(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 August 2004 (26.08.2004)

PCT

(10) International Publication Number
WO 2004/072103 A2

(51) International Patent Classification⁷: C07K

(21) International Application Number: PCT/US2004/004016

(22) International Filing Date: 11 February 2004 (11.02.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/446,960 11 February 2003 (11.02.2003) US

(71) Applicant *(for all designated States except US)*: **ACTIVX BIOSCIENCES, INC.** [US/US]; 11025 N. Torrey Pines Road, La Jolla, CA 92037 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **PATRICELLI, Matthew, P.** [US/US]; 3004 Comstock Street, San Diego, CA 92111 (US). **ZHANG, Baohong** [CN/US]; 10568 Caminito Alvarez, San Diego, CA 92126 (US).

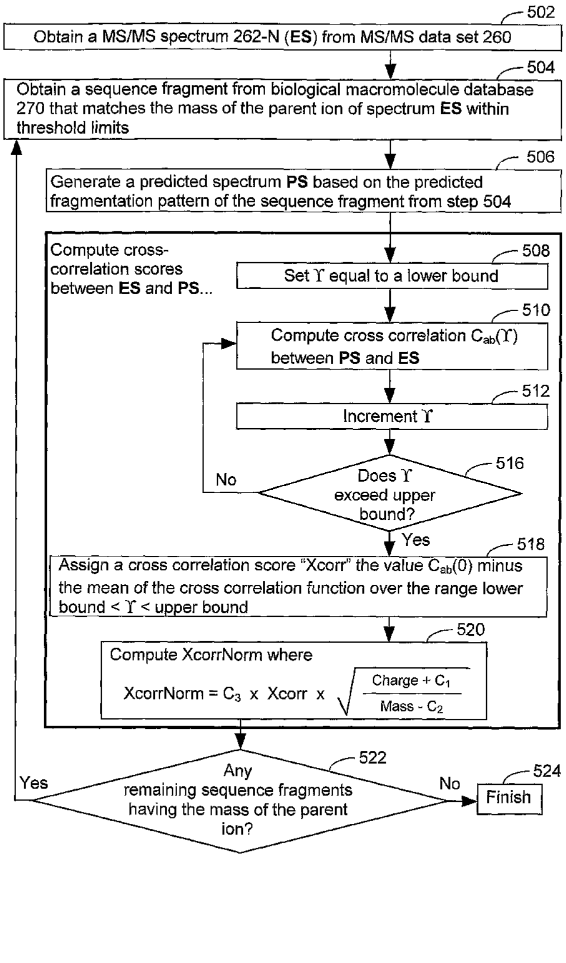
(74) Agents: **ABRAMS, Samuel, B.** et al.; Jones Day, 222 East 41st Street, New York, NY 10017-6702 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR,

[Continued on next page]

(54) Title: MACROMOLECULE IDENTIFICATION MADE BY MASS SPECTROMETRY AND DATABASE SEARCHING


502
Obtain a MS/MS spectrum 262-N (ES) from MS/MS data set 260
504
Obtain a sequence fragment from biological macromolecule database 270 that matches the mass of the parent ion of spectrum ES within threshold limits
506
Generate a predicted spectrum PS based on the predicted fragmentation pattern of the sequence fragment from step 504
Compute cross-correlation scores between ES and PS...
508
Set ϒ equal to a lower bound
510
Compute cross correlation $C_{ab}(ϒ)$ between PS and ES
512
Increment ϒ
516
Does ϒ exceed upper bound?
No
Yes
518
Assign a cross correlation score "Xcorr" the value $C_{ab}(0)$ minus the mean of the cross correlation function over the range lower bound < ϒ < upper bound
520
Compute XcorrNorm where
$XcorrNorm = C_3 \times Xcorr \times \sqrt{\frac{Charge + C_1}{Mass - C_2}}$
522
Any remaining sequence fragments having the mass of the parent ion?
Yes
No
524
Finish

(57) Abstract: A method of determining the parent ion charge state of a tandem MS spectrum. The parent ion is labeled with a probe that is cleaved as a result of the MS. A candidate parent ion charge state is chosen and the spectrum is searched for a peak having a value F, where F is the ratio between (i) the mass of the parent ion minus the mass of the cleaved probe and (ii) the difference between the candidate parent ion charge state and the charge of the cleaved probe. A method of removing a parent ion mass dependence and a parent ion charge dependence from a cross correlation score between a predicted spectrum and an experimental spectrum. A method of computing a false positive gamma distribution curve for a data set of tandem MS spectra. A method of reducing a number of false positive assignments made to tandem microcopy spectra in a tandem MS data set. In this method, the spectra are processed in a manner that is based on the presence or absence of peak fragments within the spectra.

WO 2004/072103 A2

GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

MACROMOLECULE IDENTIFICATION MADE BY MASS SPECTROMETRY AND DATABASE SEARCHING

CROSS REFERENCE TO RELATED APPLICATION

This application claims priority to United States Application no. 60/446,960, filed February 11, 2003 which is incorporated by reference herein in its entirety.

1. FIELD OF THE INVENTION

This invention pertains to systems and methods for assigning tandem mass (MS/MS) spectra to peptides using database search applications.

2. BACKGROUND OF THE INVENTION

Advances in activity-based profiling and screening as well as other technologies rely on the accurate high-throughput identification of peptides or other macromolecules from samples. Such samples typically include a complex mixture of proteins, nucleic acids, carbohydrates, or other biological macromolecules. Tandem mass spectrometry (MS/MS) has been particularly useful for determining the protein or nucleic acid components of complex mixtures. (See Link *et al.*, 1999, Nat. Biotechnol. 17, 676-682; Washburn, *et al.* 2001, Nat. Biotechnol. 19, 242-247; Gaven *et al.*, 2002, Nature 415, 141-147; and Ho *et al.*, 2002, Nature 418, 180-183).

In the case of where the sample is a proteome from an organism, the proteins in the proteome are typically first digested into peptides using an enzyme such as trypsin and then subjected to liquid chromatography tandem mass spectrometry (LCMS/MS). Liquid chromatography (LC) provides an initial separation of the peptides, which are then ionized directly into a mass spectrometer. Following an initial scan in which the mass/charge ratio of all intact (parent) ions from the peptides are measured, the mass spectrometer selects a parent ion, fragments it and obtains the mass spectrum of the generated fragments. These fragmentation patterns are called tandem mass spectra or MS/MS spectra. This process of ion selection and fragmentation is repeated throughout the LC separation, thus generating a set of time resolved MS/MS spectra, with each spectrum representing a species eluting at a particular time from the LC separation. The resolving power of the liquid chromatography step, combined with the high mass resolution of modern mass spectrometers typically assures that each MS/MS spectrum represents the fragmentation pattern of a unique peptide in the digest.

A typical MS/MS data set contains on the order of a thousand MS/MS spectra. Each spectra is collected from a different chromatography fraction of a given sample. Ideally, each spectra corresponds to a parent ion generated from a unique fragment of a protein or other biological macromolecule that is present in the sample. For proteins, the information obtained in each MS/MS spectrum, together with the charge/mass ratio of the selected parent ion can be used to determine the sequence or partial sequence of the analyzed peptide. Typically, any sequence of greater than approximately six amino acids will be unique to one particular protein, thus the data obtained in an LCMS/MS analysis can be used to identify the proteins present in the original samples. For example, consider the case in which a protein "A" is present in a sample. The sample is subjected to enzymatic degradation yielding a solution that includes two fragments of protein A, fragment A_1 and fragment A_2. The digested solution is subjected to liquid chromatography and peaks in the chromatographic eluent are subjected to tandem mass spectrometry to yield MS/MS spectra. Based on the sequences determined from the MS/MS spectra of fragments A_1 and/or A_2, the identity of the protein A can be determined. Because of the large number of spectra in a typical data set, highly automated techniques are needed to identify the biological macromolecule fragment corresponding to each of the MS/MS spectra. A biological macromolecule fragment corresponds to a particular MS/MS spectrum when the MS/MS spectrum represents the fragmentation pattern of that biological macromolecule fragment.

In one known approach, an assignment is made to an experimental MS/MS spectrum **ES** in a tandem MS data set by comparing the experimental spectrum against a sequence database to find the best matching database entry. (See, for example, Fenyo, 2000, Curr. Opin. Biotechnol. 11, 391-395). Several programs are available for performing such sequence database searches in an automated fashion, including SEQUEST (Eng *et al.*, 1994, J. Am. Soc. Mass Spectrom. 5, 976-989), Mascot (Perkins *et al.*, 1999, Electrophoresis 20, 3551-3567), and Sonar (Field *et al.*, 2002, Proteomics 2, 36-47). In the case of proteins, these applications first identify each possible peptide in a database of proteins that has a mass that is the same as the predicted mass of the parent ion (within error limits) of the experimental MS/MS spectrum **ES**. Then, the predicted spectrum, **PS**, of each peptide in this set of possible peptides is compared with the experimental MS/MS spectrum **ES**. The peptide producing the best match is assigned to the experimental MS/MS spectrum **ES**.

To illustrate how these known programs work, consider the case in which the parent ion mass for a given experimental MS/MS spectrum **ES** in the MS/MS data set is two thousand daltons. Each peptide in a protein database having a mass of two thousand daltons is identified. In this illustration, the database consists of protein sequences. To identify each peptide in the protein database having a mass of two thousand daltons, each protein sequence in the database is scanned for a sequence fragment (subsequence of the protein) that has a calculated molecular weight of two thousand daltons, within error limits. Each sequence fragment within a protein in the database that has such a molecular weight is considered a candidate peptide sequence. Each candidate peptide sequence is evaluated by generating a predicted spectrum **PS** of the candidate peptide sequence based on expected fragmentation patterns of the candidate peptide. Then, each candidate peptide sequence is scored. The scoring process involves correlating the predicted spectrum **PS** with that of the actual measured MS/MS spectrum (**ES**). Each MS/MS spectrum (**ES**) in the tandem MS data set is assigned to the peptide (sequence fragment) from the database that has the best correlation score. This correlation score reflects the fit between the experimental MS/MS spectrum **ES** and the predicted spectrum **PS** of the peptide.

The scores computed by known automated programs help discriminate between correct and incorrect assignments to MS/MS spectra and therefore have some ability to facilitate detection of false positives. A false positive (false identification) arises when the peptide identified by the sequence database search is not the true identity of the parent ion associated with the MS/MS spectrum.

While current techniques for MS/MS spectral assignments and verification of such assignments are functional in practice, they are unsatisfactory. First, the spectrometer used to collect MS/MS spectra cannot determine the parent ion charge state of each spectrum. For example, the parent ion charge state for any given MS/MS spectrum could be +1, +2, +3, +4, +5, +6, or greater. In fact, depending upon whether an acid or base is used in the chromatographic step, the parent ion charge state for any given MS/MS spectrum could be either positive or negative. Knowledge of the parent ion charge state is important because the highest charge state a fragment of the parent ion can have (or the lowest in the case of negative parent ion charge states) in the MS/MS spectra is the charge state of the parent ion. For example, if the parent ion has a charge state of +3, each fragment of the parent ion found in the MS/MS spectrum of the parent ion can have a possible charge state of +3, +2, or +1.

It is not possible to determine, *a priori*, the charge state of each of the fragments in a MS/MS spectrum because the MS/MS spectrum only provides mass to charge state ratio information (m/z ratios), not absolute masses (m) or charge states (z). Therefore, if a MS/MS spectrum indicates that the parent ion has a given m/z ratio (*e.g.*, m/z = 1000) and the charge state of the parent ion is not known, several searches must be performed. For example, a +2 search must be performed. In the +2 search, a search for peptides in the protein database that have a mass of roughly twice the m/z ratio of the parent ion (*e.g.*, 2 x 1000) is performed and the theoretical fragmentation patterns of each of these peptides is compared to the MS/MS spectrum. A +3 must also be performed. In a +3 search, a search for peptides in the database with a mass of three times the m/z ratio of the parent ion (*e.g.*, 3 x 1000) is performed and the theoretical fragmentation patterns of each of these peptides is compared to the MS/MS spectrum. For completeness, other parent ion charge states may be searched as well.

As discussed above, in the case of proteins, a completely different set of peptides from a database is compared to the MS/MS spectra for each possible parent ion charge state. This is time consuming. Furthermore, comparisons that are performed based on the incorrect parent ion charge state will result in the incorrect assignment of the MS/MS spectrum (a false positive). What is needed in the art is a way to determine the parent ion charge state. Knowledge of the parent ion charge state would eliminate needless searches and remove a source of false positives.

Consider the case of a MS/MS spectrum in which the parent ion has an m/z of 1000. If the parent ion charge state could be determined, then only a single search would have to be performed. For example, in the case where the parent ion has a m/z of 1000, if the parent ion charge state was known to be +2, then only one search of the database would need to be performed (one for macromolecules having a molecular weight of 2000 daltons).

Another problem with prior art techniques is that the scoring function used to score the correspondence between a predicted spectrum **PS** generated from a library peptide sequence to an actual MS/MS spectrum **ES** is influenced by the peptide charge and mass. Fig. 1 illustrates the problem with Xcorr, which is a correlation score between a predicted spectrum **PS** and an experimental spectrum **ES** that is calculated using the program SEQUEST (Eng *et al.*, 1994, J. Am. Soc. Mass Spectrom. 5, 976-989). Fig. 1 illustrates the analysis of a complete tandem MS data set of MS/MS spectra using SEQUEST. SEQUEST compares predicted peptide fragmentation patterns (predicted spectrum **PS**) against experimentally derived MS/MS data **ES** and assigns the library peptide having the highest correlation score to each MS/MS spectrum. A SEQUEST search was performed for each MS/MS spectrum in the tandem MS data set assuming a parent ion charge state of +3. The SEQUEST searches were repeated assuming a parent ion charge state of +4 and repeated again assuming a parent ion charge state of +5. The highest correlation score from each SEQUEST search is plotted in Fig. 1. While the Xcorr correlation value has proven useful in general, the score is highly influenced by the charge and mass of the peptide. As is evident from Fig. 1, larger peptides in general give higher Xcorr values, regardless of whether the identification is correct or false. Furthermore, the slope of the correlation decreases as the parent ion charge state increase, indicating that Xcorr has a charge dependence. This can be seen from the line fitted to the +3 analysis (line 102), the +4 analysis (line 104), and the +5 analysis (line 106). Due to the dependence of Xcorr on peptide mass and parent ion charge state, as illustrated in Fig. 1, it is difficult to determine uniform score thresholds that would allow for a statistical determination of the confidence of any given Xcorr score. Therefore it is difficult to make a statistical determination that any given assignment in Fig. 1 is, in fact, a correct assignment as opposed to a false positive. Thus, what is needed in the art are methods for removing the peptide mass and parent ion charge state dependency that is found in known mass spectra correlation algorithms.

Another limitation of known methods for comparing the MS/MS spectra against a sequence database is that known statistical methods used to determine the confidence in a given correlation score between a predicted spectrum **PS** and an actual MS/MS spectrum **ES** are unsatisfactory. In part, this is due to the influence on parent ion charge state and peptide mass as illustrated in Fig. 1. However, this is also due to a lack of knowledge about the shape of the distribution of false positives in any given tandem MS data set, such as the one illustrated in Fig. 1.

Thus, given the above background, methods for determining the parent ion charge state for a given MS/MS spectrum are needed in order to reduce the number of database searches that are performed and thereby reduce the number of false positive identifications made in a given data set. Furthermore, what is needed in the art are improved correlation functions that are not influenced by the parent ion charge state and by the size of the parent ion. Furthermore, what is needed in the art is improved statistical methods for determining a confidence value for correlation scores between predicted spectra **PS** and experimental MS/MS spectra **ES**.

3. SUMMARY OF THE INVENTION

The present invention addresses the shortcomings of the known art. The present invention provides novel methods for determining the parent ion charge state of experimental tandem MS spectra **ES**. Furthermore, the present invention provides novel methods for computing a correlation value between a predicted spectra **PS** and an actual experimental MS/MS spectra **ES**. Advantageously, the correlation values of the present invention are not dependent upon the mass of the parent ion or the parent ion charge state. The present invention also provides novel methods for determining a confidence value for correlation scores between predicted spectra and actual MS/MS spectra.

A first aspect of the present invention provides a method of determining a parent ion charge state for a tandem MS spectrum of the parent ion. In this aspect of the invention, the parent ion has been labeled with a probe. In the method, a candidate parent ion charge state is chosen. Then, the tandem MS spectrum is searched for a peak having the value **F**, where

F = [(the m/z of the parent ion) x (the candidate parent ion charge state) – (the mass of the probe) – (the mass of a portion of the parent ion that is removed upon probe cleavage)- (the mass of any protons carried on the probe)] / [(the candidate parent ion charge state) – (the carried charge state of the probe)].

When a peak appears in the tandem MS spectrum that is within a predetermined threshold value of **F**, the candidate parent ion charge state is considered the parent ion charge state for the tandem MS spectrum. In some embodiments, the predetermined threshold value is less than ±3 daltons, less than ±2 daltons, less than ±1 daltons, less than ±0.05 daltons, or less than ±0.01 daltons. In some embodiments, the probe is a fluorophosphonate and the parent ion is a peptide. Representative candidate parent ion charge states include, but are not limited to, +2, +3, +4, +5, +6, +7, +8, +9, -2, -3, -4, -5, -6, -7, -8, and -9.

As used herein, the term "probe" references a broad range of reactive chemical moieties. Such probes can, for example, be reacted with a protein mixture to label many proteins the mixture in a non-specific, or non-directed, manner providing a quantitative analysis only of protein abundance. See, for example, Aebersold, PCT/US99/19415, which discloses that there are many chemically reactive amino acid residues within a protein that are individually reactive and that can be conjugated with chemical probes to produce protein conjugates that can be quantified to yield an indication of protein abundance in a mixture of proteins. See also, for example, Wells *et al.* (PCT/US99/14267; PCT/US98/21759) which discloses methods for identifying small organic molecule ligands that bind to biological target molecules without the requirement that the ligand bind to an active site on the target molecule.

The term "probe" as used herein further encompasses "activity-based probes" or "ABPs". ABPs are molecules with a binding moiety that are directed to the active site of a given protein class (*e.g.*, serine proteases) and linked to a tag (*e.g.*, a biotin tag). ABPs are capable of differentiating active member of a protein class in a proteome from inactive members. See, *e.g.*, Liu *et al.*, Proc. Natl. Acad. Sci. USA 96: 14694-14699 (1999); Cravatt and Sorensen, Curr Opin. Chem. Biol. 4, 663-668 (2000); Patricelli *et al.*, Proteomics 1, 1067-1071 (2001); and PCT/US02/06234. The term "probe" as used herein further encompasses proves such as the adenine nucleotide-binding protein-directed affinity probes" or "ANBPs" disclosed in United States Patent application 20030134303.

The term "probe" as used herein further encompasses fluorescent inhibitors such as those used by Scholze *et al.*, Anal. Biochem. 276: 72-80 (1999) to analyze lipases, the enzyme-activated irreversible inhibitor of ornithine decarboxylase linked to a rhodamine moiety used in U.S. Patent No. 4,433,051 for cytochemical staining procedures, the fluorescent compounds used in U.S. Patent No. 6,127,134 for analysis of protein mixtures, the fluorescent activity based probes ("fABPs") used in PCT/US02/03808 for the analysis of one or more active protein components of proteomes, and the tether activity-based probes ("tABPs") used in PCT/US03/07898 to analyze protein mixtures. The term "probe" as used herein further encompasses those chemicals or tags that exploit the selective reactivity of an amino acid residue to derivatization with a light and heavy form of the chemical or tag. For example, cysteine residues in proteins have been labeled using the light ($^1H_8$) and heavy ($^2H_8$) forms of the ICAT reagent [Gygi (1999) Nat. Biotechnol. 17, 994-999], or with the light ($1H_3$) and heavy ($2H_3$) forms of acrylamide [Sechi, S. (2002) Rapid Commun. Mass Spectrom. 16, 1416-1424]. See also, for example, United States Patent Publication 20040009567.

A second aspect of the present invention provides a computer program product for use in conjunction with a computer system. The computer program product comprises a computer readable storage medium and a computer program mechanism embedded therein. The computer program mechanism comprises (i) an experimental MS/MS tandem MS data set comprising a plurality of tandem MS spectra, (ii) a biological macromolecule database for storing biological macromolecule information and (iii) a tandem MS processing module for assigning tandem MS spectra to sequence fragments in the biological macromolecule database. The tandem MS processing module includes a charge state determination module for determining a parent ion charge state for a first tandem MS spectrum of the parent ion. The first tandem MS spectrum is in the experimental MS/MS data set. Furthermore, the parent ion has been labeled with a probe.

In this second aspect of the invention, the tandem MS processing module comprises instructions, including instructions for choosing a candidate parent ion charge state and instructions for searching the first tandem MS spectrum for a peak having the value **F**, where **F** is the same as that given in the first aspect of the present invention.

A third aspect of the present invention provides a computer system for processing tandem MS data. In this aspect of the invention, the computer system comprises a central processing unit and a memory that is coupled to the central processing unit. The memory stores an experimental MS/MS data set. The data set comprises a plurality of tandem MS spectra. The memory also stores a biological macromolecule database for storing biological macromolecule information. In addition, the memory stores a tandem MS processing module for assigning tandem MS spectra to sequence fragments in the biological macromolecule database. The tandem MS processing module includes a charge state determination module for determining a parent ion charge state for a first tandem MS spectrum of the parent ion. The first tandem MS spectrum is in the experimental MS/MS data set. Furthermore, the parent ion has been labeled with a probe. The tandem MS processing module comprises instructions for choosing a candidate parent ion charge state and instructions for searching the first tandem MS spectrum for a peak having the value **F**, where **F** is given in the first aspect of the invention.

A fourth aspect of the invention provides a method of removing a parent ion mass dependence and a parent ion charge dependence from a cross correlation score between a predicted spectrum **PS** and an experimental spectrum **ES**. The experimental spectrum **ES** is the fragmentation pattern of a parent ion. The method comprises computing a value XcorrNorm, where

$$\text{XcorrNorm} = C_3 \times \text{Xcorr} \times \sqrt{\frac{\text{Charge} + C_1}{\text{Mass} - C_2}}$$

and Xcorr is the cross correlation score between **PS** and **ES**, C_1, C_2, and C_3 are constants, "Charge" is the parent ion charge state, and "Mass" is the mass of the parent ion. In some embodiments, Xcorr is computed by (i) setting a relative displacement value $n\Delta t$ to a lower bound, (ii) computing a score $C_{ab}(n\Delta t)$ in which

$$C_{ab}(n\Delta t) = \frac{1}{T}\sum_{t=0}^{T} \text{PS}(t)\text{ES}(t \pm n\Delta t)$$

Δt is a sampling interval, (iii) incrementing $n\Delta t$; and (iv) repeating steps (ii) and (iii) until $n\Delta t$ exceeds an upper bound, and (v) assigning Xcorr the value of $C_{ab}(0)$ minus the mean of $C_{ab}(n\Delta t)$ over the range lower bound < $n\Delta t$ < upper bound.

A fifth aspect of the invention provides a computer program product for use in conjunction with a computer system. The computer program product comprises a computer readable storage medium and a computer program mechanism embedded therein. The computer program mechanism comprises an experimental MS/MS data set that includes a plurality of tandem MS spectra. The computer program mechanism further includes a biological macromolecule database for storing biological macromolecule information as well as a tandem MS processing module for assigning tandem MS spectra to sequence fragments in the biological macromolecule database. The tandem MS processing module includes a correlation determination module for removing a parent ion mass dependence and a parent ion charge dependence from a cross correlation score between a predicted spectrum **PS** and an experimental spectrum **ES**. In this aspect of the invention, the experimental spectrum **ES** is the fragmentation pattern of a parent ion. The experimental spectrum **ES** is in the experimental MS/MS data set. The correlation determination module has instructions for computing a value XcorrNorm, where XcorrNorm is as defined in the fourth aspect of the present invention.

A sixth aspect of the invention provides a computer system for processing tandem MS information. The computer system comprises a central processing unit and a memory coupled to the central processing unit. The memory stores (i) an experimental MS/MS data set comprising a plurality of tandem MS spectra, (ii) a biological macromolecule database that includes biological macromolecule information, and (iii) a tandem MS processing module for assigning tandem MS spectra to sequence fragments in the biological macromolecule database. The tandem MS processing module includes a correlation determination module for removing a parent ion mass dependence and a parent ion charge dependence from a cross correlation score between a predicted spectrum **PS** and an experimental spectrum **ES**. The experimental spectrum **ES** is the fragmentation pattern of a parent ion. The experimental spectrum **ES** is in the experimental MS/MS data set. The correlation determination module comprises instructions for computing a value XcorrNorm, where XcorrNorm is as defined in the fourth aspect of the present invention.

A seventh aspect of the invention provides a method of computing a false positive gamma distribution curve for a data set of MS/MS spectra. All or a portion of the MS/MS spectra in the data set correspond to a different parent ion that has been labeled with a probe. The method includes assigning, for each MS/MS spectrum in all or a portion of the data set, a sequence fragment from a biological macromolecule database to the MS/MS spectrum using a search parameter that includes a first incorrect mass for the probe. The method also includes the step of computing, for each MS/MS spectrum in all or a portion of the data set, a cross correlation score using the assignment made for the MS/MS spectrum in the assigning step. Further, the method includes the step of fitting each cross correlation score calculated in the computing step to a gamma distribution curve thereby computing a false positive gamma distribution curve for a data set of MS/MS spectra.

In some embodiments in accordance with the seventh aspect of the invention, the method further includes assigning, for each MS/MS spectrum in said the set, a sequence fragment from the biological macromolecule database to the MS/MS spectrum. The search for this assignment uses a search parameter that includes a second incorrect mass for the probe. Cross correlation scores are computed using these new assignments to form a second set of cross correlation scores. In this embodiment, the first set and the second set of cross correlation scores is fitted to a gamma distribution curve.

