BMC Bioinformatics

BioMed Central

Research article



# A mass accuracy sensitive probability based scoring algorithm for database searching of tandem mass spectrometry data

Hua Xu[1] and Michael A Freitas*[2]

Address: [1]Department of Chemistry, the Ohio State University, Columbus 43210, OH, USA and [2]Department of Molecular Immunology Virology and Medical Genetics, the Ohio State University, Columbus 43210, OH, USA

Email: Hua Xu - xu.171@osu.edu; Michael A Freitas* - freitas.5@osu.edu

* Corresponding author

Published: 20 April 2007

*BMC Bioinformatics* 2007, **8**:133 doi:10.1186/1471-2105-8-133

Received: 7 December 2006
Accepted: 20 April 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/133



## Abstract

**Background:** Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) has become one of the most used tools in mass spectrometry based proteomics. Various algorithms have since been developed to automate the process for modern high-throughput LC-MS/MS experiments.

**Results:** A probability based statistical scoring model for assessing peptide and protein matches in tandem MS database search was derived. The statistical scores in the model represent the probability that a peptide match is a random occurrence based on the number or the total abundance of matched product ions in the experimental spectrum. The model also calculates probability based scores to assess protein matches. Thus the protein scores in the model reflect the significance of protein matches and can be used to differentiate true from random protein matches.

**Conclusion:** The model is sensitive to high mass accuracy and implicitly takes mass accuracy into account during scoring. High mass accuracy will not only reduce false positives, but also improves the scores of true positive matches. The algorithm is incorporated in an automated database search program MassMatrix.

## Background

Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) has become one of the most used tools in mass spectrometry based proteomics [1]. In shotgun proteomics, peptides are separated using liquid chromatography and introduced into a mass spectrometer via an ionization interface. In tandem mass spectrometry, the peptide precursor ions are isolated and fragmented via collision-induced dissociation (CID) [2] with inert gas, electron capture dissociation (ECD) [3], surface induced dissociation (SID) [4] and/or electron transfer dissociation (ETD) [5]. The resulting tandem MS spectra contain product ion signatures that relate back to the identity of the peptide precursor ions [2,6,7].

Various algorithms have since been developed to automate the process for modern high-throughput LC-MS/MS experiments. These algorithms fall under two categories: *de novo* sequence inference and database searching [8]. The first approach identifies peptide sequences directly from the tandem MS data [9,10]. This type of algorithm is usually computationally expensive and limited by the mass accuracy of the tandem MS data [8]. The database searching algorithms identify peptides by comparison

with a protein sequence database [11]. In this approach, all potential peptides are created from the sequence database via digestion with proteases. Theoretical spectra containing product ion series appropriate for the given fragmentation technique are created for the peptides. All tandem MS spectra in the data set are then compared with the theoretical spectra [1]. Because of their relatively lower computation expense and higher compatibility with low mass accuracy spectra, database searching programs are more commonly used at this time [12].

There are also various probability based post-search methods used to statistically curate search results from database search algorithms [13,14]. These methods estimate the accuracy of protein/peptide identifications and compare search results from different algorithms based on a common standard. However, many models involve empirical parameters such as score from correlative scoring algorithms. Therefore they may possess biases as a result of parameter optimization or model training.

The key comparison between different algorithms lies in how each approach scores a potential match between experimental and theoretical spectra [11,15-25]. We recently developed a database searching program, MassMatrix that uses a mass accuracy sensitive statistical model for scoring. This approach is separate and distinct from algorithms that filter matches based on mass accuracy. In the latter high mass accuracy can be used to filter spectra by only searching tandem mass spectra whose precursor ion falls within the stated mass tolerance, and filtering product ions by high mass accuracy can further reduce the likelihood of a random match [26,27]. However, a score sensitive model implicitly takes mass accuracy into account during scoring. The model is rigorously derived and sensitive to the searching tolerance determined by the accuracy of mass spectrometer. High accuracy improves the sensitivity and selectivity of searches. The statistical scores represent the probability that a match is a random occurrence. In addition, a novel statistically derived algorithm to rigorously calculate protein scores from the statistically based peptide scores has been developed. Thus the protein scores reflect the significance of protein hits and can be used to differentiate true protein hits from random ones. Herein we describe the statistical models.

## Results

### *Multiple scoring algorithms*

The peptide matching algorithm contains two independent scoring models, including a descriptive model and a statistical model. These models are used to calculate three distinct scores for a peptide match. Each of the scores may be independently used to ascertain the quality of the match. Because each score is distinct, the combination of scores is useful for validating each peptide match. The two models and the application to calculating peptide match scores are described in detail in the following.

