

SYSTEM AND PROCESS FOR PROCESSING OF TANDEM MASS SPECTRA USING DC-SHIFTED SELF-CONVOLUTION

Field of the Invention

[0001] The present invention generally relates to mass spectrometry technologies, and more particularly to a system and process for processing of tandem mass spectra using DC-shifted self-convolution.

Background of the Invention

[0002] Mass Spectrometry (MS) is a powerful tool in proteomics for protein identification and sequencing. A mass spectrometer is an instrument that measures the mass-to-charge ratio of individual molecules that have been converted into electrically charged gas-phase molecules, or ions. These ions are filtered in such a way as to produce an ordered separation of the ions as they pass through the instrument, ordered from lower to higher mass-to-charge ratios. The data is typically displayed as a plot of intensity vs. mass-to-charge ratio. The ionization techniques generally used to ionize sample peptides in most proteomic analyses are Matrix Assisted Laser Desorption Ionization (MALDI) and Electrospray Ionization (ESI).

[0003] With the advent of tandem mass spectrometry (MS/MS), the identification of proteins has become an easier and more accurate process. MS/MS involves two stages of mass spectrometry analysis of a sample in a single experiment. The first stage filters out the ions of interest from the sample. These selected ions are then passed into a collision cell that breaks the ions into smaller ion fragments. The second stage then separates and detects these smaller ion fragments.

[0004] The seemingly effortless MS/MS experiments have yielded voluminous data of raw mass spectra, making it impractical to manually analyze and validate the quality of the raw mass spectra. However, the quality of each mass spectrum from which a protein is identified is crucial in validating the protein identified by protein identification tools such as MASCOT and SEQUEST.

[0005] Therefore, there is an imperative need to have a process by which the quality of mass spectra can be evaluated before subjecting them to peptide/protein identification tools, increasing the validity of the peptides/proteins identified.

Summary of the Invention

[0006] In one embodiment of the present invention, there is provided a process for evaluating the quality of a MS/MS spectrum of a polypeptide. The process comprises the steps of obtaining the MS/MS spectrum of a polypeptide from a source; and performing self-convolution upon the MS/MS spectrum, and then outputting a self-convolution plot; thereby the quality of the MS/MS spectrum is evaluated on parameters of the self-convolution plot.

[0007] In another embodiment of the process, the step of performing self-convolution comprises performing the self-convolution according to the following equation: $h(x) = f(x) * g(x)$, where $g(x)$ is the 'horizontal flip' of function $f(x)$.

[0008] In another embodiment of the process, the parameters of the self-convolution plot include the position of the highest peak, the amplitudes of the peaks, and the ratio between the highest peak and other peaks.

[0009] In another embodiment of the process, the step of performing self-convolution comprises performing a DC shifted self-convolution according to the followings: performing the Discrete Fourier Transform (DFT) onto data(n) according to $D(f) = \text{DFT}(\text{data}(n))$, wherein the data(n) is a mass spectrum; performing the self convolution in frequency domain, wherein the product of the DFTs is computed and followed by the inverse DFT of the product according to $DD(f) = D(f) \times D(f)$; removing the near DC components by setting the first 10 points of the DFT to 0 according to $DD(1:10) = 0$, and $IDD(n) = \text{iDFT}(DD(f))$; and considering only the real part of the inverse DFT, and normalizing the intensity against the maximum value according to $RIDD(n) = \text{real}(IDD(n))$, and $NRIDD(n) = RIDD(n) / \max(RIDD(n))$.

[0010] In another embodiment of the process, before the step of performing self-convolution, further comprising a step of selecting n highest peaks from the MS/MS spectrum and normalizing the n highest peaks into the same amplitude so that a new mass

spectrum is generated, wherein the n is an integral; thereby the self-convolution is performed upon the newly generated mass spectrum.

[0011] In another embodiment of the present invention, there is provided a system for evaluating the quality of a MS/MS spectrum of a polypeptide. The system comprises a means for obtaining the MS/MS spectrum from a source; a computer executable storage medium in which computer programs are embedded, wherein the computer programs enable the system to perform self-convolution upon the MS/MS spectrum; thereby the quality of the MS/MS spectrum of a polypeptide can be evaluated.

[0012] In another embodiment of the system, the computer programs perform the self-convolution according to the following equation: $h(x) = f(x) * g(x)$; where $g(x)$ is the 'horizontal flip' of function $f(x)$.

[0013] In another embodiment of the system, it further comprises a means for outputting a self-convolution plot after the computer programs perform the self-convolution. In another embodiment of the system, the parameters of the self-convolution plot include the position of the highest peak, the amplitudes of the peaks, and the ratio between the highest peak and other peaks.

[0014] In another embodiment of the system, the computer programs perform a DC shifted self-convolution according to the followings: performing the Discrete Fourier Transform (DFT) onto $data(n)$ according to $D(f) = DFT(data(n))$, wherein the $data(n)$ is a mass spectrum; performing the self convolution in frequency domain, wherein the product of the DFTs is computed and followed by the inverse DFT of the product according to $DD(f) = D(f) \times D(f)$; removing the near DC components by setting the first 10 points of the DFT to 0 according to $DD(1:10) = 0$, and $IDD(n) = iDFT(DD(f))$; and considering only the real part of the inverse DFT, and normalizing the intensity against the maximum value according to $RIDD(n) = real(IDD(n))$, and $NRIDD(n) = RIDD(n)/max(RIDD(n))$.

[0015] In another embodiment of the system, before performing self-convolution, the computer programs further perform a step of selecting n highest peaks from the MS/MS spectrum and normalizing the n highest peaks into the same amplitude so that a new mass spectrum is generated, wherein the n is an integral; thereby the self-convolution is performed upon the newly generated mass spectrum.