An eight aspect of the present invention provides a computer program product for use in conjunction with a computer system. The computer program product comprises a computer readable storage medium and a computer program mechanism embedded therein. The computer program mechanism comprises an experimental MS/MS data set, a biological macromolecule database for storing biological macromolecule information, and a tandem MS processing module for assigning tandem MS spectra to sequence fragments in the biological macromolecule database. In this aspect of the invention, the tandem MS processing module includes a false data profiling module for computing a false positive gamma distribution curve for the experimental MS/MS data set. All or a portion of the MS/MS spectra in the data set correspond to a different parent ion that has been labeled with a probe. The false data profiling module comprises instructions for assigning, for each MS/MS spectrum in all or a portion of the experimental data set, a sequence fragment from the biological macromolecule database to the MS/MS spectrum using a search parameter that includes a first incorrect mass for the probe. Furthermore, the module comprises instructions for computing, for each MS/MS spectrum in all or a portion of the data set, a cross correlation score using the assignments made for the MS/MS spectrum by the instructions for assigning. The module also includes instructions for fitting each cross correlation score computed by the instructions for computing to a gamma distribution curve.

A ninth aspect of the invention provides a computer system for processing tandem MS data. The computer system comprises a central processing unit and a memory coupled to the central processing unit. The memory stores an experimental MS/MS data set comprising a plurality of tandem MS spectra, a biological macromolecule database, and a tandem MS processing module. The tandem MS processing module includes a false data profiling module for computing a false positive gamma distribution curve for the experimental MS/MS data set. All or a portion of the MS/MS spectra in the data set respectively correspond to a different parent ion that has been labeled with a probe. The false data profiling module includes instructions for assigning, for each MS/MS spectrum in all or a portion of the experimental data set, a sequence fragment from the biological macromolecule database to the MS/MS spectrum using a search parameter that includes a first incorrect mass for the probe. The false data profiling module also includes instructions for computing, for each MS/MS spectrum in all or a portion of the data set, a cross correlation score using the assignment made for the MS/MS spectrum by the instructions for assigning. Furthermore, the false data profiling module includes instructions for fitting each cross correlation score computed by the instructions for computing to a gamma distribution curve.

A tenth aspect of the invention provides a method of reducing a number of false positive assignments made to MS/MS spectra in a tandem MS data set. Each MS/MS spectrum in all or a portion of the MS/MS spectra in the tandem MS data set represents a fragmentation pattern of a different parent ion that has been labeled with a probe. In the method, a plurality of files for a MS/MS spectrum in the tandem MS data set is generated. Each file in the plurality of files includes the peaks from the MS/MS spectrum and each file in the plurality of files represents a different candidate parent ion charge state for the MS/MS spectrum. For each file in the plurality of files, a determination is made as to whether the file includes a peak corresponding to a fragment of the probe. When a file does not contain the peak corresponding to the fragment of the probe, the file is given a first designation and removed from the plurality of files. The method also includes the step of determining, for a first file in the plurality of files, whether the first file has a valid parent ion charge state.

When another file in the plurality of files has a valid parent ion charge state and the first file does not have a valid parent ion charge state, the first file is deleted and removed from the plurality of files. The method further includes the step of giving the first file a second designation when the file remains in the plurality of files. A biological macromolecule database is searched using a file given the second designation with the criterion that the probe modified a predetermined moiety class. In some embodiments, the predetermined moiety class is serine.

An eleventh aspect of the invention provides a computer program product for use in conjunction with a computer system. The computer program product comprises a computer readable storage medium and a computer program mechanism embedded therein. The computer program mechanism comprises a module capable of executing the method described in the tenth aspect of the invention.

A twelfth aspect of the invention provides a computer system for processing tandem MS data. The computer system includes a central processing unit and a memory that is coupled to the central processing unit. The memory stores an experimental MS/MS data set comprising a plurality of tandem MS spectra, a biological macromolecule database and a tandem MS processing module. The tandem MS processing module includes instructions for carrying out the method described in the tenth aspect of the invention.

4. BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates SEQUEST Xcorr values for each assignment made to tandem mass spectrometry data collected from a digested protein sample in accordance with the prior art.

Fig. 2 illustrates a computer system for database searching with tandem mass spectrometry data and for statistical analysis of such search data in accordance with one embodiment of the present invention.

Fig. 3 illustrates an MS/MS spectrum of the probe FP-Peg-TAMRA.

Fig. 4 illustrates the processing steps that are performed in order to determine the parent ion charge state for a given MS/MS spectrum **ES** in accordance with one embodiment of the present invention.

Fig. 5 illustrates the processing steps that are used to perform a cross correlation between a predicted mass spectrum **PS** and an experimentally determined mass spectrum **ES** in accordance with one embodiment of the present invention.

Fig. 6 illustrates an example of an experimental spectrum **ES** than can be evaluated in order to determine the parent ion charge state for the parent ion of the spectrum in accordance with one embodiment of the present invention.

Fig. 7A is a scatter plot of cross correlation values for a tandem mass spectrometry data set that were computed in accordance with the prior art.

Fig. 7B is a scatter plot of cross correlation values for a tandem mass spectrometry data set that were computed in accordance with one embodiment of the present invention.

Fig. 8 illustrates the distribution of XcorrNorm′ scores for a representative tandem MS data set in accordance with one embodiment of the present invention.

Fig. 9 illustrates the distribution of XcorrNorm′ scores for incorrectly assigned experimental spectra **ES** in a tandem MS data set in accordance with one embodiment of the present invention.

Fig. 10 illustrates the process steps that are used to produce a false positive gamma distribution curve fit that takes into account the type of probe attached to the parent ion, the residue modified by the probe, and biological macromolecule database size in accordance with one embodiment of the present invention.

Fig. 11 illustrates a false positive gamma distribution curve that has been fitted to incorrect cross correlation scores computed using mass differentials of ±10 daltons away from the actual mass of the parent ion of each experimental MS/MS spectra **ES** in a tandem mass spectrometry data set in accordance with one embodiment of the present invention.

Fig. 12 illustrates a false positive gamma distribution curve superimposed on Xcorr cross correlation values computed for a digested proteome. The proteome was labeled with a probe that modifies serine residues. The correct mass for the probe is used as input to SEQUEST.

Fig. 13 illustrates a false positive gamma distribution curve superimposed on XcorrNorm′ cross correlation values for a digested proteome. The proteome has been labeled with a probe that modifies serine residues. The correct mass for the probe is used as input to SEQUEST to generate the XcorrNorm′ cross correlation values.

Fig. 14 illustrates a false positive gamma distribution curve superimposed on Xcorr cross correlation values for a digested proteome. The proteome has been labeled with a probe that modifies cysteine residues. The correct mass for the probe is used as input to SEQUEST to generate the Xcorr cross correlation values.

Fig. 15 illustrates a false positive gamma distribution curve superimposed on XcorrNorm′ cross correlation values for a digested proteome. The proteome has been labeled with a probe that modifies cysteine residues. The correct mass for the probe is used as input to SEQUEST to generate the XcorrNorm′ cross correlation values.

Fig. 16A illustrates the fluorophosphonate probe FP-Peg-TAMRA.

Fig. 16B illustrates the fluorophosphonate probe FP-Peg-TAMRA complexed with a peptide to form a phosphate/serine adduct.

Figs. 17A and 17B illustrate the processing steps that are performed in order to determine the correct search parameters for a given tandem mass spectrometry spectrum in accordance with one embodiment of the present invention.

Fig. 18 is a scatter plot of cross correlation values for a tandem mass spectrometry data set that was computed in accordance with the methods of the present invention.

Like reference numerals refer to corresponding parts throughout the several views of the drawings.

5. DETAILED DESCRIPTION OF THE INVENTION

The present invention provides novel systems and methods for determining the parent ion charge state in tandem mass spectrometry data. Furthermore, the present invention provides novel systems and methods for computing a correlation value between a predicted spectrum **PS** and an actual experimental MS/MS spectrum **ES** that is not biased by the mass of the parent ion or the parent ion charge state of the **ES**. The present invention also provides novel systems and methods for determining a confidence value for correlation scores between predicted spectra **PS** and actual

MS/MS spectra **ES**. In addition, the present invention provides novel systems and methods for specifying database search criteria for experimental MS/MS spectra **ES** based on the presence or absence of probe fragment peaks in such spectra. Specification of such search criteria reduces the chances that false positive identifications of the spectra are made.

5.1 OVERVIEW OF AN EXEMPLARY SYSTEM

FIG. 2 shows a system 200 for analyzing tandem spectrometry data in accordance with one embodiment of the present invention.

System 200 preferably includes:

- a central processing unit 222;
- a main non-volatile storage unit 234, preferably including one or more hard disk drives, for storing software and data, the storage unit 234 typically controlled by disk controller 232;
- a system memory 238, preferably high speed random-access memory (RAM), for storing system control programs, data, and application programs, including programs and data loaded from non-volatile storage unit 234; system memory 238 may also include read-only memory (ROM);
- a user interface 224, including one or more input devices, such as a mouse 226, a keypad 230, and a display 228;
- an optional network interface card 236 for connecting to any wired or wireless communication network; and
- an internal bus 233 for interconnecting the aforementioned elements of the system.

Operation of system 200 is controlled primarily by operating system 240, which is executed by central processing unit 222. Operating system 240 may be stored in system memory 238. In addition to operating system 240, a typical implementation of system memory 238 includes:

- File system 242 for controlling access to the various files and data structures used by the present invention;

- MS/MS processing module 244 for associating tandem mass spectrometry data with specific biological macromolecules and statistically assessing such assignments;
- One or more experimental MS/MS data sets 260 for processing; and
- One or more biological macromolecule databases 270 for use in assigning biological macromolecules to tandem mass spectrometry data.

In a preferred embodiment, MS/MS processing module 244 includes:

- an MS/MS data sorting module 246 for sorting MS/MS data based on the presence and levels of fragment peaks resulting from modifications to a parent ion;
- a charge state determination module 248 for determining the charge state of a parent ion;
- a database searching module 250 for searching a biological macromolecule database for sequence fragments (*e.g.*, peptides) that have molecular weights within a threshold value of the mass of a parent ion of an MS/MS spectrum;
- an MS/MS spectra prediction module 252 for generating predicted spectra from sequence fragments identified by database searching module 250;
- a correlation determination module 254 for correlating MS/MS spectra generated by spectra prediction module 252 to experimentally determined MS/MS spectra; and
- a false data profiling module 256 for generating and fitting incorrect search results to a gamma distribution curve.

Each experimental MS/MS data set 260 comprises a plurality of MS/MS spectra. In some embodiments, a data set 260 is generated was follows. A probe (*e.g.*, the fluorophosphonate probe FP-Peg-TAMRA that is illustrated in Fig. 16A) was added to a sample containing the proteome from an organism at concentrations of 5-10 μM and allowed to react with the proteome for 30-60 minutes. The samples were then denatured with urea, reduced with dithiothreitol, and alkylated with iodoacetamide. The samples were then gel filtered to remove excess reagents and digested with trypsin for one hour. The probe-modified peptides were purified from the mixture using anti-probe (*e.g.* anti-TAMRA) monoclonal antibodies immobilized on agarose beads prior to analysis by liquid chromatography-tandem mass spectrometry (LCMS/MS). This analytical separation results in a plurality of fractions. A tandem mass spectrum 262 of each fraction is taken, resulting in the plurality of MS/MS spectrums 262 found in a representative data set 260.

Each MS/MS spectrum 262 typically represents a unique parent ion collected from a liquid chromatography fraction. This unique parent ion, in turn, is fragmented by the mass spectrometer to yield a plurality of fragment ions that appear in spectrum 262. In spectrums 262, each fragment ion can be represented by a bar graph whose abscissa value indicates the mass-to-charge ratio (m/z) and whose ordinate value represents intensity. A representative experimental tandem mass spectrum 262 **ES** in accordance with one embodiment of the invention is illustrated in Fig. 6. The spectrum illustrated in Fig. 6 is the fragmentation pattern of a peptide that has been labeled with the fluorophosphonate probe FP-Peg-TAMRA. The structure of FP-Peg-TAMRA is illustrated in Fig. 16A.

A number of mass spectrometers can be used to generate MS/MS spectra 262 in data set 260, including, but not limited to, a triple-quadruple mass spectrometer, a Fourier-transform cyclotron resonance mass spectrometer, a tandem time-of-flight mass spectrometer, and a quadropole ion trap mass spectrometer. Each of the spectra 262 described herein (*e.g.*, those described in Figs. 1, 3, 6, 7A, 7B, 8, 9, 11, 12, 13, 14 and 15) were analyzed using a Finnigan LCQ-DecaXP ion trap tandem mass spectrometer interfaced with an Agilent 1100 series capillary HPLC through a nano-electrospray ionization source. Samples were injected using the LC autosampler and separated on a 15 cm X 180 µm C18 reversed phase silica column at a flow rate of one µL/min with a two hour gradient from ten percent buffer B (acetonitrile + 0.1 percent formic acid), ninety percent buffer A (water + 0.1 percent formic acid) to fifty percent buffer B, fifty percent buffer A. Data was collected continuously in data dependent mode (dynamic exclusion on) with one MS scan followed by 3 MS/MS scans. However, those of skill in the art will appreciate that there are any number of different ways that MS/MS spectra 262 can be collected and all such methods are within the scope of the present invention.

Biological macromolecule database 270 includes the characteristics of biological macromolecules. In some embodiments of the present invention, the biological macromolecules under study are proteins. In such embodiments, biological macromolecule database 270 is a protein sequence database such as the nonredundant human protein database available from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). Other representative macromolecule databases 270 include, but are not limited to, the Genpept database, the GenBank database (Burks *et al.*, 1990, Methods in Enzymology 183, 3), the EMBL data library (Kahn *et al.*, 1990, Methods in Enzymology 183, 23), the Protein Sequence Database (Barker *et al.*, 1990, Methods in Enzymology 183, 31, SWISS-PROT (Bairoch *et al.*, 1993, Nucleic Acids Research 21, 3093-3096) and the PIR-International database (1993, Protein Sew Data Anal. 5, 67-192). In instances where the biological macromolecules under study are nucleic acids, carbohydrates, or other biological macromolecules, suitable biological macromolecule databases 270 can be used.

Now that an overview of a system 200 in accordance with one embodiment of the present invention has been described, various advantageous methods in accordance with the present invention will now be disclosed in the following sections.

5.2 PARENT ION CHARGE STATE DETERMINATION

A considerable problem in biology is the identification of those members of a proteome that participate in specific biological processes, assign a function to each, and devise strategies for their selective modulation. The recent development of activity-based profiling (ABP), in which a chemical probe can be used to label and isolate an enzyme from a complex mixture, provides a strategy for identifying those proteins associated with a particular biological activity, thereby taking a step toward their functional identification. See, for example, Cravatt & Sorensen, 2000, Curr. Opin. Chem. Biol. 4, 663-668 and Greenbaum *et al.*, 2000, Chem. Biol. 7, 569-581. In one form of ABP, fluorophosphonates known to be "specific" covalent inactivators of serine proteases are linked to fluorophore. In the strategy, the "bait" (a fluorophosphonate) irreversibly "hooks" (phosphonylates) the active-site serine only in those proteases that are catalytically active, thereby attaching a fluorophore to such active proteins. Isolation of the tagged serine proteases followed by mass spectral analysis allows the identification of active proteases (as assessed by the reactivity of their active-site serine). Exemplary fluorophosphonate probes are described in Patricelli *et al.*, 2001, Proteomics 1, 1067-1071. In Patricelli *et al.*, fluorophosphonate probes are complexed with an active site serine residue in the serine hydrolase super-family of enzymes.

One aspect of the present invention provides methods for determining the parent ion charge state for a given MS/MS spectrum 262 that has been labeled with such a probe. Fig. 4 illustrates one such method in accordance with this aspect of the present invention. In some embodiments, the techniques illustrated in Fig. 4 are used in instances where the parent ion is a peptide from a protein that has been labeled with a fluorophosphonate probe (*e.g.*, the probe illustrated in Fig. 16A) thereby forming a phosphate/serine adduct with a serine in the peptide under study. A phosphate/serine adduct between a peptide and the fluorophosphonate, FP-Peg-TAMRA, is illustrated in Fig. 16 B. In some embodiments, the technique disclosed in Fig. 4 is used in instances where the parent ion is a peptide that includes a phosphorylated serine or a phosphorylated threonine residue.

In some instances, probes such as fluorophosphonates are consistently cleaved from the peptide during MS/MS fragmentation. In such instances, the peptide loses the mass of the probe. In the case of fluorophoshonate probes that attach to the peptides by forming a phosphate/serine adduct, the mass of a single proton is also lost due to probe cleavage. Furthermore, the peptide loses any formal charge that is associated with the probe. For example, in the case of peptides labeled with the FP-Peg-TAMRA probe illustrated in Fig. 16A, the peptides lose a single formal negative charge that is associated with the probe fluorophore. When this is the case, the peptide missing the probe will appear in the MS/MS spectra at the m/z value:

$$\begin{aligned}&[(\text{Measured m/z of parent ion}) \times (\text{parent ion charge state}) \qquad \text{Eqn. 1}\\ &\quad - (\text{mass of probe}) - (\text{mass of proton})] \,/\, [(\text{parent ion charge state}) - 1]\end{aligned}$$

In equation 1, the m/z of the parent ion multiplied by the parent ion charge state is simply the mass of the parent ion. Thus, the numerator of Eqn. 1 is the mass of the parent ion (the peptide including the probe) minus the mass of the fluorophosphonate probe minus the mass of one proton. The denominator of Eqn. 1 is the parent ion charge state minus the charge on the probe after it has been cleaved from the peptide. For the FP-Peg-TAMRA probe, the charge on the probe after it has been cleaved from the peptide is 1.

As noted above, the parent ion charge state is not known. However, all other elements of Eqn. 1 are known. The parent ion charge state can be determined using the process steps illustrated in Fig. 4. The process steps illustrated in Fig. 4 are performed by one embodiment of charge state determination module 248 (Fig. 2). In step 402, charge state determination module 248 sets the parent ion charge state for the MS/MS spectrum of a labeled peptide to **i**, where **i** is an integer such as +2, +3, +4, +5, +6, +7, or greater. In typical embodiments, **i** is set to 2 in the first instance of step 402. In alternative embodiments, the parent ion charge state is negative and **i** is an integer such as –7, -6, -5, -4, -3, -2, or lower. In step 404, charge state determination module 248 searches the MS/MS spectra under study for a peak that has an m/z equal to Eqn. 1, where **i** is considered the "parent ion charge state."

To illustrate, consider a parent ion with a mass/charge ratio of 1101 that contains a fluorophosphonate probe with a mass of 100 daltons. As noted above, the parent ion charge state is unknown. Thus, in accordance with Fig. 4, a parent ion charge state of 2 is first attempted (Fig. 4; step 402) by charge state determination module 248. Therefore, a search of the MS/MS spectrum is performed for the following peak (Fig. 4; step 404):

$$[(1101) \times (2) - (100) - (1)] / \ [2 - 1]$$
$$= 2101$$

If a peak is found within predetermined error limits of the value of Eqn. 1 for a given value **i** (406-Yes), then a possible valid parent ion charge state is determined to be **i** (Fig. 4, step 408).

In some embodiments, a second criterion must be satisfied in order to achieve the condition 406-Yes. That is, the peak that is found within predetermined error limits of the value of Eqn. 1 must have a certain minimum intensity. In some instances, the peak has the requisite minimum intensity level when the ratio of the size of the peak found within predetermined error limits of the value of Eqn. 1 to the largest peak in the MS/MS spectrum is between 0.001 and 1, more preferably between 0.05 and 1 or between 0.05 and 0.5. In one instance, if the peak found within predetermined error limits of the value of Eqn. 1 has a peak size of 10 arbitrary units and the largest peak in the spectrum has a size of 100 arbitrary units, the ratio between the two peaks would be 10:100 or 0.1. Thus, the peak would satisfy the second criterion (406-Yes).

Continuing with the example, if a peak is found within predetermined error limits of 2101 in the MS/MS spectra, then a possible valid parent ion charge state is 2. In some embodiments, the predetermined error limit is in a range between ±0.5 and ±5, with ±1 being preferred.

Next, charge state determination module 248 determines whether the maximum possible **i** has been set (410). Step 410 is performed regardless of the outcome of step 406 (406-No or 406-Yes). This is because there can be more than one possible valid parent ion charge state. If the maximum possible **i** has been set (410-Yes) and no possible valid parent ion charge states have been identified (406-Yes), then module 248 has not determined the parent ion charge state. If the maximum possible **i** has not been set (410-No) then **i** is incremented (414) and a new search (404) is performed by charge state determination module 248.

In the example above, a peak at 2101 ± predetermined error limits is not found (406-No), and **i** is incremented from +2 to +3. Therefore, a search of the MS/MS spectrum is performed for the following peak (Fig. 4; step 404):

$$[(1101) \times (3) - (100) - (1)] / \ [3 - 1]$$
$$= 1601$$

Processing steps 404 through 414 are repeated by charge state determination module 248 until all possible parent ion charge states have been considered.

Fig. 6 provides another example of the determination of the parent ion charge state in accordance with this aspect of the present invention. The parent ion for the MS/MS spectrum illustrated in Fig. 6 has an m/z value of 1045.67. This parent ion is labeled with the probe FP-Peg-TAMRA (Fig. 16A) that has a mass of 784.3 daltons. FP-Peg-TAMRA is efficiently cleaved during the generation of the MS/MS spectrum illustrated in Fig. 6. Upon cleavage, the FP-Peg-TAMRA takes with it a single proton from the parent ion as well as a formal charge. Thus, the cleaved probe has a mass of 785.3 and a charge state of 1. Therefore, as can be seen in Fig. 6, a peak 602 appears at m/z 785.3. Peak 602 is the cleaved FP-Peg-TAMRA. Thus, if a peak corresponding to the parent ion can be found in the spectrum illustrated in Fig. 6 using Eqn. 1, the parent ion charge state for the parent ion used to produce Fig. 6 can be determined. The process described in Fig. 4 is used to attempt to find such a peak. First, it is assumed that the parent ion charge state is two. Using Eqn. 1, a search for a peak having the following value is performed:

$$[(1045.67) \text{ x } (2) - (784.3) - (1)] \;/\; [2 - 1]$$
$$= 1307.0$$

However, there is no m/z peak at the m/z value of 1307.0. Therefore, it is clear that the parent ion charge state is not two. Next, a parent ion charge state of three is attempted (Fig. 4; step 414). Accordingly, using Eqn. 1, a search for a peak having the following value is performed:

$$[(1045.67) \text{ x } (3) - (784.3) - (1)] \;/\; [3 - 1]$$
$$= 1175.85$$

The value 1175.85 is within a threshold limited of a prominent peak in the spectrum illustrated in Fig. 6 (peak 604, m/z value 1175.3) (Fig. 4; 406-Yes). Therefore, using the methods of the present invention, the parent ion charge state for the spectrum illustrated in Fig. 6 is +3. Accordingly, there is no need to search a biological macromolecule database 270 (Fig. 2) using any other parent ion charge state other than +3.

The techniques in accordance with this aspect of the present invention can be used to determine the parent ion charge state using a wide variety of probes that are efficiently cleaved from a biological macromolcule. Such cleavage will result in the appearance of a peak in the MS/MS spectra having the value:

[(m/z of parent ion) x (parent ion charge state) – (probe mass)
– (mass of the portion of parent ion that is removed upon probe cleavage)] /
[(parent ion charge state) – (probe charge)] Eqn. 2

In equation 2, the m/z of the parent ion multiplied by the parent ion charge state is the mass of the parent ion . Thus, the numerator of Eqn. 2 is the mass of the parent ion (the biological macromolecule fragment that is labeled with the probe) minus the mass of the probe that is efficiently cleaved from the parent ion minus the mass of any portion of the parent ion that is removed from the parent ion upon probe cleavage. In the case where the parent ion is a peptide having a serine that has been modified with the fluorophosphonate probe illustrated in Fig. 16A to form a phosphate/serine adduct such as the one illustrated in Fig. 16B, a proton is lost during cleavage. In this instance, the "mass of any portion of parent ion that is removed upon probe cleavage" is the mass of a proton, 1 dalton. In the case where the probe label is a cysteine residue and a sulfur atom is lost during cleavage, the "mass of any portion of parent ion that is removed upon probe cleavage" is 16 daltons.

Examples have been provided above where the probe has a charge state of "+1" (*e.g.*, the fluorophosphonate probe illustrated in Fig. 16A). However, it is possible for the probe to carry some other charge, such as +2. Furthermore, rather than eluting the digested proteome using an acid, bases could be used. In such instances, the probe may carry one or more formal negative charges away from the parent peptide. Thus, Equation 2 can be used to identify parent ions that have either a positive or a negative charge state. In one embodiment of the present invention, this parent ion charge state is determined by using the process flow illustrated in Fig. 4, with the exception that Eqn. 2 is calculated in step 404 rather than the equation illustrated in element 404 of Fig. 4.

5.3 CHARGE AND MASS INDEPENDENT CORRELATION OF PREDICTED MS/MS SPECTRA TO ACTUAL MS/MS SPECTRA

Correlation determination module 254 advantageously correlates MS/MS spectra generated by spectra prediction module 252 to experimentally determine MS/MS spectra 262 in a parent ion charge state and parent ion mass independent manner. Fig. 5 illustrates the process steps that are used to perform such a correlation in accordance with one embodiment of the present invention. In step 502, an MS/MS spectrum 262-N is selected from experimental MS/MS data set 260.

