

High-Throughput Identification of Proteins and Unanticipated Sequence Modifications Using a Mass-Based Alignment Algorithm for MS/MS de Novo Sequencing Results

Brian C. Searle,[†] Surendra Dasari,[†] Mark Turner,[†] Ashok P. Reddy,[†] Dongseok Choi,[‡] Phillip A. Wilmarth,[§] Ashley L. McCormack,^{||} Larry L. David,[§] and Srinivasa R. Nagalla^{*†}

Department of Pediatrics, Department of Public Health & Preventive Medicine, and School of Dentistry, Oregon Health & Sciences University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239-3098, and Oregon National Primate Research Center, Oregon Health & Sciences University, 505 NW 185th Avenue, Beaverton, Oregon 97006-3448

With the increasing availability of de novo sequencing algorithms for interpreting high-mass accuracy tandem mass spectrometry (MS/MS) data, there is a growing need for programs that accurately identify proteins from de novo sequencing results. De novo sequences derived from tandem mass spectra of peptides often contain ambiguous regions where the exact amino acid order cannot be determined. One problem this poses for sequence alignment algorithms is the difficulty in distinguishing discrepancies due to de novo sequencing errors from actual genomic sequence variation and posttranslational modifications. We present a novel, mass-based approach to sequence alignment, implemented as a program called OpenSea, to resolve these problems. In this approach, de novo and database sequences are interpreted as masses of residues, and the masses, rather than the amino acid codes, are compared. To provide further flexibility, the masses can be aligned in groups, which can resolve many de novo sequencing errors. The performance of OpenSea was tested with three types of data: a mixture of known proteins, a mixture of unknown proteins that commonly contain sequence variations, and a mixture of posttranslationally modified known proteins. In all three cases, we demonstrate that OpenSea can identify more peptides and proteins than commonly used database-searching programs (SEQUEST and ProteinLynx) while accurately locating sequence variation sites and unanticipated posttranslational modifications in a high-throughput environment.

Tandem mass spectrometry (MS/MS) is a commonly used tool in the high-throughput identification of proteins.¹ Several software packages^{2–5} have been developed to identify proteins present in

samples by utilizing the amino acid sequence specific information in MS/MS spectra of peptides to search protein sequence databases. These programs typically rely on a whole peptide mass filter, where candidate peptides from the database are compared to the unknown MS/MS spectra only if they match the experimental mass of the parent ion. This method is sufficiently reliable for high-throughput identification of proteins with known amino acid sequences. However, if the sample peptide differs from the database sequence due to sequence variation or database sequence errors, or if the peptide contains sites of posttranslational modifications, the calculated mass from the database sequence may no longer match the measured mass.

In these cases, other strategies can be tried. One possibility is to create a database of proteins that contains all possible combinations of common modifications and to search unknown spectra against the new database.⁶ However, with an exhaustive search, the number of combinations of modifications that must be tested can grow prohibitively large. Since it is more likely to have modified peptides of proteins already present in a sample, an efficient technique is to search for modified forms of only those proteins identified in an initial database search.^{7–9} This optimization method is used by AutoMod, a subroutine of ProteinLynx,⁵ and it can significantly reduce the search space. However, it does require the identification of at least one unmodified peptide in the initial database search and is limited to identifying only

* Corresponding author. Phone: (503) 494-1928. Fax: (503) 494-4821. E-mail: nagallas@ohsu.edu.

[†] Department of Pediatrics.

[‡] Department of Public Health & Preventive Medicine.

[§] School of Dentistry.

^{||} Oregon National Primate Research Center.

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198–207.

(2) Eng, J. K.; McCormack, A. L.; Yates, J. R., III *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.

(3) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551–3567.

(4) Field, H. I.; Fenyö, D.; Beavis, R. C. *Proteomics* **2002**, *36*–47.

(5) Denny, R.; Neeson, K.; Rennie, C.; Richardson, K.; Leicester, S.; Swainston, N.; Worroll, J.; Young, P. The Use of Search Workflows in Peptide Assignment From MS/MS Data. Association of Biomolecular Resource Facilities, ABRF '02: Biomolecular Technologies: Tools for Discovery in Proteomics and Genomics, Austin, Texas, March 9–12, 2002.

(6) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426–1436.

(7) Gatlin, C. L.; Eng, J. K.; Cross, S. T.; Detter, J. C.; Yates, J. R., III *Anal. Chem.* **2000**, *72*, 757–763.

(8) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. *Genome Res.* **2001**, *11*, 290–299.

(9) Creasy, D. M.; Cottrell, J. S. *Proteomics* **2002**, *2*, 1426–1434.

peptides modified in ways represented by the new protein database.

Another technique is either to match ion series in MS/MS spectra to peptide sequences without using a stringent parent ion mass filter^{8,10} or to match short peptide sequence motifs to features in spectra.¹¹ Using these methods, unanticipated protein modifications and sequence variations can be identified, provided that they do not alter the masses of a significant number of sequence-specific ions. However, both approaches often assign high scores to incorrect peptide identifications by chance, thereby limiting their application in high-throughput environments. As with AutoMod, the search space can be limited by identifying candidate proteins from unmodified peptides with database-searching programs; but again, extensive manual verification is often still required.

A third potentially high-throughput approach is GutenTag,¹² an automated and enhanced version of the sequence tag method^{13,14} that relies on searching for short amino acid sequences derived from tandem mass spectra in protein sequence databases. The GutenTag scoring system, which is a combination of five factors (a tag match, a mass match on either side of the tag, and a tryptic-termini match on either side of the peptide), has been shown to be extremely reliable when identifying unmodified peptides. Unfortunately, the sequence tag method can still assign high scores to incorrect matches when attempting to identify modified peptides because only three of the five scoring factors can normally be used.

The manual interpretation of spectra, called de novo sequencing, is an approach that can sequence peptides without using database-searching programs.¹⁵ MS/MS spectra commonly contain short series of fragment ions where the mass differences between these ions match the masses of amino acids in the original peptide. These mass differences can be linked together to form partial or complete peptide sequences.¹⁶ Areas of MS/MS spectra that cannot be assigned to standard amino acids may be due to incomplete peptide fragmentation or to posttranslational modifications that change the mass of amino acids. The manual interpretation of spectra is time-consuming and requires considerable expertise. Fortunately, there are several commercial^{17–19} and

freely available^{20–23} software packages that perform automated de novo sequencing. These programs take into consideration much of the possible variation in peptide fragmentation and introduce the possibility of high-throughput, objective MS/MS sequencing.

One difficulty is that de novo sequencing algorithms often report several equally well-scoring sequences for a single spectrum, as well as ambiguous regions where the order or identity of two or more amino acids in the proposed sequence is uncertain. De novo sequencing algorithms also commonly misjudge the order of two or more residues or mislabel residues as isobaric equivalents. High mass accuracy can help alleviate the difficulty of assigning isobaric amino acids correctly. However, isomers such as leucine and isoleucine cannot be differentiated via low-energy tandem mass spectrometry. Error-tolerant search engines must be used to differentiate sections of the de novo sequence that are inappropriately assigned by the sequencing algorithm from actual amino acid variations and posttranslational modifications.

In the past, existing sequence alignment algorithms^{24,25} have been modified in order to match de novo sequences to protein sequence databases. For example, MS-BLAST,²⁶ MS-Shotgun,²⁷ and FASTS²⁸ can be used to align de novo sequences to database homologues using highly efficient sequence alignment algorithms. These programs use a modified mutation matrix to account for single-residue isobars and can identify sequence differences or possible modification sites. It is possible to account for ambiguous regions by submitting a new search for every possible combination of amino acids that could add up to the summed mass of amino acids in that region. As the number of ambiguous regions in a de novo sequence grows, it quickly becomes more difficult to interpret the search results. Another program, CIDentity,²⁹ attempts to correct for de novo sequencing errors by employing a rescoring approach. After an alignment is made, unresolved mono- and dipeptides can be matched to an adjacent section of the database sequence if they are isobars. The addition of this rescoring step can resolve some common de novo sequencing errors and produce more accurate identifications.

