research articles Journal of proteome research

DBParser: Web-Based Software for Shotgun Proteomic Data Analyses

Xiaoyu Yang,[†] Vijay Dondeti,[†] Rebecca Dezube,[‡] Dawn M. Maynard,[†] Lewis Y. Geer,[§] Jonathan Epstein,[‡] Xiongfong Chen,[‡] Sanford P. Markey,[†] and Jeffrey A. Kowalak*,[†]

National Institutes of Health, 10 Center Drive, Room 3D42, Bethesda, Maryland 20892-1262

Received April 21, 2004

We describe a web-based program called ‘DBParser’ for rapidly culling, merging, and comparing sequence search engine results from multiple LC−MS/MS peptide analyses. DBParser employs the principle of parsimony to consolidate redundant protein assignments and derive the most concise set of proteins consistent with all of the assigned peptide sequences observed in an experiment or series of experiments. The resulting reports summarize peptide and protein identifications from multidimensional experiments that may contain a single data set or combine data from a group of data sets, all related to a single analytical sample. Additionally, the results of multiple experiments, each of which may contain several data sets, can be compared in reports that identify features that are common or different. DBParser actively links to the primary mass spectral data and to public online databases such as NCBI, GO, and Swiss-Prot in order to structure contextually specific reports for biologists and biochemists.

Keywords: proteomics • data analysis • peptides • mass spectrometry • software • bioinformatics

Introduction

We present herein the architecture and major features of a web-based utility, DBParser, designed to rationally organize peptide data from tandem mass spectrometry experiments into reports meaningful to biological researchers engaged in proteomics. High-throughput “shotgun” proteomics aims to identify, characterize and quantify all of the expressed proteins simultaneously in a mixture.[1] This approach subjects protein mixtures to proteolytic digestion prior to liquid chromatographic separation and MS/MS analysis of the resulting peptides.[2] Several database search engines including Mascot,[3] Sequest,[4] and OMSSA[5] assign probable peptide sequences to MS/MS spectra and infer protein precursor identities. Because the shotgun technique is robust, sensitive and efficient, it is routinely applied for high-throughput protein identification and characterization.[2] Large data sets resulting from this approach emphasize the critical roles for data processing in reporting and validating protein identifications from proteomics experiments.[2,6,7] For any given LC−MS/MS analysis, database search engines routinely generate lists of peptide and protein sequence candidates, and the length of these lists is compounded for data sets resulting from multiple chromatographic separations. Subsequent sorting, collation, and comparison of these results pose significant challenges, especially when analyzing multiple files. Complexity arises because spectra matched to candidate peptides and proteins comprise comprehensive inclusive sets, i.e., the peptide data is matched to all possible protein data records[3,4] that contain the observed possible amino acid sequences. We chose to devise a utility to sort and reductively collate peptide and protein data by applying a parsimony principle (Ockham’s razor): peptides mapping to a simple rather than a complicated set of proteins are most likely to account for the observed spectra.

Materials and Methods

Experimental Data. Tandem MS data was derived from 1D or 2D LC−Ion Trap MS and 1D LC− Quadrupole Time-of-Flight MS experiments.[8] Three aliquots of a yeast protein extract (5, 10, and 20 μg) were used in this paper to demonstrate DBParser features and are designated Sample A, B, and C, respectively. Soluble proteins were reduced with DDT, alkylated with iodoacetamide, and digested with Endoproteinase Lys-C followed by trypsin. The resulting peptides were analyzed using 2D LC (Shimadzu LC−VP HPLC; Kyoto, Japan) directly coupled to an ESI−Ion Trap mass spectrometer (LCQ Classic; Thermo Electron, San Jose, CA).

Mascot Search Engine. Experimental data were submitted for MS/MS Ions Search to a Mascot cluster, maintained by the Helix Systems, CIT, NIH. The underlying flat files (e.g., F123456.dat) from which Mascot generates its reports were retrieved from the Mascot cluster via ftp. Mascot uses a probability-based scoring algorithm based on the Mowse algorithm[9] to assign peptide sequences to MS/MS spectra. Mascot compares peak lists containing mass and intensity pairs

[*] To whom correspondence should be addressed. Tel: (301) 496 4242. Fax: (301) 480 0198. Email: jeffrey.kowalak@nih.gov.

[†] Laboratory of Neurotoxicology, National Institute of Mental Health.

[‡] Unit on Biologic Computation, National Institute of Child Health and Human Development.

[§] National Center for Biotechnology Information, National Library of Medicine.

10.1021/pr049920x Not subject to U.S. copyright. Publ. 2004 Am. Chem.Soc.

