W376–W381 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki461

DeNovoID: a web-based tool for identifying peptides from sequence and mass tags deduced from *de novo* peptide sequencing by mass spectroscopy

Brian D. Halligan*, Victor Ruotti, Simon N. Twigger and Andrew S. Greene[1]

Bioinformatics Research Center and [1]Biotechnology and Biomedical Engineering Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53213, USA

Received February 14, 2005; Revised April 4, 2005; Accepted April 13, 2005

ABSTRACT

One of the core activities of high-throughput proteomics is the identification of peptides from mass spectra. Some peptides can be identified using spectral matching programs like Sequest or Mascot, but many spectra do not produce high quality database matches. *De novo* peptide sequencing is an approach to determine partial peptide sequences for some of the unidentified spectra. A drawback of *de novo* peptide sequencing is that it produces a series of ordered and disordered sequence tags and mass tags rather than a complete, non-degenerate peptide amino acid sequence. This incomplete data is difficult to use in conventional search programs such as BLAST or FASTA. DeNovoID is a program that has been specifically designed to use degenerate amino acid sequence and mass data derived from MS experiments to search a peptide database. Since the algorithm employed depends on the amino acid composition of the peptide and not its sequence, DeNovoID does not have to consider all possible sequences, but rather a smaller number of compositions consistent with a spectrum. DeNovoID also uses a geometric indexing scheme that reduces the number of calculations required to determine the best peptide match in the database. DeNovoID is available at http://proteomics.mcw.edu/denovoid.

INTRODUCTION

Recent events in analytical chemistry and genomics have come together to make high-throughput proteomics a possibility (1). Many of the high-throughput proteomics efforts have adopted the so-called ‘bottom up’ or ‘shotgun’ approach (2,3). In the bottom up approach, the biological sample to be analyzed is converted from a complex mixture of proteins into an even more complex mixture of peptides by digestion with a protease of defined specificity, such as trypsin. Typically this mixture is then separated with liquid chromatography and introduced into a mass spectrometer by electrospray. The development of improved nanospray ionization methods allows for a high-resolution mass spectrometer to be used in-line with multidimensional liquid chromatography (2,4–6). This LC–MS/MS setup allows for the high-throughput analysis of complex mixtures of peptides produced from biological samples such as body fluids including blood and cerebral spinal fluid, or from extracts of cells or tissues.

As the genomic data for many species approaches completeness, the list of known and possible proteins produced by those species also approaches completeness (7–9). With these nearly complete proteomes in hand, it is now possible to search for matches between mass spectra and the particular proteome database with some degree of confidence that a form of the protein producing the observed peptide will be present in the database.

One of the major impediments to high-throughput proteomics continues to be problems associated with peptide identification from mass spectral data (10–12). The standard methods for comparing a spectrum to the proteome database involve the matching of the experimental spectra to theoretical spectra calculated from all of the predicted peptides. Studies measuring the accuracy of the peptide identifications under near ideal conditions find that only 7–10% of the identifications are correct (13). There are a number of reasons for this low success rate. First, many of the spectra may not be authentic peptide spectra or may be spectra of two or more co-eluting peptides. Second, the peptide may contain post-translational modifications or experimentally induced changes that were not accounted for by the search algorithm. Third, the peptide may

*To whom correspondence should be addressed. Tel: +1 414 456 8838; Fax: +1 414 456 6595; Email: Halligan@mcw.edu

© The Author 2005. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oupjournals.org

not completely match the proteome database due to allelic variation or errors in the database.

An approach that addresses the last two problems is *de novo* peptide sequencing which is an attempt to determine the complete or partial sequence of the peptide from the experimental spectra (14–19). *De novo* peptide sequencing tries to determine the mass difference between sequential *b* and sequential *y* ions in the MS/MS spectra. The mass differences can then be matched to amino acid masses and a potential sequence deduced. Several different methods using this approach have been developed (14,16,18,20–28). Often, the *de novo* peptide sequencing program only produces a limited number of ordered sequence tags; regions for which the amino acid composition and sequence is known. In addition, *de novo* peptide sequencing can also produce unordered sequence tags; regions for which the amino acid composition is known, but the order of the amino acids cannot be determined. Lastly, since the smallest and largest *b* and *y* ions are often not observed in the spectra or in regions that ionize poorly, the *de novo* peptide-sequencing program is unable to assign any sequence to that portion of the peptide and instead reports a mass tag. These mass tags often can correspond to multiple, completely independent amino acid compositions. Thus, the typical result from *de novo* peptide sequencing is a collection of ordered sequence tags, unordered sequence tags and mass tags that are consistent with large number of potential peptide sequences.

