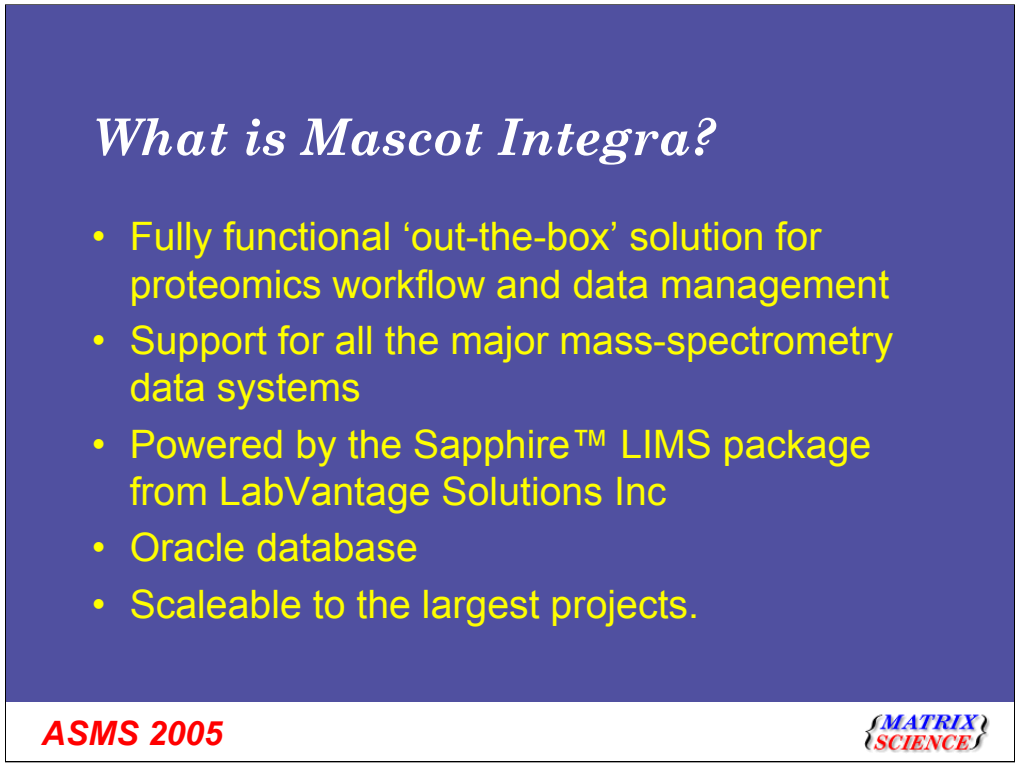
Data mining with Mascot Integra

ASMS 2005

MATRIX SCIENCE



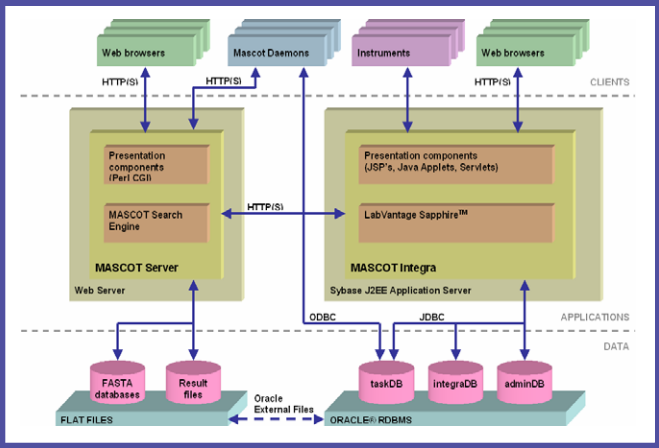
Mascot Integra is supplied as a ready to run system. It does not require the extensive setup and customisation associated with a traditional LIMS package.

Rather than re-invent the wheel, we have partnered with LabVantage Solutions Inc, (www.lims.com). Their Sapphire LIMS package provides the sample tracking and workflow modelling functionality for Mascot Integra

Using the Oracle database management system enables the database to scale efficiently as your data management requirements grow

Mascot Integra architecture

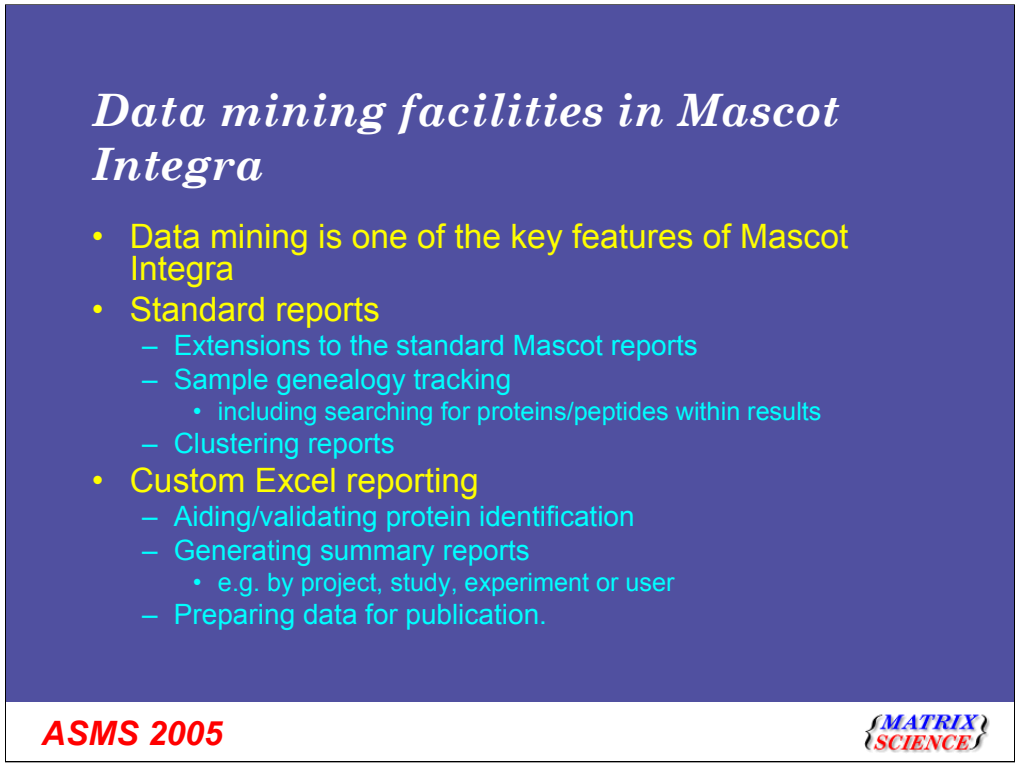
- 3 tier system
  - Oracle database server
  - Sybase Enterprise Application Server running a J2EE web application
  - All user functionality available through Internet Explorer