Published on Web 09/18/2004

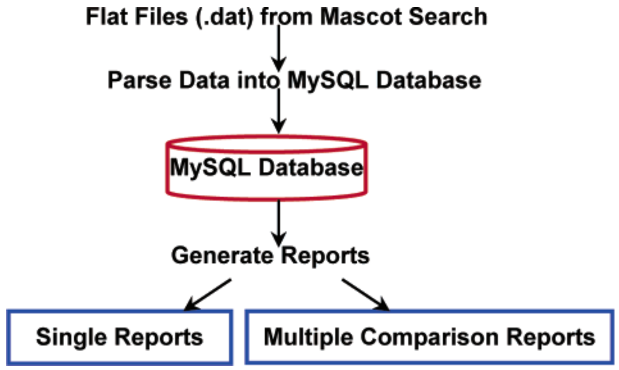
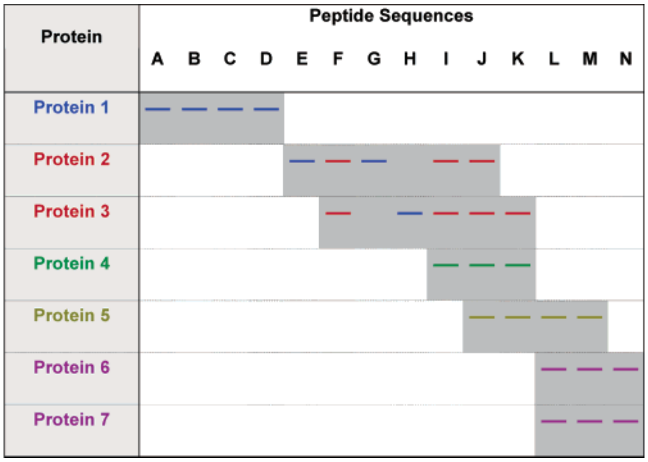

Flat Files (.dat) from Mascot Search
Parse Data into MySQL Database
MySQL Database
Generate Reports
Single Reports
Multiple Comparison Reports

Figure 1. DBParser data flowchart. DBParser has 5 steps: upload flat files, create databases, parse data, generate reports, and view reports. After each Mascot flat file is acquired by the DBParser server, the data from each flat file is parsed into a user-defined MySQL relational database. There are two main report generation procedures to summarize either a single experiment or to compare results from multiple experiments. Each of the main report generation procedures automatically produces multiple reports that emphasize specific aspects of the overall dataset.



Protein	Peptide Sequences													
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Protein 1	—	—	—	—										
Protein 2					—	—	—		—	—				
Protein 3						—		—	—	—	—			
Protein 4									—	—	—			
Protein 5										—	—	—	—	
Protein 6												—	—	—
Protein 7												—	—	—

Figure 2. Hypothetical protein parsimony analysis applied by DBParser. Letters A-to-N designate peptide sequences. Discrete peptides: A, B, C, D, E, G, and H; degenerate peptides: F, I, J, K, L, M, and N. Proteins are sorted by identified peptides as equivalent (6 and 7); subset (4 in 3); superset (3 vs 4); subsumable (5 vs.3, 5 vs 6 and 7); differentiable (2 vs 3); and distinct (1).

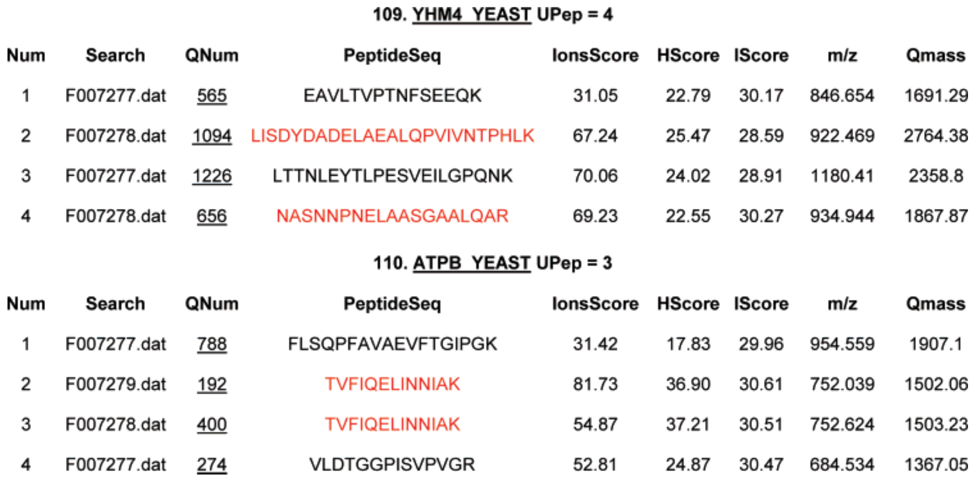
from one or more spectra with hypothetical spectra corresponding to in silico digestion of protein sequence database and assigns probable peptide sequences. On the basis of the protein sequence library database that a user selects, the Mascot peptide summary report presents a tabular summary of the most closely matching proteins.

Significance of Assignments. For each peptide, Mascot reports a probability-based Ions Score, which is defined as $-10*\log_{10}(P)$, where P is the absolute probability that the observed match between the experimental data and the database sequence is a random event. The relevance of the Ions Score is determined relative to a database-dependent Identity Score, which is a significance threshold with a $P < 0.05$ chance of a false positive.[3] This score is an absolute threshold that indicates a 5% or lower probability that the MS/MS spectrum has been randomly matched to a sequence in the database. DBParser accepts a Mascot peptide assignment only if the peptide Ions Score meets or exceeds its Identity Score. Subsequently, a protein assignment is accepted if it contains at least one significant assigned peptide. DBParser stores all of the data from Mascot files and reports significant or rejected information based upon user-defined criteria. In DBParser multiple comparison reports, only significant peptides and proteins are collated, whereas rejected peptides and proteins are shown only on rejected peptide reports. In the examples presented here, the Identity Score was used to threshold peptide assignment data.