Although *de novo* peptide sequencing is more robust with respect to peptide modifications or changes than the spectral matching approach, the degenerate results produced by *de novo* peptide sequencing do not answer the fundamental question: which proteins are present in the original sample. To do this, the results obtained from *de novo* peptide sequencing must be used in a database search. The matching algorithms used by the common search programs, BLAST (29) and FASTA (30,31), are suboptimal with small sequence queries such as the typical peptides observed in these experiments, 8–25 amino acids. Versions of these programs to deal better with these smaller queries have been produced (24,25,32,33), but fundamentally, the problem of incomplete knowledge of the sequence still exists. Additionally, all of the possible peptides consistent with the set of degenerate unordered sequence and mass tags need to be calculated and then submitted to the search program. This greatly increases the search time and also raises the expectation that a false positive identification will result.

We have developed a method for indexing and searching a peptide database that is not dependent on the amino acid sequence of the peptide, only its amino acid composition (34). This method works optimally with peptides in the size range observed in the high-throughput proteomics experiments described above. Since this method uses a hierarchical index instead of pairwise comparisons, it requires far fewer computational steps to identify the best match in a peptide database. By marrying this database search algorithm with a web-based front end that accepts the degenerate *de novo* peptide sequencing results, we have produced the DeNovoID web service. DeNovoID is capable of expanding the *de novo* peptide sequencing results, searching a pre-indexed peptide database and determining the closest peptide match in the database to the *de novo* sequencing results.

MATERIALS AND METHODS

Algorithms

The results from a *de novo* peptide sequencing experiment are used to determine the potential amino acid compositions that then become a query to the PepID indexed peptide database. The ordered and unordered sequence tags are combined and composition information is extracted. The mass tags are compared to a list of potential peptide masses to identify all possible combinations of up to three amino acids that match the mass tag within the given experimental error. A list of all possible mass tag compositions are combined with the composition deduced from the sequence tags to produce a comprehensive list of possible amino acid combinations that are consistent with the *de novo* peptide sequencing experimental results. This list of potential peptide compositions is then submitted to the PepID database search algorithm.

The PepID algorithm for constructing and searching peptide databases has been previously described by Halligan *et al.* (34) and is outlined in Figure 1A. Briefly, a file of protein sequences in fasta format is parsed and the sequences digested *in silico* using the trypsin cleavage specificity to produce a list of potential peptides. The peptides, smaller than those that would be expected to be experimentally observed, are removed, since they do not contain enough information to be uniquely mapped to a single protein in the database. This has the benefit of speeding the ensuing calculations. The peptides are then converted into 18 dimensional vectors based on their amino acid composition. Eighteen dimensions are used instead of twenty dimensions because two pairs of amino acids, glutamine and lysine or leucine and isoleucine, are difficult or impossible to distinguish based on mass, so this pair of amino acids is represented by one dimension. The vectors are then clustered based on Euclidean distance into large top-level clusters, which are then further sub-clustered into smaller clusters, each of which have an associated list of peptides (Figure 1A). To search the database, the compositions deduced from the *de novo* peptide sequencing results are converted to vectors as above and the hierarchy of the database is traversed by determining the cluster centers at each level with the shortest Euclidean distance to the experimental vector (Figure 1B).

RESULTS AND DISCUSSION

Purpose

One of the major drawbacks of *de novo* peptide sequencing is that the degenerate results that it often produces do not directly answer the fundamental question that the investigator wants to answer: what proteins are present in the biological sample being analyzed? To answer this question, the results of the *de novo* peptide sequencing must be used to query a protein or peptide database. When using the normal tools for protein database searching, BLAST and FASTA, two difficulties arise. One problem is that the short length of the query sequences do not perform well with the BLAST or FASTA search algorithms. The FASTA search algorithm has been improved to better utilize short peptides as a query (33). It does not however accept mass tags and unordered sequence tags. These must first be converted to all possible combinations of sequences and then submitted.

