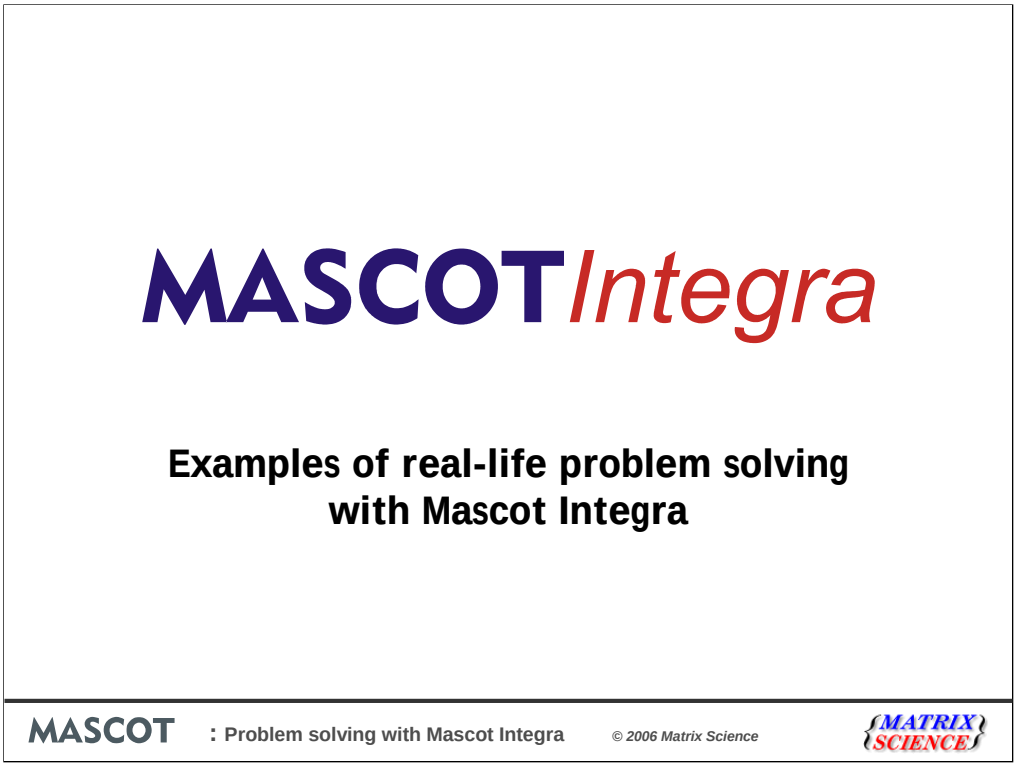
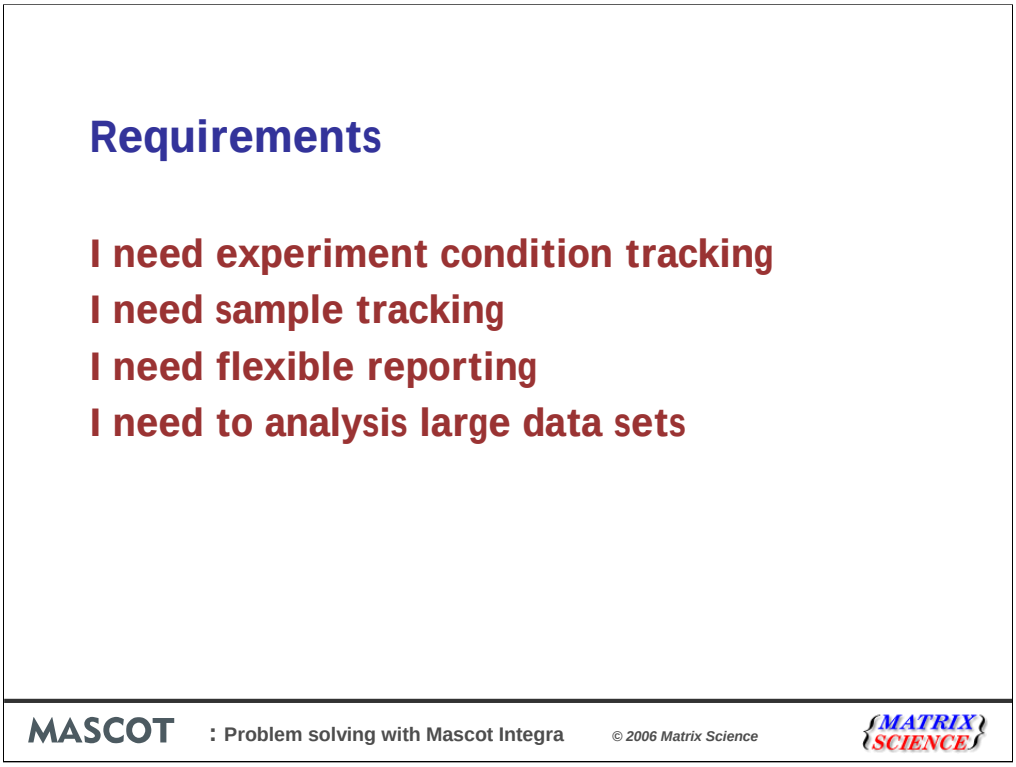


We have introduced Mascot Integra at previous ASMS user meetings so I thought that we would cover something different this time.

For those that are not familiar with Mascot Integra it is our solution to proteomics data management and you can find out more about it on our website or see it in action at our booth, number 39, during the conference.

This time I'm going to talk about solving real life problems with Mascot Integra.



What sort of requirements are we going to ask of Mascot Integra?

Firstly I need to track experiment conditions.

Secondly I want to know what happened to my samples three or more months later

Thirdly I want to create reports with the information that I'm interested in.

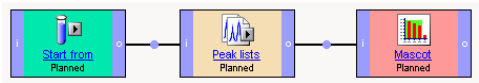
Finally I want to analysis large data sets and filter the identified proteins by different constraints

1. I need experiment condition tracking

Load the results in to the database as fast as possible

Track as many details as possible

An experiment can be as simple as this:



The need for experiment condition tracking is split in to two camps:

Those that want to load the results into the database as fast as possible and go straight to analysis and reporting

And those that would like to track as many details as possible about the experiment from within the database

We can support both camps and anywhere in-between.

I going to show some examples of minimal and detailed experiments.

First the minimal experiment.

Experiments are modelled by linking together experimental tasks selected from the library

It can be a simple as this:

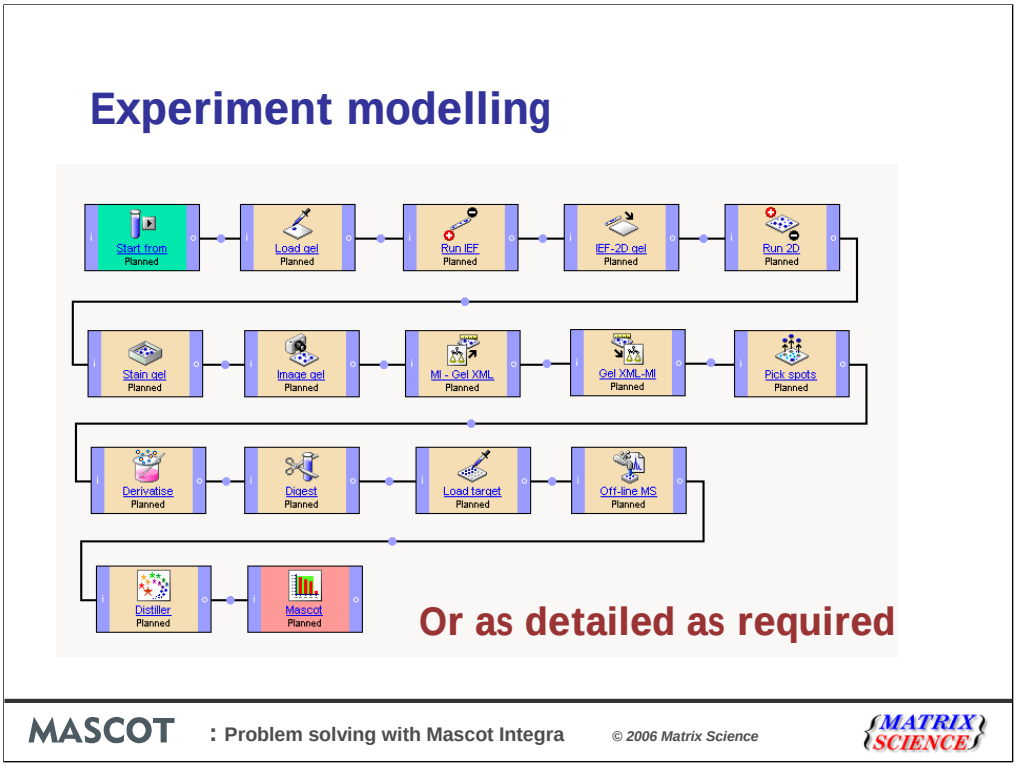
Start from a sample

Import existing MS or MS/MS peaklists

And perform a Mascot search.

In this experiment we don't capture any additional information, parameters, or conditions about the experiment.

But it allows us to search one or multiple data files quickly and capture the results, and link the search results back to the original sample.



This 2D gel protocol is an example of a more detailed experiment model.

I'm using many more different tasks to model this experiment.

Starting from our samples

we load and run the gels

Followed by staining and imaging

A link to the gel images is attached to the image gel task

Integra can interface with NonLinear Dynamics Progenesis and GE Healthcare DeCyder via xml files.

In this case Integra exports an xml file with information about the gels and their images that is imported into Progenesis

The gel images are then processed with Progenesis as per normal.

Once the gel spots have been selected for analysis we export an xml file from Progenesis.

This xml file is then imported back into Integra

and the spot list is processed generating a new gel spot sample for each selected spot.

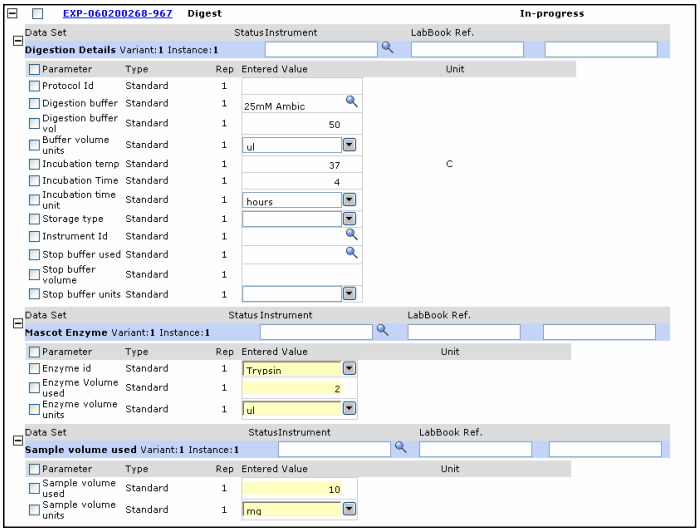
Then we continue through the experiment reducing and alkylating the proteins before in-gel digestion with trypsin

The samples are then analysed by mass spectrometry

The raw data are processed by Mascot Distiller and Mascot Daemon submits the searches to the Mascot Server.

So, in this example, we will have captured all the experimental parameters and conditions from the processing of the sample to the MS analysis and data processing.

Experimental data capture



Most tasks have a data entry page. Here is a view of the data entry page for an enzyme digest task.

The yellow boxes indicate values that are required to perform the task.

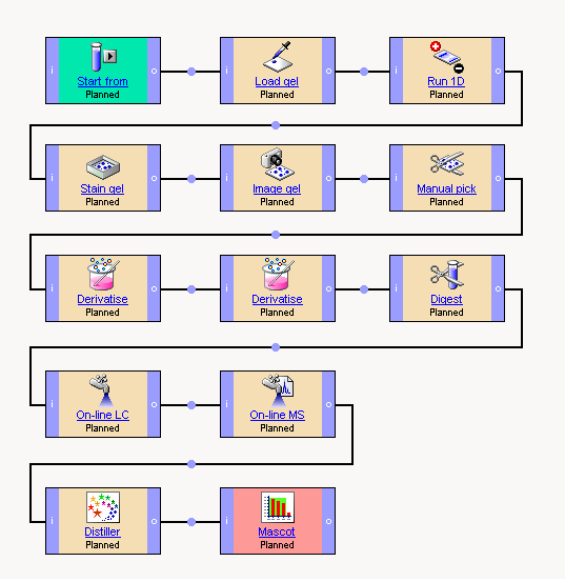
The other boxes are optional, but the more data that is entered, the more scope there will be for data mining of the experimental results.

Single tasks through to complete experiment plans can be preconfigured and saved as templates.

So, for example, if you perform all your digestions at 37C for 4 hours then these values can be predefined so that you don't have to enter them each time you run the task.

Some values are selected from a controlled, user defined list. For this task, it's the digestion and stop buffers and digestion robot or instrument.

Experiment modelling



Here is a third example of a detailed model of a protein mixture separated on a 1D gel.

The sample loaded and separated on a 1D gel

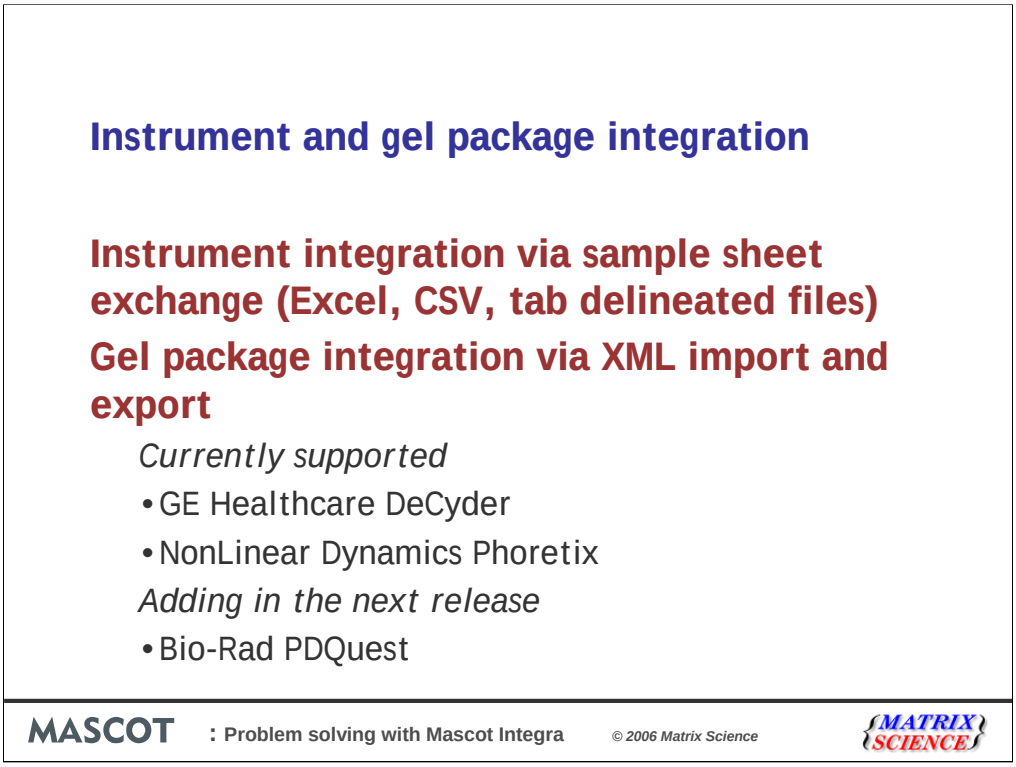
The gel is stained and the protein bands are excised

Followed by in gel digestion

LCMSMS

Peak picking and mascot database searching.

