

Using Scaffold to support additional search engines in Mascot Integra

MASCOT

{MATRIX}
{SCIENCE}

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.

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

Mascot Integra is our solution for managing and tracking proteomics workflows and results. As with Mascot and Mascot distiller it supports products from all the major mass-spectrometry vendors. 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

Supporting multiple search engines

- Many labs use more than one search engine
- Would like organise and track results from one location
- Mascot Integra only supports Mascot
- Supporting multiple search engines increases workflow options

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

Many labs have more than one search engine (e.g. both Mascot & Sequest) and would like to be able to use a package like Mascot Integra to be able to store, organise and compare/mine data from all of their search results. The current release of Mascot Integra only supports import of Mascot search result files (.dat files), but we would like to increase support for multiple search engines as this would allow us to increase the range of workflows we can support and types of analyses that can be done.

Scaffold (Proteome Software inc)

- Scaffold allows search results from multiple search engines to be imported and condensed into a single result file
- Supports all the major search engines
- Produces a protXML export file
- ScaffoldBatch allows Scaffold to be driven from Mascot Integra

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

{MATRIX}
{SCIENCE}

I'm sure most of you are familiar with Scaffold from Proteome Software inc, but I'll just run through some of the key aspects of the package.

Scaffold allows results from many different search engines (including Mascot, SEQUEST, X!Tandem etc) to be imported, compared and condensed into a single result file. In addition to its own result file, it can generate a protXML export of the results. Importantly from our point of view, you can license an additional component of Scaffold – ScaffoldBatch – which allows you to drive batch analysis of results files. This allows us to automatically run Scaffold analysis from Integra, pickup the protXML export file and then import the results in an action we can incorporate into our main experimental workflows.

Approach in Mascot Integra

Either:

1. Import an existing protXML result file (can be from any source), including TPP files or files manually exported from Scaffold
2. Use a special experiment task to pick up selected search engine result files and drive ScaffoldBatch

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

If you want to import protXML result files into Mascot Integra, you will have two options. You can either pickup and import an existing protXML file, generated from any source (including the TPP if you're running that, or manually exported from Scaffold). Note that Mascot Integra does not directly support the Trans proteomic pipeline and does not support import of pepXML results.

Or you can use a new experimental task from within an Integra experiment to pick up the search result files from whatever search engine you want to use, and then have Integra drive ScaffoldBatch to automatically generate the protXML result file and import it. For most of the rest of this talk we'll be concentrating on approach 2 – using Mascot Integra to drive ScaffoldBatch to automatically run Scaffold, generate the protXML results and then parse and import those results into the Mascot Integra database.

Example: Comparison between search databases

- **Human ovarian epithelium**
- **17,216 MS/MS spectra**
- **Searched with SEQUEST/BioWorks and Mascot**
 - Search against two databases (May 2009 releases) concatenated with random decoy sequences:
 - Human SwissProt sequences
 - IPI Human database
- **Pickup .srf & .dat result files**
- **Generate ScaffoldBatch job**
 - 2 separate jobs - one for each database searched

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

I'll be using an example dataset derived from human ovarian epithelium. This is a large dataset but we're only going to look at a subset it; a single fraction which has 17,216 MS/MS spectra.

The dataset was searched using SEQUEST from the BioWorks package and with Mascot 2.2 against two different search databases 1: The human sequences from SwissProt and 2: the IPI Human database

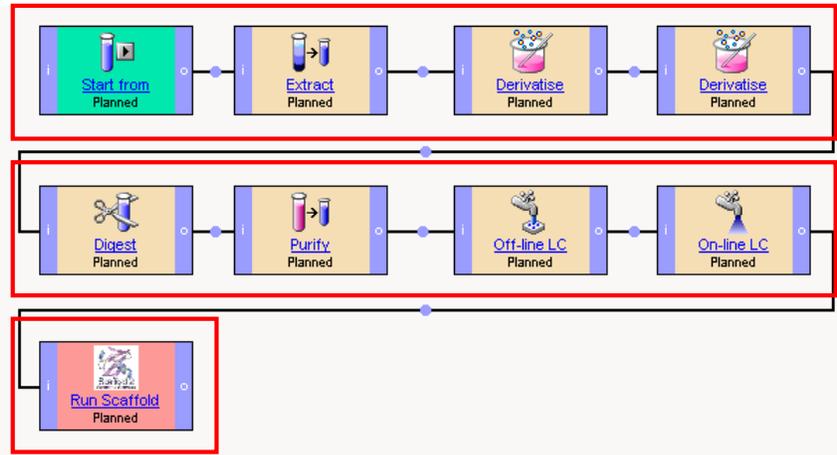
Both databases were concatenated with random sequences to give an estimate of the False Positive Rates.

We're going to do a comparison of searching between the two databases. The IPI Human database now contains just over 80,000 protein sequences compared with just over 20,000 human protein sequences in SwissProt – does the increased coverage in IPI Human result in increased coverage in our search results?

Unlike for Mascot Daemon searches, Integra doesn't try and run the BioWorks SEQUEST job – the end user needs to run the searches themselves and have Mascot Integra pick up the results files.

Once the searches have been completed, Integra can be pointed towards the result files (in this case BioWorks srf file (and also the Mascot .dat results)) and it will then generate the ScaffoldBatch job. Again, here we're going to process each search set differently, so we'll get 2 different Scaffold search results.

Workflow



MASCOT

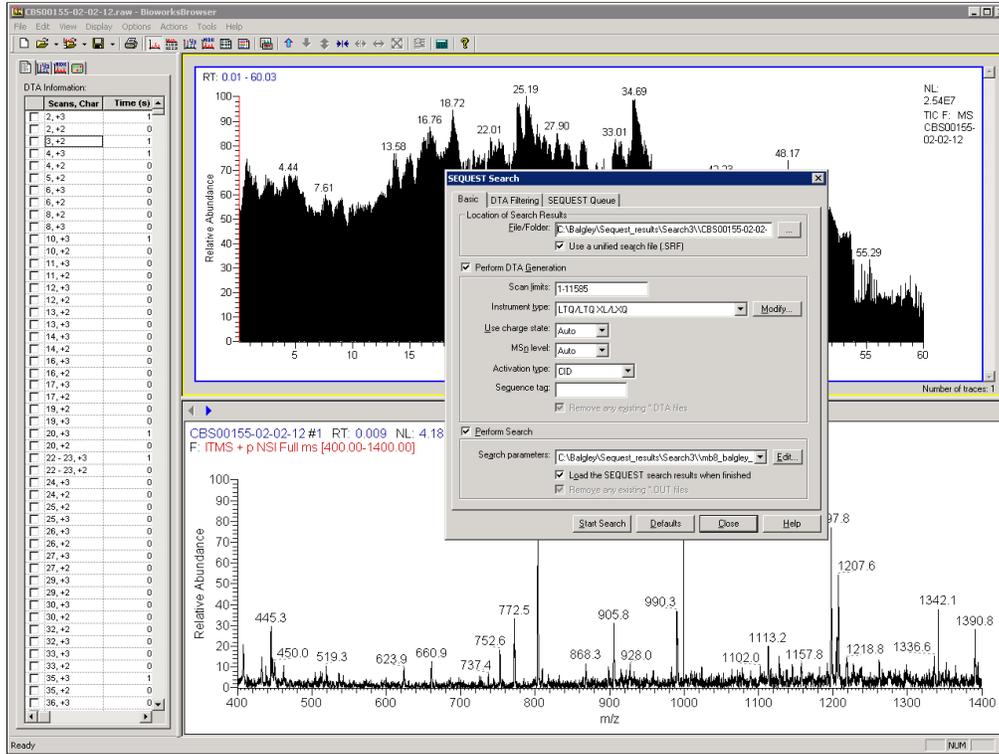
Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

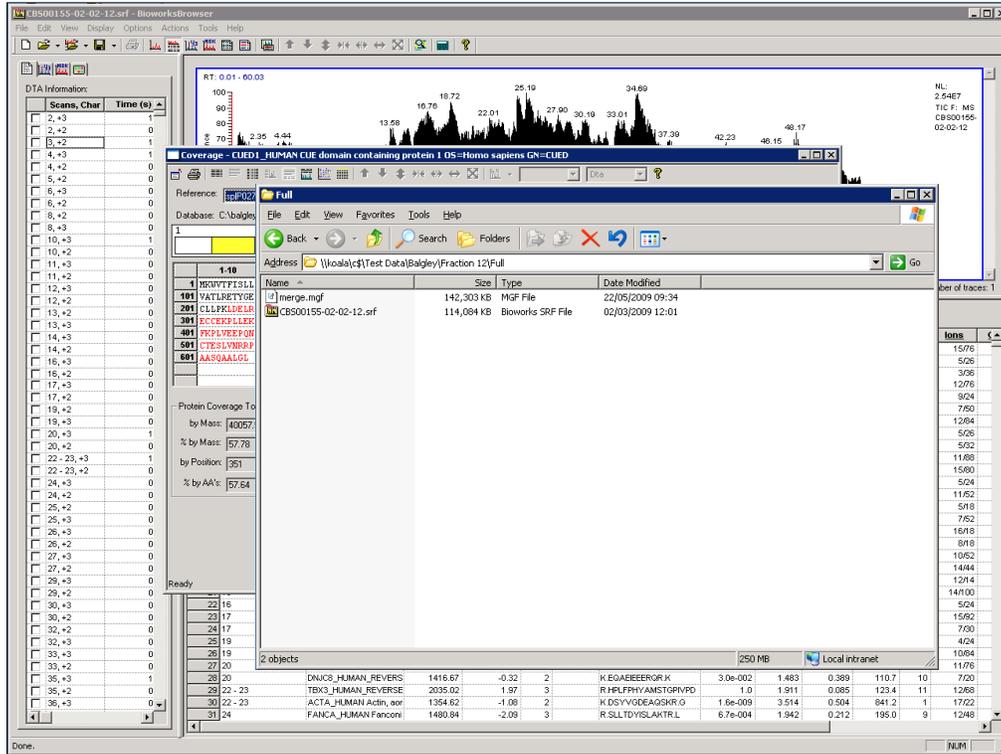
MATRIX
SCIENCE

A: Upstream workflow – this tracks the main experimental steps involved in generating the MS/MS samples, so we can track our protein extraction, reduction & alkylation steps, digestion and purification conditions and chromatography steps.