### *Descriptive peptide scoring model*

Descriptive scores do not strictly convey any statistical relevance and may be prone to bias due to the scoring parameters. However, they have proven to be useful and generally augment probability based scores [13]. The descriptive model used herein to calculate peptide match scores (S) is shown in eqn. 1.

$$S = 100 \frac{\sum_{i=1}^{n_{\text{match}}} I_i r_{\text{match}}^2 \max(0, \frac{n_{\text{match}} - 3}{n_{\text{match}}})}{\sqrt{L_{\text{pep}}}} \quad (1)$$

$I_i$ is defined as the standardized abundance of the $i^{\text{th}}$ product ion in the experimental spectrum (calculated by dividing the abundance of the $i^{\text{th}}$ product ion by the maximum abundance in the spectrum), $\sum_{i=1}^{n_{\text{match}}} I_i$ is the total standardized abundance of matched product ions, $n_{\text{match}}$ is the number of matched product ions, $r_{\text{match}}$ is the ratio of standardized abundance of matched product ions to total standardized abundance of the experimental spectrum, and $L_{\text{pep}}$ is the length of the peptide in the number of amino acids. Each of these factors contributes to the overall score as follows: $\sum_{i=1}^{n_{\text{match}}} I_i$ evaluates the quality of the match, $r_{\text{match}}^2$ introduces a penalty for unmatched product ions, $\max(0, \frac{n_{\text{match}} - 3}{n_{\text{match}}})$ is an arbitrary penalty for matches with poor fragmentation, $\sqrt{L_{\text{pep}}}$ is an additional penalty for peptides with long sequences and the constant 100 is used arbitrarily to scale the scores. By default, scores for a spectrum with less than three matched product ions will be 0 due to the arbitrary penalty. However, the minimum number of matched ions may be changed to any value. Reducing this number is especially valuable for the analysis of singly charged peptides that have characteristic C-terminal aspartic acid fragmentation [28]. The penalty for peptide length is included to normalize the scores. Peptides with longer sequences have more fragment ions and higher empirical scores than shorter sequences. The penalty results in long and short sequences both have similar scores for matches of similar quality. The choices of incorporating squared and square

root for the terms $n_{match}$ and $L_{pep}$ were empirically determined from the evaluation of tandem MS data sets collected from LCQ and LTQ-Orbitrap mass spectrometers.

### *Descriptive protein score*

For "true" matches, we assume that the scores are normally distributed with a mean of 20 and a variance of 25. This arbitrary distribution estimates the distributions observed from analysis of several datasets. The expected contribution of each match to the protein score will be $S \times \int_0^S \frac{e^{-(x-20)^2/50}}{5\sqrt{2\pi}} dx$. Thus, the protein score from the descriptively scored matches is calculated from eqn. 2.

$$\text{protein score} = \sum_i S_i \times \int_0^{S_i} \frac{e^{-(x-20)^2/50}}{5\sqrt{2\pi}} dx \qquad (2)$$

### *Probability based peptide scoring model*

In addition to the empirical score, a mass accuracy sensitive probability based scoring model was derived to evaluate peptide matches. The model determines the likelihood that an experimental spectrum match to a theoretical spectrum is a random occurrence. Consider a pair of spectra: one experimental and one theoretical. $W_e$ and $W_t$ denote their precursor masses respectively. In addition, the experimental data contains information regarding the abundance of product ions $I_i$ for each precursor, $W_e$. The model ultimately tests the following two hypotheses: the null hypothesis, $H_0$, states that a match is random, i.e. the theoretical spectrum is independent of the experimental; and the alternative hypothesis, $H_A$, states that the match is not random, i.e. the theoretical spectrum is related to the experimental one.

Scoring the match is performed in two stages: 1) match $W_e$ against $W_t$ within the specified precursor ion mass accuracy and 2) match all product ions in the experimental spectrum against the theoretical within the specified mass accuracy. Both stages rely on calculating the probability that the occurrence of an ion within a fixed mass window could be a random occurrence ( $p = \frac{\text{mass window}}{\text{mass range}}$ ).