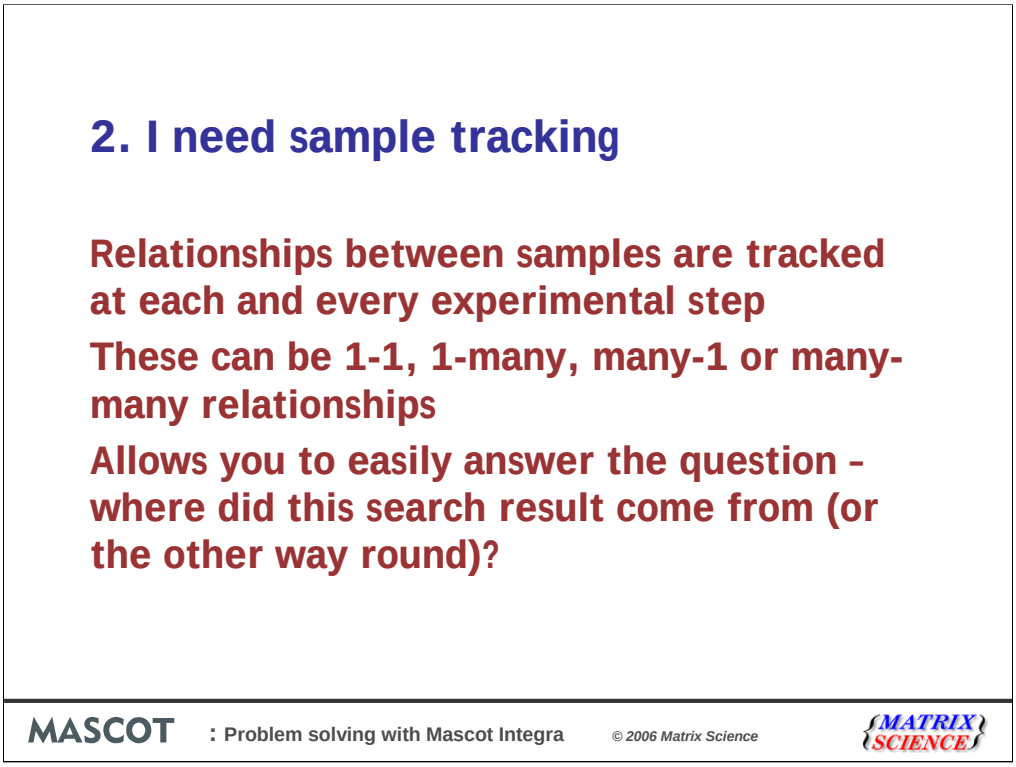
Again we have captured all the experimental parameters and conditions from the processing of the sample to the MS analysis and data processing



Rather than trying to run the instrumentation directly from Mascot Integra we have implemented a samplesheet exchange system. During the relevant step of running an experiment, you produce an Excel or CSV samplesheet from within the Integra system which you can then load into the Instrument datasystem. Nearly all instrument datasystems will import one or more of the supported formats (Excel, CSV and tab delineated).

This system is highly flexible. You can design your own samplesheets for use within the system. So if Thermo add a new column to the Xcalibur samplesheet then you can add this yourself into the system – you don't have to wait for us to add the column. Likewise if you bought a spotting robot which we had not designed a samplesheet for then you could simply add a new design into the system for it.

We also integrate with a number of Gel Packages. This is carried out using XML exchange between Mascot Integra and the Gel package software. We currently support the GE Healthcare DeCyder software and the NonLinear Dynamics family of 2D Gel analysis software and will be adding support for the Bio-Rad PDQuest package in the next release.

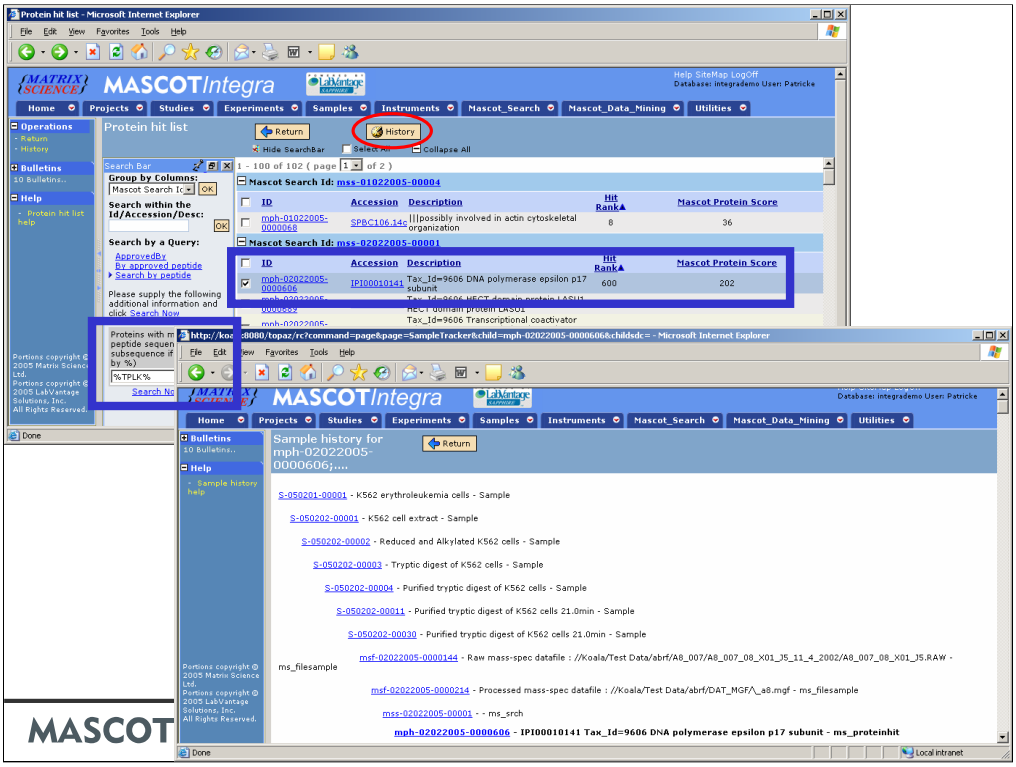


Now on to the second requirement of sample tracking.

One of the key advantages of tracking your experiments and results through Mascot Integra is that at each step of the experiment the relationships between samples are tracked and stored, readily accessible, in the database.

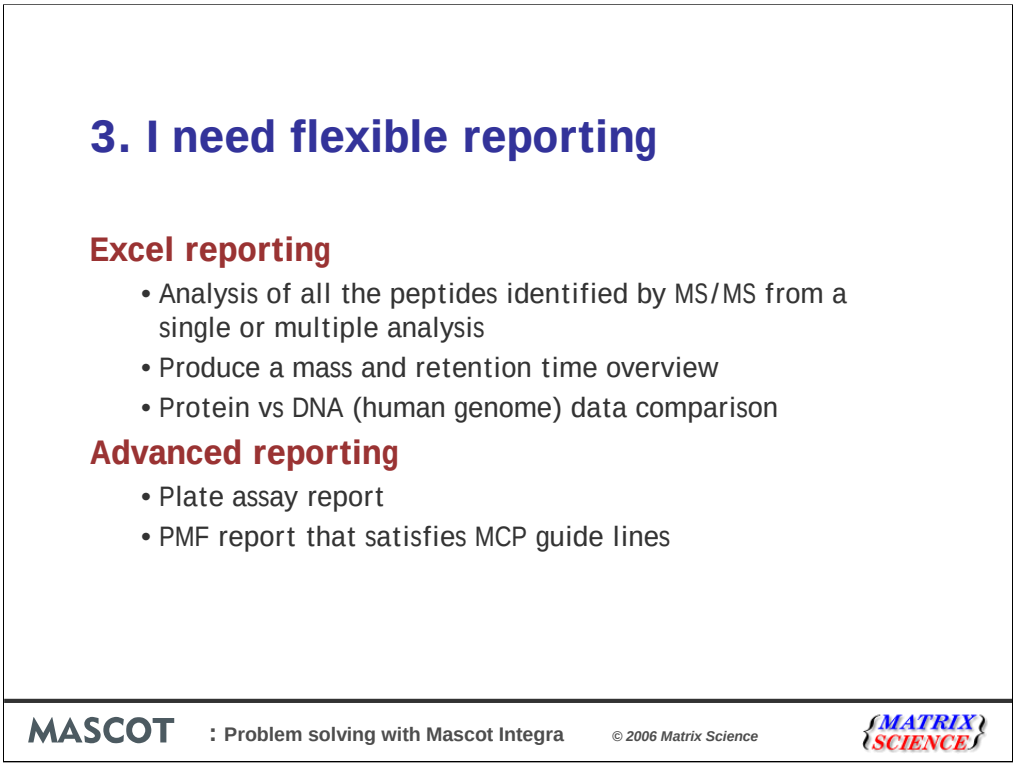
These relationships can be 1-1 (i.e. doing a simple digest to produce a child sample), 1-many (i.e. splitting a sample into multiple aliquots), many-1 (i.e. creating a single mixture from multiple starting samples) or many-many (i.e. creating multiple mixtures from multiple starting samples).

Because we have tracked all of these relationships, then answering the question 'where did this search result come from' 12 months down the line is now trivial.



Here I've done a query against our Mascot Integra system and retrieved a list of all Mascot protein hits which have a peptide match that contains the subsequence TPLK. Then to see were any one of those protein hits originates from all I have to do is select the protein hit and click on the History button.

We can see that all the intermediate steps from the protein hit back to the original cell sample.



Once that I have samples and data in the system I would like to perform some analysis and create some reports.

Mascot Integra uses Excel for custom reporting.

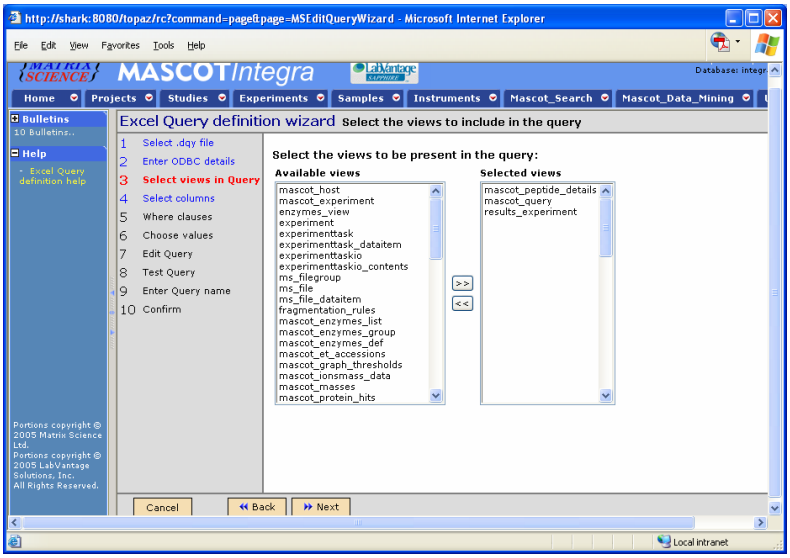
For more complex data mining, involving multiple SQL statements and further data processing, the Integra database can be mined using custom scripts or programs.

Pretty much any programming language can be used, for example Perl, Java, Visual Basic or C++. You could even use Visual Basic macros in an Excel sheet.

I'm going to show how an excel report is designed and three examples from the currently available report set.

Then I'm going to show some examples of what you might want to do with more advanced reporting.

Excel query
- choose the views



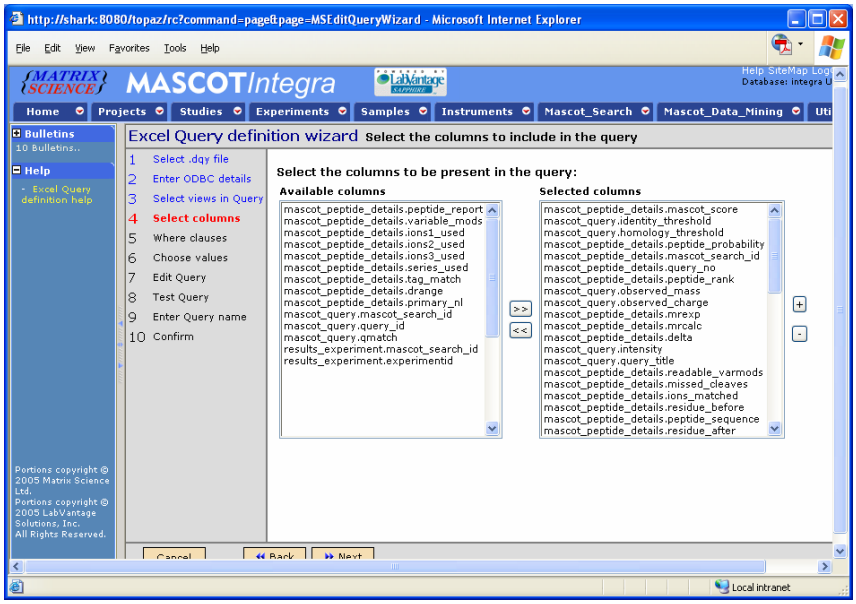
The first excel report is on the properties of the peptides identified from a MuDPIT run.

All the available views are listed in the left panel

from those I select the views related to peptides identified in a search.

The relationships between the different views and the column names for each view are documented in the online help.

Excel query - select the columns



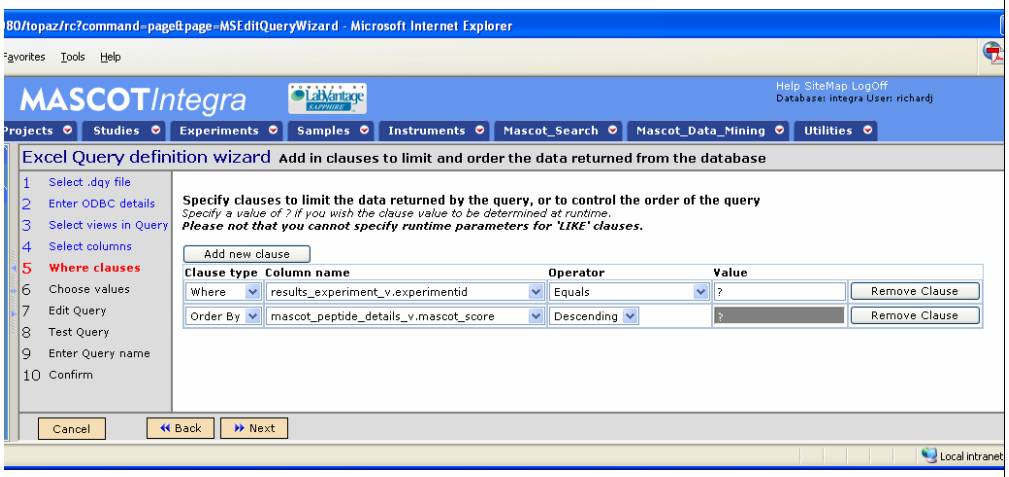
Next we select the columns to be reported

The available columns from our chosen views are displayed in the left panel.

