Published in final edited form as:

Curr Protoc Bioinformatics. 2012 March ; 0 13: Unit13.17. doi:10.1002/0471250953.bi1317s37.

Identifying proteomic LC-MS/MS data sets with Bumbershoot and IDPicker

Jerry D. Holman, Ze-Qiang Ma, and David L. Tabb*

Department of Biomedical Informatics Vanderbilt University Medical Center Nashville, TN, 37232-8575

Abstract

The identification of peptides and proteins by LC-MS/MS requires the use of bioinformatics. Tools developed in the Tabb Laboratory contribute significant flexibility and discrimination to this process. The Bumbershoot tools (MyriMatch, DirecTag, TagRecon, and Pepitome) enable the identification of peptides represented by MS/MS scans. All of these tools can work directly from instrument capture files of multiple vendors, such as Thermo RAW format, or from standard XML-based formats, such as mzML or mzXML. Peptide identifications are written to mzIdentML or pepXML format. Protein assembly is handled by the IDPicker algorithm. Raw identifications are filtered to a confident set by use of the target-decoy strategy. IDPicker arranges large sets of input files into a hierarchy for reporting, and the software applies a parsimony algorithm to report the smallest possible number of proteins to explain the observed peptides. This protocol details the use of these tools for new users.

Keywords

Shotgun proteomics; protein database search; sequence tagging; protein assembly; proteome informatics; peptide-spectrum matches

Identifying proteins from LC-MS/MS data sets

The protocols of this unit detail the process by which peptide and protein identifications may be assessed from LC-MS/MS collections (see Figure 1). Basic Protocol 1 details the use of the MyriMatch database search engine (Tabb et al. 2007) by means of the BumberDash

* (615)936-0380 voice (615)343-8372 fax david.l.tabb@vanderbilt.edu.

Internet Resources with Annotations

1. Matrix Science Data File Format page

http://www.matrixscience.com/help/data_file_help.html

Many file formats have been created to support peptide identification, and this website enumerates and diagrams some of the most common types.

2. Tabb Laboratory web page

<http://proteowizard.sourceforge.net/>

The Bumbershoot and IDPicker tools described in this protocol may be acquired from the Tabb Laboratory Team City server, which is accessible from the "Software" page at this website.

3. NIST Spectral Libraries

<http://peptide.nist.gov/>

The National Institute of Standards and Technologies has amassed spectral libraries for a large variety of samples and instruments; these collections are available from their website.

graphical user interface. Alternate Protocol 1 substitutes the TagRecon sequence tagging engine for peptide identification, which allows for more flexible recognition of post-translational modifications (Dasari et al. 2010; Dasari et al. 2011). Once raw peptide identifications have been generated, filtering and protein assembly can take place. Basic Protocol 2 supplies the steps necessary to accomplish these tasks using IDPicker (Zhang et al. 2007; Ma et al. 2009).

Strategic Planning

Initial peptide identification is the most time-intensive step of this process (Basic Protocol 1 or Alternate Protocol 1). This process is most quickly completed using a cluster of computers; this protocol, however, details the use of a Microsoft Windows PC for peptide identification. Users are encouraged to use recent desktop computers featuring multiple CPU cores for rapid results.

Researchers may be accustomed to translating raw data to MGF, PKL, or DTA format prior to search under Mascot, Spectrum Mill or Sequest protocols, respectively (see Internet Resource 1 below). This step is unnecessary in the context of MyriMatch or TagRecon because this software can read raw data from Thermo, Agilent, and Bruker instruments via ProteoWizard (Kessner et al. 2008), so long as identification is conducted in Microsoft Windows. Although Waters and AB Sciex files can also be read in this way, transforming profile data to peaklists should be conducted first by the use of vendor-supplied software.

