

# *Data mining with Mascot Integra*

**ASMS 2005**

**{MATRIX}  
{SCIENCE}**

## *What is Mascot Integra?*

- Fully functional 'out-the-box' solution for proteomics workflow and data management
- Support for all the major mass-spectrometry data systems
- Powered by the Sapphire™ LIMS package from LabVantage Solutions Inc
- Oracle database
- Scalable to the largest projects.

**ASMS 2005**

**MATRIX  
SCIENCE**

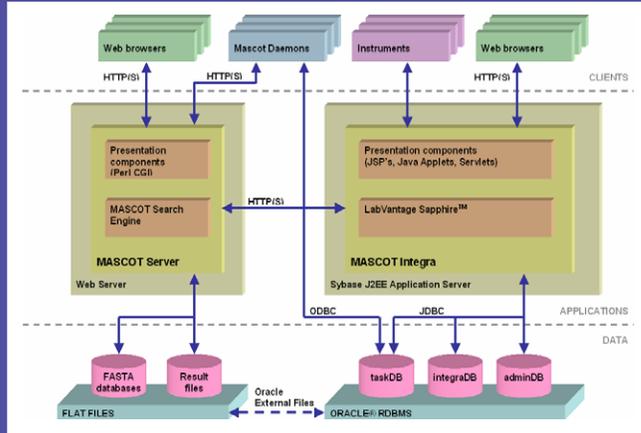
Mascot Integra is supplied as a ready to run system. It does not require the extensive setup and customisation associated with a traditional LIMS package.

Rather than re-invent the wheel, we have partnered with LabVantage Solutions Inc, ([www.lims.com](http://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



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

## *Data mining facilities in Mascot Integra*

- Data mining is one of the key features of Mascot Integra
- Standard reports
  - Extensions to the standard Mascot reports
  - Sample genealogy tracking
    - including searching for proteins/peptides within results
  - Clustering reports
- Custom Excel reporting
  - Aiding/validating protein identification
  - Generating summary reports
    - e.g. by project, study, experiment or user
  - Preparing data for publication.

**ASMS 2005**

**MATRIX  
SCIENCE**

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.

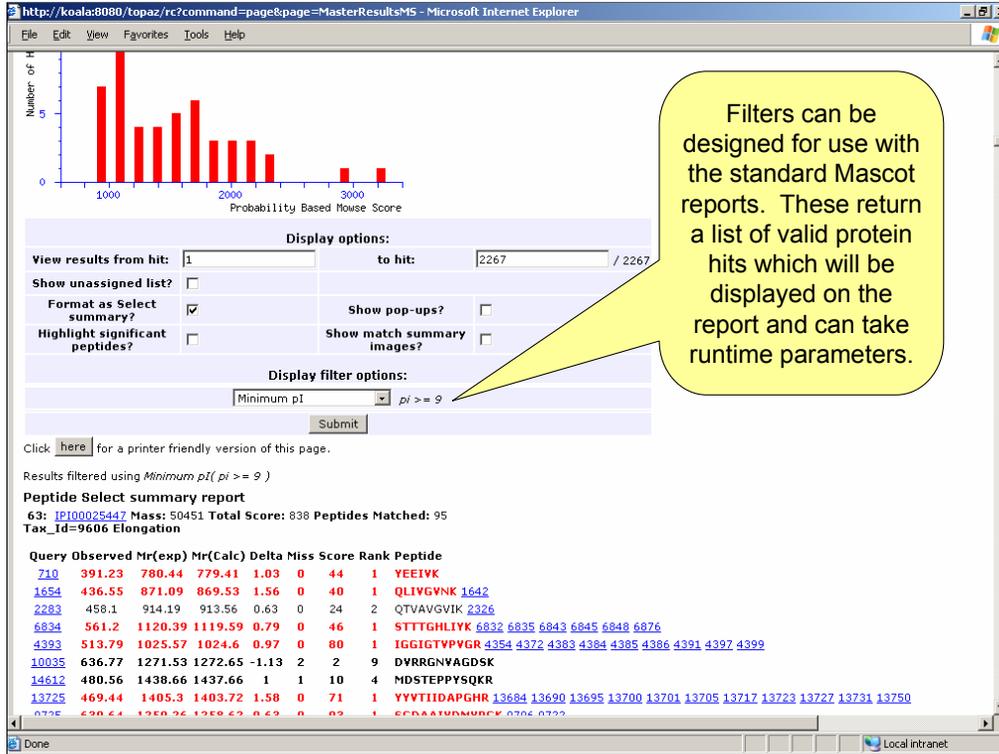
## *Extended standard reports*

- **Filtering**
  - You can set up SQL filters to exclude protein hits which do not match specified criteria
  - Can be based on any property of the protein or of its peptide matches stored in the database
  - Can take runtime parameters
    - e.g. Show me only those proteins which have a mascot score > X

**ASMS 2005**

**MATRIX  
SCIENCE**

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 value of 9 or above. 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 1<sup>st</sup> of which is hit number 63