And the selected columns are shown in the right panel.

You can order the columns as you prefer.

Excel query
– search by experiment id
and sort by peptide mascot score

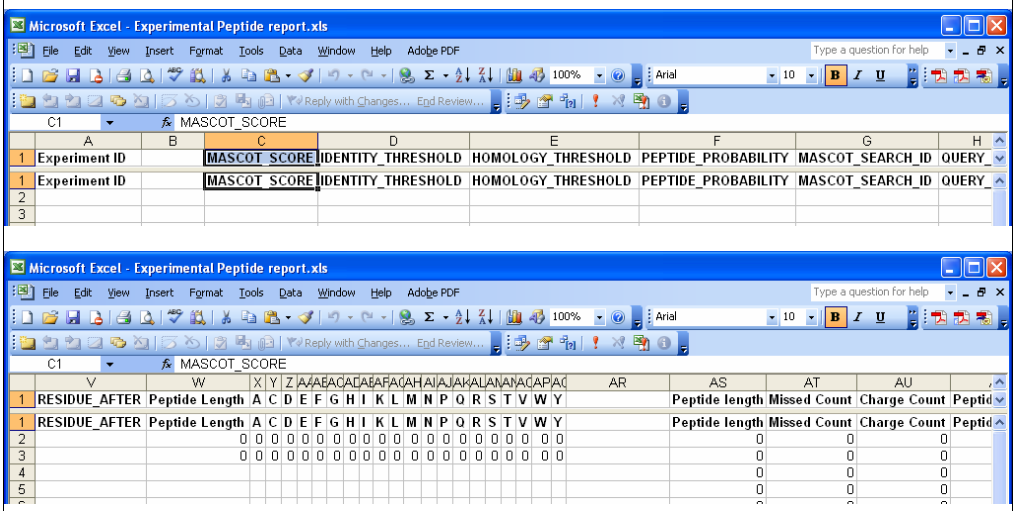


Next we specify clauses to limit the peptides to an experiment ID which will be defined at run time

And a second clause to sort the peptides by mascot score.

Mascot Integra builds the SQL statement which can then be edited (if required), tested and saved.

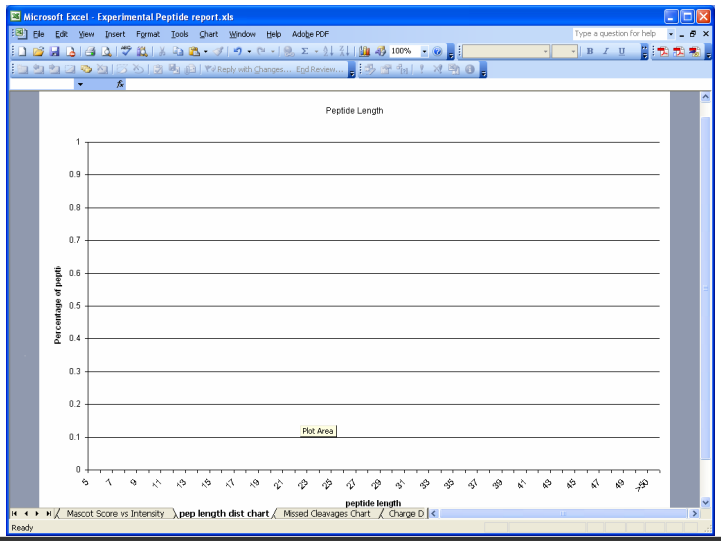
Build an Excel template



Next I open excel and build a template for the results.

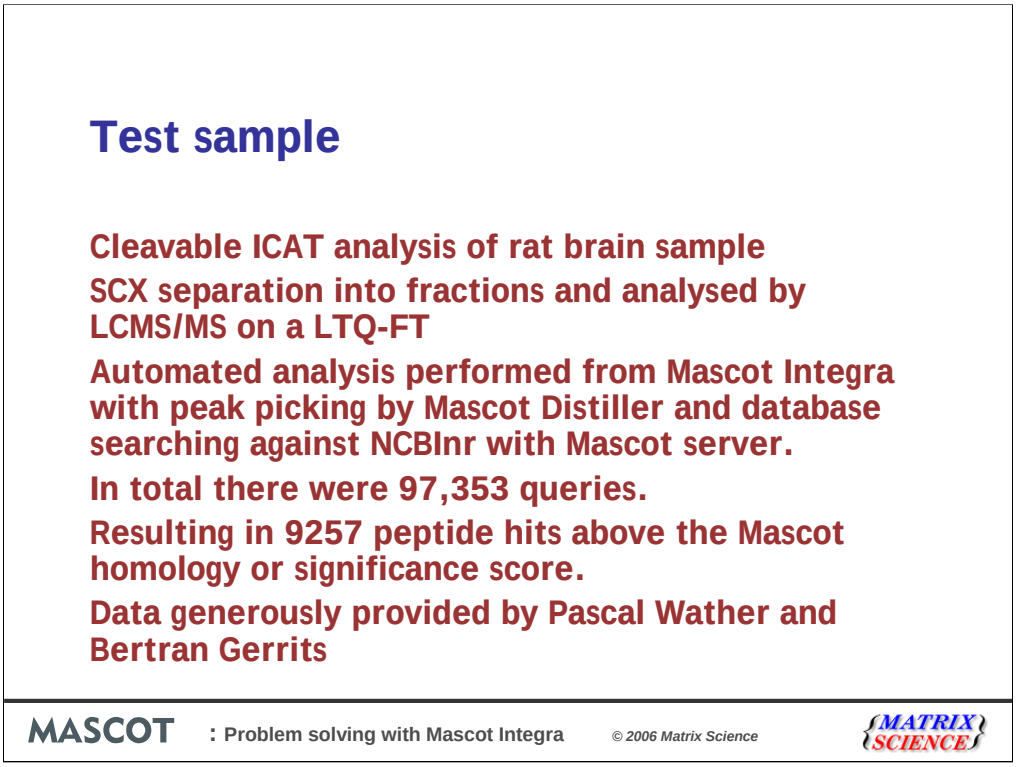
I specify where the data will be placed within the worksheet and can program additional calculations on the data and design graphs.

Build an Excel template



I can add additional graphs and calculations to the template at a later date as long as the query stays the same.

The excel sheet is then saved on the Mascot Integra server and added to the Excel report templates. Other users can then run the same report on their own searches/samples without having to know anything about the underlying SQL.



I had previously processed a data from large ICAT analysis of the rat brain proteome that had been separated by strong cation ion exchange followed by LCMS/MS on a LTQ-FT.

The data analysis was performed automatically with Mascot Distiller and Mascot Server.

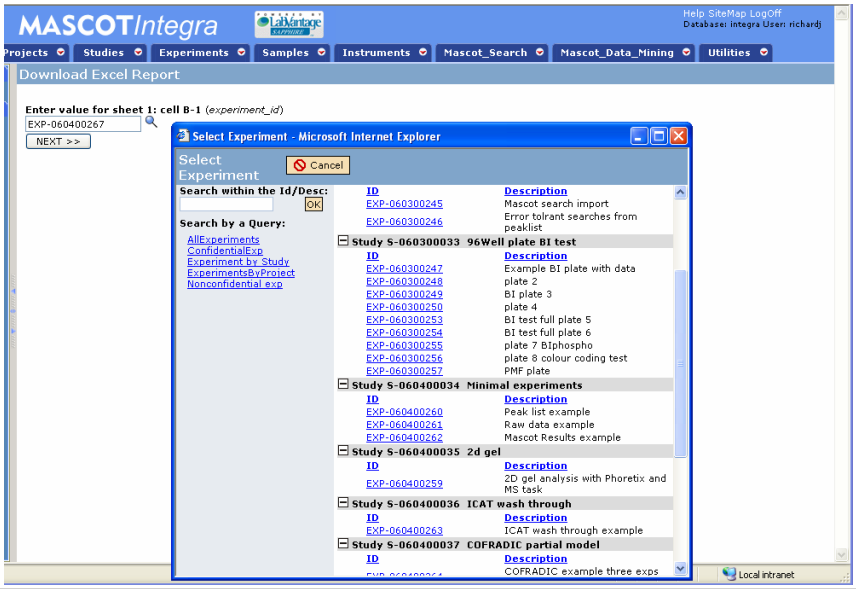
The search results were then available for reporting from Mascot Integra.

Excel query - export of custom reports

Lets export all the peptides above the homology score that have been identified by ms/ms from the experiment.

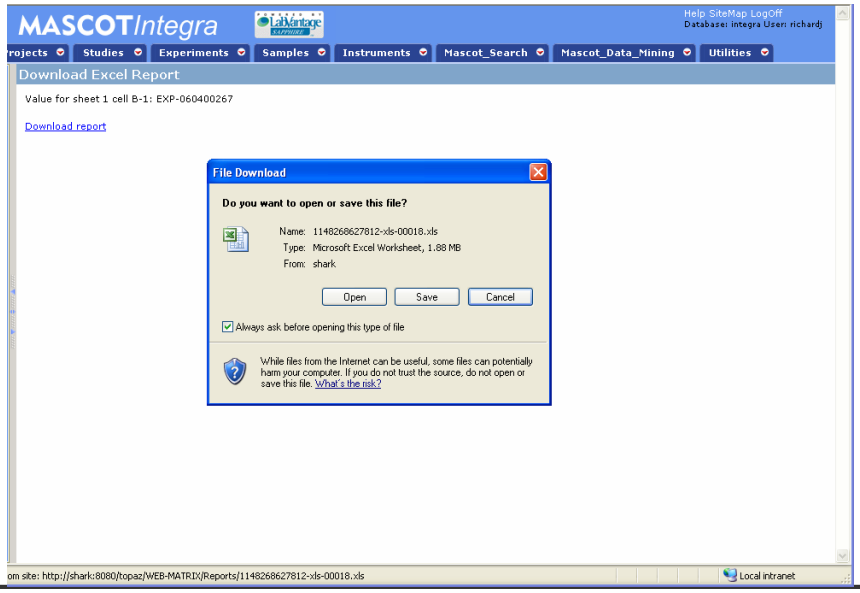
First we select the report template

Excel query - export of custom reports



Then the mascot search id

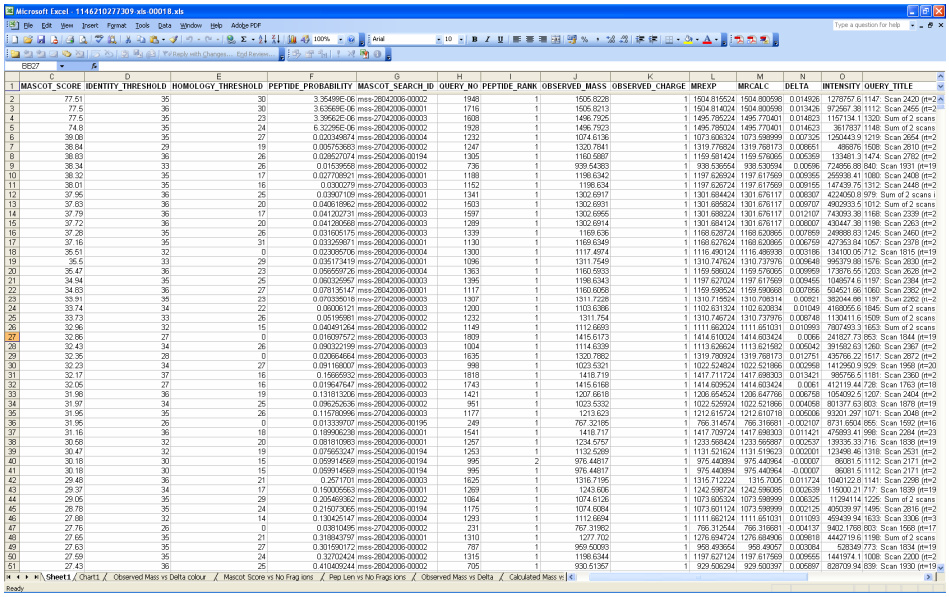
Excel query - export of custom reports



And finally save the report as an Excel document to the hard drive.

The report can then be opened and edited in Excel.