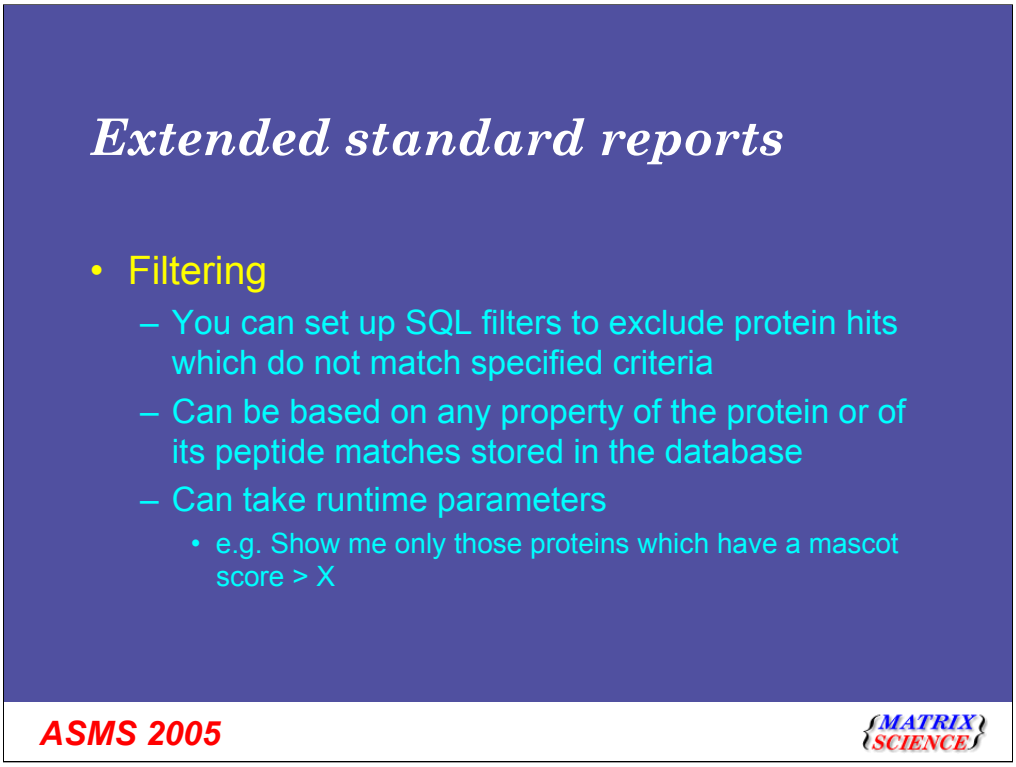
ASMS 2005

MATRIX SCIENCE

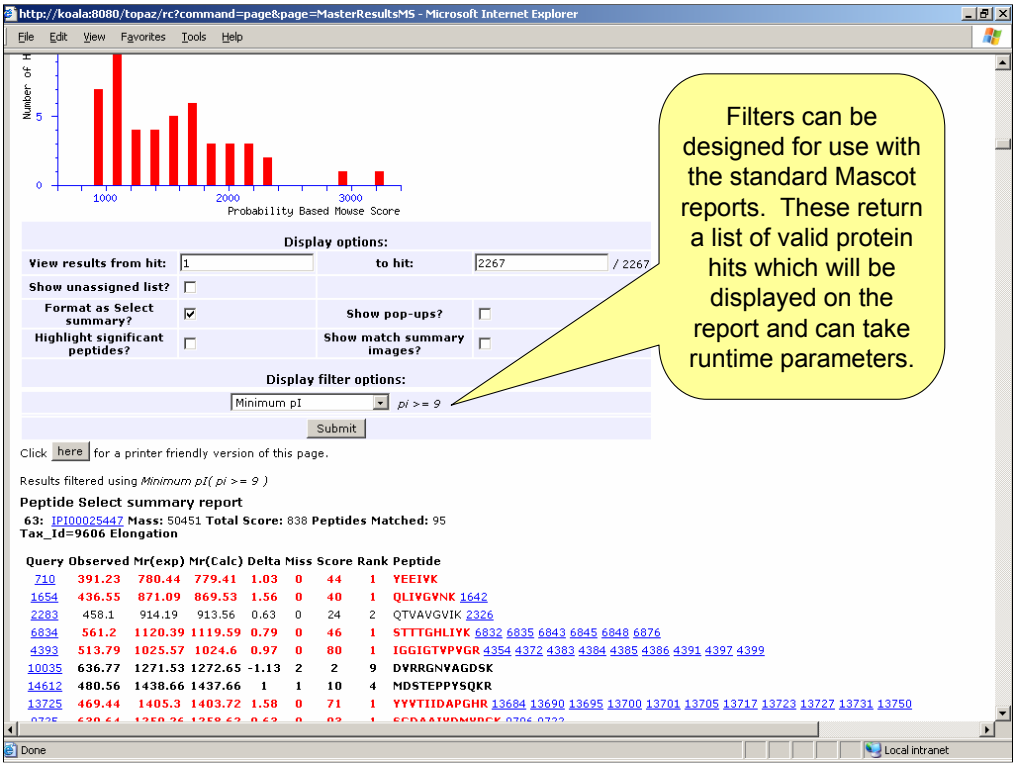
All Mascot Integra functionality is accessible through a standard web browser.



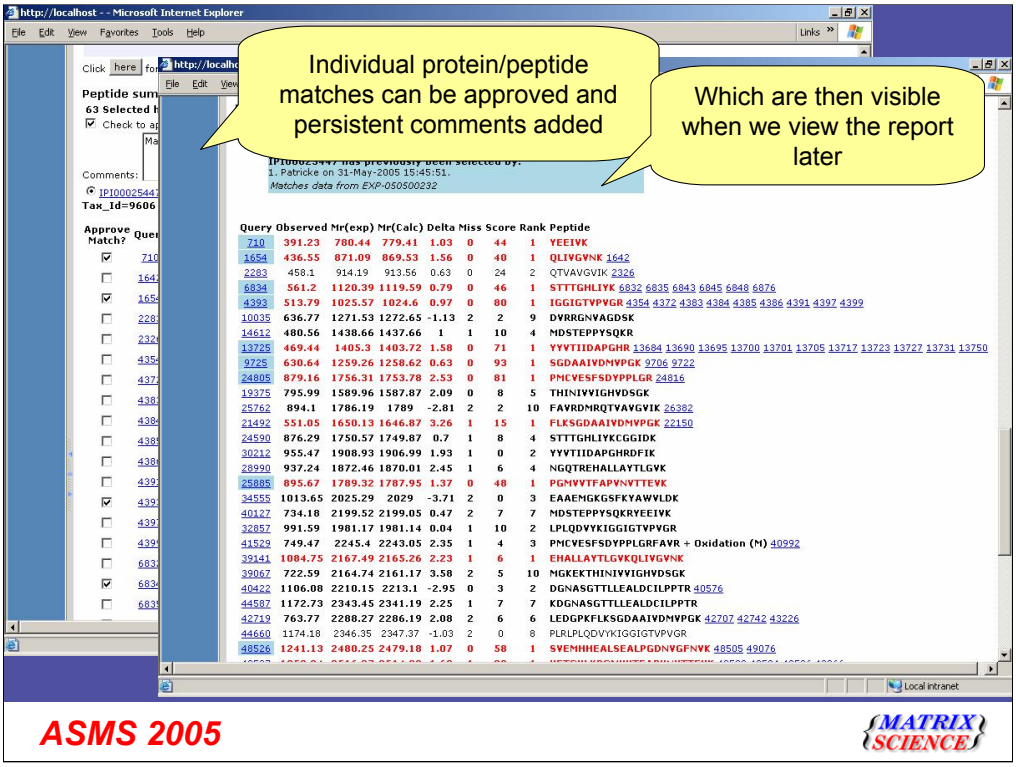
The 'raw' Mascot search results are imported into the Integra database. The schema of the database has been designed to facilitate data mining. This allows us to offer extensions to the 'Standard' Mascot reports, track sample history, offer the new clustering report and facilitates flexible ad-hoc querying of the database to produce custom reports using Microsoft Excel.



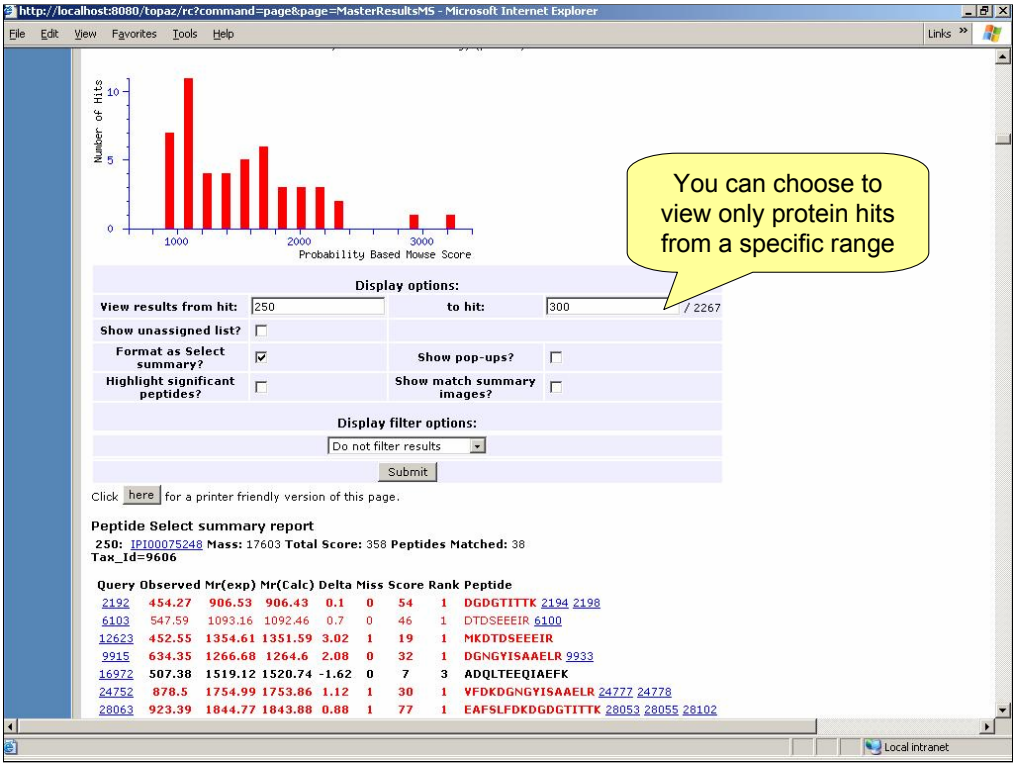
Filtering reports enables you to view only protein hits which match a range of specified criteria. They are very flexible and can be based on any property of the protein or its peptide matches which are stored within the database and can take runtime parameters, enabling the end user to specify the exact conditions after the filter has been designed - knowledge of the SQL query language is required to design the filters, but not to use them. For example, we could set up a filter which excludes all the protein hits from the report which do not contain at least one peptide match which is predicted to be phosphorylated.