Individual protein/peptide matches can be approved and persistent comments added

Which are then visible when we view the report later

Query	Observed Mr	Mr(calc)	Delta	Miss	Score	Rank	Peptide	
710	391.23	780.44	779.41	1.03	0	44	1 YEEIVK	
1654	436.55	871.09	869.53	1.56	0	40	1 QLIVGVNK 1642	
2283	458.1	914.19	913.56	0.63	0	24	2 QTVAVGVNK 2326	
6834	561.2	1120.39	1119.59	0.79	0	46	1 STTTGHLIFK 6832 6835 6843 6845 6848 6876	
4393	513.79	1025.57	1024.6	0.97	0	80	1 IGGIGTVPVGR 4354 4372 4383 4384 4385 4386 4391 4397 4399	
10035	636.77	1271.53	1272.65	-1.13	2	2	9 DYRRGNVAGDSK	
14612	480.56	1438.66	1437.66	1	1	10	4 MDSTPPYSQKR	
13725	469.44	1405.3	1403.72	1.58	0	71	1 YYYTIIDAPGHR 13684 13690 13695 13700 13701 13705 13717 13723 13727 13731 13750	
9725	630.64	1259.26	1258.62	0.63	0	93	1 SGDAALVDMVPGK 9706 9722	
24805	879.16	1756.31	1753.78	2.53	0	81	1 PNCVFSFDYPLGK 24816	
19375	795.99	1589.96	1587.87	2.09	0	8	5 THINIVIGHVDSGK	
25762	894.1	1786.19	1789	-2.81	2	2	10 FAVRDMRQTVAGVIK 26382	
4388	21492	551.05	1650.13	1646.87	3.26	1	15 1 FLKSGDAALVDMVPGK 22150	
24590	876.29	1750.57	1749.87	0.7	1	8	4 STTTGHLIYKCGGIDK	
4388	30212	955.47	1908.93	1906.99	1.93	1	0	2 YYYTIIDAPGHRDFIK
4388	28990	937.24	1872.46	1870.01	2.45	1	6	4 NGQTRHALLAYTLGVK
4399	25885	895.67	1789.32	1787.95	1.37	0	48	1 PGNVYTFAPNVYTLGVK
4399	34555	1013.65	2025.29	2029	-3.71	2	0	3 EAAEMGKGFYAVVLDK
4399	40127	734.18	2199.52	2199.05	0.47	2	7	7 MDSTPPYSQKRYEEIVK
4399	32852	991.59	1981.17	1981.14	0.04	1	10	2 LPLQVYKIGGIGTVPVGR
4399	41523	749.47	2245.4	2243.05	2.35	1	4	3 PNCVFSFDYPLGKFAVR + Oxidation (M) 40992
6833	32141	1064.75	2167.49	2165.26	2.23	1	6	1 EHALLATLVKQLIVGVNK
6833	32062	722.59	2164.74	2161.17	3.58	2	5	10 NGKEKTHINIVIGHVDSGK
6833	40422	1106.08	2210.15	2213.1	-2.95	0	3	2 DGNASGTTLEALDCILPPTK 40576
6833	44592	1172.73	2343.45	2341.19	2.25	1	7	7 KDNASGTTLEALDCILPPTK
42719	763.77	2288.27	2286.19	2.08	2	6	6 LEDGPKFLKSGDAALVDMVPGK 42707 42742 43226	
44660	1174.18	2346.35	2347.37	-1.03	2	0	8	8 PRLRPLQVYKIGGIGTVPVGR
48526	1241.13	2480.25	2479.18	1.07	0	58	1 SVENHHEALPALDGNVGFNVK 48505 49076	

ASMS 2005

MATRIX SCIENCE

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.

Number of Hits

Probability Based House Score

Display options:

View results from hit: 250 to hit: 300 / 2267

Show unassigned list?

Format as Select summary?

Highlight significant peptides?

Show pop-ups?

Show match summary images?

Display filter options:

Do not filter results

Submit

Click [here](#) for a printer friendly version of this page.

**Peptide Select summary report**  
 250: [IP100075248](#) Mass: 17603 Total Score: 358 Peptides Matched: 38  
 Tax\_id=9606

Query	Observed Mr(exp)	Mr(Calc)	Delta	Miss	Score	Rank	Peptide
<a href="#">2192</a>	454.27	906.53	906.43	0.1	0	54	1 DGGDTITTK <a href="#">2194</a> <a href="#">2198</a>
<a href="#">6103</a>	547.59	1093.16	1092.46	0.7	0	46	1 DTDSEEEIR <a href="#">6100</a>
<a href="#">12623</a>	452.55	1354.61	1351.59	3.02	1	19	1 MKDTSSEEIR
<a href="#">9915</a>	634.35	1266.68	1264.6	2.08	0	32	1 DNGYISAAELR <a href="#">9933</a>
<a href="#">16972</a>	507.38	1519.12	1520.74	-1.62	0	7	3 ADQLTEEQIAEFK
<a href="#">24752</a>	878.5	1754.99	1753.86	1.12	1	30	1 VFDKDNGYISAAELR <a href="#">24777</a> <a href="#">24778</a>
<a href="#">28063</a>	923.39	1844.77	1843.88	0.88	1	77	1 EAFSLFDKGGDTITTK <a href="#">28053</a> <a href="#">28055</a> <a href="#">28102</a>

Because the results here 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.