DBParser System. DBParser has 4 major components: a web-based interface, a relational database, a parser and a report generator. A data flowchart is shown in Figure 1. The web-based interface facilitates a multiple independent user environment and provides users various options regarding data parsing, report generation and report viewing. CGI and HTML were used to create web-based interface. The parser and report generator programs were written in Perl (ActivePerl 5.8, downloaded from http://www.activestate.com) on a Windows 2000 system. MySQL was obtained from http://www.mysql.com, and used as the relational database. The Perl package manager bundled with ActivePerl was used to install the Perl DBI database interface and the MySQL database driver. Apache server was downloaded from http://www.apache.com and used as the web server. DBParser is run on a Dell Server PE 2650 with dual Intel Xeon, 2.8 GHz processors, 2GB RAM, and dual 136 GB hard drives under Windows 2000 server operating system. DBParser utilizes 3 types of databases: a flat file database, a public reference database, and a central database. The flat file database stores parsed data from Mascot flat files. It contains 10 indexed relational tables to track the experiment, search parameters, input, search history, ions, scores, peptide hits, protein hits, peptides, and proteins. The public reference database contains entries including GO terms, GO ID etc. from the online public databases. GO termdb database was downloaded from http://www.godatabase.org, and GOASPTR and Human GOA databases were downloaded from http://srs.ebi.ac.uk. These 3 public databases were parsed into a MySQL public database locally to assign GO terms and GO IDs. The central database stores the DBParser processes (all user initiated commands, e.g., database creation, data parsing and report generation) and is manipulated by a perl program on the DBParser server.

Parsimony Analysis. Peptides are classified into two categories: discrete or degenerate. On the basis of the distribution of discrete and degenerate peptides, protein sequence database records are classified into 6 hierarchical categories: distinct, differentiable, subsumable, superset, subset, and equivalent. An example of the application of these rules is included in Figure 2.

The example outlined in Figure 2 illustrates schematically *discrete* peptides as sequences that are assigned to exactly one protein, e.g., A, B, C, D, E, G, and H. In contrast, *degenerate* peptides are assigned to more than one protein, e.g., F, I, J, K, L, M, and N. Using these peptide definitions, *equivalent* proteins are based on the same set of degenerate peptide(s), e.g., protein 6 and protein 7. It is of importance to note that no information is discarded. In any designated category, all accession numbers are retained. In the case of equivalent proteins, these records are grouped together, one record is counted and the other equivalent records are flagged. The same counting procedure is propagated throughout the parsimony analysis. Each equivalent protein data set is referred to as a



109. YHM4_YEAST UPep = 4

Num	Search	QNum	PeptideSeq	IonsScore	HScore	IScore	m/z	Qmass
1	F007277.dat	565	EAVLTVPTNFSEEQK	31.05	22.79	30.17	846.654	1691.29
2	F007278.dat	1094	LISDYDADELAEALQPVIVNTPHLK	67.24	25.47	28.59	922.469	2764.38
3	F007277.dat	1226	LTTNLEYTLPESVEILGPQNK	70.06	24.02	28.91	1180.41	2358.8
4	F007278.dat	656	NASNNPNELAASGAALQAR	69.23	22.55	30.27	934.944	1867.87

110. ATPB_YEAST UPep = 3

Num	Search	QNum	PeptideSeq	IonsScore	HScore	IScore	m/z	Qmass
1	F007277.dat	788	FLSQPFAVAEVFTGIPGK	31.42	17.83	29.96	954.559	1907.1
2	F007279.dat	192	TVFIQELINNIAK	81.73	36.90	30.61	752.039	1502.06
3	F007278.dat	400	TVFIQELINNIAK	54.87	37.21	30.51	752.624	1503.23
4	F007277.dat	274	VLDTGGPISVPVGR	52.81	24.87	30.47	684.534	1367.05

Figure 3. *Sample A* peptide report. Peptides are grouped and associated with identified proteins, which are sorted by the number of unique peptides (UPep). The tabular display includes a peptide counter (Num), the flat file name (Search), the Mascot query number (QNum), the assigned peptide sequence (PeptideSeq), Ions Score (IonsScore), Homology Score (HScore), Identity Score (IScore), mass/charge (*m*/*z*) of the precursor ion, and relative molecular mass (Qmass) of the peptide. Note that the Ions Score of each listed peptide is equal to or greater than its Identity Score, consistent with the evaluation criterion used to extract data from the Mascot files. Peptides are color coded black or red. Use of the same color for consecutive peptide sequences indicates multiple detection events, i.e., redundant identification of the same peptide.

nonredundant equivalent protein and counted only once. *Subset* proteins contain peptides common to a larger set of peptides corresponding to another protein identification, e.g. protein 4 is a subset of protein 3, which is a *superset*. A *superset* protein contains the degenerate peptides from at least one other subset protein, e.g., protein 3 is a superset that includes all the peptides used to identify protein 4. A *subsumable* protein contains degenerate peptides that can be distributed as subsets of two or more other proteins, e.g., peptides J and K from protein 5 are a subset of protein 3, whereas peptides L and M of protein 5 are a subset of proteins 6 and 7. Formally, subsumable proteins are simply another class of subsets. The position of subsumable above superset in our hierarchy is not a qualitative statement, merely a procedural order. A *differentiable* protein can be distinguished from other proteins by having at least one discrete peptide, e.g., protein 2. Since protein 3 has one discrete peptide "H", it is promoted from superset to the differentiable category. A *distinct* protein is identified by only discrete peptide(s), e.g., protein 1. In summary, parsimony analysis reduces this list of seven proteins to a list of four nonredundant proteins, i.e., proteins 1, 2, 3, and 6.