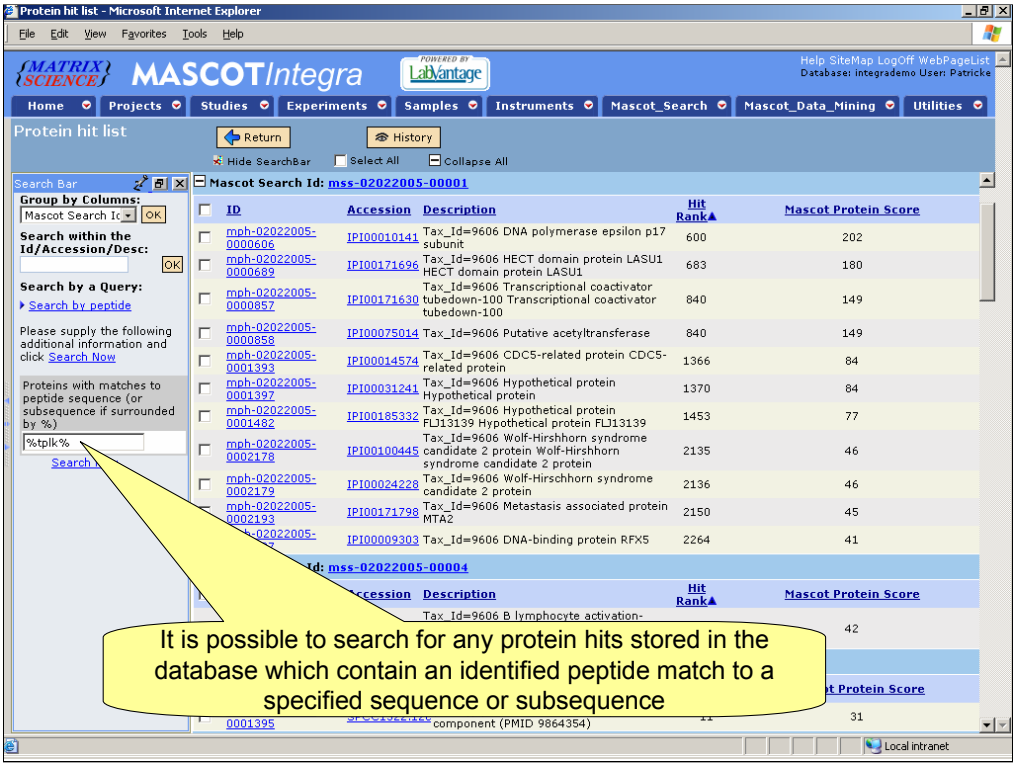
This report is from a MudPIT run and contains 2267 protein hits – a large number of results to look through. However, we may have some a priori knowledge which we can apply to the situation – for example, we may know that there is a protein of interest in the source mixture with a pI > 9. If we design a filter which only displays protein hits with a pI value of 9 or above, then this returns 474 of those 2267 hits – the 1st of which is hit number 63



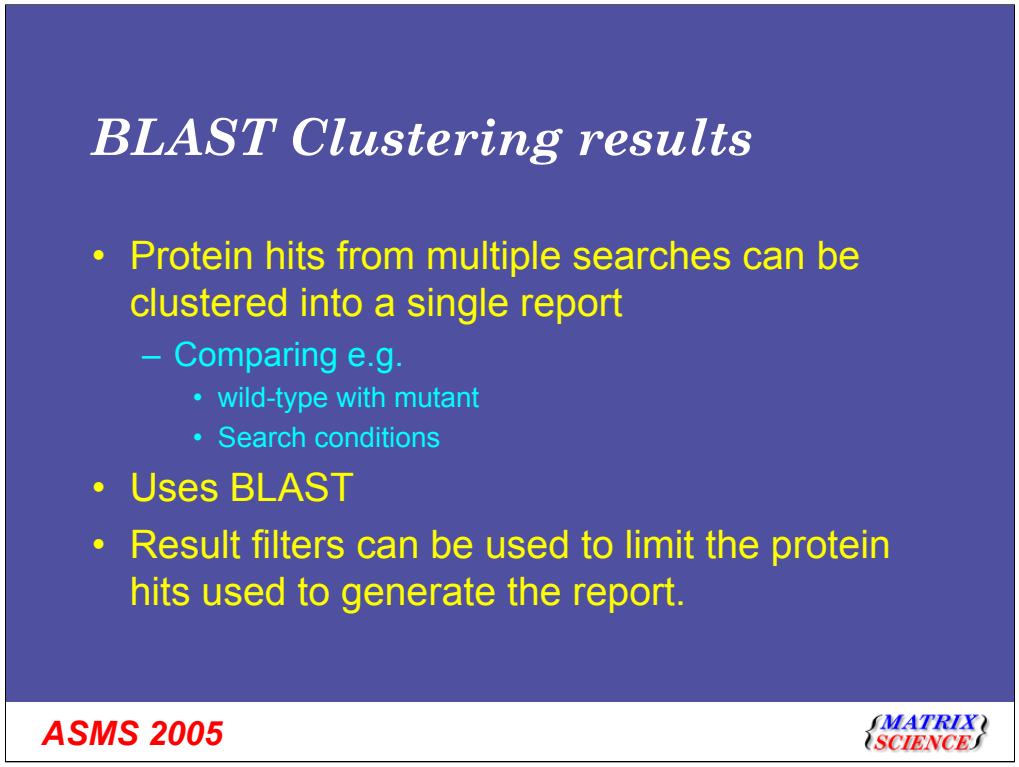
All of the standard Mascot reports can be generated from the database. In addition, individual protein and peptide matches can be annotated and approved. These annotations are persistent and will be displayed on the report when viewed at a later date. A protein/peptide match can be annotated/approved as many times as you wish, so that additional notes and corrections can be added.



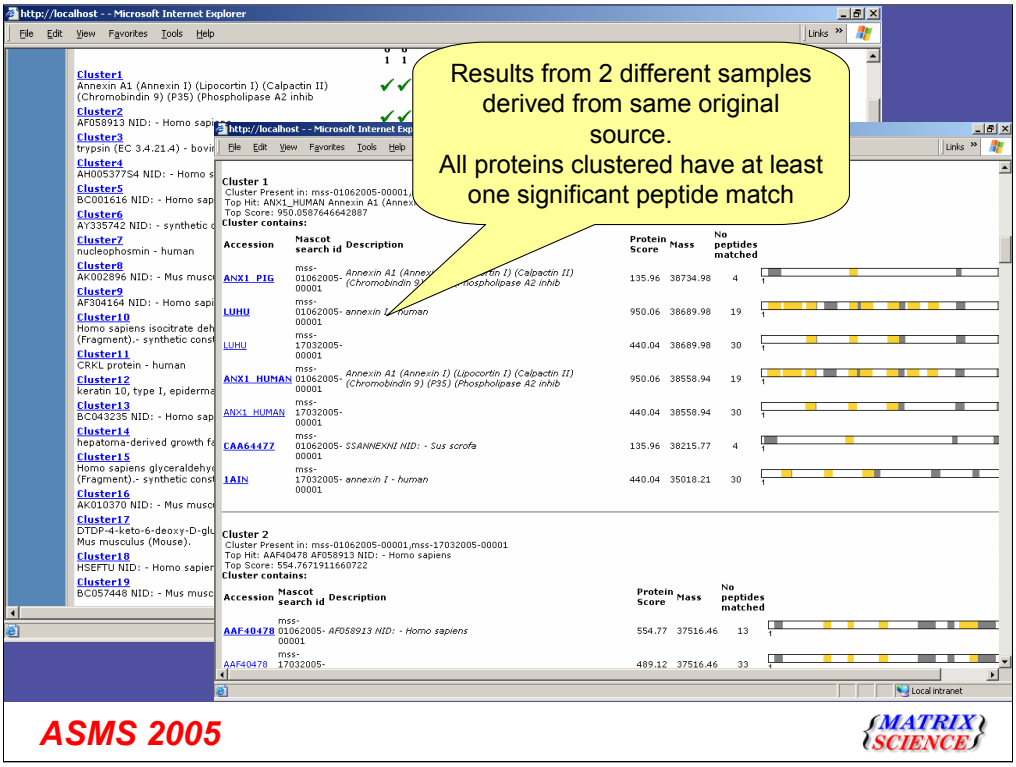
Because the results hare held in the database, you can choose to view just a specified range of search results, rather than displaying all of the protein hits in the report. This speeds up report generation and reduces problems with Internet Explorer opening very large results files, and also helps with working systematically through a report.