If IDPicker is to be used for filtering identifications (Basic Protocol 2), researchers should ensure that the FASTA protein sequence database supplied for peptide identification contains decoy sequences along with the normal protein sequences. These decoy sequences are necessary for assessing the confidence of accepted peptide identifications (Elias and Gygi 2010); most typically, the reversed version of each protein sequence is appended to the sequence database. Accession numbers for decoy sequences must contain a prefix that signifies their status as decoy. For example, FASTA databases from the Tabb Laboratory denote reversed sequences by accession numbers that start with a “rev_” string, such as “rev_sp|P0CD91.”

Basic Protocol 1: Database Search

The initial process of matching database peptides to experimental MS/MS scans is conducted by the MyriMatch algorithm. BumberDash is a graphical user interface that streamlines the use of this tool and the other “Bumbershoot” identification algorithms. The steps enumerated below illustrate the use of this interface and introduces some of the configuration decisions users make in the process of identification.

Necessary Resources

- BumberDash 1.3 (downloadable from Internet Resource 2 below)
- Spectrum source file in a compatible format (mzML, mzXML, RAW, WIFF, YEP, or MGF)
- FASTA protein sequence database

Adding new MyriMatch job to BumberDash

1. Start at the BumberDash main screen (see Figure 2).
2. Go to File→ New job or click on the last row of the job queue to load the Add Job dialog.
3. If it is not selected already, select “Database Search” from the drop down list at the top of the dialog (see Figure 3).

The configuration panel at the bottom of the panel should now contain a box labeled “MyriMatch Config.”

4. In the “Name” box enter a title for the new job.

This field can be left blank if desired and will automatically fill with a default value after the “Input Files” box is filled.
5. (Optional) By default BumberDash will write output directly to the output folder. If a new folder should be created within the output folder to hold results, select the checkbox next to the “Name” box.
6. Click the “Browse” button next to the box labeled “Input Files”. From the pop-up dialog navigate to the folder in which the spectrum source files are located.
7. Select the spectrum source file that is to be examined or hold down ctrl to select multiple files.
8. Once all files have been selected click “Open” to populate the “Input Files” box with their names and locations.

If the “Name” or “Output Files” boxes are empty, BumberDash will automatically fill them with the folder name from which the files were selected.
9. If a different output location is desired, select the “Browse” button next to the “Output Directory” box and navigate to the desired location.
10. Click the “Browse” button next to the FASTA Database box and select the protein database file that corresponds to the spectrum source files.
11. If a configuration file has previously been created for MyriMatch, select it with the “Browse” button. Otherwise click “New” next to the “MyriMatch Configuration” box.
12. Make all desired changes to the job's configuration using the Myrimatch Configuration Editor (see Figure 4).

The MyriMatch configuration editor has five options available by default; Instrument, Precursor Mass, Digestion Enzyme, Specificity, and Modifications. BumberDash comes preloaded with default values for four general types of instruments. When a user selects an instrument from the drop down list, several label colors will change to indicate which options would be changed by loading the instrument template. To actually apply

the changes click the “Load” button next to the drop down box. Modifications can be entered within the interface in the lower panel, and a few common modifications are already described in the list. Clicking on any of these list items will populate the corresponding boxes, which can then be added to the “Applied Modification” list by clicking the “>” button. Static modifications change the mass of every corresponding amino acid. With dynamic modifications, each modifiable site is evaluated both with and without the modifying mass. To view more options click the “Use Advanced Mode” box in the lower left corner of the configuration form. This will activate the grayed out options and add a new tab of advanced options.

13. Save the configuration to return to the Add Job Dialog.
 - a. If the configuration settings need to be saved for a future job click “Save As New”. Once saved the configuration form will close, and the new .cfg file will automatically be entered into the proper box.
 - b. If the configuration only needs to be used for the job at hand click “Use Once.” BumberDash will prompt for a configuration name; if this is left blank, the program will supply a default name.
14. (Optional) If the number of CPU cores needs to be limited to enable other simultaneous uses of the computer, set the maximum number to be used in the “CPUs” box in the lower right corner.
15. Once all boxes are filled out, click the “Add” button to add the job to the queue.