The screenshot shows the MascotIntegra web interface with a search for the peptide sequence '%tplk%'. The search results are displayed in a table with the following columns: ID, Accession, Description, Hit Rank, and Mascot Protein Score. The search criteria are specified in the left-hand panel under 'Search by a Query' and 'Search within the Id/Accession/Desc'.

ID	Accession	Description	Hit Rank	Mascot Protein Score
<input type="checkbox"/>	mph-02022005-000606	IPI00010141 Tax_Id=9606 DNA polymerase epsilon p17 subunit	600	202
<input type="checkbox"/>	mph-02022005-000689	IPI00171696 Tax_Id=9606 HECT domain protein LASU1 HECT domain protein LASU1	683	180
<input type="checkbox"/>	mph-02022005-000857	IPI00171630 Tax_Id=9606 Transcriptional coactivator tubedown-100 Transcriptional coactivator tubedown-100	840	149
<input type="checkbox"/>	mph-02022005-000858	IPI00075014 Tax_Id=9606 Putative acetyltransferase	840	149
<input type="checkbox"/>	mph-02022005-0001393	IPI00014574 Tax_Id=9606 CDC5-related protein CDC5-related protein	1366	84
<input type="checkbox"/>	mph-02022005-0001397	IPI00031241 Tax_Id=9606 Hypothetical protein Hypothetical protein	1370	84
<input type="checkbox"/>	mph-02022005-0001482	IPI00185336 Tax_Id=9606 Hypothetical protein FLJ13139 Hypothetical protein FLJ13139	1453	77
<input type="checkbox"/>	mph-02022005-0002178	IPI00100445 Tax_Id=9606 Wolf-Hirschhorn syndrome candidate 2 protein Wolf-Hirschhorn syndrome candidate 2 protein	2135	46
<input type="checkbox"/>	mph-02022005-0002179	IPI00024228 Tax_Id=9606 Wolf-Hirschhorn syndrome candidate 2 protein	2136	46
<input type="checkbox"/>	mph-02022005-0002193	IPI00171798 Tax_Id=9606 Metastasis associated protein MTA2	2150	45
<input type="checkbox"/>	mph-02022005-0001395	IPI00009303 Tax_Id=9606 DNA-binding protein RFX5	2264	41
<input type="checkbox"/>	mph-02022005-0001395	IPI00009303 Tax_Id=9606 B lymphocyte activation-		42
<input type="checkbox"/>	mph-02022005-0001395	IPI00009303 component (PMID 9864354)		31

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)

## *BLAST Clustering results*

- Protein hits from multiple searches can be clustered into a single report
  - Comparing e.g.
    - wild-type with mutant
    - Search conditions
- Uses BLAST
- Result filters can be used to limit the protein hits used to generate the report.

**ASMS 2005**

**MATRIX  
SCIENCE**

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).

Results from 2 different samples derived from same original source.  
All proteins clustered have at least one significant peptide match

Accession	Nascent search id	Description	Protein Score	Mass	No peptides matched
<b>Cluster 1</b> Cluster Present in: mss-01062005-00001 Top Hit: ANX1_HUMAN Annexin A1 (Annexin I) (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhib					
ANX1_PIG	mss-01062005-00001	Annexin A1 (Annexin I) (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhib	135.96	38734.98	4
LUHU	mss-01062005-00001	Annexin A1 (Annexin I) (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhib	950.06	38689.98	19
LUHU	mss-17032005-00001	Annexin A1 (Annexin I) (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhib	440.04	38689.98	30
ANX1_HUMAN	mss-01062005-00001	Annexin A1 (Annexin I) (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhib	950.06	38558.94	19
ANX1_HUMAN	mss-17032005-00001	Annexin A1 (Annexin I) (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhib	440.04	38558.94	30
CAA64477	mss-01062005-00001	Annexin A1 (Annexin I) (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhib	135.96	38215.77	4
IAIN	mss-17032005-00001	Annexin A1 (Annexin I) (Lipocortin I) (Calpactin II) (Chromobindin 9) (P35) (Phospholipase A2 inhib	440.04	35018.21	30
<b>Cluster 2</b> Cluster Present in: mss-01062005-00001,mss-17032005-00001 Top Hit: AAF40478 AF058913 NID: - Homo sapiens Top Score: 554.7671911660722					
AAF40478	mss-01062005-00001	AF058913 NID: - Homo sapiens	554.77	37516.46	13
AAF40478	mss-17032005-00001	AF058913 NID: - Homo sapiens	489.12	37516.46	33

**ASMS 2005**

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%)

## *Custom Excel Reports*

- **Why Excel?**
  - Good integration with databases via ODBC
  - Familiar interface
  - Good graphing facilities
  - Good data analysis tools
  - VB Macros allow users to extend reports.

**ASMS 2005**

**{MATRIX}**  
**{SCIENCE}**

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.