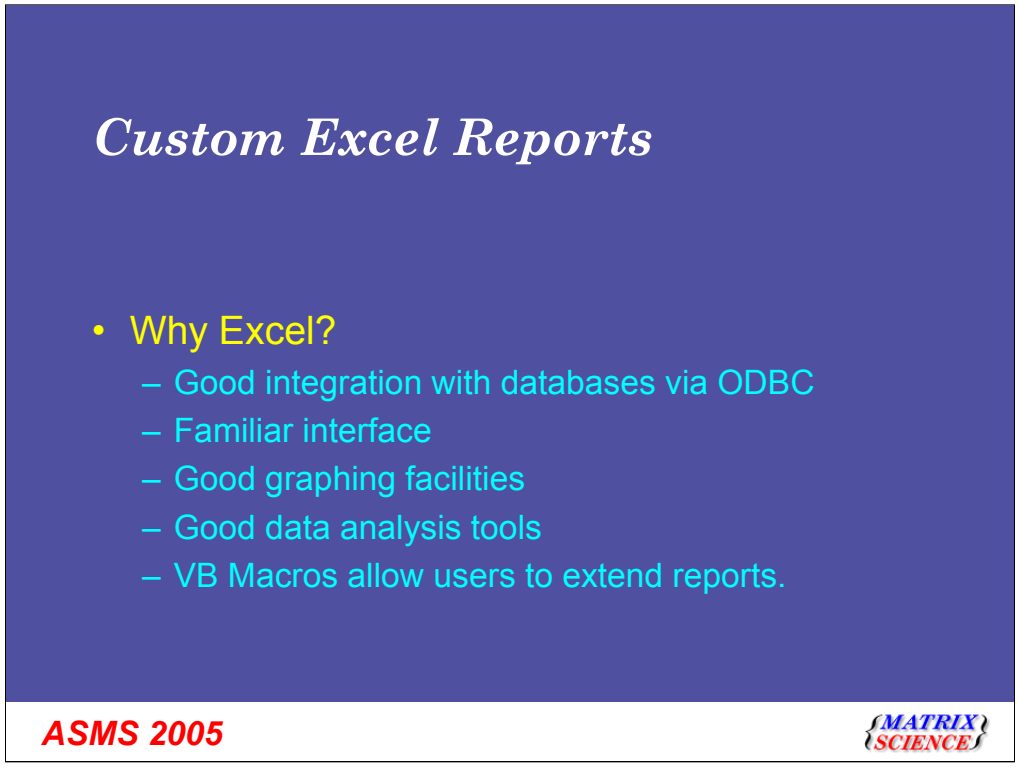
We can search for any accession or anything within the protein hit description. In addition to this, it is possible to search for any protein hits stored in the database which have an identified Mascot peptide match to a specified peptide sequence or subsequence. Here, we are looking for any protein hits in the database which have a peptide match containing the subsequence TPLK – one of the recognition sites for the p34cdc2 cell-cycle regulating kinase. As we can see, there are matches to several proteins which may be involved in the Cell-cycle (e.g. bub1, IPI00010141)



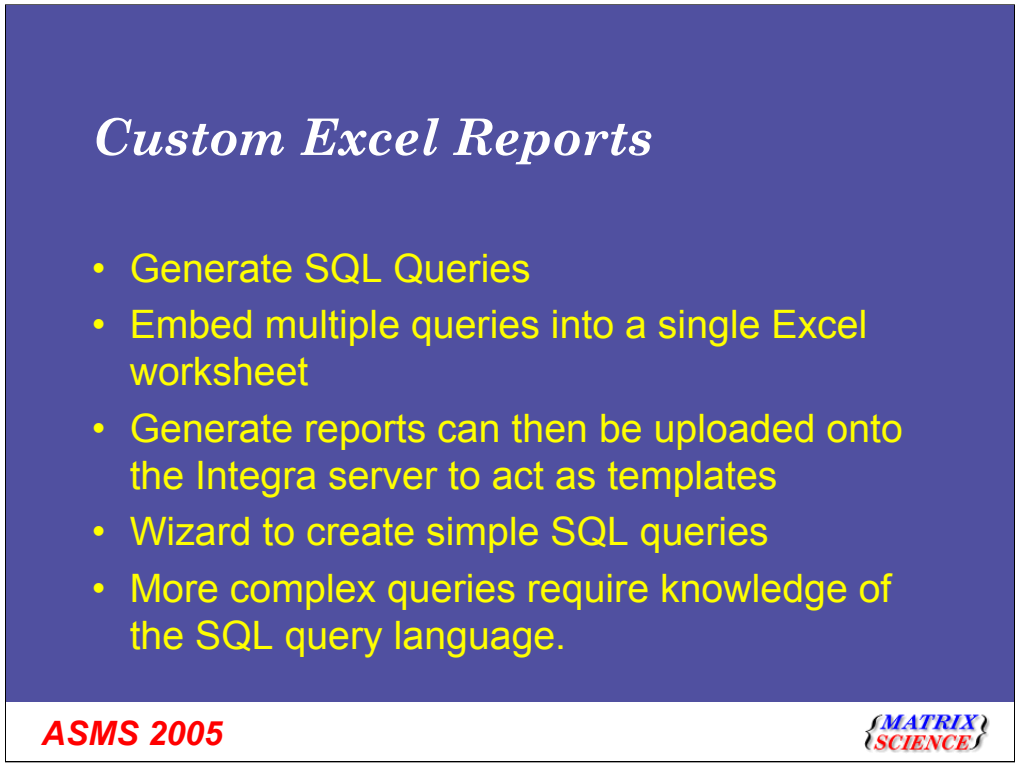
Another report groups protein hits from multiple reports, to allow comparison of the proteins present between the reports. This uses BLASTClust from NCBI and so uses the whole protein sequence to generate the clusters, not the peptide matches shared between the hits and not the protein accession – homologous proteins will appear in the same cluster. We can filter the proteins present in the report using the same filters which can be applied to the standard reports, and we can also provide protein sequences which we wish to exclude any matches to (e.g. Trypsin, Keratin).



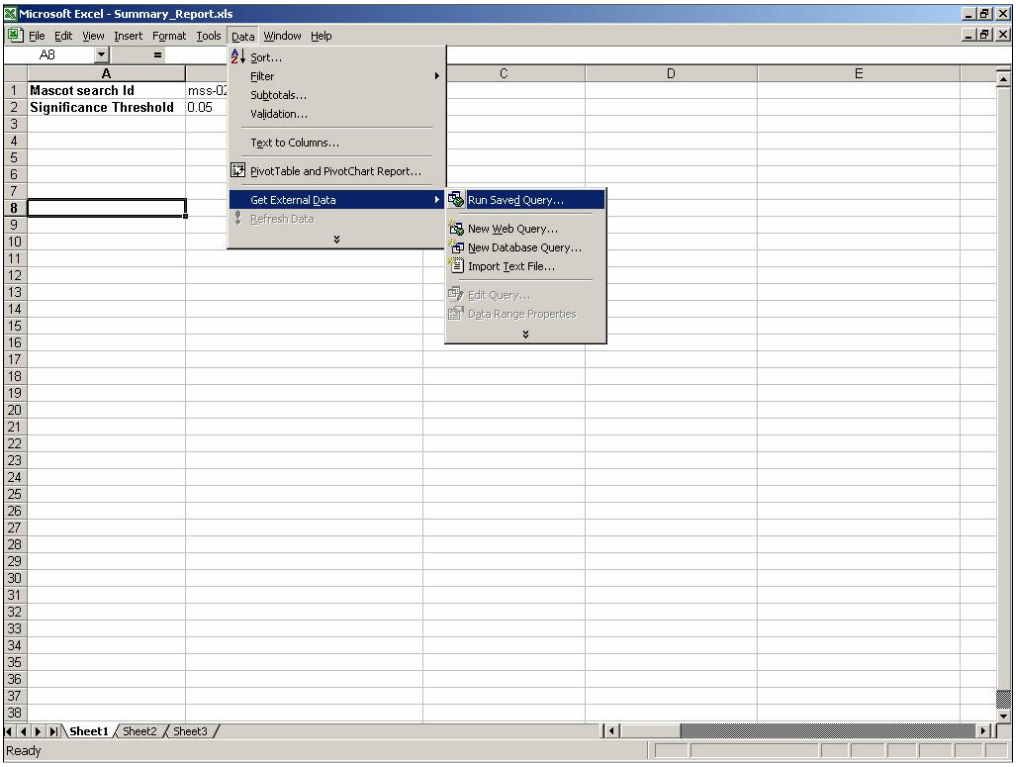
The report shows which clusters are present in the searches. We can then take a closer look at the proteins present in the clusters. In the images on the right, yellow represents significant peptide matches (0.5%)