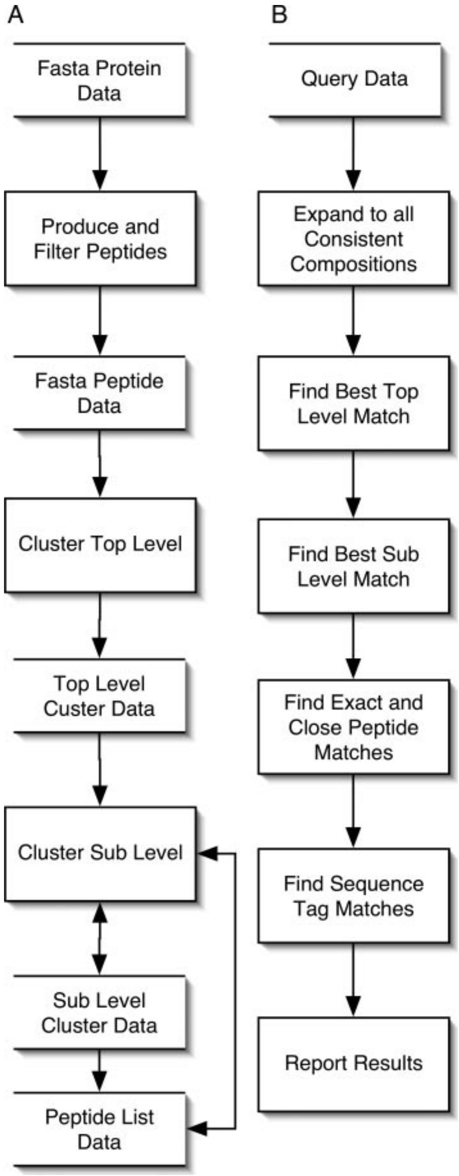

A
Fasta Protein Data
Produce and Filter Peptides
Fasta Peptide Data
Cluster Top Level
Top Level Custer Data
Cluster Sub Level
Sub Level Cluster Data
Peptide List Data
B
Query Data
Expand to all Consistent Compositions
Find Best Top Level Match
Find Best Sub Level Match
Find Exact and Close Peptide Matches
Find Sequence Tag Matches
Report Results

Figure 1. Overview of PepID/DeNovoID algorithms. (**A**) Flowchart describing creation of the indexed peptide database. (**B**) Flowchart describing the method for matching a query peptide to the closest database entry.

DeNovoID provides an important advance over previous methods in that it allows the results obtained from *de novo* peptide sequencing to be searched against an indexed peptide database.

Audience

The intended audience for the DeNovoID web service is proteomics researchers using *de novo* peptide sequencing to identify peptides by mass spectrometry. It has been found that many spectra that appear to be authentic peptide spectra do not lead to confident identifications using the standard database search methods. A number of commercial and open source programs are now available to perform *de novo* peptide sequencing with these spectra (20,22,24,32,35,36) (DeNovoX, Thermo-Finnigan). The DeNovoID web service complements these efforts by providing a means to take the output of the *de novo* peptide sequencing programs and carry out an optimized database search in a single step.

Technical description

Scoring. The algorithm used by DeNovoID is based on mapping the composition of each peptide to a vector, allowing peptides to be compared using vector algebra. Eighteen dimensions are used instead of twenty because pairs of amino acids that are difficult to distinguish by mass spectrometry (isoleucine and leucine, and lysine and glutamine) are combined. The peptide composition is mapped to each of the elements of the vector using the following formula: $\eta_i = \alpha \ln(\nu_i + 1) - \beta$, $i = 1, \ldots, 18$, where η_i is the standardized value and ν_i is the fractional composition of the amino acid in the peptide. The standardization parameters α and β were chosen to be $\alpha = 6$, $\beta = 3$ to provide an appropriate set of vector lengths for use by the clustering algorithm. The Euclidean distance between two vectors is a measure of dissimilarity between the peptides and is used as the score. A score of 0 indicates compositional identity and a score of $\leqslant 1$ indicates a high degree of compositional similarity.

