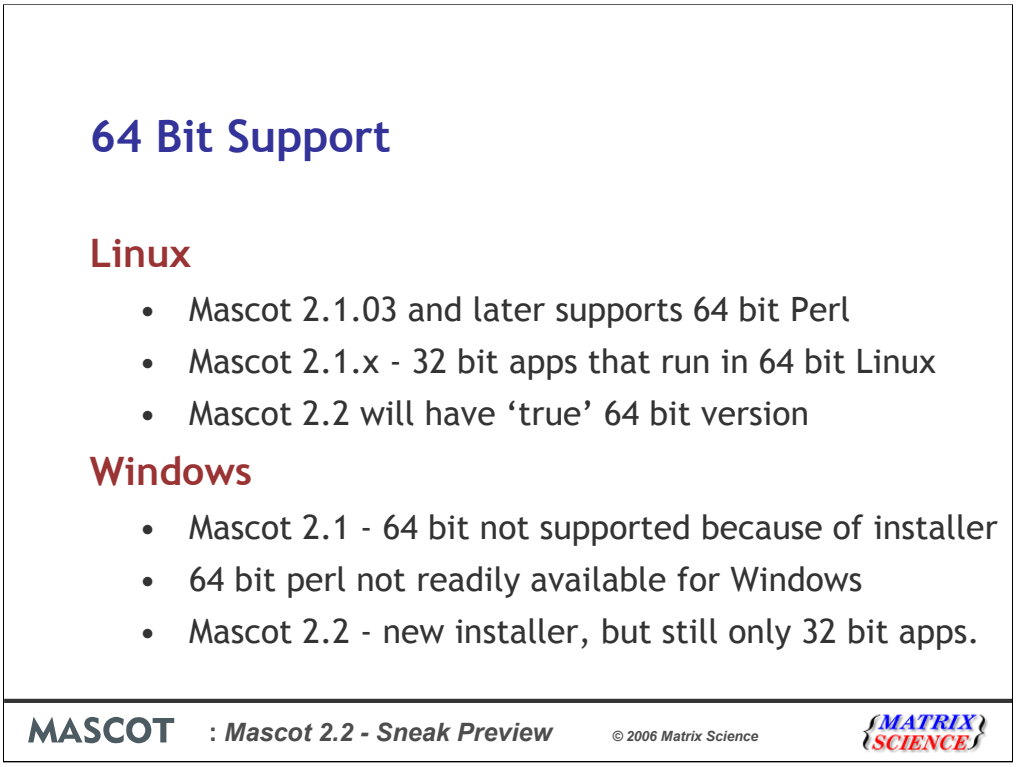
If you want to get as many identifications as possible, as efficiently as possible, you might come up with a strategy similar to this.

Enzyme file editor



MASCOT : *Mascot 2.2 - Sneak Preview* MATRIX SCIENCE

If you want to get as many identifications as possible, as efficiently as possible, you might come up with a strategy similar to this.



In Mascot 2.1.03, we added better support for 64 bit versions of Linux by adding a 64 bit version of Mascot Parser. Most 64 bit versions of Linux come with a pre-installed 64 bit perl, so it is quite difficult to install 32 bit perl on some Linux distributions. However, the search engine and all the other applications are still 32 bit which isn't really an issue. Mascot 2.2 will have the option to install 64 bit applications, and this may give a slight performance increase - possibly as much as 10%.

We currently don't support 64 bit Windows. We have confirmed that the current version of Mascot will run on 64 bit Windows XP, but it is very hard to install because we are still using an old installer.

The other issue is that we use ActiveState Perl, and they don't have a 64 bit version of perl available yet.

Mascot 2.2 will have a new installer, and 64 bit Windows will be supported, but we will only have 32 bit versions of the Mascot Parser and the applications. When a 64 bit version of ActiveState Perl is available, we will review this -(I just checked again yesterday and see that they have just released a beta version.)

Option to search randomised entries

- Check box on the search form
- Sequences based on average AA or base composition for the whole database
- For each sequence, create a randomised sequence of same length
- Option to view results from random sequences
- Summary stats:

	NCBInr	Decoy	False Positive rate
Peptide matches above identity threshold	4970	25	0.50%
Peptide matches above homology threshold	6600	330	5.00%

MASCOT : Mascot 2.2 - Sneak Preview MATRIX SCIENCE

We are delighted that the proteomics community has recently been taking an active interest in validating search results by searching a data set against a randomised or reversed database using all the same parameters. Since 1999 we have had the facility on our public web site to search a randomised database. We've also recently put up a help page on our public web site that discusses some of the issues, and has a link to a free script to create a randomised or reversed database. In Mascot 2.2, we are going to make this even easier.

There will be a checkbox on search form. If you tick this, the search will take twice as long.

We will take average AA or base composition for database as a whole from stats file. The new sequences that are generated will all have this average composition.

When searching each database entry, independently create and search a randomised sequence of the same length and of this average composition.

The results from these searches will be accumulated separately in two new sections in the result file. This means that it will be easy for us to add a link on the main results page to view the results from searching the randomised database.

The standard master results report will include summary stats as additional block in header, as shown here:

And lots more . . .

PMF searches now use intensity information

Module for top down searches from FT

Support for ETD/CID searches etc:

```
BEGIN IONS
INSTRUMENT=ETD
. . .
END IONS
BEGIN IONS
INSTRUMENT=CID
. . .
```

MASCOT : *Mascot 2.2 - Sneak Preview* MATRIX SCIENCE

If you want to get as many identifications as possible, as efficiently as possible, you might come up with a strategy similar to this.