MS/MS search   8.9

The ms/ms search of GPMAW is based on the public domain search engine 'X! Tandem'. The X! Tandem program is a professional class search engine; Although it is able to perform proteome wide searches, the GPMAW implementation is targeted towards characterizing individual proteins and small collections of proteins. It should be possible to analyze large databases, but the handling and analysis of these data has not been thoroughly tested using the GPMAW implementation, so other search engines (e.g. Mascot, Sequest etc.) can be recommended.

The focus of the GPMAW implementation is ease of use and fast analysis time. Currently all the functionality of X! Tandem is not implemented, but this is likely to improve in future versions of GPMAW.

ⓘ   **Note:** For more details on the X! Tandem search engine, particularly all the parameters that can be adjusted, see the web site of The Global Proteome Machine Organization, http://www.thegpm.org.

The MS/MS search is implemented as an external search program called 'tandem.exe'. GPMAW prepares all data and saves them in files on disk. Then Tandem is called with in-line parameters that specifies which files to load. X! Tandem then runs in a separate window. **Note:** while X! Tandem runs, you should not make changes to GPMAW. Upon completion of the search, X! Tandem closes and a message is sent to GPMAW which opens the result file created by X! Tandem.

The ms/ms search is accessed through the main menu Search | MS/MS search or by pressing the F5 key.

An example file has been included with the GPMAW installation; please see section 8.14 for details.

MS/MS data input   8.10

When you select the MS/MS search option, you are greeted with a window containing three tabs in the left-hand side:

Input: This is where you define your search: specify input mass file, search mass or database, enzyme used, output file and parameters.

Output: This is the primary output: List of all high-scoring proteins and all peptide hits.

Hit: On this tab you can analyze each individual hit on the 'Output' page.
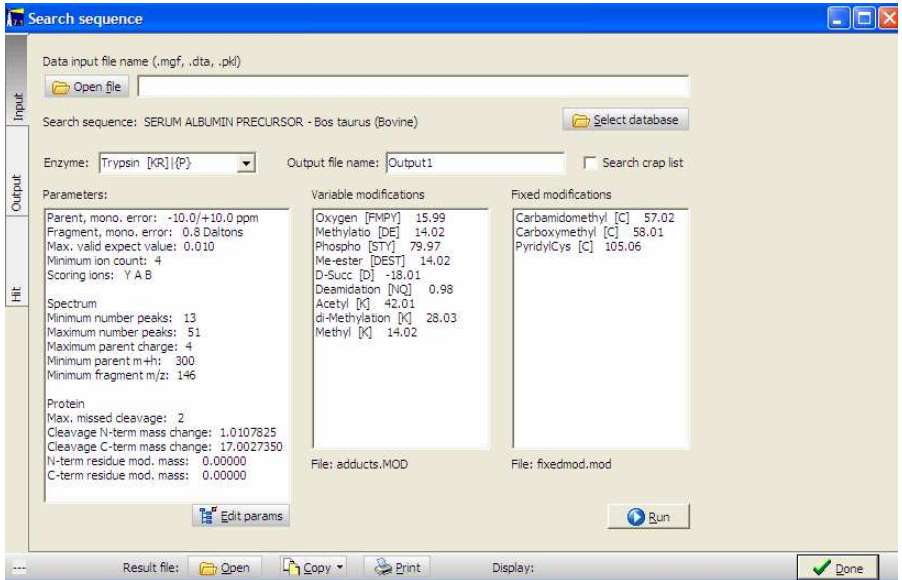
Performing a search:

1) Open your ms/ms peak list file (can be in mgf, dta or pkl format).
2) Specify sequence or database to search
3) Specify cleavage enzyme and output file name (a default name is created by GPMAW).
4) Check and set search parameters
5) Set variable and fixed modifications
6) Click the 'Run' button
7) X! Tandem opens in a command line window where you can follow the progress of the search.
8) When the search is done, GPMAW loads the result file and displays it on the 'Output' tab, see below.

Open file

Pressing the Open file button, opens a standard dialog where you select your ms/ms data file. The file has to be in .mgf, .dta or .pkl format (appendix A.3).

