

SYSTEM AND PROCESS FOR PRE-FILTERING OF TANDEM MASS SPECTRA USING PIECEWISE CONVOLUTION

Field of the Invention

[0001] The present invention generally relates to mass spectrometry technologies, and more particularly to a system and process for pre-filtering of tandem mass spectra using piecewise convolution.

Background of the Invention

[0002] Mass Spectrometry (MS) in proteomics is a powerful analytical tool that is used for obtaining amino acid sequences of peptides and protein identification. 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] Mass spectrometry is experiencing a period of rapid growth because of the applications in proteomic analyses. As a consequence of the rapid growth, the need to collect the large amounts of information produced and make it available to researchers worldwide has ended with a number of MS databases. For example, the Illinois Bio-Grid Mass Spectrometry Database (IBG-MSD) is a public database of curated and annotated empirically derived mass spectra of peptides. The goal of the database is to address the need for a public database of mass spectrometry data and to implement a useful web interface that will allow researchers to access the data and perform a variety of tasks based on their individual needs.

[0004] 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.

[0005] Currently, the actual MS/MS data have to be pre-filtered before being used for identification of amino acid sequences of peptides; this is because the actual MS/MS data contain too many noises from the systems and/or artificial operations. Many known pre-filtering algorithms are available; a few of them will be described briefly herein.

[0006] The Open Mass Spectrometry Search Algorithm (OMSSA) is an efficient search engine for identifying MS/MS peptide spectra by searching libraries of known protein sequences. OMSSA scores significant hits with a probability score developed using classical hypothesis testing, the same statistical method used in BLAST. The OMSSA noise filter has several steps. The first step is to delete background peaks whose intensity is below a percentage of the maximum intensity peak. By default, OMSSA cuts peaks below 2.5% of the maximum intensity, although this value is user adjustable and is modified dynamically in the last part of the algorithm. The second step is to delete any remaining precursor ion peaks. The third step is to eliminate peaks that are obviously not mono-isotopic. This is accomplished by examining peaks in order of intensity and deleting peaks that are within 0 to 2 Da of the m/z value of the peak being examined. The last step in the noise filter is to filter out peaks that are too close together. In this step, precursor charge 1+ and 2+ spectra are treated differently from charge 3+ spectra. It is assumed that 1+ and 2+ spectra contain mainly 1+ product ions. For these spectra, the filter examines each peak in order of intensity. In a region ± 27 Da of the peak being examined, all except the most intense peak is deleted under the assumption that in a region of this size, no other peak from the same ion series will exist and only one ion from the complementary ion series can exist. 27 Da is used as it is less than the residue mass of the smallest amino acid.

[0007] (X) of chromatography mass spectrometry (XCMS) as a preprocessing and analysis algorithm includes peak detection, peak matching, and retention time alignment. The peak detection is based on cutting the LC/MS data into slices a fraction of a mass unit (0.1 m/z) wide and then operating on those individual slices in the chromatographic time domain. Within each slice, the signal is determined by taking the maximum intensity at each time point in the slice. After filtration, peaks are selected using a signal-to-noise ratio cutoff. Because of the second-derivative transformation and the resulting negative

[0020] In another embodiment of the system, the computer programs evaluates the quality of the MS/MS spectrum based on the number of peaks obtained from the performance of piecewise convolution.

[0021] 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

[0022] Preferred embodiments according to the present invention will now be described with reference to the Figures, in which like reference numerals denote like elements.

[0023] FIG 1 illustrates various breakage points within a peptide during fragmentation.

[0024] FIG 2 shows an exemplary MS/MS spectrum of a peptide AVAGCAGAR.

[0025] FIG 3 shows a typical MS/MS spectrum data including noise and isotopic peaks.

[0026] FIG 4 shows an exemplary illustration of the principles of the piecewise convolution in accordance with one embodiment of the present invention.

[0027] FIG 5 is a figure showing a common tandem mass spectrum.

[0028] FIG 6 is a figure showing a portion of the resultant tandem mass spectrum after the noise and isotopic filtering processes, where all peaks have normalized into 100.