The goal is to correctly assign the MS/MS spectrum 262-N selected in step 502 to the biological macromolecule fragment (sequence fragment) that is represented by the spectrum. To accomplish this goal, a sequence fragment whose mass and fragmentation spectrum matches spectrum 262-N within threshold limits is obtained from biological macromolecule database 270 (Fig. 5; step 504). Typically, step 504 is performed by database searching module 250 (Fig. 2).

In some embodiments, biological macromolecule database 270 is a protein database and the sequence fragment obtained in step 504 (Fig. 5) is a peptide sequence. Furthermore, the peptide sequence is typically only a portion of the sequence of a biological macromolecule 272 in database 270 (Fig. 2). The sequence fragment that is obtained in step 504 has a calculated molecular weight that is within a threshold value of the mass of the parent ion of the MS/MS spectrum 262-N selected in step 502. This threshold value can be, for example, within ± 0.01 daltons to within ± 5 daltons of the mass of the parent ion, with a threshold value of ±1 daltons being preferred. The magnitude of the threshold value that is used in step 504 is a function of the accuracy of the ion trap mass spectrometer used to collect spectrum 262-N. In the ranges typically examined in the present invention, known ion trap mass spectrometers are typically accurate to ±0.5 daltons. Therefore, the threshold value in step 504 is set to a value larger than ±0.5 daltons in such instances. However, in the case where a more accurate tandem MS instrument is used, such as a quadrupole / time of flight (Q-TOF) tandem mass spectrometer, data collection accuracy is typically ±0.01 daltons. Thus, when this type of instrumentation is used, the magnitude of the threshold value can be set to a mass tolerance in the range of ±0.02 daltons or greater (*e.g.*, between ±0.02 daltons and ±3.0 daltons). In step 506, a predicted spectrum **PS** is generated for the candidate biological macromolecule fragment identified by database searching module 250 in step 504. The predicted spectrum **PS** is generated based on predicted fragmentation patterns. Known programs, such as SEQUEST, can be used to perform steps 502 through 506.

Once a predicted spectrum **PS** and an experimental MS/MS spectrum 262-N (ES) have been obtained (steps 502-506), steps 508 through 520 are used to compute a cross-correlation score between the two spectra. Advantageously, this cross-correlation score is not biased by the parent ion charge state or the parent ion mass of the parent ion that corresponds to the spectrum **ES** that was obtained in step 504. First, offset Υ is set to a lower bound (*e.g.*, -50, –75, *etc.*). Offset Υ is a relative displacement that is imposed between spectrum **PS** and spectrum **ES**. In other words, each abscissa value in spectrum **PS** (or, alternatively, each abscissa value in spectrum ES) is shifted by Υ in processing step 508.

In step 510 a cross correlation score is computed between spectrum **PS** and the spectrum **ES** that was obtained in step 502. The cross correlation score computed in step 510 can be done in a variety of manners. In general, the cross-correlation function $C_{ab}(\Upsilon)$ between two real waveforms a(*t*) (*e.g.*, **PS**) and b(*t*) (*e.g.*, **ES**) can be approximated by the discrete correlation:

$$C_{ab}(n\Delta t) = \frac{1}{T}\sum_{t=0}^{T} a(t)b(t \pm n\Delta t) \qquad \text{Eqn. 3}$$

where

$$n = 0, 1, 2, \ldots \frac{T}{\Delta t},$$

Δt is the sampling interval, and

$n\Delta t$ is the relative displacement (Υ) between a(*t*) and b(*t*).

See, for example, Horlick, 1973, Anal. Chem. 45, p. 319. A cross correlation function, such as Eqn. 3, measures the coherence of two signals (two spectra) by, in effect, translating one signal across the other. The displacement value $n\Delta t$ (Υ) is the amount by which the signal is offset during the translation. In other words, the displacement value $n\Delta t$ is the amount by which spectrum **PS** or spectrum **ES** was shifted during processing step 508. As described in more detail below, in processing steps 508 through 516, the displacement (Υ) is varied between a lower bound (*e.g.*, -50, –75, *etc.*) and an upper bound (*e.g.*, 50, 75, *etc.*).

The correlation between two signals a(*t*) (*e.g.*, **PS**) and b(*t*) (*e.g.*, **ES**) can also be computed through Fourier transform methods. See, for example, Horlick *et al.* (Eds.), *Contemporary Topics in Analytical and Clinical Chemistry* 3, Plenum Press, New York, 1978; Powell and Hieftje, 1978, Analytica Chemica Acta 100, 313-327; Eng *et al.*, 1994, J. Am. Soc. Mass Spectrom. 5, 976-989; United States Patent Number 6,017,693 to Yates, III *et al.*; and United States Patent Number 5,538,897 to Yates, III *et al.* Accordingly, the cross-correlation function $C_{ab}(\Upsilon)$ between **PS** and **ES** is computed any number of different ways in step 510. For example, in some embodiments, Eqn. 3, where *n* is fixed to a particular value, is used to compute the cross-correlation function $C_{ab}(\Upsilon)$ in step 510. In some embodiments, the cross-correlation function $C_{ab}(\Upsilon)$ is computed through a discrete or fast Fourier transform technique.

In step 512, Υ is incremented and, if an upper bound has not been exceeded (516-No), the cross-correlation function $C_{ab}(\Upsilon)$ between **PS** and **ES** is recomputed with the new value for Υ. In typical embodiments, the upper bound is 50 or 75. The value of $C_{ab}(\Upsilon)$ calculated during each instance of step 510 is stored for later use. If Υ exceeds the upper bound (516-Yes), control passes to step 518.

In step 518, the cross correlation score Xcorr is assigned the value $C_{ab}(0)$ minus the mean of $C_{ab}(\Upsilon)$ over the range from the lower bound through the upper bound. Thus, if the lower bound is –75 and the upper bound if +75, Xcorr is assigned the value of $C_{ab}(0)$ minus the mean value of $C_{ab}(\Upsilon)$ that was computed in all instances of processing step 510 for a given sequence fragment.

Computation of Xcorr in accordance with step 518 is known in the art. See, for example, Eng *et al.*, 1994, J. Am. Soc. Mass Spectrom. 5, 976-989. However, Xcorr is problematic because Xcorr values tend to drift upward for higher parent ion charge states and for higher parent ion charge masses as illustrated in Fig. 1 and discussed in Section 2, above. To address this shortcoming in the art, correlation determination module 254 advantageously computes the value XcorrNorm (Step 520):

$$\text{XcorrNorm} = C_3 \times \text{Xcorr} \times \sqrt{\frac{\text{Charge} + C_1}{\text{Mass} - C_2}} \qquad \text{Eqn. 4}$$

where C_1, C_2, and C_3 are constants, "Charge" is the parent ion charge state and "Mass" is the mass of the parent ion. While a number of different values for C_1, C_2 and C_3 can be used, in a preferred embodiment, C_1 has the value of about 4.26, C_2 is the mass of a probe used to label the parent ion (*e.g.*, a fluorophosphonate probe or a phosphate) plus the value 324, and C_3 is 20. As use here, the term about 4.26 means a number in a range between 4.0 and 4.4. In other words, in a preferred embodiment:

$$C_1 = 4.26\text{, and}$$
$$C_2 = \text{mass of probe} + 324$$

In some embodiments, C_1 is between –10 and +10. In some embodiments, C_2 is between –2000 and +2000. In some embodiments, C_3 is, in fact, any value. It will be appreciated that C_3 does not affect how the score functions. Rather, it is simply used to scale the score, if desired. In some embodiments, C_3 is not used.

In alternative embodiments the score XcorrNorm″ is computed rather than, or in addition to, the score XcorrNorm:

$$\mathrm{XcorrNorm}'' = \left[\frac{C_5 * \mathrm{Xcorr}}{\sqrt{(\text{Sequence fragment} + 1\ \text{dalton}) + \mathrm{Mass}_{\text{Probe modification}}} - C_6}\right] + C_7 * \left[1 - \frac{\mathrm{Xcorr}_{2\mathrm{nd}}}{\mathrm{Xcorr}_{1\mathrm{st}}}\right]$$

where

C_5, C_6, and C_7 are constants;

"Sequence fragment + 1 dalton" is the mass of the sequence fragment selected in step 504 for which the predicted spectrum **PS** was computed in step 506 and the cross correlation score Xcorr was computed in step 518;

"$\mathrm{Mass}_{\text{Probe modification}}$" is the increase in the mass of the sequence fragment as a result of the modification of the fragment by the probe;

"$\mathrm{Xcorr}_{1\mathrm{st}}$" is the Xcorr value of the predicted spectrum of the highest scoring sequence fragment in biological macromolecule database 270; and

"$\mathrm{Xcorr}_{2\mathrm{nd}}$" is the Xcorr value of the predicted spectrum of the second highest scoring sequence fragment in biological macromolecule database 270.

While a number of different values for C_5, C_6 and C_7 can be used, in a preferred embodiment, C_5 has the value of 60, C_6 has the value 667, and C_7 has the value 0.8. In other words, in a preferred embodiment:

$$\mathrm{XcorrNorm}'' = \left[\frac{60 * \mathrm{Xcorr}}{\sqrt{(\text{Sequence fragment} + 1\ \text{dalton}) + \mathrm{Mass}_{\text{Probe modification}}} - 667}\right] + 0.8 * \left[1 - \frac{\mathrm{Xcorr}_{2\mathrm{nd}}}{\mathrm{Xcorr}_{1\mathrm{st}}}\right]$$

In some embodiments, C_5 is a value between 10 and 500, a value between 30 and 300, or a value between 40 and 100. In some embodiments, C_6 is a value between 300 and 5000, a value between 350 and 2500, or a value between 400 and 1500. C_7 is a coefficient between 0.1 and 1.0. In preferred embodiments, C_6 is a value between 0.6 and 0.9.

In step 522, a determination is made as to whether there are any remaining sequence fragments having the mass of the parent ion in biological macromolecule database 270. If so (522-Yes), process control returns to step 504 where a different sequence fragment is obtained from biological macromolecule database 270 and steps 506 through 520 are repeated in order to derive an XcorrNorm score for the new sequence fragment. If not (522-No), process control passes to step 524 where the MS/MS spectrum **ES** obtained in step 502 is assigned the sequence fragment that achieved the highest XcorrNorm score.

Figs. 7A, 7B, and Fig. 18 respectively illustrate a scatter plot comparison of (i) Xcorr versus parent ion mass (Fig. 7A), (ii) XcorrNorm versus parent ion mass (Fig. 7B), and (iii) XcorrNorm″versus parent ion mass (Fig. 18) for an experimental data set 260. Due to the nature of the probe used to generate the data represented in Figs. 7A, 7B, and Fig. 18, it was possible to independently verify whether the assignments made were correct. This experimental design and verification process is described in more detail in Section 5.5.2, below. The independent verification was used to evaluate the parent ion charge state and mass dependence of Xcorr, XcorrNorm, and XcorrNorm″. In each figure, correctly assigned results are indicated by solid diamonds and incorrectly assigned results are indicated by hollow squares. The mass dependence of Xcorr (Fig. 7A) is absent in XcorrNorm (Fig. 7B) and XcorrNorm″.

The Xcorr scores in Fig. 7A were computed using SEQUEST. MS/MS spectra 262 were converted to ASCII ".dta" files and the ".dta" files were searched against biological macromolecule database 270 using SEQUEST with the search parameters:

- variable modification with probe mass on appropriate residues;
- variable +16 modification of methionine to account for oxidation;
- fixed +57 modification of cysteine to account for iodoacetamide modification;
- allowance for up to two missed trypsin cleavage sites; and
- MS mass tolerance of three daltons.

The XcorrNorm scores (Fig. 7B) were computed using Eqn. 4 in which the Xcorr scores are those calculated for Fig. 7A and where C_1 is 4.26, C_2 is the mass of the probe + 324, and C_3 is 20. The XcorrNorm″ scores (Fig. 18) wee computed using the equation for XcorrNorm given above where C_5 has the value of 60, C_6 has the value 667, and C_7 has the value 0.8. The probe used to label the proteome represented by Figs. 7A, 7B and Fig. 18 was FP-Peg-TAMRA (Fig. 16A), molecular mass 787.31.

Table 1 provides a numerical comparison of Xcorr and XcorrNorm. In Table 1, the first row, "+3, +4, +5 search average deviation," represents the percent standard deviation between the assignments made to spectra 262 in a data set 260 based on the assumption that the parent ion charge mass was respectively +3, +4, or +5. In other words, assignments were made to each spectrum 262 in a data set 260 based upon the assumption that the parent ion charge state was +3 and cross correlation scores were computed for these assignments. Similarly, assignments and cross correlation scores were computed based on the assumption that the parent ion charge mass was respectively +4 and +5. Then, the average score for the +3, +4 and +5 data set cross correlation scores was computed. The deviation of these three is provided in the first row of Table 1. The average deviation gives a measure of how well the parent ion charge (and indirectly the size) component has been removed from the cross correlation score.

In row 2 of Table 1, the number of standard deviations different from correct to incorrect identifications is equal to:

$$(\text{Average Cross Correlation Score}^{\text{Positive Data Set}} - \text{Average Cross Correlation Score}^{\text{False Positive}}) / [(\text{Standard Deviation}^{\text{Positive Data Set}} + \text{Standard Deviation}^{\text{False Positive Data Set}}) / 2].$$

where the positive data set is the cross correlation score of each correctly identified spectra 262 in a data set 260 and the false positive data set is the cross correlation score of each incorrectly assigned spectra 262 in the data set 260. As discussed above, in the case of the data set 260 used to compute Table 1, it is possible to independently verify which spectra 262 have been correctly identified and which spectra have been incorrectly identified using the experimental design that is described in more detail in Section 5.5.2, below. The number of standard deviations different from correct to incorrect identifications measures the degree of separation between the positive and false positive cross correlation score data sets. The higher the number, the less overlap there is between the two distributions. Therefore, a higher number in row 2 of Table 1 indicates a more favorable cross correlation scoring function.

To compute the "value percent incorrect identifications at eighty percent coverage of correct identifications" in row 3 of Table 1, the cross correlation scores for each spectra 262 in the data set 260 were first ranked in decreasing order. Then the

20th percentile of correctly identified scores was determined. This is the score above which eighty percent of the correctly identified spectra were contained. Then the number of incorrectly identified spectra with scores above this 20th percentile value was determined and divided by the number of correctly identified scores above the same value. In row 4, the same type of analysis was performed except that the threshold score was the score above which 90% of the correctly identified spectra were contained (or the 10th percentile score). The values in rows 3 and 4 of Table 1 provide another measure of how well the given cross correlation functions (Xcorr and XcorrNorm) distinguish between correct assignments and false positive assignments.

Table 1 – Comparison of Xcorr to XcorrNorm

Title	Xcorr	XcorrNorm
+3 search, +4 search , +5 search average deviation	8.75%	2.3%
Number of standard deviations different from correct to incorrect identifications	2.2 (+5 search) 2.17 (+4 search) 1.86 (+3 search)	2.43 (+5 search) 2.63 (+4 search) 2.63 (+3 search)
Percent incorrect identifications at eighty percent coverage of correct identifications	19.3%	7.2%
Percent incorrect identifications at ninety percent coverage of correct identifications	50.9%	16.0%

From Table 1, it can be seen that the dependence of XcorrNorm on the parent ion charge is much less than that of Xcorr. Additionally the degree of overlap between the score distributions of incorrect identification and the score distributions of correct identifications is much less for XcorrNorm. An important note about Table 1 is that even within the data for each parent ion charge state, an increased separation between incorrect and correct score distributions is observed for XcorrNorm.

Eng *et al.* (J. Am. Soc. Mass. Spectrom. 5, 976-989 (1994)) noted that the differences between the normalized cross-correlation parameter of the first and second ranked sequence fragments for any given mass spectrum have shown a trend useful for distinguishing correct identifications from false positives. Accordingly, the program SEQUEST provides a score called DCN which is the fractional difference between the Xcorr of the highest scoring sequence fragment for a particular MS/MS spectrum and the score of the second highest sequence fragment for the particular MS/MS spectrum 262.

The DCN score can be combined with XcorrNorm to further distinguish correct and incorrect spectral assignments. In one embodiment, the DCN score is combined with the XcorrNorm score to produce XcorrNorm′. In one embodiment, XcorrNorm′ equals XcorrNorm + (C_4*DCN). In some embodiments, C_4 is a number between 1.0 and 5.0. In a preferred embodiment, C_4 is equal to 1.54. Table 2 provides a comparison of XcorrNorm, and XcorrNorm′ for the tandem spectrometry data set used to generate Table 1, above. Each value in Table 2 was computed in the same manner that the corresponding value in Table 1, above, was computed.

Table 2 – Comparison of XcorrNorm and XcorrNorm′

	XcorrNorm	**XcorrNorm′**
+3 search, +4 search , +5 search average deviation	2.3%	2.2%
Number of standard deviations different from correct to incorrect identifications	2.43 (+5 search) 2.63 (+4 search) 2.63 (+3 search)	2.55 (+5 search) 2.72 (+4 search) 2.68 (+3 search)
Percent incorrect identifications at eighty percent coverage of correct identifications	7.2%	4.7%
Percent incorrect identifications at ninety percent coverage of correct identifications	16.0 %	16.7 %

Table 3 provides a comparison of the accuracy of Xcorr, XcorrNorm and XcorrNorm″ for the same data set used to generate the scatter plots illustrated in Figs. 7A and 7B.

Table 3 – Comparison of Xcorr, XcorrNorm, and XcorrNorm″

Percent Accuracy	**Xcorr**	**XcorrNorm**	**XcorrNorm″**
at 50% coverage of positives	99.8	99.4	**100**
at 60% coverage of positives	97.7	98.3	**98.6**
at 70% coverage of positives	93.2	**97.6**	97.3
at 80% coverage of	83.7	**95.3**	94.8

Percent Accuracy	Xcorr	XcorrNorm	XcorrNorm″
positives			
at 90% coverage of positives	66.1	**85.7**	82.8
at 95% coverage of positives	50.1	**73.6**	68.6
at 100% accuracy**	28.1	36.1	**50.6**

To produce Table 3, an attempt is made to determine the sequence fragment identity of each spectrum in MS/MS data set 260 using the techniques disclosed in Fig. 5. In this way, each spectrum in MS/MS data set 260 is assigned a predicted sequence fragment. There are two classes of sequence assignments, those that represent the correct (true) assignment, and those that represent an incorrect (false) assignment. Experimental conditions, such as those described in Section 5.5.2, can be used to independently determine whether a given spectrum in MS/MS data set 260 has been correctly assigned. In other words, it is possible to classify the sequence fragment assignments made using the methods described in Fig. 5 into false assignments and true assignments using independent experimental means.

For each respective spectrum in MS/MS data set 260, there is a correlation score that measures the agreement between the actual spectrum and the predicted fragmentation pattern of the sequence fragment that has been assigned to the spectrum using each of the three correlation metrics: Xcorr, XcorrNorm, and XcorrNorm″. These correlation metrics include both false assignments and true assignments.

While the sets of correlation scores include both falsely and truly (correctly) assigned spectra in data set 260, it is the distribution of the correlation scores of truly (correctly) assigned spectra that is used as the basis for defining the rows in Table 3. There are three such distributions, one for each of the three types of correlation metrics considered (Xcorr, XcorrNorm, and XcorrNorm″). Thus, there is (i) a distribution of Xcorr values from the truly (correctly) assigned spectra, (ii) a distribution of XcorrNorm values from the truly assigned spectra, and (iii) a distribution of XcorrNorm″ values from the truly assigned spectra. From these distributions, a pool of correlation values can be evaluated. The first row of Table 3 considers the pool of correlation values that exceed the 50th percentile correlation value in the distributions. For example, consider the case in which the 50th percentile

Xcorr value in the Xcorr true distribution is 0.73. In this example, the pool of correlation values that exceed the 50th percentile correlation value of Xcorr are all those Xcorr values that exceed 0.73, regardless of whether such values represent correctly or incorrectly assigned spectra.

With this background in mind, the term "percent accuracy" as used in Table 3 can be better understood. Here, percent accuracy for a given pool of correlation values is:

$$\frac{\text{Number of correlation values in pool representing correctly identified sequence fragments}}{\text{Total number of correlation values in the pool}}$$

Thus, consider the pool represented by the cell at column 1, row 1. This cell corresponds to the pool of sequences having Xcorr values that exceed the 50th percentile of the distribution of Xcorr values of correctly assigned spectra. The value reported for this cell is the number of correlation values in the pool representing correctly identified sequence fragments divided by the total number of correlation values in the pool. In this case, the value 99.8 means that 99.8 percent of the Xcorr values exceeding the 50th percentile Xcorr value in the Xcorr "true" distribution are correctly assigned.

Each value in column 1 refers to a pool of Xcorr values. Each value in columns 2 and 3 respectively correspond to pools of XcorrNorm and XcorrNorm″ values. The final row of Table 3 is the percentage of correctly assigned spectra in the pool of spectra that have Xcorr values exceeding the highest scoring false result. For example, if the highest scoring false result has an Xcorr of 0.4, the last cell in column 1 represents the percentage of correctly assigned spectra in the pool of spectra having an Xcorr value exceeding 0.4.

From Table 3, it can be seen that Xcorr gives a better overall separation of the positive and false positive (false, incorrect) result distributions (i.e. most of the percent accuracy scores are best with this score, particularly as you start to try to include the higher percentages of the positive results). But the last line of Table 3 shows clearly that XcorrNorm″ has the highest number of one hundred percent accurate hits, those that score higher than the highest scoring false positive. Moreover, what has been found when looking at real data is that there is a higher degree of confidence (i.e. >95% confidence) for identifications made based upon the XcorrNorm″ score.

XcorrNorm′ and XcorrNorm″ were developed in different manners. For XcorrNorm′, the average positive and negative scores for each charge state in a set of data was considered. The score was developed by minimizing the difference between the average positive scores across the different charge states, and maximizing the number of standard deviations between the average positive and false positive score distributions. It was determined that the charge states would self-normalize when the two distributions were separated. The one drawback with the method used to derive XcorrNorm′ is that both distributions are treated as normal distributions (*i.e.* standard deviations and averages were used). Such distributions don't actually apply for gamma distributions, which is the actual form of the distribution of the underlying data. So the skew, or tailing, of the gamma distribution was not accounted for in the process used to derive XcorrNorm′. Thus, XcorrNorm′ fails to account for the few straggling false positives with relatively high scores (*i.e.* the tail of the gamma distribution).

To address the shortcomings in XcorrNorm′, XcorrNorm″ was developed. To derive XcorrNorm″, the number of positive results with a higher score than the highest scoring false positive (the last line in Table 3) was maximized. A by-product of this effort was that the mass and charge dependence of the data disappeared from the correlation function.

If either of these approaches to computing a Xcorr value are tried with other equation formats [*e.g.* log(charge/mass), (charge/mass)^2, *etc.*) the process is not nearly as smooth]. The two goals of separating the charge/mass dependence, and increasing the distinction between positive and negative results seem to be diverging goals with different solutions.

5.4 ASSIGNING PROBABILITY VALUES TO INDIVIDUAL CROSS CORRELATION SCORES

Fig. 8 illustrates the distribution of XcorrNorm′ scores for a representative data set 260. The data set used to produce the XcorrNorm′ scores in Fig. 8 allowed for the independent determination as to whether the spectra were correctly assigned. This technique is described in more detail in Section 5.5.2, below. XcorrNorm′ scores representing correctly assigned spectra were binned into cross hatched bars and XcorrNorm′ scores representing incorrectly assigned spectra were binned into unmarked bars. The magnitude of these bars represents the number of XcorrNorm′ scores in any given bar.

Fig. 8 shows how the distribution of scores from incorrect and correctly assigned spectra 262 are overlapping but distinct distributions. The distribution of correlation scores for incorrectly assigned spectra and the distribution of correlation scores for correctly assigned spectra are resolved better with XcorrNorm′ than they are with Xcorr alone. However, it is still not possible to unambiguously determine which results in a data set 260 are correct and which results are incorrect except in the limited case where techniques such as those disclosed in Section 5.5.2 are used to independently determine which assignments in a given data set 260 are correct. Therefore, a method is needed that effectively determines the relative contribution of the correct and incorrect distributions to a data set 260, and assigns probabilities to each score range. To accomplish this, at least one of these distributions in a real data set must be determined.

Fig. 9 illustrates the distribution of XcorrNorm′ scores for incorrectly assigned spectra 262 in a data set 260. To generate Fig. 9, a probe was used to label a proteome in such a way that it was possible to experimentally determine which spectra were assigned to incorrect sequence fragments. This technique is described in more detail in Section 5.5.2, below. Then, the Xcorr values for each of these incorrectly assigned spectra were plotted in Fig. 9. In Fig. 9, the distribution of incorrect search results was fitted to a gamma distribution curve using least squares. For more information on the least squares fit of data to a gamma distribution function, see Bevington and Robinson, *Data Reduction and Error Analysis for the Physical Sciences* Second Edition, 1992, WCB McGraw-Hill, Boston Massachusetts. Fig. 9 illustrates that the gamma distribution curve can be fitted to a distribution of incorrectly assigned Xcorr values for a given data set 260. Therefore, the distribution of false positive cross correlation scores can be statistically defined.

As illustrated in Fig. 9, the fit of false positive cross correlation scores to a gamma distribution is readily accomplished when the identity of the false positives is known. However, the identity of the false positive assignments for a data set 260 is not typically known. Therefore, it is usually not possible to obtain the type of fitting illustrated in Fig. 9. Furthermore, the shape of such a false positive distribution will be different depending on several factors, including the type of probe attached to the parent ion, the residue modified by the probe, biological macromolecule database 270 size, and general data complexity. Thus, in the typical setting, using known techniques, there is not enough information available to compute a false positive gamma distribution curve such as the curve illustrated in Fig. 9.