B: Unlike with Mascot searches, Mascot Integra does not try and run the additional search engines – you run the searches as you would normally (e.g. through BioWorks), then point Mascot Integra towards the generated result files.



Here we are in BioWorks with some of the raw data loaded and ready for searching. You just have to submit the search to SEQUEST in the normal way – note that here we are going to produce .SRF unified search result files. Scaffold can handle other SEQUEST result files such as .out files etc, but if you're using BioWorks then the .SRF unified files are much easier to handle.



Once the search is completed, you'll have a set of .SRF files – the next step is to point Mascot Integra towards these and get it to setup and run ScaffoldBatch.

Running the ScaffoldBatch Experiment task

- **Should be familiar to anyone who has manually set up a Scaffold analysis**
 - Define biosample and associate result files
 - Choose analysis thresholds
 - Specify X!Tandem analysis options (if you wish).

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

{MATRIX}
SCIENCE}

Now we'll run through how you set up and run ScaffoldBatch from within a Mascot Integra experiment. The steps involved should look very familiar to anyone who has run an analysis through the main Scaffold interface.

The three main steps are:

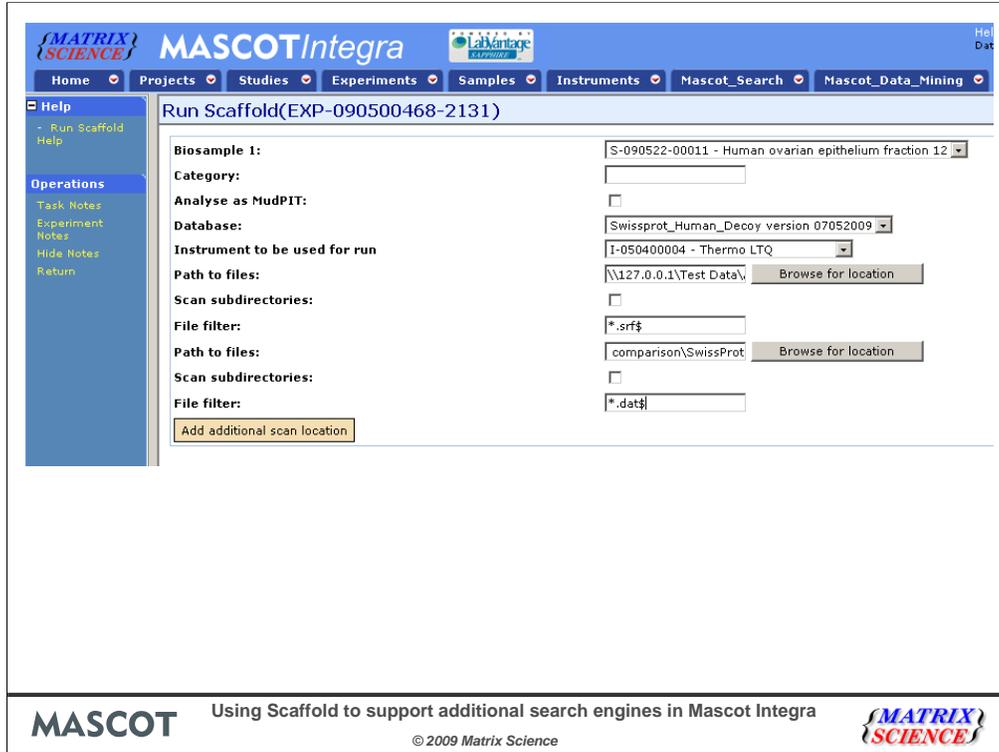
1. Pick your sample (biological sample) and associated the search result files with it
2. Choose the analysis thresholds
3. Specify any integrated X!Tandem analysis options if you want Scaffold to run an X!Tandem search for you

The screenshot displays the MASCOTIntegra web application interface. At the top, there is a navigation bar with the Matrix Science logo and the text "MASCOTIntegra". Below this, a menu contains "Home", "Projects", "Studies", "Experiments", "Samples", "Instruments", and "Mascot_Search". A sidebar on the left lists "Help" (with sub-items "Run Scaffold" and "Help") and "Operations" (with sub-items "Task Notes", "Experiment Notes", "Hide Notes", and "Return"). The main content area is titled "Run Scaffold(EXP-090500468-2131)" and contains a form with the following elements:

- A label: "Select Scaffoldbatch server for new group"
- A dropdown menu with the text: "Select Scaffoldbatch service to use:"
- A note: "Please note: only services currently available are shown" followed by a dropdown menu showing "KOALA".

At the bottom of the page, there is a footer with the MASCOT logo, the text "Using Scaffold to support additional search engines in Mascot Integra", the copyright notice "© 2009 Matrix Science", and the Matrix Science logo.

First, you select the system your copy of Scaffold and ScaffoldBatch is installed and licensed on.



Next you define the Scaffold ‘Biosample’ – including the FASTA database the search was carried out against, the instrument the MS/MS run was done on and also point the system towards the location the result files are in and specify a filter to pick out the files. You can easily specify multiple scan locations and filters for a single biosample. In this instance, this allows us to pick up both the SEQUEST .SRF and the Mascot .dat result files.

MASCOT Integra LabVantage

Home Projects Studies Experiments Samples Instruments Mascot_Search

Run Scaffold(EXP-090500468-2131)

Help
- Run Scaffold
Help

Operations
Task Notes
Experiment Notes
Hide Notes
Return

Import file?	Path
<input checked="" type="checkbox"/>	\\127.0.0.1\Test Data\ASMS20...tabase comparison\SwissProt\CBS00155-02-02-12.srf
<input checked="" type="checkbox"/>	\\127.0.0.1\Test Data\ASMS2009\Fr...J\Database comparison\SwissProt\F002616.dat
	\\127.0.0.1\Test Data\ASMS2009\Fraction 12\Full\Database comparison\SwissProt\F002616.dat

MASCOT Using Scaffold to support additional search engines in Mascot Integra **MATRIX SCIENCE**
© 2009 Matrix Science

The system then scans the specified location and picks out matching files, which you can then check and exclude and result files you're not interested in submitting to Scaffold.

MASCOTIntegra Help SiteMap LogOff Database: Integrademo User: Patricke

Home Projects Studies Experiments Samples Instruments Mascot_Search Mascot_Data_Mining Utilities

Run Scaffold(EXP-090500468-2131)

Biological sample 1: S-090522-00011

//127.0.0.1/Test Data/ASMS2009/Fraction 12/Full/Database comparison/SwissProt/CBS00155-02-02-12.srf
//127.0.0.1/Test Data/ASMS2009/Fraction 12/Full/Database comparison/SwissProt/F002616.dat

Add additional biological sample

MASCOT Using Scaffold to support additional search engines in Mascot Integra **MATRIX SCIENCE**
© 2009 Matrix Science

You can then go back and setup another Biosample if you wish.

MASCOTIntegra

Home Projects Studies Experiments Samples Instruments Mascot_Search Mascot_Data_Mining

Run Scaffold (EXP-090500468-2131)

General Minimum Thresholds:

Protein probability (%): 20

Min no. peptides: 1

Peptide probability (%): 50

Use Individual Program Thresholds

Accept Charge +1

Accept Charge +2

Accept Charge +3

Accept Charge +4 and higher

Min NTT: 0

Use Mascot scores:

Ion - Identity Score: 0.0

Ion Score (+1): 20

Ion Score (+2): 30

Ion Score (+3): 40

Ion Score (+4): 40

Use SEQUEST scores:

DeltaCn: 0.1

XCorr (+1): 1.8

XCorr (+2): 2.5

XCorr (+3): 3.5

XCorr (+4): 3.5

Use XI Tandem scores:

-Log(E-Value): 2.0

MASCOT Using Scaffold to support additional search engines in Mascot Integra **MATRIX SCIENCE**

© 2009 Matrix Science

The next step is to specify the export thresholds. The thresholds you specify here will determine which protein and peptide hits are going to be exported into the protXML result file by ScaffoldBatch. Since we can do additional filtering and comparisons once the data is in the Mascot Integra database, in this instance we want to import more than just the high confidence hits, so we're going to change the thresholds.

Now we're going to include all protein hits including one hit wonders, with a peptide probability threshold of 50% and the protein probability threshold of 20 percent. Just like when you're running Scaffold interactively, you can alternatively specify search engine specific thresholds instead. One thing to note if you're running Scaffold on Mascot search results - I've tended to find that Scaffold is quite conservative with Mascot hits. Scaffold is taking its thresholds for Mascot from the identity threshold which we tend to find is quite conservative (if you've attended many of our past user group meetings you'll have seen many talks where we've advocated using the less conservative homology threshold) – for example, for this dataset I'd be quite happy to drop the peptide probability threshold from Scaffold for Mascot matches to 0.8 as this actually gives similar results to using the homology threshold in Mascot and with a low FDR.

The screenshot displays the MascotIntegra software interface. At the top, there is a navigation bar with the following menu items: Home, Projects, Studies, Experiments, Samples, Instruments, Mascot_Search, Mascot_Data_Mining, and Utilities. The main title is "Run Scaffold(EXP-090500468-2131)".