[0016] The objectives and advantages of the invention will become apparent from the following detailed description of preferred embodiments thereof in connection with the accompanying drawings.

Brief Description of the Drawings

[0017] Preferred embodiments according to the present invention will now be described with reference to the Figures, in which like reference numerals denote like elements.

[0018] FIG 1 illustrates various breakage points within a peptide during fragmentation.

[0019] FIG 2 shows an exemplary MS/MS spectrum of a peptide AVAGCAGAR.

[0020] FIG 3 shows a plot of the b-ions and y-ions from peptide sequence **MTDQEAIQDLWQWR** processed by equation (8).

[0021] FIG 4 is the resultant plot from the application of self-convolution to the plot shown in FIG 3.

[0022] FIG 5 shows a plot of the b-ions and y-ions from peptide sequence **MTDQEAIQDLWQWR** processed by equation (9).

[0023] FIG 6 is the resultant plot from the application of self-convolution to the plot shown in FIG 5.

[0024] FIG 7 is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3

[0025] FIG 8 is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5

[0026] FIG 9 is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for the noise level being set at 25

[0027] FIG 10(a) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 2 b-ion peaks

[0028] FIG 10(b) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 2 b-ion peaks

[0029] FIG 10(c) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 4 b-ion peaks

- [0030] FIG 10(d) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 4 b-ion peaks
- [0031] FIG 10(e) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 6 b-ion peaks
- [0032] FIG 10(f) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 6 b-ion peaks
- [0033] FIG 10(g) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 8 b-ion peaks
- [0034] FIG 10(h) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 8 b-ion peaks
- [0035] FIG 10(i) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 10 b-ion peaks
- [0036] FIG 10(j) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 10 b-ion peaks
- [0037] FIG 10(k) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for no b-ion peaks
- [0038] FIG 10(l) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for no b-ion peaks
- [0039] FIG 11(a) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 2 y-ion peaks
- [0040] FIG 11(b) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 2 y-ion peaks
- [0041] FIG 11(c) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 4 y-ion peaks
- [0042] FIG 11(d) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 4 y-ion peaks
- [0043] FIG 11(e) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 6 y-ion peaks
- [0044] FIG 11(f) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 6 y-ion peaks
- [0045] FIG 11(g) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 8 y-ion peaks

- [0046] FIG 11(h) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 8 y-ion peaks
- [0047] FIG 11(i) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 10 y-ion peaks
- [0048] FIG 11(j) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 10 y-ion peaks
- [0049] FIG 11(k) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 12 y-ion peaks
- [0050] FIG 11(l) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for minus 12 y-ion peaks
- [0051] FIG 11(m) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for no y-ion peaks
- [0052] FIG 11(n) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for no y-ion peaks
- [0053] FIG 12(a) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 3 b/y-ion peaks
- [0054] FIG 12(b) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 5 b/y-ion peaks
- [0055] FIG 12(c) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 8 b/y-ion peaks
- [0056] FIG 12(d) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 11 b/y-ion peaks
- [0057] FIG 12(e) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for minus 14 b/y-ion peaks
- [0058] FIG 13(a) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for b-ions peaks reduced by 20%
- [0059] FIG 13(b) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for b-ions peaks reduced by 20%
- [0060] FIG 13(c) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for b-ions peaks reduced by 50%
- [0061] FIG 13(d) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for b-ions peaks reduced by 50%

- [0062] FIG 13(e) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for b-ions peaks reduced by 75%
- [0063] FIG 13(f) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for b-ions peaks reduced by 75%
- [0064] FIG 14(a) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for y-ions peaks reduced by 20%
- [0065] FIG 14(b) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for y-ions peaks reduced by 20%
- [0066] FIG 14(c) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for y-ions peaks reduced by 50%
- [0067] FIG 14(d) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for y-ions peaks reduced by 50%
- [0068] FIG 14(e) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3 except for y-ions peaks reduced by 75%
- [0069] FIG 14(f) is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for y-ions peaks reduced by 75%
- [0070] FIG 15 illustrates the quantitative measurement of determining the quality of any given MS/MS spectrum on the DC shifted self-convolution plots in accordance with embodiment of the present invention.
- [0071] FIG 16(a) shows a real experimental MS/MS data.
- [0072] FIG 16(b) shows the DC shifted self-convolution plot resulting from the application to the spectrum of FIG 16(a) that reveals a high mid point peak.
- [0073] FIG 17 shows the MASCOT prediction result of the real experiment data shown in FIG 16(a), confirming the presents of nearly all b and y ions.
- [0074] FIG 18(a) shows another real experimental MS/MS data that is of good quality.
- [0075] FIG 18(b) shows the DC shifted self-convolution plot resulting from the application to the spectrum of FIG 18(a) that reveals a high mid point peak.
- [0076] FIG 19 shows the MASCOT prediction results of the real experiment data of FIG 18(a), confirming the present of nearly all b and y ions.
- [0077] FIG 20(a) shows a poor quality MS/MS data.

[0078] FIG 20(b) shows the DC shifted self-convolution plot resulting from the application to the spectrum of FIG 20(a) that reveals a low or missing mid point peak.

[0079] FIG 21 shows the MASCOT prediction results of the real experiment data of FIG 20(a), confirming the poor quality of the real experiment data.

[0080] FIG 22 (a) shows a poor quality MS/MS data.

[0081] FIG 22(b) shows the DC shifted self-convolution plot resulting from the application to the spectrum of FIG 22(a) that reveals a false peak.

[0082] FIG 23 shows the MASCOT prediction results of poor quality MS/MS spectrum of FIG 22(a), showing many 'predicted' peaks that are too low to be significant.

[0083] FIG 24(a) shows a poor quality MS/MS data as shown in FIG 22(a).

[0084] FIG 24(b) shows a plot of b-ions and y-ions of 30 highest peaks from FIG 24(a).