Advantageously, one aspect of the present invention provides a method for producing a false positive gamma distribution curve fit that takes into account the type of probe attached to the parent ion, the residue modified by the probe, biological macromolecule database 270 size, and general data complexity. Furthermore, unlike the case of Fig. 9, independent experimental determination of which spectra 262 have been incorrectly assigned is not necessary in this aspect of the invention.

The false positive gamma distribution curve fit that is computed in accordance with this embodiment of the invention can be used to determine the confidence that any given cross correlation score in a data set of such scores is a correct identification rather than a false positive identification.

An embodiment in accordance with this aspect of the invention will now be described in conjunction with Fig. 10. In some embodiments, some or all of the steps illustrated in Fig. 10 are performed by false data profiling module 256 (Fig. 2). In step 1002 of Fig. 10, a search of an experimental MS/MS data set 260 is performed. Step 1002 assigns a sequence fragment from biological macromolecule database 270 to each MS/MS spectrum 262 in data set 260 and then computes a cross correlation score for these assignments. The process steps illustrated in Fig. 5 may be used to make these assignments and to compute these cross correlation scores. However, there is no requirement that XcorrNorm′ or XcorrNorm scores be computed in the embodiment illustrated in Fig. 10. Programs such as SEQUEST can perform step 1002.

In step 1004 a search of the same experimental MS/MS data set 260 that was searched in step 1002 is performed. Step 1004 assigns a sequence fragment from biological macromolecule database 270 to each MS/MS spectrum 262 in data set 260 and computes cross correlation scores for these assignments. However, in step 1004, the molecular weight of a moiety (*e.g.*, the probe used to label the proteome, a residue such as serine, *etc.*) is defined with the incorrect mass. So, for example, the process steps illustrated in Fig. 5 can be performed for each MS/MS spectrum 262 in data set 260 in which one component (*e.g.*, the probe used to label the proteome, a residue such as serine, *etc.*) is defined as having a mass that is ten daltons greater than the actual mass of that component.

In some embodiments, the component that is assigned a higher mass in step 1004 is a probe (*e.g.*, a fluorophosphonate probe) that is complexed with proteome used to generate data set 260. For example, if the probe has a mass of 500 daltons, the probe is intentionally assigned an incorrect mass of 510 daltons. However, the present invention is not limited to the assignment of a higher mass to the probe. Any component that is likely to be present in the parent ions of the spectra 262 in data set 260 can be assigned a higher mass. For example, in the case where the parent ion is a peptide, any residue type, such as serine, can be assigned a mass that is 10 daltons higher than its actual mass. Like step 1002, XcorrNorm′ or XcorrNorm scores do not need to be calculated in step 1004.

In a preferred embodiment the parent ion of each spectrum 262 in data set 260 is labeled with a probe. In such embodiments, SEQUEST is run with the mass of the probe set to a mass that is ten daltons greater than the actual mass of the probe. The same search performed in step 1004 is performed in step 1006, with the exception that the parent ion is set to ten daltons less than the actual parent ion mass (*e.g.*, a probe mass that is ten daltons less than the actual probe mass). However, as is the case in step 1004, there is no requirement that the probe be assigned an incorrect mass in step 1006. Any component that is likely to appear in the parent ion (*e.g.*, serine) can be assigned an incorrect mass in step 1006. Generally, the same component is assigned incorrect masses in steps 1004 and 1006.

In step 1008, the false positive cross correlation scores computed in steps 1004 and 1006 are used to perform a least squares fit of a gamma distribution curve. It is not necessary to fit a gamma distribution curve to the correlation scores from both steps 1004 and 1006. In other words, only one of steps 1004 and 1006 needs to be performed. However, in a preferred embodiment, false positive cross correlation scores from both steps 1004 and 1006 are used so that the average probe modification mass in the combined set of cross correlation scores from steps 1004 and 1006 is the same as the probe modification mass in the set of cross correlation scores computed in step 1002. This is important when using XcorrNorm since the mass of the probe is a factor in computing XcorrNorm. Furthermore, the present invention is not limited to the use of the ±10 dalton adjustment described in Fig. 10. In fact, any differential pair of masses can be used. For example, in some embodiments, a differential of +5, +15, +20, or +25 daltons from the parent ion mass is used in step 1004 and, correspondingly, a differential of -5, -15, -20, or –25 daltons is used in step 1006 (or vice versa). In embodiments where the parent ion in one or more spectra in data set 260 is labeled with a probe, the probe may be given a mass of more than 5 daltons, more than 15 daltons, more than 20 daltons, more than 25 daltons, or an amount that is between five and 50 daltons more than the actual mass of the probe in step 1004. Correspondingly, the probe may be given a mass that is more than 5 daltons, more than 15 daltons, more than 20 daltons, or more than 25 daltons less than the actual mass of the probe in step 1006. In some embodiments, the probe is given a probe mass that is between five and 50 daltons less that the actual mass of the probe.

In some embodiments, steps 1004 and 1006 are repeated using a second mass differential. For example, step 1004 is performed with a differential of +10, step 1006 is performed with a differential of –10, and then step 1004 is repeated with a differential of +15 and 1006 is repeated with a differential of –15. Then, least squares is used to fit a gamma distribution curve to all of the false positive cross correlation scores computed in both instances of steps 1004 and 1006.

Fig. 11 illustrates a gamma distribution curve that has been fitted to false positive cross correlation scores that were computed with mass differentials of ±10 daltons from the actual mass of the probe used to label each of the parent ions in each MS/MS spectra in a data set 260 (steps 1004 and 1006) in accordance with step 1008 of Fig. 10. The cross correlation scores were computed using the same MS/MS spectra used to generate the scores in Fig. 9. However, the gamma distribution curve fit was determined by analyzing data that was generated by performing SEQUEST searches where the mass of the probe was incorrectly entered in accordance with steps 1004 and 1006 of Fig. 10. A comparison of Fig. 9 and Fig. 11 reveals that the shape of the distribution of false positive cross correlation scores can be adequately predicted by performing searches of each data set 260 (Fig. 2) with incorrect probe mass parameters set in a program such as SEQUEST. Fig. 13 demonstrates that, with two additional searches (steps 1004 and 1006 of Fig. 10), the false positive data distribution can be determined for any data set, in any database, with any probe.

In step 1010, the scale of the gamma distribution curve is adjusted to fit the mass search results from step 1002. Then, in step 1012, the gamma distribution curve is used to determine the confidence that cross correlation scores computed in step 1002 are not within the gamma distribution curve for false positive cross correlation scores. This can be done by binning cross correlation scores and then comparing how many scores are predicted in each bin based on the false positive result gamma distribution versus how many scores are actually in each bin.

Fig. 12 illustrates the binning process that is performed in step 1012 of Fig. 10. Fig. 12 is a plot of all the Xcorr cross correlation values that were determined for a MS/MS data set 260. This data set was derived from tandem MS of probe labeled proteins. Three sets of cross correlation scores were computed using data set 260. The first set (step 1002, Fig. 10) is the true experimental cross correlation scores that were determined using SEQUEST where the probe mass was correctly entered. The second set (step 1004, Fig. 10) is a set of false positive cross correlation scores that were determined using SEQUEST where the probe mass was entered incorrectly with a mass that is 10 daltons greater than the actual mass. The third set (step 1006, Fig. 10) is a set of false positive cross correlation scores that were determined using SEQUEST where the probe mass was entered incorrectly with a mass that is 10 daltons less than the actual mass.

The false positive cross correlation scores from steps 1004 and 1006 were fitted to a gamma distribution curve and the curve was fitted to the experimental cross correlation scores derived in step 1002. The experimental cross correlation scores computed during step 1002 are plotted in bar graph form in Fig. 12. Each bar represents a small range of cross correlation values and the magnitude of each bar represents the number of cross correlation scores computed in step 1002 that fall within the range of cross correlation scores represented by the bar.

The total number of cross correlation scores plotted in Fig. 12 together with the false positive gamma distribution curve can be used to determine how many cross correlation scores should fall within each bar in Fig. 12. For example, the gamma distribution curve indicates that bar 1204 should have approximately 100 incorrectly assigned cross correlation scores. Box 1202 encompasses those bars in which there is at least a one percent confidence that a given Xcorr value falls within the incorrect cross correlation score gamma distribution curve. In other words, the confidence level that Xcorr values in bars outside of box 1202 are not part of the false positive cross correlation score gamma distribution curve is at least 99 percent.

Fig. 12 shows the advantages of the present invention. Using the false positive gamma distribution curve, it is now possible to determine how likely any given score is outside of the distribution of false positive cross correlation scores. Thus, the technique disclosed in Fig. 10 provides a method to statistically determine the confidence in any given cross correlation value, and therefore, the correctness of the sequence fragment assignment made to the spectrum 262 that corresponds to the cross correlation value.

The technique disclosed in Fig. 10 can be used to compare the effectiveness of the Xcorr and the XcorrNorm′ cross correlation functions. The tandem spectrometry data set 260 used to generate the cross correlation scores plotted in Fig. 12 was used to generate the cross correlation scores plotted in Fig. 13. The only difference between the scores in Fig. 13 and Fig. 12 is that the XcorrNorm′ cross correlation function was used to compute the ±10 incorrect cross correlation scores (steps 1004 and 1006) and true experimental cross correlation scores (step 1002) in Fig. 13 whereas the Xcorr cross correlation function was used to compute these correlation scores in Fig. 12. In Fig. 12, 312 cross correlation scores have a confidence of not being part of the false correlation score gamma distribution curve that is greater than 99 percent. In Fig. 13, 445 cross correlation scores have a confidence of not being part of the false correlation score gamma distribution curve that is greater than 99 percent, which is an increase of 43 percent over Fig. 12. Thus, in the case of the data set 260 used to generate Figs. 12 and 13, XcorrNorm′ produces better results than Xcorr.

Figs. 14 and 15 compare the use of the cross correlation function Xcorr to XcorrNorm′ using a different experimental MS/MS data set 260 than the data set 260 used for Figs. 12 and 13. The probe used to label the sample measured in Figs. 14 and 15 is different than the probe used to label the sample measured in Figs. 12 and 13. The probe used to label the proteome measured in Figs. 14 and 15 modifies cysteine residues whereas the probe used to label the proteome measured in Figs. 12 and 13 modifies serine residues. The only difference between the cross correlation scores in Fig. 15 and Fig. 14 is that the XcorrNorm′ cross correlation function was used to compute the ±10 incorrect cross correlation scores (steps 1004 and 1006) and true experimental cross correlation scores (step 1002) in Fig. 15 whereas the Xcorr cross correlation function was used to compute these correlation scores in Fig. 14. In Fig. 14, 23 cross correlation scores have a confidence of not being part of the false correlation score gamma distribution curve that is greater than 99 percent. In Fig. 15, 64 cross correlation scores have a confidence of not being part of the false correlation score gamma distribution curve that is greater than 99 percent, which is 2.8 fold improvement over Fig. 14. Thus, in the case of the data set 260 used to generate Figs. 14 and 15, XcorrNorm′ produces better results than Xcorr.

Reference has been made (*e.g.*, Fig. 10) to process steps that are used to produce a false positive gamma distribution curve fit for a probe-labeled sample. The same techniques can be applied when samples that have not been labeled with a probe are studied. For instance, the parent ion information can be altered prior to the search. Thus, another aspect of the present invention provides a method of computing a false positive gamma distribution curve for a data set of MS/MS spectra in which all or a portion of the MS/MS spectra in the data set respectively correspond to a different parent ion. In the method, for each respective MS/MS spectrum in all or a portion of the data set, a sequence fragment from a biological macromolecule database 270 is assigned to the respective MS/MS spectrum using a first search parameter that includes the mass of the parent ion of the respective MS/MS spectrum upwardly adjusted by a first amount. Then, for each respective MS/MS spectrum in all or the portion of the set, a cross correlation score is computed using the false assignment made for the respective MS/MS spectra. Each of these computed cross correlation scores is fitted to a gamma distribution curve thereby computing a false positive gamma distribution curve for the data set. In some embodiments, the first amount that each respective parent ion is upwardly adjusted by in the first search parameter is between 2 daltons and 50 daltons.

In some embodiments, each respective MS/MS spectrum in the data set is assigned another sequence fragment from a biological macromolecule database using a second search parameter that also includes an incorrect mass for the parent ion of the respective MS/MS spectrum. In some embodiments, the mass that is used in the second search parameter for the search of a respective MS/MS spectrum in the data set is the mass of the parent ion of the respective MS/MS spectrum downwardly adjusted by a second amount. The second amount can be, for example, between 2 daltons and 50 daltons. Then, for each respective MS/MS spectrum in the data set, a cross correlation score is computed using the assignments made for the MS/MS spectrum based upon the second search parameter. In embodiments were such dual assignments and such dual correlation scores are made for each respective spectra in a dataset, each of the cross correlation scores are fitted to a gamma distribution curve. In this way, a false positive gamma distribution curve is made for a data set of MS/MS spectra.

5.5 SORTING TANDEM MASS SPECTROMETRY DATA BASED ON THE PRESENCE OF PROBE FRAGMENT PEAKS

5.5.1 Sorting MS/MS data based on the presence of a probe fragment

Another aspect of the present invention processes tandem mass spectrometry data 260 based on the presence and levels of probe fragments within spectra 262 (Fig. 2). This aspect of the invention is best introduced by way of examples. In one example, the fluorophosphonate probe FP-Peg-TAMRA (Fig. 16A) was reacted with mouse liver, mouse testis, and mouse brain cytosol. The sample was processed for LC/MS/MS analysis. In addition to efficiently cleaving off the parent peptide, some of the FP-Peg-TAMRA probe will fragment into predictable fragments. The spectra 262 were sorted by MS/MS data sorting module 246 (Fig. 2) based on criterion of whether or not such probe fragments were present in the spectra. In one embodiment, each spectra 262 trace was converted to an ASCII ".dta" file in order to perform this search. Several ASCII ".dta" files were created for each spectra 262 in the data set 260. Each ".dta" file, for a given spectra 262, represented a different parent ion charge state (*e.g.*, -3, -2, -1, +1, +2, +3, *etc.*). If a given ".dta" filed contained probe fragments, then the ".dta" file was further filtered to determine the parent ion charge state using the techniques described in Section 5.2. Further, ".dta" files that did not have predicted probe fragments present or had multiple probes present were removed in order to prevent the generation of false positive data.

5.5.2 Method for sorting MS/MS data into a false positive bin and an experimental bin without the use of cross correlation scores

Fluorophosphonate probes (*e.g.*, the probe illustrated in Fig. 16A) have been shown to react selectively with the active site serine nucleophile of most serine hydrolases. For example, FP-Peg-TAMRA contains a fluorosphosphonate moiety known to selectively react with serine residues at the active site of serine hydrolase enzymes. The active site serine of serine hydrolases typically lies within a consensus GXSXG (SEQ ID NO: 1) motif. To separate search results into "positive" and "false-positive" pools, the sequences were analyzed to determine whether the peptide identified by the SEQUEST run contained the active site serine nucleophile of a serine hydrolase. If the protein was not annotated as a serine hydrolase, the sequence was searched against a non-redundant database to determine if the protein was significantly homologous to known serine hydrolases. Search results indicating the probe attached to a known or predicted serine hydrolase active site serine were segregated into the "positive" pool while all other results were labeled as "false-positive."

In one example, a set of three tissue samples (mouse liver, kidney, and submaxillary) was reacted with this probe and analyzed as described above. The results from SEQUEST searching of the MS/MS data were sorted into positive and false-positive pools. Any peptide that contained the active site serine nucleophile of a serine hydrolase enzyme was considered positive, and all other peptides were considered to be false positives. There are approximately 500 serine hydrolases in the mouse genome and approximately 15,000 protein sequences. On average, each serine hydrolase will yield roughly 40 tryptic fragment peptides, only one of which will contain the active site tryptic nucleophile. Thus the chances of randomly observing a serine hydrolase active site fragment in the data set is approximately 0.08%. In the exemplary data set here, 2443 total SEQUEST results were used, which would be expected to contain roughly two serine hydrolase active site peptides through random chance. The data set contained 938 serine hydrolase active site peptides suggesting that the data set assigned as positive results is highly accurate.

5.5.3 General method for sorting MS/MS spectrometry data.

The present invention advantageously exploits the information that can be derived from the use of probe labels to reduce the number of false positive assignments that are made in a given data set 260. Fig. 17 illustrates one such method. The method illustrated in Fig. 17 incorporates various aspects of the invention that are described in other sections. In some embodiments, one or more steps in Fig. 17 are performed by MS/MS sorting module 246 (Fig. 2).

In step 1702 each MS/MS spectrum 262 in a data set 260 is processed. To generate data set 260, a TAMRA probe was added to samples containing the proteome from an organism and allowed to react with the proteome. The samples were then denatured with urea, reduced with dithiothreitol, and alkylated with iodoacetamide. The samples were then gel filtered and digested with trypsin. The probe-modified peptides were purified from the mixture using anti-TAMRA monoclonal antibodies immobilized on agarose beads prior to analysis by liquid chromatography-tandem mass spectrometry (LCMS/MS). This analytical separation results in a plurality of fractions. A tandem mass spectrum 262 of each fraction is taken, resulting in a plurality of the MS/MS spectrums 262. Many other methods other than the one described in this section may be used to generate data set 260 and all such methods are within the scope of the present invention. The method described for generating data set 260 provided in this section is merely by way of example and not by limitation.

At the initial stages of processing, the parent ion charge state is unknown. Therefore, multiple parent ion charge states were considered for each MS/MS spectrum 262 as follows:

```
For each MS/MS spectrum 262 in experimental data set 260 {
        For each parent ion charge state under consideration {
                Generate a file from the spectrum 262 trace that includes the
                peaks from the trace and the given parent ion charge state
}}
```

For example, if the parent ion charge states under consideration are +2, +3, and +4, then +2, +3, and +4 files are created for the 262-1 spectrum, etc. Each file includes the peaks in the corresponding spectrum 262 as well as a particular parent ion charge state. Step 1702 results in the generation of a number of files. If three different parent ion charge states are under consideration, then there are three different files for each spectrum 262 in data set 260.

In step 1704 each file generated in step 1702 is screened to determine whether the file can be placed in a particular folder. Step 1704 comprises steps 1706 through 1712. Steps 1706 through 1712 are repeated for each spectrum 262 in data set 260. In step 1706 a file for the spectrum under consideration is obtained. In step 1708, a determination is made as to whether the file includes a probe fragment. The file will contain a probe fragment if the probe is efficiently cleaved off of the parent ion during the generation of the spectrum 262 that corresponds to the file. Probes used to label proteins in a proteome will generate characteristic peaks. For example, Fig. 3

illustrates the characteristic peak fragments of the probe FP-Peg-TAMRA in the absence of protein. The molecular structure of FP-Peg-TAMRA is illustrated in Fig. 16A. Fig. 6 illustrates how the FP-Peg-TAMRA probe fragments are also observed when probe-modified peptides are fragmented. Taken together, Figs. 3 and 6 illustrate that MS/MS spectra can be sorted based on the presence or absence of specific probe fragment peaks.

If a peak corresponding to the probe or to probe fragments is not observed in the file (1708-No) then the file is placed in a first folder (1710). At a later stage, each file in the first folder is searched against biological macromolecule database 260 based on the assumption that the label has labeled a first residue class thereby forming a bond that is not efficiently cleaved during the generation of the MS/MS spectrum 262 that the file represents. This residue class could be, for example, tyrosine. The label can react with the tyrosine hydroxyl group thereby forming a complex that is not readily cleaved under certain tandem mass spectrometry conditions.

If the file does contain characteristic label peaks (1708-No) then the file is not placed into the first folder. In step 1712, a determination is made as to whether there are any additional files, for the spectrum under consideration, that have not been evaluated by step 1708. If so, (1712-Yes) process control returns to 1706 where another file corresponding to the spectrum under consideration is obtained. If not (1712-No), another spectrum 262 in data set 260 is obtained and steps 1706 through 1712 are repeated using the new spectrum 262. When each spectrum 262 in data set 260 has been evaluated in this manner, process control passes to step 1760 (Fig. 17B).

Step 1760, in fact, comprises steps 1762 through 1770. Each spectrum 262 in data set 260 is individually evaluated in step 1760. For a given file that represents the spectrum under consideration, a determination is made as to whether the file has a valid parent ion charge state. To determine whether a file has a valid parent ion charge state, the parent ion charge state (*e.g.*, +3, +4, +5) that is assigned to the file is used in Eqn. 2 of Section 5.2, above. If the file includes a peak at the m/z value computed using Eqn. 2 (1762-Yes), then the file has a valid parent ion charge state. It should be noted that more than one file corresponding to a given spectrum could have a valid parent ion charge state. This is because it is possible that the spectra 262 could have a peak having the m/z value computed using Eqn. 2 because of cleavage of the peptide into a fragment rather than cleavage of the probe from the parent ion. If the file does not include a peak at the m/z value computed using Eqn. 2 (1762-No), then the a determination is made as to whether any file corresponding to the given spectrum 262 under consideration has a valid parent ion charge state (1764). If so, (1764-Yes), then the file that does not have a valid parent ion charge state is deleted (step 1768). Step 1768 is advantageous because it removes a file that is very unlikely to have a valid parent ion charge state. A file is discarded in step 1768 when it does have a probe fragments (1708-Yes), does not have a valid parent ion charge state (1762-No), and another file corresponding the same spectra 262 represented by the discarded file does have a valid parent ion charge state (1764-Yes).

In the case where the file does have a valid ion parent ion charge state or in the case where no file that represents the spectra 262 under consideration has a valid parent ion charge state, the probe fragments are removed from the files corresponding to the spectra 262 (step 1766). Step 1766 is advantageous because it removes peaks from the files that would not assist in identifying a matching sequence fragment from biological macromolecule database 270. Next, the file is placed in a second folder. At a later stage, each file in the second folder is searched against biological macromolecule database 260 based on the assumption that the label has labeled a second residue class thereby forming a bond that is efficiently cleaved during the generation of the MS/MS spectrum 262 that the file represents. This residue class could be, for example, serine.

The process steps illustrated in Fig. 17 are advantageous because they use information that can be derived from probe label fragmentation patterns, or the lack thereof, to limit the types of probe parameters that are used in subsequent searches for matching sequence fragments in database 270. This reduces the number of false positives that are identified. The files that represent invalid parent ion charge states are eliminated. The files that do not include probe fragments are not searched using probe parameters that assume efficient cleavage of the probe (*e.g.*, do not assume a serine modification). The files that include probe fragments in a valid parent ion charge state are searched using probe parameters that assume modification of a residue class that is efficiently cleaved (*e.g.*, do not assume a tyrosine modification).

One of skill in the art will appreciate that several variations of the process steps illustrate in Fig. 17 are possible and all such modifications are within the scope of the present invention. For example, fragment patterns are often different depending on the nature of the attachment point between the probe and the peptide. For example, a peptide modified on a cysteine residue often give a probe fragment that has the mass of the probe plus the mass of a sulfer atom (32 daltons), while a serine modified peptide gives a fragment corresponding to the probe plus the mass of an oxygen (16 daltons), and a peptide modified on tyrosine gives only internal probe fragments. Distinct patterns like this can be used to sort data into separate folders that subsequently get searched with the appropriate potential residue modifications.

5.6 REFERENCES CITED

All references cited herein are incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or patent or patent application was specifically and individually indicated to be incorporated by reference in its entirety for all purposes.

5.7 ALTERNATIVE EMBODIMENTS

The present invention can be implemented as a computer program product that comprises a computer program mechanism embedded in a computer readable storage medium. For instance, the computer program product could contain the program modules shown in Fig. 2. These program modules may be stored on a CD-ROM, magnetic disk storage product, or any other computer readable data or program storage product. The software modules in the computer program product may also be distributed electronically, via the Internet or otherwise, by transmission of a computer data signal (in which the software modules are embedded) on a carrier wave.

Many modifications and variations of this invention can be made without departing from its spirit and scope, as will be apparent to those skilled in the art. The specific embodiments described herein are offered by way of example only, and the invention is to be limited only by the terms of the appended claims, along with the full scope of equivalents to which such claims are entitled.

What is claimed:

1. A method of determining a parent ion charge state for a tandem MS spectrum of said parent ion, wherein said parent ion has been labeled with a probe, the method comprising:

choosing a candidate parent ion charge state; and

searching the tandem MS spectrum for a peak having the value **F**, wherein

F = [(the m/z of the parent ion) x (the candidate parent ion charge state) – (the mass of the probe) – (the mass of a portion of the parent ion that is removed upon probe cleavage)-(the mass of any protons carried on the probe] / [(the candidate parent ion charge state) – (the carried charge of the probe)];

and wherein, when a peak, having a value that is greater than a first predetermined threshold value, appears in the tandem MS spectrum that is within a second predetermined threshold value of **F**, said candidate parent ion charge state is said parent ion charge state for said tandem MS spectrum.

2. The method of claim 1 wherein said first predetermined threshold value is a ratio between the size of said peak and the size of the largest peak in said tandem MS spectrum and wherein said peak has said value that is greater than said first predetermined threshold value when said ratio is greater than 5:100 (0.05).

3. The method of claim 1 wherein said second predetermined threshold value is between ±0.1 and ±3.

4. The method of claim 1 wherein said probe is a fluorophosphonate.

5. The method of claim 1 wherein said parent ion is a peptide.

6. The method of claim 1 wherein said candidate parent ion charge state is +2, +3, +4, +5, +6, +7, +8, or +9.

7. The method of claim 1 wherein said candidate parent ion charge state is -2, -3, -4, -5, -6, -7, -8, or -9.

8. The method of claim 1 wherein the mass of a portion of the parent ion that is removed upon probe cleavage is the mass of one proton.

