DEVELOPMENT AND VALIDATION OF IN SILICO TOOLS FOR THE IDENTIFICATION OF POST-TRANSLATIONAL MODIFICATIONS AND SEQUENCE VARIATIONS IN HUMAN PROTEINS

By

Surendra Dasari

A THESIS

Presented to the Department of Medical Informatics and Clinical Epidemiology and

the Oregon Health & Science University

School of Medicine

in partial fulfillment of

Master of Science

May 2005

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is certify that the Master's thesis of

Surendra Dasari

has been approved

Advisor

Member

Member

Member

TABLE OF CONTENTS

ACKNOWLEDGEMENT

This work has been supported by National Institute of Health Grants U19ES11384 and U24DK5870 to Srinivasa Nagalla and EY07755 to Larry David. I thank Brian Searle and Phillip Wilmarth for playing a major role in this project. I also thank Ashok Reddy for assistance with the mass spectrometer, Mark Turner and Tim Sheard for insightful discussions, Melissa Standely for technical assistance and Nan Jiang for programming assistance. I also thank my advisor Christopher Dubay and my thesis committee for giving me an opportunity to write this thesis.

ABSTRACT

Mass spectrometry has become an important tool in proteomics. With the help of this technique we can study proteins at a much larger scale, which will aid further biological interpretations. However, Sophisticated algorithms that can interpret mass spectrometry data and robustly identify proteins, post-translational protein modifications and sequence variations present in a sample are needed. One of the problems with existing algorithms is that they often fail to identify unanticipated post-translational protein modifications and sequence variations. We present a novel mass-based alignment algorithm, implemented in OpenSea, which uses *de novo* sequencing results to robustly identify protein, post-translational protein modifications and sequence variations. OpenSea interprets the database protein sequences and the *de novo* peptide sequences (derived from a tandem mass spectrum) as a series of masses and the masses, rather than amino acid codes, are compared using a heuristic-based dynamic programming technique. Any unexplained mass shifts between the database peptide sequence and the *de novo* peptide sequence are automatically interpreted as post-translational protein modifications and sequence variations using an auto-interpretation routine. The identification of peptides and the location of identified post-translational protein modifications and sequence variations are validated using an independent rank-based correlation method. The performance of OpenSea was tested with three types of data: a mixture of ten known control proteins, a sample containing unknown and sequence modified human and rhesus monkey amniotic fluid proteins, and a 93 year old cataractous lens sample containing post-translationally modified proteins. In all the three cases we demonstrate that OpenSea can identify more peptides than existing algorithms

(SEQUEST, CIDentify) while accurately locating unanticipated post-translational protein modifications and sequence variations present in a high-throughput environment.

INTRODUCTION

In the post-genomics era, much of the focus is now shifting to the study of proteins (proteomics). There are several interesting aspects of a protein to study: i.e. its sequence, three-dimensional structure, or biological function etc. However, studying the sequence of a protein (protein sequencing) is still of a basic requirement in proteomics [1]. This is emphasized by the fact that a large part of the euakryote proteome remains unknown [2]. Furthermore, a protein's overall structure and functional relationships are largely determined by its sequence [3]. The function of a protein not only depends on its sequence but also on the modifications that are present in that protein. After a protein is manufactured inside the cell, it is processed in a myriad number of ways before it is released out of the cell. One of the important post-translational steps is the addition of chemical modifications to the protein sequence [4]. These post-translational modifications determine the function of a protein to a large extent. It has also been suggested that accumulation of posttranslational modifications over time can have detrimental effects on the function of protein [3-5]. Thus, the identification of post-translational modifications in proteins is an important part of protein sequencing.

A protein can also exist in several different forms. If there are polymorphisms at the genetic level, then the resulting proteins can differ from their original forms in terms of amino acid composition [3-4]. The sequence variations between the original form and the mutant form of the protein can make the mutant protein unstable. For example, numerous mutants of the hemoglobin protein have been discovered so far [6] and the majority of these mutants don't disrupt the function of the hemoglobin molecule. However, some of these mutants, like the sickle cell variant, disrupt the oxygen-carrying

function of the hemoglobin molecule [6]. When a protein's natural function is disrupted by sequence variation, then a pathological condition could occur. Thus along with post-translational modifications, identification of sequence variations in a protein is also an important aspect of protein sequencing.

Several existing methods for protein sequencing fall into two different classes, low-throughput and high-throughput, based on the throughput of the corresponding method. One of the foremost methods developed for protein sequencing is Edman degradation [7]. In this method, a pure form of a protein is sequenced from its N-terminus, amino acid after amino acid. However, this method cannot reliably identify the posttranslational modifications present in a protein. Furthermore, it requires a considerable amount of time to purify and sequence a protein using this method. The sensitivity of this method drops exponentially after sequencing the first thirty or forty amino acids of a protein, making this method less amenable to large-scale proteomics experiments.

With the advent of peptide ionization methods like electro-spray and MALDI, mass spectrometry (MS) became an important tool in proteomics for protein sequencing and quantification [8]. Figure 1 shows a typical MS-based approach for protein sequencing. Generally, the biological samples contain a large number of proteins. Before the samples are subjected to further analysis, they are purified to remove any extraneous materials. The purified complex protein mixture (Figure 1a) is simplified by separating it into smaller fractions using a SDS-PAGE gel [9]. The proteins that are present in the fractions are further fragmented into peptides using a protease (Figure 1b) [1,8]. The resulting peptide mixture is separated and ionized using reverse-phase liquid chromatography and electro spray ionization (ESI) technique (Figure 1c) [1,8]. The

ionized peptide species are then analyzed by a mass spectrometer. The mass spectrometer acquires the mass spectrum (MS) of the ionized peptide species eluting at a specific time point (Figure 1d). Using this mass spectrum, the computer controlled mass spectrometer determines the most abundant protonated peptide species and subjects it to further fragmentation inside the mass spectrometer using low energetic collision, also known as collision induced dissociation (CID), against an inert gas wall [8]. After CID, the mass spectrometer acquires the mass spectrum (MS/MS) of the resulting fragment ions (Figure 1e). The acquired MS and MS/MS spectra are stored for matching against protein sequence databases. The MS-based proteomics experiment can analyze multiple samples in a single day, making it suitable for high-throughput analysis. However, the stored unknown MS and MS/MS spectra have to be interpreted to find the peptides that are present from the proteins in the sample. Manual interpretation of such unknown spectra, on such a large scale, is not practical for high-throughput analysis. However, computer automation of such tasks is very effective.

Software packages (such as SEQUEST) have been developed to identify the peptides (and therefore the proteins) present in the sample by matching the amino acid sequence information present in an unknown MS/MS spectrum against the protein sequence database [10-11]. These programs take a protein sequence database and generate a list of candidate peptides by performing a theoretical digestion of proteins present in the database using the same experimental protease. The candidate peptide sequences for the unknown MS/MS spectrum are selected from the list of theoretical peptides by using a whole peptide mass filter, where the masses of the candidate peptide sequences should match the mass of the experimental peptide with in a specified tolerance. A set of hypothetical spectra is generated for the selected candidate peptide

sequences. These hypothetical spectra are compared to the unknown experimental MS/MS spectra using cross-correlation. A match score for each hypothetical spectrum is generated based on the degree of correlation between itself and the unknown MS/MS spectra. The peptide associated with the hypothetical spectrum that matches the unknown experimental MS/MS spectrum is assumed to be in the sample. A list of proteins that are most likely to be present in the sample is compiled from the list of identified peptides. However, this type of search fails to identify the sample peptides if they differ from the database peptides due to sequence variations or posttranslational modifications.

Such problems can be avoided by using alternative strategies. One such strategy is to create a database of proteins that contains all possible combinations of commonly occurring posttranslational modifications and sequence variations and to search the unknown spectra against the new database [12]. However, the size of a typical proteome and the number of possible posttranslational modifications and sequence variations make the search space prohibitively large. This technique is more effective when looking for a few user-defined modifications rather than unknown modifications. Since it is more probable that the modifications and sequence variations are present in the proteins that are already present in the sample, we can reduce the size of the new database by only considering the proteins identified in an initial database search [13-14]. This type of strategy has been implemented in the AutoMod subroutine of ProteinLynx [15]. However, this method requires the identification of at least one unmodified peptide in the initial database search. Furthermore, this method can only identify the modifications that are represented in the new database.

In order to identify peptides with unanticipated modifications, several different search strategies have been suggested. One such strategy derives the fragment ion series directly from the MS/MS spectrum [14,16]. The derived fragment ion series is then compared to peptide sequences without using any stringent parent ion mass filters. Another such strategy takes a short peptide motif and matches it against the experimental MS/MS spectrum [17]. Both these strategies can identify unknown posttranslational modifications, provided the modifications do not shift the sequence specific fragment ions of the ion series and the sequence motif. Because there is only a partial matching of the peptide sequence with the experimental MS/MS spectrum, both these approaches give a significant number of false positive matches. These techniques often only report mass shifts because of their inherent inability to locate the modification site. Hence, manual validation of the results is a requirement and these programs are not amenable for use in high-throughput proteomics experiments.

An effective way to identify peptides with unanticipated modifications or sequence variations is direct interpretation of MS/MS spectrum, known as *de novo* sequencing, without the help of a protein sequence database [18]. Typically, an MS/MS spectrum contains a series of sequence specific fragment ions where the mass difference between these ions corresponds to the amino acids in the original peptide (Figure 1e). These amino acids can be linked together to form a complete peptide sequence. Due to the irregular fragmentation of the peptide or presence of posttranslational modifications on some amino acids, some areas of the MS/MS spectrum cannot be linked directly to the mass of a single known amino acid [19]. This results in a partial peptide sequence rather than a complete peptide sequence, as it is a common occurrence. Manual interpretation of the MS/MS spectrum is time consuming and error prone. However, there are many

commercial and free programs that automate the *de novo* sequencing process [20-23]. These programs consider the irregularities in the fragmentation patterns and the possibility of occurrence of posttranslational modifications thus making the process amenable for high-throughput proteomics experiments.

De novo sequencers often produce several high scoring similar candidate peptide sequences for an experimental MS/MS spectrum. They also report sequence ambiguities for parts of MS/MS spectra in which either sequence specific fragment ion information is absent or an unanticipated modification is present. When using experimental MS/MS spectrum acquired on a low mass accuracy instrument such as a LCQ, these programs often fail to distinguish between isobaric amino acid residues. Also, structural isomers cannot be distinguished with low energy CID spectrum [21]. Hence, we need an error-tolerant search engine that can resolve sequencing errors and posttranslational modifications by matching the *de novo* sequences to a protein database using a homology based search.

Several search protocols for the BLAST sequence homology search engine have been developed to align *de novo* sequences with the proteins in the protein databases. For example, MS-BLAST uses a modified amino acid substitution matrix to account for single isobaric mistakes, sequence variations and possible modifications [24]. This method can be extended to account for ambiguous regions in the *de novo* sequences by further modifying the amino acid substitution matrix to include all possible combinations of amino acids that can add up to a single mass. However, as the number of ambiguous regions in a *de novo* sequence grows the number of combinations that need to be tested increase exponentially. This leads to increase time spent validating the results from such a search strategy and reduce throughput.

The popular FASTA sequence homology search algorithm has also been modified, known as FASTF, to search *de novo* sequences with ambiguous sites against protein sequence databases [25]. This algorithm determines the positional composition of the peptide using multiple similar *de novo* sequences. It uses the FASTA algorithm to find the peptide from the protein sequence database that has similar positional composition as the query sequences. This approach is useful in matching the ambiguous regions of a *de novo* sequence to a peptide sequence from the database. However, this approach cannot differentiate between *de novo* sequence errors and posttranslational modifications. Another program, CIDentify, attempts to correct for the *de novo* sequencing errors by employing a rescoring approach [26]. It initially matches the high confidence regions of the *de novo* sequence to a protein sequence using FASTA algorithm. After the initial alignment is made, unresolved consecutive mono and dipeptides can be matched to an adjacent section of the database sequence if they are isobars. This additional rescoring step helps to explain some common *de novo* sequencing errors and produce more accurate results.

The sequence homology approaches discussed above is limited in several ways when used to align *de novo* sequences containing ambiguous regions to database protein sequences. First, they often fail to match *de novo* sequences that are derived from marginal MS/MS spectrum to a database sequence. Second, they also fail to distinguish between *de novo* sequence errors and sequence modifications. Third, their results often require extensive manual validation for assembling the identified peptides in to proteins. These disadvantages make the sequence homology based programs very ineffective to use in a high-throughput protein identification environment like an MS-based proteomics experiment.