[0029] FIG 7 is a figure showing the result of piecewise convolution performed on the tandem mass spectrum shown in FIG 6.

[0030] FIG 8 is a figure showing the scatter plot of MF Score versus Mascot score.

[0031] FIG 9 is a figure showing the scatter plot of MF Score versus SONAR score.

[0032] FIG 10 is a plot showing the sensitivity and the selectivity of the present invention.

[0033] FIG 11 shows the pseudo code of the pre-filtering algorithm using piecewise convolution that was used in the tests of the present invention.

[0034] FIG 12 shows the scores obtained from MASCOT, SONAR and the algorithm of the present invention.

[0035]

Detailed Description of the Invention

[0036] The present invention may be understood more readily with reference to the following detailed description of certain embodiments of the invention.

[0037] 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.

[0038] 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.

[0039] The present invention teaches various embodiments of the system and process that pre-filtering MS/MS spectrum data by employing the principles of convolution. The general principles of convolution and fragmentation are well known in their relevant arts; nonetheless, for better understanding of the present invention brief descriptions of convolution and fragmentation are provided herein.

[0040] Now there is provided a brief description of peptide fragmentation and resultant mass spectra.

[0041] 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.

[0042] 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.

[0043] 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:

[0044] $b(n) = [171.11 \ 242.15 \ 299.17 \ 473.22 \ 544.26 \ 601.28 \ 672.31]$.

[0045] The weights of the y-ions are as follows:

[0046] $y(n) = [775.39 \ 676.32 \ 605.28 \ 548.26 \ 374.22 \ 303.18 \ 246.16 \ 175.12]$.

[0047] 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.

[0048] 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. This observed symmetry proved to be a useful feature that is suitable for the application of convolution in pre-filtering MS/MS spectrum data.

[0049] Now there is provided a brief description of the fast Fourier transform and convolution.

[0050] The fast Fourier transform (FFT) is a discrete Fourier transform algorithm which reduces the number of computations needed for N points from $2N^2$ to $2N \lg N$, where $\lg$ is the base-2 logarithm. If the function to be transformed is not harmonically related to the sampling frequency, the response of an FFT looks like a sinc function (although the integrated power is still correct). Aliasing (leakage) can be reduced by apodization using a tapering function. However, aliasing reduction is at the expense of broadening the spectral response.

[0051] Fast Fourier transform algorithms generally fall into two classes: decimation in time, and decimation in frequency. The Cooley-Tukey FFT algorithm first rearranges the input elements in bit-reversed order, and then builds the output transform (decimation in time). The basic idea is to break up a transform of length N into two transforms of length $N/2$ using the identity according to the following equation (1):

[0052]

$$\begin{aligned} \sum_{n=0}^{N-1} a_n e^{-2\pi i n k / N} &= \sum_{n=0}^{N/2-1} a_{2n} e^{-2\pi i (2n) k / N} + \sum_{n=0}^{N/2-1} a_{2n+1} e^{-2\pi i (2n+1) k / N} \\ &= \sum_{n=0}^{N/2-1} a_n^{\text{even}} e^{-2\pi i n k / (N/2)} + e^{-2\pi i k / N} \sum_{n=0}^{N/2-1} a_n^{\text{odd}} e^{-2\pi i n k / (N/2)}, \end{aligned} \quad (1)$$

[0053] The algorithm for `ifft(x)` is the same as the algorithm for `fft(x)`, except for a sign change and a scale factor of $n = \text{length}(x)$. As for `fft`, the execution time for `ifft` depends on the length of the transform. It is fastest for powers of two. It is almost as fast for lengths that have only small prime factors. It is typically several times slower for lengths that are prime or which have large prime factors.

[0054] Convolution was performed using two blocks of 8 intensities each at any one time, starting from both ends of the spectrum, till the entire spectrum is being processed. 8 peaks was chosen because FFT was used to perform the convolution, hence the length is best to be of 2^n . In addition, the precision of this length would be 0.8 Dalton.