The sequence homology approach discussed above is limited in several ways when trying to match de novo sequences containing ambiguous regions to database sequences. First, this approach can only consider a small number of specific isobaric equivalences, making it difficult to separate de novo sequencing errors from actual sequence modifications. Second, it is often impossible to analyze marginal de novo sequences derived from

- (10) Clauser, K. R.; Baker, P.; Burlingame, A. L. Peptide Fragment-Ion Tags from MALDI/PSD for Error-tolerant Searching of Genomic Databases. *Proceedings of the 44th ASMS Conference on Mass Spectrometry and Allied Top.ics*, Portland, OR, May 12–16, 1996.
- (11) Liebler, D. C.; Hansen, B. T.; Davey, S. W.; Tiscareno, L.; Mason, D. E. *Anal. Chem.* **2002**, *74*, 203–210.
- (12) Tabb, D. L.; Saraf, A.; Yates, J. R., III *Anal. Chem.* **2003**, *75*, 6415–6421.
- (13) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390–4399.
- (14) Pappin, D. J. C.; Rahman, D.; Hansen, H. F.; Bartlett-Jones, M.; Jeffery, W.; Bleasby, A. J. *Mass Spectrom. Biol. Sci.* **1996**, 135–150.
- (15) Johnson, R. S. *How to sequence tryptic peptides using low energy CID data*. <http://www.abrf.org/ResearchGroups/MassSpectrometry/EPosters/ms97quiz/SequencingTutorial.html>.
- (16) McCormack, A. L.; Eng, J. K.; Yates, J. R., III *Methods Companion Methods Enzymol.* **1994**, *6*, 284–303.
- (17) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337–2342.
- (18) Scigelova, M.; Maroto, F.; Dufresne, C.; Vazquez, J. High-Throughput De Novo Sequencing. *14th Meeting Methods of Protein Structure Analysis*, Valencia, Spain, September 8–12, 2002.
- (19) Langridge, J. I.; Millar, A.; Young, P.; O'Malley, R.; Swainston, N.; Skilling, J.; Hoyes, J.; Richardson, K. A Fully Automated Hierarchical Software Strategy for De Novo Sequencing of Whole Q-ToF Electrospray LC-MS/MS Datasets. *Proceedings of the 50th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, Florida, June 2–6, 2002.

- (20) Fernandez-de-Cossio, J.; Gonzalez, J.; Betancourt, L.; Besada, V.; Padron, G.; Shimonishi, Y.; Takao, T. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 1867–1878.
- (21) Taylor, J. A.; Johnson, R. S. *Anal. Chem.* **2001**, *73*, 2594–2604.
- (22) Uttenweiler-Joseph, S.; Neubauer, G.; Christoforidis, S.; Zerial, M.; Wilm, M. *Proteomics* **2001**, *1*, 668–682.
- (23) Lu, B.; Chen, T. *J. Comput. Biol.* **2003**, *10*, 1–12.
- (24) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
- (25) Pearson, W. R.; Lipman, D. J. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444–2448.
- (26) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917–1926.
- (27) Huang, L.; Jacob, R. J.; Pegg, S. C.; Baldwin, M. A.; Wang, C. C.; Burlingame, A. L.; Babbitt, P. C. *J. Biol. Chem.* **2001**, *276*, 28327–28339.
- (28) Mackey, A. J.; Haystead, T. A. J.; Pearson, W. R. *Mol. Cell. Proteomics* **2002**, *1*, 139–147.
- (29) Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067–1075.

poor quality spectra. Third, these alignment programs cannot easily find posttranslational modifications, nor is it possible to search for particular modifications of interest to the researcher. Finally, significant manual interpretation of BLAST²⁴ and FASTA²⁵ results is often required to group peptide hits into likely protein identifications, rendering these programs difficult to use in high-throughput environments.

In this paper, we describe a new mass-based alignment algorithm that matches *de novo* sequences to homologous proteins and overcomes many of the limitations of other sequence homology search algorithms. This algorithm was designed to align *de novo* sequences from all MS/MS spectra for a given experiment to database protein sequences, even in situations when *de novo* sequencing algorithms cannot account for an entire peptide sequence. The implementation of this algorithm, a program named OpenSea, can rapidly identify proteins from complex mixtures of peptides using *de novo* sequences derived via high mass accuracy tandem mass spectrometry. The performance of OpenSea was tested against other common database-searching and *de novo* sequence alignment programs using three common types of data.

EXPERIMENTAL SECTION

Mass-Based Alignment Algorithm. The search algorithm used by OpenSea was designed to align ambiguous MS/MS *de novo* sequences to protein database sequences. *De novo* sequencing algorithms generally interpret MS/MS spectra as accurate sequences of three to five amino acids that are linked by unresolved mass regions. To take advantage of short but accurate regions, OpenSea first identifies a list of “tags” in a *de novo* sequence that are all possible combinations of three amino acids not broken by ambiguous mass regions. OpenSea then identifies the tags that are common to both the *de novo* sequence and a given database sequence via a series of string searches where isobaric single amino acids (I/L and K/Q) are replaced with a representative character, similar to the sequence tag method.¹³

Candidate alignments found by the tag search are then subjected to mass-based alignment (Figure 1a). For each candidate alignment, amino acids encompassing the short tag match in both the *de novo* and database sequences are converted into their corresponding monoisotopic masses. A series of consecutive local alignments on either side of the tag match are made to form a complete alignment. For each local alignment, all possible combinations of the next three masses in each sequence are compared sequentially with a “breadth-first search” algorithm, as shown in Figure 1b. Initially, OpenSea compares the masses of each of the next residues in the sequences within a fixed mass tolerance. If the masses are unequal, the sequences are compared one “level” deeper, where the mass of one database residue is compared to the mass of two query residues, followed by two database residues versus one query residue, and finally, two database residues versus two query residues. The breadth-first search continues through three levels deep until it finds a mass match. For example, when aligning the isobaric residue combinations of threonine–leucine and valine–aspartic acid, first the mass of Thr (101.0 amu) is compared to the mass of Val (99.1 amu), then Thr (101.0 amu) to sum of Val + Asp (214.1 amu), and finally the sum of Thr + Leu (214.1 amu) to the sum of Val + Asp (214.1 amu), representing a mass match. The comparison of the mass of Thr + Leu (214.1 amu) to the mass of Val (99.1 amu) does not

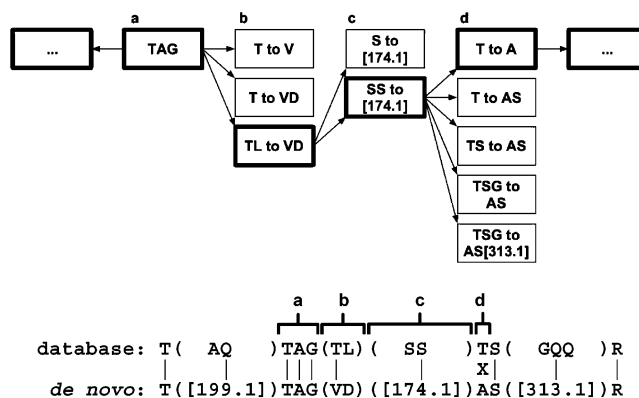


Figure 1. Mass-based *de novo* sequence alignment scheme. Here, the *de novo* sequence T[199.1]TAGVD[174.1]AS[313.1]R is aligned to a database peptide sequence TAQTAGTLSSTSGQQR using the “breadth-first search” method using a depth of three amino acids. A tag match (a) is used to initiate a mass-based alignment where the boxes in each column (b–d) represent the mass-based comparisons that must be computed to make a single local alignment. Bold bordered boxes signify the path through the search space taken by the mass-based alignment algorithm to produce a complete alignment. Accurate mass-based local alignments are signified by “|”, whereas mass mismatches are signified by “X”. OpenSea can group residues and masses in the *de novo* sequence (indicated in parentheses) if the mass of the group matches the mass of a corresponding sequence in the database. See text for further discussion.

need to be considered, because it has already been established by the Thr to Val comparison that Thr by itself weighs more than Val.