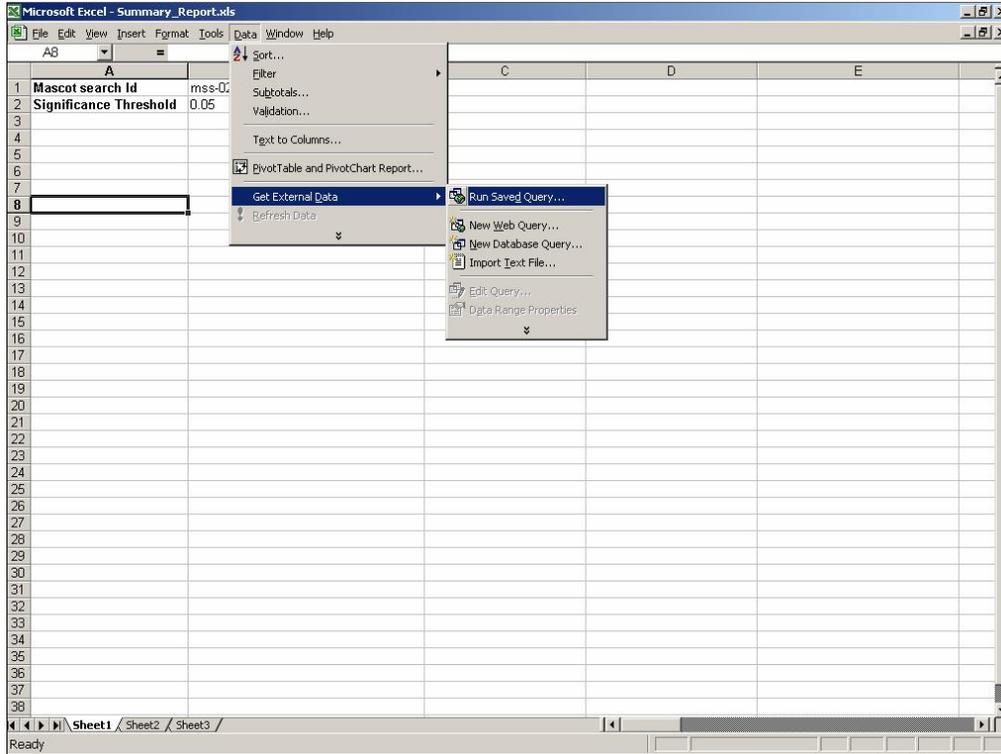
## *Custom Excel Reports*

- Generate SQL Queries
- Embed multiple queries into a single Excel worksheet
- Generate reports can then be uploaded onto the Integra server to act as templates
- Wizard to create simple SQL queries
- More complex queries require knowledge of the SQL query language.

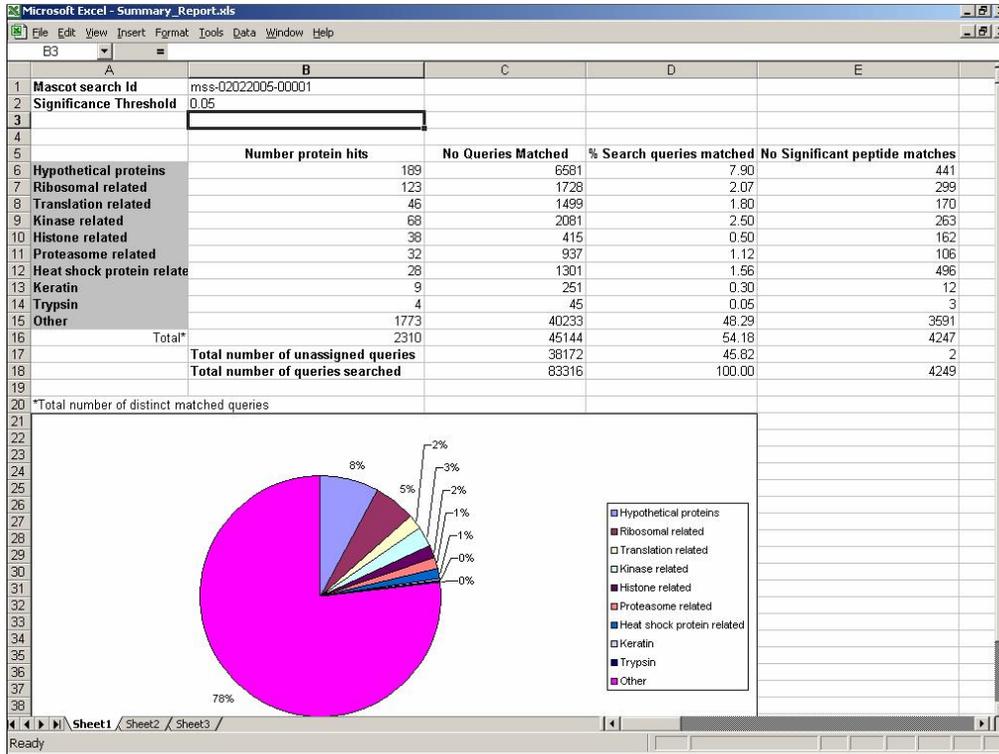
**ASMS 2005**



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 download 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.



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.

## *S.pombe 2D Gel example*

- 2D Gel -> PMF search
- 2 different spot processing protocols
  - Manual in-gel digest
  - Automated in-gel digest on novel platform
- Compare results from the two different processing methods.