The following rules were used for protein parsimony analysis: 1. Parsimony analysis is applied in ascending hierarchy: equivalent, subset, superset, subsumable, differential, and distinct proteins. 2. Each protein is counted in exactly one category. It is possible for a protein to be listed in more than one category, in which case it is counted in the highest category in which it occurs where distinct is the highest category and equivalent the lowest.

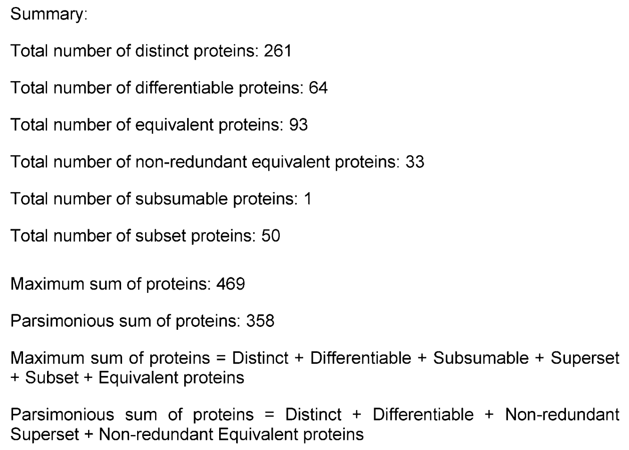
Testing, Validation, and Performance. DBParser was tested using Apache Server and Internet Explorer 6.0 running on Windows 2000 workstation, Windows 2000 server, Windows XP and Linux systems separately. Peptide and protein assignments from DBParser were confirmed with manual validation and Mascot search results from yeast Samples A, B, C as noted above. Parsing and report generation times were measured for flat files of various sizes and are reported in the Supplemental Data.

Results

The results of applying DBParser can best be illustrated with data files typical of complex mixture analyses. In the course of evaluating and optimizing a 2D-LC MS/MS system, many analyses of tryptic digests of yeast soluble proteins were obtained.[8] Peptides derived from yeast digests provide very complex mixtures that are readily available analytical test references. Multiple datasets containing tens of thousands of spectra were searched using Mascot. The resulting DBParser reports fall into two categories: (1) single sample summaries, applying the principles of parsimony to peptides observed from concatenated chromatographic separations, and (2) multiple sample summaries to permit comparisons of multiple concatenated datasets. We describe examples of four of the formats: peptide, protein parsimony analysis, unique peptide, and protein links reports for single samples, and a protein comparison report for three different samples.

Single Sample Summary. The report summary produced for a single sample may contain data ranging from a single LC–MS/MS experiment to multiple 2D LC–MS/MS runs. In 2D LC–MS/MS experiments, output files are summed from the analysis of each ion exchange fraction in order to generate a composite file representing all of the analytical data from one mixture of proteins. It is possible to concatenate multiple mass spectral datasets prior to using a sequence search engine. Alternatively, DBParser allows this concatenation after database searching, significantly reducing aggregate search times, and permitting selection of files for concatenation based upon individual file reports.

In all DBParser reports, proteins are ordered relative to the number of unique peptides assigned. Each spectrum has an assigned Mascot query number (Qnum) that provides an active link to the Mascot mass spectrum display, facilitating visual inspection of data records. Within each protein, peptide sequences are listed alphabetically, and then by Ions Score. Although parsimony dictates concise list generation, all data is retained and is accessible; a user may generate a rejected



Summary:

Total number of distinct proteins: 261

Total number of differentiable proteins: 64

Total number of equivalent proteins: 93

Total number of non-redundant equivalent proteins: 33

Total number of subsumable proteins: 1

Total number of subset proteins: 50

Maximum sum of proteins: 469

Parsimonious sum of proteins: 358

Maximum sum of proteins = Distinct + Differentiable + Subsumable + Superset + Subset + Equivalent proteins

Parsimonious sum of proteins = Distinct + Differentiable + Non-redundant Superset + Non-redundant Equivalent proteins

Figure 4. *Sample A* parsimony analysis report−summary section.

peptide report (a report listing peptides with Ions Score values below that of the Identity Score), or a rejected protein links report.

Example 1−Peptide Report. *Sample A* was analyzed using 2D LC−MS/MS, resulting in 6 flat files (corresponding to LC−MS/MS analysis of 6 ion exchange fractions). The peptide summary is the primary report of the significant peptides and proteins for combined data sets (see portion of report in Figure 3).

Example 2−Parsimony Analysis Reports. The Parsimony Analysis Report format is comparable to that described for the peptide report, sub-divided by category. A summary lists the number of proteins in each category and includes the equation used for calculating the reported totals (Figure 4). The format graphically presents all of the distinct proteins first, and then nondistinct proteins. This format emphasizes the relationships between nondistinct proteins sharing peptide sequences displayed as bars (Figure 5).

Example 3−Unique Peptide Report. There are two format options. The first option presents the unique peptides organized by protein identification. The second format displays a list of peptides along with the accession number of the proteins to which the peptide can be mapped (Figure 6).