Figure 1c illustrates that masses or groups of amino acids that were unresolved in the *de novo* sequence are treated as if they were single residues that commonly align to two or more residues in the database sequence. If no mass match can be found by searching through three levels, as in Figure 1d, an amino acid substitution is assumed to have occurred. When a mass match is made or a substitution is assumed, the breadth-first search is stopped and a new local alignment is initiated starting from the next amino acid in each sequence. OpenSea continues making local alignments until the entire *de novo* sequence is accounted for. However, only one consecutive substitution is allowed, and the alignment process is terminated if more consecutive substitutions are required to make a match.

OpenSea can be configured to search for residue-specific variable modifications by assigning both the modified and unmodified masses to that residue. Variable N- and C-peptide termini modifications are accounted for in a similar way. Special database amino acid characters, such as B (either asparagine or aspartic acid), Z (either glutamine or glutamic acid), and X (any amino acid) are also implemented, for instance, by assigning the mass of both asparagine and aspartic acid to B. Unknown posttranslational protein modifications can be deduced from the shifted masses of specific amino acids, as well as the N- and C-peptide termini.

This approach can find short, isobaric equivalences of an arbitrary residue length, in this case, three consecutive residues or masses, within a given mass tolerance. Although the program execution time grows when more levels are searched, some algorithmic and heuristic-based optimizations have been used to reduce the search time. On average, it takes 9 s to search one *de*

novo sequence against the 127 873 protein sequences contained in the SwissProt database³⁰ (release 41.11) on a single Intel Pentium 4 2.0 GHz processor.

Scoring of Alignments and Resulting Protein Identifications. Each local alignment is scored separately, and the scores are summed to create a score for the overall peptide alignment. If a mass match is made in a local alignment, the local alignment score is the average value of the Blosum-90 substitution matrix³¹ identities for the database residues in that local alignment. If an amino acid substitution is made, the local alignment score is the matrix substitution score (S) between the database residue (i) and the de novo sequence residue (j).

$$\text{mass match} = \frac{\sum_{i=\text{database residues}}^n S_{ii}}{n}, \quad \text{substitution} = S_{ij} \quad (1)$$

If i contains a residue-specific variable modification, then S_{ii} for that residue is the average identity value (AIV) for the matrix. Similarly, if j is a mass, then S_{ij} for that mass is the average nonidentity value (ANV). Gapped matches, which are only allowed at the beginning and end of the database sequence, are scored as substitutions.

Local alignment mass matches are broken into three categories: one-to-one, one-to-many or many-to-one, and many-to-many matches, which refer to the number of amino acids in the database and de novo sequences, respectively. Local alignment substitutions are also broken into two categories: common substitutions (with score matrix scores >0) and uncommon substitutions (with score ≤ 0). The peptide alignment score is a linear combination of the summed local alignment scores from these groups

$$\begin{aligned} \text{alignment score} = & \alpha \left(\sum_{\text{matches}}^{1\text{-to-}1} \right) + \beta \left(\sum_{\text{matches}}^{1\text{-to-}m} \right) + \\ & \chi \left(\sum_{\text{matches}}^{m\text{-to-}m} \right) + \delta \left(\sum_{\text{substitutions}}^{\text{common}} \right) - \\ & \epsilon \left(\sum_{\text{substitutions}}^{\text{uncommon}} \right) - \phi \left(\sum_{\text{matches}}^{\text{gapped}} \right) \quad (2) \end{aligned}$$

where α has been assigned to 1.2, β to 1.1, χ to 0.9, δ to 1.0, ϵ to 5.0, and ϕ to 5.0. These values were empirically derived by analyzing MS/MS spectra derived from human amniotic fluid proteins. In the future, these weights can be statistically tuned for greater resolving power. For reference, the first four terms are always positive, while the last two terms are always negative.

As with CIDentify,²⁹ information about the enzymatic digestion is used to modify alignment scores. With trypsin, for example, the alignment score is augmented by $3.0 \times \text{AIV}$ for each terminus of the candidate peptide that matches a tryptic cleavage site (at lysine or arginine). If the candidate peptide indicates a nontryptic cleavage, the alignment score is decreased by $1.5 \times \text{ANV}$ for each unmatched terminus. Similarly, the score is decreased by ANV for each lysine or arginine present inside the matched database sequence, representing missed cleavage sites. Other enzymes can be considered in a similar fashion.

Peptide matches with alignment scores over 85 are accepted as correct identifications. Example peptide matches with their

corresponding alignments and alignment scores can be found in a supplementary file on the Web.³² Peptides with long sequences typically have larger scores; however, due to the requirements placed on the actual generation of the alignments, long sequences are generally more difficult to match, justifying their higher score. We've found that factoring the peptide length into the scoring function does not significantly improve the separation of correct from incorrect matches.

OpenSea contains an automatic results compiler that assists in protein identification. The results compiler is similar to ProteinProphet,³³ another algorithm developed for database-searching programs that detects proteins using "Occam's Razor" to combine complex peptide identifications into protein hits. The Occam's Razor approach assumes that the simplest combination of proteins that explains the spectral data is the correct interpretation. To find the simplest explanation, OpenSea first identifies a list of spectra that can be uniquely assigned to a single protein. This is done by ranking each peptide with an alignment score above 85 by a "delta score", which is the difference between the scores of the first and second best alignments for that spectrum. The spectrum with the largest delta score is assigned to the protein corresponding to its best alignment. Two alignments for the same de novo sequence with a score difference of <20 are considered to match equally well. Therefore, all other spectra that match to the protein in question with a delta score of <20 are assigned to that protein. Of the remaining spectra, the spectrum with the next largest delta score is then considered and assigned to the protein it matches best. This process is repeated through all of the uncontested identifications. This way, peptides that match multiple proteins equally well are assigned to the protein with the strongest single peptide evidence (greatest delta score). Two proteins that match the same peptides with the same scores are considered "degenerate" and are grouped together.

OpenSea scores each protein as the sum of the scores of the alignments that match independent regions of that protein. De novo sequences from MS/MS spectra that match the same region of a protein but have different precursor masses (often representing modified peptides) or have different charges are also considered independent. Otherwise, if two de novo sequences align to the same region of a single protein, only 10% of the alignment score for the second sequence is added to the protein score, as these additional identifications often do not provide any new evidence for the protein.

Once the proteins have been identified from the spectra, the remaining unmatched de novo sequences (with alignment scores below 85) are then realigned to only the identified proteins. The alignments are made using different parameters tuned specifically to find peptides that were poorly sequenced. In particular, five mass levels are searched to identify isobaric equivalent regions for each local alignment, while the length of tags required to initiate an alignment is decreased to two. Furthermore, two consecutive substitutions are allowed. Realignment matches with alignment scores above 85 are accepted, and matches with scores between 85 and 60 are flagged for manual interpretation or verification by a cross-correlation method (such as SEQUEST).

(32) Additional results and analysis can be found in the supplementary file on the Web at <http://medir.ohsu.edu/~geneview/publication/opensea/>.

(33) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 4646–4658.

(30) Balroch, A.; Boechmann, B. *Nucleic Acids Res.* **1991**, *19*, 2247–2249.

(31) Henikoff, S.; Henikoff, J. G. *Proc. Natl. Acad. Sci.* **1992**, *89*, 10915–10919.