**ASMS 2005**

**{MATRIX}**  
**{SCIENCE}**

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

## *Data analysis approach*

- Data from the two different methods searched and automatically imported into Integra (258 Mascot searches)
- Pull out all protein hits from all searches in the experiment exceeding specified criteria for each processing method into Excel
  - $p < 0.05$
  - 25% sequence coverage
  - Matches 6 or more peptides
    - (Single SQL query)
- Analyse.

**ASMS 2005**

**MATRIX  
SCIENCE**

After the results 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) can be specified within a single, simple SQL query against the protein hit table of the database.

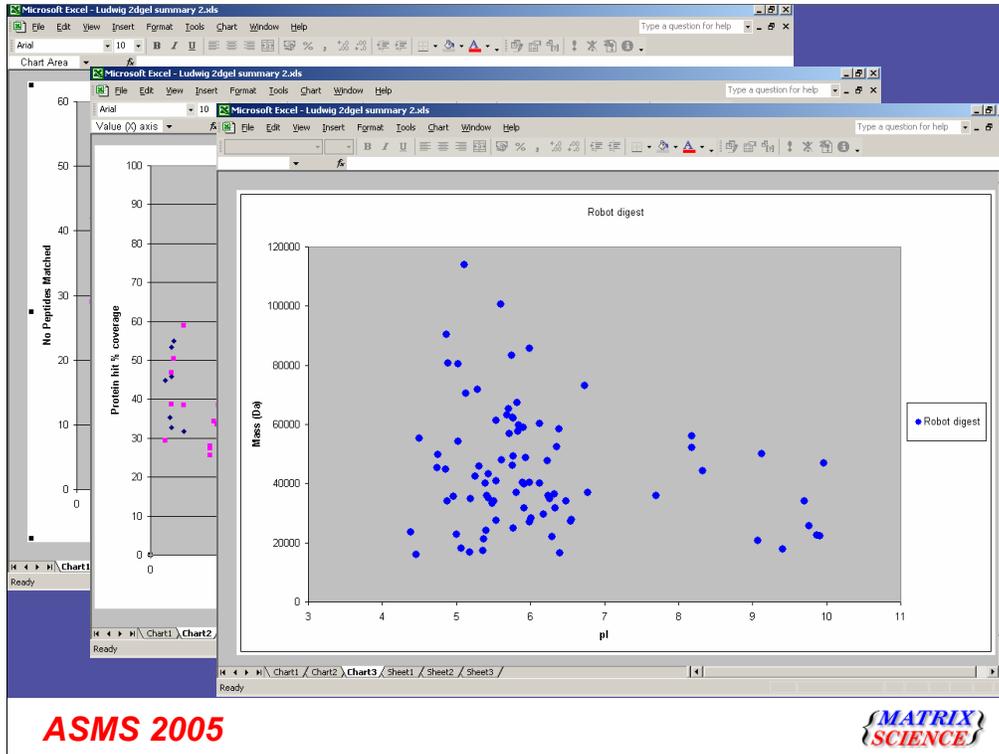
We will then do some analysis of these data in Excel

SPOT_NO	ACCESSION	MASS	DESCRIPTION	MASCOT_PROTEIN_SCORE	PROBABILITY	NUMBER_PEPTIDE
35	SPBC1604.21c	113960.23	lpt3ubp1, SPBC211.09 ubiquitin-activating enzyme e1	205	1.58272E-17	
48	SPAC1565.08	90353.78	l SPAC6F12.01 AAA family ATPase	361	3.97561E-33	
49	SPAC1565.08	90353.78	l SPAC6F12.01 AAA family ATPase	372	3.15794E-34	
55	SPBC16010.08c	100616.57	l AAA family ATPase	417	9.98629E-39	
77	SPAC95.03	93414.87	l eu2 putative 3-isopropylmalate dehydratase	425	1.58272E-39	
78	SPAC57A7.04c	71752.71	l pab1 pab1poly(A) binding protein	243	2.50844E-21	
140	SPAC57A7.04c	71752.71	l pab1 pab1poly(A) binding protein	130	5.005E-10	
140	SPAC9.09	65685.48	l putative homocysteine methyltransferase (S-methyltet	114	1.99253E-08	
140	SPAC926.04c	80717.18	l swo1 hsp90 molecular chaperone	65.2	0.001511466	
149	SPAC57A7.04c	71752.71	l pab1 pab1poly(A) binding protein	173	2.50844E-14	
155	SPAC57A7.04c	71752.71	l pab1 pab1poly(A) binding protein	204	1.99253E-17	
160	SPAC57A7.04c	71752.71	l pab1 pab1poly(A) binding protein	294	1.99253E-26	
173	SPCC1739.13	70474.82	l heat shock protein 70 family	217	9.98629E-19	
179	SPAC13G7.02c	70384.82	l heat shock protein 70 family	114	1.99253E-08	
213	SPCC1739.13	70474.82	l heat shock protein 70 family	138	7.9329E-11	
213	SPAC13G7.02c	70384.82	l heat shock protein 70 family	62.2	0.003019811	
223	SPCC1739.13	70474.82	l heat shock protein 70 family	164	1.99253E-15	
223	SPAC13G7.02c	70384.82	l heat shock protein 70 family	111	3.97561E-08	
242	SPAC926.04c	80717.18	l swo1 hsp90 molecular chaperone	84.2	1.90265E-05	
267	SPBC1709.05	67449.01	l sks2 hsc1 heat shock protein 70 family	215	1.58272E-18	
273	SPAC1F5.02	55244.45	l putative protein disulphide isomerase	296	7.9329E-27	
302	SPBC1709.05	67449.01	l sks2 hsc1 heat shock protein 70 family	159	6.30092E-13	
322	SPAC12G12.04	62413.98	l mcg60 hsp60 putative mitochondrial chaperonin 60 (Pv	367	9.98629E-34	
322	SPBC3B9.08c	17370.56	l limgo-nashi homolog	52.4	0.02880769	
326	SPAC12G12.04	62413.98	l mcg60 hsp60 putative mitochondrial chaperonin 60 (Pv	439	6.30092E-41	
328	SPBC646.11	58853.75	l ctc6 t-complex protein 1, zeta subunit	236	1.2572E-20	
330	SPAC1420.02c	63796.32	l ctc5 t-complex protein 1, epsilon subunit	139	6.30092E-11	
340	SPCRJ732.02c	62157.56	l putative xylose kinase	187	9.98629E-16	