To match the experimental precursor with that of theoretical peptide we first define the variable $q$:

$$q = \begin{cases} 1 & W_e \text{ and } W_t \text{ match} \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

Under $H_0$, the possibility that any precursor ion match ($q$ = 1) could be random is given in eqn. 4.

$$p_1 = \frac{2\tau_{pep}}{\Pi} \qquad (4)$$

In the above equation, $\tau_{pep}$ is the mass accuracy of the precursor ion and $\Pi$ is the detection range for the precursor ion. For each precursor ion the mass window is defined as $\pm\ \tau_{pep}$ ($2 \times \tau_{pep}$). Thus $q$ has a bernoulli ($p_1$) distribution under $H_0$. If the precursor ion masses of the pair of spectra do not match ($q$ = 0) then the second stage is skipped. If $q$ = 1 we proceed to stage 2 where we test the match of the experimental product ion spectrum against the theoretical spectrum.

The variable $b_i$ is defined for each product ion, $i$, in the experimental spectrum as follows:

$$b_i = \begin{cases} 1 & \text{peak } i \text{ matches} \\ 0 & \text{otherwise} \end{cases}. \qquad (5)$$

Under $H_0$, all matched product ions are random and independent occurrences. The probability that a product ion randomly matches any of the product ions in the theoretical spectrum is:

$$p_2 = \frac{\Pi_{theo}}{\Pi} \qquad (6)$$

where $\Pi_{theo}$ is the total coverage of the detection range for all product ions in the theoretical spectrum and $\Pi$ is the MS/MS detection range. It is assumed that $\Pi$ is the same as the precursor ion mass range. However, for instruments that have a dynamic detection range assuming a fixed value $\Pi$ will result in more conservative scores. For each product ion in the theoretical spectrum, the mass window is $\pm\ \tau_{msms}$ ($2 \times \tau_{msms}$). If we assume there is no overlap in the product ion mass windows, then $\Pi_{theo}$ is calculated using the following equation

$$\Pi_{theo} = 2\ m \times \tau_{msms}. \qquad (7)$$

The probability that any single matched product ion ($b_i$ = 1) could be random can be calculated using the eqn. 8

$$p_2 = \frac{2m \times \tau_{msms}}{\Pi} \qquad (8)$$

where $\tau_{msms}$ is the product ion mass accuracy and $m$ is the number of product ions within the detection range in the theoretical spectrum. Because the theoretical spectrum is independent of the experimental under $H_0$, all $b_i$ ($i$ = 1, 2 ..., n) are assumed to have an identical and independent bernoulli ($p_2$) distribution under $H_0$. The model is then used to perform two distinct tests. Each uses a different approach to evaluate the quality of a match: number of

matched product ions $x$ and total abundance of matched product ions $Y$.

### *The pp score*

The model is used to evaluate whether the number of matched product ions in an experimental spectrum could be a random occurrence. For all spectra whose precursor ion masses match, i.e. $q = 1$, the variable $x$ is defined as the number of product ions in the experimental spectrum that match the theoretical spectrum (eqn. 9) where $b_i$ ($i$ = 1, 2 ..., n) is defined in eqn. 5 and $n$ is the number of product ions in the experimental spectrum.

$$x = \sum_{i=1}^{n} b_i \tag{9}$$

Under $H_0$, all $b_i$ have an identical and independent bernoulli ($p_2$) distribution. Therefore, $x$ will have a binomial ($n$, $p_2$) distribution. Consequently the probability mass function for $x$ is:

$$p(x) = \frac{n!}{x!(n-x)!} p_2^x (1-p_2)^{n-x} \tag{10}$$

where $p_2$ is calculated from eqn. 6. The p-value, $\alpha$, is defined as the probability that the quality of a random match between a pair of spectra is greater than or equal to a match observed under $H_0$. The pp value, $\beta$, is defined as the negative common logarithm of the p-value:

$$\beta = -\log(\alpha) \tag{11}$$

We use $x$ to evaluate the quality of a match, such that the p-value is the probability that $x$ for a random match between the pair of theoretical and experimental spectra is greater than or equal to that of the actual match, $x = n_{\text{match}}$, under $H_0$. The p-value is:

$$\alpha = \sum_{x=n_{\text{match}}}^{n} p(x) = \sum_{x=n_{\text{match}}}^{n} \frac{n!}{x!(n-x)!} p_2^x (1-p_2)^{n-x} \tag{12}$$

and the pp value is

$$\beta = -\log(\alpha) = -\log\left( \sum_{x=n_{\text{match}}}^{n} \frac{n!}{x!(n-x)!} p_2^x (1-p_2)^{n-x} \right) \tag{13}$$