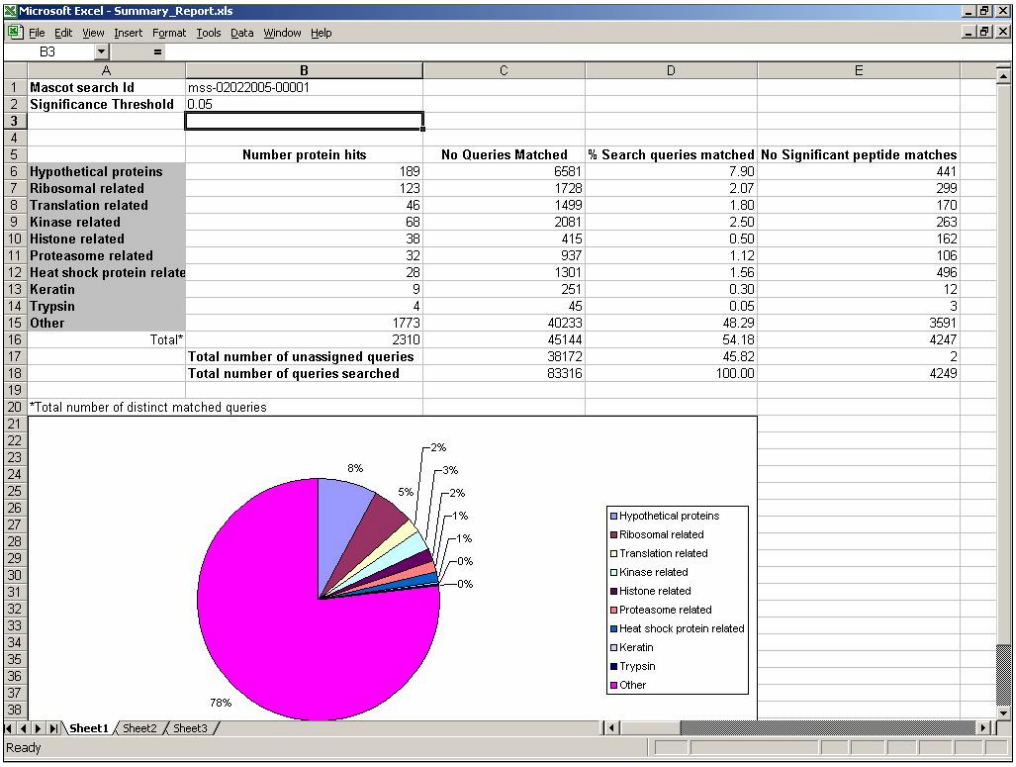
One of the main advantages of holding the raw mascot search results in a database is the ability to do ad hoc querying and generate custom reports. The database schema allows searches to be grouped by experiment, study, project or across the whole database, enabling complex cross search queries to be generated easily. The interface we have chosen to use to generate custom reports is Microsoft Excel.



To generate an Excel report requires knowledge of the SQL Query language and knowledge of how to get the best out of Excel. However, once the lab expert has designed an Excel report, it can be uploaded into Mascot Integra as a report template. Then the individual users can downloaded the report to use for their own search results/experiments/studies/projects.



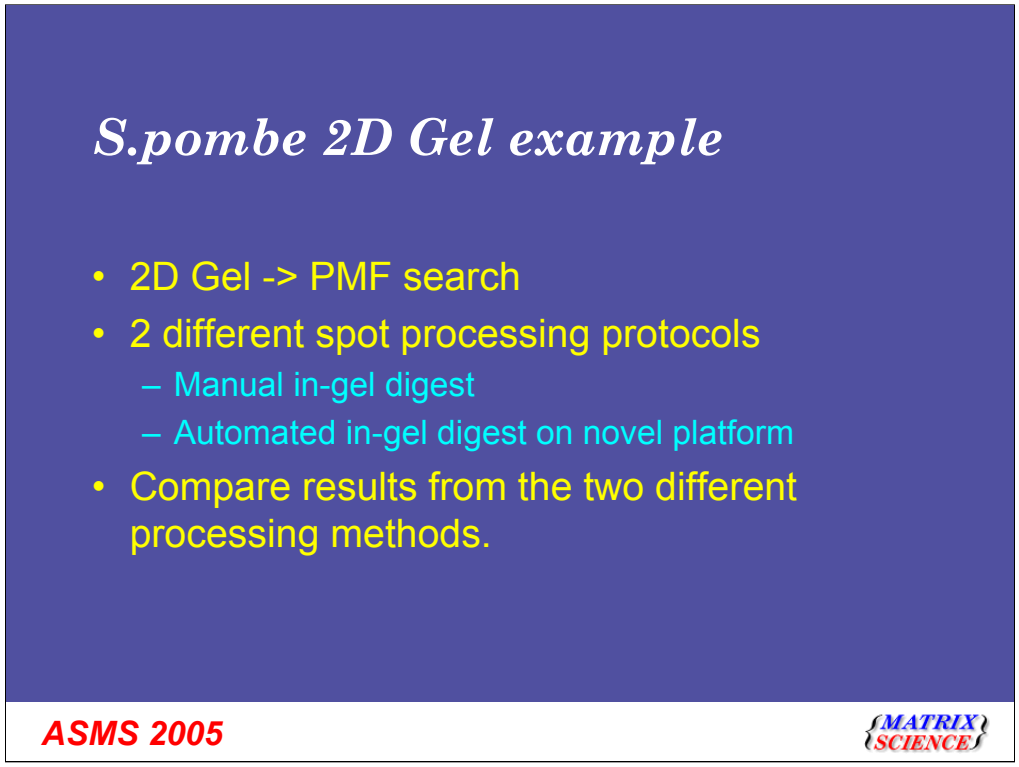
Here we have generated a query which generates summary data for a specified search using a specified peptide significance threshold. After setting up cells in the Excel worksheet which the query will take these values from, you import the Saved Query.



	A	B	C	D	E
1	**Mascot search Id**	mss-02022005-00001			
2	**Significance Threshold**	0.05			
3					
4					
5		**Number protein hits**	**No Queries Matched**	**% Search queries matched**	**No Significant peptide matches**
6	**Hypothetical proteins**	189	6581	7.90	441
7	**Ribosomal related**	123	1728	2.07	299
8	**Translation related**	46	1499	1.80	170
9	**Kinase related**	68	2081	2.50	263
10	**Histone related**	38	415	0.50	162
11	**Proteasome related**	32	937	1.12	106
12	**Heat shock protein relate**	28	1301	1.56	496
13	**Keratin**	9	251	0.30	12
14	**Trypsin**	4	45	0.05	3
15	**Other**	1773	40233	48.29	3591
16	Total*	2310	45144	54.18	4247
17		**Total number of unassigned queries**	38172	45.82	2
18		**Total number of queries searched**	83316	100.00	4249
19					
20	*Total number of distinct matched queries				