In this thesis, we describe a new mass-based alignment algorithm that matches *de novo* sequences to proteins in the sequence databases and overcomes many limitations of the other homology-based search algorithms. This algorithm was designed to align *de novo* sequences of all experimental MS/MS spectrum of a MS-based proteomics experiment to database protein sequences, even when the *de novo* sequences cannot account for the entire peptide sequence. The algorithm can also identify the post-translational modifications and sequence variations by interpreting the mass difference between the parent ion present in the experimental MS/MS spectrum and the identified peptide. The implementation of this algorithm is known as OpenSea, which can align *de novo* sequences derived from both high and low mass accuracy mass spectrometers. The performance of the algorithm was tested against other common database searching software using three types of datasets: a dataset of a known control proteins mixture, a dataset of unknown proteins with sequence variations, and a dataset of proteins with many known types of posttranslational modifications.

EXPERIMENTAL SECTION

OpenSea search algorithm aligns the *de novo* sequences derived from the MS/MS spectra to a protein sequence database. OpenSea search algorithm has six essential components.

1. A mass-based alignment algorithm that aligns the *de novo* sequences derived from the MS/MS spectrum to a protein database using amino acid masses present in the *de novo* and database sequences.
2. A scoring system that scores the peptides identified by the mass-based alignment between *de novo* and database sequence (mass-based alignment) to identify the

correct peptides present in the sample. It also contains a protein hits compiler that collates the peptide hits into a short and succinct list of proteins that may be present in the sample.

3. A post-translational modification and sequence variation identification algorithm that identifies and localizes any modifications or sequence variations present in a peptide hit by considering the mass difference between the corresponding *de novo* peptide sequence and the theoretical database peptide sequence.
4. An algorithm to cross-validate the identified peptide sequences with their corresponding MS/MS spectrum.
5. A unified scoring system that ties the peptide identification and the validation into a single scoring system.
6. An automated protein identification workflow that integrates the peptide, protein, post-translational modification, and sequence variation identification into a single step.

Mass-Based Alignment Algorithm. OpenSea search algorithm uses a mass-based alignment algorithm to align the *de novo* sequences derived from the MS/MS spectra to a protein sequence database. In order to rapidly locate the candidate protein sequences, OpenSea uses a tag-based search, similar to the FASTA method [27]. During the tag-based search OpenSea extracts the amino acids from short but high confidence regions of the *de novo* sequence. OpenSea locates the candidate proteins by using these short three to five amino acid tags to scan the protein database via a string search. *De novo* sequencers often fail to correctly determine isobaric amino acids from low energy CID MS/MS spectrum. Hence, they often report a representative amino acid residue for all isobaric amino acid residues (I/L and K/Q) instead of a correct amino acid residue. In

order to remedy this effect in the tag search, OpenSea replaces all isobaric amino acids in both the database sequence and the tag with a single representative symbol.

The matches found in the tag search are further subjected to a mass-based alignment (Figure 2a). For each match, all the amino acids in both the database sequence and the *de novo* sequence are converted into strings of either mono-isotopic or average masses (according to the mass spectrometer used in the experiment). A series of consecutive local alignments are made on either side of the tag match to generate a complete alignment. At each local alignment all possible combinations of the next three masses are compared using a "breadth-first search algorithm" (Figure 2b). OpenSea first tries to compare the single next mass using a user specified mass tolerance. If the masses are unequal then the comparison is repeated one level deeper, where the mass of a single database residue is compared to the combined mass of two successive *de novo* sequence residues, followed by combined mass of two database residues verses one *de novo* sequence residue and combined mass of two database residues verses the combined mass of two *de novo* sequence residues. For example, when aligning the amino acid residues, threonine (T), leucine (L), valine (V), and aspartic acid (D), first the mass of T (101.0 AMU) is compared to the mass of V (99.1 AMU). Since this is not a mass match, the combined mass of V + L (214.1 AMU) is compared to the mass of T (101.0 AMU), and finally the combined mass of T + L (214.1 AMU) to the combined mass of V + D (241.1 AMU). The algorithm doesn't try to perform all possible combinations of mass comparisons. For example, it precludes the possibility of comparing the combined mass of T + L (214.1 AMU) to the mass of V (99.1 AMU), since comparing the mass of T (101.0 AMU) to the mass of V (99.1 AMU) establishes the former comparison as a potential failure. This breadth-first search continues up to three levels deep until it finds

a mass match. However, the user can alter the depth of the search by changing a "search depth" parameter.

Figure 2b illustrates that the algorithm can correct for errors in the *de novo* sequence, when the sequencer failed to determine the correct order of the amino acids in a particular section of the MS/MS spectrum due to lack of proper fragment ion information. Figure 2c demonstrates that the algorithm can also correct for the errors in the *de novo* sequence, when the sequencer failed to sequence sections of the MS/MS spectrum due to noise or insufficient fragment ion information. When a mass match can't be made after three levels of search depth, OpenSea algorithm assumes that either a modification or substitution has occurred at that position (Figure 2d) and reports a mass mismatch error. OpenSea continues to make local alignments until it can account for the entire section of the *de novo* sequence. Although the search time grows when more levels are searched, some algorithmic and heuristic-based optimizations have been employed to reduce the search time. OpenSea can be configured to search for residue specific variable modifications by adding both the modified and unmodified mass to the residue.

Scoring of Mass-based Alignments and Identification of Proteins. Each local alignment is scored separately and the scores are summed to create an overall peptide alignment score. If a mass match has been made in a local alignment, then the score for that local alignment is the average identity scores for the database amino acid residues taken from the Blosum-90 substitution matrix [28]. If a variable modification has been found on any of the database amino acid residues, in the mass match, then the log odds score for that modification is used in the local alignment score calculation. The scoring algorithm for mass mismatches assumes that all the mismatches are due to substitutions.

Thus, if a mass mismatch has been made in a local alignment then the score for that local alignment is the substitution matrix score between the database residue and the *de novo* sequence residue. On the other hand, if the *de novo* sequence region in the mass mismatch has a mass instead of an amino acid symbol, then the average non-identity value (ANV) of the substitution matrix is used as the local alignment score. Gapped matches are also scored as mass mismatches. Since the peptide alignments are small, when compared to the whole protein alignments, gaps are allowed only at the beginning and end of the database sequence.

Local alignment mass matches are broken into primarily 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 the *de novo* sequences, respectively. Local alignment mass mismatches are broken into two categories: common substitutions (with score > 0) and uncommon substitutions (with score ≤ 0). The peptide alignment score, also known as OpenSea Alignment Score (OSAS), is a linear combination of the summed local alignment scores for the above-mentioned groups

(1)
$$\begin{aligned} \mathrm{OSAS} = {} & \alpha\left(\sum\nolimits_{matches}^{1\text{-}to\text{-}1}\right) + \beta\left(\sum\nolimits_{matches}^{1\text{-}to\text{-}m}\right) + \chi\left(\sum\nolimits_{matches}^{m\text{-}to\text{-}m}\right) \\ & + \delta\left(\sum\nolimits_{substitutions}^{common}\right) - \varepsilon\left(\sum\nolimits_{substitutions}^{uncommon}\right) - \phi\left(\sum\nolimits_{matches}^{gapped}\right) \end{aligned}$$

where the parameters $\alpha, \beta, \chi, \delta, \varepsilon, \phi$ are determined empirically, as 1.2, 1.1, 0.9, 1.0, 5.0, and 5.0 respectively, based on the optimal separation of correct and incorrect protein hits in the known control protein mixture. The candidate peptide alignment scores are adjusted for the presence or absence of the tryptic cleavage site (at lysine and arginine). If the candidate peptide alignment represents a perfect tryptic peptide then the score is augmented by 3.0 * average identity value (AIV) of the substitution score matrix.

Otherwise, the score is decreased by 4.0 * ANV of the substitution matrix. If the candidate peptide alignment has internal tryptic residues then the alignment score is decreased by ANV of the substitution matrix for each residue. Peptide matches with an alignment score over a particular score threshold (discussed in results) are accepted as correct identifications.

OpenSea uses an automated results compiler to identify the most probable proteins present in the sample. The results compiler, which is similar to ProteinProphet, uses an "Occum's Razor" approach to combine complex peptide identifications into protein hits [29]. The underlying principle of the Occum's Razor approach is to find the simplest possible combination of proteins that can explain most of the peptide hits (Figure 3). To find the simplest possible combination of the proteins, OpenSea first identifies the list of spectra that can be assigned to a list of proteins unambiguously. This is achieved by ranking the peptide hits above the score threshold by a "delta score" measure, which is the difference between the scores of the first and second best alignments for that spectrum. The spectrum with largest delta score is assigned to the protein corresponding to its best alignment. However, if the delta score is less than 20 percent of the score threshold then it is assumed that the proteins of both first best alignment and second best alignment for that spectrum are equivalent, and the spectrum is assigned to both of them. All other spectra that matched to the protein(s) in question within the delta score threshold are assigned to that protein(s). Of the remaining spectra, the spectrum with the next largest delta score is considered and the above procedure is repeated. This process is repeated until there are no further spectra for consideration. Due to this procedure, peptides that match multiple proteins equally well are assigned to the protein with strongest evidence (greatest delta score). This procedure also has an

inherent ability to identify proteins that have only a single good peptide hit (delta score above the delta score threshold). A cumulative protein score is calculated by adding the scores of the alignments that match independent regions of the protein. If more than one *de novo* sequence matches to a single region of the protein then only 10% of their scores are added to the protein score. Proteins that match the same set peptides with identical scores are considered as "degenerate" and are grouped together.

Identification of Post-translational Modifications and Sequence Variations. Any mass mismatches between the *de novo* sequence and the database sequence in a peptide alignment are processed by an auto-interpretation procedure in order to identify post-translational modifications or sequence variations. It is assumed that having multiple modifications on successive amino acids is very unlikely. Hence, the auto-interpretation procedure starts by grouping the successive mass mismatches into a single mass mismatch (Figure 4a, 4b). It uses the unexplained mass shift and the amino acids in the database sequence to determine the possible modification or sequence variation through a table lookup (Figure 4c). This lookup table contains the list of known posttranslational modifications and sequence variations, indexed by their corresponding masses, amino acids that can have that modification or sequence variation, and the likelihood of finding that modification or sequence variation in the biological sample. The log odds for the modifications are estimated based on the frequency of occurrence of sample processing artifacts and previously known human lens modifications. The Blosum90 substitution matrix provides the log odds score for substitutions. The sequence variation or modification suggested by the table lookup is inserted into the alignment. The alignment is rescored as if a new amino acid were identified at the location of the identified modification or sequence variation (Figure 4d). If the table

lookup fails to suggest a modification or a sequence variation then the alignment is rescored as if a mass mismatch has occurred in the alignment and an unknown mass shift is reported. Any ambiguity in the location of the suggested modification or sequence variation (Figure 4c) is resolved using a Smooth Rank Scoring (SRS) algorithm.

Validation of Peptide Identifications and Localization of Identified Modifications. The peptide identifications made by the OpenSea mass-based alignment algorithm are validated against the corresponding experimental spectrum using a rank-based scoring mechanism. This rank-based scoring scheme computes a quantity similar to the covariance between the ranked peaks of the experimental tandem mass spectrum and a hypothetical tandem mass spectrum derived from the identified peptide and the fragmentation model suggested in Table 1 [10]. To derive the ranks of peaks in the experimental tandem mass spectrum, the precursor ion is removed, and each peak in the mass spectrum is normalized by dividing the peak intensity by the mean of all visible peaks in a +/- 100 AMU range. The top 50 normalized peaks of the tandem mass spectrum are retained and ranked by their intensities, where the most intense peak is assigned the rank of 50 and the least intense peak is assigned the rank 1. A vector X corresponding to the integer values from 1 to MH^+ (mass of peptide with an additional proton) is created. The value of the vector X is zero except for the m/z values of the top 50 ranked peaks, where it is equal to the rank of the peak. The generated vector is normalized to the value one.

A hypothetical tandem mass spectrum is generated using the candidate peptide sequence and the fragmentation rules listed in Table 1. The hypothetical tandem mass spectrum vector, Y, is obtained in a same way as the experimental tandem mass spectrum

vector X. The normalized vectors X and Y are used to compute the Smooth Rank Score (SRS) given as:

$$(2) \qquad SRS = \left[\sum_{i=1}^{MH^+} [X_i - \overline{X}][Y_i - \overline{Y}] \right] + \begin{cases} 1 & for\ n = 0 \\ 1 - (n-1)^2 & for\ n > 0 \end{cases}$$

where X_i and Y_i are the normalized tandem mass spectrum rank value and hypothetical spectrum rank value, respectively, $\overline{X}$ and $\overline{Y}$ are the mean rank values in the respective rank vectors, and n is the total number of modifications and substitutions. The first term in the SRS score computes the covariance between the ranked peaks of the experimental tandem mass spectrum and a hypothetical tandem mass spectrum derived from the candidate peptide sequence. The second term is used to penalize the SRS score depending upon the number of modifications (n) identified in the candidate peptide sequence. The advantage of using a rank-based covariance, over a traditional cross-correlation method, is its robustness for non-normally distributed data like the intensities of the MS/MS peaks. Peptide identifications that score above a particular score threshold (discussed in results section) are accepted as valid identifications.