### *The pp2 score*

The second approach evaluates whether the total abundance of matched product ions in the experimental spectrum could be a random occurrence. $Y$ is defined as the total abundance of experimental product ions that match product ions in a given theoretical spectrum:

$$Y = \sum_{i=1}^{n} I_i b_i \tag{14}$$

where $I_i$ is the standardized abundance of the $i^{\text{th}}$ product ion in the experimental spectrum and $b_i$ is defined in eqn. 5. For clarity we define $\gamma_i = I_i b_i$ to give eqn. 15.

$$Y = \sum_{i=1}^{n} \gamma_i \tag{15}$$

However, to complete the test we must know the inherent distribution of $Y$. This distribution is unknown and thus pp2 values can not be precisely calculated as were the pp values based on the total number of matched product ions. In order to estimate the pp2 value, three assumptions are needed:

1. $I_i$ is identically and independently distributed across product ions in the experimental spectrum,

2. $b_i$ is uncorrelated with $I_i$ in the experimental spectrum,

3. the number of product ions, $n$, in the experimental spectrum is large ($n > 30$).

Under assumption 1, the mean $\mu_I$ and variance $\sigma_I^2$ for the distribution of $I_i$ are estimated by:

$$\begin{cases} \hat{\mu}_I = \dfrac{1}{n} \sum_{i=1}^{n} I_i \\ \hat{\sigma}_I^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (I_i - \hat{\mu}_I)^2 \end{cases} \tag{16}$$

Since $\gamma_i = I_i b_i$, assumption 2 yields eqn. 17 under $H_0$,

$$\begin{cases} \mu_\gamma = E_\gamma(\gamma_i) = E_I(E_{\gamma|I}(\gamma_i \mid I_i)) = E_I(p_2 I_i) = p_2 \mu_I \\ \sigma_\gamma^2 = E_\gamma(\gamma_i^2) - E_\gamma(\gamma_i)^2 = E_I(E_{\gamma|I}(\gamma_i^2 \mid I_i)) - (p_2 \mu_I)^2 = p_2(1-p_2)\mu_I^2 + p_2\sigma \end{cases} \tag{17}$$

Thus, $\mu_\gamma$ and $\sigma_\gamma^2$ can be estimated as:

$$\begin{cases} \hat{\mu}_\gamma = p_2 \hat{\mu}_I \\ \hat{\sigma}_\gamma^2 = p_2(1-p_2)\hat{\mu}_I^2 + p_2 \hat{\sigma}_I^2 \end{cases} \tag{18}$$

According to the central limit theorem, $Y$ is approximately a normal distribution with the following parameters under assumption 3, i.e. when $n$ is large ($n > 30$)

$$\begin{cases} \mu_Y = n\mu_\gamma \\ \sigma_Y^2 = n\sigma_\gamma^2 \end{cases} \tag{19}$$

The resulting probability density function is given in eqn. 20.

$$f_Y(Y) = \frac{e^{-(Y-\mu_Y)^2/(2\sigma_Y^2)}}{\sqrt{2\pi}\sigma_Y} \quad (20)$$

And $\mu_Y$ and $\sigma_Y^2$ are estimated by eqn. 21.

$$\begin{cases} \hat{\mu}_Y = n\,\hat{\mu}_\gamma = n\,p_2\,\hat{\mu}_I \\ \hat{\sigma}_Y^2 = n\,\hat{\sigma}_\gamma^2 = n\,\{p_2(1-p_2)\hat{\mu}_I^2 + p_2\hat{\sigma}_I^2\} = n\,p_2(1-p_2)\hat{\mu}_I^2 + n\,p_2\hat{\sigma}_I^2 \end{cases} \quad (21)$$

The p-value, $\alpha$, is the probability that $Y$ for a random match is greater than or equal to that of the actual match, $I_{\text{match}}$, under $H_0$. The p-value becomes:

$$\alpha = \int_{I_{\text{match}}}^{+\infty} f_Y(x)dx = \int_{I_{\text{match}}}^{+\infty} \frac{e^{-(x-\mu_Y)^2/(2\sigma_Y^2)}}{\sqrt{2\pi}\sigma_Y} dx \quad (22)$$

and is estimated by:

$$\alpha = \int_{I_{\text{match}}}^{+\infty} \frac{e^{-(x-\mu_Y)^2/(2\sigma_Y^2)}}{\sqrt{2\pi}\sigma_Y} dx \quad (23)$$

resulting in the pp2 value, $\beta$, as follows:

$$\beta = -\log(\alpha) = -\log\left(\int_{I_{\text{match}}}^{+\infty} \frac{e^{-(x-\mu_Y)^2/(2\sigma_Y^2)}}{\sqrt{2\pi}\sigma_Y} dx\right) \quad (24)$$

The pp2 value can be estimated by equation 17 very efficiently. However, the real distribution of $Y$ is more tailed to larger values than the normal distribution. Therefore, pp2 values are overestimated when they are large.