This approach is similar to the retroactive search done by ProteinLynx via the AutoMod subroutine.

Sample Preparation and LC/MS/MS Spectra Acquisition.

Three types of samples were used to test OpenSea. The known protein control mixture was obtained by combining 10 purified proteins of varying molecular weight and physiochemical properties. *Bos taurus* insulin, ubiquitin, cytochrome *c*, superoxide dismutase, β -lactoglobulin A, serum albumin, and immunoglobulin G, as well as *Equus caballus* myoglobin, *Armoracia rusticana* peroxidase, and *Gallus gallus* conalbumin were obtained from CIPHERGEN (Fremont, CA). The proteins were combined with urea, reduced with dithiothreitol, and alkylated with iodoacetamide. The mixture was then digested overnight at 37 °C with 1 μ g of modified trypsin (Promega)/50 μ g protein. The resulting peptide mixture was dissolved in 5% formic acid to 2 pmol of total protein/ μ L of solution. Twelve 1-pmol samples, 22 2-pmol samples, and a single 4-pmol sample were analyzed with MS/MS.

Homo sapiens and *Macaca mulatta* amniotic fluid samples containing unknown, sequence-modified proteins were obtained from the Oregon Health & Sciences University with Institutional Review Board approval. Proteins were separated by one-dimensional gel electrophoresis and were visualized by Coomassie staining. Bands from each sample were excised and in-gel digested with trypsin, and the peptides were extracted from the gel matrix, filtered (0.22 μ m), evaporated, and dissolved in 5% formic acid. One high-molecular-weight band from each sample was chosen for MS/MS analysis.

A lens sample from a 55-year-old *H. sapiens* containing posttranslationally modified proteins was also obtained from the Oregon Lyons Eye Bank with Institutional Review Board approval from the Oregon Health & Sciences University. A 10- μ g portion of total protein was reduced, alkylated, and trypsin-digested. The resulting peptides were diluted with 5% formic acid, and 10 μ g of total protein was analyzed by MS/MS.

All MS/MS spectra were acquired with a Micromass Q-TOF-2 (Milford, MA) quadrupole/time-of-flight hybrid mass spectrometer with an online capillary LC (Waters, Milford, MA). Samples were desalted with an in-line C18 trap cartridge (LC Packings, San Francisco, CA) and separated on a 75 μ m \times 15 cm C18 IntegraFrit column (Waters, Milford, MA). Peptides were injected into the online mass spectrometer through a nanospray source.

De novo Sequencing and Database Searching. All MS/MS spectra acquired were de novo sequenced. Peaks 1.3^{17,34} (Bioinformatics Solutions Inc., Waterloo, ON Canada) and Lutefisk 1900 1.3.2^{21,35} De novo sequencing programs were used to test the performance of OpenSea. Both programs were configured to assume that all cysteines were alkylated and that all peptides were tryptically digested. Unlike Lutefisk, Peaks reports full amino acid sequences without unknown mass regions, but does assign each amino acid in the sequence a confidence score. Sequence regions where amino acids had confidence scores below 50% were replaced by the combined mass of those amino acids. Lutefisk reports as many as five de novo sequences for each spectrum. All of these sequences were used to produce a match. Only the

top scoring sequence reported by Peaks was used, as generally all of the top five Peaks sequences could be represented by the 50% consensus sequence.

Two database-searching programs, TurboSEQUEST 2.0² (Thermo Finnigan, San Jose, CA) and ProteinLynx 2.0⁵ (Waters, Milford, MA), and one de novo sequence alignment program, CIDentify 1.0.8,³⁶ were used to benchmark OpenSea. All samples of the control mixture were searched against the SwissProt database³⁰ (release 41.11) that was modified to include sequences for the control proteins that were selected from the nonredundant reference protein database³⁷ (PIR-NREF, release 1.25). The human and rhesus monkey amniotic fluid samples, as well as the human lens sample, were searched against the SwissProt database selected for human proteins.

SEQUEST and ProteinLynx were configured to identify tryptic peptides and search for variably alkylated cysteines. DTASelect³⁸ was used to identify protein matches from SEQUEST results. Protein matches were accepted with multiple peptide hits having cross-correlation scores (Xcorr) of greater than 1.8, 2.5, and 3.5 for singly, doubly, and triply charged peptides, respectively. In ProteinLynx, protein hits having multiple positive peptide match scores were accepted, and the AutoMod subroutine of ProteinLynx⁵ was used on all samples to find modified peptides belonging to the identified proteins.

CIDentify assumed fixed alkylations and results with *E* values $<10^{-4}$ were accepted. A version of CIDentifyRC³⁹ that was modified to process over 100 de novo sequences at a time was used to identify successfully matched proteins. OpenSea was configured to search for the variable alkylation of cysteines, and protein hits with multiple peptide matches having alignment scores of >85.0 were accepted. Both CIDentify and OpenSea were configured to preferentially identify tryptic peptides. In all searches, matches to keratins and trypsin were ignored as contaminants.

RESULTS AND DISCUSSION

Identification of the Control Mixture Proteins. A mixture of 10 tryptically digested proteins was used to evaluate OpenSea. A total of 10 685 tandem mass spectra from 35 LC/MS/MS runs of the control mixture were processed with Peaks and then OpenSea. In Figure 2, a de novo sequence generated by Peaks from one MS/MS spectrum is shown to align to bovine serum albumin with significant homology. Peaks accurately identified a three-amino-acid sequence tag, ADE. From that tag, OpenSea was able to interpret two incorrect regions in the de novo sequence as isobaric equivalents of regions in the protein database sequence, as indicated in parentheses. Variations found by OpenSea represent localized mass discrepancies, which imply the presence of unanticipated modifications or substitutions. In this case, a variation from threonine in the database sequence (101.0 amu) to an unresolved section of the de novo sequence (144.1 amu)

(36) Current versions of CIDentify are available for download at <ftp://ftp.virginia.edu/fasta/CIDentify/>.

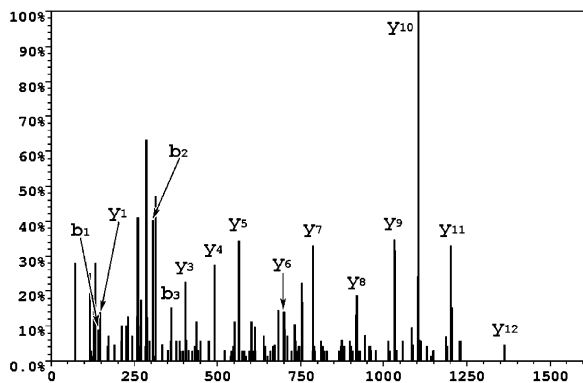
(37) Wu, C. H.; Huang, H.; Arminski, L.; Castro-Alvear, J.; Chen, Y.; Hu, Z.; Ledley, R. S.; Lewis, K. C.; Mewes, H.; Orcutt, B. C.; Suzek, B. E.; Tsugita, A.; Vinayaka, C. R.; Yeh, L. L.; Zhang, J.; Barker, W. C. *Nucleic Acids Res.* **2002**, *30*, 35–37.

(38) Tabb, D. L.; McDonald, W. H.; Yates, J. R., III *J. Proteome Res.* **2002**, *1*, 21–26.

(39) Johnson, R.; Taylor, J. In *Methods in Molecular Biology: Mass Spectrometry of Proteins and Peptides*; Chapman, J., Ed.; Humana Press: Totowa, NJ, 2000; Vol. 146, pp 41–62.

(34) Ma, B.; Zhang, K.; Liang, C. An Effective Algorithm for the Peptide De Novo Sequencing from MS/MS Spectrum. The 14th Symposium on Combinatorial Pattern Matching, March 2003, 266–278.

(35) Current versions of Lutefisk are available for download at <http://www.hairy-fatguy.com/Lutefisk/>.