The job will be run as soon as it has reached the next available spot in the queue (see Figure 5). As it is running BumberDash will provide status updates on the progress bar, the recent events log box, and in the full log (see Figure 6). Jobs are run at a below-normal priority, so they should not interfere with computer performance, and the BumberDash form minimizes to a tray icon by default to conserve task bar space. Double-clicking the BumberDash icon anytime it is minimized will restore the form. A job can be double-clicked to open the output folder in Windows Explorer, and IDPicker, if installed, can be launched from the File menu to start the next step in spectra analysis.

Alternate Protocol 1: Sequence Tagging

While database search is the most conventional strategy for peptide identification, researchers interested in post-translational modifications benefit from sequence tag-based identification instead. The TagRecon algorithm supporting this strategy is also supported within BumberDash. It begins by inferring partial sequences from MS/MS scans by the DirecTag algorithm and then reconciling these partial sequences to peptide sequences from the database.

Necessary Resources

- BumberDash 1.3 (downloadable from Internet Resource 2 below)
- Spectrum source file in a compatible format (mzML, mzXML, RAW, WIFF, YEP, or MGF)
- FASTA protein database

Adding new DirecTag/TagRecon job to BumberDash

1. Start at the main screen.
2. Open Add Job dialog from the menu or by clicking on the last row in the queue.
3. Select the Tag Sequencing option from the radio buttons at the top of the dialog.

The configuration panel should now display boxes for both DirecTag and TagRecon.
4. Fill out the “Name”, “Input Files”, “Output Directory”, “FASTA Database” (required for identification after tags have been created), and “CPUs” boxes as in Basic Protocol 1.
5. If a DirecTag configuration file has been previously created, select it using the “Browse” button next to the DirecTag configuration box; otherwise click the button labeled “New.”
6. Make all desired changes to the job's configuration using the DirecTag Configuration Editor.

Only Instrument and modifications are available in basic mode. To show all options, check the Advanced Mode box.

7. Once all changes are made, save the configuration as a file or use it as a temporary configuration.

The new settings should appear in the DirecTag configuration box.

8. As before, if a TagRecon configuration file has been previously created, select it using the “Browse” button next to the TagRecon configuration box, otherwise click the button labeled “New.”
9. Make all desired changes to the job's configuration using the TagRecon Configuration Editor.

The TagRecon configuration editor looks much like the MyriMatch editor (Figure 4); however, there are a few key differences that should be noted. The most important of these differences are visible within the “Modifications” panel. TagRecon has the ability to find mass shifts in a spectrum and try to explain them using a predefined method. By default the option is blank, however the different strategies can be utilized by selecting an option in the box labeled “Explain Unknown Mass Shifts As.” The “mutations” option attempts to substitute one amino acid in for another that would be a better fit, and is detailed in the first TagRecon

publication (Dasari et al. 2010). Selecting “preferredptms” enables the user to select a set of preferred modifications to be sought, but allowing for much greater speed than in standard database search (Dasari et al. 2011). The final option, “blindptms”, is also detailed in this publication. In brief, this option allows TagRecon to add an arbitrary mass shift to any single amino acid in a peptide, allowing for unexpected modification discovery. Due to the increase in search time, blind PTM search is recommended for use with FASTA databases that contain only the sequences of proteins known to be in a sample (plus decoys).

10. After all changes are complete, save the configuration as a new file or click Use Once to create a temporary configuration.
11. Once all boxes are filled out in the Add Job form, click Run to add the job to the queue.

Once the queue reaches a Tag Sequencing job, BumberDash will first infer partial sequences with DirecTag. After that is done the tag files will automatically be reconciled to full peptide sequences via TagRecon. As with MyriMatch jobs, these jobs will be run at below average process settings, so they should not create undue system lag.

Basic Protocol 2: Running IDPicker

The raw identifications generated by MyriMatch and TagRecon provide the information needed to generate confident lists of peptides. The IDPicker algorithm filters identifications from pepXML or mzIdentML files by an aggregate FDR strategy. It also organizes multiple LC-MS/MS data sets into a user-specified experimental hierarchy and produces parsimonious protein sets to explain the observed peptides. These steps detail the IDPicker graphical user interface.