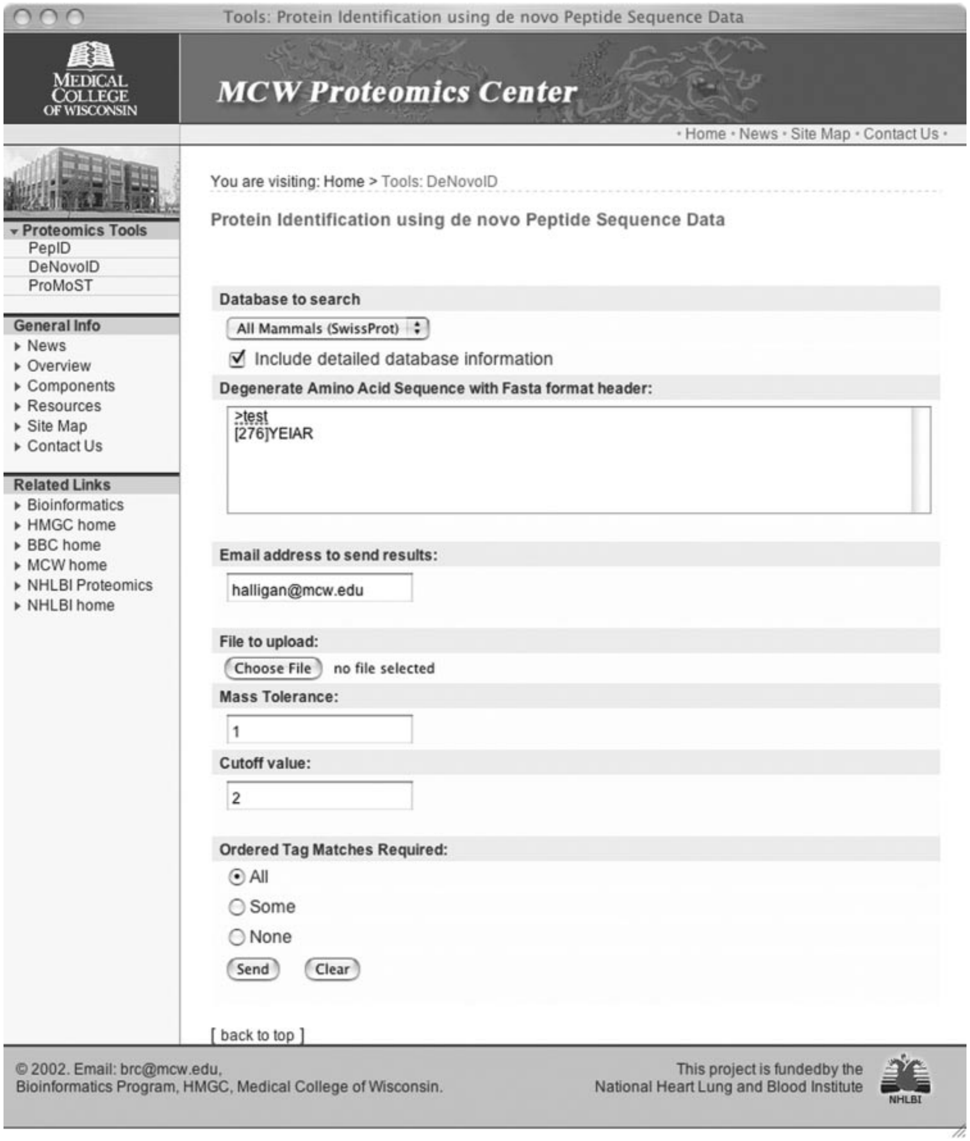
Mass tags. Mass tags are converted to amino acid compositions by including all possible combinations of up to four amino acids that match the mass tag within the specified mass tolerance. The default tolerance for the accuracy of the mass determination is 1 Da and the mass tolerance can be altered by the user using the web interface. Additionally, the user can have the mass of the tag automatically adjusted for ion type. Since a *b* ion carries an additional N-terminal H, the mass of the tag should be reduced by 1 Da before searching. DeNovoID automates this process. If the user includes a # symbol at the end of the mass tag, DeNovoID will automatically subtract 1 Da from the *b* ion mass tag before generating the matching amino acid combinations. Similarly, a *y* ion carries an N-terminal H and a C-terminal OH. Including '@' symbol at the end of the mass tag causes DeNovoID to automatically subtract 19 Da from the mass tag.