ASMS 2005



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.



**ASMS 2005**



We've pulled out a lot of data relating to the quality of these matches from the database. Using Excel's 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.

The screenshot shows an Excel spreadsheet with two sections of summary statistics. The first section is for 'Hand digest' and the second is for 'Robot digest'. Each section compares 'NUMBER\_PEPTIDES' and 'PERCENTAGE\_COVERAGE' across various statistical measures.

	A	B	C	D	E	F	G	H	I
2	<b>Hand digest</b>								
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	<b>Robot digest</b>								
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					

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...*

**ASMS 2005**

**MATRIX**  
**SCIENCE**

The screenshot displays two overlapping Excel windows. The background window, titled 'Mudpit summary.xls', shows a list of protein accessions and descriptions. The foreground window, titled 'Mudpit summary.xls', shows a detailed table of search results for a specific protein hit (E980235).

**Merged hit list from all fractions of a MudPIT run**

ACCESSION	DESCRIPTION
AA063401	AY3
AA068393	AY3
BAB20776	ABD
E980235	H.S.
HHHUB4	heat
HS9B_HUMAN	Heat sh
AAA37866	MUSHSH
HHMS94	heat sh
HS9B_MOUSE	heat sh
T46243	hypothet
AAA37866	MUSHSH
BAC62488	AB07236
AAB23369	S45392
HS9B_RAT	Heat sh
G9TBA7	Hypothet
AAC48718	SSU9438
BAB20777	AB04367
BAC62457	AB07236
CAC18967	Sequenc
HHHUB6	heat sh
HS9A_HUMAN	Heat sh
HS9A_PIG	Heat sh
A29170	phospho
AAP36132	Homo sa
CAA59331	HS2PPH
ENDA_HUMAN	Alpha en
G7ZNV6	Hypothet
HHMS96	heat sh
HS9A_MOUSE	Heat sh
G80Y52	Hspca pr
G91XW0	Heat sh
A35922	dnaK-ty
G95LNB	Hypothet
G9UV14	Heat sh
A27077	dnaK-ty
BAB19615	AB03498