Necessary Resources

- IDPicker 2.6 (downloadable from Internet Resource 2 below)
- Database search result in pepXML format
- FASTA protein sequence database
- (Optional) Spectrum source file in a compatible format (mzML, mzXML, RAW, WIFF, YEP, or MGF)

Setup Default Configurations

1. Start at the IDPicker main screen.
2. Click Tools→Options in the menu bar.
3. In the “Options” dialog as shown in Figure 7, change “Decoy prefix” to the prefix string of decoy sequences in the protein database.

In this example, all decoy protein sequences start with a prefix “rev_”. IDPicker requires a target-decoy search to enable peptide validation by

False Discovery Rate (FDR). Simple threshold-based filtering is not supported.

4. Select the “Database” tab and click “Add Path” to add the directory that includes the protein database in FASTA format.
5. (Optional) Select “Source” tab and click “Add Path” to add spectral source files.

IDPicker provides a built-in spectrum viewer for manual validation of peptide-spectrum-matches. This setting directs IDPicker to find source files in the specified directories.

Select Input Files for a New Project

6. Click File→New Report in the menu bar to start a new project.

IDPicker also provides a “clone” function to start a project from a previously run report. To clone a report, right click on the previously run report on “My Reports” page and select “Clone.” IDPicker will copy all the settings of this report into a new report form.

7. Using the “Load and Qonvert pepXML Search Results” dialog, configure the project as shown in Figure 8.

IDPicker requires all input files in a report to use the same FASTA protein database.

8. Click “List Files” button and select all input files.

(Optional) Setup Advanced Options

9. Click the “Advanced” button to see the “Advanced Options” dialog as Shown in Figure 9. 10.

11. Click “Score names and weights” to configure search score combination options.

By default, IDPicker will work with pepXML files from MyriMatch, X! Tandem, Sequest and Mascot. However, it is able to read arbitrary score names from input files and assign them weights to produce the “total score” which is used to sort the results from each spectrum. IDPicker combines multiple database search scores as a weighted summation. Weights may either be user-defined (static) or automatically determined using a Monte Carlo simulation method (dynamic). In the dynamic mode, IDPicker tests random score weights to determine which maximizes the total number of confident identifications for the specified FDR. This feature is disabled by default; enable it by checking the “Apply score optimization” checkbox. The permutation count only has meaning when score optimization is enabled.

12. Click “Modifications” to configure distinct/indistinct modifications.

By default, IDPicker treats peptides with a modification as distinct from the unmodified peptide in protein assembly. To override this setting, enter amino acids and modification masses, then select “Indistinct.”

Configure Data Grouping and Filters

13. Click the “Next” button to access the “Configure Data Groupings and Filters” dialog.
14. Setup group hierarchy as shown in Figure 10.

IDPicker supports protein assembly and analysis with arbitrarily complex hierarchies. Right click on each group to add a new group or remove/rename this group.
15. Configure peptide and protein level filters.

Setting “minimum distinct peptides per protein” filter to 2 efficiently removes “one hit wonders” that are only identified by one peptide. For data sets searched against multi-species databases, setting “minimum additional peptides per protein group” filter to 2 reduces the falsely identified orthologous proteins.
16. (Optional) check the “Automatically export TSV” checkbox. This enables IDPicker to generate reports in text format.
17. Click “Run Report” to start generating the report.

View IDPicker Report

18. A report is opened as it is created. Reports can be reopened at a later time from the “My Reports” page. The report can also be viewed using a Web browser by double clicking the “index.html” file in the output directory. Figure 11 shows the summary page of the report.

(Optional) Export Report

19. Right click on a report on the “My Reports” page and click “Export.” The export dialog allows for configuration of which files get included in the ZIP, TSV or XML format.

Guidelines for Understanding Results

Determining whether or not peptide identification has completed is straightforward; Basic Protocol 1 and Alternate Protocol 1 each generate a pepXML file that is written after this step has completed for each LC-MS/MS file. Determining whether or not these identifications are worthwhile or not, however, is more complex; although peptide identification yields the best matches it can for a given database and configuration, flawed identifications will be produced if the researcher has specified the wrong type of instrument or supplied a FASTA for the wrong species.