The SRS scoring system is also used to localize the ambiguous modification sites that are identified by the auto-interpretation procedure. Several alternatives are possible when the suggested modification can occupy more than one possible site in the alignment. These alternative interpretations differ from each other in the location of the suggested modification in the candidate peptide sequence. These alternative interpretations are validated by computing the SRS score between the experimental tandem mass spectrum and the corresponding candidate peptide sequences. The interpretation that has the highest SRS score is reported as the best interpretation, thus effectively localizing the modification to a single location. For example, an auto-

interpretation table lookup (Figure 4c) for the human gamma-S crystallin peptide shown in Figure 4b suggested a methylation (+14 AMU) modification on one of the Cysteine amino acid residues. However, there are two possible locations for the methylation modification (m), C24 and C26, in the local alignment. In this case, OpenSea algorithm generated two candidate peptide sequences, "YDCD[Cm]DCADFHTYLSR" (for methylation on C24) and "YDCDCD[Cm]ADFHTYLSR" (for methylation on C26), and validated each of them against the experimental tandem mass spectrum using the SRS score. The candidate peptide sequence with the highest SRS score (12.5 for "YDCD[Cm]DCADFHTYLSR") was selected as the best interpretation thus, in this case, localizing the methylation modification to C24.

Combined Protein Identification Score and Protein Identification Workflow. OpenSea algorithm combines the peptide identification score (OSAS) and the validation score (SRS) into a Combined Alignment Score (CAS). OSAS measures the extent of mass-based sequence homology between the *de novo* sequence and the database sequence, where as the SRS score measures the similarity between the normalized experimental tandem mass spectrum and the hypothetical fragmentation model of the identified peptide. The two scores are linearly combined, as CAS, for the optimal separation of the correct and incorrect peptide hits based on the analysis of a known control protein mixture:

$$CAS = \frac{0.9 * OSAS + 6.0 * SRS}{100} \qquad (3)$$

OpenSea uses a generalized three step automated workflow for high-throughput identification of proteins, sequence variations and posttranslational modifications that are present in a sample. In the first step of the workflow, OpenSea performs a fast scan of

the database, with stringent scoring parameters, to find all the identifiable proteins present in the sample. In the second step of the workflow, OpenSea performs an in-depth search of the remaining unidentified *de novo* sequences against a much smaller database of proteins that were identified in the first step. The search space for the second search is enlarged by decreasing the length of tag used in the tag search and increasing the depth of the "breadth-first search". The enlarged search space in the second step helps in the identification of peptides derived from poor quality tandem mass spectra. In the third step of the workflow, OpenSea searches the same smaller database of proteins looking specifically for the peptides that have similar modifications that were identified in the first and the second step of the workflow. This third step aids in the identification of peptides that have ambiguous mass gaps that were not identified in the second step. Although the scoring system for all the three steps is kept constant, the enlarged search space for the second and the third step of the workflow could help in the identification of new peptides. On a single Intel Pentium 4 2.0 GHz processor with 64 MB RAM, OpenSea takes approximately 14 seconds to search a single *de novo* sequence against the SwissProt database (Version 4.11 containing 127,873 entries) and to interpret any modifications or substitutions that may be present.

Sample Preparation and MS/MS Spectra Acquisition. OpenSea was tested using three sets of samples.

1. A mixture of ten known control proteins that was used to test and compare the peptide and protein identification of OpenSea with other existing programs.

2. Homo sapiens and Macca mulatta amniotic fluid samples, containing unknown, sequence-modified proteins, which were used to test the sequence variation identification capabilities of OpenSea algorithm.
3. A lens sample from a 93-year-old male human with nuclear cataracts containing post-translationally modified proteins that was used to test the post-translational modification identification capabilities of OpenSea algorithm.

The known protein control mixture was obtained by combining ten purified proteins of varying molecular weight and physicochemical properties including, Bos *taurus* insulin, ubiquitin, cytochrome c, superoxide dismutase, beta-lactoglobulin A, serum albumin, and immunoglobulin G, as well as *Equus caballus* myoglobin, *Armoracia rusticana* peroxidase, and *Gallus gallus* conalbumin (obtained from Ciphergen, Fremont, CA). The sample preparation and digestion protocols have been described elsewhere [30-31].

Homo sapiens and Macaca mulatta amniotic fluid samples were obtained from the Oregon Health & Sciences University. Proteins were separated by one-dimensional gel electrophoresis and were visualized by Coomassie staining. A high-molecular weight band from each sample was chosen for MS/MS analysis. Precise protocols used for sample preparation, gel electrophoresis and digestion have been described elsewhere [30-31].

The lens sample was obtained from the Oregon Lyons Eye Bank with Institutional Review Board approval from the Oregon Health & Sciences University. The sample preparation, fractionation and digestion protocols have also been described elsewhere

[30-31]. All samples were obtained with proper approval from Institutional Review Board.

Q-TOF MS/MS spectra were acquired with a MicroMass QTOF-2 (Waters) quadrupole/time-of-flight hybrid mass spectrometer with an online capillary liquid chromatography (Waters). Peptides were injected in to the QTOF mass spectrometer using a nanospray source. Ion trap MS/MS spectra were acquired on a LCQ classic ion trap mass spectrometer (ThermoFinnigan, San Jose, CA) with an Agilent 1100 series capillary liquid chromatography system (Agilent Technologies, Palo Alto, CA). Peptides were injected into the ion trap mass spectrometer using a standard electro spray source modified with a 34 G metal needle. Detailed descriptions of the equipment used for desalting the peptide mixture have been described elsewhere [30-31].

***De novo* Sequencing and Database Searching:** All Q-TOF MS/MS spectra were *de novo* sequenced using Peaks Batch (Version 2.2, Bioinformatics Solutions, Waterloo, ON Canada) *de novo* sequencer using a mass accuracy of 0.1 AMU. Peaks software reports full sequences without unknown mass regions [20]. However, it reports the confidence levels of individual amino acids in the sequence. Sequence regions in which the amino acids have a confidence level below 50% are replaced by their combined mass. However, if the average confidence of the entire sequence is below 50% then only the amino acid regions that have a confidence level below the average confidence are replaced by their combined mass. All of the sequences were searched with the OpenSea algorithm using mono-isotopic masses with a fragment ion mass tolerance of 0.25 AMU.

Ion trap MS/MS spectra were *de novo* sequenced with LutefiskXP using parent ion and fragment ion mass tolerances of 1.2 AMU and 0.4 AMU [32], respectively.

LutefiskXP was configured to consider a maximum of 500 subsequences and 2000 final sequences for speed considerations. Further, LutefiskXP was configured to report a maximum of five sequences with a score above 0.01 Pr(c) per MS/MS spectrum. All sequences were searched with OpenSea using average masses and a fragment ion mass tolerance of 0.5 AMU.

All MS/MS spectra that were generated from the control protein mixture were searched against the Swissprot database (Version 41.11 containing 128055 entries) that was modified to include the control protein sequences from PIR-NREF database (release 39) [33,34]. OpenSea analysis was compared to both SEQUEST (ThermoFinnigan) and CIDentify results [10,26]. Considering our sample processing protocols, SEQUEST was configured to search for variable oxidation and carbamylation. Peptide identifications with cross correlation scores greater than 1.8, 2.5, and 3.5 for singly, doubly, triply charged peptides respectively, and DeltaCN greater than 0.08 were accepted as correct SEQUEST identifications [10]. CIDentify was configured to search with fixed alkylations of cysteine and peptide identifications with a Z-score, which is defined as -log(E-score)/log(10), greater than 3.5 were accepted as correct identifications [26]. OpenSea was configured to look for fixed alkylations of cysteine. OpenSea was not configured to look specifically for any protein modifications. Proteins with multiple peptide hits having CAS scores greater than 0.9 were accepted as correct identifications.

OpenSea was used to analyze the data from the 93-year-old human lens sample by searching against the SwissProt database selected for human proteins (9615 entries). Peptide identifications with CAS scores greater than 0.9 and protein identifications with a protein score greater than 1.5 were accepted for further analysis. Posttranslational modifications identified by OpenSea in the lens sample were accepted only if they were

identified in at least two MS/MS spectra. Furthermore, all identified modifications were manually confirmed by at least one independent researcher.

RESULTS AND DISCUSSION

Identification of Proteins in Control Protein Mixture: A mixture of ten known proteins was used to test OpenSea. The control protein mixture was run on both high mass accuracy instrument (QTOF) and a low mass accuracy instrument (LCQ). A total of 10,685 tandem mass spectra from 35 LC/MS/MS runs of the control proteins were obtained from QTOF, where as a total of 14,469 tandem mass spectra from 20 LC/MS/MS runs of the same mixture were obtained from LCQ. Both datasets were processed with Peaks (QTOF) or LutefiskXP(LCQ) and then OpenSea.

One of the major requirements of high-throughput MS/MS analysis is an accurate peptide scoring system that can distinguish between correct and incorrect peptide identifications. The accuracy of the OSAS scoring system, for both QTOF and LCQ, was estimated by searching the *de novo* sequences generated from all LC/MS/MS runs of control protein mixture, acquired on both QTOF and LCQ, against the SwissProt protein database (version 4.11 containing 127,863 entries) modified to include control protein sequences from PIR-NREF database. Peptide assignments to the ten control proteins were considered as correct identifications and the rest of the peptide identifications were considered as incorrect identifications. As shown in Figure 5a, for high mass accuracy data, OpenSea mass-based alignment scoring system (OSAS) clearly distinguishes correct identifications from incorrect identifications when compared to other database searching software like SEQUEST (Figure 5b) and CIDentify (Figure 5c). At a cutoff of 85 OSAS scoring system has a sensitivity of 89% and specificity of 95%. For

comparison, SEQUEST has a sensitivity of 80% and specificity of 86% at Xcorr score cutoff of 2.5 (Figure 5b), where as CIDentify has a sensitivity of 76% and specificity of 70% at a Z-Score cutoff of 3.5 (Figure 5c). OpenSea mass-based scoring system (OSAS) was also tested for its accuracy when using low mass accuracy data from an ion trap instrument. As shown in Figure 6a, OpenSea mass-based alignment scoring (OSAS) system clearly distinguishes between correct and incorrect identifications even when using low mass accuracy data. At a cutoff of 85 OSAS scoring system has a sensitivity of 78% and specificity of 98%. For comparison, SEQUEST has a sensitivity of 74% and specificity of 91% at Xcorr score cutoff of 2.5 (Figure 6b), where as CIDentify has a sensitivity of 61% and specificity of 80% at a Z-Score cutoff of 3.5 (Figure 6c). The drop in the sensitivity of OSAS is not unexpected. When using LCQ data, low mass accuracy, interspersion of b and y fragment ions, and lack of both low m/z and high m/z peaks introduces more errors in the *de novo* sequences, which are beyond the correction capabilities of OpenSea algorithm. OpenSea uses an independent scoring (SRS) mechanism to validate the peptide identifications made by the OSAS. Both OSAS and SRS scoring systems were linearly combined in to a Combined Scoring System (CAS) based on the optimal separation of correct and incorrect peptide assignments (Figure 7a and 7b) of the peptides in the control protein mixture. When using the high mass accuracy data, the linear combination is justified by the higher sensitivity (97%) and specificity (98%) values of the CAS when compared to the sensitivity and specificity of OASS (89% and 95%) and SRS (99% and 71%) alone. The same CAS scoring parameters can be used for analyzing low mass accuracy data from ion traps. Since the sensitivity (97%) and specificity (87%) values of the CAS score were reasonable, no further tuning of the parameters was performed for ion trap data. This makes OpenSea

amenable for analyzing MS/MS data from both high mass accuracy and low mass accuracy instruments without any further tuning of the program.

A second requirement for high-throughput MS/MS analysis is accuracy and ease of interpreting protein identifications from peptide matches. The Occum's razor approach used by OpenSea algorithm to derive candidate protein hits from unambiguous spectra identifications has several advantages. One of the advantages is that a single spectrum is assumed to match to a single protein. In case the peptide matches several proteins equally well then it is assigned to the protein with greatest evidence. This reduces a lot of degeneracy in protein hits, which is a very desirable feature in high-throughput MS/MS analysis programs, especially in the case of large experiments. Furthermore, protein evidence is generated based on how exclusively a single spectrum matches, using the delta score, to that protein. Another advantage of the approach used by OpenSea is that if a peptide matches an over-all low coverage protein with a high confidence (very high delta score) then the low coverage protein is reported, even though the same peptide matched to other high coverage proteins. This helps in the identification of low abundance proteins that often have only one good peptide hit. However, if the sample contains homologous proteins then OpenSea can be configured to report all the degenerate proteins without any filtering.