Example 4−Protein Links Report. A representative protein links report (see Figure 7) displays related information about the significant proteins, links to the supporting peptide data as well as to relevant public databases.

Multiple Sample Summary. We have designed comparison reports for 2 to 6 data sets, where each data set may contain one or more flat files. Reports indicate the peptides, proteins and nonredundant proteins unique and common in each data set. For simplicity, the procedure makes pairwise comparisons between data sets as elements in a matrix, for example:

1 vs 2, 1 vs 3, 1 vs.4, 1 vs 5, 1 vs 6

2 vs 3, 2 vs 4, 2 vs 5, 2 vs 6

3 vs 4, 3 vs 5, 3 vs 6

4 vs 5, 4 vs 6

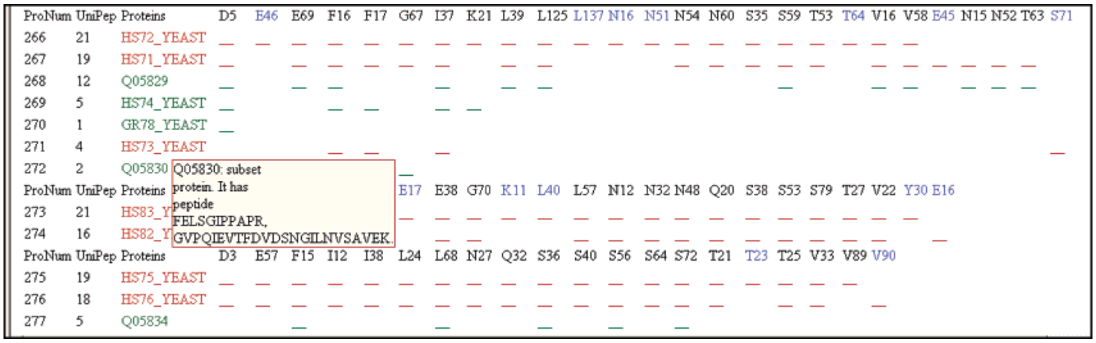
5 vs 6

From this matrix, DBParser tabulates which peptides are unique to each dataset, and which peptides are common to all datasets. These tabulations are available as pairwise output and aggregate output. There are four reports: peptide comparison, protein comparison, nonredundant protein comparison, and protein links. As with the single sample report, a user may choose to regenerate the Mascot search report or a rejected peptide report and rejected protein links report.

Example 5−Protein Comparisons for Yeast Samples A vs B vs C. For this example, DBParser compared 3 data sets of Mascot results. Each data set has 6 flat files. The report shows the proteins unique in A, B, and C samples separately, and then lists proteins common to all 3 data sets. There is also pairwise comparison between A and B, B and C, and A and C. Within each pairwise comparison, for example, A vs B, proteins unique in A, proteins unique in B, and proteins common in A and B are listed. This comparison report contains results in list and detail formats with active protein links. A summary accounting of proteins is included at the end of the report (Figure 8). Multiple peptide and nonredundant protein comparison reports are also available.

Discussion

Peptide analyses by LC/MS/MS can generate large datasets, imposing labor-intensive efforts to consolidate peptide and/or protein identification information into meaningful reports. Software bioinformatics tools can facilitate and accelerate high-throughput proteomic data analyses. Several software utilities for data processing have been reported such as DTASelect & Contrast,[10] and CHOMPER.[11] The utilities PeptideProphet,[12] ProteinProphet,[13] and SEQUEST-NORM[14] have been designed to improve the accuracy and validity of peptide and protein data analysis. DBParser differs from these utilities both in method and function, but it does share many of their objectives described previously.



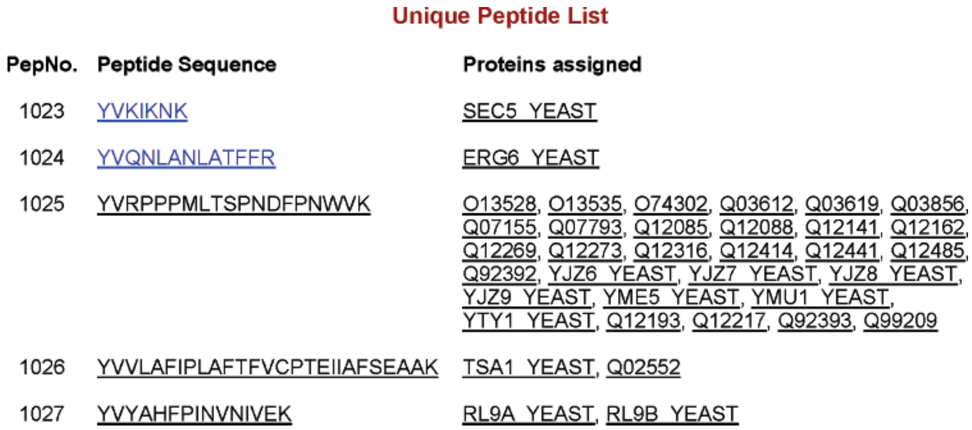
ProNum	UniPep	Proteins	D5	E46	E69	F16	F17	G67	I37	K21	L39	L125	L137	N16	N51	N54	N60	S35	S59	T53	T64	V16	V58	E45	N15	N52	T63	S71
266	21	HS72_YEAST	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—					
267	19	HS71_YEAST	—		—	—	—	—	—		—	—				—	—	—	—	—		—	—	—	—	—	—	
268	12	Q05829	—		—	—			—		—	—							—			—	—		—	—	—	
269	5	HS74_YEAST	—			—	—		—	—																		
270	1	GR78_YEAST	—																									
271	4	HS73_YEAST				—	—		—																			—
272	2	Q05830						—																				

Q05830: subset protein. It has peptide FELSGIPPAPR, GVPQIEVTFDVDSNGILNVSAVEK.