Search sequence

By default GPMAW wants to analyze the topmost sequence window on the desktop. This is selected in the 'Search sequence' line below the data file input. If this is used for analysis, it will be saved in FastA format in the xtandem directory with the name 'searchSeq.fasta'. As this is a temporary file, it will always overwrite previous copies of the same name.

If you want to search a database instead, you select this by pressing the 'Select database' button. **Note:** this database has to be in FastA format, a standard GPMAW file does not work.

ⓘ **Hint:** If you want to make a FastA formatted file consisting of a limited number of sequences, an easy way is to load the relevant sequences in GPMAW and save them as a single FastA formatted file using the File | Export sequence command (Chapter 2.8).

Enzyme:

In this box you select the enzyme used for the digestion of the protein(s). Currently only the most common enzymes are implemented.

If the enzyme of your choice is not on the list, you can select the last option '- user defined –'. This opens an input box where you can enter any most kinds of enzyme specificity.

The enzyme specificity is a formatted string that contains three elements: two specifiers and a separating vertical bar.

The specifiers are either accepting residues in square brackets [ ] or inhibiting residues in curly brackets { }.

The character representing any residue is X.

This means that the cleavage of the enzyme trypsin is specified as [RK]|{P} - cleavage takes place after Arg or Lys, but not if the following residue is Pro. Endoproteinase Asp-N is defined as [X]|[D] - cleavage takes place after any residue but before Asp.

Multiple rules can be strung after each other separated by a comma e.g. [RK]|{P},[X]|[D] - cleavage following either a tryptic cleavage site or Asp-N cleavage site.

[X]|[X] - cleavage after any residue (remember to increase 'missed cleavages').

Output file name:

An output file name is suggested by GPMAW based on the input file name combined with the current date and time (hours and minutes). The file name can be freely edited. Note that if you run multiple searched with the same input file but varying the parameters, you have to change this file name as GPMAW does not check for an existing file of the same name.

This file is saved in the xtandem directory.

Search crap list:

The cRAP list is a file, 'crap.fasta.pro', containing a number of advantageous proteins, primarily keratin, bsa and trypsin, maintained by the Global Proteome Machine Organization (www.thegpm.org). These proteins are typical contaminants in proteolytic experiments and may in some cases obscure your expected data. When this option is checked, your data file will additionally be searched against these proteins. If the file 'crap.fasta.pro' is not present in the xtandem directory, the option will be grayed.

> ⓘ **Note:** The .pro format is a FastA file compiled with the fasta_pro.exe program in order to make a more compact file that is faster to search by X!Tandem. The fasta_pro.exe program is included on the disk and can be used to compact your own FastA files. The current cRAP file can be downloaded from ftp://ftp.thegpm.org/fasta/cRAP, in FastA format. A list of proteins in the file can be found at http://www.thegpm.org/crap/index.html. Here you will also be able to download updated versions of the list.
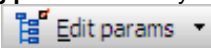
Parameters:

If you need to change the search parameters, you select the 'Edit params' button, which opens a dialog box where all parameters can be edited (see section 8.11 below). Alternatively, sets of parameters can be saved and re-loaded from the drop-down menu activated by the down-arrow on the button.

Select variable and fixed modifications from the two list boxes by click once on each modification to include in the search. You deselect a choice by clicking once again.

> The **variable modifications** are loaded from the currently selected modification file (standard GPMAW modification file), see Chapter 4.3 for details on how construct and edit this file.

> The **Fixed modifications** is also a standard GPMAW modification file, but with the name 'fixedmod.mod'. You can load and edit this file in the 'Edit modification dialog', Chapter 4.3. The first time you run the MS/MS search function, this file will be created automatically.

Saving parameters: If you click on the down-arrow on the 'Edit params' button ![Edit params] , you will get a menu that enables you save and load up to five sets of parameters. This is handy when using different instruments or experimental settings.

Refinement

The refinement option ![Enable refinement] makes X! Tandem perform a second pass search through the identified protein hits. This makes it possible to define a second round of modifications, particularly useful if you search for multiple modifications on the same residue (e.g. methylation and acetylation of lysine residues).