9. The method of claim 1 wherein the probe charge is +1.

10. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

an experimental MS/MS data set comprising a plurality of tandem MS spectra;

a biological macromolecule database for storing biological macromolecule information;

and a tandem MS processing module for assigning tandem MS spectra to sequence fragments in said biological macromolecule database, wherein

(i) said tandem MS processing module includes a charge state determination module for determining a parent ion charge state for a first tandem MS spectrum of said parent ion,

(ii) said first tandem MS spectrum is in said experimental MS/MS data set, and

(iii) said parent ion has been labeled with a probe,

the tandem MS processing module comprising:

instructions for choosing a candidate parent ion charge state; and

instructions for searching said first tandem MS spectrum for a peak having the value **F**, wherein

F = [(the m/z of the parent ion) x (the candidate parent ion charge state) – (the mass of the probe) – (the mass of a portion of the parent ion that is removed upon probe cleavage)-(the mass of any protons carried on the probe] / [(the candidate parent ion charge state) – (the carried charge of the probe)];

and wherein, when a peak, having a value that is greater than a first predetermined threshold value, appears in the tandem MS spectrum that is within a second predetermined threshold value of **F**, said candidate parent ion charge state is said parent ion charge state for said tandem MS spectrum.

11. The computer product of claim 10 wherein said first predetermined threshold value is a ratio between the size of said peak and the size of the largest peak in said tandem MS spectrum and wherein said peak has said value that is greater than said first predetermined threshold value when said ratio is greater than 5:100 (0.05).

12. A computer system comprising:

a central processing unit;

a memory, coupled to the central processing unit, the memory storing:

an experimental MS/MS data set comprising a plurality of tandem MS spectra;

a biological macromolecule database for storing biological macromolecule information;

and a tandem MS processing module for assigning tandem MS spectra to sequence fragments in said biological macromolecule database, wherein

(i) said tandem MS processing module includes a charge state determination module for determining a parent ion charge state for a first tandem MS spectrum of said parent ion,

(ii) said first tandem MS spectrum is in said experimental MS/MS data set, and

(iii) said parent ion has been labeled with a probe,

the tandem MS processing module comprising:

instructions for choosing a candidate parent ion charge state; and

instructions for searching said first tandem MS spectrum for a peak having the value **F**, wherein

F = [(the m/z of the parent ion) x (the candidate parent ion charge state) – (the mass of the probe) – (the mass of a portion of the parent ion that is removed upon probe cleavage)-(the mass of any protons carried on the probe] / [(the candidate parent ion charge state) – (the carried charge of the probe)];

and wherein, when a peak, having a value that is greater than a first predetermined threshold value, appears in the tandem MS spectrum that is within a second predetermined threshold value of **F**, said candidate parent ion charge state is said parent ion charge state for said tandem MS spectrum.

13. The computer system of claim 12 wherein said first predetermined threshold value is a ratio between the size of said peak and the size of the largest peak in said tandem MS spectrum and wherein said peak has said value that is greater than said first predetermined threshold value when said ratio is greater than 5:100 (0.05).

14. A method of computing a cross correlation score between a predicted spectrum **PS** and an experimental spectrum **ES**, wherein the experimental spectrum **ES** is the fragmentation pattern of a parent ion, the method comprising

computing a value XcorrNorm, wherein

$$\mathrm{XcorrNorm} = \mathrm{Xcorr} \times \sqrt{\frac{\mathrm{Charge} + C_1}{\mathrm{Mass} - C_2}}$$

wherein

Xcorr is the cross correlation score between **PS** and **ES**;

C_1 is a first constant;

C_2 is a second constant;

Charge is the parent ion charge state; and

Mass is a mass of the parent ion.

15. The method of claim 14 wherein

C_1 is between −10 and +10, and

C_2 is between −2000 and +2000.

16. The method of claim 14 wherein said parent ion is labeled with a probe and C_2 is equal to the mass of the probe + 324 daltons.

17. The method of claim 14 wherein C_1 is about 4.26.

18. The method of claim 14 wherein said probe is a fluorophosphonate.

19. The method of claim 14 wherein Xcorr is computed by:

setting a relative displacement value $n\Delta t$ to a lower bound;

computing a score $C_{ab}(n\Delta t)$ wherein

$$C_{ab}(n\Delta t) = \frac{1}{T}\sum_{t=0}^{T} PS(t)ES(t \pm n\Delta t)$$

wherein Δt is a sampling interval;

incrementing $n\Delta t$;

repeating said computing an incrementing until $n\Delta t$ exceeds an upper bound; and

assigning Xcorr the value of $C_{ab}(0)$ minus the mean of $C_{ab}(n\Delta t)$ over the range lower bound < $n\Delta t$ < upper bound.

20. The method of claim 19 wherein said lower bound is –75 and said upper bound is 75.

21. The method of claim 14 wherein Xcorr is computed using a Fourier transform.

22. The method of claim 14 wherein the method further comprises:

computing a value XcorrNorm′, wherein the value XcorrNorm′ equals XcorrNorm + (C_4 * DCN) wherein DCN is the fractional difference between:

(i) a cross correlation score Xcorr_1 of the highest scoring sequence fragment in a biological macromolecule database for **ES**, and

(ii) a cross correlation score Xcorr_2 of the second highest sequence fragment in the biological macromolecule database for **ES**.

23. The method of claim 22 wherein C_4 is a number between 1.0 and 5.0.

24. The method of claim 22 wherein C_4 is 1.54.

25. A method of computing a cross correlation score between (i) a first predicted spectrum corresponding to a first sequence fragment and (ii) an experimental

spectrum, wherein the experimental spectrum is the fragmentation pattern of a parent ion, the method comprising:

computing a value XcorrNorm″, wherein

$$\text{XcorrNorm}'' = \left[\frac{C_5 * \text{Xcorr}}{\sqrt{(\text{Sequence fragment} + 1\ \text{dalton}) + \text{Mass}_{\text{probe modification}} - C_6}}\right] + C_7 * \left[1 - \frac{\text{Xcorr}_{2\text{nd}}}{\text{Xcorr}}\right]$$

wherein

Xcorr is the cross correlation score between the first predicted spectrum and the experimental spectrum;

C_5, C_6, and C_7 are each constants;

"Sequence fragment + 1 dalton" is the mass of the first sequence fragment incremented by 1 dalton;

$\text{Mass}_{\text{probe modification}}$ is an increase in the mass of the first sequence fragment as a result of a modification of the first sequence fragment by a probe; and

$\text{Xcorr}_{2\text{nd}}$ is the cross correlation score between the experimental spectrum and a second predicted spectrum corresponding to a second sequence fragment.

26. The method of claim 25 wherein

$$\text{XcorrNorm}'' = \left[\frac{60 * \text{Xcorr}}{\sqrt{(\text{Sequence fragment} + 1\ \text{dalton}) + \text{Mass}_{\text{probe modification}} - 667}}\right] + 0.8 * \left[1 - \frac{\text{Xcorr}_{2\text{nd}}}{\text{Xcorr}}\right]$$

27. The method of claim 25 wherein C_5 is a value between 10 and 500.

28. The method of claim 25 wherein C_6 is a value between 300 and 5000.

29. The method of claim 25 wherein C_7 is a coefficient between 0.1 and 1.0.

30. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a

computer program mechanism embedded therein, the computer program mechanism comprising:

an experimental MS/MS data set comprising a plurality of tandem MS spectra;

a biological macromolecule database for storing biological macromolecule information;

and a tandem MS processing module for assigning tandem MS spectra to sequence fragments in said biological macromolecule database, wherein said tandem MS processing module includes a correlation determination module for computing a cross correlation score between a predicted spectrum **PS** and an experimental spectrum **ES**, wherein the experimental spectrum **ES** is the fragmentation pattern of a parent ion and wherein said experimental spectrum **ES** is in said experimental MS/MS data set, said correlation determination module comprising

instructions for computing a value XcorrNorm, wherein

$$\mathrm{XcorrNorm} = \mathrm{Xcorr} \times \sqrt{\frac{\mathrm{Charge} + \mathrm{C}_1}{\mathrm{Mass} - \mathrm{C}_2}}$$

wherein

Xcorr is the cross correlation score between **PS** and **ES;**

C_1 is a first constant;

C_2 is a second constant;

Charge is the parent ion charge state; and

Mass is the mass of the parent ion.

31. A computer system for processing tandem MS data, the computer system comprising:

a central processing unit;

a memory, coupled to the central processing unit, the memory storing:

an experimental MS/MS data set comprising a plurality of tandem MS spectra;

a biological macromolecule database for storing biological macromolecule information;

and a tandem MS processing module for assigning tandem MS spectra to sequence fragments in said biological macromolecule database, wherein said tandem MS processing module includes a correlation determination module for computing a cross correlation score between a predicted spectrum **PS** and an experimental spectrum **ES**, wherein the experimental spectrum **ES** is the fragmentation pattern of a parent ion and wherein said experimental spectrum **ES** is in said experimental MS/MS data set, said correlation determination module comprising

instructions for computing a value XcorrNorm, wherein

$$\text{XcorrNorm} = \text{Xcorr} \times \sqrt{\frac{\text{Charge} + \text{C}_1}{\text{Mass} - \text{C}_2}}$$

wherein

Xcorr is the cross correlation score between **PS** and **ES;**

C_1 is a first constant;

C_2 is a second constant;

Charge is the parent ion charge state; and

Mass is the mass of the parent ion.

32. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

an experimental MS/MS data set comprising a plurality of tandem MS spectra;

a biological macromolecule database for storing biological macromolecule information;

and a tandem MS processing module for assigning tandem MS spectra to sequence fragments in said biological macromolecule database, wherein said tandem MS processing module includes a correlation determination module for computing a cross correlation score between (i) a first predicted spectrum corresponding to a first sequence fragment in said biological macromolecule database and (ii) an experimental spectrum, wherein the experimental spectrum is the fragmentation pattern of a parent ion, said correlation determination module comprising:

instructions for computing a value XcorrNorm″, wherein

$$\text{XcorrNorm}'' = \left[\frac{\text{C}_5 * \text{Xcorr}}{\sqrt{(\text{Sequence fragment} + 1\ \text{dalton}) + \text{Mass}_{\text{probe modification}}} - \text{C}_6}\right] + \text{C}_7 * \left[1 - \frac{\text{Xcorr}_{2\text{nd}}}{\text{Xcorr}}\right]$$

wherein

Xcorr is the cross correlation score between the first predicted spectrum and the experimental spectrum;

C_5, C_6, and C_7 are each constants;

"Sequence fragment + 1 dalton" is the mass of the first sequence fragment incremented by 1 dalton;

$\text{Mass}_{\text{probe modification}}$ is an increase in the mass of the first sequence fragment as a result of a modification of the first sequence fragment by a probe; and

Xcorr_{2nd} is the cross correlation score between the experimental spectrum and a second predicted spectrum corresponding to a second sequence fragment in said biological macromolecule database.

33. A method of computing a false positive gamma distribution curve for a data set of MS/MS spectra wherein all or a portion of the MS/MS spectra in the data set respectively correspond to a different parent ion that has been labeled with a probe, the method comprising:

(A) assigning, for each respective MS/MS spectrum in all or a portion of said data set, a sequence fragment from a biological macromolecule database to said respective MS/MS spectrum using a search parameter that includes a first incorrect mass for said probe;

(B) computing, for each respective MS/MS spectrum in all or said portion of said data set, a cross correlation score using the assignment made for said respective MS/MS spectrum in step (A); and

(C) fitting each cross correlation score computed in step (B) to a gamma distribution curve thereby computing a false positive gamma distribution curve for a data set of MS/MS spectra.

34. The method of claim 33 wherein said first incorrect mass for said probe is the mass of said probe plus or minus an amount that is between 2 daltons and 50 daltons.

35. The method of claim 33, further comprising:

(D) for each respective MS/MS spectrum in said data set, assigning a sequence fragment from a biological macromolecule database to said respective MS/MS spectrum using a search parameter that includes a second incorrect mass for said probe;

(E) for each respective MS/MS spectrum in said data set, computing a cross correlation score using the assignment made for said respective MS/MS spectrum in step (D); and wherein

said step (C) comprises fitting each cross correlation score computed in step (B) and step (E) to a gamma distribution curve thereby computing a false positive gamma distribution curve for a data set of MS/MS spectra.

36. The method of claim 35 wherein said first incorrect mass for said probe is the mass of said probe plus an amount and said second incorrect mass for said probe is the mass of said probe minus an amount.

37. The method of claim 36 wherein said amount is between 2 daltons and 50 daltons.

38. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

an experimental MS/MS data set comprising a plurality of tandem MS spectra;

a biological macromolecule database for storing biological macromolecule information;

and a tandem MS processing module for assigning tandem MS spectra to sequence fragments in said biological macromolecule database, wherein said tandem MS processing module includes a false data profiling module for computing a false positive gamma distribution curve for said experimental MS/MS data set wherein all or a portion of the MS/MS spectra in the data set respectively correspond to a different parent ion that has been labeled with a probe, the false data profiling module comprising:

(A) instructions for assigning, for each MS/MS spectrum in all or a portion of said experimental data set, a sequence fragment from said biological macromolecule database to said MS/MS spectrum using a search parameter that includes a first incorrect mass for said probe;

(B) instructions for computing, for each MS/MS spectrum in all or said portion of said data set, a cross correlation score using the assignment made for said MS/MS spectrum by said instructions for assigning; and

(C) instructions for fitting each cross correlation score computed by said instructions for computing (B) to a gamma distribution curve.

39. A computer system for processing tandem MS data, the computer system comprising:

a central processing unit;

a memory, coupled to the central processing unit, the memory storing:

an experimental MS/MS data set comprising a plurality of tandem MS spectra;

a biological macromolecule database for storing biological macromolecule information; and

a tandem MS processing module for assigning tandem MS spectra to sequence fragments in said biological macromolecule database, wherein said tandem MS processing module includes a false data profiling module for computing a false positive gamma distribution curve for said experimental MS/MS data set wherein all or a portion of the MS/MS spectra in the data set respectively correspond to a different parent ion that has been labeled with a probe, the false data profiling module comprising:

(A) instructions for assigning, for each MS/MS spectrum in all or a portion of said experimental data set, a sequence fragment from said biological macromolecule database to said MS/MS spectrum using a search parameter that includes a first incorrect mass for said probe;

(B) instructions for computing, for each MS/MS spectrum in all or said portion of said data set, a cross correlation score using the assignment made for said MS/MS spectrum by said instructions for assigning; and

(C) instructions for fitting each cross correlation score computed by said instructions for computing (B) to a gamma distribution curve.

40. A method of reducing a number of false positive assignments made to MS/MS spectra in a tandem MS data set, wherein each MS/MS spectrum in all or a portion of the MS/MS spectra in said tandem MS data set represents a fragmentation pattern of a different parent ion that has been labeled with a probe, the method comprising:

(A) generating a plurality of files for a MS/MS spectrum in said tandem MS data set, each file in said plurality of files including the peaks from the MS/MS spectrum and each file in said plurality of files representing a different candidate parent ion charge state for said MS/MS spectrum;

(B) determining, for each file in said plurality of files, whether the file includes a peak corresponding to a fragment of said probe, wherein when a file does not contain said peak corresponding to said fragment of said probe, said file is given a first designation and removed from said plurality of files;

(C) determining, for a first file in said plurality of files, whether said first file has a valid parent ion charge state, wherein

when another file in said plurality of files has a valid parent ion charge state and said first file does not have a valid parent ion charge state, said first file is deleted and removed from said plurality of files;

(D) giving said first file a second designation when said file remains in said plurality of files after step (C); and

(E) searching a biological macromolecule database using a file given said second designation with the criterion that the probe modified a predetermined moiety class.

41. The method of claim 40 wherein said predetermined moiety class is a residue of a predetermined amino acid.

42. The method of claim 40 wherein said predetermined amino acid is serine.

43. The method of claim 40, the method further comprising:

(F) searching said biological macromolecule database using a file given said first designation with the assumption that the probe modified a different predetermined moiety class.

44. The method of claim 43 wherein said different predetermined moiety is a tyrosine residue.

45. The method of claim 40 wherein said fragment of said probe represents the probe upon cleavage from the parent ion corresponding to said MS/MS spectrum in said tandem MS data set.

46. The method of claim 40 wherein said steps (A) through (E) are repeated for said all or said portion of the MS/MS spectra in said tandem MS data set.

47. The method of claim 40 wherein said first file has said valid parent ion charge when there is a peak having the value **F** in said first file, wherein

F = [(the m/z of the parent ion of said MS/MS spectrum) x (the parent ion charge state assigned to said first file) – (the mass of the probe) – (the mass of a portion of the parent ion that is removed upon probe cleavage)] / [(the parent ion charge state assigned to said first file) – (the carried charge of the probe)].

48. A computer program product for use in conjunction with a computer system, the computer program product comprising a computer readable storage medium and a computer program mechanism embedded therein, the computer program mechanism comprising:

an experimental MS/MS data set comprising a plurality of tandem MS spectra;

a biological macromolecule database for storing biological macromolecule information; and

a tandem MS processing module for assigning tandem MS spectra to sequence fragments in said biological macromolecule database, wherein said tandem MS processing module includes a MS/MS data sorting module for reducing a number of false positive assignments made to tandem MS spectra in said experimental tandem MS data set, wherein each tandem microcopy spectrum in all or a portion of the tandem microcopy spectra in said tandem MS data set represents a fragmentation pattern of a different parent ion that has been labeled with a probe, wherein said MS/MS data sorting module comprises:

(A) instructions for generating a plurality of files for a tandem MS spectrum in said experimental tandem MS data set, each file in said plurality of files including the peaks from the tandem MS spectrum and each file in said plurality of files representing a different candidate parent ion charge state for said tandem MS spectrum;

(B) instructions for determining, for each file in said plurality of files, whether the file includes a peak corresponding to a fragment of said probe, wherein when a file does not contain said peak corresponding to said fragment of said probe, said file is given a first designation and removed from said plurality of files;

(C) instructions for determining, for a first file in said plurality of files, whether said first file has a valid parent ion charge state, wherein

when another file in said plurality of files has a valid parent ion charge state and said first file does not have a valid parent ion charge state, said first file is deleted and removed from said plurality of files;

(D) instructions for giving said first file a second designation when said file remains in said plurality of files after said instructions for determining (C); and

(E) instructions for searching said biological macromolecule database using a file given said second designation with the criterion that the probe modified a predetermined moiety class.

49. A computer system for processing tandem MS data, the computer system comprising:

a central processing unit;

a memory, coupled to the central processing unit, the memory storing:

an experimental MS/MS data set comprising a plurality of tandem MS spectra;

a biological macromolecule database for storing biological macromolecule information; and

a tandem MS processing module for assigning tandem MS spectra to sequence fragments in said biological macromolecule database, wherein said tandem MS processing module includes a MS/MS data sorting module for reducing a number of false positive assignments made to tandem MS spectra in said experimental tandem MS data set, wherein each tandem microcopy spectrum in all or a portion of the tandem microcopy spectra in said tandem MS data set represents a fragmentation pattern of a different parent ion that has been labeled with a probe, wherein said MS/MS data sorting module comprises:

(A) instructions for generating a plurality of files for a tandem MS spectrum in said experimental tandem MS data set, each file in said plurality of files including the peaks from the tandem MS spectrum and each file in said plurality of files representing a different candidate parent ion charge state for said tandem MS spectrum;

(B) instructions for determining, for each file in said plurality of files, whether the file includes a peak corresponding to a fragment of said probe, wherein when a file does not contain said peak corresponding to said fragment of said probe, said file is given a first designation and removed from said plurality of files;

(C) instructions for determining, for a first file in said plurality of files, whether said first file has a valid parent ion charge state, wherein

when another file in said plurality of files has a valid parent ion charge state and said first file does not have a valid parent ion charge state, said first file is deleted and removed from said plurality of files;

(D) instructions for giving said first file a second designation when said file remains in said plurality of files after said instructions for determining (C); and

(E) instructions for searching said biological macromolecule database using a file given said second designation with the criterion that the probe modified a predetermined moiety class.

50. A method of computing a false positive gamma distribution curve for a data set of MS/MS spectra wherein all or a portion of the MS/MS spectra in the data set respectively correspond to a different parent ion, the method comprising:

(A) assigning, for each respective MS/MS spectrum in all or a portion of said data set, a sequence fragment from a biological macromolecule database to said respective MS/MS spectrum using a first search parameter that includes the mass of the parent ion of said respective MS/MS spectrum adjusted by a first amount;

(B) computing, for each respective MS/MS spectrum in all or said portion of said data set, a cross correlation score using the assignment made for said respective MS/MS spectrum in step (A); and

(C) fitting each cross correlation score computed in step (B) to a gamma distribution curve thereby computing a false positive gamma distribution curve for a data set of MS/MS spectra.

51. The method of claim 50 wherein the first amount is between 2 daltons and 50 daltons.

52. The method of claim 50 wherein the first amount is between -2 daltons and -50 daltons.

53. The method of claim 50, further comprising:

(D) for each respective MS/MS spectrum in said data set, assigning a sequence fragment from a biological macromolecule database to said MS/MS spectrum using a second search parameter that includes the mass of the parent ion of said respective MS/MS spectrum adjusted by a second amount;

(E) for each respective MS/MS spectrum in said data set, computing a cross correlation score using the assignment made for said respective MS/MS spectrum in step (D); and wherein

said step (C) comprises fitting each cross correlation score computed in step (B) and step (E) to a gamma distribution curve thereby computing a false positive gamma distribution curve for a data set of MS/MS spectra.

54. The method of claim 53 wherein said first incorrect mass for said parent ion is the mass of said parent ion plus an amount and said second incorrect mass for said parent ion is the mass of said parent ion minus an amount.

55. The method of claim 54 wherein said amount is between 2 daltons and 50 daltons.

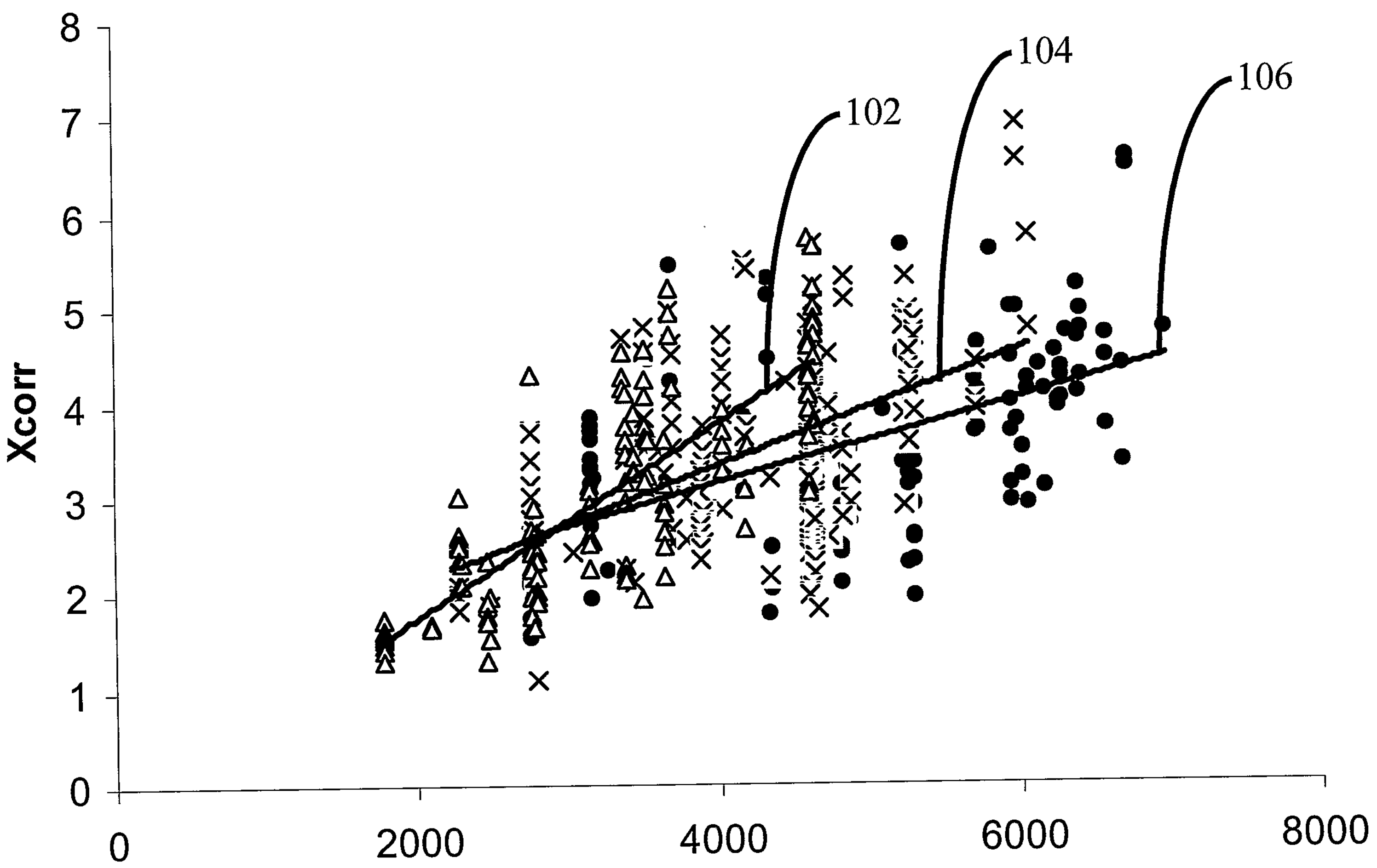


Fig. 1
(Prior Art)