The interface is divided into several sections:

- Batch description:** Search 1: SEQUEST & Mascot vs Human SwissProt
- Options:**
 - Condense data while loading?
 - Analyze with X!Tandem?
 - Search subset database?
- Add extra modification:** A list of modifications including 15dB-biotin, 2HPG, 4-ONE, a-type-ion, AcQTag, Acetyl, Acetyl:2H(3), ADP-Ribosyl, AEBs, AEC-MAEC, Ala->Asp, Ala->Glu, Ala->Gly, Ala->Pro, Ala->Ser, Ala->Thr, Ala->Val, Amidated, Amidine, and Amidino. There are "Add >>", "New +", and "<< Remove" buttons.
- Selected variable modifications:** Carbamidomethyl, Oxidation

The footer contains the Mascot logo, the text "Using Scaffold to support additional search engines in Mascot Integra", the Matrix Science logo, and the copyright notice "© 2009 Matrix Science".

The next step is to define any search conditions for use with X!Tandem. Scaffold comes with X!Tandem incorporated into it and you can (optionally) carry out an X!Tandem search for incorporation into the Scaffold results; either against the whole database or just against those protein hits identified in the source results files. You can also specify any additional variable modifications, or define a custom variable modification 'on the fly'.

The screenshot shows the Mascot Integra web interface. At the top, there is a navigation bar with menus for Home, Projects, Studies, Experiments, Samples, Instruments, Mascot_Search, Mascot_Data_Mining, and Utilities. The main content area is titled 'Run Scaffold(EXP-090500468-2131)'. Below this title, there is a section for 'Result import status' which contains a table:

Description	Status	Comments
Search 1: SEQUEST & Mascot vs Human_SwissProt	Queuing	Job queuing on Scaffoldbatch server

Below the table, there is a 'Refresh' button and a warning: 'Do not use the browser refresh button as you will lose any task and experiment notes you have added. This page will automatically refresh every 60 seconds.'

The footer of the page contains the Mascot logo, the text 'Using Scaffold to support additional search engines in Mascot Integra', and the Matrix Science logo.

Now, just as with carrying out Mascot searches from Mascot Integra, you can sit back and leave the system to run everything in the background – you don't need to stay logged into the system. It'll run ScaffoldBatch automatically, pick up the resulting files, parse and import the proXML results and attach the Scaffold .sfd result file to the dataset in the database (so any user with access to the results in Integra can easily also access the original Scaffold results).

Result Import

- **Additional values not present in protXML are calculated during result parsing/import including:**
 - Protein mass
 - Protein pI
 - Peptide start/end positions and flanking residues
 - Missed enzyme cleavage points
- **Results then appear in lists alongside Mascot results**

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

During the parsing phase, Mascot Integra calculates and stores some additional information about the protein hits not included in the protXML file, including the protein mass and pI, peptide start/end & flanking residues, missed enzyme cleavage points in the peptide. Once the results have been imported, then they are accessed in Mascot Integra in the same way as imported Mascot search results.

Results overview

- **SwissProt human sequences**
 - 1074 protein hits (990 target, 84 decoy)
 - 1018 protein hit ranks (84 decoy)
- **IPI Human**
 - 2266 protein hits (2166 target, 100 decoy)
 - 981 protein hit ranks (100 decoy)
- **20 of the 53 protein hit ranks 'missing' from IPI Human are '1 hit wonders'.**

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

After processing the results with Scaffold we can see that we have many more protein hits from IPI Human with only a slight increase in the number of decoy hits. However, the number of individual groups of proteins has actually slightly decreased (there are 53 fewer target protein hit ranks). This suggests that many of the additional protein hits from IPI Human could be as a result of increased redundancy in IPI Human compared with SwissProt human.

Viewing results 1: HTML reports

- Similar to the standard protXML XSLT result view
- Shows additional information and formats the results in the Mascot style
 - Results can be filtered using user defined SQL filters
- Lacks some of the additional information from Scaffold (e.g. more like looking at a result from the TPP)

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

This means that once the protXML results are imported into Mascot Integra, you can use them in much the same way you would any Mascot results imported into the system, including being able to manually validate and approve protein and peptide matches. There are four main ways to view results from within Mascot Integra. Option 1 is to view relatively static HTML based reports. For those of you familiar with the Trans Proteomic Pipeline, these reports show a similar level of information to the standard XSLT reports generated from that. It does include some additional information so that the view is also a bit like looking at a standard Mascot HTML report (master_results report). However, Integra adds some additional functionality such as being able to filter the reports with user definable filters, approve protein hits and view different chunks of the report rather than the whole report in one go.

Select/deselect all protein hits
 1 Selected hit: sp|P02768|ALBU_HUMAN Serum albumin OS?Homo sapiens GN?ALB
 Check to approve protein and peptide matches.

Comments:
 sp|P02768|ALBU_HUMAN Mass: 69348 Probability: 1.0 Spectrum ids: 0.0% Coverage: 50.7%
 Serum albumin OS?Homo sapiens GN?ALB

Approve

Match#	Weight	p	nsp	adj p	nsp	Total Peptide group	Mr(Calc)	Start End	Miss	Charge	Peptide
<input checked="" type="checkbox"/>	1	0.95	0	1.9	2		1370.480	187 198	0	1	K.AAFTEC[57.02][57.02]QAADK.A
<input checked="" type="checkbox"/>	1	0.942	0	1.874	2		1370.505	187 198	0	1	K.AAFTEC[57.02][57.02]QAADK.A
<input checked="" type="checkbox"/>	1	0.649	0	0.649	1		1372.484	187 198	0	1	K.AAFTEC[57.02][57.02]QAADK.A
<input checked="" type="checkbox"/>	1	0.5	0	0.5	1		1372.495	187 198	0	1	K.AAFTEC[57.02][57.02]QAADK.A
<input checked="" type="checkbox"/>	1	0.95	0	12.35	13		879.434	250 257	0	1	K.AEFAEVSK.L
<input checked="" type="checkbox"/>	1	0.95	0	4.75	5		879.570	250 257	0	1	K.AEFAEVSK.L
<input checked="" type="checkbox"/>	1	0.95	0	1.9	2		880.537	250 257	0	1	K.AEFAEVSK.L
<input checked="" type="checkbox"/>	1	0.861	0	1.685	2		1304.687	301 310	0	1	K.EC[57.02][57.02]KPLLEK.S
<input checked="" type="checkbox"/>	1	0.95	0	4.75	5		1025.637	35 44	1	1	R.FKDLGGENPK.A
<input checked="" type="checkbox"/>	1	0.95	0	0.95	1		1227.605	35 44	1	1	R.FKDLGGENPK.A
<input checked="" type="checkbox"/>	1	0.95	0	3.8	4		1348.040	66 75	0	1	K.LVNEVTEFAK.T
<input checked="" type="checkbox"/>	1	0.95	0	4.75	5		1348.712	66 75	0	1	K.LVNEVTEFAK.T
<input checked="" type="checkbox"/>	1	0.95	0	3.8	4		1349.654	66 75	0	1	K.LVNEVTEFAK.T
<input checked="" type="checkbox"/>	1	0.512	0	0.512	1		1350.602	66 75	0	1	K.LVNEVTEFAK.T
<input checked="" type="checkbox"/>	1	0.765	0	0.765	1		788.649	258 264	0	1	K.LVTDLTLV
<input checked="" type="checkbox"/>	1	0.95	0	5.728	7		982.875	376 383	0	1	K.TYETLEK.C
<input checked="" type="checkbox"/>	1	0.925	0	0.348	10		983.527	376 383	0	1	K.TYETLEK.C
<input checked="" type="checkbox"/>	1	0.639	0	0.639	1		984.502	376 383	0	1	K.TYETLEK.C
<input checked="" type="checkbox"/>	1	0.95	0	0.95	1		1370.153	187 198	0	2	K.AAFTEC[57.02][57.02]QAADK.A
<input checked="" type="checkbox"/>	1	0.95	0	7.6	8		1371.275	187 198	0	2	K.AAFTEC[57.02][57.02]QAADK.A
<input checked="" type="checkbox"/>	1	0.95	0	0.95	1		1371.643	187 198	0	2	K.AAFTEC[57.02][57.02]QAADK.A
<input checked="" type="checkbox"/>	1	0.95	0	2.85	3		1373.350	187 198	0	2	K.AAFTEC[57.02][57.02]QAADK.A
<input checked="" type="checkbox"/>	1	0.95	0	0.95	1		1373.917	187 198	0	2	K.AAFTEC[57.02][57.02]QAADK.A
<input checked="" type="checkbox"/>	1	0.95	0	1.824	2		1374.577	187 198	0	2	K.AAFTEC[57.02][57.02]QAADK.A
<input checked="" type="checkbox"/>	1	0.95	0	1.9	2		879.132	250 257	0	2	K.AEFAEVSK.L
<input checked="" type="checkbox"/>	1	0.95	0	14.25	15		880.263	250 257	0	2	K.AEFAEVSK.L
<input checked="" type="checkbox"/>	1	0.95	0	0.95	1		880.533	250 257	0	2	K.AEFAEVSK.L
<input checked="" type="checkbox"/>	1	0.95	0	0.95	1		882.076	250 257	0	2	K.AEFAEVSK.L
<input checked="" type="checkbox"/>	1	0.95	0	0.95	1		1553.404	384 396	0	2	K.C[57.02][57.02]AAADPHEC[57.02]YAK.V

Here we have an example of an HTML report generated from within Integra. As you can see it looks fairly similar to a standard Mascot result report, but we have the additional protXML associated columns, such as Weight, nsp. We also have checkboxes available for manual validation of the protein and peptide hits.