Researchers can determine whether peptide identification has worked by continuing to Basic Protocol 2, which attempts to build protein-level information from peptide identifications. As shown in Figure 11, IDPicker reports a series of empirical FDRs for the assembly. Typically, researchers will choose settings that keep the protein-level FDR below 10%, though this value may be allowed to go higher, especially if proteins evidenced by a single peptide are included in the report.

Next, computing the percentage of identified spectra is valuable in evaluating the experiment. The number of spectra identified throughout the set of files can be found to the right of the '/' symbol in the "Confident IDs" column (see Figure 11). Dividing this number by the total number of tandem mass spectra in all of the raw files will produce an identification rate. As many as half of the tandem mass spectra collected for concentrated samples from small genomes may be identified if they have been analyzed by instruments that can determine the charge state of peptides prior to MS/MS collection. When samples with large dynamic range and complex genomes are identified on low-resolution ion traps, however, identification rates tend to fall. An identification rate below 5% may suggest either low data quality or improperly configured peptide identification.

Commentary

Background Information

Tandem mass spectra (MS/MS) record the fragment ions generated as ions of a particular peptide dissociate through collision-induced dissociation (CID) (Wysocki et al. 2000), electron transfer dissociation (ETD) (Swaney et al. 2007), or another technique. Peptides can produce a pair of sequence-specific fragments for each peptide bond in their structure, generating b-y fragment ion pairs by breaking at peptide bonds in CID or generating c-z fragment ion pairs by breaking N-terminal to alpha carbons in ETD (Roepstorff and Fohlman 1984). Other fragment ions may be generated through the neutral loss of small molecules such as water, ammonia, or carbon monoxide (Paizs and Suhai 2005). The mass analyzer used to collect the tandem mass spectrum determines the accuracy by which the fragment ion mass-to-charge ratios are known, just as the mass analyzer used to catalog peptides in mass spectrometry determines how closely the observed peptide mass-to-charge ratio conforms to the true peptide mass-to-charge.

Interpreting a tandem mass spectrum manually to determine its sequence is both challenging and error-prone (Hunt et al. 1986). Several algorithms have been designed to automate the process of matching a sequence to observed tandem mass spectra. The database search algorithm was introduced by the publication of Sequest in 1994 (Eng et al. 1994). Sequest (UNIT 13.3) generates peptide sequences from a database of proteins, determines which peptides have a mass close to that observed for a tandem mass spectrum, and compares that set to the MS/MS by predicting their fragments and scoring them against the observed fragments. The sequence tagging algorithm was also introduced in 1994 (M Mann and Wilm 1994), but fully automating this technique was delayed by several years (Tabb et al. 2003; Tanner et al. 2005). Sequence tags are partial sequences inferred directly from MS/MS scans that can be used to determine which peptides from a sequence database should be compared to a given MS/MS. Spectral library search algorithms, as embodied by SpectraST (Lam et al.

2007), match newly acquired spectra to existing identified spectra. At present, database search is by far the most popular way to identify LC-MS/MS collections, while use of the other techniques is growing in response to improved software availability.

All of these techniques have been implemented as tools in the “Bumbershoot” suite from the Tabb Laboratory. MyriMatch was published in 2006, introducing a statistical match scoring system called “MVH,” based on the multivariate hypergeometric distribution (Tabb et al. 2007). DirecTag, published in 2008, introduced a high-discrimination strategy for inferring partial sequences from tandem mass spectra (Tabb et al. 2008). TagRecon, which leverages these sequence tags to identify the full-length sequences for spectra, has been published both for hunting mutations (Dasari et al. 2010) and for recognizing modifications of unknown mass and specificity (Dasari et al. 2011). The Pepitome algorithm matches spectra from the NIST spectral libraries (see Internet Resource 3) to experimental sets to identify peptides (submitted, Dasari et al.); an interface to Pepitome will soon be incorporated in Bumberdash. All three strategies are intended to find the best peptide sequence interpretation for each MS/MS scan.