De Novo Sequence:

[144.1] SATADESHAGM [158.1] K

OpenSea Alignment:

ALBU_BOVIN Serum Albumin Precursor
 a-**C** FAK (T) (CcV) ADESHAG (CcE) KSLH
 X
 b-**C** ([144.1]) (SAT) ADESHAG (M [158.1]) K
 c d

Figure 2. An example mass-based alignment of a peptide in the control mixture to bovine serum albumin. A MS/MS spectra derived from a tryptic peptide present in the control mixture was de novo-sequenced by Peaks. An OpenSea alignment showing how a bovine serum albumin sequence from the SwissProt database (a) was matched to the de novo sequence (b) is also illustrated. Isobaric equivalent regions of the de novo and database sequences are indicated in parentheses. The localized mass discrepancy (144.1–101.1) of 43.0 amu (c) suggests that the N terminus is carbamylated. Lower case letters in the database sequence represent user-defined protein modifications inserted into the sequence by OpenSea, such as Cc, which represents an alkylated cysteine (d). Three residues on either side of the database peptide sequence are reported to provide context within the protein.

was identified. The mass shift of 43.0 amu suggested that the peptide was carbamylated at the N terminus. This peptide was one of eight from a single LC/MS/MS run that were found to contain this mass shift, which was most likely the result of using urea as a protein denaturant.⁴⁰

One major requirement for high-throughput MS/MS analysis is an accurate peptide scoring system that can reliably distinguish between correct and incorrect peptide assignments. The accuracy of the default alignment scoring system was estimated by searching de novo sequences generated from all 35 LC/MS/MS runs of the control mixture against the SwissProt protein database (release 41.11), which contained 127 863 proteins from various species. Peptide assignments to the 10 control proteins were considered unlikely to have occurred by chance, and were therefore assumed to be correct. Conversely, assignments to any other protein were considered incorrect. As shown in Figure 3a, the alignment scoring system used by OpenSea clearly separates correct from incorrect peptide assignments. OpenSea's default alignment score cutoff of 85 identified 94% of the correct assignments (sensitivity) and eliminated 97% of the incorrect assignments (specificity). For comparison, the sensitivity of the Xcorr score used by SEQUEST was 77%, while the specificity was 85%

(40) Stark, G. R.; Stein, W. H.; Moore, S. *J. Biol. Chem.* **1960**, *235*, 3177–3181.

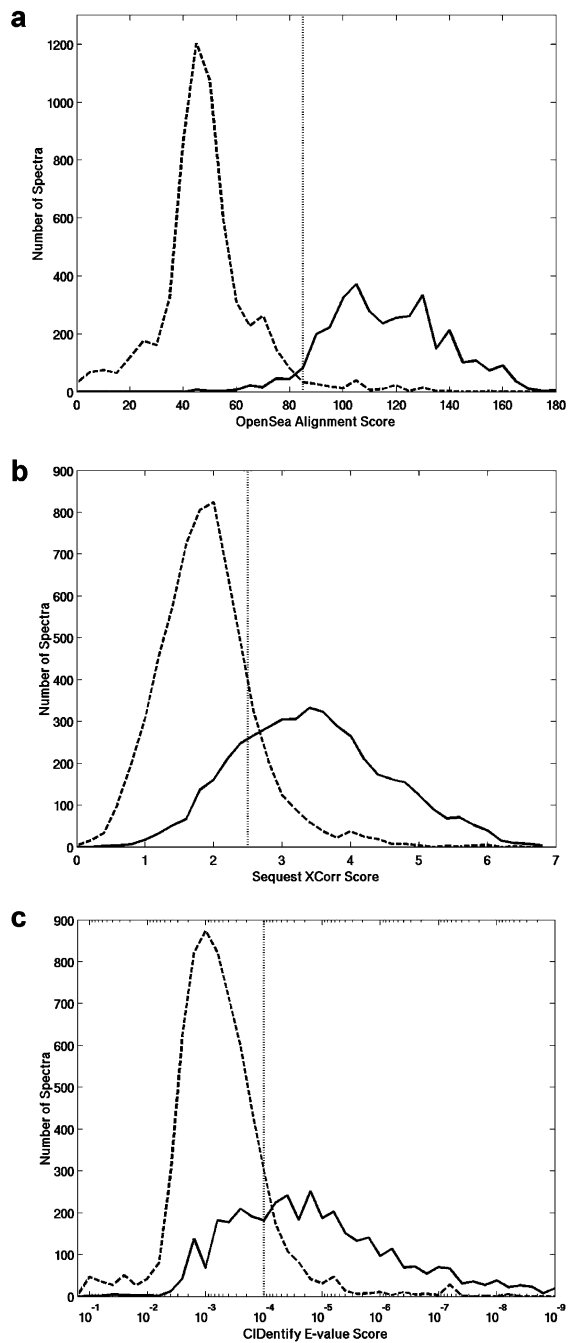


Figure 3. Score distributions for OpenSea/Peaks, Sequest, and CIDentify/Peaks when analyzing 10611 MS/MS spectra from peptides in the control protein mixture. Correct spectra matches are shown as a solid line, whereas incorrect matches are shown as a dashed line. (a) The OpenSea alignment score distributions were obtained by placing matches in bins of 5.0 alignment score points. The default alignment score cutoff of 85 is shown as a dotted vertical line. (b) Sequest Xcorr scores have been normalized across parent ion charge to reflect the differences in score thresholds that were used (+0.7 for singly and -1.0 for triply charged peptides). The distributions were obtained by placing matches in 0.2 Xcorr score bins, and the normalized Xcorr score cutoff of 2.5 is shown as a dotted vertical line. (c) CIDentify *E*-value score distributions were obtained by being placed in match bins of $10^{0.2}$ *E*-value scores, and the score distributions are shown in reverse log scale. 10^{-4} was used as the *E*-value score cutoff and is drawn as a dotted vertical line. The improved separation of the OpenSea alignment score over other identification techniques allows for the accurate identification of low-scoring, but correct peptide matches.

Table 1. The Number of MS/MS Spectra Identified as Control Mixture Proteins

protein name ^a	OpenSea/ Peaks ^b	OpenSea/ Lutefisk ^b	CIDentify/ Peaks ^b	CIDentify/ Lutefisk ^b	SEQUEST ^b	ProteinLynx/ AutoMod ^b
bovine serum albumin	48	14	26	11	40	29
chicken conalbumin	27	8	22	4	29	17
bovine immunoglobulin G	13	0	7	2	11	14
equine myoglobin	9	3	4	2	6	8
bovine β -lactoglobulin	8	2	6	2	9	4
bovine superoxide dismutase	5	2	5	2	9	4
bovine cytochrome <i>c</i>	5	0	5	0	4	2
bovine ubiquitin	4	0	2	0	3	4
horseradish peroxidase	3	0	2	0	6	2
bovine insulin	0	0	0	0	0	0
total	122	29	79	23	117	84

^a The number of spectra correctly assigned to proteins in the control mixture by various programs. ^b OpenSea, and CIDentify are de novo sequence alignment programs, whereas SEQUEST and ProteinLynx are common database-searching software packages. Peaks and Lutefisk were used to generate de novo sequences. ProteinLynx was used with a workflow including the AutoMod posttranslational peptide modification subroutine. At least two spectra hits for each protein were required for identification.

using minimum Xcorr values of 1.8, 2.5, and 3.5 for peptides of +1, +2, and +3 charge, respectively (Figure 3b). Similarly, the sensitivity of the CIDentify *E* value score was 70%, and the specificity was 89% with a minimum score cutoff of 10^{-4} (Figure 3c). Statistical analysis of the OpenSea alignment score distributions can be found in the supplementary file on the Web.³²

A second requirement for high-throughput MS/MS analysis is accurate and easy to interpret protein identifications from peptide matches. The Occam's Razor approach used by OpenSea to identifying protein candidates from the most unambiguous spectral evidence has many benefits, one of which is that a single spectrum is assumed to match only one protein. In the case in which the spectrum matches multiple proteins equally, it is assigned to the protein with the greatest evidence for existing in the sample. This is critical to high-throughput analysis because it removes degenerate peptide hits in the case of homologous proteins, which often confound results in large studies. Another benefit is that protein evidence is generated on the basis of how exclusively a single MS/MS spectra can be assigned to that protein on the basis of the delta score, and not on the overall score for that protein. For example, if a single spectrum can be assigned with high confidence to a protein with low overall coverage, the low-coverage protein will be reported. This allows low-abundance proteins with poor coverage to be found, even if proteins with higher coverage dwarf them. Alternatively, if homologous proteins are expected, OpenSea can be configured to report degenerate peptide matches in proteins with amino acid sequence similarity.