The screenshot shows the MascotIntegra web interface. The main content area displays 'Protein View' for a search result. The protein identified is 'Serum albumin OS/Homo sapiens GN7ALH'. The search parameters include a nominal mass of 69321.5 and a calculated pI value of 5.92. The sequence coverage is 50.7%. Matched peptides are listed with their positions in the protein sequence, with the correct peptide highlighted in red: **FAFYLQKCF**. Below the peptide list, there is a table of predicted peptides with columns for Start, End, Weight, p, nsp, adj, p, nsp, and Miss Sequence.

Start	End	Weight	p	nsp	adj	p	nsp	Miss Sequence
35	44	1	0.95	4.75	✓		1225.637	1 R.FKDLGEEFK.A
35	44	1	0.95	0.95	✓		1227.605	1 R.FKDLGEEFK.A
35	44	1	0.95	2.85	✓		1225.255	1 R.FKDLGEEFK.A
35	44	1	0.95	10.45	✓		1226.44	1 R.FKDLGEEFK.A
35	44	1	0.95	3.8	✓		1227.414	1 R.FKDLGEEFK.A
35	44	1	0.95	0.95	✓		1228.347	1 R.FKDLGEEFK.A
35	44	1	0.95	1.9	✓		1228.513	1 R.FKDLGEEFK.A
35	44	1	0.95	1.9	✓		1225.539	1 R.FKDLGEEFK.A
35	44	1	0.95	9.004	✓		1227.016	1 R.FKDLGEEFK.A
35	44	1	0.95	0.95	✓		1227.585	1 R.FKDLGEEFK.A
37	65	1	0.95	0.95	✓		3423.823	1 K.DLGEENFKALVIAFAQYLQCC[57.02]PFEDHVL.L

In addition to the main protein hit list, we can get a protein view for any of the hits, showing protein coverage, peptide match positions etc.

What we can't generate though is a peptide fragmentation view though as protXML doesn't include enough information to generate this.

Doesn't contain peptide level information. If you want to see the fragmentation details you'll need to open up the original Scaffold sfd file. Since the ScaffoldBatch task automatically attached/associated the sfd, that file is available for downloading from the Browser and if you have the free Scaffold viewer installed on your client PC, you can open up and view the Scaffold file.

Viewing results 2: Mascot Integra Results Applet

- More dynamic results view
- Primary results interface in Mascot Integra
- Faster to use for large datasets - only fetches the data you are currently viewing
- Shows additional information from Scaffold
- Can link out to Scaffold sfd file if Scaffold/free viewer installed

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

Our recommended result interface in Mascot Integra is the Java results applet which offers a more dynamic result interface and is more like using the main Scaffold viewer than the static HTML reports.

The applet is much faster to use when viewing large datasets as it only fetches the current slice of results you are looking at, making examining large datasets a far more practical prospect. It also offers more advanced and dynamic filtering options (at both a protein and peptide level).

More of the additional information from Scaffold which has been included in the protXML export is shown.

As with the HTML reports, we can't generate a peptide fragmentation view. You can open up the original Scaffold sfd file easily though if you have either Scaffold or the free Scaffold viewer installed on your client PC.

Mascot Integra Results Applet

File Processing Help

Protein centric view Approve Unapprove

Hit rank (a)	Accession	Description	Mass	p	Peptide sequences matched	% Spectrum ids	S-090522-00011	CS00195-02-0
1	sp P02768 ALBU_HUMAN	Serum albumin OS?Homo sapiens GNFALE	66549	1.0	8	3.009	✓	✓
2	sp P12111 COXA3_HUMAN	Collagen alpha 2(I) chain OS?Homo sapiens GNFCOL6A3	142534	1.0	1	0.63	✓	✓
3	sp P60709 ACTB_HUMAN	Actin, cytoplasmic 1 OS?Homo sapiens GNACTB	41719	1.0	0	0.54	✓	✓
4	sp P62736 ACTA_HUMAN	Actin, axonic smooth muscle OS?Homo sapiens GNACTA2	41791	1.0	0	0.438	✓	✓
5	sp P12110 COX6C_HUMAN	Collagen alpha 2(II) chain OS?Homo sapiens GNFCOL6A2	138555	1.0	0	0.365	✓	✓
6	sp P08670 VIME_HUMAN	Vimentin OS?Homo sapiens GN2VIM	53502	1.0	4	0.339	✓	✓
7	sp Q05707 COXA1_HUMAN	Collagen alpha 1(OV) chain OS?Homo sapiens GNFCOL1A1	193467	1.0	7	0.334	✓	✓
8	sp P35779 MYH9_HUMAN	Myosin heavy chain, nonmuscle type A OS?Homo sapiens GNMYH9	226514	1.0	0	0.319	✓	✓
9	sp P12109 COXA1_HUMAN	Collagen alpha 1(OV) chain OS?Homo sapiens GNFCOL6A1	193629	1.0	0	0.291	✓	✓
10	sp P12133 FLNA_HUMAN	Filamin A OS?Homo sapiens GNFLNA	280743	1.0	0	0.265	✓	✓
11	sp P35555 FN1_HUMAN	Fiblin-1 OS?Homo sapiens GNFN1	312294	1.0	7	0.233	✓	✓
12	sp P00330 LDHA_HUMAN	Lactate dehydrogenase A chain OS?Homo sapiens GNLDHA	36540	1.0	0	0.228	✓	✓
13	sp P00450 SERU_HUMAN	Caruloplasm OS?Homo sapiens GNQC	122187	1.0	2	0.217	✓	✓
14	sp P19627 ITIH1_HUMAN	Inter-alpha-trypsin inhibitor heavy chain H1 OS?Homo sapiens GNITIH1	101371	1.0	0	0.212	✓	✓
15	sp P01024 C3_HUMAN	Complement C3 OS?Homo sapiens GNFC3	187146	1.0	7	0.201	✓	✓

Query No.	Weight	Charge	PK(Ca)	Start	End	Ms	Probability	NSP	Total	Rank	Mod.	Peptide	Cte...	Variable mods	S-090522-00011	CS00195-02-0
358	1.0	1	168.84	129	138	0	0.95	0.95	1	58	K	ILLAELFLK	G		✓	0.95
359	1.0	1	310.544	390	400	0	0.95	0.95	1	59	K	MLDLEIATYR	K	[L.S.99]	✓	0.95
360	1.0	2	319.032	186	195	1	0.95	1.9	2	60	R	EKLQEFMQR	E	[L.S.99]	✓	0.95
361	1.0	2	319.534	186	195	1	0.95	0.95	1	61	R	EKLQEFMQR	E	[L.S.99]	✓	0.95
362	1.0	2	668.047	424	438	0	0.95	0.95	1	62	R	ETNDSLPLVDTHSK	Q		✓	0.95
363	1.0	2	65.198	334	341	0	0.95	6.281	7	63	K	GTNESLER	Q		✓	0.95
364	1.0	2	160.511	129	138	0	0.95	5.7	6	64	K	ILLAELFLK	G		✓	0.95
365	1.0	2	660.755	401	409	1	0.95	4.75	5	65	R	KLLLEGEESR	K		✓	0.95
366	1.0	2	660.564	401	409	1	0.95	1.9	2	66	R	KLLLEGEESR	K		✓	0.95
367	1.0	2	660.946	400	408	1	0.95	0.95	1	67	R	KLLLEGEESR	K		✓	0.95
368	1.0	2	536.825	222	234	1	0.95	0.95	1	68	R	KVI			✓	0.95
369	1.0	2	536.785	222	234	1	0.646	0.646	1	69	R	KVI			✓	0.95
370	1.0	2	733.155	364	377	1	0.786	0.786	1	70	R	LQR			✓	0.95
371	1.0	2	734.712	364	377	1	0.95	0.95	1	71	R	LQR			✓	0.95

Peptide match was also found in the following proteins:

- sp|P05787|K2C8_HUMAN
- sp|P08729|K2C7_HUMAN

Protein View | Taxonomy | Family | Annotation | Spectrum | Ions matched | Error distribution | Top ten query matches | Search Summary

Protein view for sp|P08670|VIME_HUMAN Vimentin OS?Homo sapiens GN2VIM

```

1 STRSYSSSY RRMFGGOTA RPFSSRSYV ITTQTYSLG SALRPFSTSR
51 LYASSFGQYV ATRSSAVLR SYPQVRLQ DSYDFSLADA INTDFPMTFR
101 NERVDELGR DEKANTDVF PFLQKQL LRLQLRQD QSRSLQTE
151 ERMDELSPQV DQLTNSRNVY EYKDNLAED INLRLKLE EHLGEEKAKH
201 FLQSPQDQV NASTLARLDE PVESSLDEI RFLKLLHEZE TQELQALQE
251 GPQIDVDYV KFLTALLED VROQESYVA RNLQEEVY KRSFADLSEA
301 ADRHMDALQ AKQSTETYSR QVSLICVDY RKLQTHESLE RQRPDEDFP
351 AVALRYDQV IGRGRQTH RQKRRHMLR EYGLLAVR RMLDLYAVD
401 KLLLEGEESR SLFLNFFSL RLRFNLDLQ PLVDTHKST LLIKVYEDD
451 GQVZMETQR HDGLE
  
```

Sequence coverage: 29.5%

This is one of our Scaffold/protXML results viewed in the Results Applet. Here we can see some of the additional columns available. These include:

For protein hits – the assigned probability value and the percentage of spectra assigned to the protein hit. For peptide hits – the weighting associated with the peptide and the assigned probability. The other common protXML values such as the number of sibling peptide (NSP) etc are also available – the applet lets you dynamically add an remove columns from the report.

The applet also allows you to view the additional information that comes back from Scaffold as an extension to the standard protXML export. This includes which biological sample each protein or peptide was found associated with, and also details about the presence of the protein or peptide match within each of the source peaklist files the Scaffold results were generated from.