ProNum	UniPep	Proteins	E17	E38	G70	K11	L40	L57	N12	N32	N48	Q20	S38	S53	S79	T27	V22	Y30	E16
273	21	HS83_Y	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	
274	16	HS82_Y		—	—			—	—	—	—	—	—	—	—	—	—		—

ProNum	UniPep	Proteins	D3	E57	F15	I12	I38	L24	L68	N27	Q32	S36	S40	S56	S64	S72	T21	T23	T25	V33	V89	V90
275	19	HS75_YEAST	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	
276	18	HS76_YEAST	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—		—	—		—
277	5	Q05834			—				—			—		—		—						

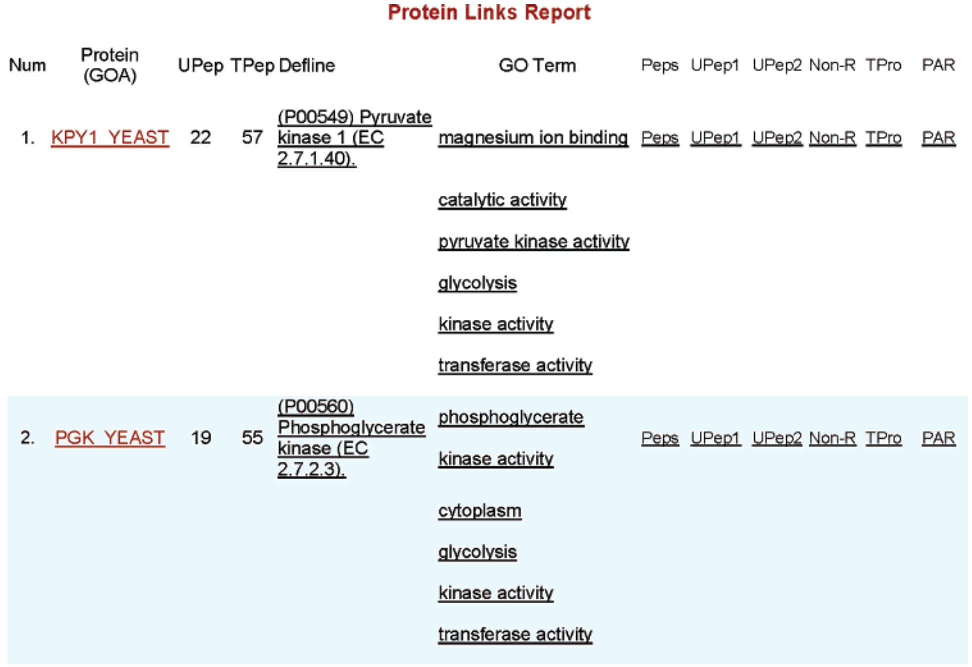
Figure 5. *Sample A* parsimony analysis display report−nondistinct protein section. Peptide sequence and protein details can be viewed using the mouse to reveal a tool tip window for any specific entry.



Unique Peptide List

PepNo.	Peptide Sequence	Proteins assigned
1023	YVKIKNK	SEC5_YEAST
1024	YVQNLANLATFFR	ERG6_YEAST
1025	YVRPPPMLTSPNDFPNWVK	O13528, O13535, O74302, Q03612, Q03619, Q03856, Q07155, Q07793, Q12085, Q12088, Q12141, Q12162, Q12269, Q12273, Q12316, Q12414, Q12441, Q12485, Q92392, YJZ6_YEAST, YJZ7_YEAST, YJZ8_YEAST, YJZ9_YEAST, YME5_YEAST, YMU1_YEAST, YTY1_YEAST, Q12193, Q12217, Q92393, Q99209
1026	YVVLAFIPLAFTFVCPTEIIAFSEAAK	TSA1_YEAST, Q02552
1027	YVYAHFPINVNIVEK	RL9A_YEAST, RL9B_YEAST

Figure 6. *Sample A* unique peptide report. Peptide sequences are listed alphabetically. Each discrete peptide (blue) is assigned to a single protein; each degenerate peptide (black) maps to more than one protein. Peptide sequences link to the report portion containing detailed information about each peptide.