In addition you can search for
'**Semi cleavage**': This option searches for peptides that fulfill cleavage requirements only at one end of the peptide. Trimming at the N- and C-terminally end of a protein can typically be found using this option.
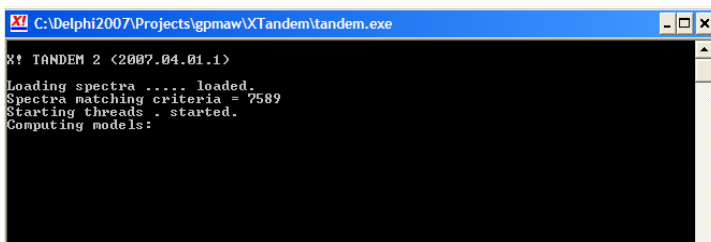'**Spectrum synthesis**': If checked, the following bond cleavages are favored relative to other cleavages (e.g. results in a higher score): N-terminal to Pro and C-terminal to Asp, Glu, Asn, Gln, Val, Leu and Ile.
'**N-terminal acetylation**': searches for N-terminal acetylation.

Please note that the run time for 'refinements' is often much longer than the initial search.

Run

When you press the 'Run' button, the parameters are saved in a file called 'input.xml'. This file works in conjunction with the file called 'default_input.xml' which contains all the available parameters for the X! Tandem search engine. The parameters in 'input.xml' override the corresponding values found in 'default_input.xml'. A file called 'taxonomy.xml' is also saved, which allows for the use of databases with taxonomy information, however, this function is not currently used by GPMAW.

GPMAW then calls the XTandem program with the given parameters. The XTandem program opens in a separate command window, where you can follow the progress of the search.

Upon completion of the search, the window closes and the results are saved to a file on disk with the given 'output file name' and the extension '.xml'.

This file is then read by GPMAW, parsed and displayed on the 'Results' tab (see section 8.12 below). The saved file can at any time later be opened in GPMAW by selecting the 'Result file: Open' button at the bottom of the ms/ms search window.

ⓘ **Hint:** The .xml result file can also be viewed on the Proteome Machine home page. Go to www.thegpm.org, select one of the search sites in the left-hand margin and on the search page, you can select 'view saved xml data' in order to see another representation of your data. This can also be used to view data that you ship to colleagues.

When you edit the parameters, the following dialog box is shown:

Parameters:

Parent mass error: This is the precision of the parent ion, and is set as a plus and minus error to accommodate the (few) mass spectrometers that have an asymmetric mass profile. The precision can be set as either ppm (parts per million) or Da. (Dalton).

Fragment mass error: The precision of the fragment ions. The error can be set as ppm or Da.

Total # peaks used: Maximum number of peaks used for identification of the target peptide.

Minimum # of peaks used: Minimum number of peaks required for a spectrum to be considered.

Minimum parent mz: Minimum M+ H+ required for a spectrum to be considered.

Minimum fragment mz: Minimum fragment m/z to be considered.

Maximum parent charge: Highest charge of parent ion to be considered.

Max. e-value: Highest e-value for peptides to be recorded in the output list.

Minimum ion count: Sets the minimum number of ions required for a peptide to be scored.

Max. missed cleavages: Largest number of missed cleavages considered for a given peptide.

Include ions: Allow the checked ions to be used in scoring.

Refine search: Controls whether the refinement module of X! Tandem is used. Setting of these parameters is not currently implemented in GPMAW, but the user can set them manually in the 'default_input.xml' file.

N-terminal mass change: Moiety added to the peptide N-terminus upon cleavage.

C-terminal mass change: Moiety added to the peptide C-terminus upon cleavage.

N-term modif. mass: Moiety added to the N-terminus of the protein.

C-term modif. mass: Moiety added to the C-terminus of the protein.

ⓘ **Note:** For more details on the X! Tandem parameters see the web site of The Global Proteome Machine Organization, http://www.thegpm.org in the API section.

MS/MS search results 8.12

The output tab is divided into three panels:

Sequence hits are listed at the top. This is a list of all the models in the database that fit to the current search. Each hit is listed with log of e-value, number of peptide hits, unique hits and name of protein. In the above example, a test run on bovine albumin, the protein has been determined with a score of log e at -189.7, 40 hits, none of which are unique, as the same protein is part of the crap file, which of course have the same hits. In addition

porcine trypsin has been found with a single hit, and human albumin with 6 hits, all of which are shared with bovine albumin.

To the right of the protein names are listed a model number, this can be correlated with the model numbers in the peptide list below.

Double-click on a model name to view the peptide hit results on the 'Hit' tab.