Some additional features – Scaffold tries to unambiguously assign peptide matches to a specific protein hit. If Scaffold has assigned a peptide to the current hit (which means it has been used in calculating the protein hit details such as the protein probability, % spectrum ids etc), then it is shown in bold face as we can see here. However, if there is some ambiguity about the assignment then the applet flags this up by showing the hit in italic. Hovering over the match will bring up a tool tip which includes the information about which additional proteins the peptide could have been associated with. Therefore in this case, these peptides were assigned to VIME_HUMAN by Scaffold, but the italic face is telling us that there are alternative assignments. If we now take a look at one of those possible matches – K2C8_HUMAN

MassSpec Integra Results Applet

File Processing Help

Explore

Protein-centric view Approve Unapprove

Searches

HR rank (a)	Accession	Description	Mass	p	Peptide sequences matched	% Spectrum ids	S-090522-00011	CS000155-02-0
133	sp I16553 EPPL2_HUMAN	Hydroxymethylase related protein-2 OS=Homo sapiens GN?EPPL2	62276	1.0	4	0.032	✓	✓
134	sp P15744 GPII_HUMAN	Glycerol-phosphate isomerase OS=Homo sapiens GN?GPII	63998	1.0	4	0.032	✓	✓
135	sp P23284 PP1B_HUMAN	Peptidyl-prolyl cis-trans isomerase B OS=Homo sapiens GN?PP1B	22724	1.0	3	0.032	✓	✓
136	sp P13489 RINI_HUMAN	Placental ribonuclease inhibitor OS=Homo sapiens GN?RINI	49824	1.0	3	0.032	✓	✓
137	sp P08227 K1C19_HUMAN	Keratin, type I cytoskeletal 19 OS=Homo sapiens GN?K1C19	44066	1.0	3	0.032	✓	✓
138	sp P62244 R515A_HUMAN	40S ribosomal protein S15a OS=Homo sapiens GN?RPS15A	14600	0.998	2	0.032	✓	✓
139	sp P18124 RL7_HUMAN	60S ribosomal protein L7 OS=Homo sapiens GN?RPL7	29208	0.998	2	0.032	✓	✓
140	sp P11766 ADH9C_HUMAN	Alcohol dehydrogenase class III ch1 chain OS=Homo sapiens GN?ADH9C	39575	0.998	2	0.032	✓	✓
141	sp P08723 IASP_HUMAN	Brain acid soluble protein 1 OS=Homo sapiens GN?IASP1	22544	0.998	2	0.032	✓	✓
142	sp Q9NWS9 ZNF46_HUMAN	Zinc finger protein 446 OS=Homo sapiens GN?ZNF446	48939	0.837	1	0.032	✓	✓
143	sp Q9R6P7 G35T4_HUMAN	Galactose 3-O-sulfotransferase 4 OS=Homo sapiens GN?GAL35T4	54140	0.837	1	0.032	✓	✓
144	sp P04243 H3_HUMAN	Histone H3.3 OS=Homo sapiens GN?H3F3A	15179	0.837	1	0.032	✓	✓
145	sp P02763 A1AG1_HUMAN	Alpha-1-acid glycoprotein 1 OS=Homo sapiens GN?ORM1	23494	0.837	1	0.032	✓	✓
146	sp P55071 TERA_HUMAN	Transitional endoplasmic reticulum ATPase OS=Homo sapiens GN?VCP	89173	1.0	5	0.026	✓	✓
147	sp P05767 K2C8_HUMAN	Keratin, type II cytoskeletal 8 OS=Homo sapiens GN?KRT8	53525	1.0	6	0.064	✓	✓

Query No. Weight Charge pI (pI) Start End Mass Probability NSP Total Rank Ntc... Peptide Cite... Variable mods S-090522-00011 CS000155-02-0

✓	1365	1.0	2	1060.353	392	460	1	0.95	4.75	5	365	R	K	KLLEGEESR
✓	1366	1.0	2	1060.564	392	460	1	0.95	5.9	2	366	R	K	KLLEGEESR
✓	1940	1.0	2	1185.338	275	284	0	0.95	0.95	1	1940	R	A	KLLEGEESR
✓	1941	1.0	2	1083.761	158	175	1	0.95	0.95	1	1941	K	L	KLLEGEESR
✓	1942	1.0	2	1080.42	316	324	0	0.95	0.95	1	1942	R	L	KLLEGEESR
✓	1943	1.0	2	1128.914	352	361	0	0.95	0.95	1	1943	K	L	KLLEGEESR
✓	1944	1.0	2	1137.538	285	294	0	0.95	0.95	1	1944	K	Y	KLLEGEESR

Peptide match was also found in the following proteins:
 sp|P08670|VIME_HUMAN*
 sp|P08729|K2C7_HUMAN*
 * peptide was assigned by Scaffold to indicated accession

Protein View | Taxonomy | Family | Annotation | Spectrum | Ions matched | Error distribution | Top ten query matches | Search Summary

Protein view for sp|P05767|K2C8_HUMAN Keratin, type II cytoskeletal 8 OS=Homo sapiens GN?KRT8

```

1 SIRVTRSYK VSTSGPFAFS SRSTYSGFOS RLSSSSFVY GSSMFRGGIG
51 GPGGASGSGS GITAYTWSG LLSPLYLEVD PMIQAVRTVE KEQIKTLANK
101 PASTYDTRFP LEQNSDELT DSGLLQDQRT ARSNENRTE STIHLRPGIL
151 ETLGQEDLM EEARLGRGGL VEDNKHQVSD EINDSTFENR EPLYLSDSDVD
201 EAYNNVYELR SRLLEGLTDEI NFPLQYEEH IRELQSQISD TSYVLSNDNS
251 RSLNDSDIIA EYKAYVEYIA NRSFAEAESE YDIXEDELQS LAGHSGDLR
301 FRKTEISEDN PMTIRPQREI DALRQDRAEL EAAIADAEQR GELAIADANA
351 RLSRERAKLQ RAGQDAPQL RYDQELRHKV LAETETATVY NELLRQKRSR
401 LESQHMMSI HTKTKGTYAG GLSSATYGLT SFGLSYSLGS SFQSGAGSSS
451 FSRSTSSSAV VYQKLETRGQ KLYSESSDVL PK
  
```

Sequence coverage: 13.7%

We can see those matches again, but they're not being shown in bold which means Scaffold did not assign the match to this protein. If we take a look at the tooltip, we're now getting the additional information that Scaffold in fact assigned these peptides to VIME_HUMAN.

The image shows two dialog boxes from the Mascot software interface. The top dialog, titled "Peptide centric filter", contains a text field with the SQL query "Prot:ML peptide prob >= 0.95". Below this, a dropdown menu is set to "Probability", followed by a comparison operator ">=" and the value "0.95". There are "Add" and "Reset" buttons. The main area of the dialog displays the SQL query "PEPTIDE_PROBABILITY >= '0.95'". At the bottom are "Apply", "Clear", and "Cancel" buttons.

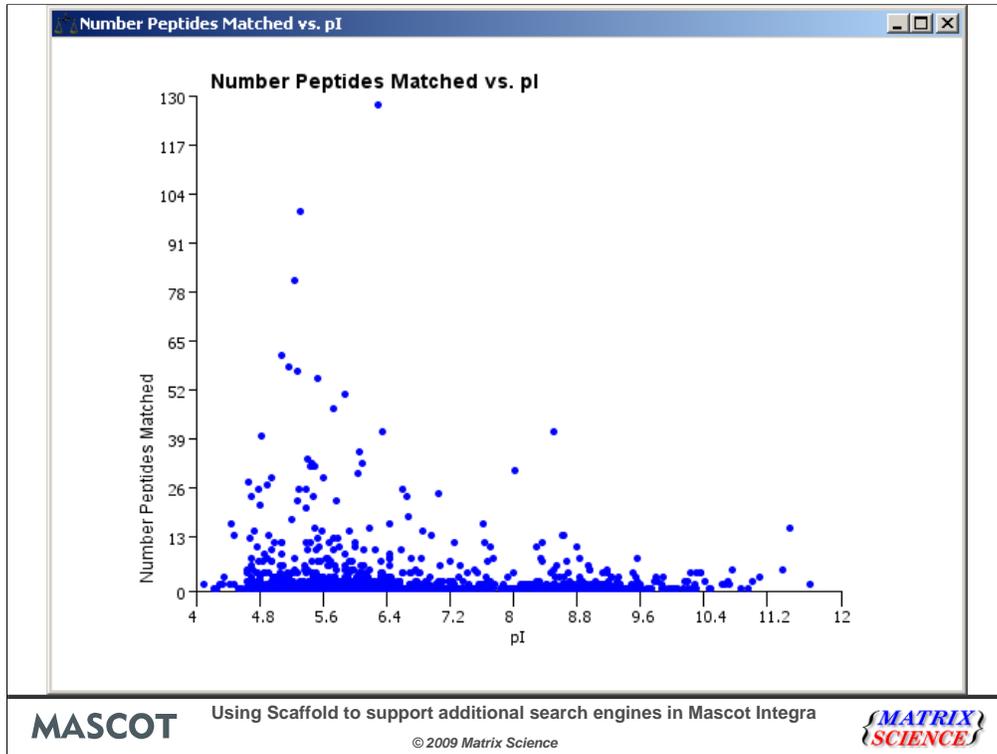
The bottom dialog, titled "Protein hit filters*", is a table with the following columns: filter name, comparison operator, and value. The "No peptides matched after filter" row has the value "2".

Filter Name	Operator	Value
Protein Mass	<	
Probability	<	
No peptides matched	<	
Peptide sequences matched	<	
No peptides matched after filter	>=	2
Percentage coverage	<	
Protein length	<	
Protein pI	<	
% Spectrum Ids	<	

At the bottom of the "Protein hit filters*" dialog are "Apply", "Clear", and "Cancel" buttons.