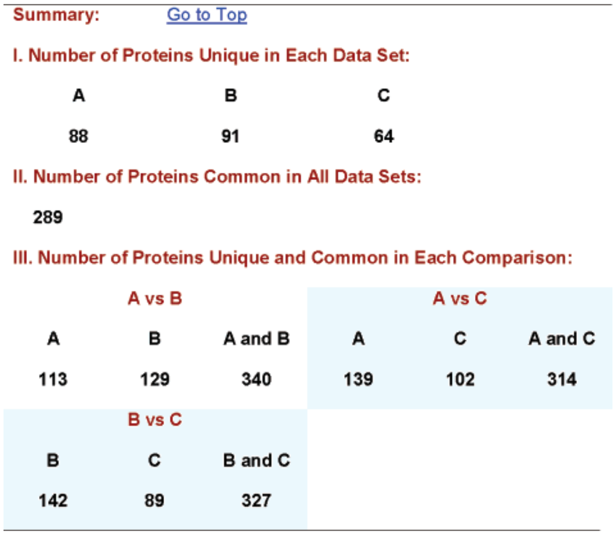


Protein Links Report

Num	Protein (GOA)	UPep	TPep	Defline	GO Term	Peps	UPep1	UPep2	Non-R	TPro	PAR
1.	KPY1_YEAST	22	57	(P00549) Pyruvate kinase 1 (EC 2.7.1.40).	magnesium ion binding catalytic activity pyruvate kinase activity glycolysis kinase activity transferase activity	Peps	UPep1	UPep2	Non-R	TPro	PAR
2.	PGK_YEAST	19	55	(P00560) Phosphoglycerate kinase (EC 2.7.2.3).	phosphoglycerate kinase activity cytoplasm glycolysis kinase activity transferase activity	Peps	UPep1	UPep2	Non-R	TPro	PAR

Figure 7. *Sample A* protein links report. The protein links report includes a protein counter (Num), accession reference (Protein), the number of unique peptides (UPep), the number of total peptides (TPep), definition line (Defline), and GO terms. If an entry is nonredundant, then the protein accession is colored maroon; redundant proteins are black. Protein accessions link to the corresponding entry at the GOA search website at http://srs.ebi.ac.uk. Definition lines link to the corresponding entry at the Swiss-Prot search website at http://us.expasy.org. Each GO term links to the corresponding entry at AmiGO website at http://www.godatabase.org. If the data contains human proteins, then protein name links are active to the corresponding NCBI LocusLink at http://www.ncbi.nlm.nih.gov/LocusLink/. Active links connect to the primary data reports for each experiment (Peps, peptide; UPep1, unique peptide format 1; UPep2, unique peptide format 2; TPro, total proteins; Non-R, nonredundant protein; PAR, parsimony analysis).

Protein sequence databases, e.g., NCBInr, contain a host of entries containing nearly identical sequence information. The apparent redundancies arise from natural sequence diversity, e.g, isoforms, as well as sequenced protein fragments and sequencing errors.[15] A significant consequence of shotgun proteomics is that the connectivity between peptides and their precursor proteins is lost. The exercise of assembling protein level information is complicated by these sequence database redundancies. For example, there are two isoforms of yeast enolase, ENO1 and ENO2, whose sequences are 95.4% identical. Complete tryptic digestion produces 30 and 29 peptides 6 residues in length or greater, respectively. Seventeen of these peptides are identical. One would not be able to discriminate whether ENO1 or ENO2 or both were present based on the detection of any of these 17 degenerate peptides. Current database search engines do not address the difficulty of counting the number of unique proteins identified based on observed peptides.[16] In an attempt to reconcile this problem, a nomenclature is defined for categorization of peptide and protein assignments. A central tenet of this nomenclature is that peptides do not identify proteins per se; they identify protein sequence database records. Extant database search engines output lists of proteins, or more correctly, lists of protein sequence database records, each with a unique accession number, which represent the maximum possible number of proteins that could account for all observed peptides.



Summary: Go to Top

I. Number of Proteins Unique in Each Data Set:

A	B	C
88	91	64

II. Number of Proteins Common in All Data Sets:

289

III. Number of Proteins Unique and Common in Each Comparison:

A vs B			A vs C		
A	B	A and B	A	C	A and C
113	129	340	139	102	314

B vs C		
B	C	B and C
142	89	327

Figure 8. Sample multiple protein comparison report−Summary section. It summarizes the number of proteins unique and common in sample A, B, and C and pairwise comparisons.

DBParser analyzes protein identifications based on peptide sequences to produce a parsimonious protein analysis report, generating a concise set of protein sequence database records that account for all of the observed peptides in the experimental data sets. Distinct and differentiable protein sequence database records are identified, having discrete peptides not found in other protein sequence database records. Because of the complexity of protein sequence databases, some protein sequence database records cannot be distinguished by the observed peptides, and are categorized into equivalent, subset and subsumable categories. Consequently, DBParser defines not only a parsimonious nonredundant protein sequence database record set, but also a complete list of nonexcludable, possible proteins in sets of data. DBParser results are consistent with those reported by Nesvizhskii et al.[13] who developed a statistical model to compute probabilities that proteins are present in a sample on the basis of peptides assigned to tandem mass spectra. Using only the logic of parsimony, DBParser analysis of Nesvizhskii's examples[13] produces identical results.

DBParser collates identification information such as the Mascot peptide Ions, Identity and Homology scores and merges other Mascot flat files using the parsimony principle to compile results from protein samples that may have been fractionated prior to mass spectral analysis. For example, proteolytic digests analyzed using 2D LC−MS/MS generate one file for each ion exchange fraction. With off-line protein fractionation, one sample can be further sub-divided, and upon mass spectral analysis produce multiple Mascot result files. Alternatively, data from samples analyzed several times in varying concentrations or using variable mass windows may be combined. It is possible to concatenate the MS/MS data from individual runs either prior or following database searching. We have found it convenient to concatenate files after searching, both for speed and to avoid reaching the Mascot limit of approximately 30 000 queries. DBParser rapidly merges the Mascot search results to generate a single composite report.