A third desirable feature of a high-throughput MS/MS analysis program is to automatically interpret the mass differences between the parent ion of an experimental tandem mass spectrum and the identified peptide as posttranslational modifications or sequence variations. The auto-interpretation routine of OpenSea combined with its mass-based alignment algorithm gives it the capability of identifying and interpreting unanticipated posttranslational modifications and sequence variations present in

experimental tandem mass spectra from both high and low mass accuracy instruments. OpenSea can also localize the identified modifications and sequence variations whenever such discriminating information is present, to correct locations of the MS/MS spectrum, thus reducing the demands on the manual interpretation and validation of the identified modifications.

Comparison of OpenSea to Other MS/MS Protein Identification Software. One LC/MS/MS run of the control protein mixture acquired on both QTOF (328 MS/MS spectra) and LCQ (707 spectra) was examined in order to benchmark the number of accurately identified spectra by OpenSea, as compared to the other common protein identification software. Protein identifications from the QTOF and LCQ data sets were made by SEQUEST, and by two *de novo* sequence alignment programs, OpenSea and CIDentify. Peaks and LutefiskXP were used to provide *de novo* sequences for both OpenSea and CIDentify. The total number of visually verified spectra, in both LCQ and QTOF data sets, identified by various programs (or combination of programs) was tabulated in Table 2.

In the control protein mixture sample run acquired on the QTOF instrument, OpenSea/Peaks, CIDentify/Peaks and SEQUEST were able to identify one of the two tryptic peptides from bovine insulin that are in the mass range of the experiment (data not shown). However, the matches were difficult to verify because only one peptide from insulin was found. In the same data set, OpenSea, when using *de novo* sequences derived by Peaks, identified 48% more spectra than CIDentify. However, SEQUEST was able to identify more MS/MS spectra than either OpenSea or CIDentify alone, as it is more sensitive towards poor quality MS/MS spectra. In contrast, most of the *de novo* sequencers fail to produce even acceptable *de novo* sequences for poor quality spectra,

which results in the lower number of spectra identifications by both OpenSea and CIDentify. CIDentify identified more MS/MS spectra when using *de novo* sequences from LutefiskXP, rather than Peaks. This is due to the fact that CIDentify is less sensitive to the lower quality *de novo* sequences (due to its assumption that *de novo* sequences are often correct and sequence variations thus found in the database search are often real) coupled by the fact that LutefiskXP produces better *de novo* sequences than Peaks [26,32].

When analyzing the control protein mixture run acquired on LCQ, none of the programs found any peptide from bovine insulin. This might be because both tryptic peptides in the bovine insulin are long peptides with multiple basic residues in them. This might result in higher charge states for their parent ions, whose MS/MS spectra are difficult to analyze through *de novo* sequencing, especially when acquired on a low mass accuracy instrument such as a LCQ. OpenSea, when using *de novo* sequences derived from LutefiskXP, identified 73% more spectra than CIDentify from the LCQ data. Further confirming the superior sensitivity of SEQUEST towards poor quality spectra, the SEQUEST algorithm identified more MS/MS spectra from the LCQ data set than OpenSea and CIDentify. Confirming previous reports [26-32], OpenSea identified 18% more MS/MS spectra when using *de novo* sequences from LutefiskXP, rather than Peaks, in the database search. CIDentify identified twice as many MS/MS spectra when *de novo* sequences from LutefiskXP, rather than Peaks, were used in the database search.

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 the alignments to single or pairs of residues, and the further interpretation often results in higher alignment scores for correct matches. Second, all alignments must pass

stringent, empirically derived criteria requiring that the entire *de novo* sequence should be accounted for, allowing only one consecutive sequence modification, and the requirement that each alignment must contain at least one accurately matching sequence tag. Third, OpenSea scoring system separates correct identifications from incorrect identifications more accurately than CIDentify, leading to the better identification *de novo* sequences derived from poor quality MS/MS spectrum without introducing false positives in the identifications.

Identification Of Unknown Homologous Proteins And Sequence Variations. OpenSea can identify proteins from organisms whose proteome has not been completely sequenced, provided they share reasonable close sequence homology to other well-sequenced proteomes. To test this hypothesis, OpenSea was used to identify proteins from two unknown samples, human and rhesus monkey amniotic fluid, using an adult human protein database. Human amniotic fluid contains proteins that are homologous to adult human proteins with sequence modifications. For example, the γ chain of fetal human hemoglobin contains 39 sites of sequence variations from the adult β chain [35]. The rhesus monkey proteome was also not well sequenced. However, most of the rhesus monkey proteins are also homologous to adult human proteins with some sequence modifications.

One LC/MS/MS run of Homo sapiens amniotic fluid sample from a high-molecular weight 1D gel band was analyzed on QTOF (311 MS/MS spectra). The spectra were *de novo* sequenced using Peaks software and the resulting sequences were used by OpenSea to search the Swiss-Prot protein database selected for human proteins (9426 spectra). The same spectra were also processed using CIDentify/Peaks and Sequest. Protein identifications from each search engine were manually validated (Table

3a). Sequence variations suggested by OpenSea/Peaks were accepted only if the corresponding MS/MS spectra showed a strong presence of sequence-specific fragment ions around the site of the variation. As expected, Sequest was able to find a few unmodified peptides from these samples. CIDentify identified a subset of peptides identified by OpenSea. The OpenSea algorithm, with the help of the mass-based alignment and Auto-Interpretation routine, was able to identify a greater number of proteins along with any sequence variations that were present between the adult and fetal form of the proteins.

To further test this hypothesis, one LC/MS/MS run of Maraca mulatta amniotic fluid sample from a corresponding high-molecular weight band was analyzed on QTOF (315 MS/MS spectra). The spectra were analyzed and the identified proteins and sequence variations were validated in a similar fashioned outlined for the Homo sapiens amniotic fluid sample run (Table 3b). As expected, OpenSea routinely identified peptides with sequence variations from their human homologues, again outperforming both Sequest and CIDentify. A major advantage of OpenSea algorithm over other similar programs is its ability to distinguish between *de novo* sequence errors and actual sequence variations. Because OpenSea's mass-based alignment can align isobaric residues of arbitrary length, it can account for common errors found in *de novo* sequences, while at the same time identifying real sequence variations present in the peptide.

Identification Of Co- and Posttranslational Modifications In Human Cataractous Lens Tissue. Eleven crystallin proteins represent the majority of total human lens tissue protein by mass [36]. Pristine crystallins form highly regular aggregates, which produce the requisite optical properties (transmittance, reflectance, and

refraction) of the lens [37]. Crystallin proteins in human lens do not turn over and, as the tissue ages, they accumulate substantial modifications. As they accumulate age-related modifications, crystallins dissociate from the highly regular aggregates and start to form irregular conglomerations. It is suspected that these conglomerations change the optical properties of the human lens, by increasing reflectance and decreasing refraction and changing other optical properties, leading to cataractogenesis [38]. An extensive catalog of protein modifications in human lens includes carbamylations, acetylations [39], phosphorylations [40], oxidations [41-43], methylations [38,44], and deamidations [41-43]. In particular, deamidations and carbamylations of the crystallins have been linked with aging and cataractogenesis [37-39]. As human lens tissue has relatively low protein complexity and many modifications in the lens proteins have already been well characterized, the lens tissue is an ideal choice to test our techniques that can identify co- and posttranslational modifications.

In this study, a 93-year-old human male lens tissue containing an age-related nuclear cataract was separated into water-soluble and water-insoluble fractions. Both of these fractions were digested and fractionated via cation exchange HPLC. All of the fractions were reverse phase separated and analyzed on both QTOF and LCQ instruments. The tandem mass spectra from both the instruments were analyzed for modifications using the OpenSea algorithm, as outlined in the experimental section.

OpenSea was able to identify 80 sites of modifications in eleven crystallins, as cataloged in Table 4. All of the reported modification sites were confirmed with both QTOF and LCQ data. Table 4a confirms 44 sites of phosphorylation, acetylation, oxidation, methylation and deamidation that were previously reported in literature [37-44]. Table 4b reports 36 new modifications discovered in this study. Although there is a

significant difference in fragmentation of peptides between the instruments, 85% of the reported modifications in table 4 were identified in both instruments, which gives credibility to our results. Furthermore, 210 MS/MS spectra from the soluble fraction and 587 spectra from the insoluble fraction were assigned to deamidated peptides. Using the spectra count as a rough estimate for relative abundance [45], this suggests a three-fold increase in the overall extent of deamidation in the insoluble fraction as compared to the soluble fraction, which is consistent with previously reported values, lending further credibility to our results [41-42].

Along with the modifications listed in Table 4, seventy-five N-terminal carbamylation sites were found represented by 526 QTOF and LCQ spectra. However, no lysine carabamylation sites were identified in this study, as suggested in previous studies, suggesting a strong N-terminal preference [38]. Furthermore, twenty-three sites of pyroglutamic acid [46] and N-terminal S-carbamyl methyl cysteine cyclization [47] sites were identified. These cyclized forms are common *in vitro* modifications. There was a +38 AMU mass shift identified in seventeen acidic sites. Most of the spectra with this modification showed a charged neutral loss of +38 AMU from the parent ion. As 350 mM KCL was used during the strong cation exchange separation, it is possible that the potassium cation was bound to the acidic residues in some peptides.

The +28 AMU mass shift reported in Table 4b was unanticipated and was found on serine and histidine residues. This modification could be either a dimethylation modification as suggested by Hansen *et al.* [41] in a previous lens study or an *in vitro* formylation. Although 5% formic acid was use to stop the digestion, formylation was not observed in any of the control protein mixture samples, which were processed using a similar protocol. Further analysis is needed to characterize the unanticipated

modifications. Further studies are also needed to test any biological significance of the new modifications identified in this study.

CONCLUSION

Mass-based alignment of *de novo* sequences can accurately identify sequence modifications, thus allowing the approach to be used in high-throughput proteomics experiments. No previous study has been able to identify as many new posttranslational modifications in human lens as this study, which demonstrates the effectiveness of the mass-based alignment approach in high-throughput proteomics experiments. The major advantages of the OpenSea algorithm are in the sophistication of its scoring function for mass-based alignments, cross validation of the peptide identifications, automated interpretation of mass-mismatches between *de novo* sequence and the database peptide sequence as posttranslational modifications and sequence variations using a lookup table, and a simple Occum's razor based protein results compiler. OpenSea can simultaneously search for over 75 euakaryote co- and posttranslational modifications. An equivalent search is either impossible or unproductive when using other protein database search engines.

OpenSea is batch-scriptable and the protein identification results are reported as XML files, which facilitates further processing of results. OpenSea was written in Java and it will run on any platform that can run Java Runtime Environment (1.3). OpenSea has been tested on Windows and Linux platforms.

REFERENCE

1. Aebersold, R.; Goodlett, D. R.; *Chem. Rev.;* (Review); **2001**; *101*(2); 269-296.
2. Paul, M. H; Anuj, K; Ning, L; Michael, S; Mark G; *Neu. Acid. Res.;***2002**;30(5);1083-1090.
3. Todd, C. W.; William, R. P,; *J. Mol. Bio.;* **1999**; 291; 977-995.
4. Molecular Biology of the Cell.; Garland Publishing.; 3rd edition; March **1994**;
5. Posttranslational Modification of Proteins:Tools for Functional Proteomics.; Humana Press.; 1st edition.; April **2002**.;
6. Basset, P.; Beuzard. Y.; Garel, M.C.; Rosa. J.; *J. Blood.;* **1978;** 51; 971-982.
7. Edman, P.; *Acta. Chem. Scand.;* **1950**; 4; 283.
8. Aebersold, R.; Mann, M. *Nature* **2003**, 422, 198-207.
9. Scheffler, K. N.; Falick, A. M.; Hall, S. C.; Ray, W. C.; Post, D. M.; Munson, R. S. Jr.; Gibson, B. W.; *J. Prot. Res.;* **2003**; 2; 523-533.
10. Eng, J. K.; McCormack, A. L.; Yates, J. R. III.; *J. Am. Soc. Mass Spectrom.*; **1994**;5; 976-989.
11. Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrel.; J. S. *Electrophoresis.;* **1999**; 20; 3551-3567.
12. Yates, J. R. III; Eng, J. K.; McCormack, A. L.; Schieltz, D.; *Anal. Chem.;* **1995**; 67;1426-1436.
13. Gatlin, C. L.; Eng, J. K.; Cross, S. T.; Detter, J. C.; Yates, J. R.; *Anal. Chem.;* **2000**; 72; 757-763.
14. Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. *Genome Research* **2001**, 11, 290-299.
15. 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.
16. 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 Topics.; Portland, Oregon; May 12–16; 1996.
17. Liebler, D. C.; Hansen, B. T.; Davey, S. W.; Tiscareno, L.; Mason, D. E.; *Anal. Chem.;* **2002**; 74; 203-210.
18. Biemann, K.; *Meth. Enzy.;* **1990**; 193; 455-479.
19. Richard, S. J.; Biemann, K.; *Biomed. and Environ. Mass Spec.*; **1989** ;18; 945-957.