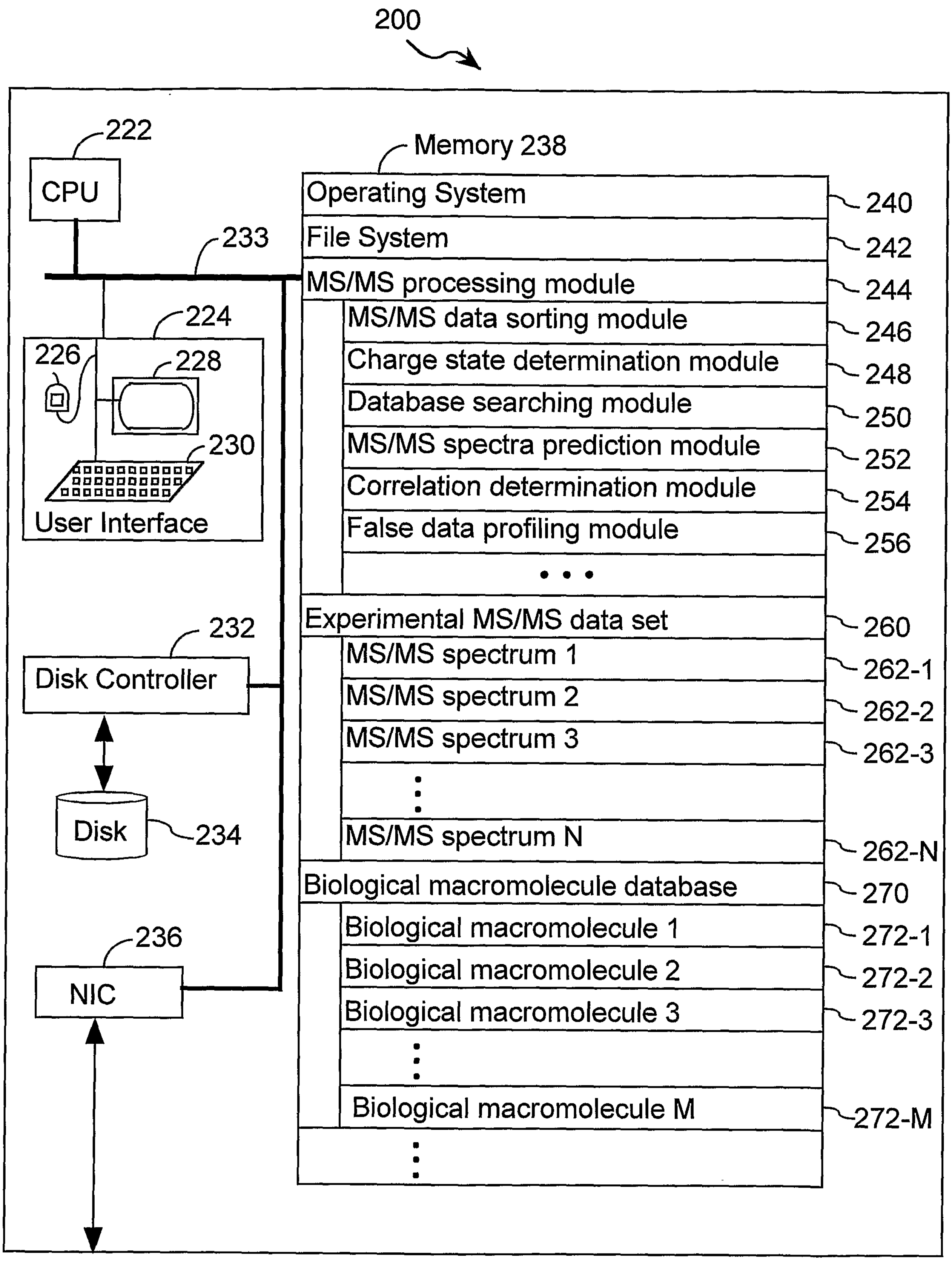

200
222
CPU
233
224
226
228
230
User Interface
232
Disk Controller
Disk
234
236
NIC
Memory 238
Operating System
240
File System
242
MS/MS processing module
244
MS/MS data sorting module
246
Charge state determination module
248
Database searching module
250
MS/MS spectra prediction module
252
Correlation determination module
254
False data profiling module
256
Experimental MS/MS data set
260
MS/MS spectrum 1
262-1
MS/MS spectrum 2
262-2
MS/MS spectrum 3
262-3
MS/MS spectrum N
262-N
Biological macromolecule database
270
Biological macromolecule 1
272-1
Biological macromolecule 2
272-2
Biological macromolecule 3
272-3
Biological macromolecule M
272-M

FIG. 2

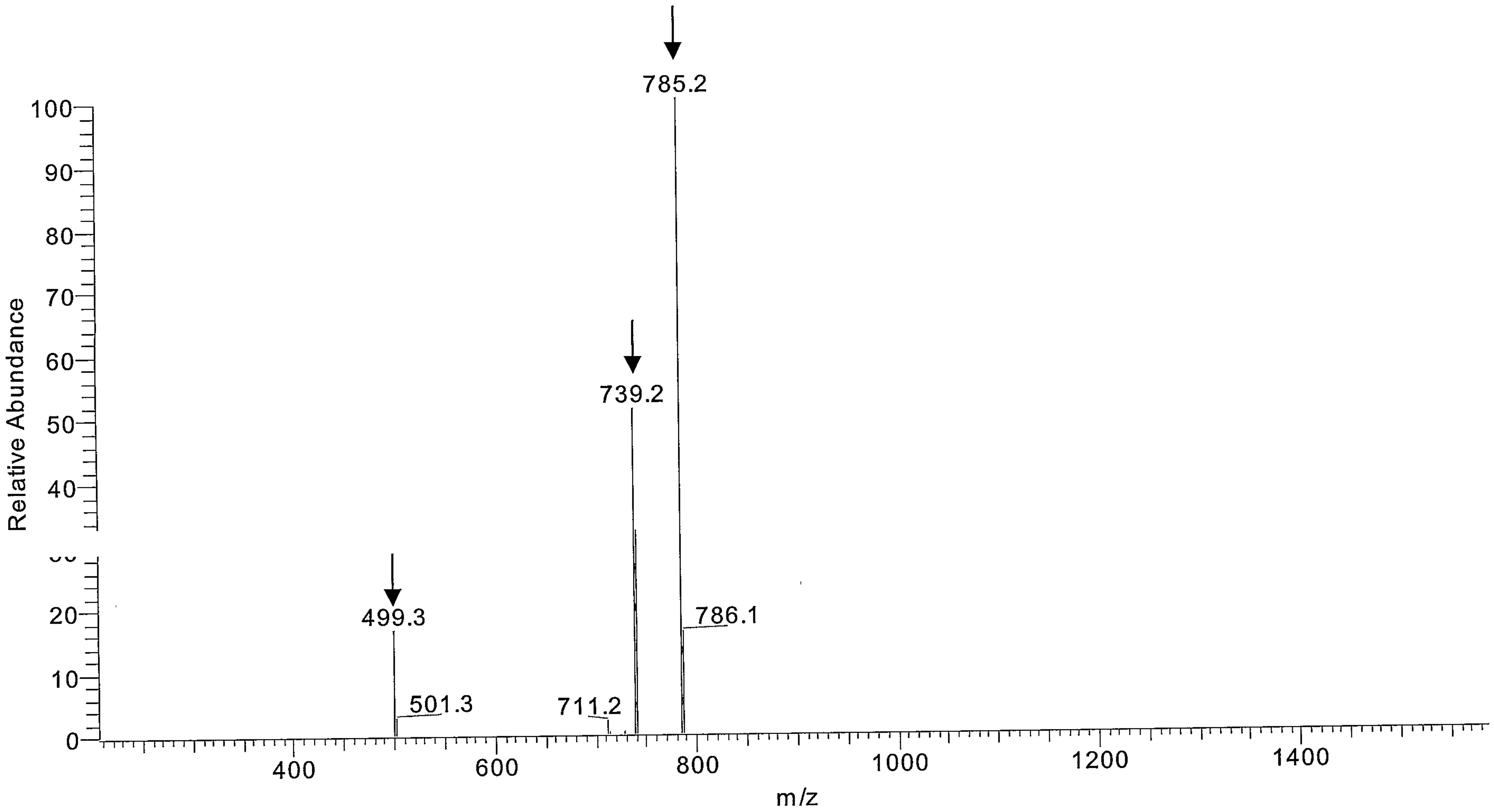

785.2
739.2
499.3
501.3
711.2
786.1
Relative Abundance
100
90
80
70
60
50
40
20
10
0
400
600
800
1000
1200
1400
m/z

Fig. 3

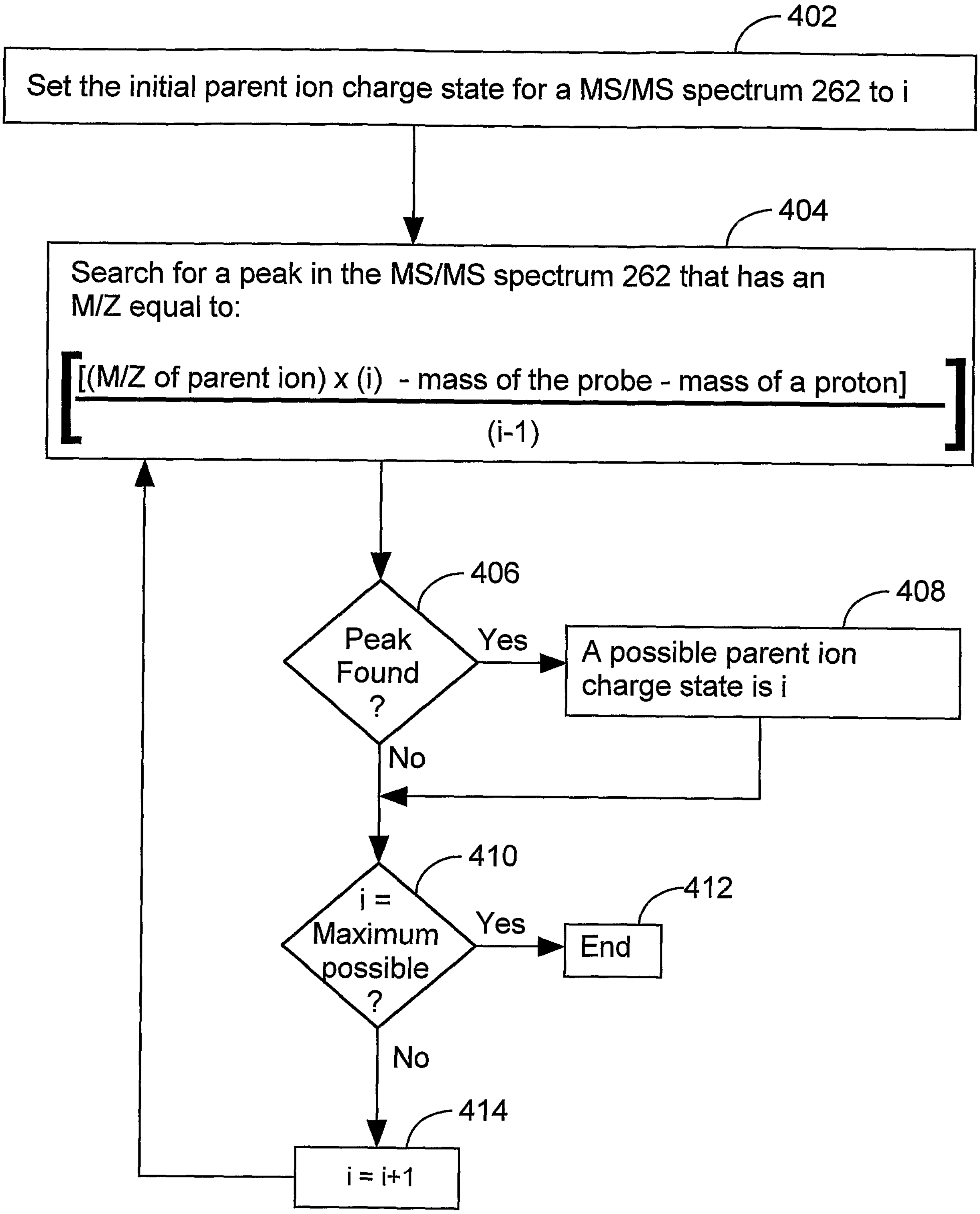

402
Set the initial parent ion charge state for a MS/MS spectrum 262 to i
404
Search for a peak in the MS/MS spectrum 262 that has an M/Z equal to:
[[(M/Z of parent ion) x (i) - mass of the probe - mass of a proton] / (i-1)]
406
Peak Found ?
Yes
408
A possible parent ion charge state is i
No
410
i = Maximum possible ?
Yes
412
End
No
414
i = i+1

FIG. 4

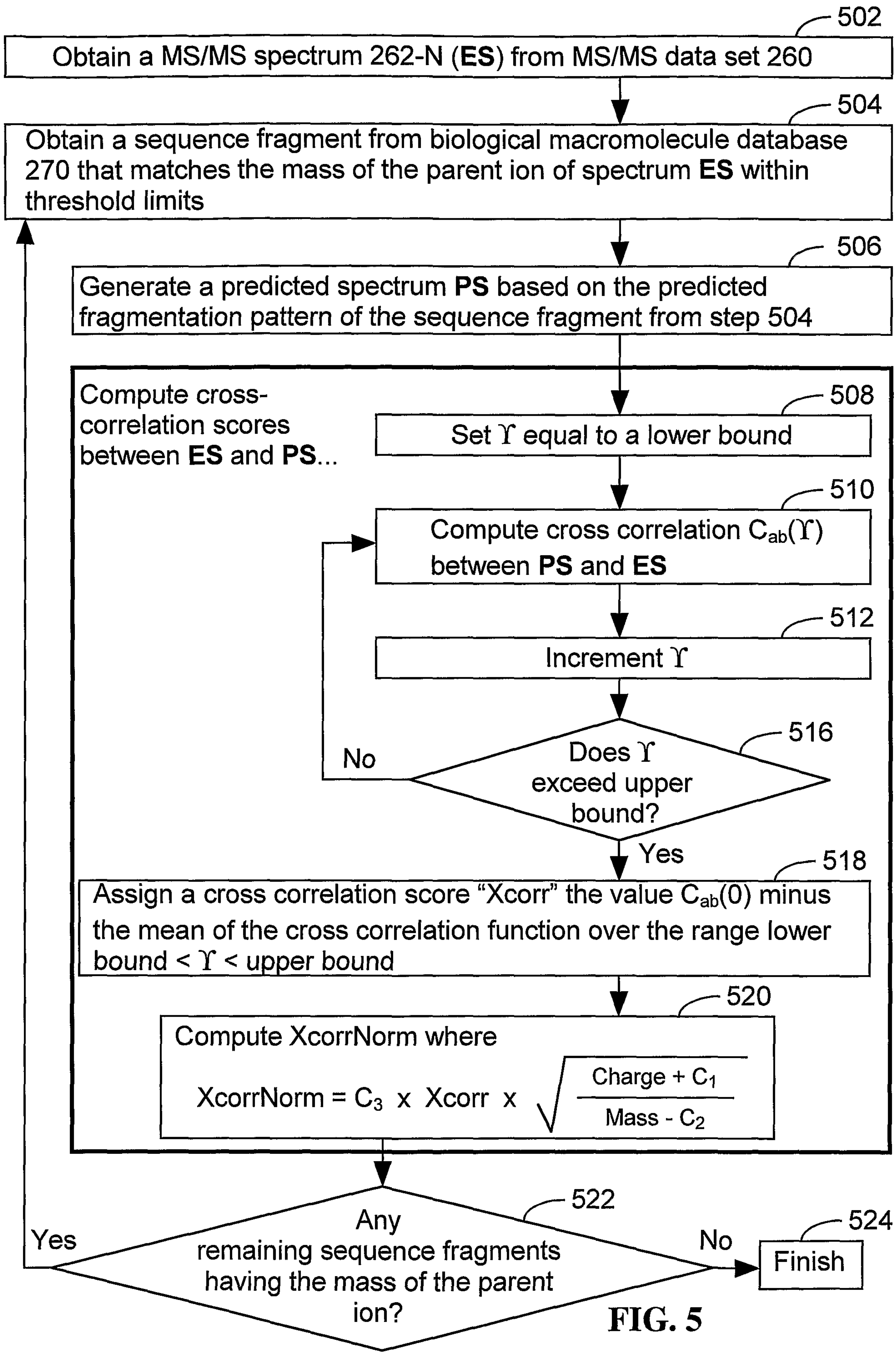

502
Obtain a MS/MS spectrum 262-N (ES) from MS/MS data set 260
504
Obtain a sequence fragment from biological macromolecule database 270 that matches the mass of the parent ion of spectrum ES within threshold limits
506
Generate a predicted spectrum PS based on the predicted fragmentation pattern of the sequence fragment from step 504
Compute cross-correlation scores between ES and PS...
508
Set ϒ equal to a lower bound
510
Compute cross correlation $C_{ab}(ϒ)$ between PS and ES
512
Increment ϒ
516
Does ϒ exceed upper bound?
No
Yes
518
Assign a cross correlation score "Xcorr" the value $C_{ab}(0)$ minus the mean of the cross correlation function over the range lower bound < ϒ < upper bound
520
Compute XcorrNorm where
$XcorrNorm = C_3 \times Xcorr \times \sqrt{\frac{Charge + C_1}{Mass - C_2}}$
522
Any remaining sequence fragments having the mass of the parent ion?
Yes
No
524
Finish

FIG. 5

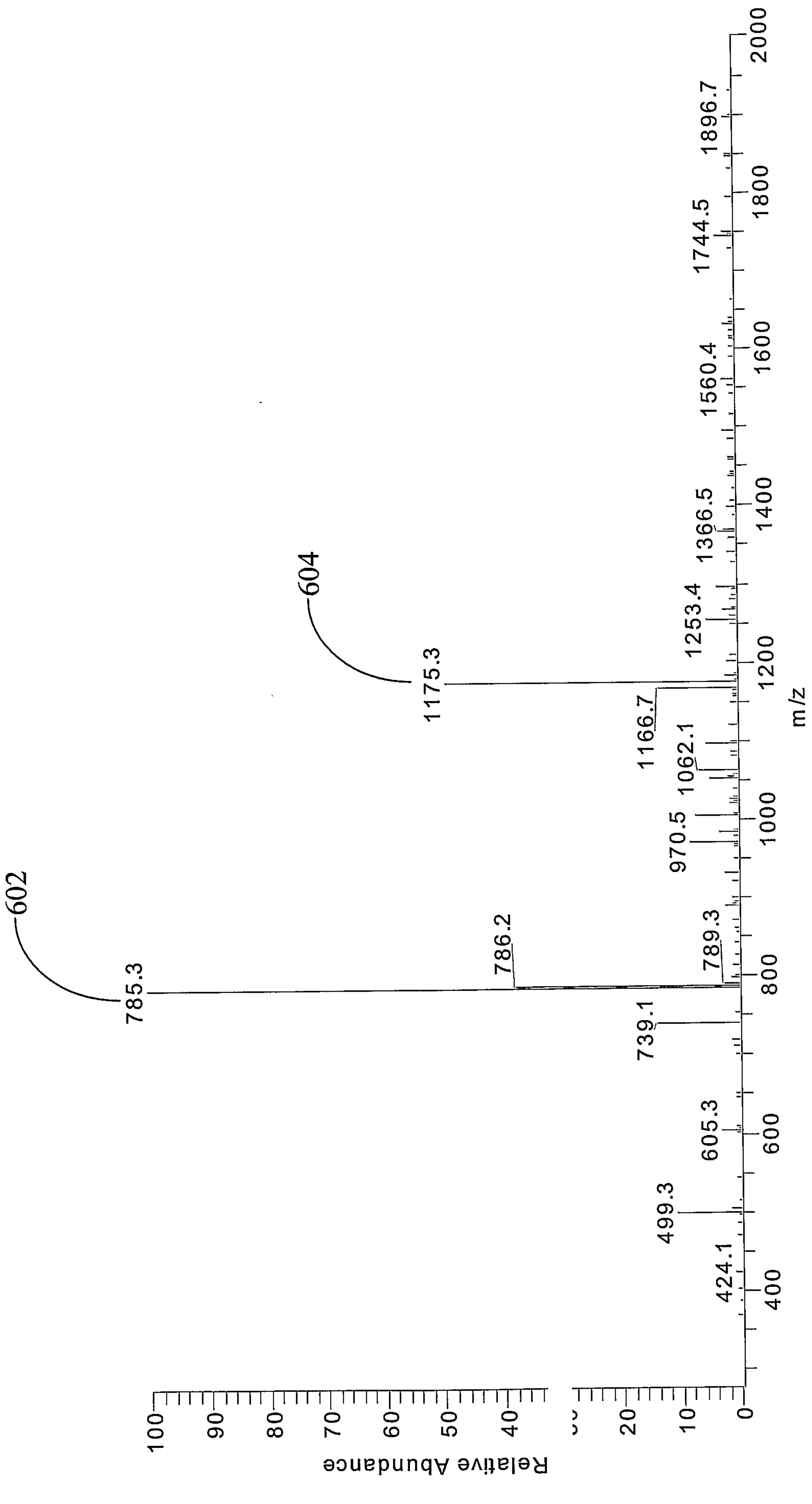

602
604
785.3
786.2
789.3
739.1
1175.3
1166.7
1062.1
970.5
1253.4
1366.5
1560.4
1744.5
1896.7
605.3
499.3
424.1
400
600
800
1000
1200
1400
1600
1800
2000
m/z
Relative Abundance
0
10
20
40
50
60
70
80
90
100

Fig. 6

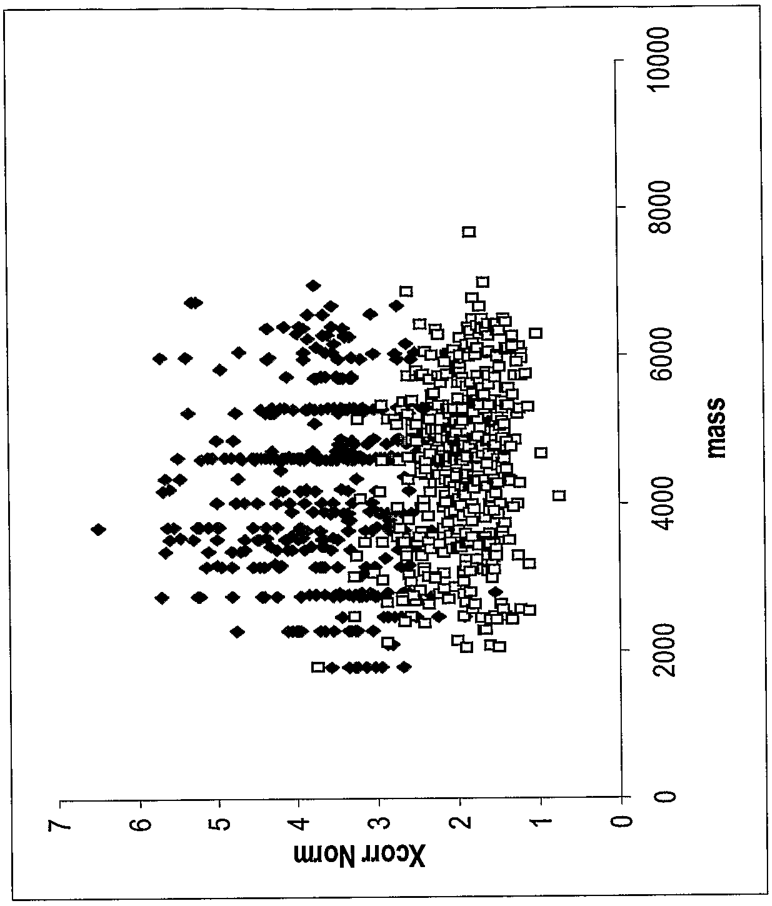


Fig. 7B

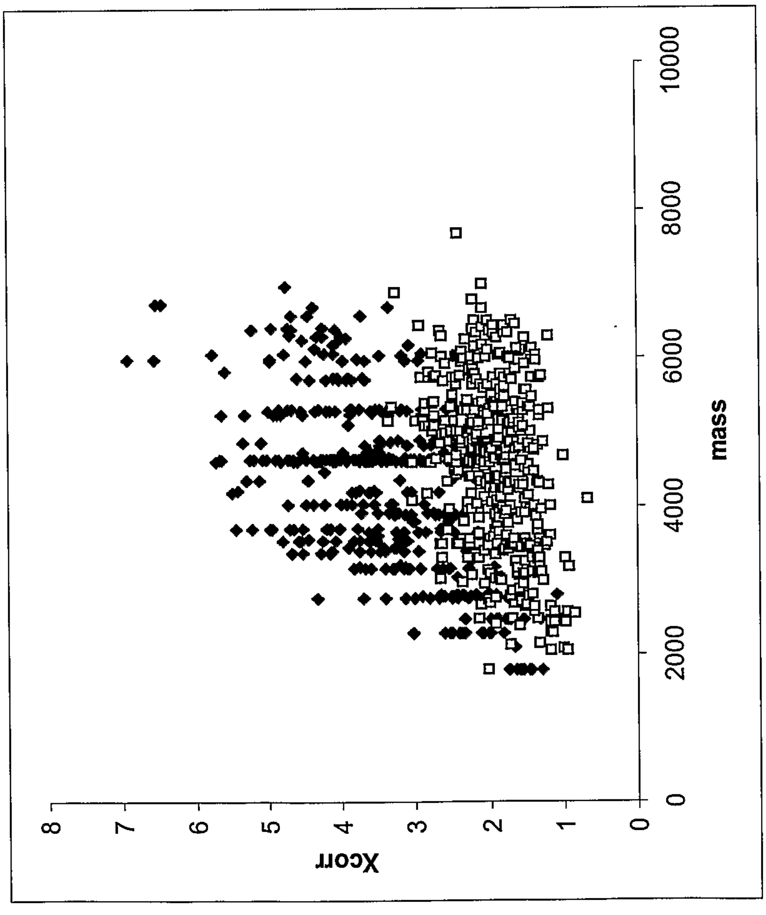
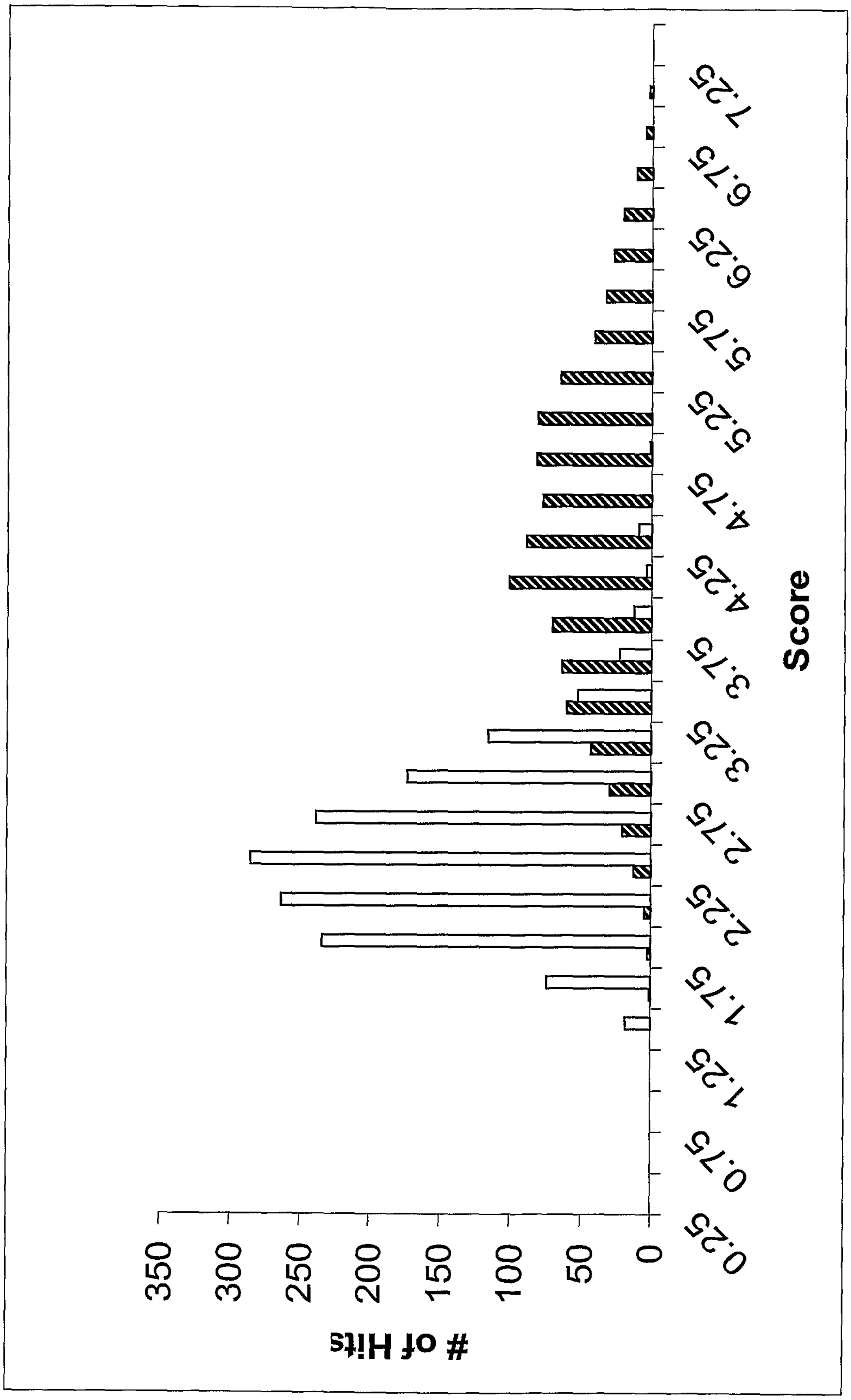