A graph at the right-hand side of the window shows the precision of all the peptide hits of the currently selected protein model. The graph has the x-axis in ppm (auto scaled) and mass values along the y-axis (500 – 4000).

The bottom table shows all the peptide hits as they have scored during the search:

Spec id: Spectrum id, location in the input mass list.

Top model hit: Name of protein hit.

e-value: The score of the peptide hit.

Z: Charge of parent ion.

MH+ exp: Experimental mass.

MH+ theo: Theoretical mass value calculated based on the database hit.

Delta: Deviation between experimental and theoretical mass.

From – To: Peptide location in the model sequence.

Pre: Residues prior to the peptide hit.

Seq: Sequence of peptide hit.

Post: Residues (up to 4) following the peptide hit.

Miss: Number of missed cleavages in the peptide hit.

Modifs: Modifications found in the peptide hit. These are labeled as position in the sequence followed by '@' and finally the mass of the modification. Multiple modifications can be listed after each other, all of which have been found in the same peptide hit.

Models: Number of the model(s) that have been allocated this peptide hit. In this case all of the first hits have model number 1 and 2 (bovine serum albumin, first and third in the model protein list above).

Check box: When copying the list to the clipboard, the menu option 'checked peptides only' will only copy those lines where this check box is checked. On the 'Hit' tab only the top hit for each peptide in the protein sequence will be selected when the 'unique' option has been selected.

ⓘ **Note:** The results of ms/ms searches are always saved in .xml files in the search directory. You can at any point load a previous result file by clicking

on the 'Result file: Open' button and select an older file. Result file: 📂 Open
If you perform multiple searches on the same input data file and you want to keep the intermediate results, please remember to change the output file name as GPMAW only sets the name when you load the input file.

When you have double-clicked on a protein name on the output tab, GPMAW switches to the 'Hit' tab and shows the details of the peptides making up the hit.


The Hit tab shows a three window layout:

At the very top is a peptide information line which shows details on the currently selected peptide.

Below the peptide line is a two-tabbed window, which initially shows the protein sequence with the total coverage and each identified peptide highlighted in red.

At the bottom left is a list of all peptides belonging to this protein, and to the right is a window showing the corresponding ms/ms spectrum.

When you select a peptide in the table, the topmost tabbed list changes to show the corresponding fragment ions:
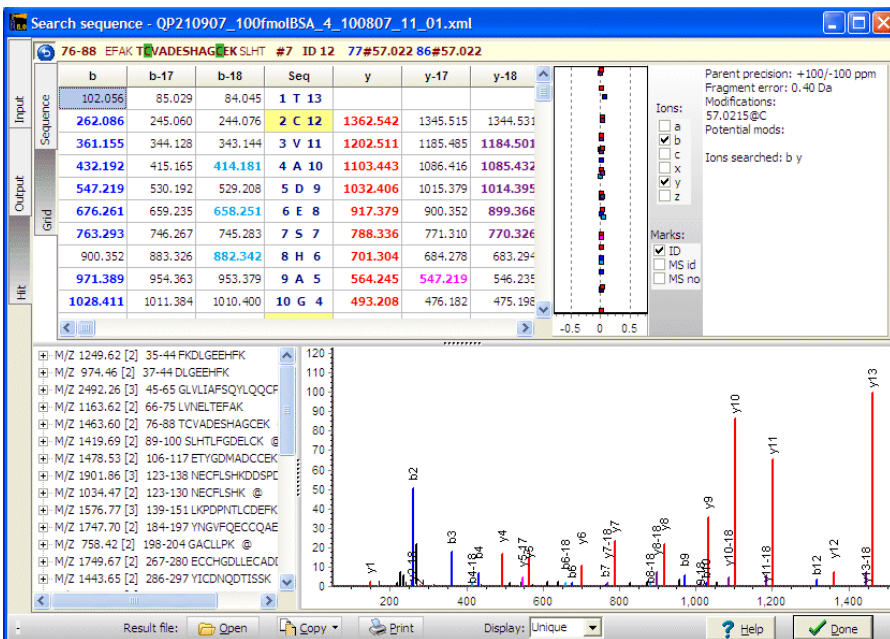
The tabbed window at the top changes to a grid, displaying the theoretical fragments that can be generated. Fragments that are found in the mass spectrum are highlighted and colored. Each fragment type has its own color. The same colors are used to display the ms/ms spectrum, bottom right.