Comparison of OpenSea to Additional MS/MS Protein Identification Software. One LC/MS/MS run of a 2-pmol control mixture sample was examined in detail to benchmark the number of spectra accurately identified by OpenSea, as compared to common database-searching programs. Protein identifications of 328 spectra were made by two commonly used database-searching programs, SEQUEST and ProteinLynx, and by two de novo sequence alignment programs, OpenSea and CIDentify. Peaks and Lutefisk were used to provide de novo sequences for both OpenSea and CIDentify. The number of visually verified spectra matching each control protein was tabulated for all of the programs (or combination of programs), and shown in Table 1. Sequences derived by ProteinLynx automated de novo sequenc-

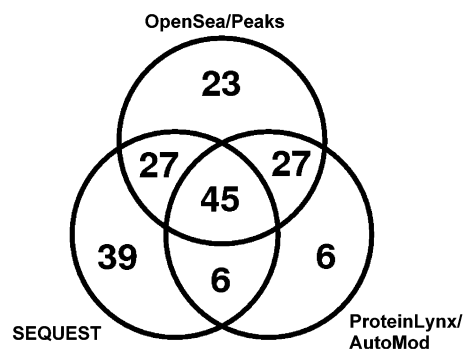


Figure 4. A comparison of MS/MS spectra identifications in the control sample between OpenSea/Peaks, SEQUEST, and ProteinLynx/AutoMod. It is clear that although there is significant overlap among the three programs, no single approach can identify every spectrum.

ing¹⁹ were also tested, but both OpenSea and CIDentify generally produced fewer identifications with these sequences than with sequences generated by either Peaks or Lutefisk (data not shown). OpenSea/Peaks and CIDentify/Peaks were the only analysis methods that found one of the two tryptic peptides from bovine insulin that were within the mass range of the experiment (not shown in table). However, the match would be difficult to verify because only one peptide from insulin was found.

OpenSea, using de novo sequences derived by Peaks, identified 4% more MS/MS spectra than SEQUEST and 45% more MS/MS spectra than the ProteinLynx search engine using the AutoMod subroutine. A breakdown of the identifications made by OpenSea/Peaks, SEQUEST, and ProteinLynx/AutoMod is shown in Figure 4. OpenSea, like CIDentify, identified a comparably low number of MS/MS spectra when using Lutefisk-derived de novo sequences. Although both programs identified significantly more peptides when using Peaks de novo sequences versus Lutefisk sequences, OpenSea identified 54% more MS/MS spectra than CIDentify. Only three matches of the identifications made by CIDentify/Peaks were not found by OpenSea/Peaks.

OpenSea's increased accuracy in deciphering de novo sequences, as compared to CIDentify, could be due to three major factors. First, OpenSea does not limit the length of its local alignments to single or pairs of residues, and the further interpretation often results in higher alignment scores for correct

Table 2. The Number of MS/MS Spectra from Human (A) and Rhesus Monkey (B) Amniotic Fluid Samples that Were Assigned to Adult Human Proteins

protein name	OpenSea/ Peaks ^a	CIDentify/ Peaks ^a	SEQUEST ^a	Protein- Lynx/ AutoMod ^a	confirmed/unconfirmed amino acid variants found by OpenSea/Peaks ^b
A					
lactotransferrin	22	13	5	18	12/1
glia-derived nexin	11	5	5	10	1/1
serotransferrin	6	4	2	3	2/0
serum albumin	4	0	5	6	0/1
α-1-acid glycoprotein	3	2	2	0	1/0
moesin	3	0	2	0	0/0
myeloperoxidase	3	0	2	0	0/0
histidine-rich glycoprotein	2	0	0	0	1/0
α-1 antichymotrypsin	2	0	2	3	0/0
α-1 antitrypsin	2	0	2	2	1/0
total	58	24	27	42	18/3
B					
lactotransferrin	25	13	5	16	12/1
glia-derived nexin	8	5	5	3	1/1
collagen α 2(I) chain	8	3	0	0	17/2
α-1 antitrypsin	4	2	2	2	2/2
serum albumin	4	0	3	2	0/0
gelsolin	3	0	0	2	0/0
92-kDa type IV collagenase	2	0	2	0	0/0
α-1 antichymotrypsin	0	2	0	0	0/0
total:	54	25	17	25	32/6

^a OpenSea/Peaks, CIDentify/Peaks, SEQUEST, and ProteinLynx were used to identify homologous adult human proteins when searching with proteins in human and rhesus monkey amniotic fluid samples. ProteinLynx was used with a workflow including the AutoMod subroutine to identify sequence variations. At least two spectra hits for each protein were required for identification with every protein identification program. Protein identifications to trypsin and human keratins were omitted. ^b The number of sites of amino acid sequence variation found between adult and amniotic fluid proteins by OpenSea/Peaks and confirmed by SEQUEST searching a modified sequence database, as well as the number of unconfirmed sequence variations reported by OpenSea/Peaks.

matches. Second, all alignments must pass stringent, empirically developed criteria requiring that the entire de novo sequence be accounted for, allow for only one consecutive sequence modification, and require that each alignment contain at least one accurately matching sequence tag. Third, the OpenSea scoring function separates correct from incorrect matches more reliably than CIDentify, which allows OpenSea to accurately identify lower scoring peptides without introducing a significant number of false positives. The two programs have very distinct approaches to sequence alignment: CIDentify assumes that de novo sequences are generally correct and tries to match them against protein sequences in databases while presuming that sequence variations are often real. OpenSea, on the other hand, assumes that de novo sequences must be verified, and uses protein databases to correct as much of the sequence variation as possible. As a result, the CIDentify algorithm tends to identify distant protein homologies better than OpenSea, whereas OpenSea tends to make more complete and more robust interpretations of the actual de novo sequences.

Identification of Unknown, Homologous Proteins. OpenSea can identify proteins that have not been completely sequenced, provided that proteins with close sequence homology are present in the searched databases. Human amniotic fluid was used to represent a mixture of unknown proteins. The amniotic fluid contains fetal proteins that are known to have amino acid variances with their adult homologues. For example, the γ chain of fetal hemoglobin contains 39 sites of amino acid sequence variation from the adult β chain.⁴¹

A LC/MS/MS run of *H. sapiens* amniotic fluid proteins from a high-molecular-weight 1D gel band, generating 416 MS/MS spectra, was analyzed. The spectra were sequenced using Peaks, and the resulting sequences were aligned and identified with OpenSea by searching against human proteins in the SwissProt database (9436 proteins). The same spectra were also processed with CIDentify/Peaks, SEQUEST, and ProteinLynx/AutoMod. Protein identifications for each spectrum were manually validated and reported in Table 2A. Sequence variations identified by OpenSea/Peaks were confirmed in 18 of the 21 cases by modifying the human protein database to include those sequence variations and searching the MS/MS spectra against the new database with SEQUEST. For example, OpenSea/Peaks identified 12 sites of single amino acid variance in amniotic fluid lactotransferrin relative to the human SwissProt sequence (accession number P02788) obtained from nonamniotic fluid samples. ProteinLynx's AutoMod subroutine is an effective modification and sequence variance identification tool and found many of the sequence variant peptides in lactotransferrin that OpenSea/Peaks reported. However, AutoMod cannot find proteins that have not been identified in the initial database search. OpenSea, using Peaks de novo sequences, had a significantly higher peptide and protein identification rate than ProteinLynx/AutoMod. As with the control sample, CIDentify/Peaks found a subset of the peptides identified by OpenSea/Peaks along with two original peptide matches. SEQUEST, as expected, could only find a few unmodified peptides from these proteins (Table 2A).