Fig. 7A
(Prior Art)



# of Hits
350
300
250
200
150
100
50
0
0.25
0.75
1.25
1.75
2.25
2.75
3.25
3.75
4.25
4.75
5.25
5.75
6.25
6.75
7.25
Score

Fig. 8

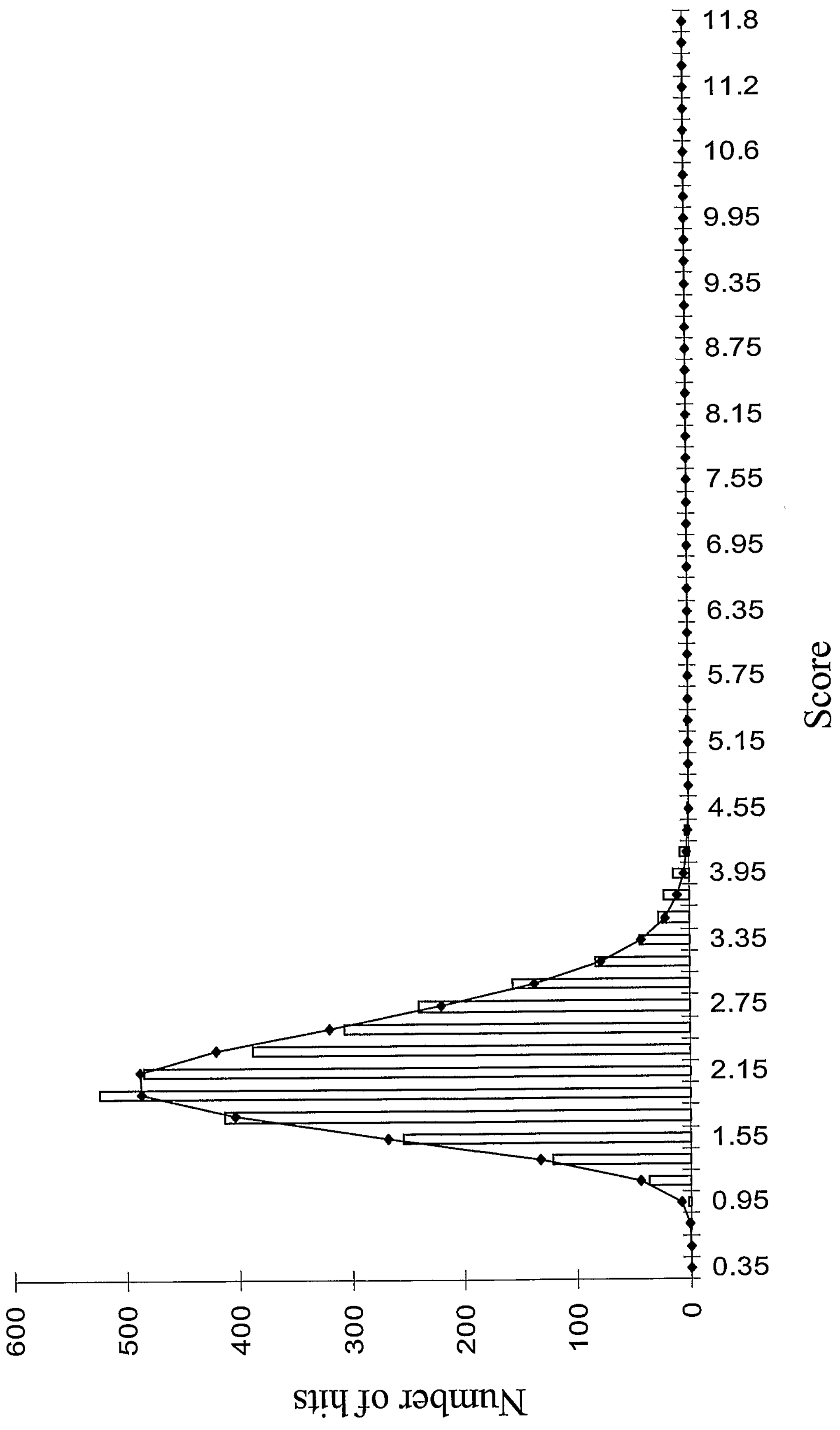

Number of hits
600
500
400
300
200
100
0
Score
0.35
0.95
1.55
2.15
2.75
3.35
3.95
4.55
5.15
5.75
6.35
6.95
7.55
8.15
8.75
9.35
9.95
10.6
11.2
11.8

Fig. 9

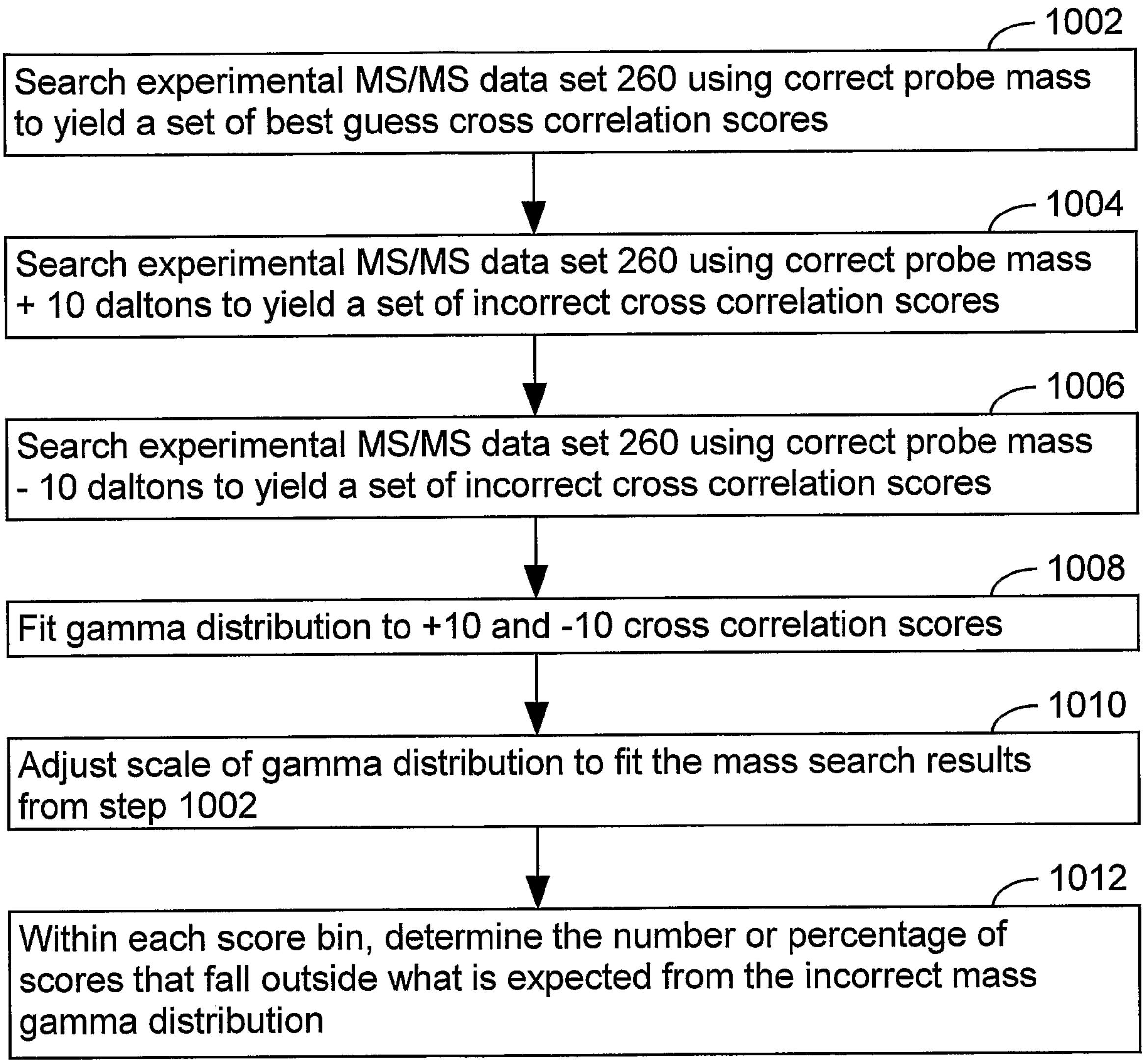

1002
Search experimental MS/MS data set 260 using correct probe mass to yield a set of best guess cross correlation scores
1004
Search experimental MS/MS data set 260 using correct probe mass + 10 daltons to yield a set of incorrect cross correlation scores
1006
Search experimental MS/MS data set 260 using correct probe mass - 10 daltons to yield a set of incorrect cross correlation scores
1008
Fit gamma distribution to +10 and -10 cross correlation scores
1010
Adjust scale of gamma distribution to fit the mass search results from step 1002
1012
Within each score bin, determine the number or percentage of scores that fall outside what is expected from the incorrect mass gamma distribution

FIG. 10

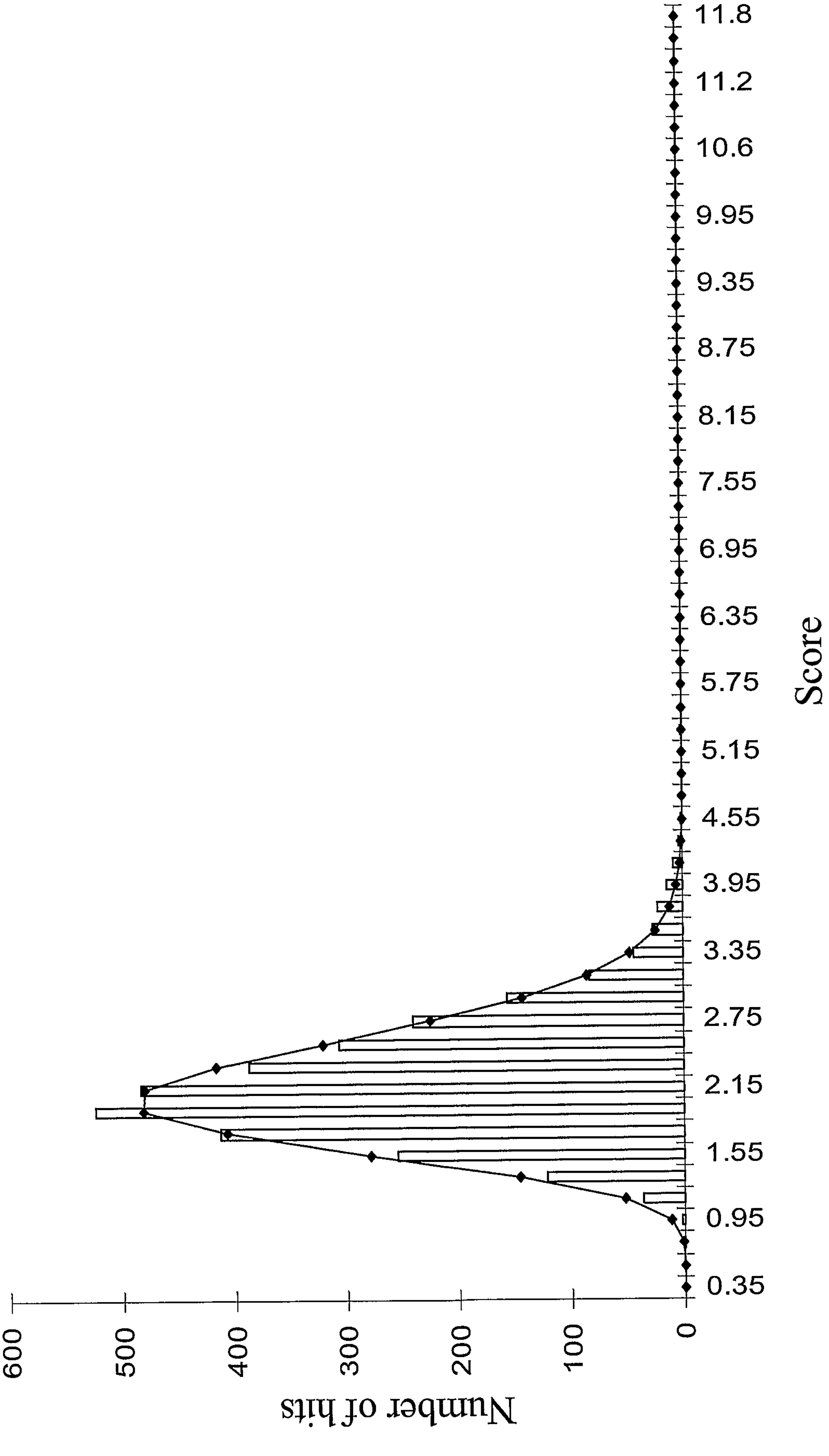

Number of hits
0
100
200
300
400
500
600
Score
0.35
0.95
1.55
2.15
2.75
3.35
3.95
4.55
5.15
5.75
6.35
6.95
7.55
8.15
8.75
9.35
9.95
10.6
11.2
11.8

Fig. 11

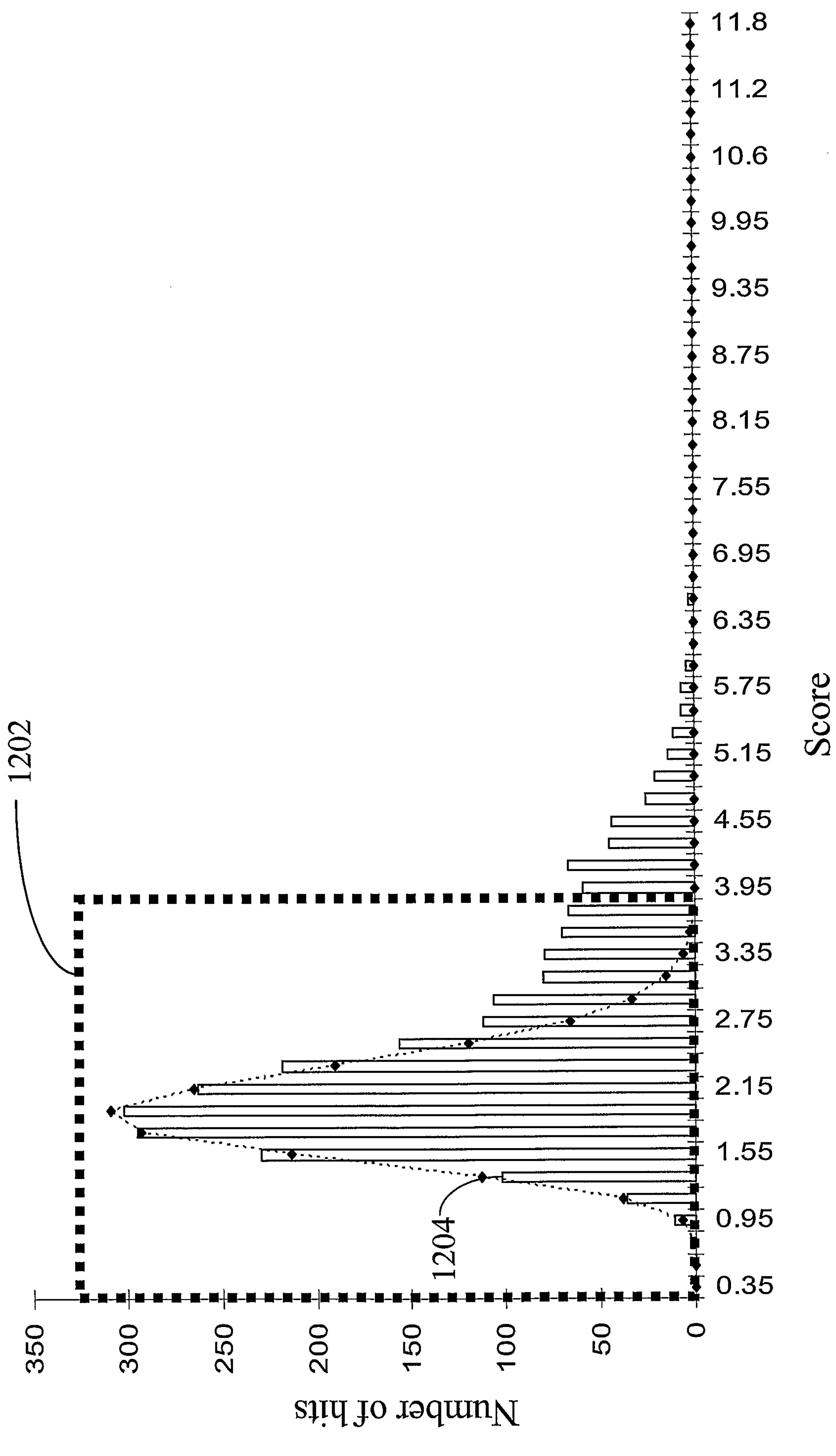

1202
1204
Number of hits
350
300
250
200
150
100
50
0
Score
0.35
0.95
1.55
2.15
2.75
3.35
3.95
4.55
5.15
5.75
6.35
6.95
7.55
8.15
8.75
9.35
9.95
10.6
11.2
11.8

Fig. 12

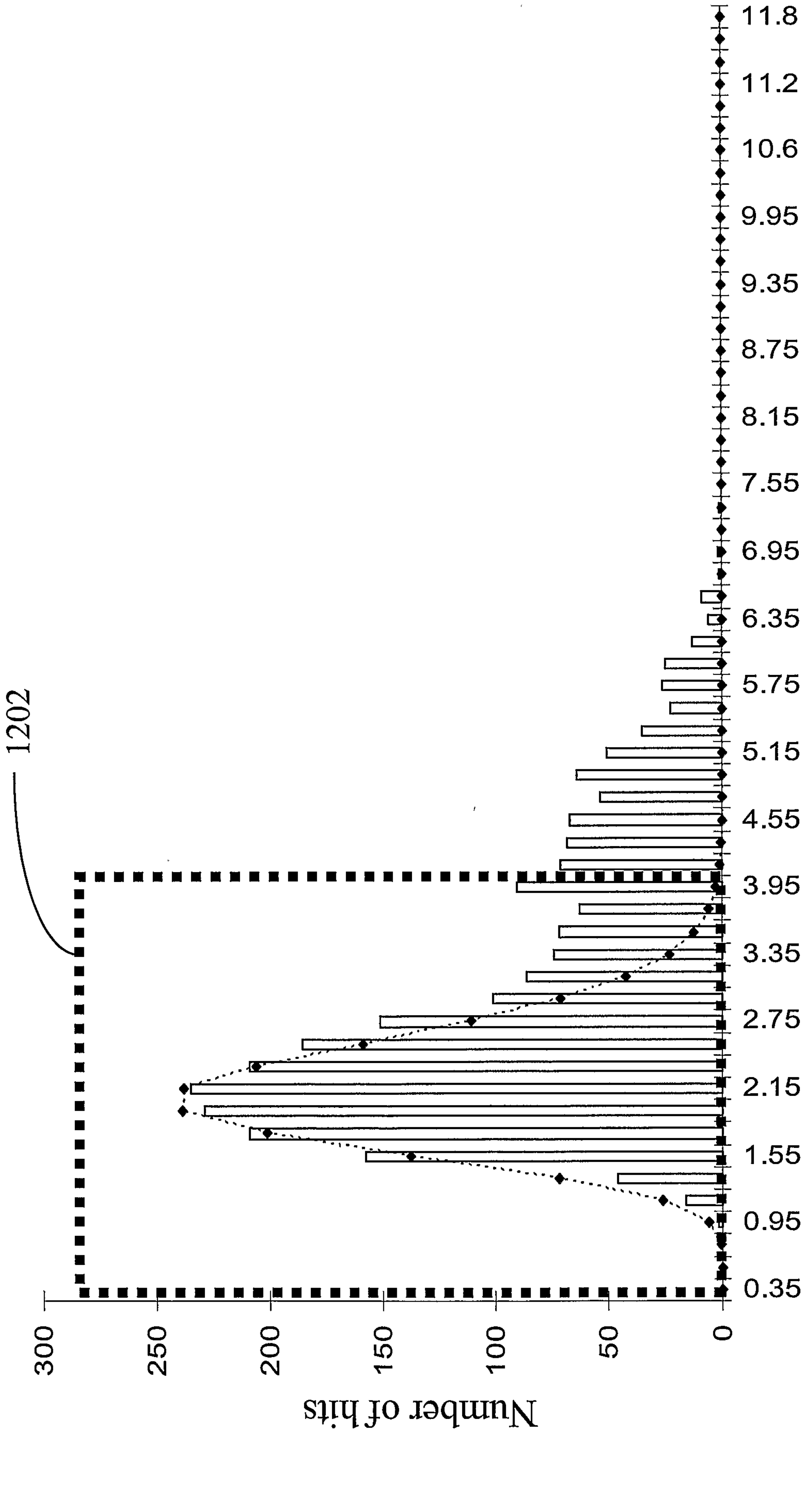

1202
Number of hits
300
250
200
150
100
50
0
0.35
0.95
1.55
2.15
2.75
3.35
3.95
4.55
5.15
5.75
6.35
6.95
7.55
8.15
8.75
9.35
9.95
10.6
11.2
11.8