Excel query - view the results



	C	D	E	F	G	H	I	J	K	L	M	N	O
1	MASCOT_SCORE	IDENTITY_THRESHOLD	HOMOLOGY_THRESHOLD	PEPTIDE_PROBABILITY	MASCOT_SEARCH_ID	QUERY_NO	PEPTIDE_RANK	OBSERVED_MASS	OBSERVED_CHARGE	MREXP	MRCALC	DELTA	INTENSITY QUERY_TITLE
2	77.51	35	30	3.35499E-06	mss-28042006-00002	1948	1	1505.8228	1	1504.815524	1504.800598	0.014926	1278767.6 1147: Scan 2420 (rt=2
3	77.5	36	30	3.63669E-06	mss-28042006-00001	1716	1	1505.8213	1	1504.814024	1504.800598	0.013426	972567.38 1112: Scan 2455 (rt=2
4	77.5	35	23	3.39662E-06	mss-27042006-00003	1608	1	1496.7925	1	1495.785224	1495.770401	0.014823	1157134.1 1320: Sum of 2 scans
5	74.8	35	24	6.32295E-06	mss-28042006-00002	1928	1	1496.7923	1	1495.785024	1495.770401	0.014623	3617837.1 148: Sum of 2 scans
6	39.08	35	27	0.020349874	mss-28042006-00004	1232	1	1074.6136	1	1073.606324	1073.598999	0.007325	1292443.9 1219: Scan 2664 (rt=2
7	38.84	29	19	0.005753683	mss-27042006-00002	1247	1	1320.7841	1	1319.776824	1319.768173	0.008651	486876 1508: Scan 2810 (rt=2
8	38.83	36	26	0.028527074	mss-25042006-00194	1305	1	1160.6987	1	1159.691424	1159.676095	0.015329	133481.3 1474: Scan 2702 (rt=2
9	38.34	33	26	0.01539958	mss-28042006-00002	736	1	939.54383	1	938.536964	938.530694	0.00596	724956.88 640: Scan 1931 (rt=19
10	38.32	35	17	0.027708921	mss-28042006-00001	1189	1	1198.6342	1	1197.626924	1197.617569	0.009355	265938.41 1080: Scan 2408 (rt=2
11	38.01	35	16	0.0300279	mss-27042006-00003	1152	1	1198.634	1	1197.626724	1197.617569	0.009155	1474397.75 1312: Scan 2448 (rt=2
12	37.96	36	25	0.03907109	mss-28042006-00001	1341	1	1302.6917	1	1301.684424	1301.676117	0.008307	4204650.8 979: Sum of 2 scans i
13	37.83	36	20	0.040618962	mss-28042006-00002	1503	1	1302.6931	1	1301.685824	1301.676117	0.009707	4902933.6 1012: Sum of 2 scans
14	37.79	36	17	0.041202731	mss-28042006-00003	1597	1	1302.6955	1	1301.688224	1301.676117	0.012107	743093.38 1168: Scan 2339 (rt=2
15	37.72	36	20	0.041280568	mss-27042006-00003	1289	1	1302.6914	1	1301.684124	1301.676117	0.008007	430447.38 1198: Scan 2263 (rt=2
16	37.28	35	26	0.031605175	mss-28042006-00003	1339	1	1169.636	1	1168.628724	1168.620865	0.007859	249888.83 1245: Scan 2460 (rt=2
17	37.16	35	31	0.033259871	mss-28042006-00001	1130	1	1169.6349	1	1168.627624	1168.620865	0.006759	427353.94 1057: Scan 2378 (rt=2
18	35.51	32	0	0.023085706	mss-28042006-00004	1300	1	1117.4974	1	1116.490124	1116.486938	0.003186	134100.05 712: Scan 1815 (rt=19
19	35.5	33	29	0.035173419	mss-27042006-00001	1096	1	1311.7549	1	1310.747624	1310.737976	0.009648	995379.88 1576: Scan 2630 (rt=2
20	35.47	36	23	0.056559726	mss-28042006-00004	1363	1	1160.5933	1	1159.586024	1159.576065	0.009959	173876.55 1203: Scan 2626 (rt=2
21	34.94	35	25	0.060329957	mss-28042006-00003	1395	1	1198.6343	1	1197.627024	1197.617569	0.009455	1048574.6 1197: Scan 2384 (rt=2
22	34.83	36	27	0.078135147	mss-28042006-00001	1117	1	1160.6058	1	1159.598524	1159.590668	0.007856	504521.66 1060: Scan 2382 (rt=2
23	33.91	35	23	0.070335018	mss-27042006-00003	1307	1	1311.7228	1	1310.715524	1310.706314	0.00921	382044.66 1197: Scan 2262 (rt=2
24	33.74	34	22	0.06006121	mss-28042006-00003	1200	1	1103.6386	1	1102.631324	1102.620834	0.01049	4168055.6 1845: Sum of 2 scans
25	33.73	33	26	0.05195981	mss-27042006-00002	1232	1	1311.754	1	1310.746724	1310.737976	0.008748	1130411.6 1509: Sum of 2 scans
26	32.96	32	15	0.040491264	mss-28042006-00002	1149	1	1112.6693	1	1111.662024	1111.651031	0.010993	7807493.3 1663: Sum of 2 scans
27	32.86	27	0	0.016397572	mss-28042006-00003	1809	1	1416.6173	1	1414.610024	1414.603424	0.0066	241827.73 853: Scan 1844 (rt=19
28	32.43	34	26	0.090322199	mss-27042006-00003	1004	1	1114.6339	1	1113.626624	1113.621582	0.005042	391582.63 1260: Scan 2367 (rt=2
29	32.35	28	0	0.020664664	mss-28042006-00003	1635	1	1320.7882	1	1319.780924	1319.768173	0.012751	435766.22 1517: Scan 2872 (rt=2
30	32.23	34	27	0.091168007	mss-28042006-00003	988	1	1023.5221	1	1022.524624	1022.521896	0.002968	1412960.9 929: Scan 1966 (rt=20
31	32.17	37	16	0.15995932	mss-28042006-00003	1818	1	1418.719	1	1417.711724	1417.698303	0.013421	985756.5 1181: Scan 2360 (rt=2
32	32.05	27	16	0.019647647	mss-28042006-00002	1743	1	1415.6168	1	1414.609524	1414.603424	0.0061	412119.44 728: Scan 1763 (rt=18
33	31.98	36	19	0.131813206	mss-28042006-00003	1421	1	1207.6618	1	1206.654524	1206.647766	0.006758	1054092.5 1207: Scan 2404 (rt=2
34	31.97	34	25	0.096252636	mss-28042006-00002	961	1	1023.5332	1	1022.525924	1022.521896	0.004068	801377.63 803: Scan 1976 (rt=19
35	31.96	35	26	0.115780996	mss-27042006-00003	1177	1	1213.623	1	1212.615724	1212.610716	0.005008	93201.297 1071: Scan 2048 (rt=2
36	31.95	26	0	0.013339707	mss-25042006-00195	249	1	767.32165	1	766.314574	766.316681	-0.002107	8731.6504 655: Scan 1592 (rt=16
37	31.16	36	18	0.189606238	mss-28042006-00001	1541	1	1418.717	1	1417.709724	1417.698303	0.011421	476893.41 998: Scan 2284 (rt=23
38	30.58	32	20	0.081810983	mss-28042006-00001	1257	1	1234.5757	1	1233.568424	1233.565887	0.002537	139335.33 716: Scan 1838 (rt=19
39	30.47	32	19	0.075953247	mss-25042006-00194	1263	1	1132.5269	1	1131.521624	1131.519623	0.002001	123496.46 1318: Scan 2531 (rt=2
40	30.18	30	15	0.059914569	mss-25042006-00194	995	2	976.44817	1	975.440894	975.440864	-0.00007	86081.5 1112: Scan 2171 (rt=2
41	30.18	30	15	0.059914569	mss-25042006-00194	995	1	976.44817	1	975.440894	975.440864	-0.00007	86081.5 1112: Scan 2171 (rt=2
42	29.48	36	21	0.257101	mss-28042006-00003	1635	1	1316.7195	1	1315.712224	1315.7005	0.011724	1040122.8 1141: Scan 2206 (rt=2
43	29.37	34	17	0.160005563	mss-28042006-00001	1269	1	1243.606	1	1242.598724	1242.590695	0.008029	115000.21 717: Scan 1839 (rt=19
44	29.05	35	29	0.205469362	mss-28042006-00002	1064	1	1074.6126	1	1073.605324	1073.599999	0.005325	11294114 1225: Sum of 2 scans
45	28.78	35	24	0.215073065	mss-25042006-00194	1175	1	1074.6084	1	1073.601124	1073.599999	0.001125	409039.97 1496: Scan 2816 (rt=2
46	27.98	32	14	0.130425147	mss-28042006-00004	1293	1	1112.6694	1	1111.662124	1111.651031	0.011093	459439.94 1633: Scan 3306 (rt=3
47	27.76	36	0	0.03910495	mss-27042006-00002	231	1	767.31982	1	766.312544	766.316681	-0.004137	9402.1768 803: Scan 1568 (rt=17
48	27.66	35	21	0.318843797	mss-28042006-00001	1310	1	1277.702	1	1276.694724	1276.684906	0.009818	4442719.6 1198: Sum of 2 scans
49	27.63	35	27	0.301590172	mss-28042006-00002	787	1	959.50093	1	958.493654	958.49057	0.003084	528349.773 Scan 1834 (rt=19
50	27.59	35	24	0.32702424	mss-28042006-00002	1315	1	1198.6344	1	1197.627124	1197.617569	0.009555	1441974.1 1008: Scan 2200 (rt=2
51	27.43	36	25	0.410409244	mss-28042006-00002	705	1	930.51357	1	929.506294	929.500397	0.005897	828709.94 839: Scan 1930 (rt=19

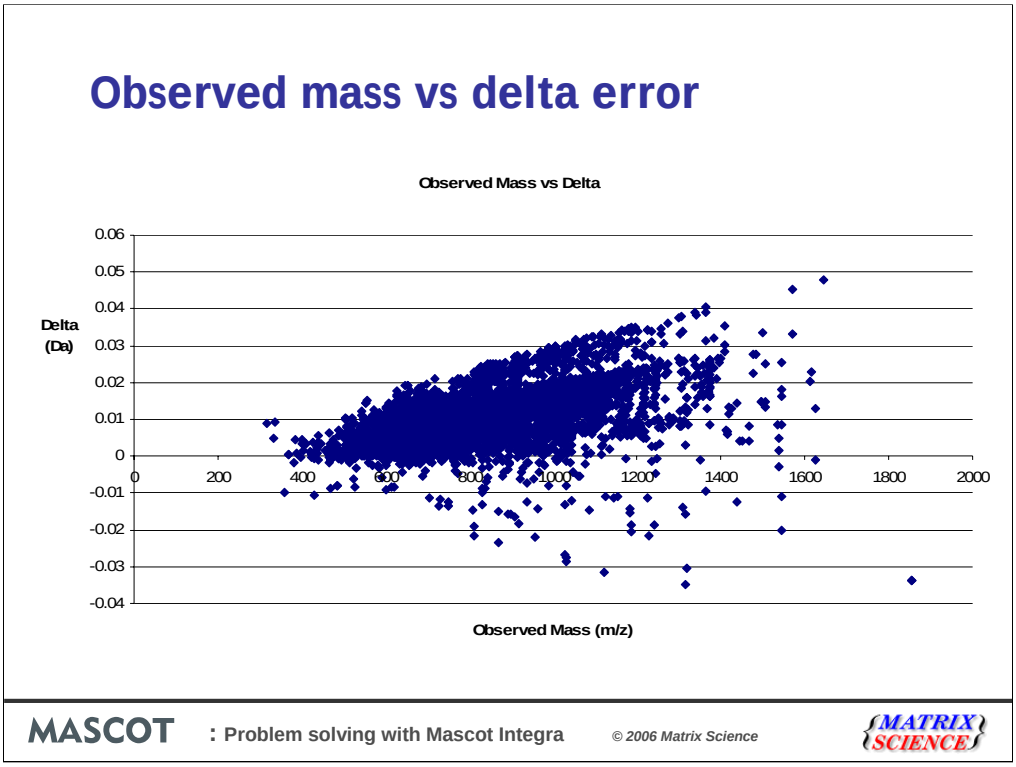
That 5 minutes of work designing a query produced a table with information on all the peptides identified in an experiment. In this case data on 9257 peptides.

Excel query
-what shall we do with the data

- Observed mass vs. delta error
- Peptide length
- Missed cleavages distribution
- Charge state distribution
- AA frequency
- Mascot score vs. delta error
- Mascot score vs. number of fragment ions

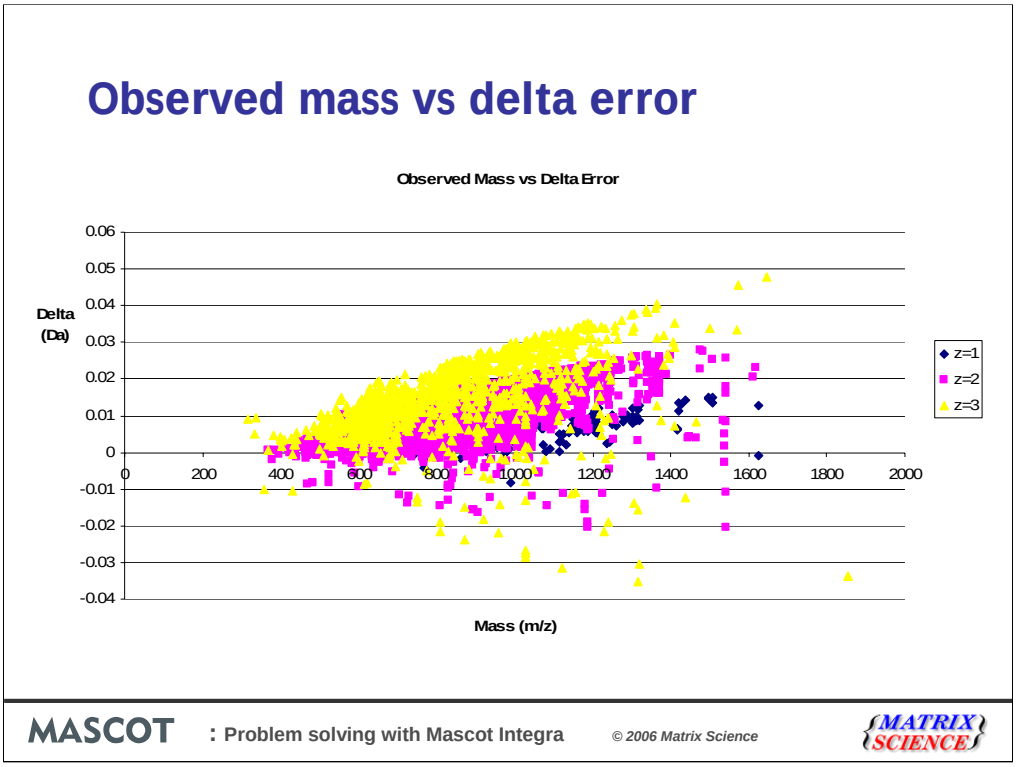
Most of you will be familiar with Excel and already have enough knowledge to generate sophisticated calculations and graphs.

Here are some of the calculations and graphs that I built.

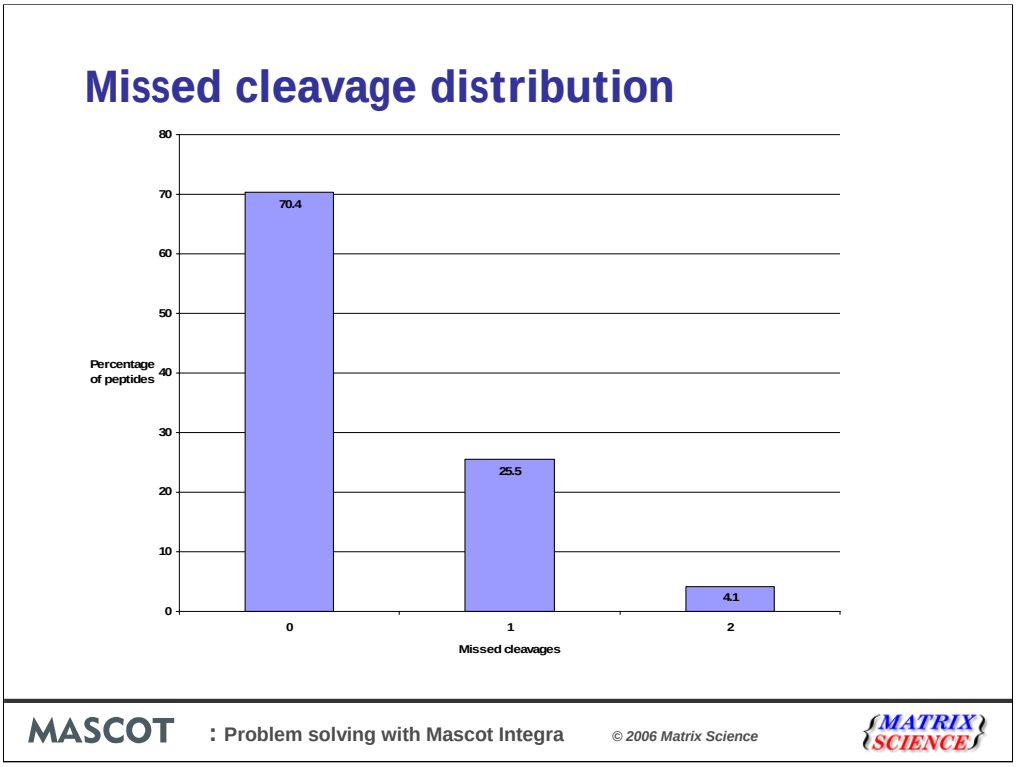


Looking at the observed mass vs delta error graph there seemed to be multiple populations probably relating to the charge state.