MASCOT Using Scaffold to support additional search engines in Mascot Integra **MATRIX SCIENCE**
© 2009 Matrix Science

The applet also allows extensive filtering of the results at both the peptide level (via user definable SQL filters) and the protein level



Within the Applet you can also plot graphs from either the protein or peptide level information – here we’re plotting the number of peptides matched against the calculated protein isoelectric point; we can see that there is a cluster of proteins with a larger number of peptide matches around the pI range of 4.5-6.4.

Viewing results 3: BlastCluster

- Uses NCBI BLAST to cluster protein hits together based on sequence homology/identity
- Can use to compare between imported Mascot and protXML results
- Can group together searches based on shared protein clusters

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

Mascot Integra can produce a cluster report using NCBI BLAST Cluster to group together protein hits based on the primary protein sequence and not on shared peptide matches.

This allows us to generate a report which directly compares between Mascot and protXML searches which have been imported into the system. The source searches can be grouped together based on their shared protein clusters.

BLAST clustering percentage identical residues threshold: 40
 BLAST clustering minimum length coverage threshold: 0.5
 Specified thresholds required on one strand only

Mascot Searches used:

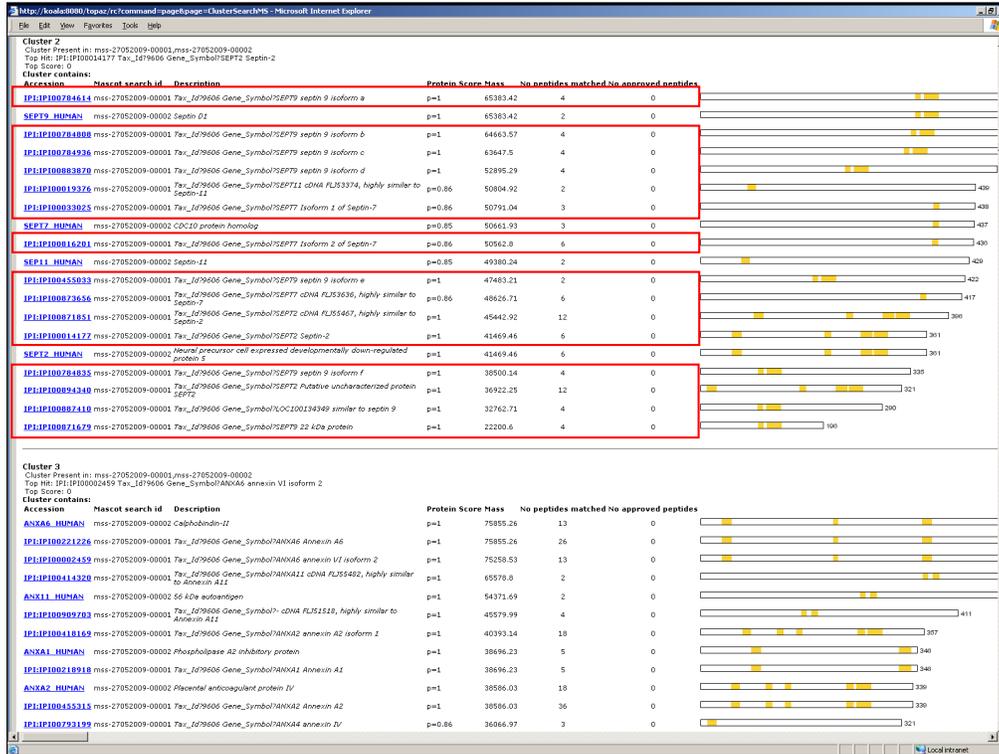
Search Title	Search Database	Sample id:
Group 1		
ms-27052009-00002	C:\Databases for talkiprot_human_plus_random_05082009.fasta	S:090522-00011
ms-27052009-00001	C:\Databases for talkipi_HUMAN_plus_decoy_07052009.fasta	S:090522-00011

Comparison table

	m	m
	4	4
	4	4
	-	-
	2	2
	7	7
	0	0
	5	5
	2	2
	0	0
	0	0
	9	9
	-	-
	0	0
	0	0
	0	0
	0	0
	2	1

Cluster	Gene Symbol	Gene Name	Match
Cluster1	Gene_Symbol7FGA	Isoform 1 of Fibrinogen alpha chain	✓✓
Cluster2	Gene_Symbol7SEPT2	Septin-2	✓✓
Cluster3	Gene_Symbol7ANXA6	annexin V1 isoform 2	✓✓
Cluster4	Gene_Symbol7ITIH3	Isoform 1 of Inter-alpha-trypsin inhibitor heavy chain H3	✓✓
Cluster5	Gene_Symbol7C4A	Complement C4-A	✓✓
Cluster6	Gene_Symbol7PLEC1	Isoform 1 of Plectin-1	✓✓
Cluster7	Gene_Symbol7POSTN	Isoform 1 of Periostin	✓✓
Cluster8	Gene_Symbol7STAT1	Isoform Alpha of Signal transducer and activator of transcription 1-alpha/beta	✓✓
Cluster9	Gene_Symbol7HSPAS	HSPAS protein	✓✓
Cluster10	Gene_Symbol7BGN	Biglycan	✓✓
Cluster11	Gene_Symbol7HSP90B1	Endoplasmic	✓✓
Cluster12	Gene_Symbol7PP1A	Peptidyl-prolyl cis-trans isomerase A	✓✓
Cluster13	Gene_Symbol7NID1	Isoform 1 of Nidogen-1	✓✓
Cluster14	Gene_Symbol7FTUD2	116 kDa US small nuclear ribonucleoprotein component	✓✓

Here we've clustered together the results from the two different searches – as we can see the top ranked clusters are present in both searches.



If we take a closer look at cluster 2, we can see that there are far more protein hits from IPI Human in the cluster than there are from the SwissProt human search. The SwissProt search results has matches to Septin 2, 7, 9 and 11. The IPI Human search results have the members of the same four Septin family members, but has additional members of each Septin subgroup present with 6 different isoforms of Septin 9 for example.

Viewing results 4: Custom reports

- Full access to the results schema
- Can generate report based on Excel template or from a Java custom report class
- Allows for generation of more complex, customised reports

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

The final reporting option is to generate your own custom reports. The Mascot Integra results schema is fully described allowing you to put together your own SQL based reports. These can be generated from either an Excel template or from a custom Java class. Generating the report from Java is the most powerful option, but either method allows you to generate your own, complex and customised reports.

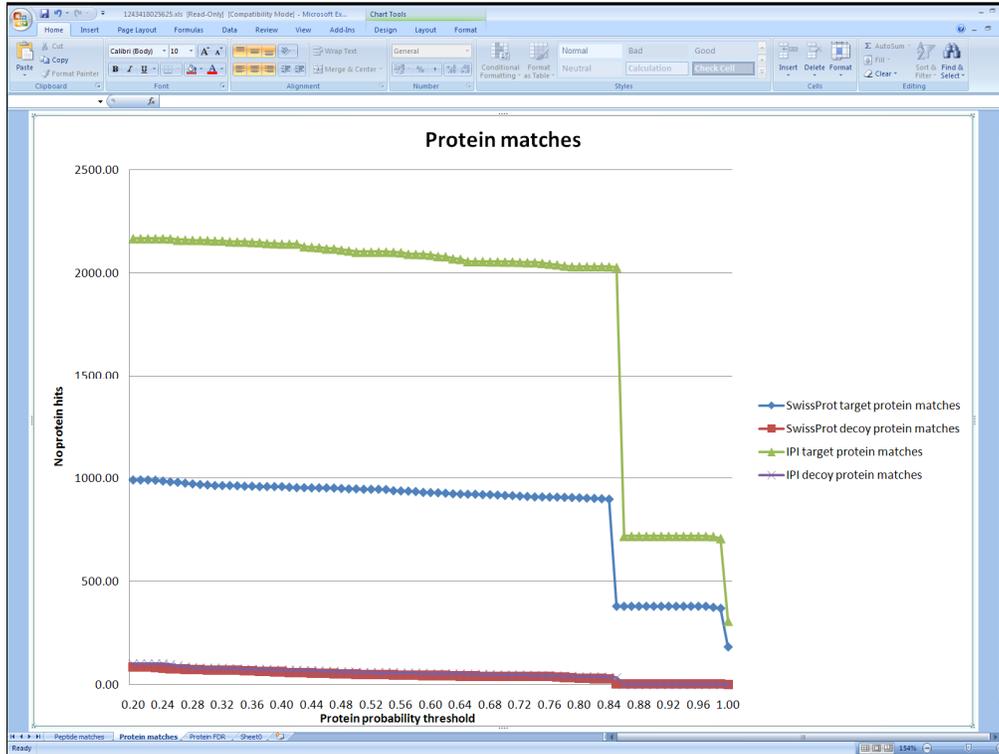
The screenshot shows an Excel spreadsheet with the following data structure:

experiment_id	Search: mss-27052009-00002 Human SwissProt sequences				Search: mss-27052009-00001 IPI Human									
	p threshold	No. protei	No. decoy	FDR (%)	No. peptid	No. decoy	FDR (%)	p threshold	No. protei	No. decoy	FDR (%)	No. peptid	No. decoy	FDR (%)
5	0.20	990.00	84.00	8.48				0.20	2166.00	100.00	4.62			
6	0.21	990.00	84.00	8.48				0.21	2166.00	100.00	4.62			
7	0.22	990.00	84.00	8.48				0.22	2166.00	100.00	4.62			
8	0.23	989.00	83.00	8.39				0.23	2166.00	100.00	4.62			
9	0.24	988.00	79.00	8.02				0.24	2166.00	99.00	4.57			
10	0.25	981.00	77.00	7.85				0.25	2166.00	96.00	4.43			
11	0.26	975.00	75.00	7.66				0.26	2159.00	90.00	4.17			
12	0.27	975.00	74.00	7.59				0.27	2159.00	84.00	3.89			
13	0.28	971.00	73.00	7.52				0.28	2158.00	83.00	3.85			
14	0.29	968.00	72.00	7.44				0.29	2158.00	80.00	3.71			
15	0.30	968.00	70.00	7.25				0.30	2157.00	80.00	3.71			
16	0.31	963.00	70.00	7.27				0.31	2155.00	80.00	3.71			
17	0.32	963.00	69.00	7.17				0.32	2155.00	80.00	3.71			
18	0.33	963.00	69.00	7.17				0.33	2151.00	77.00	3.58			
19	0.34	962.00	69.00	7.17				0.34	2150.00	75.00	3.49			
20	0.35	960.00	67.00	6.98				0.35	2150.00	73.00	3.40			
21	0.36	960.00	66.00	6.88				0.36	2146.00	73.00	3.40			
22	0.37	958.00	64.00	6.68				0.37	2147.00	73.00	3.40			
23	0.38	958.00	63.00	6.56				0.38	2143.00	71.00	3.31			
24	0.39	958.00	62.00	6.47				0.39	2142.00	70.00	3.27			
25	0.40	958.00	62.00	6.47				0.40	2140.00	69.00	3.22			
26	0.41	953.00	58.00	6.07				0.41	2140.00	68.00	3.18			
27	0.42	953.00	57.00	5.98				0.42	2140.00	68.00	3.18			
28	0.43	953.00	57.00	5.98				0.43	2127.00	67.00	3.15			
29	0.44	952.00	56.00	5.88				0.44	2126.00	67.00	3.15			
30	0.45	952.00	55.00	5.78				0.45	2123.00	65.00	3.06			
31	0.46	952.00	54.00	5.67				0.46	2117.00	63.00	2.98			
32	0.47	951.00	53.00	5.57				0.47	2117.00	62.00	2.93			
33	0.48	949.00	51.00	5.37				0.48	2111.00	61.00	2.89			
34	0.49	947.00	51.00	5.39				0.49	2107.00	60.00	2.85			
35	0.50	947.00	50.00	5.28	4388.00	106.00	2.42	0.50	2101.00	59.00	2.81	4347.00	112.00	2.58
36	0.51	945.00	49.00	5.19	4382.00	104.00	2.37	0.51	2101.00	58.00	2.76	4343.00	109.00	2.61
37	0.52	945.00	49.00	5.19	4380.00	101.00	2.31	0.52	2101.00	57.00	2.71	4341.00	105.00	2.42
38	0.53	945.00	49.00	5.19	4371.00	99.00	2.26	0.53	2101.00	56.00	2.67	4334.00	102.00	2.36
39	0.54	944.00	48.00	5.08	4368.00	97.00	2.22	0.54	2101.00	55.00	2.67	4327.00	96.00	2.22
40	0.55	938.00	47.00	5.01	4365.00	96.00	2.20	0.55	2100.00	56.00	2.67	4324.00	95.00	2.20
41	0.56	937.00	46.00	4.91	4359.00	96.00	2.20	0.56	2098.00	55.00	2.62	4323.00	91.00	2.11
42	0.57	936.00	45.00	4.81	4353.00	95.00	2.18	0.57	2091.00	55.00	2.63	4320.00	91.00	2.11
43	0.58	934.00	45.00	4.82	4348.00	94.00	2.16	0.58	2090.00	55.00	2.63	4316.00	91.00	2.11
44	0.59	930.00	43.00	4.62	4345.00	94.00	2.16	0.59	2089.00	53.00	2.54	4312.00	91.00	2.11
45	0.60	929.00	43.00	4.63	4338.00	92.00	2.12	0.60	2086.00	52.00	2.49	4310.00	91.00	2.11
46	0.61	928.00	43.00	4.63	4335.00	91.00	2.10	0.61	2080.00	52.00	2.50	4304.00	88.00	2.04
47	0.62	926.00	43.00	4.64	4334.00	91.00	2.10	0.62	2079.00	52.00	2.50	4302.00	86.00	2.00
48	0.63	923.00	42.00	4.55	4332.00	91.00	2.10	0.63	2069.00	52.00	2.51	4299.00	84.00	1.95
49	0.64	922.00	41.00	4.45	4327.00	91.00	2.10	0.64	2066.00	52.00	2.52	4297.00	84.00	1.95
50	0.65	921.00	40.00	4.34	4321.00	89.00	2.06	0.65	2065.00	49.00	2.38	4292.00	83.00	1.93
51	0.66	921.00	40.00	4.34	4315.00	85.00	1.97	0.66	2065.00	49.00	2.38	4285.00	81.00	1.89
52	0.67	919.00	39.00	4.24	4313.00	85.00	1.97	0.67	2065.00	48.00	2.34	4280.00	80.00	1.87
53	0.68	919.00	39.00	4.24	4309.00	84.00	1.95	0.68	2064.00	48.00	2.34	4275.00	79.00	1.85
54	0.69	917.00	39.00	4.25	4304.00	83.00	1.93	0.69	2064.00	46.00	2.24	4273.00	78.00	1.83
55	0.70	915.00	39.00	4.26	4300.00	81.00	1.88	0.70	2063.00	45.00	2.19	4271.00	78.00	1.83
56	0.71	914.00	38.00	4.16	4295.00	79.00	1.84	0.71	2063.00	44.00	2.14	4263.00	77.00	1.81
57	0.72	913.00	38.00	4.07	4293.00	78.00	1.82	0.72	2061.00	44.00	2.15	4260.00	76.00	1.78

In this example, I've written a custom Java class to generate an Excel sheet which pulls out the target and decoy protein and peptide matches for each search at different probability thresholds. Because the data has been exported from Scaffold, unassigned peptide matches are not included – only peptide matches assigned to a protein hit are present. Scaffold does try and assign a spectrum unambiguously to a protein hit – where ambiguity was reported (as we saw in the results earlier), then this will be taken into account (if a peptide is assigned to both a target and decoy sequence for example).

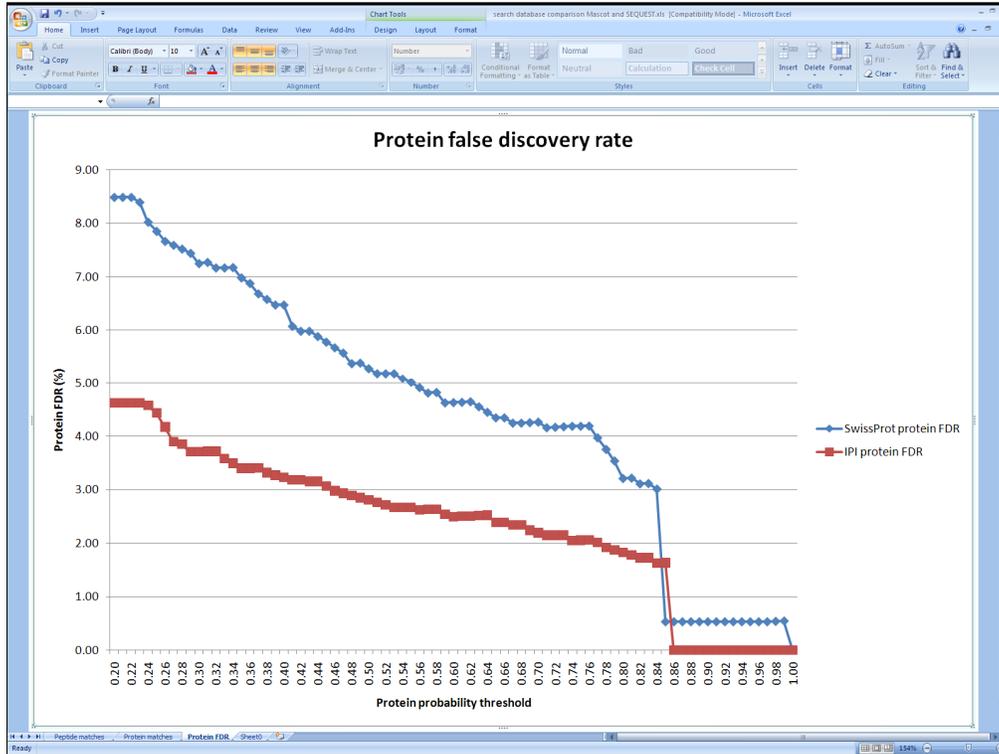
To remind you of the two sets of search conditions:

1. Search against the human protein sequences in SwissProt, approx 20,000 sequences
2. Search against the IPI Human database, approx 80,000 sequences



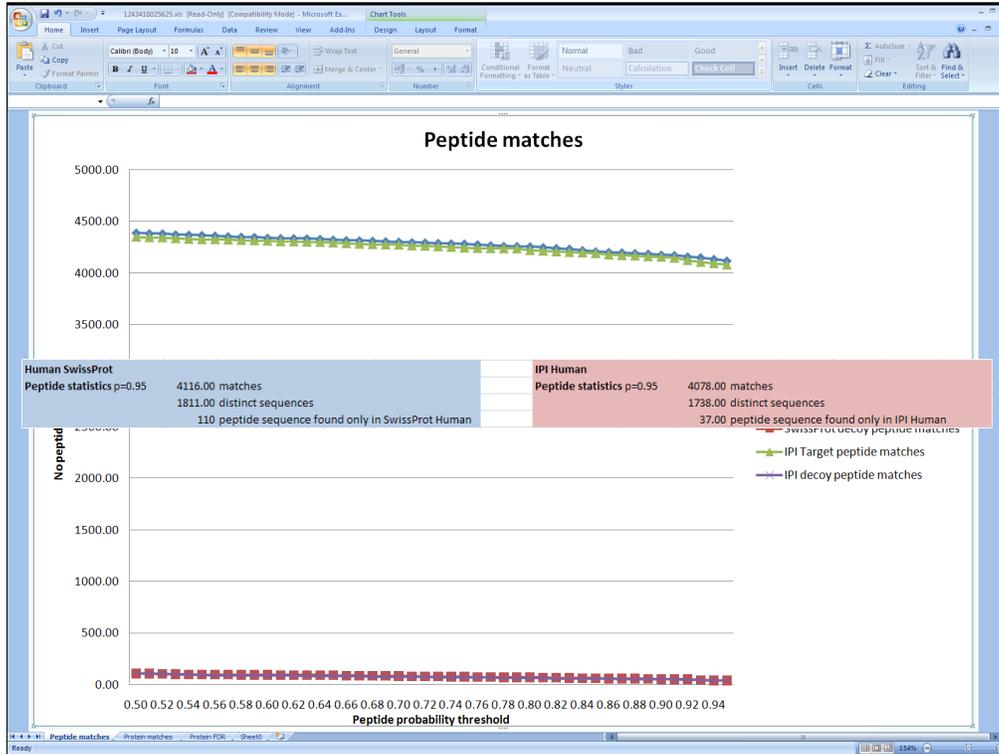
Please note that this is not a ROC plot.