Fig. 13

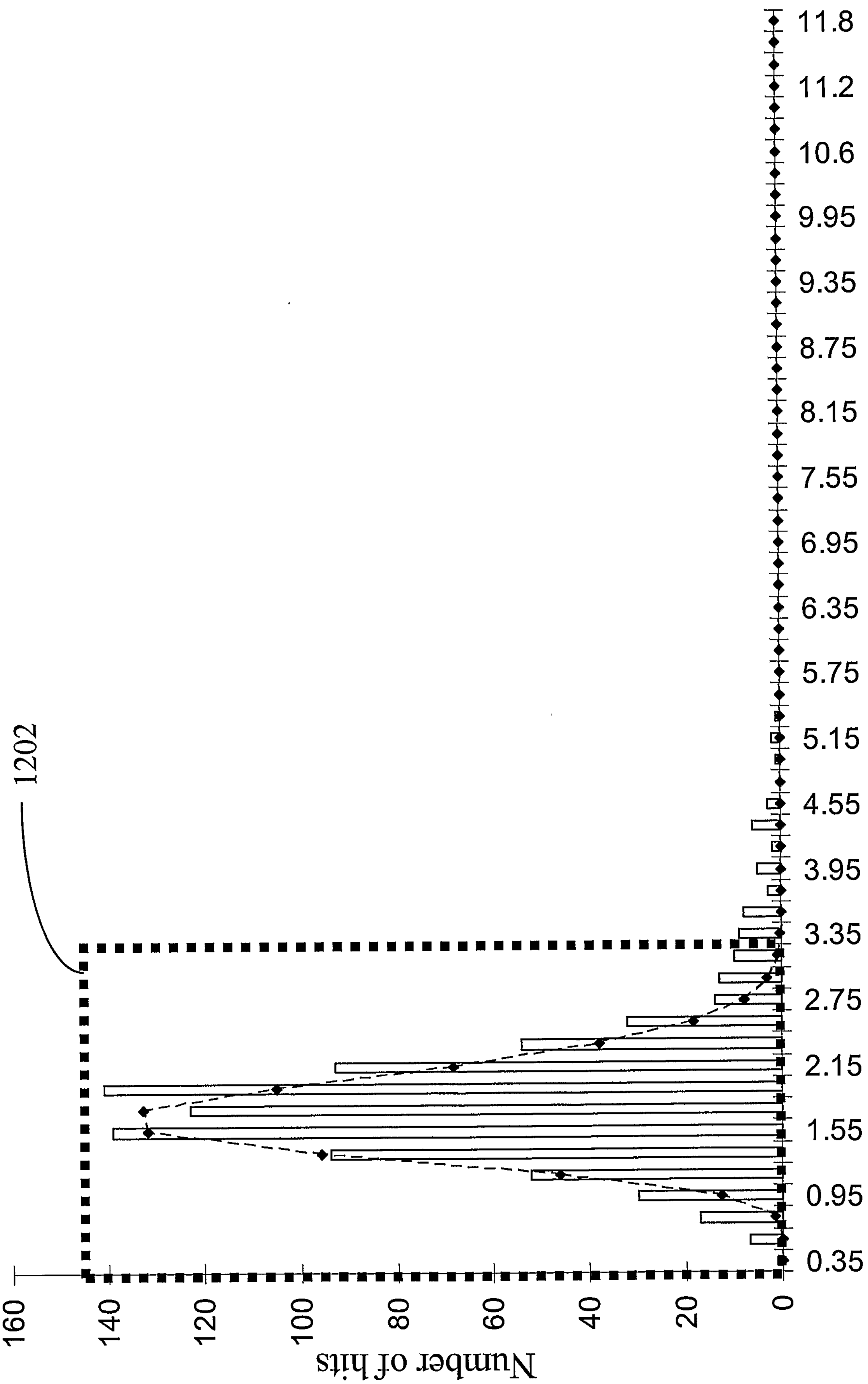

1202
Number of hits
160
140
120
100
80
60
40
20
0
0.35
0.95
1.55
2.15
2.75
3.35
3.95
4.55
5.15
5.75
6.35
6.95
7.55
8.15
8.75
9.35
9.95
10.6
11.2
11.8

Fig. 14

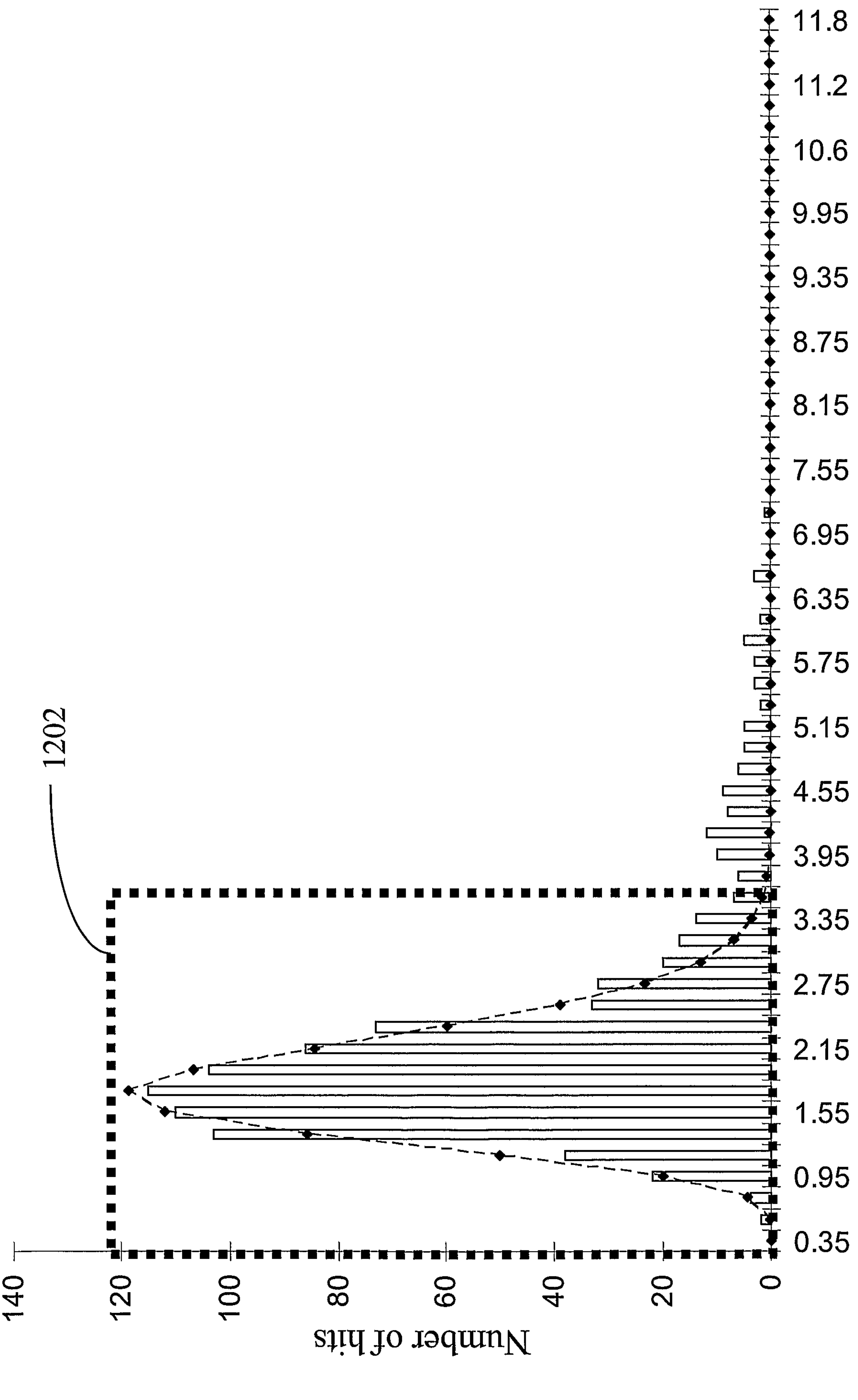

1202
140
120
100
80
60
40
20
0
Number of hits
0.35
0.95
1.55
2.15
2.75
3.35
3.95
4.55
5.15
5.75
6.35
6.95
7.55
8.15
8.75
9.35
9.95
10.6
11.2
11.8

Fig. 15

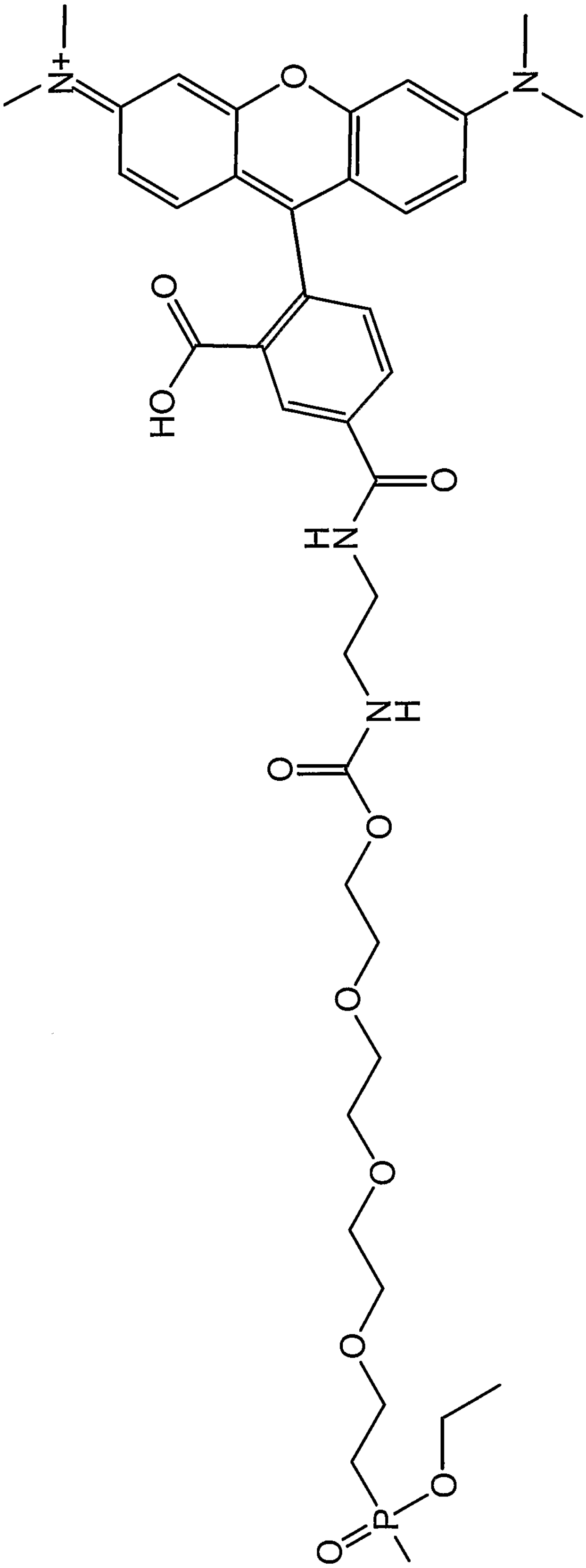


Fig. 16A

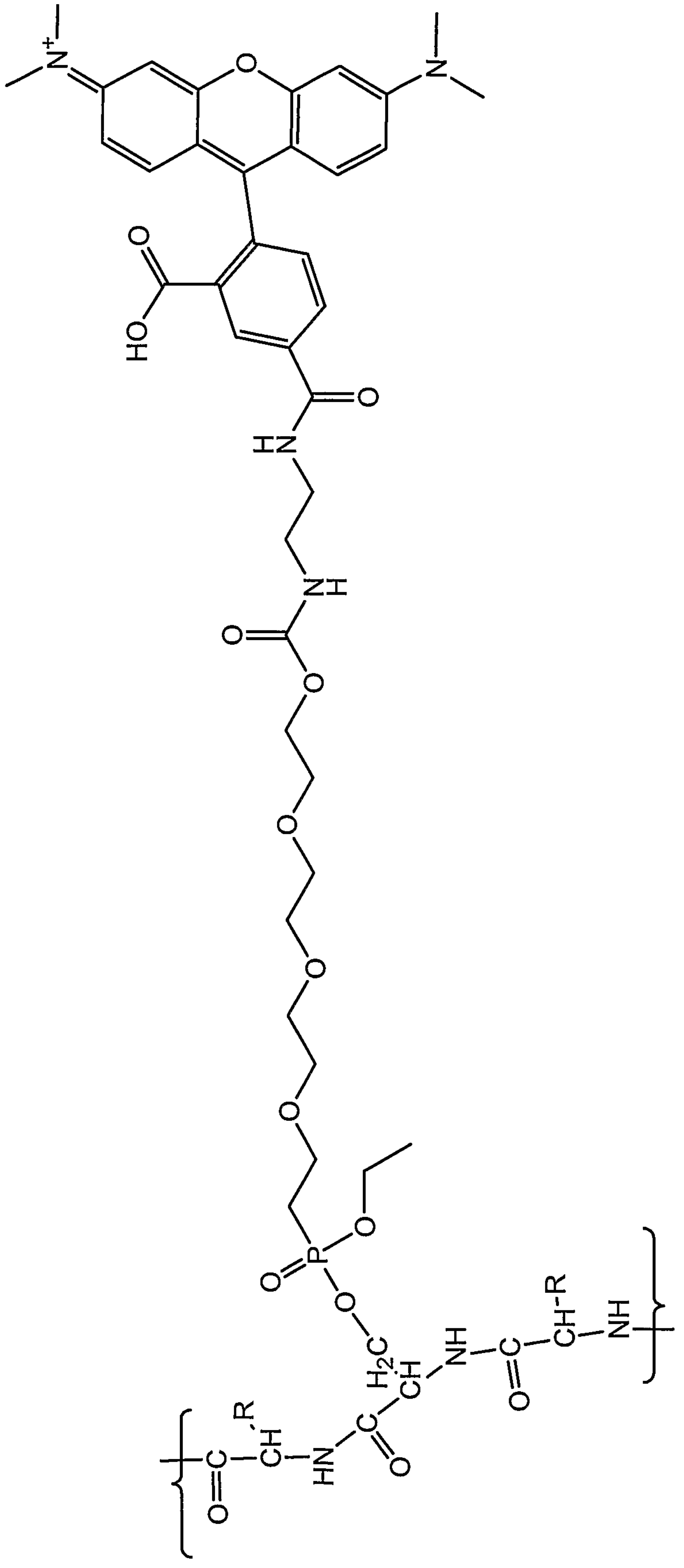


Fig. 16B

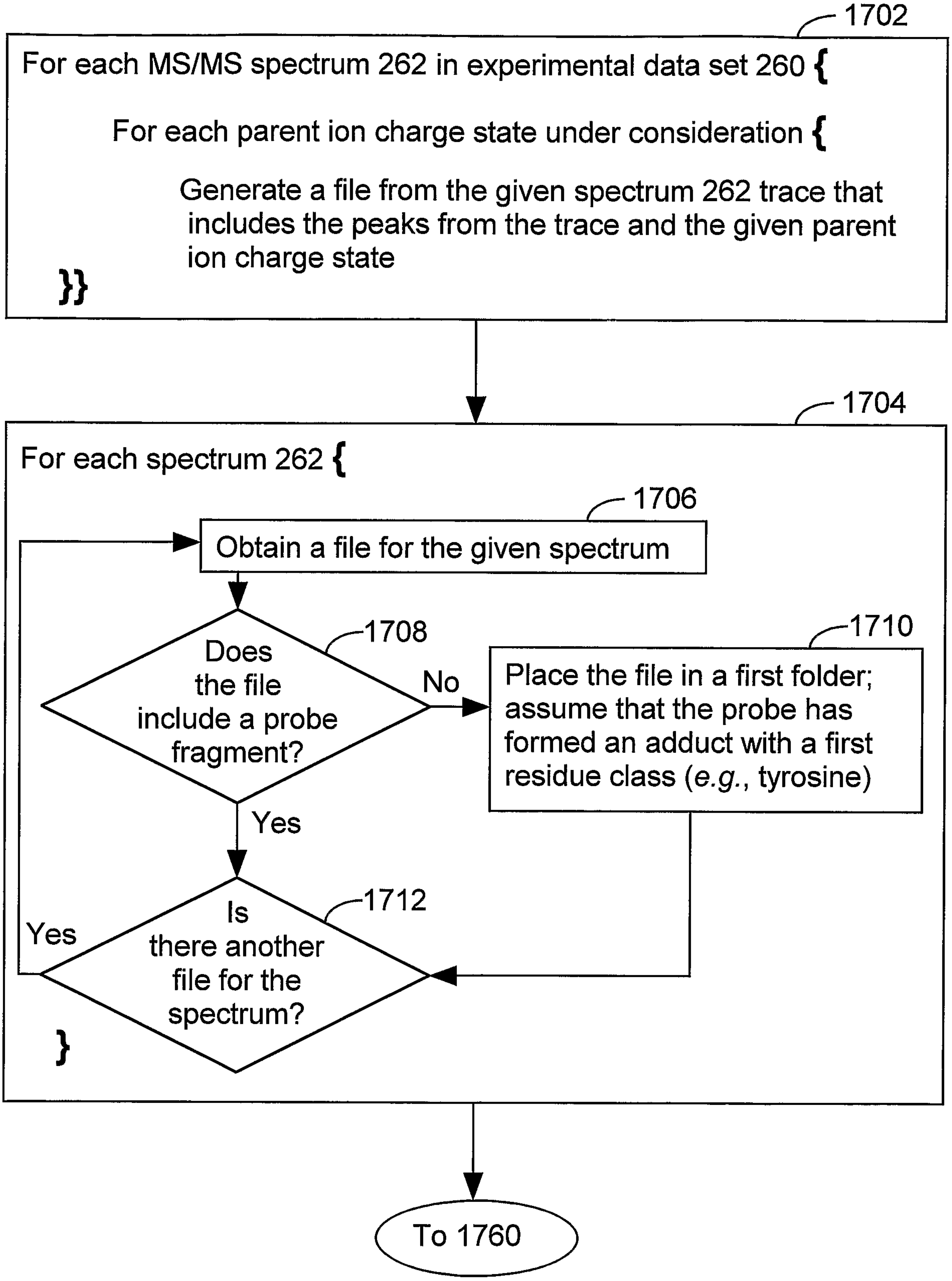

1702
For each MS/MS spectrum 262 in experimental data set 260 {
For each parent ion charge state under consideration {
Generate a file from the given spectrum 262 trace that includes the peaks from the trace and the given parent ion charge state
}}
1704
For each spectrum 262 {
1706
Obtain a file for the given spectrum
1708
Does the file include a probe fragment?
No
1710
Place the file in a first folder; assume that the probe has formed an adduct with a first residue class (e.g., tyrosine)
Yes
1712
Is there another file for the spectrum?
Yes
}
To 1760

FIG. 17A

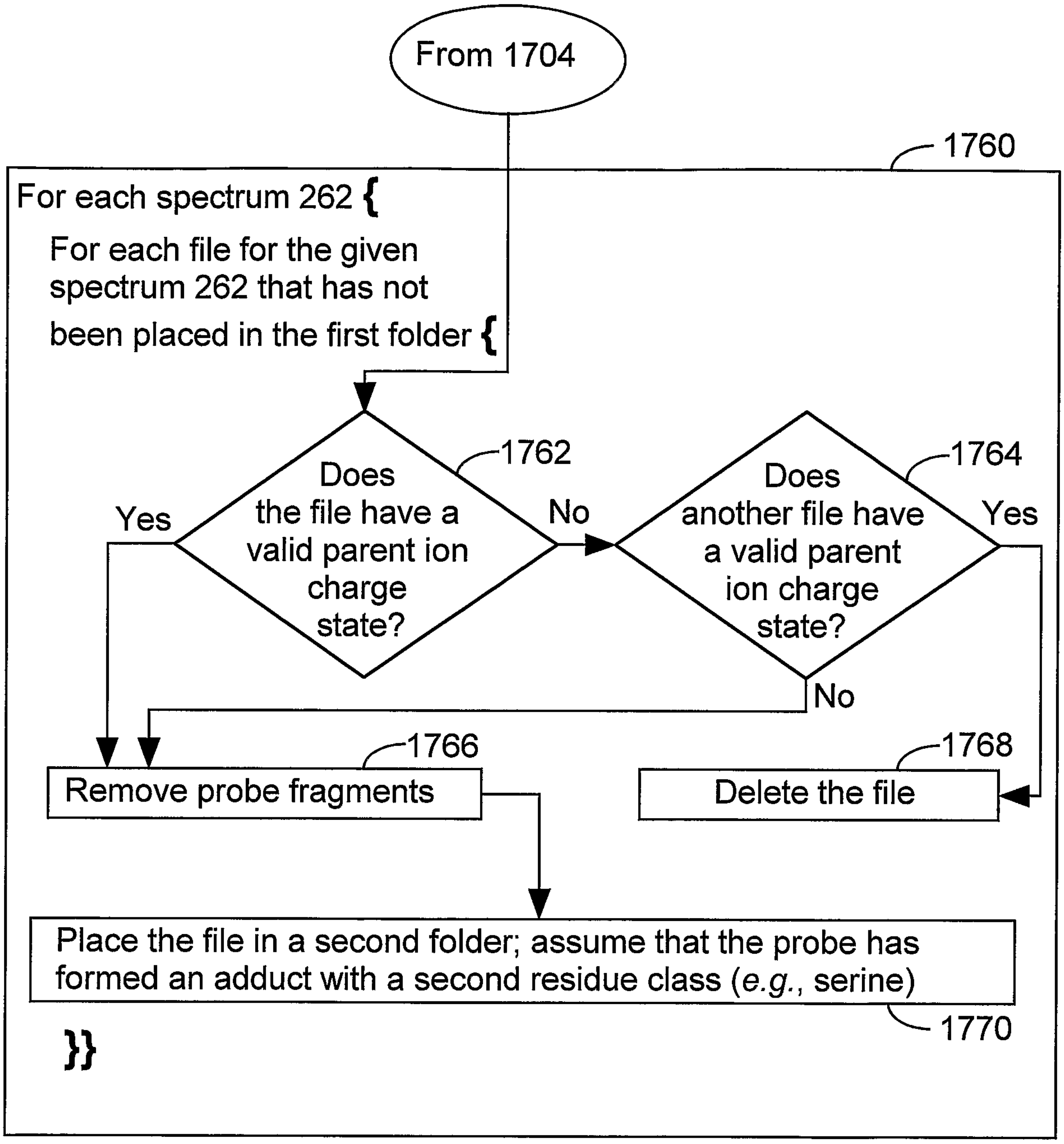

From 1704
1760
For each spectrum 262 {
For each file for the given spectrum 262 that has not been placed in the first folder {
Does the file have a valid parent ion charge state?
1762
Yes
No
Does another file have a valid parent ion charge state?
1764
Yes
No
Remove probe fragments
1766
Delete the file
1768
Place the file in a second folder; assume that the probe has formed an adduct with a second residue class (e.g., serine)
1770
}}

FIG. 17B

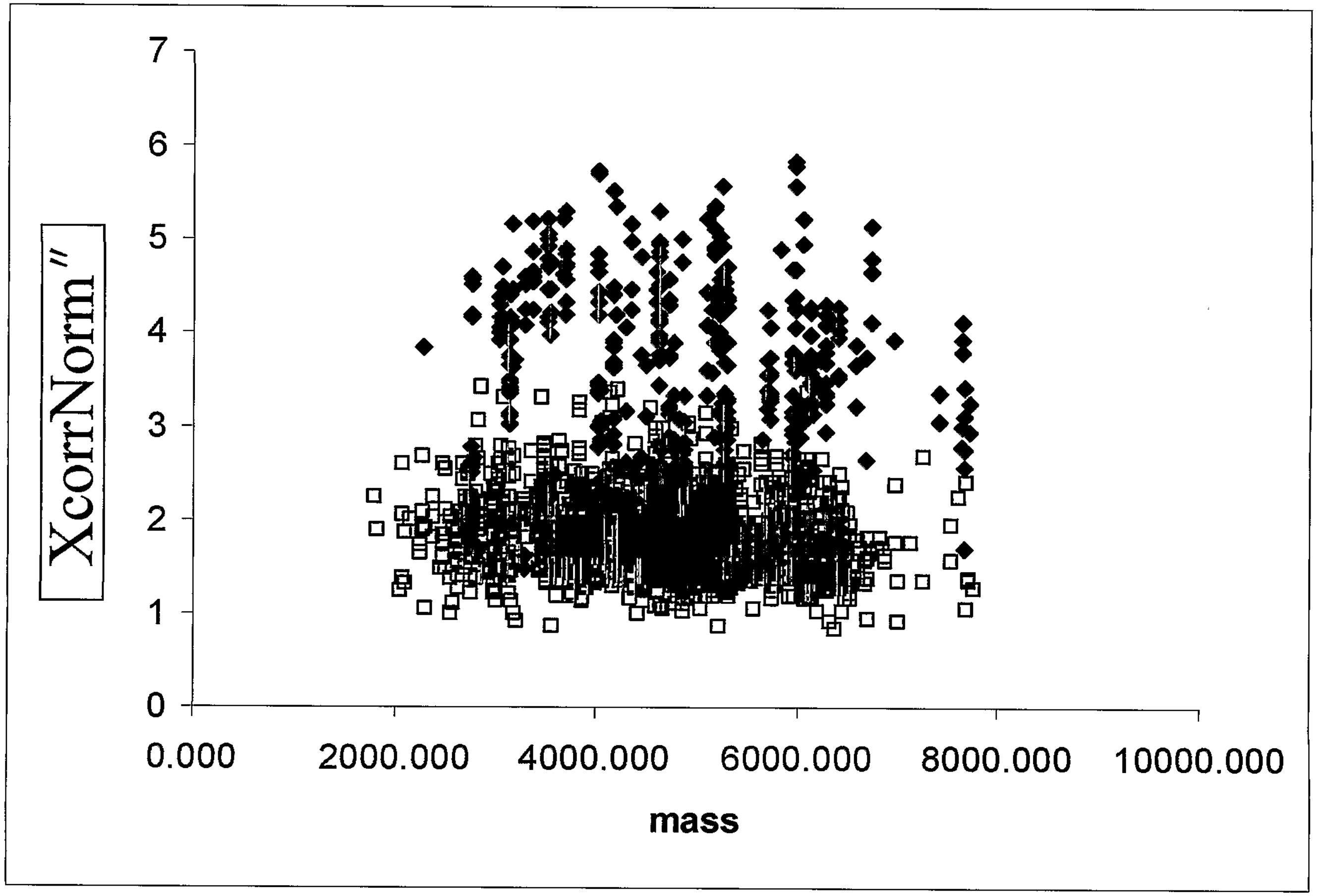

XcorrNorm''
7
6
5
4
3
2
1
0
0.000
2000.000
4000.000
6000.000
8000.000
10000.000
mass

Fig. 18