**Merged peptide match details for a protein hit from all MudPIT fractions/searches**

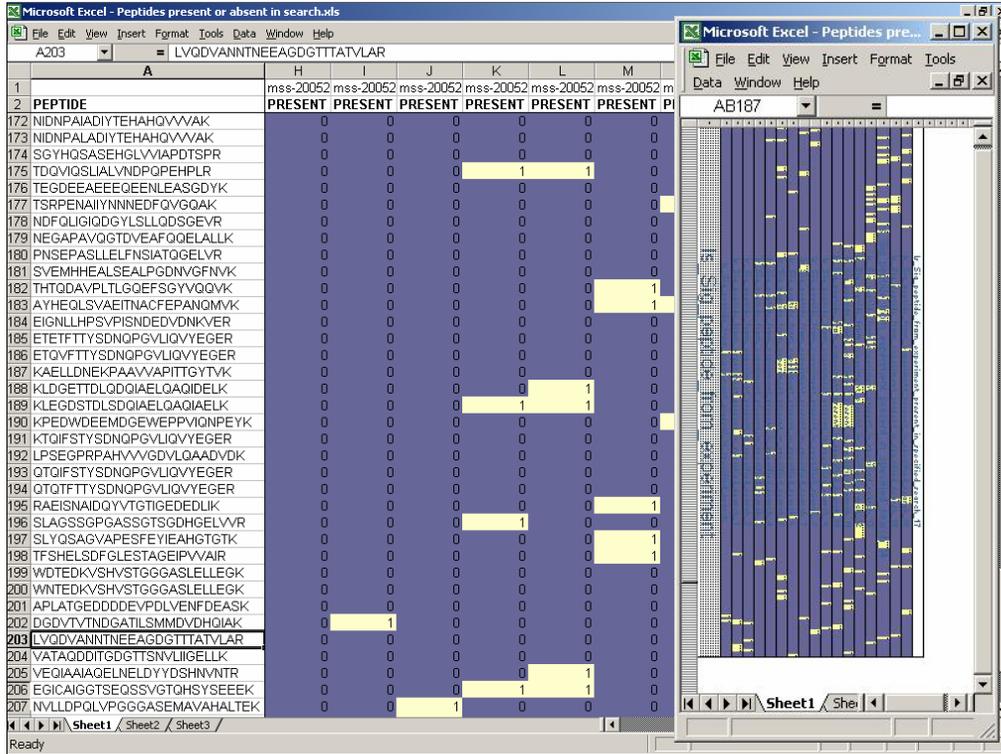
QUERY	RANK	SCORE	EXPECT	MREXP	MRCALC	DELTA	MISSED	PEPTIDE
352917.H.SAPIENS.HSP90	31							
E980235	22	41.88	1.736105736	730.673448	729.438446	1.235002	0	LSELLR
AA063401	18	22.97	107.0429508	1039.632724	1038.486892	1.146801	0	YESLTPSPSK
AA068393	55	51.86	0.146724233	1143.784172	1140.55231	3.231865	0	LGIHEDSTNR
BAB20776	53	29.58	24.38973262	1143.193448	1140.55231	2.641141	0	LGIHEDSTNR
E980235	49	38.53	2.959642326	1140.672724	1140.55231	0.120417	0	LGIHEDSTNR
mss-2005-2005-00021	1	40.19	2.028007079	1140.702724	1140.55231	0.150417	0	LGIHEDSTNR
mss-2005-2005-00021	52	29.1	26.6628154	1141.983448	1140.55231	1.431141	0	LGIHEDSTNR
mss-2005-2005-00014	97	33.34	9.586363644	1151.112724	1150.56057	0.562156	0	YIDGQELNKK
mss-2005-2005-00014	104	6.35	4937.128193	1161.692724	1159.57605	2.116674	0	SIYYITGESK
mss-2005-2005-00014	103	23.44	101.1598663	1160.152724	1159.57605	0.576674	0	SIYYITGESK
mss-2005-2005-00011	45	30.68	17.38799637	1194.732724	1193.6404	1.092328	0	IDIDNPNQGR
mss-2005-2005-00010	51	33.1	10.50421287	1241.92724	1241.69789	0.274833	0	ADLNINLGTIAK
mss-2005-2005-00011	54	24.03	84.71995008	1242.512724	1241.69789	0.814833	0	ADLNINLGTIAK
mss-2005-2005-00010	55	48.57	0.315816002	1275.222724	1274.63536	0.587363	0	ELISNASDALK
mss-2005-2005-00011	60	52.76	0.120346885	1275.222724	1274.63536	0.587363	0	ELISNASDALK
mss-2005-2005-00011	68	34.91	6.951738625	1311.212724	1310.56259	0.650132	0	EDDOTEYLEER
mss-2005-2005-00010	63	54.18	0.079721127	1349.402724	1348.72717	0.675551	0	TLTLVDTGIGMTK
mss-2005-2005-00011	82	31.29	15.50788679	1349.462724	1348.72717	0.735551	0	TLTLVDTGIGMTK
mss-2005-2005-00011	83	9.6	2256.077364	1361.502724	1348.72717	2.775551	0	TLTLVDTGIGMTK
mss-2005-2005-00010	68	29.88	21.54172172	1416.242724	1415.63031	0.612414	0	ESLELPEDEEEK
mss-2005-2005-00011	100	52.63	0.109215049	1416.022724	1415.63031	0.392414	0	ESLELPEDEEEK
mss-2005-2005-00020	99	37.54	3.72573359	1530.083448	1526.73648	3.346967	0	SLTNDWEDHLAVK
mss-2005-2005-00021	118	63.72	0.008862511	1529.743448	1526.73648	3.006967	0	SLTNDWEDHLAVK
mss-2005-2005-00021	116	32.29	12.33101215	1529.314172	1526.73648	2.577691	0	SLTNDWEDHLAVK
mss-2005-2005-00021	117	59.63	0.022750908	1529.383448	1526.73648	2.646967	0	SLTNDWEDHLAVK
mss-2005-2005-00020	96	29.33	24.82778342	1529.844172	1526.73648	3.107691	0	SLTNDWEDHLAVK
mss-2005-2005-00020	96	55.06	0.065163272	1529.443448	1526.73648	2.706967	0	SLTNDWEDHLAVK
mss-2005-2005-00011	124	72.76	0.000962907	1848.633448	1846.78967	1.843775	0	NPPDITQEEYGEFYK
mss-2005-2005-00011	123	46.25	0.445640404	1846.693448	1846.78967	-0.10623	0	NPPDITQEEYGEFYK
mss-2005-2005-00020	159	25.28	40.80005132	2179.423448	2175.93764	3.485612	0	YHTSGSGDEMSTLSEYVSR
mss-2005-2005-00021	171	8.1	2670.523823	2177.173448	2175.93764	1.235612	0	YHTSGSGDEMSTLSEYVSR