This graph shows the numbers of target and decoy protein hits against each search database. The number of protein hits for both search conditions (for both target and decoy) show a rapid fall in the number of protein hits at a probability of around 0.84, while an increase in probability threshold from 0.99 to 1 results in a large decrease in the number of target hits without a significant decrease in the number of decoy hits. Looking at the results at a probability threshold of 0.9, we can see that IPI human is giving us around twice as many protein hits as searching SwissProt.



Please note that this is not a ROC plot.

Which is more clearly shown on this graph which shows the protein hit FDR for each search. As you can see, the protein false discovery rate shows the same pattern from searching either database. However, the protein FDR from searching IPI Human does appear to be consistently lower than from searching human SwissProt. This probably reflects the fact that the decoy segment of the databases don't have the redundancy present in the target portion since the decoy sequences are randomised sequences; hence with the larger database the increase in matches to the random sequences in the decoy section of the database isn't as great as the increase in matches to redundant protein sequences in the target database. This shows one of the difficulties in trying to obtain reliable protein false discovery rates.



Here we have the absolute number of target and decoy peptide matches from both databases. Despite the fact that IPI Human is giving us twice as many protein hits, we're getting pretty much identical numbers of peptide matches from both databases for both target and decoy. In fact the we have very slightly lower numbers from IPI Human. This probably reflects the larger database size tightening up the statistics – for example the Mascot identity thresholds would be higher which could remove some more borderline matches. Therefore, the peptide discovery and false discovery rates are in fact approximately equal for the two databases. Taking a closer look at the peptide dataset, we've pulled out some additional information for each search. Taking the probability threshold at $p=0.95$, the SwissProt Human search has given us 4116 high confidence target matches, representing 1811 distinct peptide sequences, while IPI Human has 4078 high confidence matches representing 1738 distinct sequences – so the numbers are very similar. The SwissProt dataset has 110 peptide sequences unique to it, while the IPI Human dataset has 37 sequences unique to it, so searching both databases has given us a total of 147 peptide sequences which might otherwise have been missed.

Database comparison summary

- **No significant change in the number of peptides identified**
- **More protein hits identified in IPI Human**
 - Increased redundancy
 - Protein threshold $p=0.9$
 - Swissprot: 357 target protein hit ranks
378 protein hits
 - IPI Human: 349 target protein hit ranks
718 protein hits

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

So to summarise; The peptide set identified by searching both databases was very similar, despite the much larger size of the IPI Human database. Searching IPI Human did increase the total possible protein hit list though. This was largely due to an increase in redundancy of identified protein hits (such as the increase in the number of Septins matched). Depending on what you are looking for, searching SwissProt may well have been sufficient for this dataset.

Custom report 2: High confidence matches in both SEQUEST and Mascot

- SwissProt dataset (same peaklist)
- Mascot .dat result file imported normally
- SEQUEST .srf file processed with Scaffold
- Generate shared high confidence list
 - SEQUEST - Scaffold p of 0.95
 - Mascot - Score > homology threshold

MASCOT

Using Scaffold to support additional search engines in Mascot Integra

© 2009 Matrix Science

MATRIX
SCIENCE

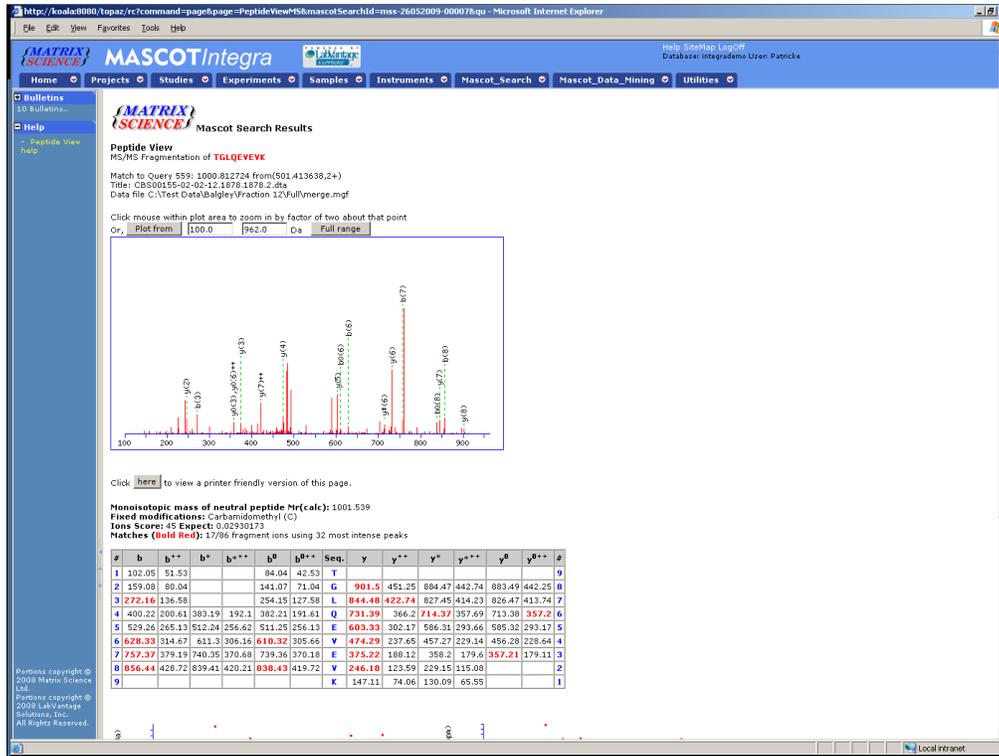
Something else you would want to be able to do easily is to compare results between search engines and runs from within the Integra database, allowing for large scale comparisons.

Using the same search results (just the SwissProt results), I've imported the Mascot result files into Integra (so they're present separately as standard Mascot results), and processed and imported the SEQUEST results with Scaffold separately.

I've then put together a report which pulls out the peptide matches which have a high confidence in both search engines (p 0.95 for SEQUEST/Scaffold and a score above the homology threshold for Mascot results).

Peptide sequence	Mascot peptide view
DTLMISR	Q125 Q132
RLEELR	Q404
TLYGFGG	Q20 Q21 Q22 Q23
LYEGR	Q367
AFPEVSK	Q183 Q184 Q186 Q187 Q188 Q189 Q190 Q191 Q192 Q193 Q194 Q195 Q196 Q197 Q198 Q199 Q200
AVLTIDEK	Q250 Q253 Q259 Q263
EKVESELR	Q518
FMETVAEK	Q436
GTNESLR	Q320
IREEYPDR	Q839 Q840 Q841
LLASVEER	Q350
LSKGEER	Q521
YLGEVYVK	Q553
VSEEVYNR	Q869 Q870 Q871 Q872 Q873 Q875
TYETILEK	Q465 Q479 Q480 Q482 Q490 Q492
MLVDELK	Q587
LYDAYELK	Q631
LVFDEYK	Q669
LQEEMLQR	Q773 Q774 Q775 Q776
LALLEEAR	Q333
LALLEEAR	Q333
FTFEEAAK	Q393 Q396 Q397
FHVEEGK	Q444 Q446 Q447 Q451
EKIETELK	Q642 Q643 Q646
AQLVIDEK	Q1083
CCTESLVNR	Q1081
LDKENALDR	Q801 Q802 Q805 Q806 Q807 Q809 Q810 Q812 Q814 Q816 Q820 Q824 Q827
KLLEGEER	Q762 Q763 Q765 Q767 Q768 Q770
KLLEGEER	Q575
KEEGEAFAR	Q695
IISNEGYR	Q862 Q863
HEQEYMEVR	Q1599 Q1600
GYLQEGDR	Q731 Q733 Q734 Q735
FMVQEEFSR	Q1325 Q1326 Q1332
FOESEERPK	Q1130
TGLQEVK	Q559 Q563 Q566 Q567 Q570 Q883
SFPEVYVDR	Q879
RYDQEVYK	Q112
RLVDDIK	Q97
MSTEEIQR	Q102
MEIEGILK	Q1105
LSYEGEVYK	Q667
LSPVGEEMR	Q682 Q712 Q713 Q714 Q715 Q718
LSEELSGGR	Q410
LQAEIDNIK	Q702
...	...

Here we have our list of shared high confidence peptide matches from SEQUEST and Mascot. Because we've looked at the source Mascot results and compared those with the SEQUEST/Scaffold protXML export we have access to the full peptide information from the Mascot search. Next to each peptide sequence I've incorporated a reference to the source Mascot query match – which is actually a link to the peptide view from Integra. Click on the cell to follow the link (here we're going to look at query 559) and the peptide view will open in a browser window



And here we have our match from Mascot.

Acknowledgements

Brian Searle (Proteome Software inc)
LabVantage Solutions Inc