To further this argument, a corresponding LC/MS/MS run containing 411 MS/MS spectra of *M. mulatta* amniotic fluid

(41) Lorkin, P. A. *J. Med. Genet.* **1973**, *10*, 50–64.

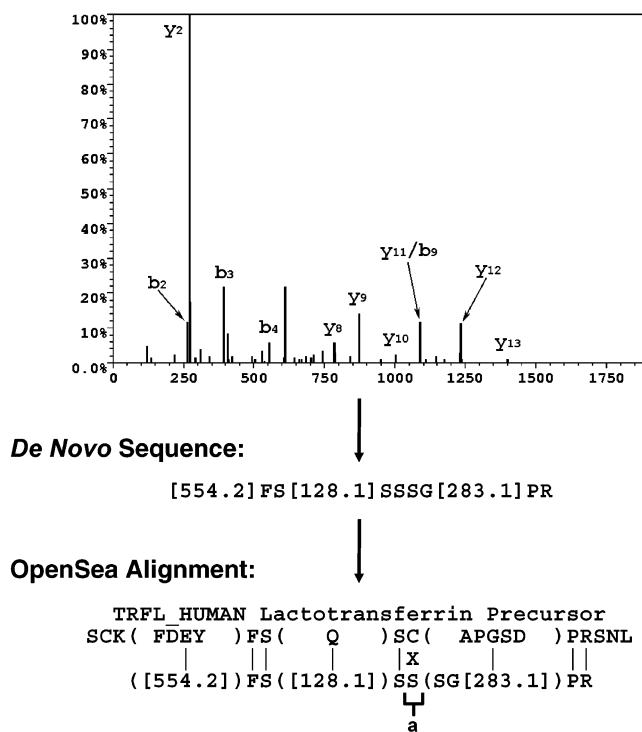


Figure 5. A homologous match of a human amniotic fluid peptide to the SwissProt database sequence of adult human lactotransferrin protein. A poor quality MS/MS spectrum derived from a tryptic peptide from the human amniotic fluid was de novo-sequenced by Peaks and database-searched by OpenSea. 50% of the residues in the peptide were ambiguously or incorrectly sequenced. OpenSea is able to correctly interpret all of the ambiguous regions and determine a single amino acid difference from cysteine in the database sequence to serine in the amniotic fluid protein (a).

proteins was analyzed in a similar fashion (Table 2B). Although very few rhesus monkey proteins have known sequences, the few known proteins have high sequence homology to their human counterparts. As with the human amniotic fluid sample, sequence variant amino acid sites identified by OpenSea/Peaks were confirmed with SEQUEST. OpenSea/Peaks routinely identified peptides with sequence variation from their human analogues and again out-performed CIDentify/Peaks, SEQUEST, and ProteinLynx/AutoMod at peptide and protein identification. For example, only OpenSea/Peaks and CIDentify/Peaks could identify collagen α 2(I) chain protein, because seven of the eight peptides identified by OpenSea had at least one single amino acid variation.

Many other sequence search engines^{26–29} can identify sequence variations between de novo sequenced peptides and their corresponding sequences in protein databases. One major difficulty is identifying actual sequence variation in the presence of de novo sequencing errors. Because OpenSea's mass-based search algorithm can identify isobaric equivalences of an arbitrary length, it can account for many of the common errors found in sequences generated by Peaks. For example, a poor-quality MS/MS spectrum of a human amniotic fluid peptide was de novo-sequenced, and while the resulting sequence contained many ambiguous regions, OpenSea could align it to the lactotransferrin protein (Figure 5). The OpenSea algorithm was able to assign every ambiguous amino acid region to the database sequence, regardless of length. With the unknown regions of the sequence accounted for, a single amino acid variation can be observed at

residue 513 in the SwissProt lactotransferrin precursor sequence. The human SwissProt database was modified to reflect this variation, and the spectrum was searched against this database with SEQUEST, which confirmed the match ($z = 2$, $Xcorr = 3.6$, $dCn = 0.37$). Additionally, OpenSea/Peaks assigned the single large peak at $272.2 m/z$ to a proline–arginine fragment representing a bond cleavage between aspartic acid and proline, which is expected to have enhanced cleavage over other residue pairs in the peptide.⁴² This enhanced cleavage helped support the peptide identification from an otherwise poor-quality spectrum.

Identification of Posttranslational Protein Modifications.

An alternative method of using OpenSea to identify unanticipated in vivo and in vitro protein modifications involves an iterative process in which mass differences between the de novo sequence and the database that are associated with particular protein modifications are fed back into OpenSea. The previously unmatched de novo sequences are then searched with OpenSea against the entire database to identify any other peptides that have the same modifications. This two-step process mines information from poor-quality de novo sequences or peptides with multiple modifications that could not otherwise be identified by mass shift alone.

A human lens sample from a 55-year-old male, containing proteins with known posttranslational modifications, was used to illustrate this method. Approximately 95% of the protein in the human lens consists of just 12 crystallins that do not turn over.⁴³ These crystallins undergo posttranslational modifications over time, and because of their long life spans, many tryptic peptides can accumulate two or more modifications per peptide. An initial OpenSea/Peaks search of the 305 LC/MS/MS spectra generated from this sample generated 85 matches, while identifying 16 peptides with mass variations consistent with either carbamylation, methylation of cysteine, acetylation, oxidation of methionine, or the loss of ammonia or water from a carboxylic acid containing amino acid. Once these identifications were confirmed, OpenSea was configured to specifically find other peptides with these modifications, and six new modification sites were found from 12 new MS/MS matches. All together, OpenSea found six different types of modifications, which are listed in Table 3, and many of the actual modification sites confirm previous reports. For comparison, the AutoMod feature of ProteinLynx identified three types of modifications.

Cysteines at residues 24 and 26 in γ -crystallin S,⁴⁴ as well as cysteine 82 in β -crystallin A3,⁴⁵ were confirmed as methylated in some peptides. Cysteine 185 in β -crystallin A3 was also methylated, and SEQUEST verified this previously unidentified methylation site ($z = 2$, $Xcorr$ 3.6, $dCn = 0.58$). Similarly, N-terminal acetylation of α -crystallin A and β -crystallin B2 were confirmed⁴⁶ and the first methionine in α -crystallin A was variably oxidized.⁴⁶ An asparagine

(42) Breci, L. A.; Tabb, D. L.; Yates, J. R., III; Wysocki, V. H. *Anal. Chem.* **2003**, *75*, 1963–1971.

(43) MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R. III *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7900–7905.

(44) Lapko, V. N.; Smith, D. L.; Smith, J. B. *Biochem.* **2002**, *41*, 14645–14651.

(45) Lapko, V. N.; Smith, D. L.; Smith, J. B. S-Methylation and glutathionylation of human lens beta crystallins. Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics, Montreal, Canada, June 8–12, 2003.

(46) Lampi, K. J.; Ma, Z.; Shih, M.; Shearer, T. R.; Smith, J. B.; Smith, D. L.; David, L. L. *J. Biol. Chem.* **1997**, *272*, 2268–2275.