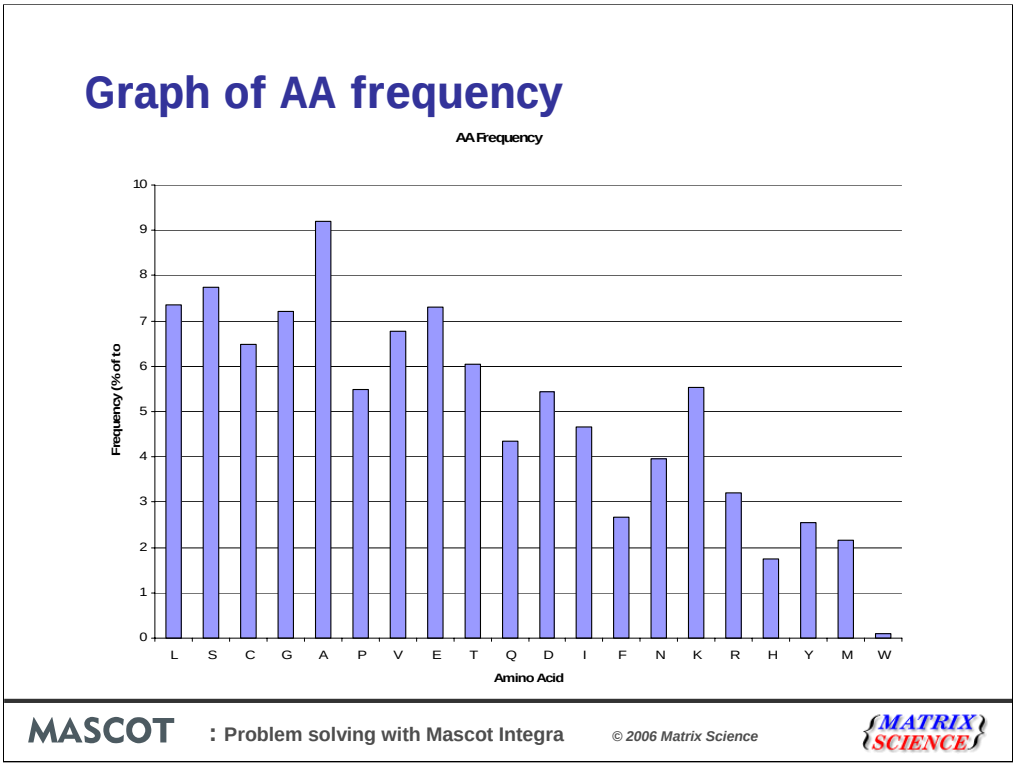
I sorted the data by charge state and re-plotted the graph with each charge state as a different series.



And you can see that the three populations are indeed loosely related to charge state.



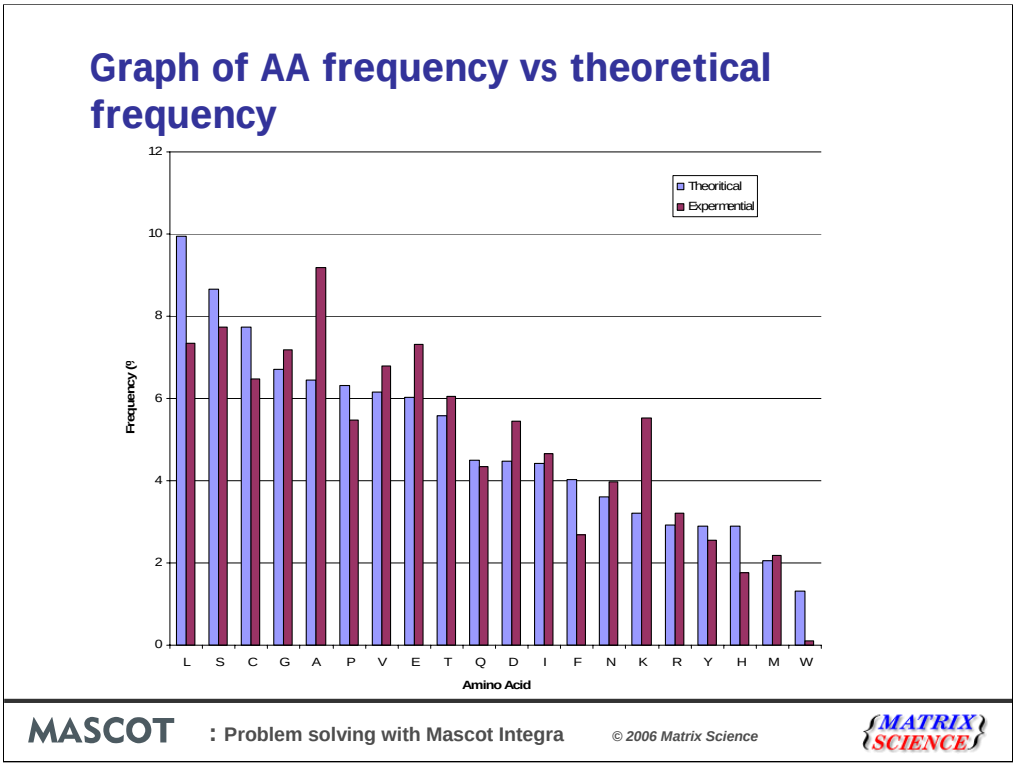
It was easy to build a missed cleavage histogram.



Next I calculated the amino acid frequency across all the peptides.

As this was an ICAT experiment the frequencies are going to be different compared to a complete proteome.

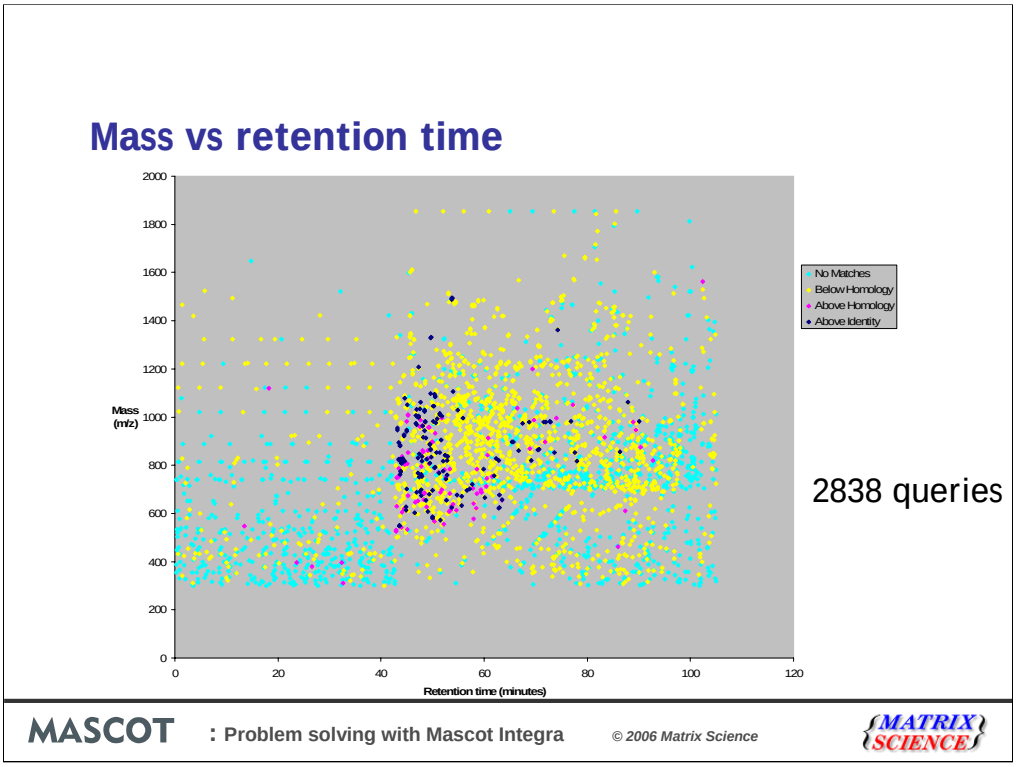
Therefore I generated the theoretical AA frequencies for all the tryptic Cysteine containing peptides and added them to the plot.



And we can see a number of differences between what was expected and what was measured.

Of particular interest was the under representation of Tryptophan.

I ran a number of error tolerant searches to see if I could identify more Tryptophan containing peptides with a consistent modification for example oxidation but the results have so far been inconclusive.



The second example is an overview plot of the mass and retention time of all the peptides analyzed by MS/MS from a single MuDPIT fraction.

There were 2838 ms/ms queries generated in the analysis of the fraction.

Four SQL queries have been combined on to one excel sheet to generate the four series shown in the plot.

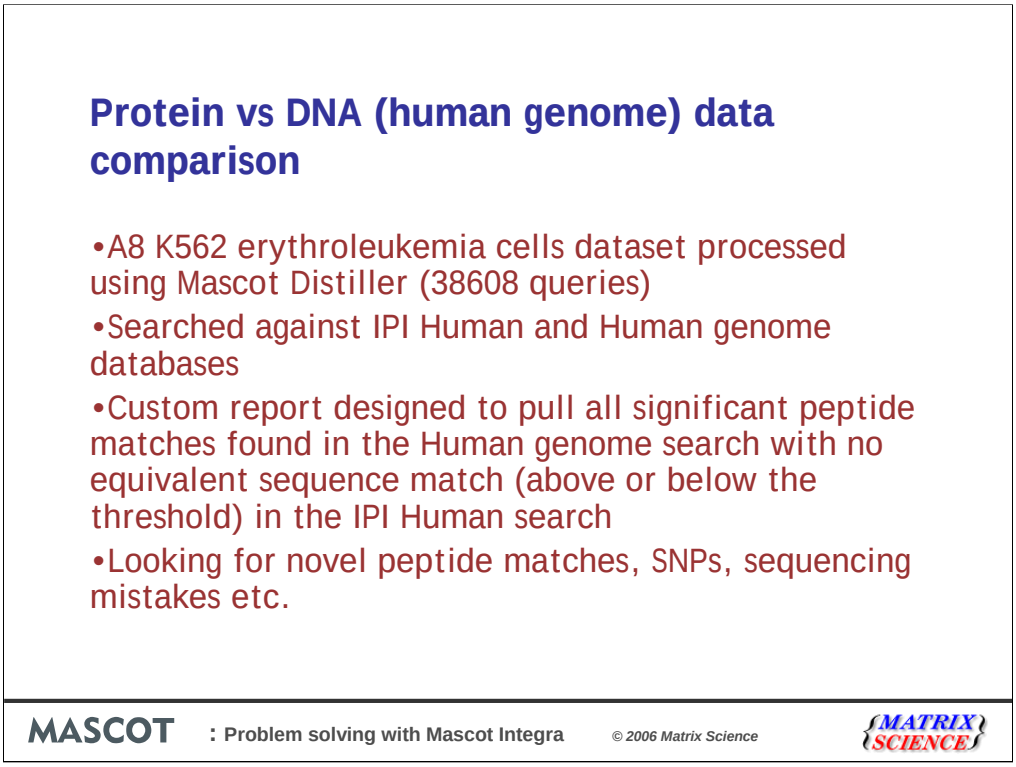
The queries with a significant Mascot score are shown in blue.

The queries with a Mascot score greater than homology but less than the significant cut off are shown in purple

The queries with an insignificant match are shown in yellow

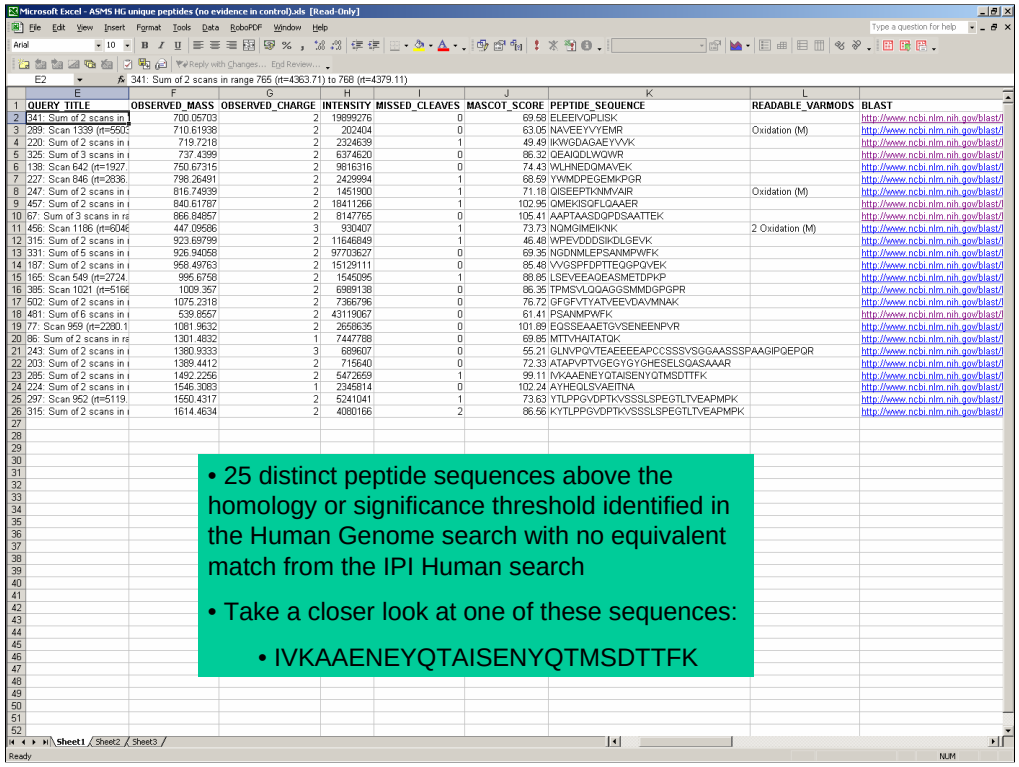
While the queries that did not generate a match are shown in cyan.

As you can see the majority of identifications eluted in the 40 to 80 minute period.