20. Ma, B.; Zhang, K.; Lajoie, G.; Doherty-Kirby, A.; Liang, C.; Li, M. Doherty-Kirby, A; Lajoie, G.; *Rapid. Commun. Mass. Spectrom.;* **2003**; 17; 2337-2342.

21. Taylor, J. A.; Johnson, R. S.; *Anal. Chem.;* **2001;** 73; 2594-2604.

22. Fernandez-de-Cossio, J.; Gonzalez, J.; Betancourt, L.; Besada, V.; Padron, G.; Shimonishi, Y.; Taka.; T. *Rapid Commun. Mass Spectrom.;* **1998;** 12; 1867-1878.

23. Lu, B.; Chen, T. *J. Comp. Biol.;* **2003;** 10; 1-12.

24. 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.

25. Mackey, A. J.; Haystead, T. A. J.; Pearson, W. R.; *Mol. Cell. Proteomics;* **2002**; 1; 139-147.

26. Taylor, J. A.; Johnson, R. S.; *Rapid Commun. Mass Spectrom.;* **1997**;11;1067-1075.

27. Pearson, W. R.; Lipman, D. J.; *Proc. Natl. Acad. Sci.* **1998**; 85; 2444-2448.

28. Henikoff, S.; Henikoff, J. G.; *Proc. Natl. Acad. Sci.;* **1992**; 89; 10915-10919.

29. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R.; *Anal. Chem.;* **2003;** 75; 4646-4658.

30. Searle, B. C.; Dasaris, S.; Turner, M.; Reddy, A. P.; Choi, D.; Wilmarth, P. A.; McCormack, A. L.; David, L. L.; Nagalla, S. R. *Anal. Chem.* **2004**, 76, 2220-2230.

31. Searle, B. C.; Dasaris, S.; Wilmarth, P. A.; Turner, M.; Reddy, A. P.; David, L. L.; Nagalla, S. R.; J. *Prot. Res.;* (In Press).

32. Johnson, R. S. "Lutefisk1900 vs Peaks: A comparison of automated *de novo* sequencing programs"; ABRF '04: Integrating Technologies in Proteomics and Genomics; Portland, Oregon; February 28-March 2; 2004.

33. Lorkin, P. A.; *J. Med. Genet.;* **1973**; 10; 50-64.

34. Cathy H. W.; Lai-Su L. Y.; Hongzhan H.; Leslie A.; Jorge C.; Yongxing C.; Zhang-Zhi H.; Robert S. L.; Panagiotis K.; Baris E. S.; Vinayaka C. R.; Jian Z.; Winona C. B.; *Nuc. Acids Res.*; **2003**; 31;345-347.

35. Lorkin, P. A.; *J. Med. Genet.;* **1973;** 10; 50-64.

36. National Advisory Eye Council, *Report of the Retinal Diseases Panel: Vision Research: A National Plan, 1994-1998.* United States Department of Health and Human Services, Bethesda, MD, 1993; Publication NIH 93-3186; pp 175-183.

37. Lapko, V. N.; Smith, D. L.; Smith, J. B.; *Prot. Sci.* **2001**; 10; 1130-1136.

38. Lapko, V. N.; Smith, D. L.; Smith, J. B.; *Protein Sci.;* **2003;** 12; 1762-1774.

39. Lapko, V.N., Smith, D.L., Smith, J.B. *Protein Sci.* **2001**, 10, 1130-1136.

40. Lund, A. L.; Smith, J. B.; Smith, D. L. *Exp. Eye Res.* **1996**, 63, 661-672.

41. Hanson, S. R.; Hasan, A.; Smith, D. L.; Smith, J. B.; *Exp. Eye Res.;* **2000;** 71; 195-207.

42. Zhang, Z.; Smith, D. L.; Smith, J. B.; *Exp. Eye Res.;* **2003;** 77; 259-272.

43. Harrington, V.; McCall S.; Huynh, S.; Srivastava, K.; Srivastava, O. P.; *Mol. Vis.* **2004;** 10; 476-489.

44. Lapko, V. N.; Smith, D. L.; Smith J. B. *Biochemistry*, **2002**, 41, 14645-14651.

45. Liu, H.; Sadygov, R. G.; Yates, J. R. III.; *Anal. Chem.;* **2004;** 76; 4193-4201.

46. Khandke, K. M.; Fairwell, T.; Chait, B. T.; Manjula, B. N.; *Int. J. Peptide Protein Res.;* **1989;** 34; 118-123.

47. Geoghegan, K. F.; Hoth, L. R.; Tan, D. H.; Borzilleri, K. A.; Withka, J. M.; Boyd, J. G.; *J. Proteome Res.;* **2002;** 1; 181-187.

TABLE 1

Ion Type [a]	Intensity [b]
b- or y-type	50
b- or y-type +1 AMU	25
b- or y-type -1 AMU [c]	25
b- or y-type -H2O	10
b- or y-type -NH3	10
a-type	10
b- or y-type +2H [d]	25
Neutral loss of modification [e]	50

[a] Ion types used in the SRS peptide fragmentation model. [b] The intensity of each ion type. This model is similar to the fragmentation model described in ref 10. [c] b- or y-type -1 AMU are only incorporated into the model when average masses are used to calculate the mass of ions. [d] +2 charged b- and y-type ions are incorporated only for peptides of +3 charge or higher. [e] If a protein modification is expected and that modification has a known labile bond, the neutral loss of the modification from the parent ion is also included in the fragmentation model.

TABLE 2

The number of correctly assigned MS/MS spectra from the control protein mixture acquired on the QTOF or LCQ.

	QTOF [b]					LCQ [c]				
	OpenSea [d]		CIDentify [e]		SEQUEST [f]	OpenSea		CIDentify		SEQUEST
Protein [a]	Peaks [g]	LutefiskXP [h]	Peaks	LutefiskXP		Peaks	LutefiskXP	Peaks	LutefiskXP	
Bovine Serum Albumin	54	49	24	35	57	49	57	17	31	85
Bovine Cytochrome C	5	4	5	2	6	4	0	1	2	8
Bovine Immunoglobulin G	29	22	15	19	32	30	35	6	16	34
Bovine Beta-Lactoglobulin	11	10	5	7	10	6	7	2	6	9
Equine Myoglobin	11	11	4	8	9	9	7	5	3	11
Chicken Conalbumin	31	28	14	21	30	28	36	8	20	44
Horseradish Peroxidase	4	5	3	5	6	7	9	3	6	9
ovine Superoxide Dismutase	5	7	1	4	9	5	9	1	6	10
Bovine Ubiquitin	4	3	1	3	3	6	9	0	6	10
Bovine Insulin	0	0	0	0	0	0	0	0	0	0
Total Peptide Identifications	154	139	72	104	162	144	169	43	96	220

[a] Proteins present in the control protein mixture. Tandem MS/MS spectra from both a high mass accuracy instrument [b] (328 MS/MS spectra) and a low mass accuracy instrument [c] (707 MS/MS spectra) were used to compare the programs. OpenSea [d] and CIDentify [e] are two *de novo* sequence alignment programs, where as SEQUEST [f] is a commonly used database search program. Peaks [g] and LutefiskXP [h] were used to provide *de novo* sequences for OpenSea and CIDentify. At least two peptide hits for a protein were required for identification.

TABLE 3

The number of MS/MS spectra from human (a) and rhesus monkey (b) amniotic fluid samples that were assigned to adult human proteins.

a

Protein Name	OpenSea/ Peaks[a]	CIDentify/ Peaks[a]	Sequest[a]	Verified Substitutions Found by OpenSea/Peaks[b]
Lactotransferin	21	13	5	5
Glia-derived nexin	17	5	5	0
Serotransferin	8	4	2	2
Serum Albumin	7	0	5	3
Alpha-1-acid Glycoprotein	3	2	2	0
Alpha-1-antichymo Trypsin	5	0	2	2
Histidine Rich Glycoprotein	2	0	0	1
Meosin	2	0	2	0
total	**65**	**24**	**23**	**13**

b

Protein Name	OpenSea/ Peaks[a]	CIDentify/ Peaks[a]	Sequest[a]	Verified Substitutions Found by OpenSea/Peaks[b]
Lactotransferin	26	13	5	10
Glia-derived nexin	13	5	5	1
Collagen Alpha 2(I) chain	7	3	0	9
Alpha-1-antitrypsin	6	2	2	4
Serum Albumin	4	0	3	0
92 kDa type IV collagenase	2	0	2	0
Alpha-1-antichymo Trypsin	0	2	0	0
Total	**58**	**25**	**17**	**24**

[a] OpenSea/Peaks, OpenSea/CIDentify, and Sequest were used to identify analogous human adult proteins when searching with proteins in human and rhesus monkey amniotic fluid samples. At least two spectra identifications were required for the identification of the protein by every identification program. [b] The number of amino acid sequence variations found between adult proteins and homologous amniotic fluid proteins by OpenSea/Peaks and confirmed manually by visual inspection of all the MS/MS spectra with corresponding sequence variations.

TABLE 4

Summary of identified modifications and sites in 93-year old cataractous human lens.

A: Confirmation of Previously Reported Protein Modifications

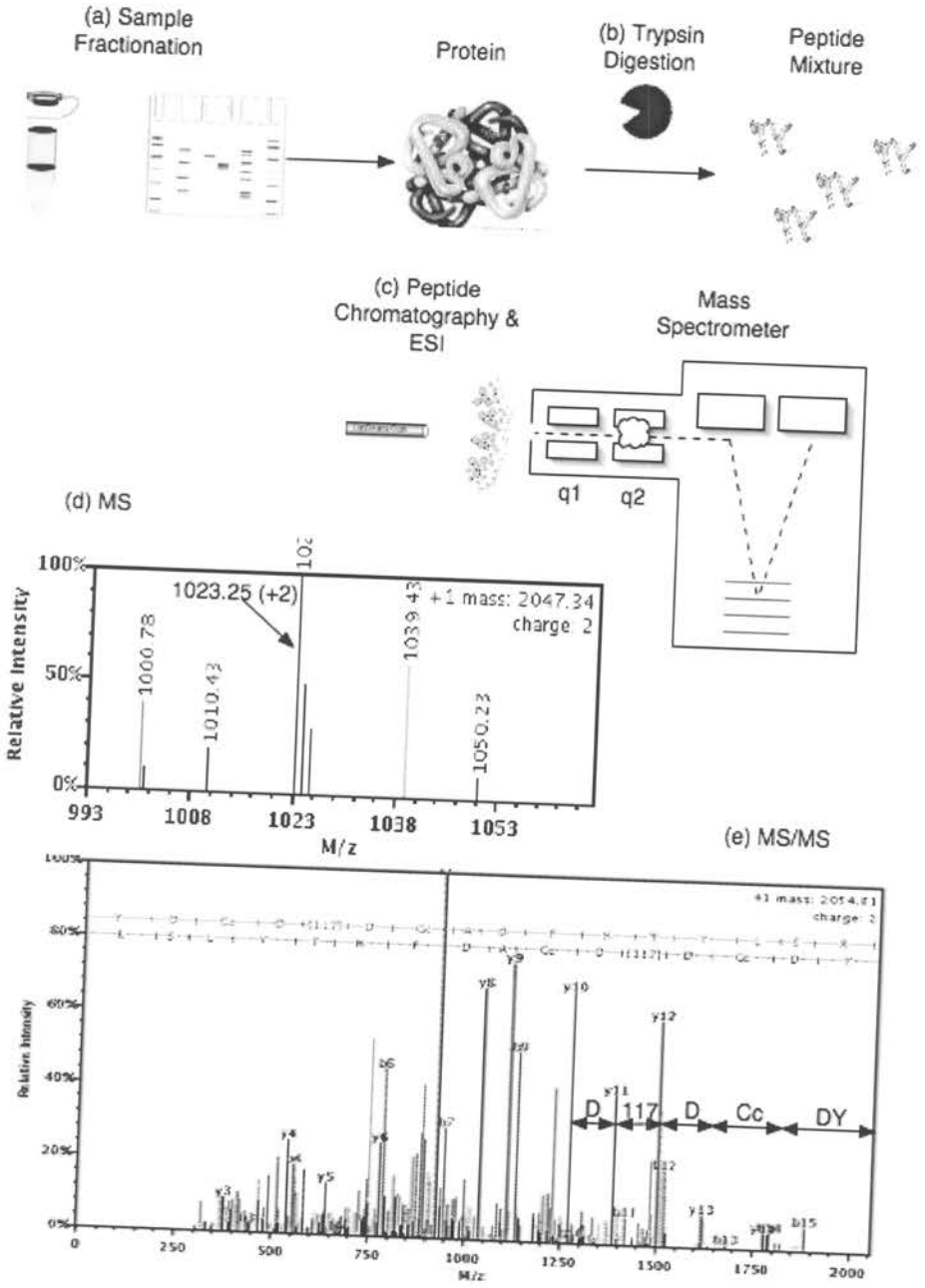
Protein / Accession Number [a]	Deamidation	Oxidation	Methylation	Phosphorylation	Acetylation
Crystallin, αA chain (P02489)	Q6, Q90, Q147	M1		S122	n1
Crystallin, αB chain (P02511)	N146	M1, M68[b]		S59	n1
Crystallin, βA3 (P05813)	Q42, N54, N103, N120, Q164	M126[b]	C82, C117, C185		n1
Crystallin, βA4 (P53673)		M13			n1[b]
Crystallin, βB1 (P53674)	N157, N161	M112, M136, W192, M225[b]			
Crystallin, βB2 (P43320)		M121, W150			n1
Crystallin, βB3 (P26998)					n1[b, d]
Crystallin, γB (P07316)					
Crystallin, γC (P07315)			C22		
Crystallin, γD (P07320)			C110		
Crystallin, γS (P22914)	N14, Q16, Q63, N76, Q120	M58, M73	C24, C26		