[0085] FIG 24(c) shows the DC shifted self-convolution plot of FIG 24(a) that reveals a false peak, resulting in wrong prediction of poor quality as good quality as shown in FIG 22(b).

[0086] FIG 24(d) shows the DC shifted self-convolution plots of the newly generated spectrum of FIG 24(b).

[0087] FIG 25(a) shows the same MS/MS plot of data as shown in FIG 16(a) of Example 1.

[0088] FIG 25(b) shows the plots of b-ions and y-ions of 30 highest peaks from FIG 25(a) according to the methods as described in Example 5.

[0089] FIG 25(c) shows the same DC shifted self-convolution plot of FIG 25(a).

[0090] FIG 25(d) shows the DC shifted self-convolution plot of the newly generated spectrum of FIG 25(b).

[0091] FIG 26(a) shows the same MS/MS plot of data as shown in FIG 18(a) of Example 2.

[0092] FIG 26(b) shows the plots of b-ions and y-ions of 30 highest peaks from FIG 26(a) according to the methods as described in Example 5.

[0093] FIG 26(c) shows the same DC shifted self-convolution plot of FIG 26(a).

[0094] FIG 26(d) shows the DC shifted self-convolution plot of the newly generated spectrum of FIG 26(b).

[0095] FIG 27(a) shows the same MS/MS plot of data as shown in FIG 20(a) of Example 3.

[0096] FIG 27(b) shows the plots of b-ions and y-ions of 30 highest peaks from FIG 27(a) according to the methods as described in Example 5.

[0097] FIG 27(c) shows the same DC shifted self-convolution plot of FIG 27(a).

[0098] FIG 27(d) shows the DC shifted self-convolution plot of the newly generated spectrum of FIG 27(b).

Detailed Description of the Invention

[0099] The present invention may be understood more readily with reference to the following detailed description of certain embodiments of the invention.

[00100] Throughout this application, where publications are referenced, the disclosures of these publications are hereby incorporated by reference, in their entireties, into this application in order to more fully describe the state of art to which this invention pertains.

[00101] In the following detailed description, specific details are set forth in order to provide a thorough understanding of the invention. However, it will be understood by those skilled in the relevant art that the present invention may be practiced without these specific details. In other instances, well-known methods, procedures, components, and materials have not been described in detail so as not to obscure the present invention.

[00102] The present invention teaches various embodiments of the system and process that processes MS/MS spectrum data by employing the principles of convolution. Convolution and fragmentation in principle are well known in their relevant arts; thus for better understanding of the present invention brief descriptions of convolution and fragmentation are provided herein.

[00103] Convolution is a very basic but important mathematical operation in the field of signal processing. Briefly, the convolution of two functions $f(t)$ and $g(t)$ can be expressed in equation (1) below:

[00104]

$$h(t) \equiv \int_{-\infty}^{\infty} f(u) g(t-u) du \quad (1)$$

[00105] where $h(t)$ denotes values of quantity h as a function of time (t); $f(u)$ denotes values of quantity f as a function of time (u); and $g(t-u)$ denotes values of quantity g as a function of time ($t-u$).

[00106] Equation (1) is often written as equation (2) below:

[00107]

$$h(t) = f(t) * g(t) \quad (2)$$

[00108] where $h(t)$, $f(t)$ and $g(t)$ are functions of time (t).

[00109] For discrete time series, the convolution is a sum instead of an integral as expressed in equation (3) below:

[00110]

$$h_i = \sum_{j=0}^m f_j g_{i-j} \quad (3)$$

[00111] where i , j and m are integers.

[00112] Self-convolution refers to the convolution applied onto the same series of data. It follows the same principles of the convolution as discussed above. In the present application, self-convolution is specifically referred to as convolution of a function with the 'horizontal flip' of the same function. Thus self-convolution is given by equation $h(x) = f(x) * g(x)$ where $g(x)$ is the 'horizontal flip' of function $f(x)$.

[00113] Now there is provided a brief description of the convolution theorem for a better understanding of the present invention.

[00114] Suppose that $F(\omega)$ and $G(\omega)$ are the Fourier Transforms (FT) of $f(x)$ and $g(x)$, then equation (4) is as follows:

[00115]

$$\begin{aligned} F(\omega) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{i\omega x} dx, & G(\omega) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} g(x) e^{i\omega x} dx, \\ f(x) &= \int_{-\infty}^{\infty} F(\omega) e^{-i\omega x} d\omega, & g(x) &= \int_{-\infty}^{\infty} G(\omega) e^{-i\omega x} d\omega. \end{aligned} \quad (4)$$

[00116] Let $H(\omega) = F(\omega) G(\omega)$. The Inverse Fourier Transform (IFT) of $H(\omega)$ is:

[00117]

$$h(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(s) g(x-s) ds = \frac{1}{2\pi} \int_{-\infty}^{\infty} g(s) f(x-s) ds \quad (5)$$

[00118] The first integral in is called the convolution of $f(x)$ and $g(x)$. In other words, the IFT of the product of two FTs is the $1/2\pi$ times the convolution of the two functions.

[00119] For discrete time series, the Discrete Fourier Transform used is defined as equation (6) below:

[00120]

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \quad k = 0, \dots, N-1 \quad (6)$$

[00121] And the inverse Discrete Fourier Transform is defined as equation (7) below:

[00122]

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{\frac{2\pi i}{N} kn} \quad n = 0, \dots, N-1. \quad (7)$$

[00123] For faster implementation, Fast Fourier Transform can be used.

[00124] Now there is provided a brief description of peptide fragmentation and resultant mass spectra.