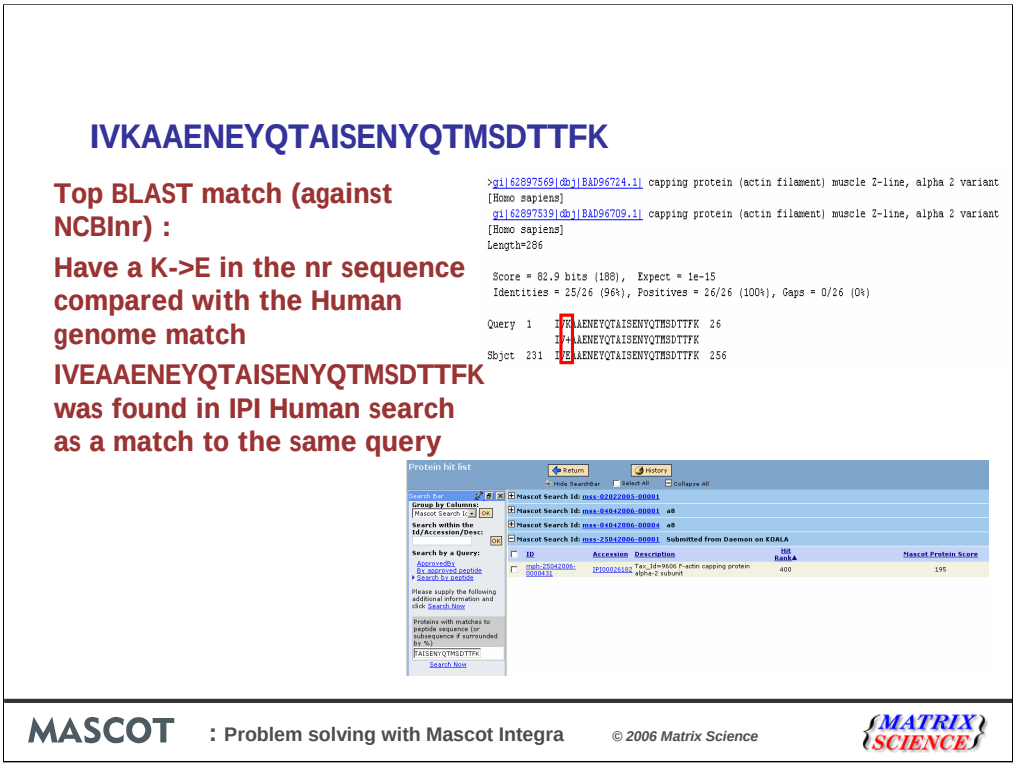
And the third example is a comparison between searching the IPI database and Human genome DNA database

We used the A8 Dataset from Katheryn Resing, and processed the raw data using Mascot Distiller to produce a dataset with 38608 queries. We then searched this against the IPI Human and Human Genome databases, and created a custom reports to identify all peptide matches from the Human genome search above the homology threshold with no equivalent sequence match (whether above the homology threshold or not) in the IPI Human search. By doing this we produce a list of peptides unique to the human genome search which we can use to look for novel peptide matches, possible sequencing errors, SNPs and other polymorphisms



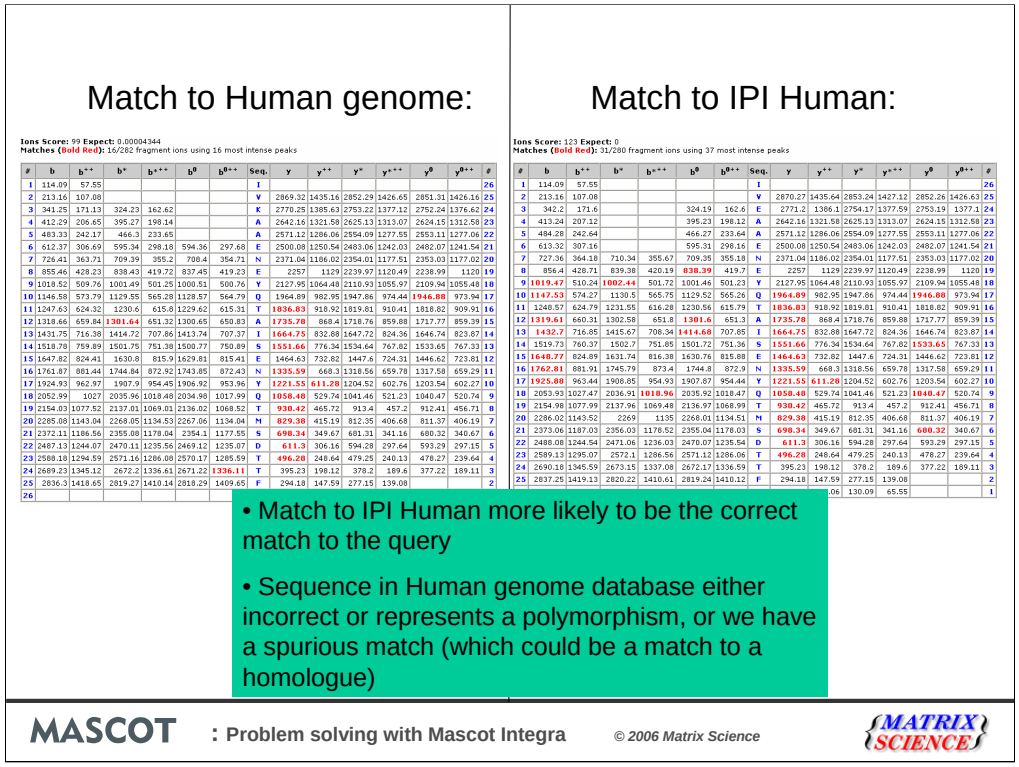
Running the report on these samples identified 25 distinct peptide sequences which were above the homology threshold in the Human Genome search which were absent from the IPI Human search. Doing the comparison by hand would have taken many hours (or days), but by having all of the information stored inside Mascot Integra, we were able to generate this hit list in a couple minutes.

The next step is, of course, to see if any of these peptides are telling us anything interesting – do any of them represent novel sequences, polymorphisms or sequencing errors? We'll take a closer look at one of the peptides identified.



The first step was to BLAST the peptide sequence against NCBInr and see if there is a match in the database. In this case we retrieved a good match to an actin capping protein – the only difference between the sequence from the Human Genome search against the sequence in NCBInr is the K->E at position 3 in the NCBInr sequence compared with the Human genome match.

Since the published sequence of this peptide/protein has Glutamic Acid at position 3, then any match from the IPI Human search would have been to this sequence. If we do a search in Integra for any protein hits containing this sequence, then we do indeed find a match to the peptide from the same query in the IPI Human search.



If we take a closer look at the peptide matches...Here we have the match from the Human Genome search – its a pretty good match with a reasonable run of y ions and gets a good score (98). However, the match to IPI Human (with Glutamic Acid at position 3) is clearly the stronger match – we have additional b series matches and a higher ions score.

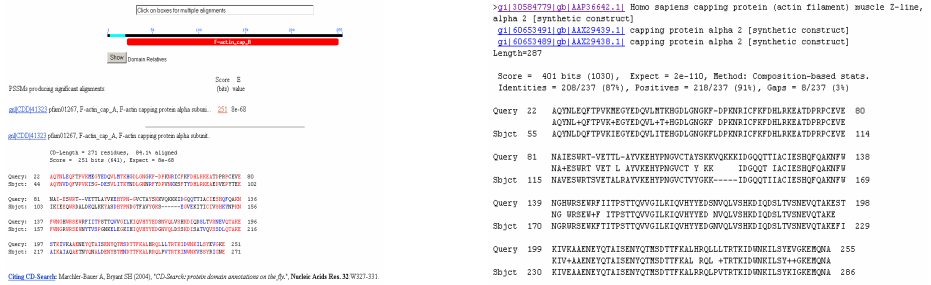
Therefore, we would conclude that the match to the sequence published in IPI Human and NCBInr is probably the correct match to this query. The sequence in the Human Genome database is either incorrect (i.e. a sequencing error) or it represents a polymorphism, or we have a spurious match. The spurious match could be a genuine 'false positive', or it could be a match to an homologous protein sequence. If it is to a homologous sequence then the analysis could have identified a match of interest.

Match was on chromosome 15

Actin capping protein on chromosome 7

BLAST translated sequences for this section of chromosome 15

Find a very strong match to Actin capping protein (but not identical) on Frames 4 and 5:



>gi|30584779|gb|AAP36642.1| Homo sapiens capping protein (actin filament) muscle Z-line, alpha 2 [synthetic construct]
gi|60653491|gb|AAZ29439.1| capping protein alpha 2 [synthetic construct]
gi|60653489|gb|AAZ29438.1| capping protein alpha 2 [synthetic construct]
Length=287

Score = 401 bits (1030), Expect = 2e-110, Method: Composition-based stats.
Identities = 208/237 (87%), Positives = 218/237 (91%), Gaps = 8/237 (3%)

Query 22 AQYNLEQFTPVKMEGYEDQVLNTKHGDLGNGKF-DPKNRICFKFDHLRKEATDPRPCEVE 80
AQYNL+QFTPVK+EGYEDQVL+T+HGDLGNGKF DPKNRICFKFDHLRKEATDPRPCEVE
Sbjct 55 AQYNLDQFTPVKIEGYEDQVLITEHGDLGNGKFLDPKNRICFKFDHLRKEATDPRPCEVE 114

Query 81 NAIESWRT-VETTL-AYVKEHYPNGVCTAYSKKVQKKKIDGQQTTIACIESHQFQAKNFW 138
NA+ESWRT VET L AYVKEHYPNGVCT Y KK IDGQQT IACIESHQFQAKNFW
Sbjct 115 NAVESWRTSVETALRAYVKEHYPNGVCTVYGKK-----IDGQQTIIACIESHQFQAKNFW 169

Query 139 NGHWRSEWRFIITPSTTQVVGILKIQVHYYEDSNVQLVSHKDIQDSLTVSNEVQTAKEST 198
NG WRSEW+F ITPSTTQVVGILKIQVHYYED NVQLVSHKDIQDSLTVSNEVQTAKE
Sbjct 170 NGRWRSEWKFTITPSTTQVVGILKIQVHYYEDGNVQLVSHKDIQDSLTVSNEVQTAKEFI 229

Query 199 KIVEAAENEVQTAISENTVQTNSDTTFKALHRQLLLTRTKIDWNKILSYEVGKEMQNA 255
KIV+AAENEVQTAISENTVQTNSDTTFKAL RQL +TRTKIDWNKILSY+VGKEMQNA
Sbjct 230 KIVEAAENEVQTAISENTVQTNSDTTFKALRRQLPVTRTKIDWNKILSYKIGKEMQNA 286

Conclusion: Match found in Human Genome could have identified a novel homologue on chromosome 15

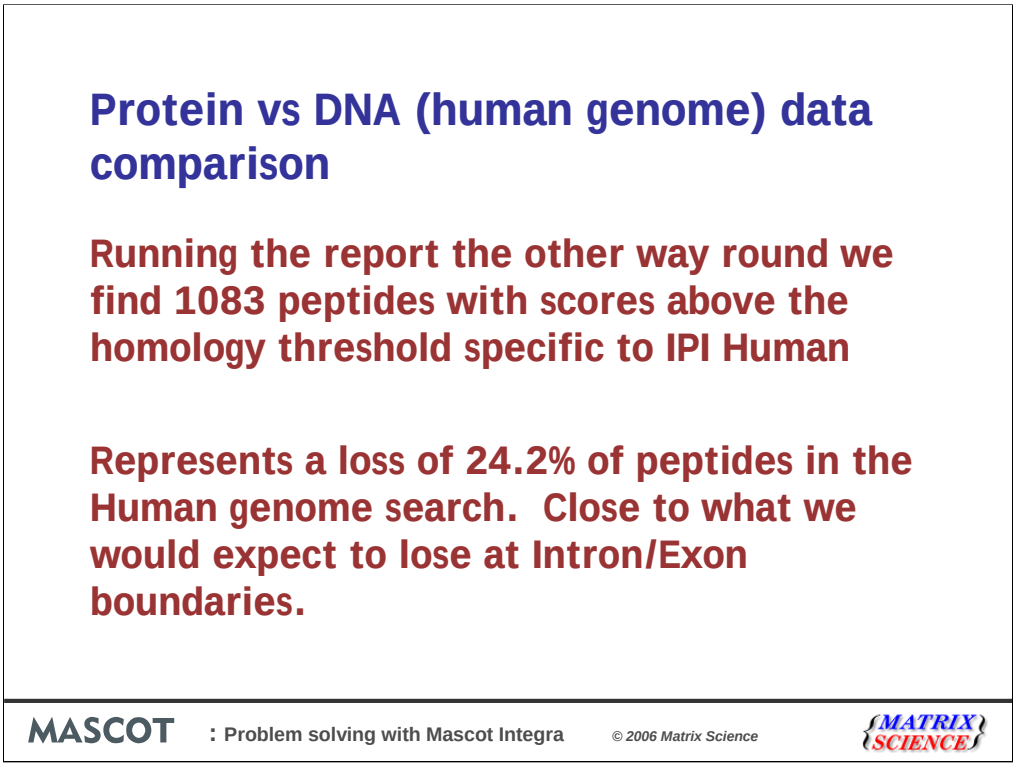
The human genome peptide match was found on chromosome 15.

The published location of human Acting capping protein is on chromosome 7.

To see if we have found a possible homologue to Actin capping protein we did a BLAST search of the translated protein sequences from frames 4 and 5 of this region of chromosome 15.

The result of this is that there is a very strong (85% identity) match to Actin capping protein – but it isn't an identical sequence.

Therefore we may have a novel gene encoding a homologue of Actin capping protein on chromosome 15 (nothing in Ensembl to confirm this though), or there may be a pseudogene present there (again there are no pseudogene matches to the sequence from chromosome 15 in Ensembl).



The report can, of course, be run the other way around, to identify peptide matches unique to the IPI Human search. If we do this the we find 1766 peptides from the IPI Human search with scores above the homology threshold which do not have an equivalent match in the Human genome search.

The IPI Human search identified a total of 4461 possible peptide matches with scores above the homology threshold, so the 1083 'missing' peptides represent approximately 24% of the total number of peptides identified from IPI Human – close to the approx 20% of peptide matches we would expect to loose because of Intron/Exon boundaries.

Advanced reporting

96 well plate assay report.

PMF report that satisfies MCP guidelines.