Overview

Input. The input to the DeNovoID web service can take several different formats. The simplest use is to paste a query sequence in the sequence entry box on the form. The query is in the fasta format sequence composed of a header line beginning with the '>' character followed by an optional description of the query, followed by one or more lines containing the query. The query can consist of a mixture of ordered sequence tags, disordered sequence tags (sequence contained within square brackets) and mass tags (numeric values contained within square brackets). DeNovoID takes all three types of information and creates a set of possible peptide compositions for the search (Figure 2).

Multiple sequences can be submitted in batch by inputting multiple queries in the sequence entry box on the form. Each query begins with a fasta header line. A text file containing multiple queries in the same format can also be uploaded by selecting the file using the browse function of the form. Entering a filename for upload takes precedence over data entered in the sequence entry box on the form.

While many searches with simple single queries can be completed rapidly, complex searches with multiple mass tags



You are visiting: Home > Tools: DeNovoID

Protein Identification using de novo Peptide Sequence Data

Database to search

All Mammals (SwissProt)

☑ Include detailed database information

Degenerate Amino Acid Sequence with Fasta format header:

```
>test
[276]YEIAR
```

Email address to send results:

halligan@mcw.edu

File to upload:

no file selected

Mass Tolerance:

1

Cutoff value:

2

Ordered Tag Matches Required:

● All

○ Some

○ None

© 2002. Email: brc@mcw.edu,
Bioinformatics Program, HMGC, Medical College of Wisconsin.

This project is fundedby the
National Heart Lung and Blood Institute

Figure 2. DeNovoID User Interface showing a query containing a mass tag.

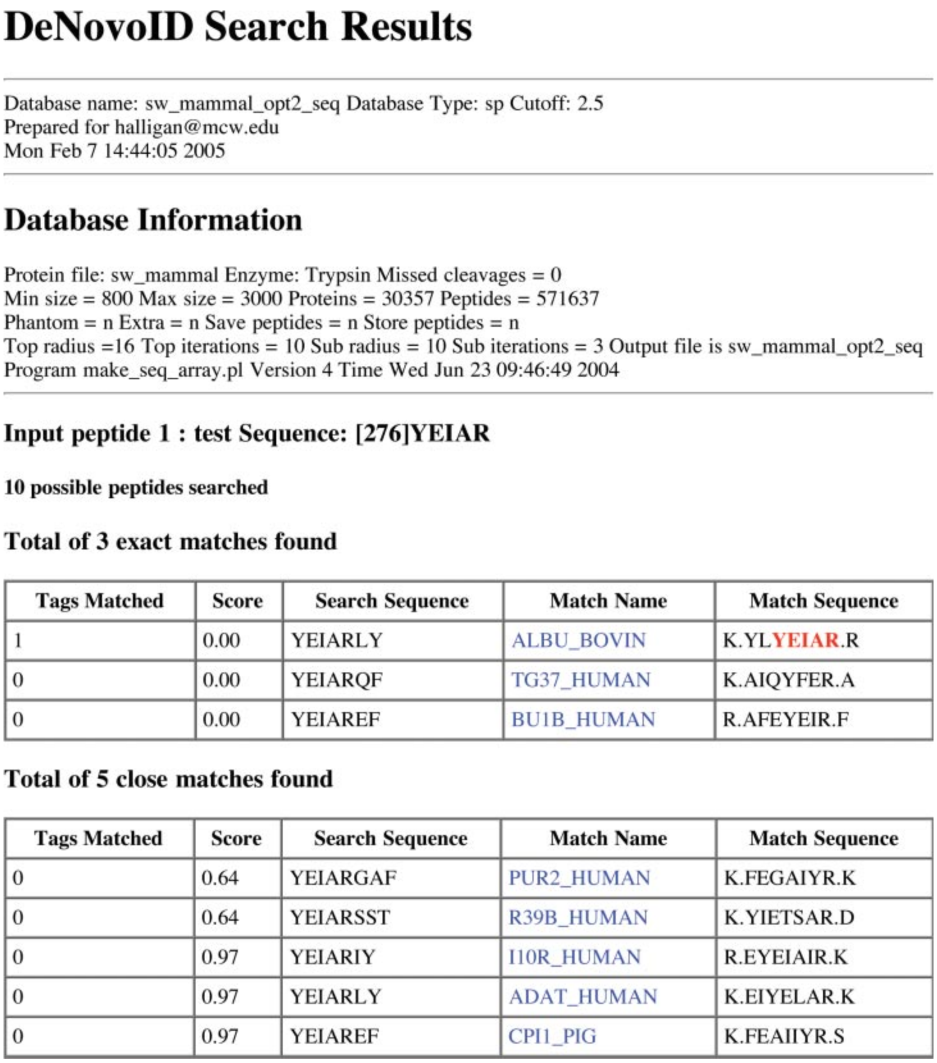
or batch searches with multiple queries often take longer to complete than the time-out limit (∼1 min) of some web browsers. For this reason, the results of the DeNovoID search are returned by email as an attached hypertext markup language (html) document. A box on the form is provided to enter the email address to which the results should be returned.