[00125] When a peptide is subjected to fragmentation by collision with neutral gas molecules, the fragmentation is induced by transfer of kinetic energy from the neutral gas molecules to the peptide. While the breakage can occur between any bonds in the peptide, it commonly occurs at the peptide bond. When a peptide is fragmented at a single peptide bond between the carbonyl and nitrogen, two fragments are formed. In the case where one peptide fragment retains the positive charge at the C-terminus of the peptide ion, it is called a y-ion. If the fragment retains the positive charge at the N-terminus, it is known as a b-ion. A b-ion usually has a complementary y-ion. When a singly charged peptide is fragmented, the charge is retained only at one terminus and only the fragment containing the charge is detected while the other fragment is lost as a neutral fragment. Doubly charged peptides tend to produce two singly charged ions, though sometimes doubly charged ions can also

be formed. b-ions and y-ions are usually formed when fragmentation occurs under low energy conditions.

[00126] As shown in FIG 1, fragmentation may cause various breakage points within a peptide, resulting in different fragments. In addition to y-ion and b-ion, other ions including a-ions and z-ions (complementary pairs) and c-ions and x-ions (complementary pairs) are also formed. These ions are formed when fragmentation occurs at high-energy conditions since higher amounts of energy are required to break these bonds.

[00127] When mass spectra contain mainly of y-ions and b-ions, they can be used for protein sequencing and identification. FIG 2 shows an exemplary MS/MS spectrum of a peptide AVAGCAGAR. The weights of the b-ions are as follows:

[00128] $b(n) = [171.11 \ 242.15 \ 299.17 \ 473.22 \ 544.26 \ 601.28 \ 672.31]$.

[00129] The weights of the y-ions are as follows:

[00130] $y(n) = [775.39 \ 676.32 \ 605.28 \ 548.26 \ 374.22 \ 303.18 \ 246.16 \ 175.12]$.

[00131] To derive the peptide sequence from the above spectrum, the weight differences between the b-ion peaks and the weight differences between the y-ion peaks are calculated. For example, from the spectrum above, the weight difference between b_2 and b_3 ions is 71.04 Daltons, which corresponds to the amino acid 'Alanine', and the weight difference between b_3 and b_4 ions is 57.02 Daltons, which corresponds to the amino acid 'Glycine', and so on, thus yielding the final peptide sequence of AVAGCAGAR.

[00132] The inventors of the present application discovered that there is symmetry in mass spectra data around the mid point of a tandem MS/MS spectrum, where the mid point equals to (mass-to-charge ratio of peptide) / 2. For example, as shown in FIG 2, the weight difference between b_1 and b_2 is equivalent to that between y_8 and y_7 as they represent the same amino acid 'Alanine' at 71.04 Dalton and b_1 corresponds to y_8 while b_2 corresponds to y_7 . Likewise, the weight difference between b_2 and b_3 is equivalent to that which is between y_7 and y_6 as they represent the same amino acid 'Glycine' at 57.02 Dalton and b_3 corresponds to y_6 , and so on.

[00133] This observed symmetry proved to be a useful feature that is suitable for the application of self-convolution. As shown below, the application of self-convolution to symmetrical MS/MS spectra data can be used to determine the quality of the spectra generated from the mass spectrometer.

[00134] As convolution is an integral that expresses the amount of overlap of one function as it is shifted over another function, the self-convolution of the mass spectra should produce the highest peak when all the corresponding b-ions and y-ions of one graph are aligned with the y-ions and b-ions of the ‘horizontally flipped’ graph. For example, for the spectrum shown in FIG 2, the highest peak occurs at the mid point where $y_7, y_6, y_5, y_4, y_3, y_2$ of MS/MS spectrum correspond respectively to $b_2, b_3, b_4, b_5, b_6, b_7$ of the ‘horizontally flipped’ MS/MS spectrum, and vice versa. Thus, the more corresponding b and y ions that are present in a MS/MS spectrum, the higher the midpoint peak.

[00135] To verify the observation, the molecular weights of the theoretical b-ions and y-ions were generated for peptide sequence **MTDQEAIQDLWQWR**, using **MS-Digest** from The Regents of the University of California.

[00136] The b-ions thus obtained were:

[00137] $b(n) = [233.10, 348.12, 476.18, 605.22, 676.26, 789.35, 917.40, 1032.43, 1145.51, 1331.59, 1459.65, 1645.73]$.

[00138] The y-ions generated were:

[00139] $y(n) = [1688.80, 1587.76, 1472.73, 1344.67, 1215.63, 1144.59, 1031.51, 903.45, 788.42, 675.34, 489.26, 361.20, 175.12]$.

[00140] A time series $data(n)$ is then setup such that the starting mass is 0 Da and the ending mass is 1819.84, the monoisotopic peptide mass (MH^+) generated by MS-Digest, with an interval of 0.01 Da. The following conditions are used to set the time series $data(n)$ as equation (8) below:

$$[00141] \quad data(n) = \begin{cases} 100 & \text{if } mass = b(n) \text{ or } mass = y(n) \\ \text{else (noise level)} & 1 \end{cases} \quad (8)$$

[00142] FIG 3 shows a plot of the b-ions and y-ions from peptide sequence **MTDQEAIQDLWQWR** processed by equation (8).

[00143] The conventional self-convolution is used to compute the output:

$$h_i = \sum_{j=0}^m f_j g_{i-j} \quad (3)$$

[00144] where g is the ‘horizontal flip’ of f .

[00145] FIG 4 is the resultant plot from the application of self-convolution to the plot shown in FIG 3. As shown in FIG 4, a prominent peak occurs at the mid point of the self-convolution plot as the b-ions ($b_n, b_{n-1}, b_{n-2}, \dots b_2$) were aligned with their corresponding y-ions ($y_2, y_3, y_4, \dots y_n$). However, the cumulating sum of the product of all the points steadily increases from 0 to the midpoint and then reduces thereof, forming a triangle below the peaks. This is potentially damaging to the detection of the peaks especially in non-ideal cases where significant noise levels are present, compounded by low b-ions and y-ions peaks and missing peaks.