Because many spectra cannot be correctly identified by these techniques, additional refinement is necessary. The IDPicker engine employs the target-decoy approach (Elias and Gygi 2010) to select a set of identifications that achieve a user-specified false discovery rate (FDR) (Ma et al. 2009). The software enables users to combine the identifications among hundreds or thousands of LC-MS/MS experiments into a single analysis by applying an experimental hierarchy to organize the results. Finally, the software chooses a minimal set of proteins to explain the observed peptides by constructing a bipartite graph relating potential proteins to observed peptides and picking a set cover (Zhang et al. 2007). IDPicker is able to process results from all three identification algorithms interchangeably.

The chief advantages of the Bumbershoot / IDPicker approach include:

- ProteoWizard provides wide LC-MS/MS data format compatibility.
- The MVH scoring system yields high discrimination, especially when paired with XCorr.
- Sequence tagging allows for aggressive post-translational modification discovery.
- IDPicker reports can sensibly organize identifications from hundreds of LC-MS/MS experiments.
- IDPicker produces parsimonious protein lists and shows which peptides are shared.

Critical Parameters and Troubleshooting

Identification quality depends heavily upon the selection of several settings at both peptide and protein levels. The first of these is precursor mass-to-charge tolerance. If set too broadly, tandem mass spectra are compared with many extra candidate peptides that have little chance of being correct matches. If set too narrowly (assuming too high a mass accuracy for precursors), the correct sequence may be excluded from the set of comparisons, making a correct match impossible. Bumberdash recommends several standard mass settings for major mass analyzer types to forestall this type of error.

Enzyme specificity is also critical for identification quality. A given peptide may conform to trypsin specificity on both termini, on one terminus, or on neither terminus. Some authors have argued that because identified peptides are predominantly “fully tryptic,” one should require all candidate peptides extend to trypsin cutting sites on both termini (Olsen et al. 2004). This is also convenient because only about 1% of all peptides that can be generated from a protein sequence database are fully tryptic, leading to a fast search. Bioinformaticists have argued that conducting a “semi-tryptic” search, one that evaluates peptides even when only one end is a tryptic cutting site, makes it far easier to recognize correctly identified peptides in downstream analysis (Keller et al. 2002). These searches, however, typically take roughly an order of magnitude more time to complete.

Controlling FDRs to attain manageable error rates is a continuing challenge for the field. If a researcher limits the Peptide-Spectrum Match (PSM) FDR to 5%, he or she could draw the erroneous conclusion that any identified peptide has a one-in-twenty probability of error. In fact, the error rate for the highest-scoring peptides is considerably lower than the lowest-scoring peptides within the collection. Researchers need to keep in mind that peptides bearing multiple post-translational modifications should be viewed with a very skeptical eye; sets of modifications may induce a false fit between an unrelated peptide sequence and the observed tandem mass spectrum. Popularity can be a useful trait; peptides for which many spectra have been identified are far less likely to be erroneous than those corresponding to a single spectrum (though it is certainly possible for many spectra to be matched to the same, incorrect sequence). Likewise, researchers will do well to place more trust in proteins for which many peptides appear than the proteins for which only two peptides appear (or worse, proteins supported by only a single peptide). Manually examining a peptide-spectrum match is appropriate to protect against errors in critical identifications (Tabb et al. 2006).

The most likely problem sources associated with the use of this software include decoy sequences, .NET libraries, and hardware challenges:

- If a user has not added decoy sequences to a database, peptide identification will complete, but protein assembly will fail. Decoy sequences must be flagged by a prefix (such as “rev_”), and this prefix must be correctly specified for IDPicker.
- Bumberdash and IDPicker algorithms both employ the .NET libraries. Users should install the 3.5SP1 version of these libraries or later before installing the proteomic software to ensure that the needed tools are in place.
- Peptide identification is taxing on hardware, particularly when low mass accuracy precursors are paired with semi-tryptic searches and large sequence databases. Users will benefit from recent microprocessors with multiple cores and high clock speeds. When raw data come from high scan rate instruments such as the “Velos” series from Thermo, the memory load for conducting peptide identification may be substantial. Computing servers with at least 4GB of RAM are appropriate for handling these large data sets.