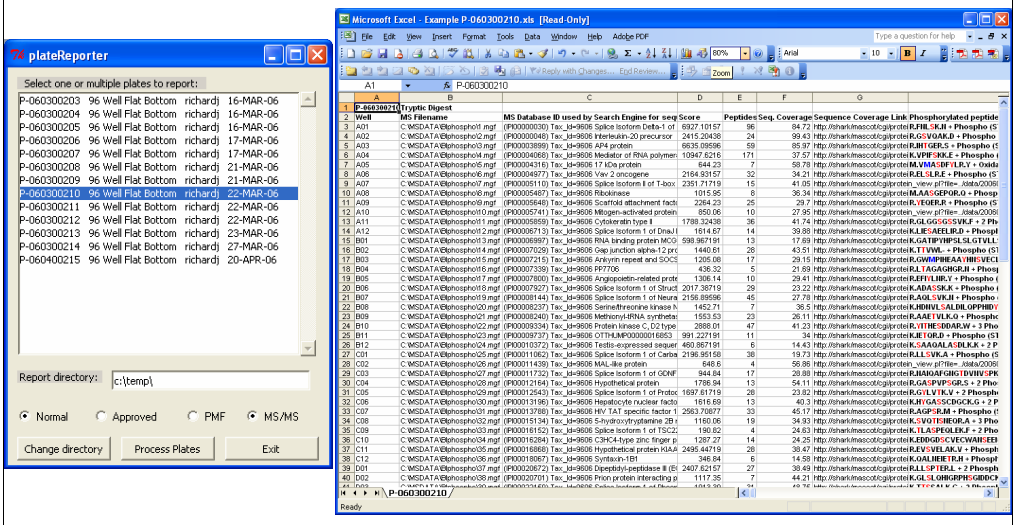
- Blast cluster of proteins to determine nearest non homologous match
- Meet all of the MCP criteria

Now on to more advanced reporting. As I said earlier we could use many different computer languages to do this but for these examples I used the perl scripting language.

The first example is of an assay report for a 96 well plate.

The second is a report for the proteins identified from a 2D gel by peptide mass fingerprinting and it has to satisfy the Molecular and Cellular proteomics guide lines

96 well plate assay report



MASCOT : Problem solving with Mascot Integra © 2006 Matrix Science MATRIX SCIENCE

Imagine that you have an assay with a single protein per a well of 96 well plate and that you are interested in the modified peptides for each protein.

First we select the plate.

Then advanced reporting application pulls out the protein result for each well of the plate and list the modified peptides for the protein with colour coded amino acids and save the results into an excel sheet.

Molecular Cellular Proteomics publication guidelines

Guideline 1: Supporting information

- The method and/or program (including version number) used to create the "peak list" from the raw data and the parameters used in the creation of this peak list.
- The name and version of the program(s) used for database searching and the values of search parameters.
- The name and version of the sequence database(s) used.

This next example generates a peptide mass fingerprint report that meets the Molecular Cellular Proteomics publication guidelines.

I'm only going to review the guidelines that pertain to general reporting or are specifically for PMF data.

The Molecular Cellular Proteomics publication guideline 1 specifies what supporting information should be reported with data.

The data analysis program and parameters used

The database searching program and parameters used

And the name and version of the database used.

Molecular Cellular Proteomics publication guidelines

Guideline 2: Information for each protein sequence identified should specify the following:

- Accession number and database source;
- Score(s) and any associated statistical information obtained for searches conducted;
- Sequence coverage
- Total number of peptides assigned to the protein.

Guideline 3: Additional potentially valuable information

- Retention time of each peptide
- Observation of multiple charge states
- Multiple observations of the same peptide
- Flanking residues
- Start and end positions of peptides in proteins

The second guideline specifies the protein information that should be reported while the third guide line is concerned with the peptides.

Molecular Cellular Proteomics publication guidelines

Guideline 6 Peptide Mass Fingerprinting

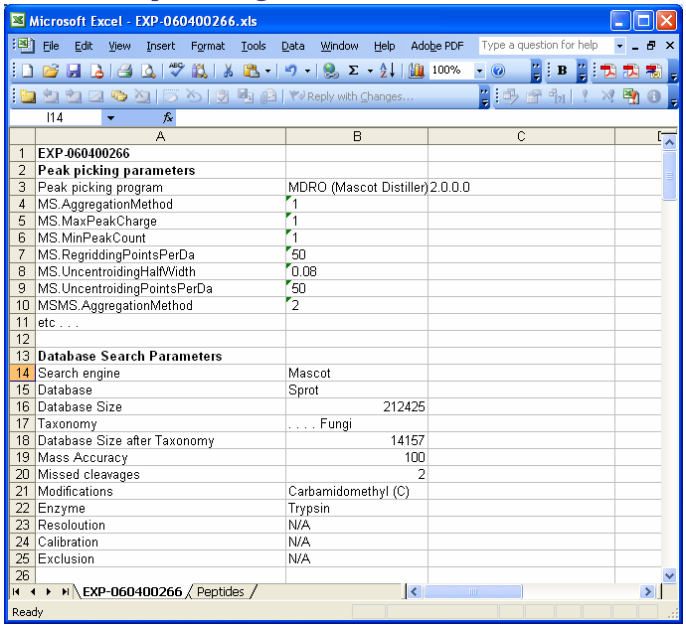
- Number of matched peaks
- Number of unmatched peaks
- Sequence coverage
- In addition to the score for the top match we must also show the score for the highest ranked hit to a non-homologous protein

Guideline 6 is for peptide mass fingerprinting and asks for the number of matched and unmatched peaks, sequence coverage and the score of the nearest non-homologous protein hit.

Depending on the redundancy of the database this may not necessarily be the second ranked protein hit.

We determine the score for the highest ranked hit to a non-homologous protein with BLAST cluster.

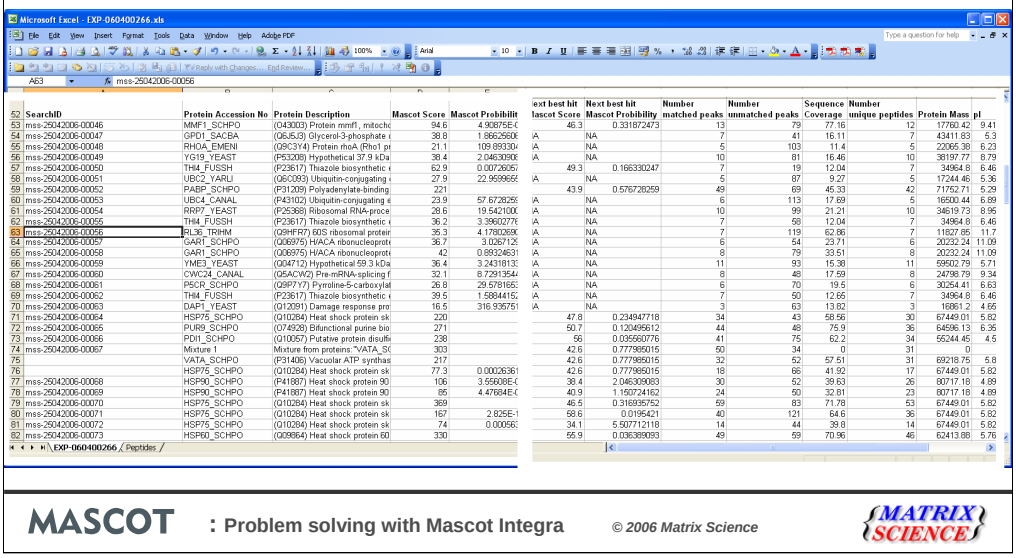
MCP PMF report guideline 1



	A	B	C
1	**EXP-060400266**		
2	**Peak picking parameters**		
3	Peak picking program	MDRO (Mascot Distiller)	2.0.0.0
4	MS.AggregationMethod	1	
5	MS.MaxPeakCharge	1	
6	MS.MinPeakCount	1	
7	MS.RegriddingPointsPerDa	50	
8	MS.UncentroidingHalfWidth	0.08	
9	MS.UncentroidingPointsPerDa	50	
10	MSMS.AggregationMethod	2	
11	etc . . .		
12			
13	**Database Search Parameters**		
14	Search engine	Mascot	
15	Database	Sprot	
16	Database Size	212425	
17	Taxonomy	. . . . Fungi	
18	Database Size after Taxonomy	14157	
19	Mass Accuracy	100	
20	Missed cleavages	2	
21	Modifications	Carbamidomethyl (C)	
22	Enzyme	Trypsin	
23	Resoloution	N/A	
24	Calibration	N/A	
25	Exclusion	N/A	
26			

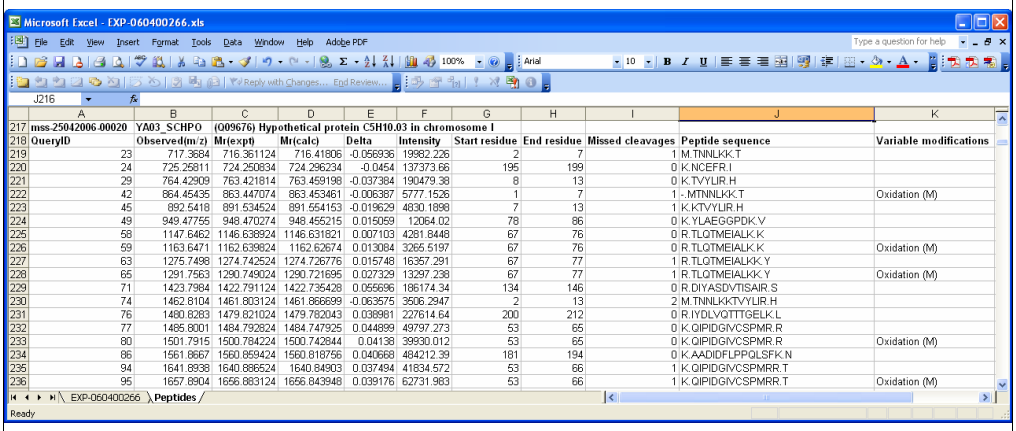
After selecting the 2D gel experiment the application extracts the processing and database searching parameters from Mascot Integra.

MCP PMF report
guidelines 2 and 6 - proteins



Then reports the top ranked protein hit along with information required by guidelines 2 and 6 to the first worksheet.

MCP PMF report guidelines 3 - peptides



Microsoft Excel - EXP-060400266.xls

	A	B	C	D	E	F	G	H	I	J	K
217	mss-25042006.00020	YA03_SCHPO	(Q09676) Hypothetical protein C5H10.03 in chromosome I								
218	QueryID	Observed(m/z)	Mr(expt)	Mr(calc)	Delta	Intensity	Start residue	End residue	Missed cleavages	Peptide sequence	Variable modifications
219	23	717.3684	716.361124	716.41806	-0.056936	19982.226	2	7	1	M.TNNLKK.T	
220	24	725.25811	724.250834	724.296234	-0.0454	137373.66	195	199	0	K.NCEFR.I	
221	29	764.42909	763.421814	763.459198	-0.037384	190479.38	8	13	0	K.TVYLIR.H	
222	42	864.45435	863.447074	863.453461	-0.006387	5777.1526	1	7	1	-.MTNNLKK.T	Oxidation (M)
223	45	892.5418	891.534524	891.554153	-0.019629	4830.1898	7	13	1	K.KTVYLIR.H	
224	49	949.47755	948.470274	948.455215	0.015059	12064.02	78	86	0	K.YLAEGGPDK.V	
225	58	1147.6462	1146.638924	1146.631821	0.007103	4281.8448	67	76	0	R.TLQTMEIALK.K	
226	59	1163.6471	1162.639824	1162.62674	0.013084	3265.5197	67	76	0	R.TLQTMEIALK.K	Oxidation (M)
227	63	1275.7498	1274.742524	1274.726776	0.015748	16357.291	67	77	1	R.TLQTMEIALKK.Y	
228	65	1291.7563	1290.749024	1290.721695	0.027329	13297.238	67	77	1	R.TLQTMEIALKK.Y	Oxidation (M)
229	71	1423.7984	1422.791124	1422.735428	0.055696	186174.34	134	146	0	R.DIYASDVTISAIR.S	
230	74	1462.8104	1461.803124	1461.866699	-0.063575	3506.2947	2	13	2	M.TNNLKKTVYLIR.H	
231	76	1480.8283	1479.821024	1479.782043	0.038981	227614.64	200	212	0	R.IYDLVQTTTGELK.L	
232	77	1485.8001	1484.792824	1484.747925	0.044899	49797.273	53	65	0	K.QIPIDGIVCSPMR.R	
233	80	1501.7915	1500.784224	1500.742844	0.04138	39930.012	53	65	0	K.QIPIDGIVCSPMR.R	Oxidation (M)
234	86	1561.8667	1560.859424	1560.818756	0.040668	484212.39	181	194	0	K.AADIDFLPPQLSFK.N	
235	94	1641.8938	1640.886524	1640.84903	0.037494	41834.572	53	66	1	K.QIPIDGIVCSPMRR.T	
236	95	1657.8904	1656.883124	1656.843948	0.039176	62731.983	53	66	1	K.QIPIDGIVCSPMRR.T	Oxidation (M)

EXP-060400266 | Peptides

Ready

MASCOT : Problem solving with Mascot Integra MATRIX SCIENCE

On a second worksheet the peptide information for each protein hit is reported and that satisfies guideline 3.

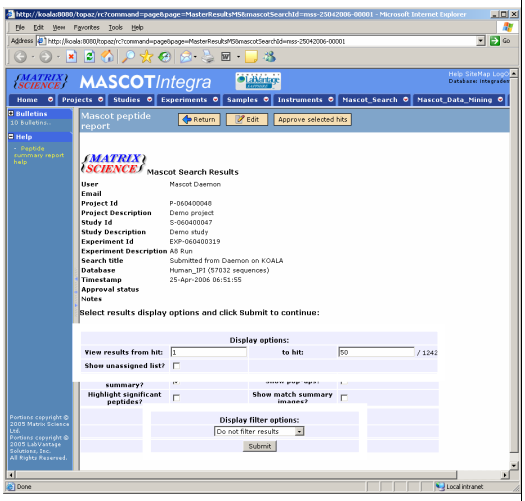
4. Large search viewing

Paging of results

- Can select range of protein hit ranks to view
- Far easier to handle than all 1242 protein hit ranks at once
- Avoids client side browser memory issues
- Much faster to generate report
  - 5 hit ranks ~ 9seconds
  - 50 hit ranks ~ 20 seconds
  - 100 hit ranks ~ 30 seconds
  - All 1242 ranks ~ 11 minutes*

Filtering of results

- Integra ships with a range of filters which can be applied to reports
- Can write your own SQL filter to limit proteins returned
- Can filter on any aspect of the protein, peptide or original query.