[00146] To determine the effects of increasing noise levels, a plot of noise level set to 10 is generated below.

$$[00147] \quad data(n) = \begin{cases} 100 & \text{if } mass = b(n) \text{ or } mass = y(n) \\ \text{else } (noise \text{ level}) & 10 \end{cases} \quad (9)$$

[00148] FIG 5 shows a plot of the b-ions and y-ions from peptide sequence **MTDQEAIQDLWQWR** processed by equation (9).

[00149] FIG 6 is the resultant plot from the application of self-convolution to the plot shown in FIG 5. As shown in FIG 6, the self-convolution plot, while the noise level is only 10% of the ions intensity, has a significantly reduced mid-point peak in comparison to the increased overall overlapping convolution values. In addition, the other observable peaks in FIG 4 are also lost in view of the greatly increased overlapping convolution values due to an increase in noise levels.

[00150] In order to solve this problem, the Convolution Theorem is introduced with a DC component shift at the transformed time series. By applying the convolution theorem, the Discrete Fourier Transform (DFT) was first performed onto the data(n).

$$[00151] \quad D(f) = DFT(data(n)) \quad (10)$$

[00152] To perform the self convolution in frequency domain, the product of the DFTs is computed and followed by the inverse DFT of the product:

$$[00153] \quad DD(f) = D(f) \times D(f) \quad (11)$$

[00154] The near DC components are removed by setting the first 10 points of the DFT to 0:

$$[00155] \quad DD(1:10) = 0 \quad (12)$$

$$[00156] \quad IDD(n) = iDFT(DD(f)) \quad (13)$$

[00157] Only the real part of the inverse DFT will be considered herein. And the intensity is normalized against the maximum value:

$$[00158] \quad RIDD(n) = \text{real}(IDD(n)) \quad (14)$$

$$[00159] \quad NRIDD(n) = RIDD(n)/\max(RIDD(n)) \quad (15)$$

[00160] FIG 7 is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 3. FIG 8 is the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5. As shown in FIGS 7 and 8, the maximum peak at the mid point and the other observable peaks as compared with FIGS 4 and 6 are preserved, so that the detrimental effects of noise level have been eliminated. We have therefore established a method when a region of the DC-Shifted Self-Convolution can be used to detect the existence of the corresponding b-ions and the y-ions, hence be able to determine the completeness and quality of the tandem mass spectra.

[00161] In real experiments, MS/MS spectra data are not perfect; instead they suffer from one or more of the defects including: Noise interference; Lost of b & y ions; and Diminishing b & y ions. The robustness of the DC shifted self-convolution of the present invention will be demonstrated hereinafter by hypothetical and real examples.

[00162] Referring to FIG 9, there is provided the resultant plot from the application of DC shifted self-convolution to the plot shown in FIG 5 except for the noise level being set at 25. At noise level = 25, the mid point peak is still detectable as shown in FIG 9.

[00163] Referring to FIGS 10(a)-(l), there is provided a panel of 12 resultant plots from the application of DC shifted self-convolution to the plots shown in either FIG3 or FIG 5, where each plot shows the effects of progressively randomly removing the b-ions with noise level 1 and 10 on the mid-point peak. As the number of b-ion peaks decreases, the relative height difference between the mid-point peak and the rest of the peaks decreases until eventually the mid point is over taken by the peaks nearer the right of the plots as shown in FIGS 10(i)-(j) with minus 10 b-ion peaks.

[00164] Referring to FIGS 11(a)-(n), there is provided a panel of 14 resultant plots from the application of DC shifted self-convolution to the plots shown in either FIG 3 or FIG 5 where each plot shows the effects of progressively randomly removing the y-ions with noise level 1 and 10 on the mid-point peak. Like the case of removing b-ion peaks, as the number of y-ion peaks decreases, the relative height difference between the mid-point

peak and the rest of the peaks decreases until eventually the mid point is over taken by the other peaks.

[00165] Referring to FIGS 12(a)-(e), there is provided a panel of 5 resultant plots from the application of DC shifted self-convolution to the plot shown in FIG 3 where each plot shows the effects of progressively randomly removing both b- and y-ions with noise level 1 on the mid-point peak. As the number of ion peaks that are removed increases, it is noted that the relative height difference between the highest mid point peak and the rest of the peaks reduces until eventually when the number of peaks removed are significant enough, the midpoint peak is no longer the highest peak.

[00166] Referring to FIGS 13(a)-(f), there is provided a panel of 6 resultant plots from the application of DC shifted self-convolution to the plots shown in either FIG 3 or FIG 5 where each shows the effects of progressively diminishing the intensity of b-ions peaks on the mid-point peak. As in the case of removing b-ion peaks, as the intensity of b-ion peaks diminishes, the relative height between the mid point highest peak and the rest of the peaks diminishes proportionally until eventually the peaks to the right of the mid point increase beyond the mid point peak.

[00167] Referring to FIGS 14(a)-(f) there is provided a panel of 6 resultant plots from the application of DC shifted self-convolution to the plots shown in either FIG 3 or FIG 5 where each shows the effects of progressively diminishing the intensity of y-ions peaks on the mid-point peak. As in the case of diminishing b-ion peaks, as the intensity of y-ion peaks diminishes, the relative height between the mid point highest peak and the rest of the peaks diminishes proportionally until eventually the peaks to the left of the mid point increase beyond the mid point.