Literature Cited

- Dasari S, Chambers MC, Codreanu SG, Liebler DC, Collins BC, Pennington SR, Gallagher WM, Tabb DL. Sequence Tagging Reveals Unexpected Modifications in Toxicoproteomics. *Chemical Research in Toxicology*. 2011
- Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham A-JL, Tabb DL. TagRecon: high-throughput mutation identification through sequence tagging. *Journal of Proteome Research*. 2010; 9(4):1716–1726. [PubMed: 20131910]
- Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in Molecular Biology* (Clifton, N.J.). 2010; 604:55–71.
- Eng JK, McCormack AL, Yates JR III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*. 1994; 5(11):976–989. [PubMed: 24226387]
- Hunt DF, Yates JR 3rd, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*. 1986; 83(17):6233–6237. [PubMed: 3462691]
- Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*. 2002; 74(20):5383–5392. [PubMed: 12403597]
- Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* (Oxford, England). 2008; 24(21):2534–2536.
- Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*. 2007; 7(5):655–667. [PubMed: 17295354]
- Ma Z-Q, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, Tabb DL. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *Journal of Proteome Research*. 2009; 8(8):3872–3881. [PubMed: 19522537]
- Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*. 1994; 66(24):4390–4399. [PubMed: 7847635]
- Olsen JV, Ong S-E, Mann Matthias. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics: MCP*. 2004; 3(6):608–614. [PubMed: 15034119]
- Paizs B, Suhai S. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews*. 2005; 24(4):508–548. [PubMed: 15389847]
- Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry*. 1984; 11(11):601. [PubMed: 6525415]
- Swaney DL, McAlister GC, Wirtala M, Schwartz JC, Syka JEP, Coon JJ. Supplemental activation method for high-efficiency electron-transfer dissociation of doubly protonated peptide precursors. *Analytical Chemistry*. 2007; 79(2):477–485. [PubMed: 17222010]
- Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *Journal of Proteome Research*. 2007; 6(2):654–661. [PubMed: 17269722]
- Tabb DL, Friedman DB, Ham A-JL. Verification of automated peptide identifications from proteomic tandem mass spectra. *Nature Protocols*. 2006; 1(5):2213–2222. [PubMed: 17406459]
- Tabb DL, Ma Z-Q, Martin DB, Ham A-JL, Chambers MC. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *Journal of Proteome Research*. 2008; 7(9):3838–3846. [PubMed: 18630943]
- Tabb DL, Saraf A, Yates John R. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Analytical Chemistry*. 2003; 75(23):6415–6421. [PubMed: 14640709]
- Tanner S, Shu H, Frank A, Wang L-C, Zandi E, Mumby M, Pevzner PA, Bafna V. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Analytical Chemistry*. 2005; 77(14):4626–4639. [PubMed: 16013882]

- Wysocki VH, Tsapralis G, Smith LL, Brei LA. Mobile and localized protons: a framework for understanding peptide dissociation. *Journal of Mass Spectrometry: JMS*. 2000; 35(12):1399–1406. [PubMed: 11180630]
- Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *Journal of Proteome Research*. 2007; 6(9):3549–3557. [PubMed: 17676885]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

1.

This high-level view shows how these protein identification tools connect. Bumpershoot tools (in orange) generate raw peptide identifications. These are fed to IDPicker for filtering and parsimonious protein enumeration.