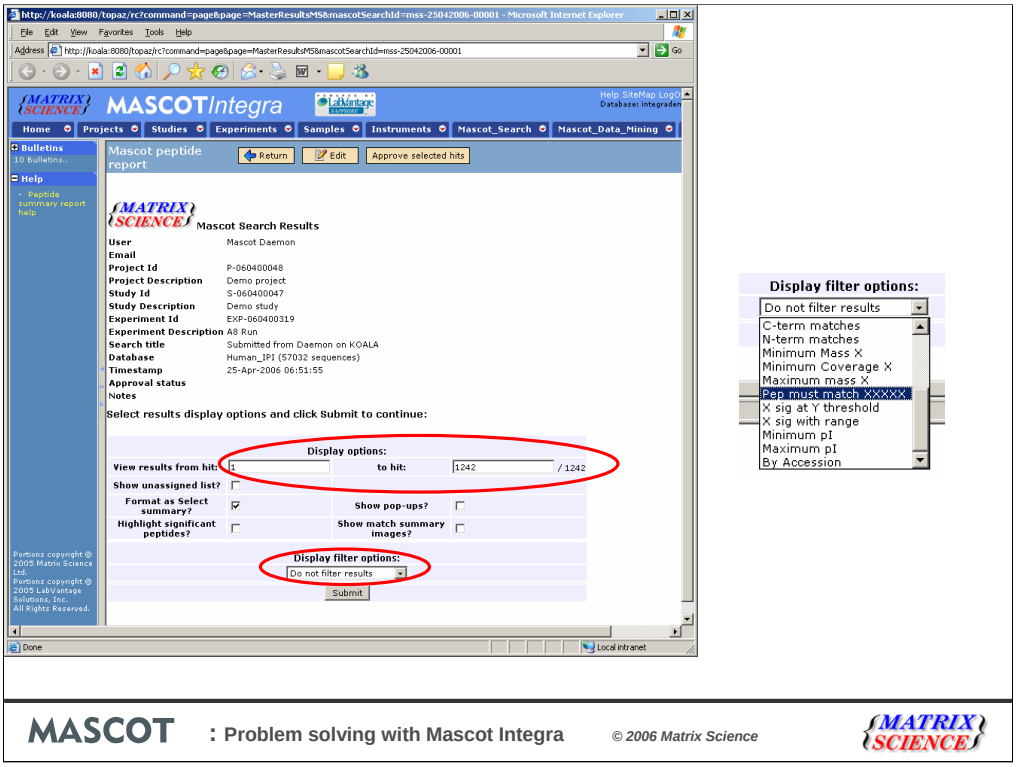
* when opened either from Mascot Integra or from Mascot server

Another advantage of using Mascot Integra is that we can simplify viewing large result sets in the standard reports.

For example because we are pulling the protein hits out of the Integra database we can specify a range of protein hit ranks to view in the report. For example here I'm only going to open up the first 50 protein hit rank from the A8 Dataset – a report which contains 1242 protein hit ranks in total. This means that we only have to deal with a subset of the results at any one time, and that Internet Explorer isn't going to run out of memory on your client PC – which can be a major problem when opening large result sets – and the report is much faster to generate than opening all of the results on the Mascot server. For example, opening just 5 hit ranks from the A8 dataset takes 9 seconds, 20 seconds to open 50, 30 seconds to open 100 but approximately 11 minutes to open all 1242 hit ranks. Opening the report from the Mascot server also took approximately 11 minutes (the results started displayed earlier, but it was 11 minutes before IE was accepting user input)

We can also write SQL filters which produce a valid list of protein hits to display. We can filter on any aspect of the protein hit

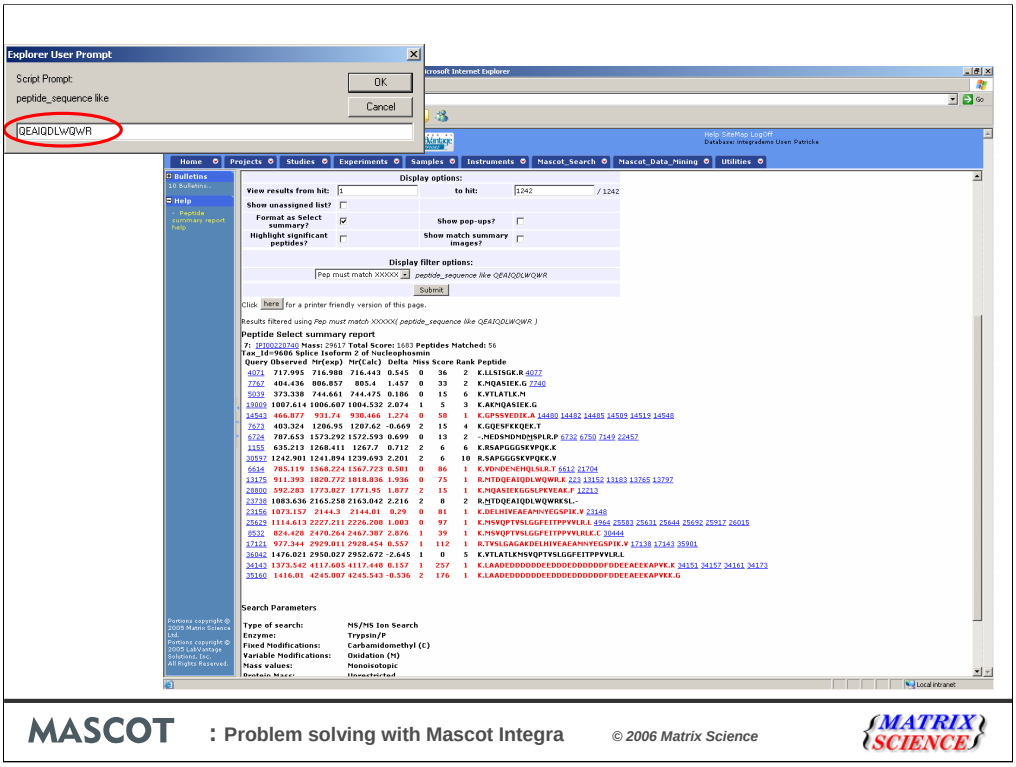
– i.e. the protein length, mass or pI – on the peptides which the protein has matched – i.e. require that the hit contains at least X distinct peptide matches above the homology threshold – or on the original query – i.e. pull out all the proteins which contain a good match to Query 53



This is how we apply a filter

Note that I've chosen to display all 1242 hit ranks. This is because the result filter and view to/from ranks work in conjunction. If I had chosen to display only the 1st 50 protein hit ranks, then the filter would only work on those ranks.

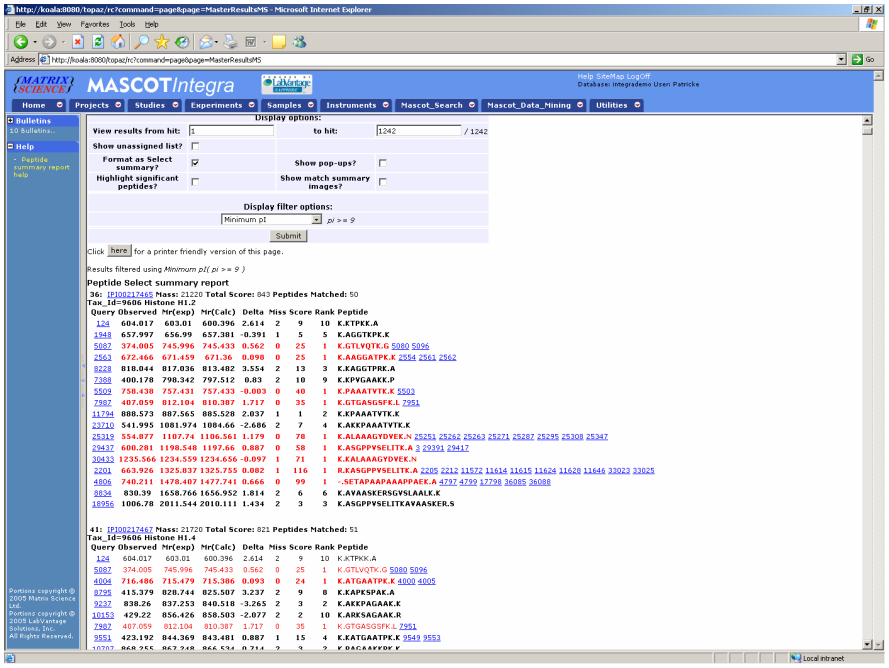
1st we select our filter from the drop down list. Here we're using one of the filters that comes with Mascot Integra, which allows us to filter out any proteins which do not contain a Mascot peptide match to a peptide sequence/subsequence.



When we select the filter, we are presented with a popup window into which we must enter the peptide subsequence we are interested in. In this case we'll use one of the peptide sequences identified in the A8 dataset.

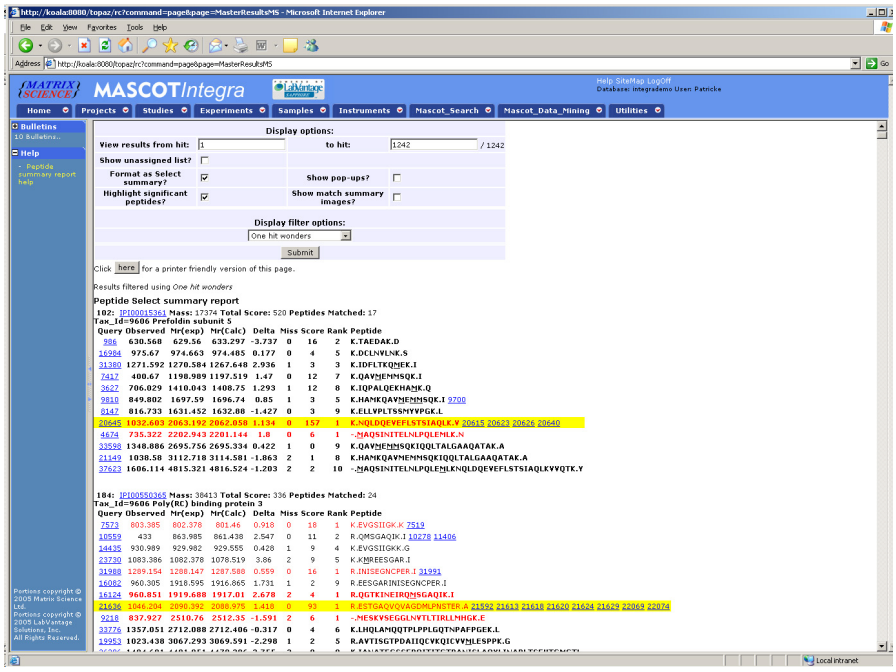
Click OK and then Submit.

This is the resulting report. Out of all 1242 protein hit ranks, only hit number 7 – Nucleophosmin – contains a match to the required peptide subsequence.

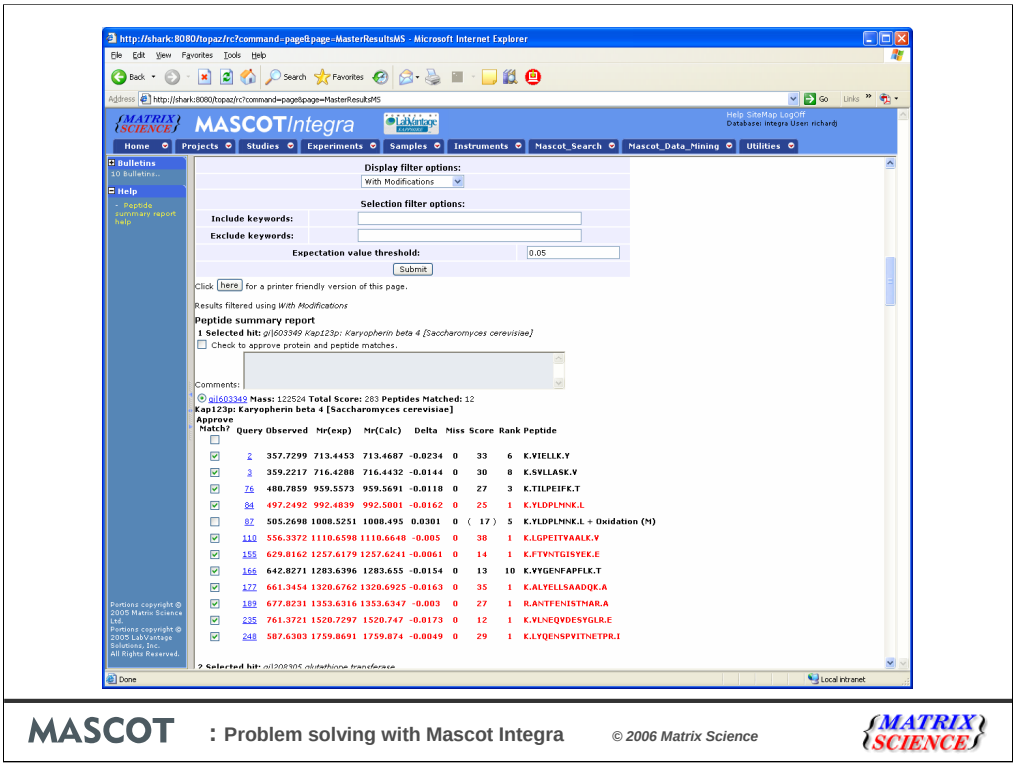


Some more quick examples of the type of things we can do with filtering.

1: Here we are filtering for pI and only viewing protein hits with a predicted pI of > 9



2: Here we filtering for only those proteins that contain one significant peptide match so called one hit wonders.

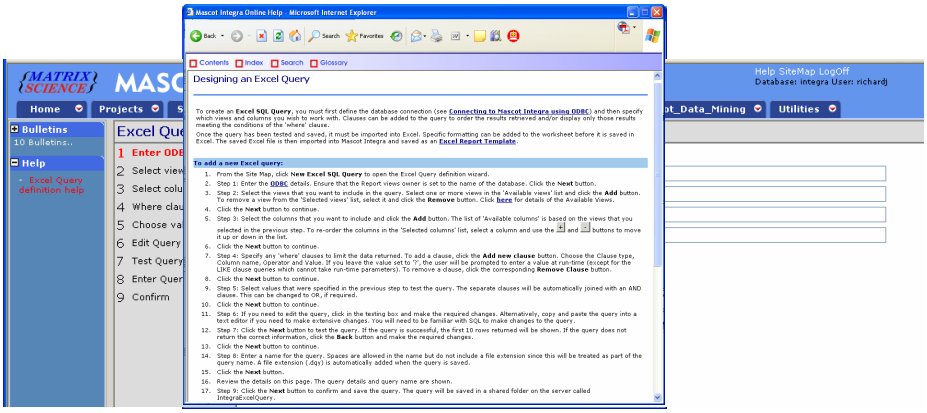


3: Here we are filtering out any protein matches that contain any peptide with any variable modification.

It would, of course, be possible to generate a filter which acted on any, or all, of these criteria in one go, so it offers a very flexible and powerful way of mining larger Mascot search results

Help

Online context sensitive help



MASCOT : Problem solving with Mascot Integra

Don't worry if some of that looked complicated, online context sensitive help is at hand.

You can access help on each of the menu options from the tramline page or open the relevant help section directly from your current location in Integra.

Shown here is a walk through for designing new excel queries

Acknowledgments

Pascal Wather

- Department of Biochemistry, University of Zurich, Switzerland

and Bertran Gerrits for mouse ICAT dataset

- Functional Genomics Center Zurich, Uni|ETH Zurich, Switzerland

Katheryn Resing for A8 dataset

- University of Colorado, USA

I would like to thank two labs for the data sets,

Pascal Wather and Bertran Gerrits for the use of the Mouse ICAT data and

Katheryn Resing for making the A8 dataset publicly available on the Peptide Atlas website.