[00168] Now there is provided a more detailed description of quantitative measurement of the quality of any given MS/MS spectra in accordance with one embodiment of the present invention. Referring now to FIG 15, the quantitative measurement is generally performed by the following operations: (1) first find the maximum peak in spectrum and determine that the maximum peak occurs at mid point of the MS/MS spectra (± 2 Da, error window for MS/MS fragment ion mass values); (2) find 5 highest peaks to the left of and 5 highest peaks to the right of the maximum mid point peak; and (3) calculate the ratio of the maximum mid point peak to the average of the 10 highest peaks to the left and right of the maximum mid point peak.

[00169] Table 1 below shows the results of the different scenarios as described above for theoretical spectrum of protein sequence MTDQEAIQDLWQWR.

Protein Sequence : MTDQEAIQDLWQWR					
S/No	Description	Max mid point peak	Max mid point peak value	Average of 10 highest peaks	Ratio of max mid point to 10 highest peaks
1	noise level 1	910.92	1.00	0.22	4.52
2	noise level 10	910.92	1.00	0.22	4.47
3	add 10 random peaks, noise level 1	910.92	1.00	0.24	4.15
4	add 20 random peaks, noise level 1	910.92	1.00	0.24	4.14
5	b-ions peaks reduced by 10%, noise level 1	910.92	1.00	0.23	4.30
6	b-ions peaks reduced by 20%, noise level 1	910.92	1.00	0.25	4.04
7	b-ions peaks reduced by 30%, noise level 1	910.92	1.00	0.27	3.73
8	b-ions peaks reduced by 40%, noise level 1	910.92	1.00	0.29	3.40
9	b-ions peaks reduced by 50%, noise level 1	910.92	1.00	0.33	3.02
10	b-ions peaks reduced by 60%, noise level 1	910.92	1.00	0.41	2.43
11	b-ions peaks reduced by 70%, noise level 1	-	-	-	-
12	b-ions peaks reduced by 80%, noise level 1	-	-	-	-
13	b-ions peaks reduced by 90%, noise level 1	-	-	-	-
14	y-ions peaks reduced by 10%, noise level 1	910.92	1.00	0.22	4.46
15	y-ions peaks reduced by 20%, noise level 1	910.92	1.00	0.23	4.36
16	y-ions peaks reduced by 30%, noise level 1	910.92	1.00	0.24	4.20
17	y-ions peaks reduced by 40%, noise level 1	910.92	1.00	0.25	3.99
18	y-ions peaks reduced by 50%, noise level 1	910.92	1.00	0.27	3.65
19	y-ions peaks reduced by 60%, noise level 1	910.92	1.00	0.33	2.99
20	y-ions peaks reduced by 70%, noise level 1	910.92	1.00	0.44	2.25
21	y-ions peaks reduced by 80%, noise level 1	-	-	-	-
22	y-ions peaks reduced by 90%, noise level 1	-	-	-	-
23	minus 2 b-ions peaks, noise level 1	910.92	1.00	0.25	4.06
24	minus 4 b-ions peaks, noise level 1	910.92	1.00	0.27	3.64
25	minus 6 b-ions peaks, noise level 1	910.92	1.00	0.32	3.13
26	minus 8 b-ions peaks, noise level 1	910.92	1.00	0.48	2.09
27	minus 10 b-ions peaks, noise level 1	-	-	-	-
28	minus 2 y-ions peaks, noise level 1	910.92	1.00	0.26	3.79
29	minus 4 y-ions peaks, noise level 1	910.92	1.00	0.26	3.80
30	minus 6 y-ions peaks, noise level 1	910.92	1.00	0.30	3.32
31	minus 8 y-ions peaks, noise level 1	910.92	1.00	0.26	3.80
32	minus 10 y-ions peaks, noise level 1	910.92	1.00	0.44	2.29
33	minus 12 y-ions peaks, noise level 1	-	-	-	-

[00170] As shown in Table 1, when all b and y ions peaks are present, the ratio of 'Mid point Maximum peak' to 'the Average of 10 highest peaks' is 4.52 (noise level 1). As b or y ion peaks reduce in intensity, the ratio drops to 2.25 and 2.09 respectively until eventually the max mid points are not found at reduced 70% of b ions and reduced 80% of y ions (see S/No 5 to 13 and 24 to 22 in Table 1). Similar trends are observed when removing b and y ion peaks. As the number of b or y peaks removed increases, the ratio values reduced until the max mid point no longer exist (see S/no 23 to 27 and 28 to 33 in Table 1).

[00171] In addition, the robustness of the system can be shown when random peaks were added to the spectra. The ratios of spectra with 10 and 20 random peaks added (at 4.15 and 4.14 respectively) differ only slightly from the ratio for noise level 1 spectrum (4.52).

[00172] Similar trends are observed in another theoretical spectrum HNPTSNVLFLGR of as shown in Table 2.