B: Newly Identified Protein Modifications

Protein / Accession Number [a]	Deamidation	Oxidation	Methylation	+28 AMU[e]
Crystallin, αA chain (P02489)	N123	W9[b]		S20, H79
Crystallin, αB chain (P02511)		W9		H83[b]
Crystallin, βA3 (P05813)	N40, N62, N133	M46, W96, W99,[b] W168		
Crystallin, βA4 (P53673)	N82, N113			
Crystallin, βB1 (P53674)	N57, N67, Q69, N124	W100[c]		S151, H214[b]
Crystallin, βB2 (P43320)	N115, Q162	M192[b]		
Crystallin, βB3 (P26998)	N155			
Crystallin, γB (P07316)		M102		
Crystallin, γC (P07315)	N24, Q66	M101 or M102[f]		
Crystallin, γD (P07320)	Q12, N49,[c] N160	W156		
Crystallin, γS (P22914)		W162	C114	

[a] Protein sequences are referenced to their SwissProt database accession number. [b] Modification observed only in ion trap data. [c] Modification observed only in Q-TOF data. [d] N-terminally acetylated without initial methionine residue. [e] The +28 AMU mass shift modification does not agree with any previously reported lens modifications. [f] Modification could be either M101 or M102.

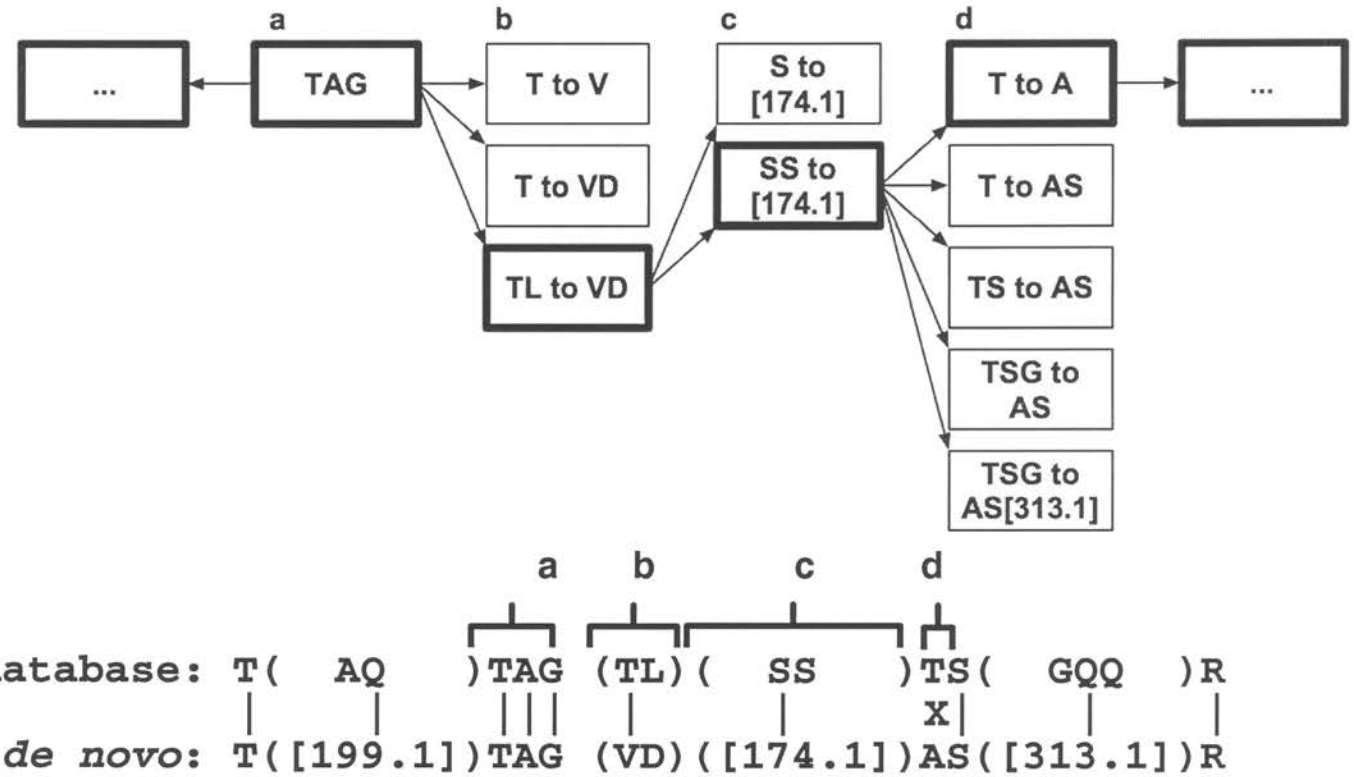
FIGURE 1



A schematic diagram of a typical MS-based proteomics experiment. The complex protein mixture **(a)** is separated into small fractions using either SDS-PAGE or liquid chromatography. The proteins in each fraction are subjected to trypsin digestion **(b)**. The peptides in the resulting peptide mixture are separated and ionized using reverse-phase liquid chromatography and electro spray ionization technique (ESI) **(c)**. After the

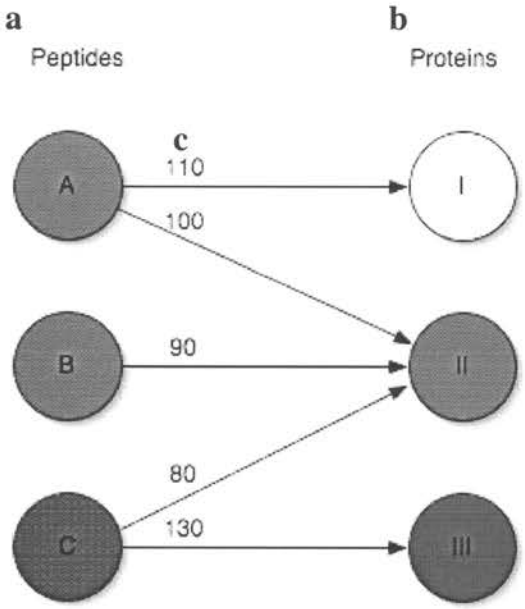
multiply-protonated peptides enter the mass spectrometer, a mass spectrum of those peptides is acquired **(d)**. A highly abundant protonated peptide species is selected and subjected to further fragmentation by a low energetic collision with an inert gas. Finally, a tandem mass spectrum (MS/MS) of the generated fragment ions is acquired **(e)**. The MS and MS/MS spectrum are stored for matching against protein sequence databases. The database matching will identify the peptides and therefore the proteins that are present in the fraction.

FIGURE 2



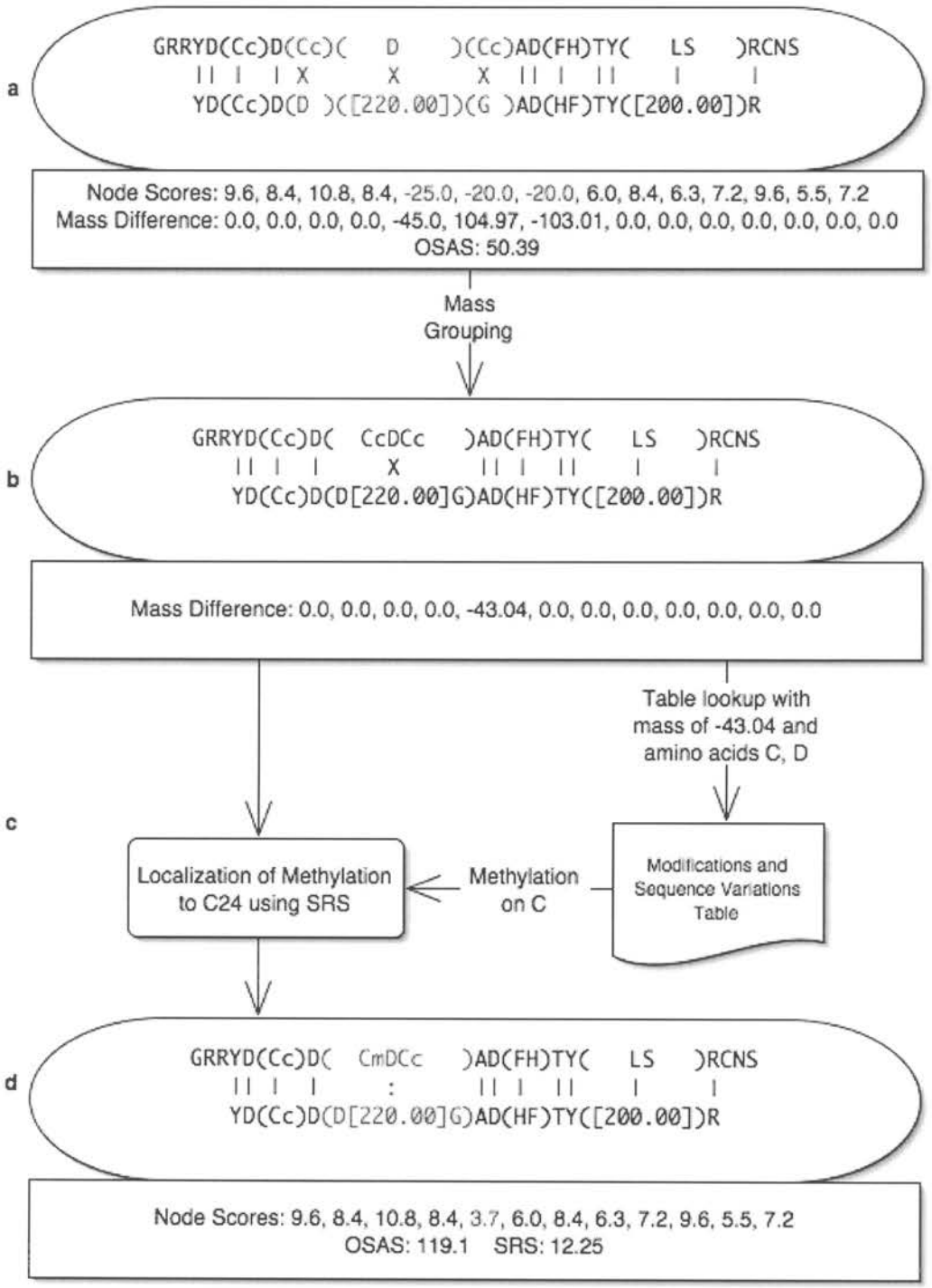
Mass-based *de novo* sequence alignment scheme. Here the *de novo* sequence "T[199.1]TAGVD[174.1]AS[313.1] R" is aligned to a database peptide sequence "TAQTAGTLSSTSGQQR" using the "breadth-first search" method with a depth of two amino acids. A tag match (a) is used to initiate a mass-based alignment 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.

FIGURE 3



A schematic diagram of OpenSea results compiler. The peptides **(a)** A, B, and C are matched to the proteins **(b)** I, II and III. The alignment scores **(c)** for each peptide are also shown in the figure. The peptides A and B are assigned to protein II and the peptide C is assigned to protein III using Occum's Razor approach (See text for further discussion). The peptide score threshold and the delta score threshold were assumed as 75 and 15. OpenSea results compiler can detect proteins with at least one significant peptide hit.