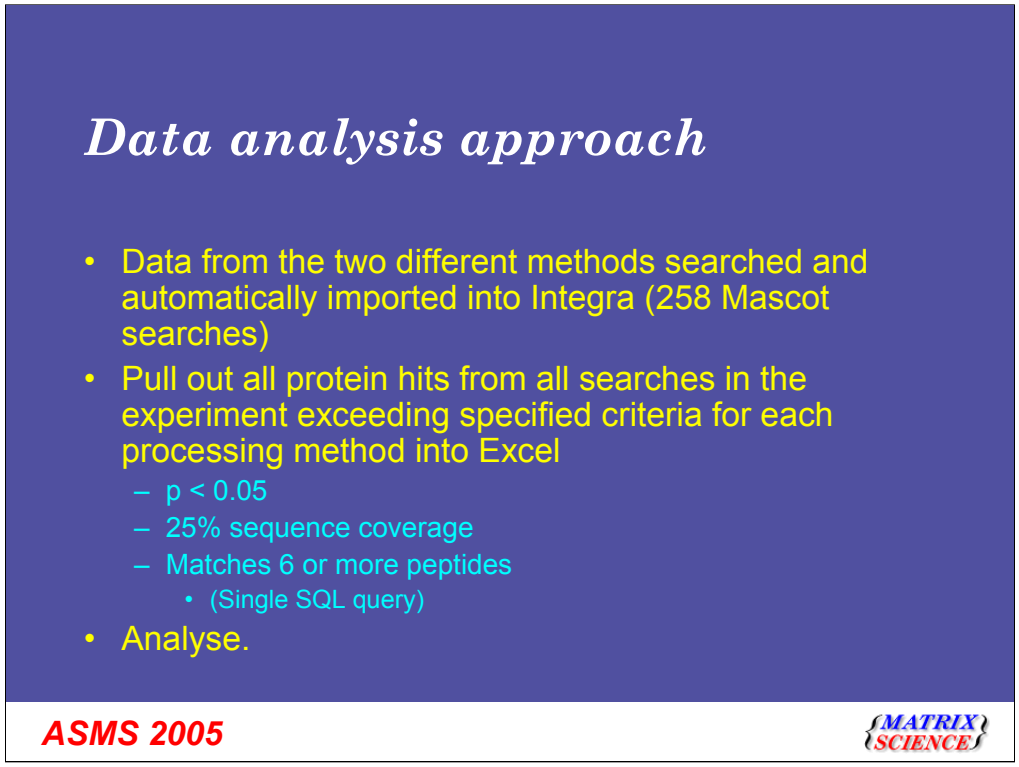
After some formatting....

This report could then be uploaded onto Integra as a template. When a user comes to download the template they will be prompted for the Mascot search Id and Significance Threshold they wish to use for their report, and the downloaded report will be based on the new values.



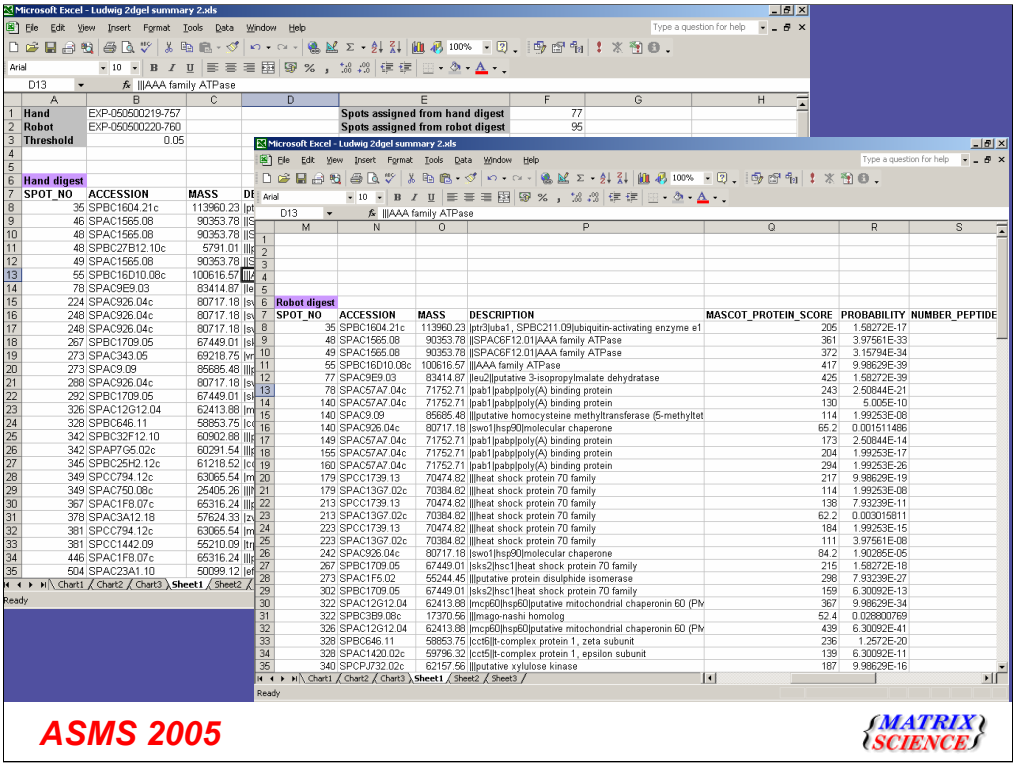
We'll take a closer look at how we can use these data mining tools to generate a report in a 'real world' example.

Running 2D Gel analysis from S.pombe protein extracts and then comparing two different gel spot processing protocols (manual or automated in-gel digest). We then want to compare the results obtained from the two methods to see if there are any differences between them

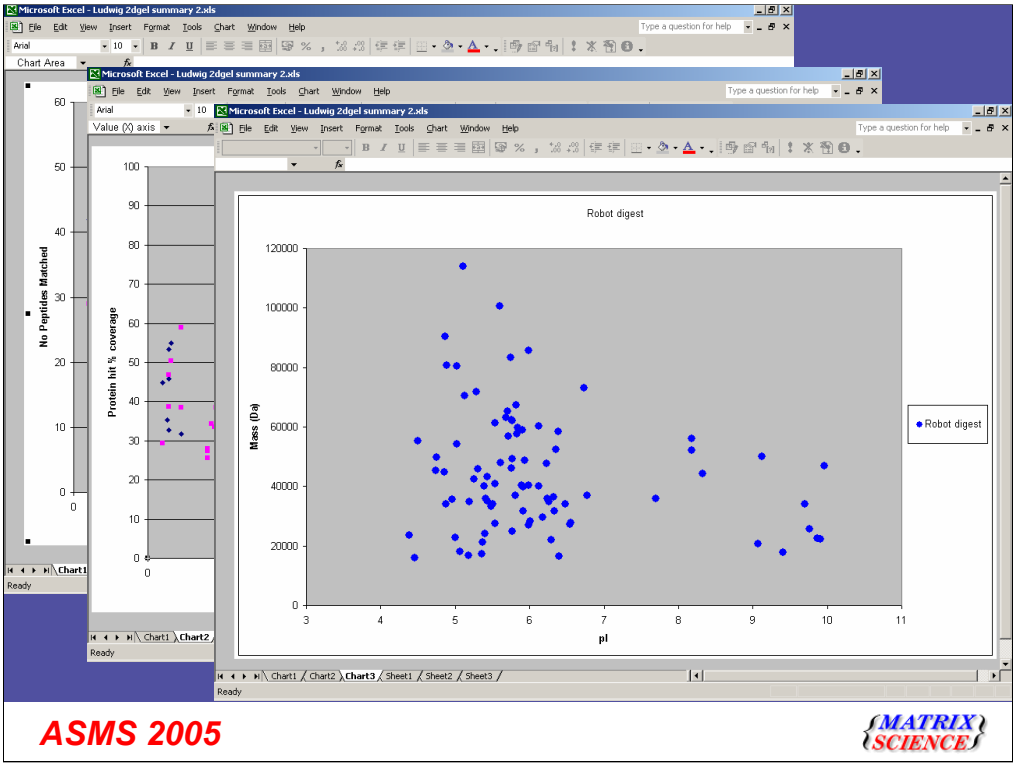


After the result are imported into the database we can generate an sql query to pull out the protein hit details from the two sets of searches (manual and automatic). The criteria (as are above) and can be specified within a single, simple SQL query against the protein hit table of the database.