Protein Sequence : HNPTSNVLFLGR					
S/No	Description	Max mid point peak	Max mid point peak value	Average of 10 highest peaks	Ratio of max mid point to 10 highest peaks
1	noise level 1	677.95	1.00	0.19	5.29
2	noise level 10	677.95	1.00	0.23	4.36
3	add 10 random peaks, noise level 1	677.95	1.00	0.19	5.26
4	add 20 random peaks, noise level 1	677.95	1.00	0.20	4.96
5	b-ions peaks reduced by 10%, noise level 1	677.95	1.00	0.20	4.95
6	b-ions peaks reduced by 20%, noise level 1	677.95	1.00	0.22	4.61
7	b-ions peaks reduced by 30%, noise level 1	677.95	1.00	0.24	4.22
8	b-ions peaks reduced by 40%, noise level 1	677.95	1.00	0.26	3.78
9	b-ions peaks reduced by 50%, noise level 1	677.95	1.00	0.30	3.28
10	b-ions peaks reduced by 60%, noise level 1	677.95	1.00	0.38	2.61
11	b-ions peaks reduced by 70%, noise level 1	677.95	1.00	0.51	1.97
12	b-ions peaks reduced by 80%, noise level 1	-	-	-	-
13	b-ions peaks reduced by 90%, noise level 1	-	-	-	-
14	y-ions peaks reduced by 10%, noise level 1	910.92	1.00	0.22	4.46
15	y-ions peaks reduced by 20%, noise level 1	910.92	1.00	0.23	4.36
16	y-ions peaks reduced by 30%, noise level 1	910.92	1.00	0.24	4.20
17	y-ions peaks reduced by 40%, noise level 1	910.92	1.00	0.25	3.99
18	y-ions peaks reduced by 50%, noise level 1	910.92	1.00	0.27	3.65
19	y-ions peaks reduced by 60%, noise level 1	910.92	1.00	0.33	2.99
20	y-ions peaks reduced by 70%, noise level 1	910.92	1.00	0.44	2.25
21	y-ions peaks reduced by 80%, noise level 1	-	-	-	-
22	y-ions peaks reduced by 90%, noise level 1	-	-	-	-
23	minus 2 b-ions peaks, noise level 1	677.95	1.00	0.19	5.32
24	minus 4 b-ions peaks, noise level 1	677.95	1.00	0.26	3.81
25	minus 6 b-ions peaks, noise level 1	677.95	1.00	0.44	2.28
26	minus 8 b-ions peaks, noise level 1	-	-	-	-
27	minus 10 b-ions peaks, noise level 1	-	-	-	-
28	minus 2 y-ions peaks, noise level 1	677.95	1.00	0.22	4.48
29	minus 4 y-ions peaks, noise level 1	677.95	1.00	0.25	3.96
30	minus 6 y-ions peaks, noise level 1	677.95	1.00	0.25	3.96
31	minus 8 y-ions peaks, noise level 1	677.95	1.00	0.34	2.95
32	minus 10 y-ions peaks, noise level 1	-	-	-	-

[00173] EXAMPLES

[00174] The following examples are for the sole purpose of illustrating the present invention. It by no means intends to limit the scope of the present invention.

[00175] The DC shifted self-convolution was tested on real MS/MS spectra. The quality of b-ions and y-ions in spectra was confirmed by subjecting the MS/MS spectra to Mascot program, analyzing the actual ion peaks versus theoretical ones.

[00176] Example 1

[00177] FIG 16 (a) shows a real experimental MS/MS data, a spectrum of good quality, that was used for the test.

[00178] As shown in FIG 16(b), the DC shifted self-convolution plot reveals a high mid point peak, indicating of a good spectrum with most if not all the b-ion and y-ion peaks present. FIG 17 shows the MASCOT prediction result of the above real experiment data, confirming the presence of nearly all b and y ions.

[00179] Example 2

[00180] FIG 18(a) shows another real experimental MS/MS data that is of good quality. As shown in FIG 18(b), the DC shifted self-convolution plot reveals a high mid point peak, indicating of another good spectrum with most if not all the b-ion and y-ion peaks present. FIG 19 shows the MASCOT prediction results of the above real experiment data, confirming the presence of nearly all b and y ions.

[00181] Example 3

[00182] FIG 20(a) shows a poor quality MS/MS data. As shown in FIG 20(b), the DC shifted self-convolution plot reveals a low or missing mid point peak, indicating of a poor spectrum. FIG 21 shows the MASCOT prediction results of the above real experiment data, confirming the poor quality of the real experiment data.

[00183] Example 4

[00184] FIG 22 (a) shows a poor quality MS/MS data. As shown in FIG 22(b), the DC shifted self-convolution plot reveals a false mid point peak, resulting in wrong prediction of a poor quality spectrum as good quality. This is due to an abnormally high peak in the center of the spectrum which when convoluted falsely gives a high mid point peak. FIG 23 shows the MASCOT prediction results of poor quality MS/MS spectrum showing many 'predicted' peaks that are too low to be significant.

[00185] Example 5

[00186] As shown in Example 4, a mass spectra data with poor quality may be classified as one with good quality in exceptional cases. To alleviate this problem, in one

embodiment of the present invention, a peak selection step is performed before subjecting spectrum to DC shifted self-convolution.

[00187] In order to eliminate the effects of high mid point peak in MS/MS spectra affecting the DC shifted self-convolution method of prediction, 30 maximum peaks of each spectrum is ferreted and a new spectrum is then generated with the following criteria before using the DC shifted self-convolution method.

$$[00188] \quad data(n) = \begin{cases} 100 & \text{if one of 30 highest peaks} \\ & else \quad 0 \end{cases} \quad (16)$$

[00189] FIG 24(a) shows a poor quality MS/MS data as previously shown in FIG 22(a). FIG 24(b) shows a plot of b-ions and y-ions of 30 highest peaks from FIG 24(a). As shown in FIG 24(c), the DC shifted self-convolution plot of FIG 24(a) reveals a false mid point peak, resulting in wrong prediction of a poor quality spectrum as good quality, as previously shown in FIG 22(b). FIG 24(d) shows the DC shifted self-convolution plots of the newly generated spectrum of FIG 24(b), where the spurious high mid point in Fig 24(c) is eradicated, thus correctly predicting the poor quality of spectrum in Fig 24(a).

[00190] By picking the 30 highest peaks and giving them a constant value, the effects of abnormally high peaks in spectra are removed. The number 30 is the number of b and y ions that is believed to be found in the spectrum, but this number can be increased without much effect to the prediction method as the proposed method is proven previously to be robust even when random spurious b or y ion peaks are added. The number should increase with increasing mass-to-charge ratio of the peptide as a longer peptide with more b and y ions are expected.

[00191] Example 6

[00192] FIG 25(a) shows the same MS/MS plot of data as shown in FIG 16(a) of Example 1. FIG 25(b) shows the plots of b-ions and y-ions of 30 highest peaks from FIG 25(a) according to the methods as described in Example 5. FIG 25(c) shows the same DC shifted self-convolution plot of FIG 16(b). FIG 25(d) shows the DC shifted self-convolution plot of the newly generated spectrum of FIG 25(b).