FIGURE 4

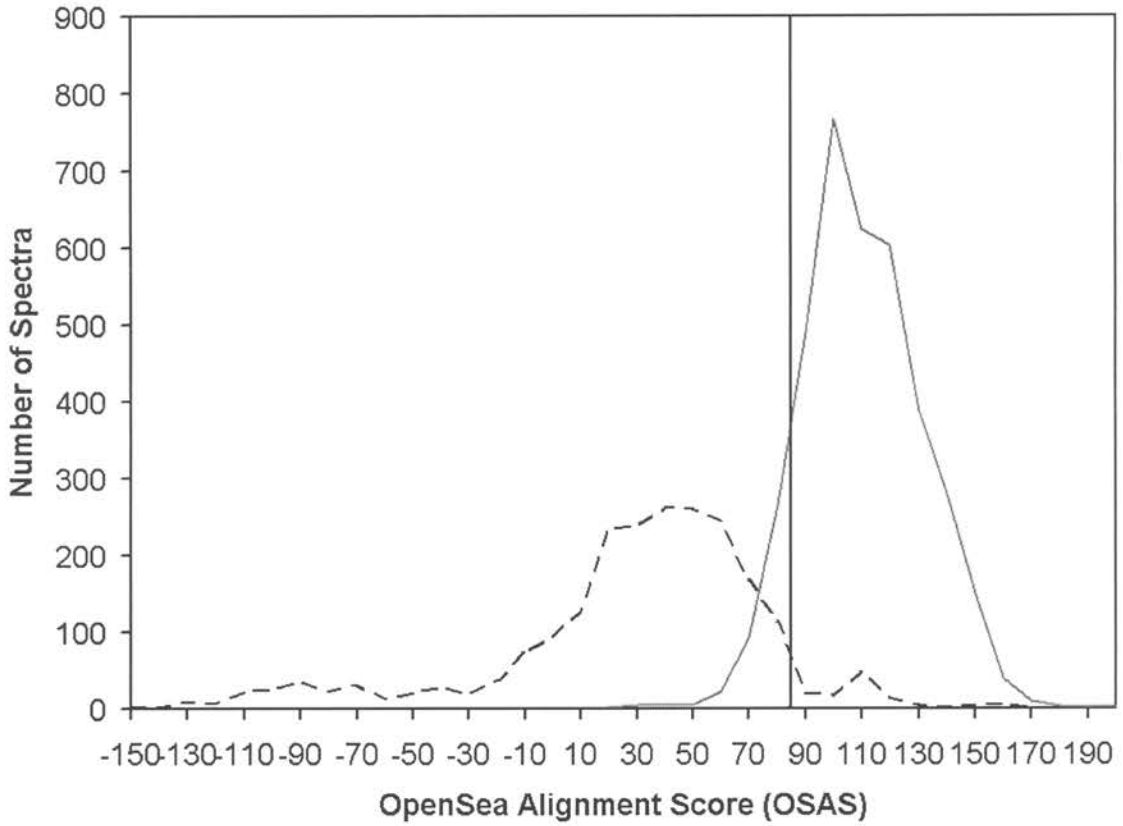


A schematic example of an auto-interpretation procedure, implemented in OpenSea, using a methylated peptide of human gamma crystallin S protein. The mass-based alignment with mass mismatch errors **(a)** is considered and consecutive mass mismatches

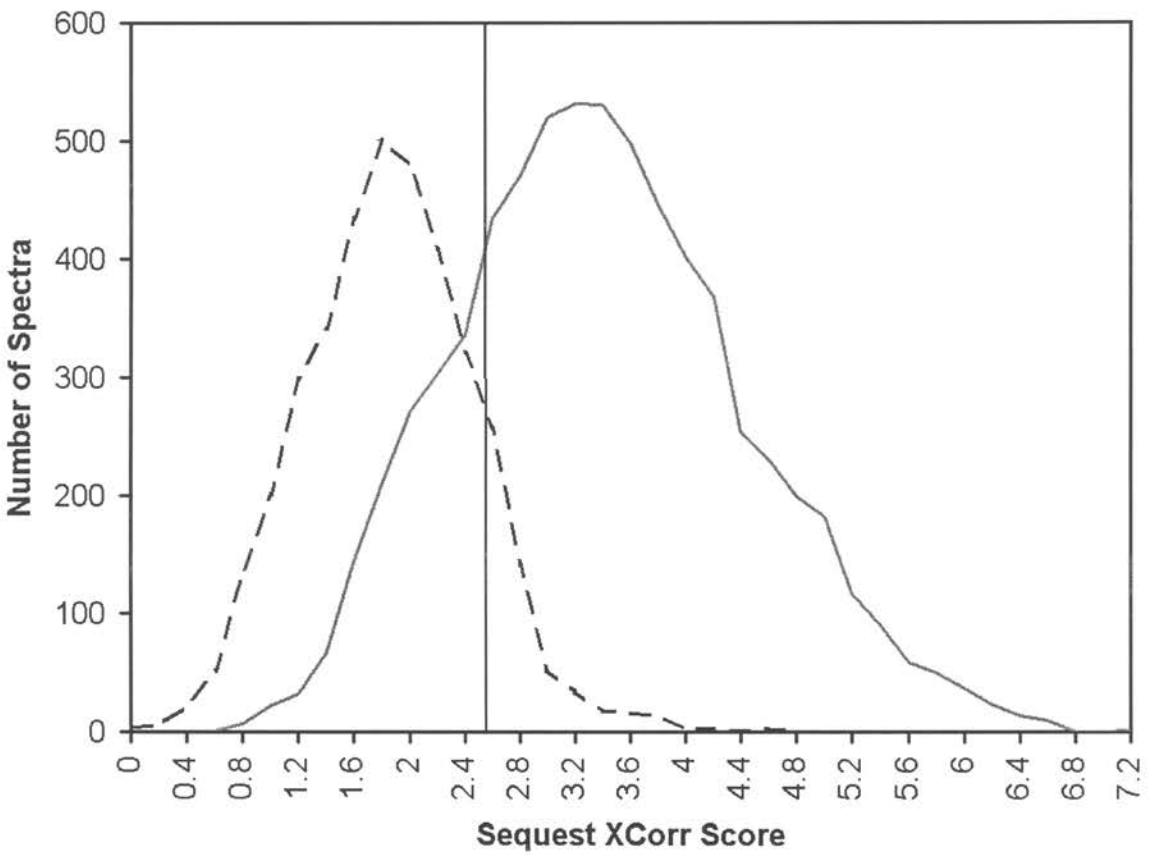
are grouped together as a single mass mismatch **(b).** The local alignment node scores for the alignment before and after the mass mismatch **(a,b)** are also shown in the figure. The mass difference between the database sequence region and the *de novo* sequence region of the alignment nodes are calculated **(b)**. The resulting unexplained mass shift and the database amino acids Cystiene and Aspartic acid are used a determine the possible modification as a methylation on Cystiene, represented as "Cm" in the alignment **(c)**. The ambiguous location of the modification is localized using Smooth Rank Score (SRS) algorithm **(c)**. The resulting alignment **(d)** is rescored as if a new amino acid was identified at the site of the modification using the log odds score for methylation modification.

FIGURE 5

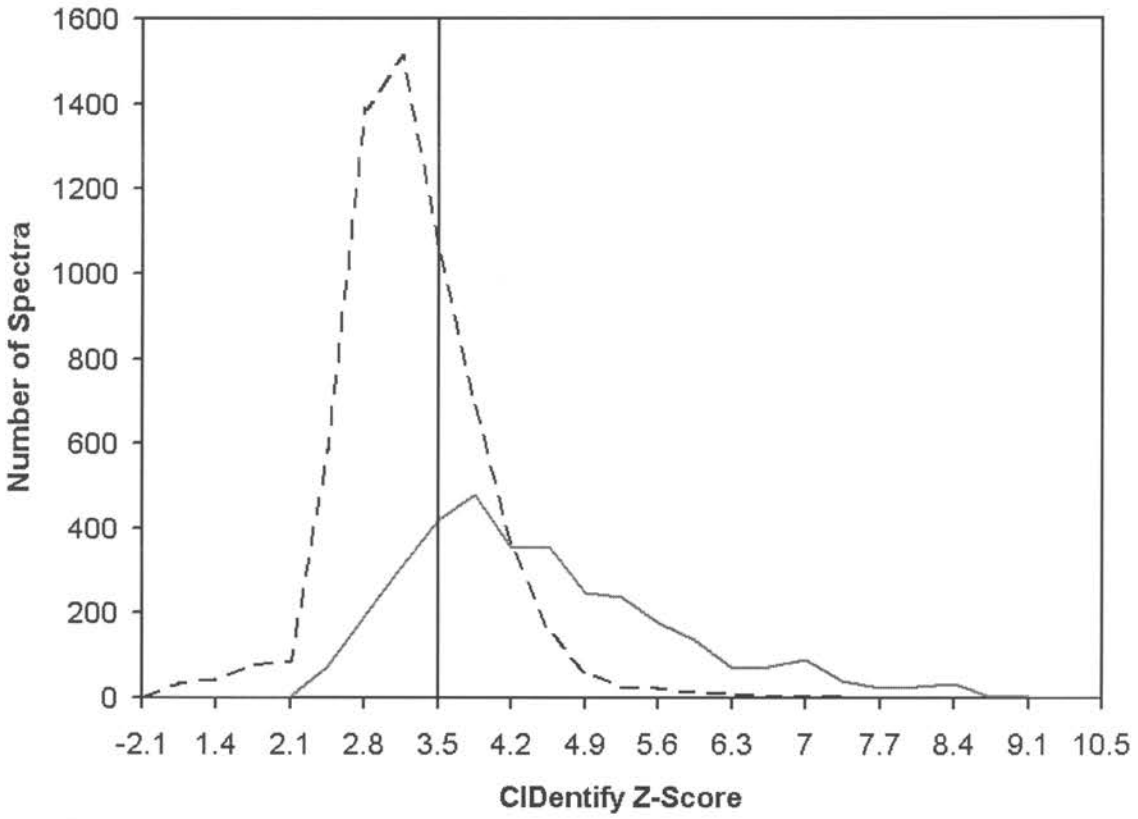
a


900
800
700
600
500
400
300
200
100
0
Number of Spectra
-150 -130 -110 -90 -70 -50 -30 -10 10 30 50 70 90 110 130 150 170 190
OpenSea Alignment Score (OSAS)

b


600
500
400
300
200
100
0
Number of Spectra
0 0.4 0.8 1.2 1.6 2 2.4 2.8 3.2 3.6 4 4.4 4.8 5.2 5.6 6 6.4 6.8 7.2
Sequest XCorr Score