[0055] Let $f(t)$ and $g(t)$ be arbitrary functions of time t with Fourier transforms so as to have the following equations (2) and (3):

[0056]

$$f(t) = \mathcal{F}_v^{-1}[F(v)](t) = \int_{-\infty}^{\infty} F(v) e^{2\pi i v t} dv \quad (2)$$

$$g(t) = \mathcal{F}_v^{-1}[G(v)](t) = \int_{-\infty}^{\infty} G(v) e^{2\pi i v t} dv, \quad (3)$$

[0057] Where $\mathcal{F}_v^{-1}(t)$ denotes the inverse Fourier transform (where the transform pair is defined to have constants $A = 1$ and $B = -2\pi$). Then the convolution is performed according to the following equation (4):

[0058]

$$\begin{aligned} f * g &= \int_{-\infty}^{\infty} g(t') f(t - t') dt' \\ &= \int_{-\infty}^{\infty} g(t') \left[\int_{-\infty}^{\infty} F(v) e^{2\pi i v (t - t')} dv \right] dt', \end{aligned} \quad (4)$$

[0059] Interchange the order of integration according to the following equation (5):

[0060]

$$\begin{aligned} f * g &= \int_{-\infty}^{\infty} F(v) \left[\int_{-\infty}^{\infty} g(t') e^{-2\pi i v t'} dt' \right] e^{2\pi i v t} dv \\ &= \int_{-\infty}^{\infty} F(v) G(v) e^{2\pi i v t} dv \\ &= \mathcal{F}_v^{-1}[F(v) G(v)](t). \end{aligned} \quad (5)$$

[0061] Now, there is provided a more detailed description of the pre-filtering process in accordance with one embodiment of the present invention. FIG 3 shows a typical MS/MS spectrum data including noise and isotopic peaks. The pre-filtering process comprises noise peaks filtering, isotopic peaks removing, and piecewise convolution.

[0062] The noise peaks filtering is the initial step of the pre-filtering process. In one embodiment, the noise peaks in MS/MS spectrum data are removed by their intensities. A peak can be treated as a noise one if it falls below a pre-determined intensity threshold that can be set by a user according to certain parameters. The noise peaks are simply removed from the MS/MS spectrum data for any further processing.

[0063] The isotopic peaks result from the existence of isotopes for certain atoms, interfering with the identification of amino acid sequences of peptides; thus the following step of pre-filtering process is to remove the isotopic peaks. In one embodiment, only the

first high peak is kept while the isotopic peaks are removed, where the isotopic peaks are within 3 Dalton from the first peak.

[0064] After the MS/MS spectrum data have been cleaned by removing the noise and isotopic peak, they are ready for the piecewise convolution. As discussed above, the tandem mass spectrum owns the symmetry property. Now referring to FIG 4, there is provided an illustration of the principles of the piecewise convolution in accordance with one embodiment of the present invention. The algorithm first picks two “pieces” of spectra from the given spectrum; the first piece, “left piece”, is extracted from the lowest m/z end of the spectrum, and the second piece, “right piece”, is extracted from the highest m/z end of the spectrum, which also contains the precursor mass of the peptide for charge $z=1$. The algorithm then performs convolution on the two “pieces” of spectra. The peak values of this convolution are kept as the representative peaks of the selected piece. Then the “left piece” is represented by the next inner piece of the spectrum at its right side; likewise the next “right piece” is replaced by the inner piece of the spectrum at its left side. The peak values of this convolution are again kept as the representative peaks of the second piece. This process is iterated until the final “left piece” and “right piece” have been convolved, and the representative peaks of the last piece have been obtained.

[0065] Now referring to FIG 5, there is provided a figure showing a common tandem mass spectrum.

[0066] Now referring to FIG 6, there is provided a figure showing a portion of the resultant tandem mass spectrum after the noise and isotopic filtering processes, where all peaks have normalized into 100.

[0067] Now referring to FIG 7, there is provided a figure showing the result of piecewise convolution performed on the tandem mass spectrum shown in FIG 6. This pre-filtered mass spectrum can be sent for protein identification, either by de novo methods or database search methods.