[00193] Example 7

[00194] FIG 26(a) shows the same MS/MS plot of data as shown in FIG 18(a) of Example 2. FIG 26(b) shows the plots of b-ions and y-ions of 30 highest peaks from FIG 26(a) according to the methods as described in Example 5. FIG 26(c) shows the same DC

shifted self-convolution plot of FIG 18(b). FIG 26(d) shows the DC shifted self-convolution plot of the newly generated spectrum of FIG 26(b).

[00195] Example 8

[00196] FIG 27(a) shows the same MS/MS plot of data as shown in FIG 20(a) of Example 3. FIG 27(b) shows the plots of b-ions and y-ions of 30 highest peaks from FIG 27(a) according to the methods as described in Example 5. FIG 27(c) shows the same DC shifted self-convolution plot of FIG 20(b). FIG 27(d) shows the DC shifted self-convolution plot of the newly generated spectrum of FIG 27(b).

[00197] In certain embodiments of the present invention, there is provided a system by which MS/MS spectra data can be processed by the DC shifted self-convolution as described above. The system comprises electronic means for receiving the MS/MS spectra data, and computer executable storage medium in which the computer programs are embedded. The embedded programs can perform the DC shifted self-convolution on MS/MS spectra data. The system further comprises an output means for outputting and displaying the results. The system may be any electronic device including PC, Notebook, or the like.

[00198] While the present invention has been described with reference to particular embodiments, it is understood that the embodiments are illustrative and that the invention scope is not so limited as such. Alternative embodiments of the present invention will become apparent to those having ordinary skill in the art to which the present invention pertains. Such alternate embodiments are considered to be encompassed within the spirit and scope of the present invention. Accordingly, the scope of the present invention is described by the appended claims and is supported by the foregoing description.

CLAIMS

What is claimed is:

1. A process for evaluating the quality of a MS/MS spectrum of a polypeptide, said process comprising the steps of:

obtaining the MS/MS spectrum of a polypeptide from a source; and
performing self-convolution upon the MS/MS spectrum, and then outputting a self-convolution plot;

thereby the quality of the MS/MS spectrum is evaluated on parameters of the self-convolution plot.

2. The process of claim 1, wherein the step of performing self-convolution comprises performing the self-convolution according to the following equation:

$$h(x) = f(x) * g(x)$$

where $g(x)$ is the 'horizontal flip' of function $f(x)$.

3. The process of claim 1, wherein the parameters of the self-convolution plot include the position of the highest peak, the amplitudes of the peaks, and the ratio between the highest peak and other peaks.

4. The process of claim 1, wherein the step of performing self-convolution comprises performing a DC shifted self-convolution according to the followings:

performing the Discrete Fourier Transform (DFT) onto $data(n)$ according to $D(f) = DFT(data(n))$, wherein the $data(n)$ is a mass spectrum;

performing the self convolution in frequency domain, wherein the product of the DFTs is computed and followed by the inverse DFT of the product according to $DD(f) = D(f) \times D(f)$;

removing the near DC components by setting the first 10 points of the DFT to 0 according to $DD(1:10) = 0$, and $IDD(n) = iDFT(DD(f))$; and

considering only the real part of the inverse DFT, and normalizing the intensity against the maximum value according to $RIDD(n) = \text{real}(IDD(n))$, and $NRIDD(n) = RIDD(n)/\max(RIDD(n))$.

5. The process of claim 1, wherein before the step of performing self-convolution, further comprising a step of selecting n highest peaks from the MS/MS spectrum and normalizing the n highest peaks into the same amplitude so that a new mass spectrum is generated, wherein the n is an integral; thereby the self-convolution is performed upon the newly generated mass spectrum.

6. A system for evaluating the quality of a MS/MS spectrum of a polypeptide, said system comprising:

a means for obtaining the MS/MS spectrum from a source; and

a computer executable storage medium in which computer programs are embedded, wherein the computer programs enable the system to perform self-convolution upon the MS/MS spectrum; thereby the quality of the MS/MS spectrum of a polypeptide can be evaluated.

7. The system of claim 6, wherein the computer programs perform the self-convolution according to the following equation:

$$h(x) = f(x) * g(x)$$

where $g(x)$ is the 'horizontal flip' of function $f(x)$.

8. The system of claim 6, further comprising a means for outputting a self-convolution plot after the computer programs perform the self-convolution.

9. The system of claim 8, wherein the parameters of the self-convolution plot include the position of the highest peak, the amplitudes of the peaks, and the ratio between the highest peak and other peaks.

10. The system of claim 6, wherein the computer programs perform a DC shifted self-convolution according to the followings:

performing the Discrete Fourier Transform (DFT) onto data(n) according to $D(f)=DFT(data(n))$, wherein the data(n) is a mass spectrum;

performing the self convolution in frequency domain, wherein the product of the DFTs is computed and followed by the inverse DFT of the product according to $DD(f)=D(f) \times D(f)$;

removing the near DC components by setting the first 10 points of the DFT to 0 according to $DD(1:10)=0$, and $IDD(n)=iDFT(DD(f))$; and

considering only the real part of the inverse DFT, and normalizing the intensity against the maximum value according to $RIDD(n)=real(IDD(n))$, and $NRIDD(n)=RIDD(n)/max(RIDD(n))$.

11. The system of claim 6, wherein before performing self-convolution, the computer programs further perform a step of selecting n highest peaks from the MS/MS spectrum and normalizing the n highest peaks into the same amplitude so that a new mass spectrum is generated, wherein the n is an integral; thereby the self-convolution is performed upon the newly generated mass spectrum.