To the right of the grid, is displayed the precision with which the mass spectrum fits the theoretical mass values, again the same coloring is used for each fragment type.

Then a band is displayed with control options for fragment ions to display in the table, default are the ions used during the search. Below this control is another for the ms/ms labels: **ID** displays the ion type for the corresponding identified ion (e.g. y10, b7); **MS id** displays the mass values for identified ions; **MS no** displays the mass values for non-identified ions.

Finally is a table for the most important search parameters.

At the bottom left is a list of all peptides identified during the search. The actual peptides displayed depends on the setting of the 'Display' option in the

bottom toolbar: Display: Unique. Three options are available: **Unique** only the highest scoring peptide of each kind is displayed. Different modifications count as different peptides, even same modification in different locations.
Modified only modified peptides are displayed. **All** all peptides that fit the search criteria are shown.

In the peptide list each line shows the experimental mass of the parent ion,
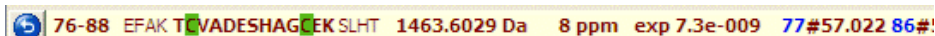
number of charges, location of the peptide, the peptide sequence and whether the peptide is modified (@).

If you click on the '+' sign, the list expands with additional information for the given peptide: ID is location in the peptide table (output tab). Theoretical mass, difference between theoretical and experimental mass, e-value, missed cleavages (mc) and lastly the modifications found in the peptide (listed as location@mass change).

Whenever a peptide is selected, the top line of the display is updated with peptide information:

The information line shows: peptide position (from-to), sequence with four residues before and after (modified residues are highlighted), experimental mass, deviation from theoretical mass, e-value and modified residues in the format: location@mass of modification.

If you click on the left-most blue button, the display changes to show the location of the hit in the result table (#7) and the location of the spectrum in the input file list (ID 12).

The ms/ms graph at the bottom right shows the mass data of the peptide as listed in the input file. Identified peptide fragments are colored along the same scheme as the fragment table. The label of each peak is controlled in the gray band above. The graph can be zoomed by click-and-drag right and down, you un-zoom by click-and-drag left and up. You can scroll the display by right-click and drag left/right.

Running the BSA example   8.14

The BSA example included with GPMAW is a short calibration run performed on an Orbitrap machine. It is present in the \gpmaw\bin\xtandem\ directory. Two files are included: The data file "BSA_100fmol.pkl" and an example of a search "BSA_100fmol_120107_1729.xml".

The ms/ms search has a lot of parameters to set, but once you have 'trimmed' the search parameters to your instruments/experiments, there will usually only be the input file, perhaps the output file, and the modifications to specify for each search.

Before you begin

Before you call the ms/ms search function, you load bovine serum albumin on the desktop – the accession number is P02769 if you need to download it from the Internet).

Download by typing number in toolbar window and click the 'Web' button (needs Internet access).

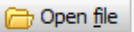Make sure that the BSA sequence is the topmost window before proceeding.

Setting up the search

Begin by selecting 'MS/MS search' from the 'Search' menu, press the F5 key

or click the magnifier [image] in the toolbar, this opens the MS/MS search window with the 'Input' tab selected.

Start by selecting the 'Open file' button. In the 'Open MS file ' dialog you navigate to \gpmaw\bin\xtandem\ and select the "BAS_100fmol.pkl" file.

Data input file name (.mgf, .dta, .pkl)

📁 Open file | C:\Dephi2007\Projects\gpmaw\XTandem\BSA_100fmol.pkl

As you already have the BSA sequence opened as the topmost window on the desktop it will be selected by the ms/ms search:

Search sequence: Serum albumin precursor (Allergen Bos d 6) (BSA). - Bos taurus (Bovine).

If you will rather search a file, you need to have the sequence(s) saved in FastA format. This is also the way to go if you need to search multiple sequences.

You now need to select the enzyme used to cleave the protein. Select from the drop-down box labeled 'Enzyme':

Enzyme: Trypsin [KR]|{P}   ▼

If the enzyme of your choice is not available on the list, select the last option: '- user defined –'. This will open a dialog box where you can specify your enzyme (for nomenclature see section 8.11 above).