Table 3. Modifications Identified in the Human Lens Crystallin

Modification ^a	Nominal Mass Shift ^b	OpenSea/Peaks Identified Sites ^c	ProteinLynx/AutoMod Identified Sites ^d	Example OpenSea Alignment ^e
N-Terminal Carbamylation	43	12	7	<pre> NYR(L)VVFELNFQGRRAE X ([156.1])VVFELNFQGR </pre>
Methylation of Cysteine	14	4	0	<pre> GRR(YD) (Cc)D(Cc)DCADFHTYLSRCNS XX ([278.1]) (Cc)D(Cc)TMADFHTYLSR </pre>
N-Terminal Acetylation	42	2	2	<pre> MDIAIHH(PW)IRRPF X: SSNLALHH(APD)LR </pre>
Formation of Pyroglutamic acid	-17/-18	2	0	<pre> VKVQDDFVEIHGKHNE :X EPDFVELHGK </pre>
Formation of Succinimide	-17	1	1	<pre> NYRLVVFELNF(Q)GRRRAE X LVVFELEPF([128.1])GR </pre>
N-Terminal Acetylation and Oxidation of Methionine	42 and 16	1	0	<pre> MD(V)TI(Q)HP(W)FKRTL X ([403.2])TL([128.1])HP([186.1])FK </pre>

^a Modifications identified in human lens crystallin proteins. ^b The mass differences between the MS/MS spectra and the database peptide. ^c The number of residue sites OpenSea identified with that modification. ^d The number of sites found by ProteinLynx using AutoMod. ^e Peaks de novo sequence (bottom)-aligned by OpenSea to the database sequence (top) with sequence mismatches ("X" and ":"), representing mass discrepancies. The mass discrepancies representing each modification are shown in bold.

in β -crystallin B1 had an apparent loss of ammonia to form succinimide, a likely intermediate in nonenzymatic deamidation.⁴⁷ An N-terminal glutamine in a peptide from α -crystallin A was identified as having lost ammonia, and an N-terminal glutamic acid in a peptide from α -crystallin B had similarly lost water. These residues have likely undergone cyclization with the amino terminus during digestion to form pyroglutamic acid.⁴⁸

All of the modifications were identified without any prior knowledge of the posttranslational modifications that are commonly found in lens proteins. An effort is currently being made to automate this search method to mine protein samples for unanticipated posttranslational modifications without the manual interpretation of OpenSea alignment results. More detailed information on the specific sequence variations that were found in the amniotic fluid samples, as well as information on the protein and peptide posttranslational modifications found in the lens sample, can be found in our supplementary file on the Web.³²

CONCLUSION AND PERSPECTIVES

Mass-based alignment of de novo sequences can accurately identify sequence variations and posttranslational protein modifications, thus allowing for these types of searches to succeed in a high-throughput environment. An effort has been made to allow the batch-scripting of OpenSea, including the ability to search any number of databases consecutively. XML result files facilitate automatically adding OpenSea alignments into relational databases for the cataloging of protein sequence variations and sites of posttranslational modifications. Although we have shown that OpenSea can already differentiate correct from incorrect hits in the control mixture with a 95% success rate using default

parameters, various intermediate score multipliers and score thresholds can be adjusted. This will allow for the future statistical tuning of OpenSea and potentially eliminate the need for the manual validation of protein identifications.

The major advantage presented here of the mass-based alignment algorithm used by OpenSea over the mass-based bootstrapping method used by CIDentify is in the allowed complexity of the scoring system. Because OpenSea uses the same approach to make every local alignment, and that that approach can be broken into subclasses of alignments, scored separately, and linearly combined to create an optimal score, OpenSea can more accurately separate correct identifications from incorrect. Although this is important, there are other significant aspects to using the mass-based alignment algorithm to make every local alignment. For instance, OpenSea can be used to align two de novo sequences from the same peptide to create more accurate consensus sequences, as well as to identify modifications in completely unknown proteins by using other de novo sequences as references. Essentially, this approach allows sequences of unknown proteins to be built from fragments of de novo sequences (including ambiguous mass regions) and those previously unsequenced proteins to be used for accurate peptide identification. Furthermore, protein sequences can be annotated with site-specific modifications, which will allow for the future utilization of known protein modifications already being cataloged in databases such as the Human Reference Protein Database.⁴⁹ Future publications will discuss the utility of this work.

Current de novo sequencing programs coupled with OpenSea can be operated as stand-alone protein identification packages or can be used in conjunction with database-searching programs for independent verification of protein identifications. In the control sample, the results from OpenSea/Peaks, SEQUEST, and Pro-

(47) Wright, H. T. *CRC Crit. Rev. Biochem.* **1991**, *26*, 1–52.

(48) Khandke, K. M.; Fairwell, T.; Chait, B. T.; Manjula, B. N. *Int. J. Pept. Protein Res.* **1989**, *34*, 118–123.

teinLynx/AutoMod could be combined to identify over 40% more MS/MS spectra than any single method alone while independently confirming over 60% of the total spectra identified. Although Peaks is tuned to only process data from high mass accuracy tandem mass spectrometers, it is possible to use other programs to generate de novo sequences. For example, an adjustable mass tolerance makes it possible for OpenSea to search protein databases using low mass accuracy de novo sequences generated by DenovoX¹⁸ (Thermo Finnigan, San Jose, CA) from ion-trap tandem mass spectra (data not shown).

A powerful new approach to de novo sequencing could be to implement a mass-based alignment algorithm, such as the one used by OpenSea, to automatically verify sequencing results against protein sequences in databases. In this approach, the mass-based alignment algorithm could help a de novo sequencing

program make choices between potential sequence candidates as well as to direct the de novo sequencing program in making more empirically driven decisions. This versatility suggests that mass-based alignment presented here can be used for a wide number of applications involving the identification of proteins.

OpenSea was written in Java, and it will run on any platform that can run the Java Runtime Environment (version 1.3). OpenSea has been tested on Windows 2000 and Linux platforms. OpenSea binaries can be obtained for independent analysis via a Material Transfer Agreement with the Oregon Health & Sciences University. DTA files from the control sample are also available for independent analysis in our supplementary file on the Web.³²

ACKNOWLEDGMENT

This work was supported by National Institute of Health Grants U19ES11384 and U24DK5870 to Srinivasa Nagalla. We thank Jean O'Malley and Tim Sheard for insightful conversations. We also acknowledge Melissa Standley for technical assistance, and Nan Jiang for programming assistance.

Received for review October 24, 2003. Accepted January 21, 2004.

AC035258X

(49) Peri, S.; Navarro, J. D.; Amanchy, R.; Kristiansen, T. Z.; Jonnalagadda, C. K.; Surendranath, V.; Niranjana, V.; Muthusamy, B.; Gandhi, T. K. B.; Gronborg, M.; Ibarrola, N.; Deshpande, N.; Shanker, K.; Shivashankar, H. N.; Prasad, R. B.; Ramya, M. A.; Chandrika, K. N.; Padma, N.; Harsha, H. C.; Yatish, A. J.; Kavitha, M. P.; Menezes, M.; Choudhury, D. R.; Suresh, S.; Ghosh, N.; Saravana, R.; Chandran, S.; Krishna, S.; Joy, M.; Anand, S. K.; Madavan, V.; Joseph, A.; Wong, G. W.; Schiemann, W. P.; Constantinescu, S. N.; Huang, L.; Khosravi-Far, R.; Steen, H.; Tewari, M.; Ghaffari, S.; Blobe, G. C.; Dang, C. V.; Garcia, J. G. N.; Pevsner, J.; Jensen, O. N.; Roepstorff, P.; Deshpande, K. S.; Chinnaiyan, A. M.; Hamosh, A.; Chakravarti, A.; Pandey, A. *Genome Res.* **2003**, *13*, 2363–2371.