We will then do some analysis of these data in Excel

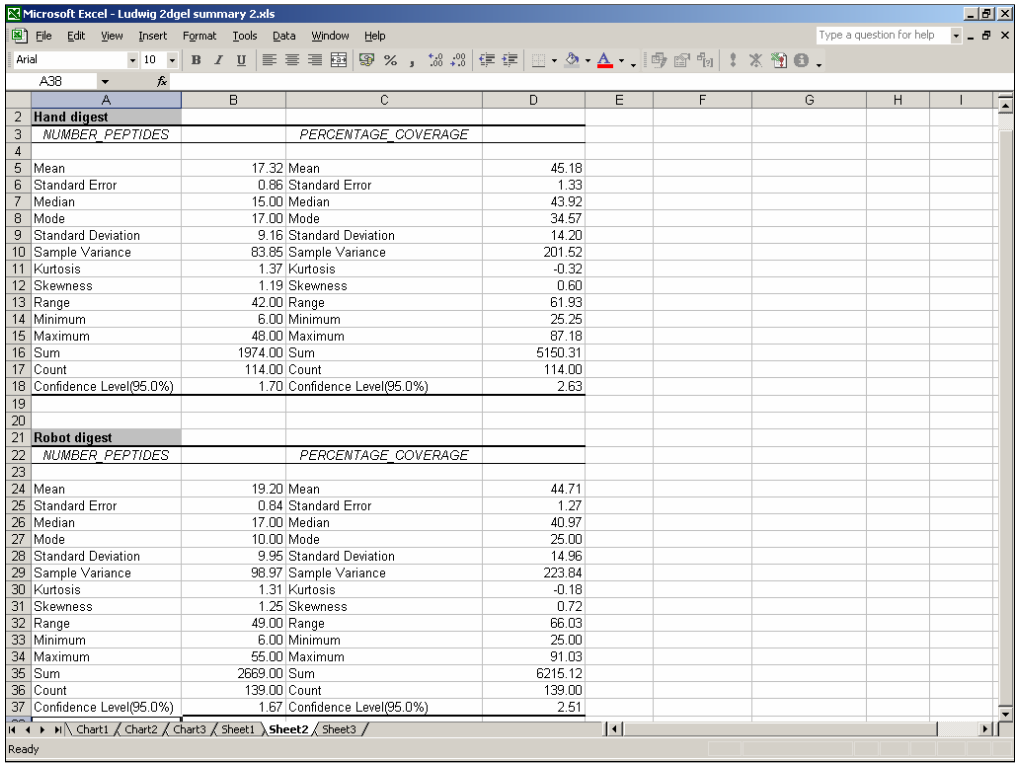


Using the specified criteria, we can see that 95/110 spots have potential protein assignments from the automated digest method, compared with 77 from the manual in gel digest.



We've pulled out a lot of data relating to the quality of these matches from the database. Using Excels graphing tools we can take a closer look at the results from the two datasets to see if there are any overall differences in the data quality:

Overall quality of the data is similar – The distribution of no peptides matches and % coverage being similar for both the hand and robot (automated) datasets (the association with spot no was also expected as the lower mass proteins have been assigned the higher spot numbers). We can also plot the protein mass against the pI value for the potentially assigned protein hits to compare this with the source 2D gel information.



Microsoft Excel - Ludwig 2dgel summary 2.xls

File Edit View Insert Format Tools Data Window Help

Type a question for help

Arial 10

A38

	A	B	C	D	E	F	G	H	I
2	**Hand digest**								
3	*NUMBER_PEPTIDES*		*PERCENTAGE_COVERAGE*						
4									
5	Mean	17.32	Mean	45.18					
6	Standard Error	0.86	Standard Error	1.33					
7	Median	15.00	Median	43.92					
8	Mode	17.00	Mode	34.57					
9	Standard Deviation	9.16	Standard Deviation	14.20					
10	Sample Variance	83.85	Sample Variance	201.52					
11	Kurtosis	1.37	Kurtosis	-0.32					
12	Skewness	1.19	Skewness	0.60					
13	Range	42.00	Range	61.93					
14	Minimum	6.00	Minimum	25.25					
15	Maximum	48.00	Maximum	87.18					
16	Sum	1974.00	Sum	5150.31					
17	Count	114.00	Count	114.00					
18	Confidence Level(95.0%)	1.70	Confidence Level(95.0%)	2.63					
19									
20									
21	**Robot digest**								
22	*NUMBER_PEPTIDES*		*PERCENTAGE_COVERAGE*						
23									
24	Mean	19.20	Mean	44.71					
25	Standard Error	0.84	Standard Error	1.27					
26	Median	17.00	Median	40.97					
27	Mode	10.00	Mode	25.00					
28	Standard Deviation	9.95	Standard Deviation	14.96					
29	Sample Variance	98.97	Sample Variance	223.84					
30	Kurtosis	1.31	Kurtosis	-0.18					
31	Skewness	1.25	Skewness	0.72					
32	Range	49.00	Range	66.03					
33	Minimum	6.00	Minimum	25.00					
34	Maximum	55.00	Maximum	91.03					
35	Sum	2669.00	Sum	6215.12					
36	Count	139.00	Count	139.00					
37	Confidence Level(95.0%)	1.67	Confidence Level(95.0%)	2.51					

Chart1 Chart2 Chart3 Sheet1 Sheet2 Sheet3

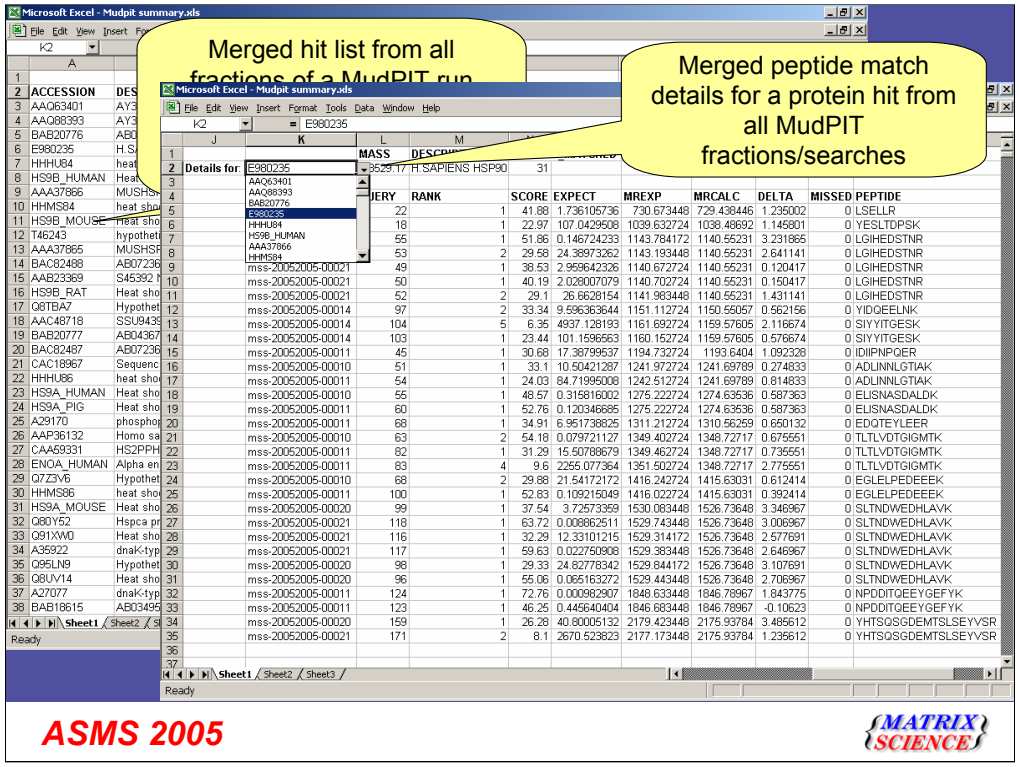
Ready

It is very easy to generate summary statistics using standard Excel features – again similar results for the number of peptides matches and percentage coverage from the two methods.