Then enter the output (result) file name. By default gpmaw selects a name consisting of the input file name with date and time added.

Output file name: BSA_100fmol_120807_1700

The field can be freely edited. Note that if you perform multiple searches on the same input file without closing the ms/ms window in between, you have to change the name manually, as gpmaw will overwrite without asking permission.

You now have to select the parameters to use for the search. Please see section 8.11 for details on the various settings.

The most important settings for this search are +/- 10 ppm as the parent ion error and +/- 0.8 Da as the fragment error.
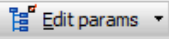
Click on the 'Edit params' button to set the parameters, you should end with settings similar to this:

Parameters:

Parent, mono. error: -10.0/+10.0 ppm
Fragment, mono. error: 0.8 Daltons
Max. valid expect value: 0.010
Minimum ion count: 4
Scoring ions: Y A B

Spectrum
Minimum number peaks: 13
Maximum number peaks: 51
Maximum parent charge: 4
Minimum parent m+h: 300
Minimum fragment m/z: 146

Protein
Max. missed cleavage: 2
Cleavage N-term mass change: 1.0107825
Cleavage C-term mass change: 17.0027350
N-term residue mod. mass: 0.00000
C-term residue mod. mass: 0.00000

File: Edit params ▾

Note that by clicking the down-arrow on the 'Edit params' button, you can save and retrieve up to five different settings to use on different instruments/ experiments.

Now select the variable modifications:

Variable modifications [1]

Oxygen [FMPY] 15.99
Methylatio [DE] 14.02
Phospho [STY] 79.97

Each modification can be selected / deselected by clicking with the mouse. If you need to load a different modification file or edit one of the modifications, click on the 'S' button [S] below the box. Oxygenation of Met is the usual choice, but a number of others like deamidation of Asn and Gln are also typically seen for most runs.
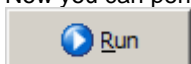
For the fixed modifications you select carbamidomethylation of cysteines (Cys has been alkylated with iodoacetamide).

Fixed modifications [1]

Carbamidomethyl [C] 57.02
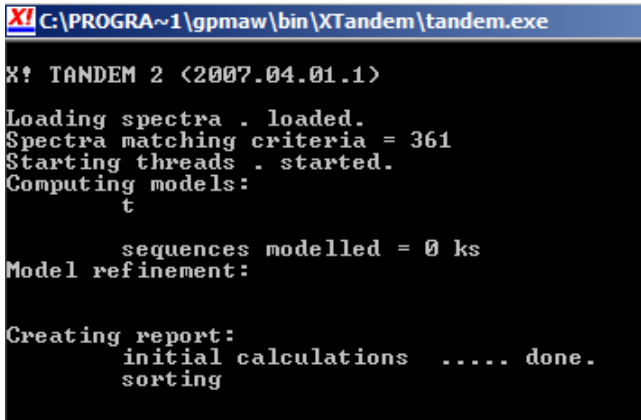Carboxymethyl [C] 58.01

You can select to the perform 'refinements' (second pass search) on your data, please see section 8.11 for details.

Now you can perform the search by clicking the Run button

▶ Run

.

The run

If parameters are correctly set and XTandem is correctly installed, a command line window will open, showing the progress of the search.

Without the refinement option set, this search should take less than two seconds.

Results

When the search is over, XTandem saves the results to disk and closes. GPMAW is informed, loads the result file and displays the contents on the 'Output' page.

As we only searched a single protein, there will be only one hit in the list of proteins found:

#	log (e)	hits	uniq	name
1	-172.6	39	39	P02769|Serum albumin precursor (A

Below the protein list is a list of all the peptides identified in the search which fall inside the parameters set (particularly below the e-value of 0.01).

#	spec id	top model hit	e-value	z	MH+ exp	MH+ theo	delta	from	to	pre	seq
1	50	P02769|Serum albumin	8.5e-005	3	1249.6121	1249.6240	-0.0120	35	44	IAHR	FKDLGEI
2	263	P02769|Serum albumin	6.3e-006	2	1249.6223	1249.6241	-0.0018	35	44	IAHR	FKDLGEI
3	42	P02769|Serum albumin	3.3e-003	2	974.4563	974.4607	-0.0044	37	44	HRFK	DLGEEH

For details on the peptide parameters, please see section 8.12.