The PepID algorithm determines the quality of match between a query and a database peptide by measuring the Euclidean distance between the vectors representing the query and database peptides. The cutoff score box provides the user an opportunity to examine matches that are close, as well as those that are exact. If zero or no value is entered in the cutoff score box, only exact matches are reported. A cutoff value of $\leq$2.5 is usually sufficient to detect all biologically relevant matches and allows for the identification of peptides with one or two amino acid substitutions. Amino acid substitution is an important reason for failure of spectral correlation based database search algorithms.

Output. The output of the DeNovoID search is an html file that is returned to the user by email as an attached file. The html file has three sections: a header, a description of the database used for the search and the results of the search.

The header contains information about which database was selected, the type of the database, the cutoff value to be used for reporting the results of the search, the email address to which the results are to be returned, and the time and date of the search. The database description section contains the name of the fasta file used to produce the database, the enzyme cleavage used and the number of missed cleavages allowed, the minimum and maximum peptide sizes for inclusion in the databases, the number of proteins and peptides included in the database, the options used in generating the database and the program version used to generate the database, and date on which it was generated (Figure 3).

The results section of the html file is organized into blocks corresponding to each query. The header and query are



DeNovoID Search Results

Database name: sw_mammal_opt2_seq Database Type: sp Cutoff: 2.5
Prepared for halligan@mcw.edu
Mon Feb 7 14:44:05 2005

Database Information

Protein file: sw_mammal Enzyme: Trypsin Missed cleavages = 0
Min size = 800 Max size = 3000 Proteins = 30357 Peptides = 571637
Phantom = n Extra = n Save peptides = n Store peptides = n
Top radius =16 Top iterations = 10 Sub radius = 10 Sub iterations = 3 Output file is sw_mammal_opt2_seq
Program make_seq_array.pl Version 4 Time Wed Jun 23 09:46:49 2004

Input peptide 1 : test Sequence: [276]YEIAR

10 possible peptides searched

Total of 3 exact matches found

Tags Matched	Score	Search Sequence	Match Name	Match Sequence
1	0.00	YEIARLY	ALBU_BOVIN	K.YL**YEIAR**.R
0	0.00	YEIARQF	TG37_HUMAN	K.AIQYFER.A
0	0.00	YEIAREF	BU1B_HUMAN	R.AFEYEIR.F

Total of 5 close matches found

Tags Matched	Score	Search Sequence	Match Name	Match Sequence
0	0.64	YEIARGAF	PUR2_HUMAN	K.FEGAIYR.K
0	0.64	YEIARSST	R39B_HUMAN	K.YIETSAR.D
0	0.97	YEIARIY	I10R_HUMAN	R.EYEIAIR.K
0	0.97	YEIARLY	ADAT_HUMAN	K.EIYELAR.K
0	0.97	YEIAREF	CPI1_PIG	K.FEAIIYR.S