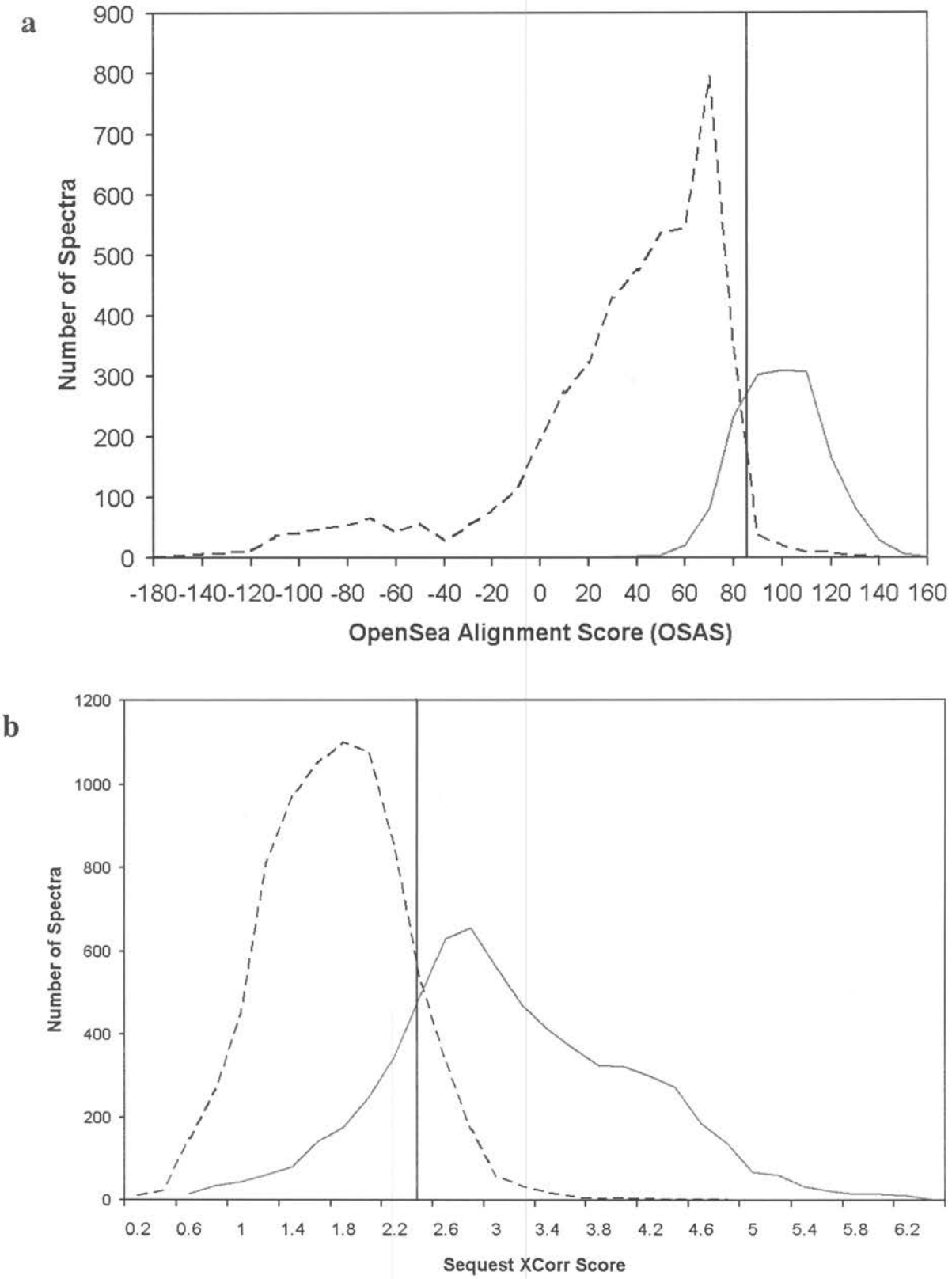
c



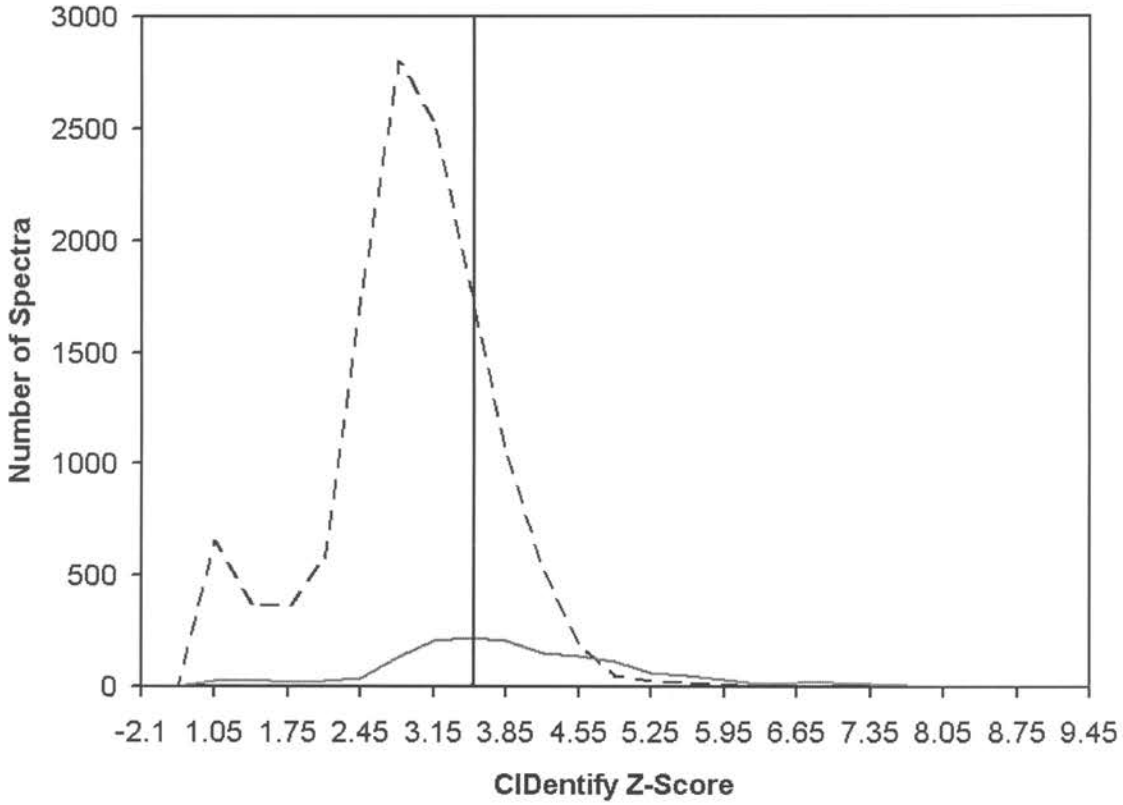
Score distributions for Sequest, CIDentify/Peaks, and OpenSea/Peaks (Mass-based Alignment Score-OSAS) when analyzing 10611 MS/MS spectra from the control protein mixture, acquired on a high mass accuracy instrument (QTOF). Correct spectra identifications were shown in solid line, whereas incorrect spectra identifications are shown in broken line. OpenSea mass-based alignment score (OSAS) distribution **(c)** was obtained by placing the matches in OSAS score bins of width 10. An OSAS score of 85 was used as cutoff as shown in the figure with a vertical solid line. Sequest Xcorr scores **(b)** were normalized for different parent ion charge states (+0.7 for singly charged and -1.0 for triply charged) to reflect the differences of scoring thresholds that were used. The distributions were obtained by placing the matches in Xcorr score bins of 0.2. The normalized Xcorr cutoff is shown in the figure by a vertical solid line. CIDentify score distribution **(c)** was obtained by placing the matches in Z-Score bins of width 0.35. A Z-score of 3.5 was used as a cutoff as shown in the figure with a vertical solid line. The

improved separation of correct and incorrect matches by OSAS score on other techniques helps OpenSea in the accurate identification of low-scoring but correct peptide matches.

FIGURE 6


a
900
800
700
600
500
400
300
200
100
0
Number of Spectra
-180 -140 -120 -100 -80 -60 -40 -20 0 20 40 60 80 100 120 140 160
OpenSea Alignment Score (OSAS)
b
1200
1000
800
600
400
200
0
Number of Spectra
0.2 0.6 1 1.4 1.8 2.2 2.6 3 3.4 3.8 4.2 4.6 5 5.4 5.8 6.2
Sequest XCorr Score

c

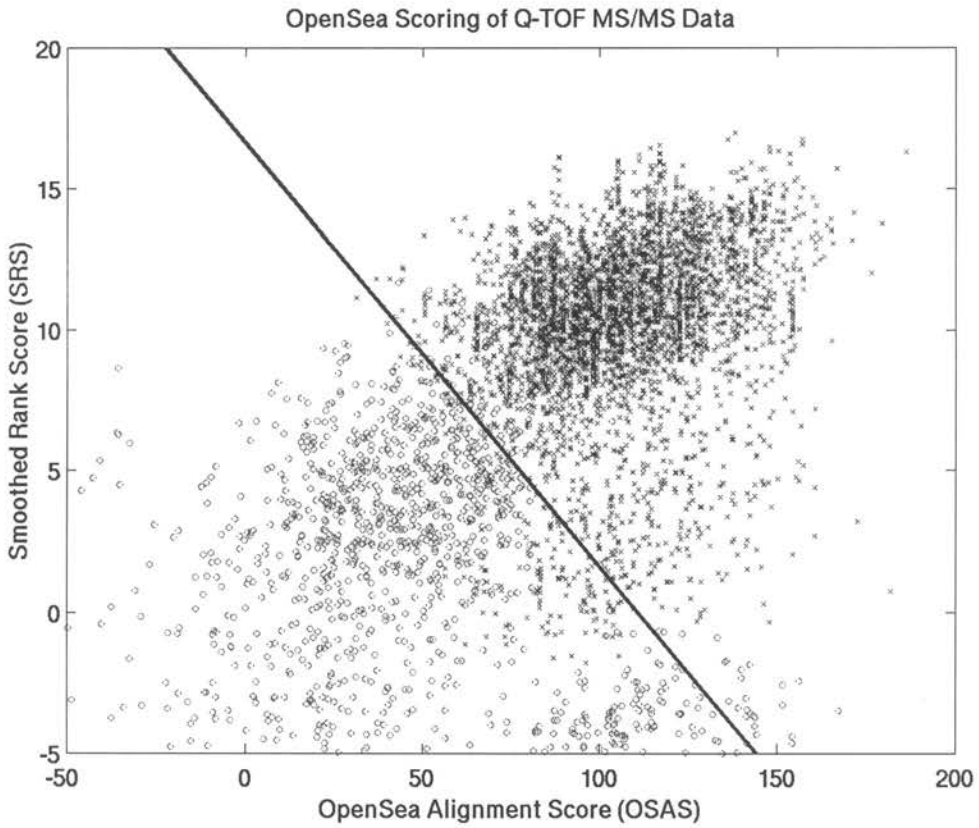


Scoring distributions of Sequest, CIDentify/Peaks and OpenSea/Peaks (OpenSea Mass-based Alignment Score-OSAS) when searching 14469 MS/MS spectra of control protein mixture, acquired on a low mass accuracy instrument (LCQ). Correct spectra identifications were shown in solid line, whereas incorrect spectra identifications are shown in broken line. OpenSea mass-based alignment score (OSAS) distribution **(a)** was obtained by placing the matches in OSAS score bins of width 10. An OSAS score of 85 was used as cutoff as shown in the figure with a vertical solid line. Sequest Xcorr scores **(b)** were normalized for different parent ion charge states (+0.7 for singly charged and -1.0 for triply charged) to reflect the differences of scoring thresholds that were used. The distributions were obtained by placing the matches in Xcorr score bins of 0.2. The normalized Xcorr cutoff is shown in the figure by a vertical solid line. CIDentify score distribution **(c)** was obtained by placing the matches in Z-Score bins of width 0.35. A Z-score of 3.5 was used as a cutoff as shown in the figure with a vertical solid line.

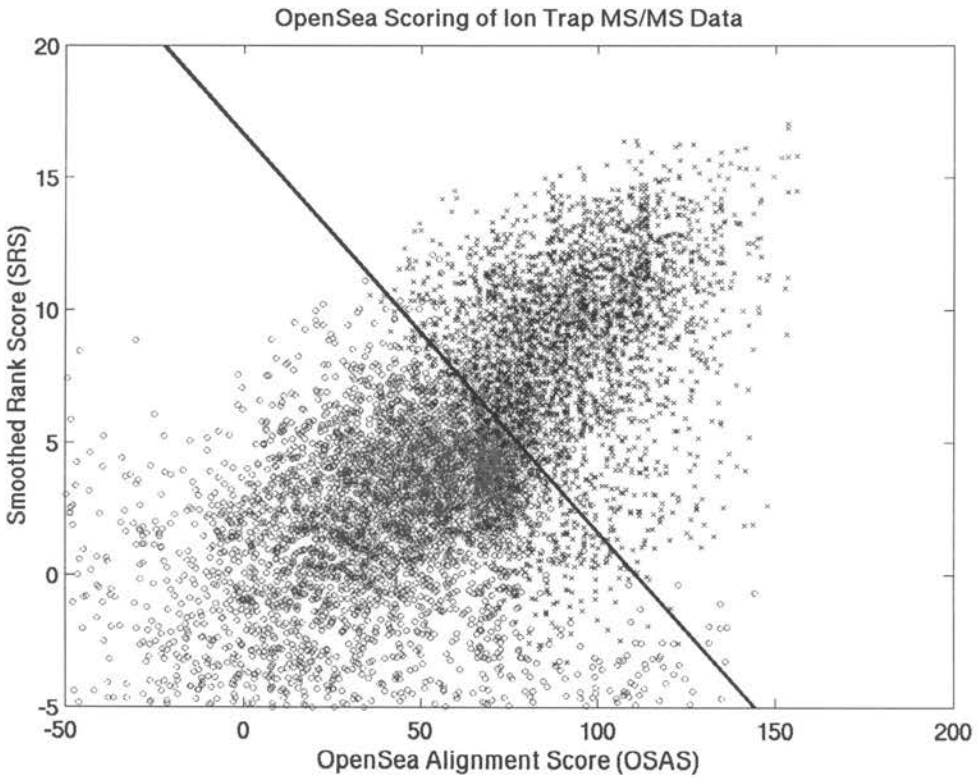
Although the separation of correct and incorrect spectra identifications is not very precise, as with the case of QTOF data, OSAS was able to distinguish the correct spectra identifications from incorrect spectra identifications without any further tuning of the OSAS score for low mass accuracy data (LCQ).

FIGURE 7

a


OpenSea Scoring of Q-TOF MS/MS Data
Smoothed Rank Score (SRS)
OpenSea Alignment Score (OSAS)
20
15
10
5
0
-5
-50
0
50
100
150
200

b


OpenSea Scoring of Ion Trap MS/MS Data
Smoothed Rank Score (SRS)
OpenSea Alignment Score (OSAS)
20
15
10
5
0
-5
-50
0
50
100
150
200

Scoring plots for Combined Alignment Score (CAS) when searching using both high mass accuracy data (10 611 MS/MS spectra from 35 LC/MS/MS runs) and low mass accuracy data (14 469 MS/MS spectra from 20 LC/MS/MS runs). Scatter plots between OSAS and SRS scoring system was obtained by analyzing both QTOF MS/MS data **(a)** and LCQ MS/MS data **(b)** by placing the OSAS sores in bins of width 5.0 and SRS scores in bins of width 0.1. The correct identifications are shown in blue color circles, whereas the incorrect identifications are shown in red color circles. A score threshold of 1.0 for the combined alignment score (CAS) was also shown in the figure with a solid line.