Now we want to see the details of the search, so we double-click on the Serum albumin precursor name in the protein list.

Hit (protein results)

This opens the protein details tab (called 'Hit') which is a quite crowded display:

The top left display shows the protein coverage (i.e. the parts of the protein that has been identified by peptides in the ms/ms run):

```
MKWVTFISLL  LLFSSAYSRG  VFRRDTHKSE  II
DEHVKLVNEL  TEFAKTCVAD  ESHAGCEKSL  H
ERNECFLSHK  DDSPDLPKLK  PDPNTLCDEF  KI
ANKYNGVFQE  CCQAEDKGAC  LLPKIETMRE  KI
```

Below on the left is a table of all unique peptides identified:

⊞ M/Z 1249.62 [2] 35-44 FKDLGEEHFK
⊞ M/Z 974.46 [2] 37-44 DLGEEHFK
⊞ M/Z 2492.26 [3] 45-65 GLVLIAFSQYLQQCF
⊞ M/Z 1463.60 [2] 76-88 TCVADESHAGCEK

Click on one of the peptides will switch the coverage display to show the fragment ions of the identified peptides.

| a	b	b-17	b-18	Seq	y	y-17
101.071	129.066	112.040	111.056	1 Q 8		
215.114	243.109	226.083	225.099	2 N 7	940.383	923.35
375.145	403.140	386.113	385.129	3 C 6	826.341	809.31
490.172	518.167	501.140	500.156	4 D 5	666.310	649.28

This table shows all the possible fragments, with the ones identified in the spectrum shown highlighted in a specific color for each fragment type. The identified residues (sequence) are displayed in the 'Seq' column. If a residue is identified as modified, is shown with a yellow background.

The top line in the window shows the identified peptide sequence with modified residues highlighted and the most important parameters:

↺ 413-420 NLIK QNCDQFEK LGEY 1051.4171 Da -1 ppm exp 3.9e-003 415#57.022

By clicking on the left-hand blue button, the display changes to show the peptide ID parameters (location in the table on the 'Output' tab and location of the ms/ms spectrum in the input file).

The bottom right window shows the corresponding ms/ms spectrum
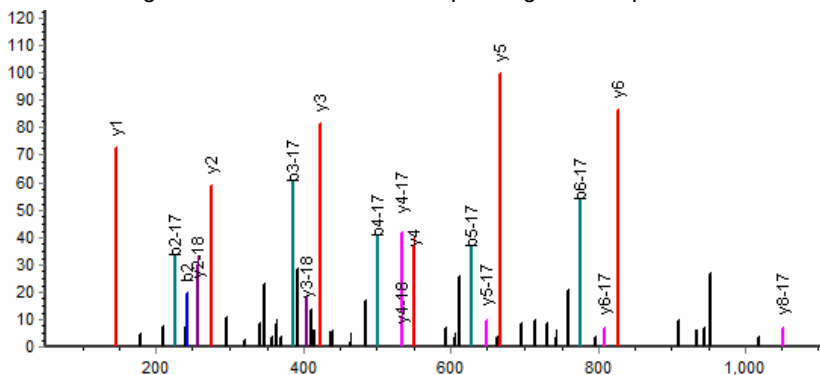
The colors of the fragments correspond to the colors in the table above. The label of the peaks can be set in the control bar to the right of the table, which also controls the table columns to display. The top central graph shows the deviation of the ms/ms fragments from the theoretical mass values, and the right-hand list shows the search parameters.

The results are saved on the computer as an xml file in the same directory as the input mass file. This file can at any point be recalled (click the bottom open button Result file: Open) and viewed again. You can also go to the Global Proteome Machine (www.thegpm.org) and load the file for an alternative view of your data.

On the disk is a file 'BSA_100fmol_120107_1729.xml' which contains a search performed by Lighthouse data. You can compare your results with this file.

References:

TANDEM: matching proteins with mass spectra, Robertson Craig and Ronald C. Beavis, *Bioinformatics*, 2004, 20, 1466-7.

A Method for Reducing the Time Required to Match Protein Sequences with Tandem Mass Spectra, Robertson Craig and Ronald C. Beavis; Rapid *Commun. Mass Spectrom.*, 2003, 17: 2310-2316.