ASMS 2005

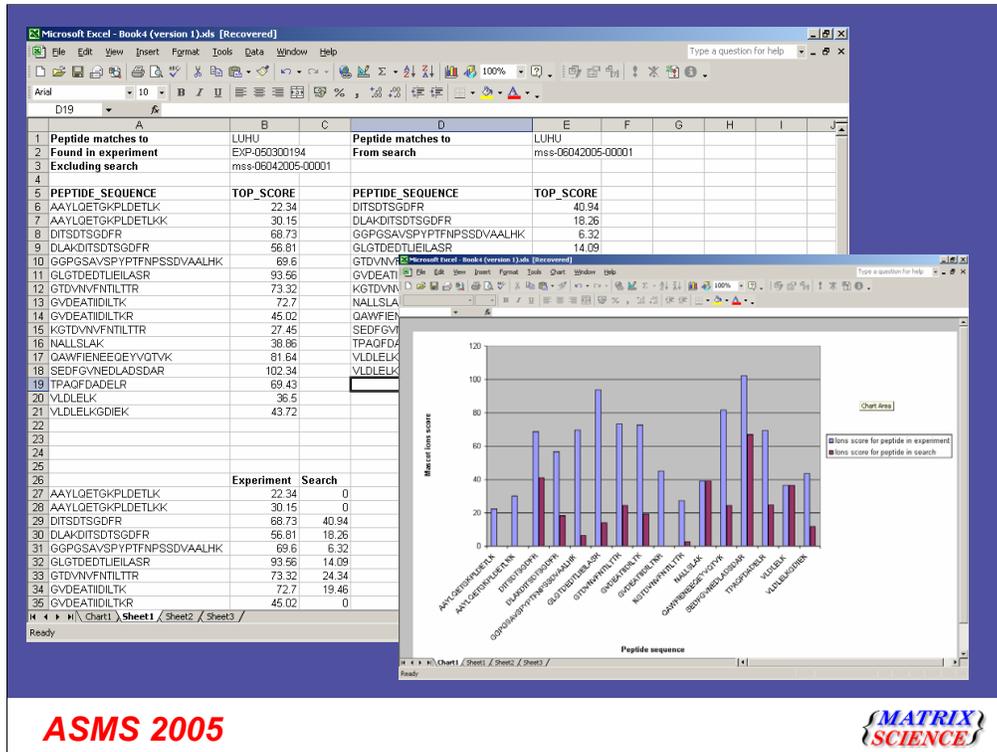
MATRIX SCIENCE

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).



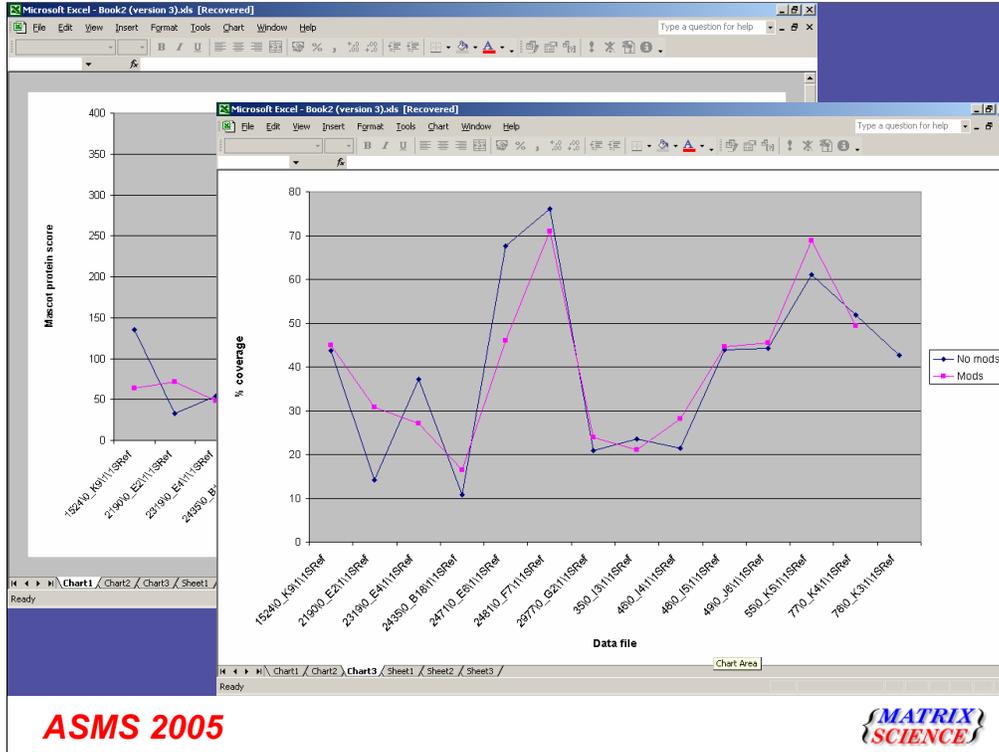
So we can check the quality of our 1<sup>st</sup> dimension chromatography...this generated from the same 17 MudPIT fractions. All of the peptides with a e value below the 0.05 threshold from all of the fractions have been identified and then cross checked against the search results from each fraction 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.



**ASMS 2005**

**MATRIX SCIENCE**

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.



**ASMS 2005**



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 21900\_E2 which might warrant further investigation (sig threshold 55)

Missing final point for 780\_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

**ASMS 2005**

**{MATRIX}  
{SCIENCE}**

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

**ASMS 2005**

**{MATRIX}**  
**{SCIENCE}**