Figure 3. DeNovoID Results Output A peptide from bovine serum albumin (YLYEIAR) is used to demonstrate DeNovoID's ability to use mass tags as input. The first two amino acids (YL) are replaced with their approximate molecular weight (276 Da) and the query is submitted to the DeNovoID. Results are returned by email as an html file attachment. Links to sequence databases are shown in blue and matched tags are shown in red.

reported as well as the number of possible peptide compositions generated corresponding to the query. A table of exact matches is reported showing the number of ordered sequence tags that were matched in the query, the peptide composition used for the search, the name of the protein matched and the peptide sequence that matched with the flanking amino acids indicated. The protein name is also a link to the database entry for that protein. Ordered sequence tags that are found in the matched peptide are indicated by the use of bold red text. If a non-zero value is supplied for the cutoff value, a table of close matches, with a score less than or equal to the cutoff value is also reported.

CONCLUSIONS AND FUTURE DIRECTIONS

As high-throughput proteomics matures as a field, it is clear that amino acid variation and modification are important factors in the analysis of proteomic data. By their inherent nature, database search algorithms that rely on spectral matching are limited by these types of differences between the database and actual peptides. To overcome these limitations, *de novo* peptide sequencing algorithms can be used to determine sequence tags directly from the spectral information. This approach is also limited in that incomplete sequence determination often leads to results containing disordered sequence and mass tags. These degenerate elements make using these results in standard sequence search algorithms difficult. Another webservice, Spider (http://bif.csd.uwo.ca/spider), which also searches protein databases with a query composed of sequence and mass tags has been recently released, further demonstrating the need for this type of analysis.

DeNovoID provides a method for searching a peptide database using amino acid composition and is optimized for peptides of the typical length observed in high-throughput proteomics experiments. By using composition instead of sequence, as well as an algorithm that is peptide length independent, many of the problems associated with standard sequence search methods are avoided. Currently, DeNovoID takes text queries in fasta format. In the future, we would like to extend DeNovoID to accept output files from the major *de novo* peptide sequencing programs and to provide a facility so that other programs can easily submit queries to DeNovoID as a link in their results.

ACKNOWLEDGEMENTS

This work was supported in part by NHLBI Proteomics Center Grant (NIH-BAA-HL-02-04). Funding to pay the Open Access publication charges for this article was provided by the Biotechnology and Bioengineering Center at the Medical College of Wisconsin.

Conflict of interest statement. None declared.