S.pombe conclusions

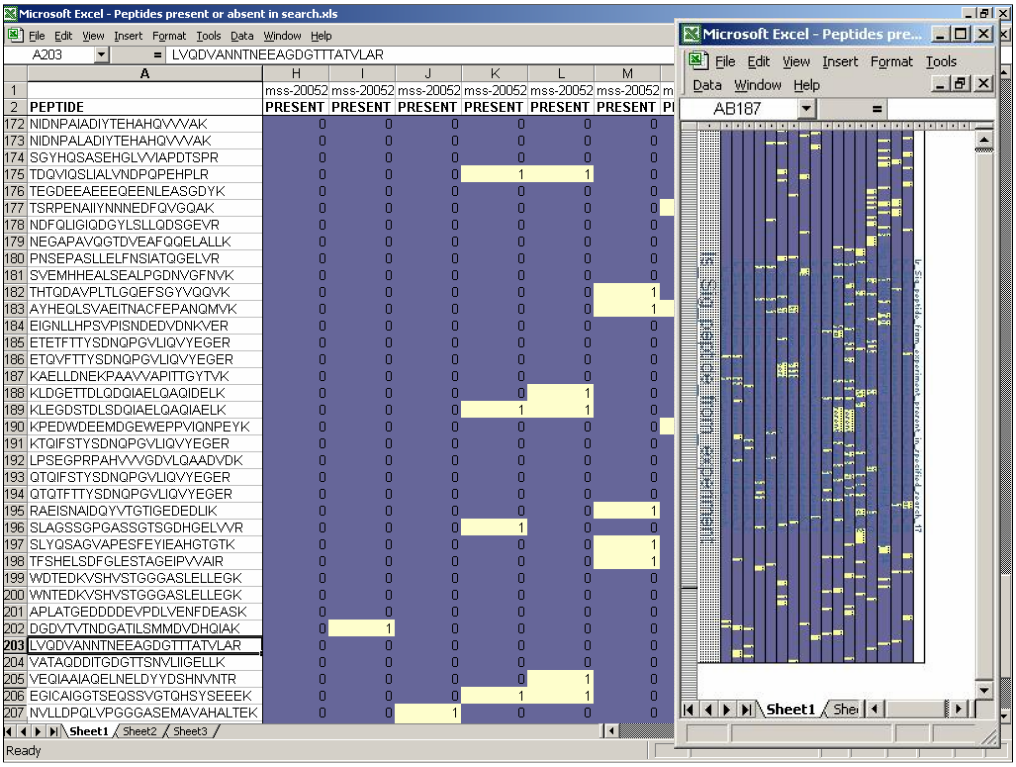
- Where we have matches, the data in both datasets is of similarly high quality
- Robot dataset identified matches for more spots within the specified criteria
  - Extraction quality was more consistent
- Use of EXCEL reports allows us to query and present these data quickly and easily.

Some other examples...

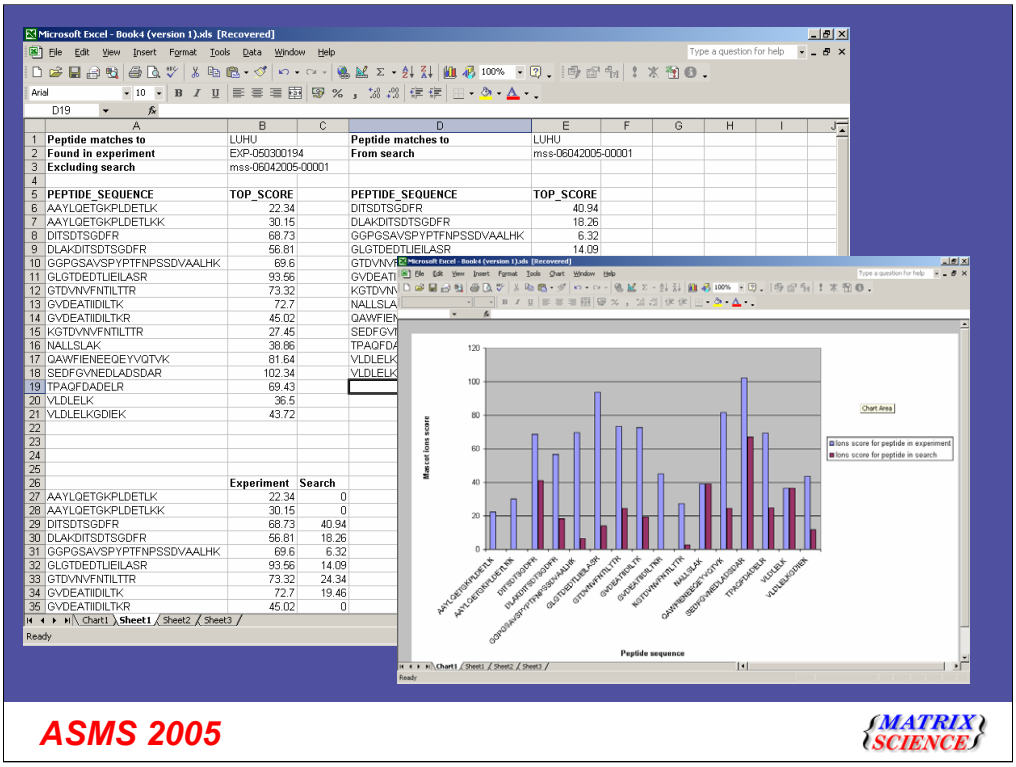


Some additional examples of the types of reports it is easy to generate from Integra but very hard to produce from the standard reports....

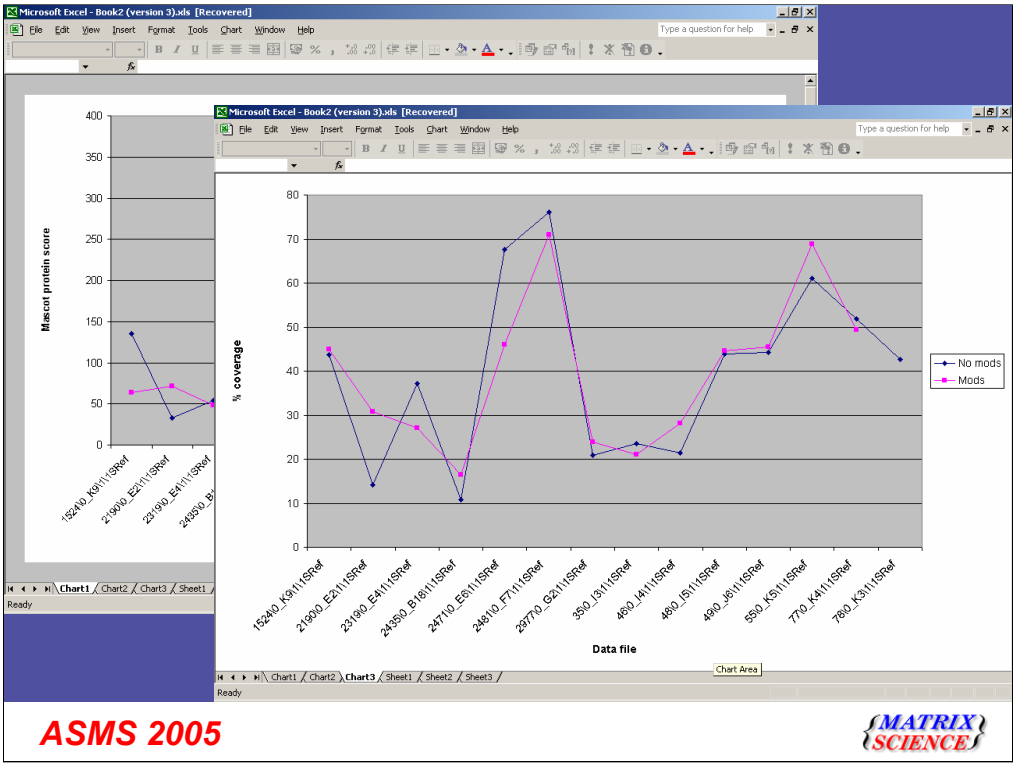
Here we have combined search results from 17 fractions of a MudPIT run, generating a merged hit list. We can also generate a merged peptide match list for each protein hit identified, combining the peptide matches to the protein from all of the Mascot searches (and hence source MudPIT fractions).



So we can check the quality of our 1st dimension chromatography…this generated from the same 17 MudPIT fractions. All of the peptides with a e value below the 0.05 threshold from all of the fractions have been identified and then cross checked against the search results from each fraction to see if the peptide is present (the yellow background shows the peptide is present, blue means it was absent from the search) – we can see that many of the peptides are present across multiple (usually adjacent) fractions.



A protein hit we're not sure of – have we identified it elsewhere in the same experiment? and if so, what peptides did we match. In this experiment we have used the same source data, searched with different search parameters. The hit LUHU (an annexin) from a particular search is of interest but the scores are borderline. We can see from the other searches in the experiment that we have previously matched this hit and obtained a similar range of peptides.



Comparison between two PMF search strategies for a series of datafiles, then looking at the Mascot protein score and % coverage from the two strategies for the top hit for each source file.

Whatever Mods were picked they overall didn't help, except possibly for 2190\0_E2 which might warrant further investigation (sig threshold 55)

Missing final point for 78/0_K3 for % coverage is because a possible PMF mixture was picked up by the Mods search. One of the proteins in the mixture was the same as that picked up by the No mods search.

Summary

- 'Raw' search results are imported into RDB tables in a logical structure, organised by Project, Study and Experimental relationships
  - Extended reports
    - Hit approval/persistent comments
    - Filtered reports
  - Data mining
    - Flexible ad hoc query
    - Excel reporting

Some knowledge of SQL required to generate the custom reports and filters for the standard reports. However, these only need to be done once and can then be used as templates by any other user.

Acknowledgements

- Mark Weeks, John Sinclair, Richard Jacob
  - Ludwig Institute for Cancer Research, London UK