[0068] There is another novel use of the result. The number of the peaks left with high intensity can be used as an indication of the quality of the mass spectrum. A spectrum producing higher number of peaks should be better than the one that produces fewer peaks. In one embodiment, the MS Filtered Score (MF Score) is defined as the quality measurement of the mass spectrum, by counting the number of high intensity peaks remain after the filtering process.

[0069] EXAMPLES

[0070] 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.

[0071] The pseudo code of the pre-filtering algorithm using piecewise convolution that was used in the tests of the present invention is shown in FIG 11.

[0072] The test was performed with 557 unpublished MS/MS datasets. Two software programmes were used to verify the results; the first one is MASCOT, and the second one is SONAR. The scores obtained from MASCOT, SONAR and the algorithm of the present invention are shown in FIG 12.

[0073] Now referring to FIG 8, there is provided a figure showing the scatter plot of MF Score versus Mascot score. There is a linear correlation between these scores, i.e. as the Mascot score increases, MF score increases linearly, which implies that good mass spectra produce high matching peak counts.

[0074] Now referring to FIG 9, there is provided a figure showing the scatter plot of MF Score versus SONAR score. There is a linear correlation between these scores, i.e. as the SONAR score increases, MF score increases linearly, which again implies that good mass spectra produce high matching peak counts.

[0075] Now referring to FIG 10, there is provided a plot showing the sensitivity and the selectivity of the present invention. A good MS is defined as the one that protein can be found in a database search. A bad MS is the one that does not yield a result after database search. A cautious note: there could be proteins which are not found in the current protein database that makes the search fails to identify any protein while the spectrum could be a good one. The plot of the good-MS and bad-MS curves shows an intersection at a score=0.3788. This threshold of score helps in differentiating the passing MS from the failing MS judging from the score itself.

[0076] Table 1. Results from SONAR

	Good-MS (SONAR)	Bad-MS (SONAR)
Higher than the intersection Point (0.3788)	243	38
Lower than the intersection Point (0.3788)	48	228

[0077] So the sensitivity and selectivity could be calculated:

[0078] Sensitivity = $243 / (243 + 48) = 0.835052 = 83.5052\%$

[0079] Selectivity = $223 / (231 + 30) = 0.854406 = 85.4406\%$

[0080] In certain embodiments of the present invention, there is provided a system by which MS/MS spectra data can be processed by the piecewise 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 piecewise 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.

[0081] 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 pre-filtering 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 piecewise convolution upon the MS/MS spectrum, and then outputting the piecewise convoluted MS/MS spectrum to an algorithm for identifying the amino acid sequence of the polypeptide;thereby the amino acid sequence of the polypeptide can be identified.
2. The process of claim 1, wherein before the step of performing piecewise convolution, further comprising a step of removing the noise peaks that have intensities lower than a pre-determined threshold.
3. The process of claim 1, wherein before the step of performing piecewise convolution, further comprising a step of removing isotopic peaks.
4. The process of claim 1, wherein before the step of performing piecewise convolution, further comprises a step of normalization, where all selected peaks are normalized into a pre-set scale.
5. The process of claim 1, wherein the step of performing piecewise convolution evaluates the quality of the MS/MS spectrum based on the number of peaks.
6. A system for pre-filtering a MS/MS spectrum of a polypeptide, said system comprising:
 - a computer executable storage medium in which computer programs are embedded, wherein the computer programs enable the system to obtain the MS/MS spectrum from a

source, and to perform piecewise convolution upon the MS/MS spectrum; thereby the amino acid sequence of the polypeptide can be generated.

7. The system of claim 6, wherein the computer programs further enables the system to remove the noise peaks that have intensities lower than a pre-determined threshold.

8. The system of claim 6, wherein the computer programs further enables the system to remove isotopic peaks.

9. The system of claim 6, wherein the computer programs further enables the system to normalize all selected peaks into a pre-set scale.

10. The system of claim 6, wherein the computer programs evaluates the quality of the MS/MS spectrum based on the number of peaks obtained from the performance of piecewise convolution.