REFERENCES

1. Patterson,S.D. and Aebersold,R.H. (2003) Proteomics: the first decade and beyond. *Nature Genet*., **33**, 311–323.
2. Wu,C.C. and MacCoss,M.J. (2002) Shotgun proteomics: tools for the analysis of complex biological systems. *Curr. Opin. Mol. Ther*., **4**, 242–250.
3. McDonald,W.H. and Yates,J.R.,III (2003) Shotgun proteomics: integrating technologies to answer biological questions. *Curr. Opin. Mol. Ther*., **5**, 302–309.
4. Washburn,M.P., Wolters,D. and Yates,J.R.,III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol*., **19**, 242–247.
5. Washburn,M.P., Ulaszek,R., Deciu,C., Schieltz,D.M. and Yates,J.R.,III (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal. Chem*., **74**, 1650–1657.
6. Liu,H., Lin,D. and Yates,J.R.,III (2002) Multidimensional separations for protein/peptide analysis in the post-genomic era. *Biotechniques*, **32**, 898, 900, 902 passim.
7. Yates,J.R.,III (2000) Mass spectrometry. From genomics to proteomics. *Trends Genet*., **16**, 5–8.
8. Pandey,A. and Mann,M. (2000) Proteomics to study genes and genomes. *Nature*, **405**, 837–846.
9. Cagney,G., Amiri,S., Premawaradena,T., Lindo,M. and Emili,A. (2003) *In silico* proteome analysis to facilitate proteomics experiments using mass spectrometry. *Proteome Sci*., **1**, 5.
10. Fenyo,D. (2000) Identifying the proteome: software tools. *Curr. Opin. Biotechnol*., **11**, 391–395.
11. Pevzner,P.A., Dancik,V. and Tang,C.L. (2000) Mutation-tolerant protein identification by mass spectrometry. *J. Comput. Biol*., **7**, 777–787.
12. Pevzner,P.A., Mulyukov,Z., Dancik,V. and Tang,C.L. (2001) Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res*., **11**, 290–299.
13. Keller,A., Purvine,S., Nesvizhskii,A.I., Stolyar,S., Goodlett,D.R. and Kolker,E. (2002) Experimental protein mixture for validating tandem mass spectral analysis. *OMICS*, **6**, 207–212.
14. Dancik,V., Addona,T.A., Clauser,K.R., Vath,J.E. and Pevzner,P.A. (1999) *De novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol*., **6**, 327–342.
15. Gevaert,K. and Vandekerckhove,J. (2000) Protein identification methods in proteomics. *Electrophoresis*, **21**, 1145–1154.
16. Chen,T., Kao,M.Y., Tepel,M., Rush,J. and Church,G.M. (2001) A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol*., **8**, 325–337.
17. Cagney,G. and Emili,A. (2002) *De novo* peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat. Biotechnol*., **20**, 163–170.
18. Spengler,B. (2004) *De novo* sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by Fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom*., **15**, 703–714.
19. Lu,B. and Chen,T. (2003) A suboptimal algorithm for *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol*., **10**, 1–12.
20. Clauser,K.R., Baker,P. and Burlingame,A.L. (1999) Role of accurate mass measurement (+/− 10 p.p.m.) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem*., **71**, 2871–2882.
21. Shevchenko,A., Chernushevich,I., Wilm,M. and Mann,M. (2000) *De novo* peptide sequencing by nanoelectrospray tandem mass spectrometry using triple quadrupole and quadrupole/time-of-flight instruments. *Methods Mol. Biol*., **146**, 1–16.
22. Lu,B.W. and Chen,T. (2003) A suboptimal algorithm for *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol*., **10**, 1–12.
23. Tabb,D.L., Saraf,A. and Yates,J.R.,III (2003) GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem*., **75**, 6415–6421.
24. Taylor,J.A. and Johnson,R.S. (2001) Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal. Chem*., **73**, 2594–2604.
25. Shevchenko,A., Sunyaev,S., Loboda,A., Shevehenko,A., Bork,P., Ens,W. and Standing,K.G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time of flight mass spectrometry and BLAST homology searching. *Anal. Chem*., **73**, 1917–1926.
26. Ma,B., Zhang,K., Hendrie,C., Liang,C., Li,M., Doherty-Kirby,A. and Lajoie,G. (2003) PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom*., **17**, 2337–2342.
27. Sunyaev,S., Liska,A.J., Golod,A. and Shevchenko,A. (2003) MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal. Chem*., **75**, 1307–1315.
28. Searle,B.C., Dasari,S., Turner,M., Reddy,A.P., Choi,D., Wilmarth,P.A., McCormack,A.L., David,L.L. and Nagalla,S.R. (2004) High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS *de novo* sequencing results. *Anal. Chem*., **76**, 2220–2230.
29. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
30. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
31. Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*., **183**, 63–98.
32. Johnson,R.S. and Taylor,J.A. (2002) Searching sequence databases via *de novo* peptide sequencing by tandem mass spectrometry. *Mol. Biotechnol*., **22**, 301–315.
33. Mackey,A.J., Haystead,T.A.J. and Pearson,W.R. (2002) Getting more from less—algorithms for rapid protein identification with multiple short peptide sequences. *Mol. Cell Proteomics*, **1**, 139–147.
34. Halligan,B.D., Dratz,E.A., Feng,X., Twigger,S.N., Tonellato,P.J. and Greene,A.S. (2004) Peptide identification using peptide amino acid attribute vectors. *J. Proteome Res*., **3**, 813–820.
35. Chen,T., Kao,M.Y., Tepel,M., Rush,J. and Church,G.M. (2001) A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *J. Comput. Biol*., **8**, 325–337.
36. Fernandez-de-Cossio,J., Gonzalez,J., Betancourt,L., Besada,V., Padron,G., Shimonishi,Y. and Takao,T. (1998) Automated interpretation of high-energy collision-induced dissociation spectra of singly protonated peptides by ‘SeqMS’, a software aid for *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom*., **12**, 1867–1878.