DBParser allows the user to specify the inclusion of only those peptides with Ions Scores greater than or equal to the Identity Score. In contrast, the Mascot Peptide Summary Report includes both significant and nonsignificant peptides for each assigned protein. With very large datasets, we have observed Mascot protein assignments that are based upon several peptides with low Ions Scores (i.e., Ions Score less than the Identity Score). In these instances, we also observe some peptides with significant Ion Scores (i.e., Ions Score > Identity Score) appearing in the Mascot unassigned queries list. DBParser lists peptides with an Ions Score lower than the Identity Score separately in a Rejected Peptides report. Significant peptides and proteins are more readily recognized in DBParser reports, especially when multiple files have been combined.

DBParser reports link to the corresponding entries in NCBInr, AmiGO, Swiss-Prot, and GOA public database websites, enabling researchers to determine the published characteristics and functionality of proteins identified in their experiments. The Gene Ontology Annotation (GOA) database utilizes "a dynamic controlled vocabulary that can be applied to all organisms even as protein knowledge is accumulating and changing."[17] Biologists that comprise the GO Consortium have "developed three separate ontologies: molecular function, biological process, and cellular component, which help to describe gene products in a standardized way and allow the annotation of molecular characteristics across species."[17]

Because DBParser is a web-based application, users only require web browser access. Multiple users can parse data and generate reports simultaneously as DBParser assigns each process a separate identification code. DBParser first parses the data from Mascot flat files into a database, increasing the efficiency of retrieving the data (such as significant peptide sequences) while generating reports. DBParser dramatically improves high-throughput proteomic data analysis for Mascot MS/MS Ions Search results. Further development of DBParser will integrate search results from other search algorithms such as Sequest and OMSSA, facilitating comparison of search results from different search algorithms, and providing an efficient data analysis tool for data mining in proteomic studies.

Conclusions

We have developed software that enhances the significance of data collected in large-scale shotgun tandem mass spectrometry experiments using automated peptide sequence search engines. DBParser utilizes parsimony analysis to filter lists of possible proteins and determine a nonredundant protein set that includes all assigned peptide sequences, providing a rational accounting of proteins identified from experimental data. DBParser then generates user-friendly html output reports for peptide and protein analyses. It provides comparison reports from multiple or concatenated experiments, significantly increasing the value of archived data.

DBParser is freely and publicly available from the authors under the terms of a Mozilla Open Source License agreement (http://www.mozilla.org/MPL/).

Acknowledgment. LNT staff, especially Dr. Kuniaki Saito, Dr. Suwako Fujigaki, Dr. Wenyao Shi, Dr. Marjan Gucek, Anthony Makusky, Junichi Masuda contributed useful suggestions and discussions as did Dr. Ming Xu of NCBI.

Supporting Information Available: Parsing flat file and generating report time (Table 1 and Table 2). This material is available free of charge at http://pubs.acs.org.

References

(1) Hunt, D. F. Personal commentary on proteomics. *J Proteome Res.* **2002**, *1*, 15−19.

(2) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17*, 676−682.

(3) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551−3567.

(4) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. An approach to correlate tandem mass-spectral data of peptides with amino acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(5) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. Open Mass Spectrometry Search Algorithm. *J Proteome Res.* **2004**, in press.

(6) Fenyo, D. Identifying the proteome: software tools. *Curr. Opin. Biotechnol.* **2000**, *11*, 391−395.

(7) Gomez, S. M.; Noble, W. S.; Rzhetsky, A. Learning to predict protein−protein interactions from protein sequences. *Bioinformatics* **2003**, *19*, 1875−1881.

(8) Maynard, D. M.; Masuda, J.; Yang, X.; Kowalak, J. A.; Markey, S. P. Characterizing Complex Peptide Mixtures Using a Multidimensional LC−MS System: *S. cerevisiae* as a Model System. *J. Chromatogr. B* **2004**, *810*(1), 69−76.

(9) Pappin, D. J.; Hojrup, P.; Bleasby, A. J. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **1993**, *3*, 327−332.

(10) Tabb, D. L.; McDonald, W. H.; Yates, J. R., III. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002**, *1*, 21−26.

(11) Eddes, J. S.; Kapp, E. A.; Frecklington, D. F.; Connolly, L. M.; Layton, M. J.; Moritz, R. L.; Simpson, R. J. CHOMPER: a bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies. *Proteomics.* **2002**, *2*, 1097−1103.

(12) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383−5392.

(13) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646−4658.

(14) MacCoss, M. J.; Wu, C. C.; Yates, J. R. III. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **2002**, *74*, 5593−5599.

(15) Rappsilber, J.; Mann, M. What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* **2002**, *27*, 74−78.

(16) Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; and Nesvizhskii, A. The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data. *Mol Cell Proteomics.* **2004**, *3*, 531−533.

(17) Camon, E.; Magrane, M.; Barrell, D.; Binns, D.; Fleischmann, W.; Kersey, P.; Mulder, N.; Oinn, T.; Maslen, J.; Cox, A.; Apweiler, R. The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS−PROT, TrEMBL, and InterPro. *Genome Res.* **2003**, *13*, 662−672.

PR049920X