2. BumberDash is a common graphical interface to launch all of the Bumbershoot tools.

3. Adding a job in BumberDash requires the user to specify several information sources and a base configuration.

The screenshot shows the 'ConfigForm' application window with the following sections:

- General**
 - Instrument Specific**
 - Instrument: New (dropdown) [Load]
 - Precursor Mass: auto (dropdown)
 - Digestion**
 - Enzyme: ? Trypsin/P (dropdown)
 - Specificity: Fully-Specific (dropdown)
 - Max Missed Cleavages: ∞ (spinbox)
 - m/z Tolerance**
 - Avg Precursor tolerance: 1.5 (spinbox) m/z (dropdown)
 - Mono Precursor tolerance: 10 (spinbox) ppm (dropdown)
 - Fragment tolerance: ? 0.5 (spinbox) m/z (dropdown)
 - Modifications**
 - Mod List:**
 - Description
 - Carboxymethylated Cysteine
 - Oxidized Methionine
 - Phosphorylated Serine (highlighted)
 - N terminal deamidation of glutamine
 - Applied Mod List:**

Motif	Mass	Type
C	57	Static
 - Buttons: > and <
 - Redidue Motif: S (text) Mod Mass: 79.966 (text) Mod Type: ? Dynamic (dropdown)
 - Max Dynamic Mods: ? 2 (spinbox)
- Legend:**
 - Non-Default Value (green)
 - Program Default, Non-Template Value (blue)
 - Non-Default or Template Value (purple)
- Buttons:**
 - ☐ Use Advanced Mode
 - Use Once
 - Save As New
 - Cancel

4. BumberDash enables the interactive development of configuration files.

5. Multiple jobs may be scheduled to run consecutively on processing servers.

6. A runtime log enables explicit tracking of the current identification job.

7. IDPicker options define the default locations for information used in protein assembly.

8. Arranging peptide identifications in subdirectories is no hindrance to import in IDPicker.

9.

If multiple scores are produced for peptide-spectrum matches, IDPicker can combine them dynamically.

10.

Users can rearrange LC-MS/MS files to their own liking in the Data Groupings dialog.

[Summary](#)
[Group association table](#)
[Index by protein](#)
[Index by spectrum](#)
[Index by modification](#)
[Sequences per protein by group](#)
[Spectra per protein by group](#)
[Spectra per peptide by group](#)
[Data processing details](#)

Hide cluster table of contents

Id	Protein Groups	Unique Results	Spectra
1	2	558	9075
2	3	189	2111
3	2	170	12477
4	3	156	1059
5	38	151	1991
6	3	149	1777
7	2	99	760
8	5	81	347
9	1	71	258
10	1	70	864
11	2	69	1317
12	1	68	1071
13	2	67	733
14	7	66	278
15	2	58	423
16	1	55	439
17	1	52	207
18	1	47	472
19	2	44	246
20	1	42	240
21	1	42	221

Hide analysis parameters

Parameter	Value
Protein database	/hactar/fasta/20080729-IP1-HUMAN-347-Trypsin-reverse.fasta
Maximum FDR per result	0.05
Maximum ambiguous IDs per result	2
Minimum peptide length per result	5
Minimum distinct peptides per protein	2
Minimum additional peptides per protein group	1
Minimum spectra per protein	5
Indistinct modifications	M 15.994 C 57

Overall protein-level FDR: 0.018
 Overall result-level FDR (all peptides): 0.001
 Overall result-level FDR (unmodified peptides): 0.001
 Overall result-level FDR (modified peptides): 0.002

Hide groups summary

Group	Confident Ids	Peptides	Peptide Groups	Proteins	Protein Groups
/	40126	3107	248	224	169
/cancers	24280	2452	215	223	168
/controls	15846	2656	224	219	167
- /cancers/sample1	14182	1048	140	174	133

Source	Confident Ids	Peptides	Peptide Groups	Proteins	Protein Groups
kde_815_EB_071306e_1	5576	736	66	113	87
kde_815_EB_071306e_2	5367	332	70	101	83
kde_815_EB_071306e_3	3239	420	117	156	122
+ /cancers/sample2	4866	1831	177	204	158
+ /cancers/sample3	5232	1885	177	210	159
+ /controls/sample4	6000	1976	167	202	153
+ /controls/sample5	5049	1567	170	197	152
+ /controls/sample6	4797	1742	179	195	152

11.

IDPicker reports are Javascript-enhanced HTML. The summary page shows the overall identification efficiency across the experimental hierarchy.