### *Distribution of pp value for random matches*

When $q = 0$, the algorithm always assigns pp value, $\beta = 0$ because the experimental and theoretical precursor ions do not match. The cumulative distribution function for pp value when $q = 0$ is shown in eqn. 25.

$$F_{\beta|q=0}(\beta) = \begin{cases} 1 & \beta = 0 \\ 0 & \beta > 0 \end{cases} \quad (25)$$

In statistical hypothesis testing, a p-value for a null hypothesis $H_0$ is always a uniform distribution on the interval [0, 1]. Therefore, the cumulative distribution function for p-value of a random match is continuously distributed as

$$F_{\alpha|q=1}(\alpha) = \alpha \quad (0 \le \alpha \le 1) \quad (26)$$

when $q = 1$. According to the definition of pp value (eqn. 11), the cumulative distribution function for pp value when $q = 1$ is

$$F_{\beta|q=1}(\beta) = 1 - 10^{-\beta} \quad (\beta \ge 0) \quad (27)$$

and the probability density function is

$$f_{\beta|q=1}(\beta) = \frac{d}{d\beta}F_{\beta|q=1}(\beta) = \frac{d}{d\beta}(1-10^{-\beta}) = \ln(10)10^{-\beta} \quad (\beta \ge 0). \quad (28)$$

Matches with pp or pp2 values under a critical value $\beta_c > 0$ are discarded, i.e. their pp values are assigned 0. Thus the distribution of pp value for random matches returned by the algorithm is

$$F^*_{\beta|q=1}(0) = \int_0^{\beta_c} f_{\beta|q=1}(x)dx = \int_0^{\beta_c} \ln(10)10^{-x}dx = 1-10^{-\beta_c} \quad (29)$$

and for $\beta \ge \beta_c > 0$,

$$F^*_{\beta|q=1}(\beta) = F_{\beta|q=1}(\beta) = 1-10^{-\beta} \quad (30)$$

Thus when $q = 1$

$$F^*_{\beta|q=1}(\beta) = \begin{cases} 1-10^{-\beta_c} & \beta = 0 \\ 1-10^{-\beta} & \beta \ge \beta_c \end{cases}. \quad (31)$$

Likewise we can specify the unconditional distribution of pp values for random matches as follows. Since $q$ has a bernoulli ($p_1$) distribution, we have

$$P(q) = \begin{cases} 1-p_1 & q = 0 \\ p_1 & q = 1 \end{cases}, \quad (32)$$

For $\beta = 0$ the cumulative distribution function becomes,

$$\begin{aligned} F_\beta(0) &= F_{\beta|q=0}(0)\times P_q(0) + F^*_{\beta|q=1}(0)\times P_q(1) \\ &= 1\times(1-p_1) + (1-10^{-\beta_c})\times p_1 = 1-p_1 10^{-\beta_c} \end{aligned} \quad (33)$$

and for $\beta \ge \beta_c > 0$, it becomes,

$$\begin{aligned} F_\beta(\beta) &= F_{\beta|q=0}(\beta)\times P_q(0) + F^*_{\beta|q=1}(\beta)\times P_q(1) \\ &= 0\times(1-p_1) + (1-10^{-\beta})\times p_1 = 1-p_1 10^{-\beta} \end{aligned} \quad (34)$$

The combined cumulative distribution function is thus,

$$F_\beta(\beta) = \begin{cases} 1 - p_1 10^{-\beta_c} & \beta = 0 \\ 1 - p_1 10^{-\beta} & \beta \geq \beta_c \end{cases} \quad (35)$$

When $\beta \geq \beta_c$, $F_\beta(\beta)$ is continuous and the probability density function of pp value for random matches is

$$f_\beta(\beta) = \frac{d}{d\beta} F_\beta(\beta) = \frac{d}{d\beta}(1 - p_1 10^{-\beta}) = p_1 \ln(10) 10^{-\beta} \quad (\beta \geq \beta_c) \quad (36)$$

### *Confidence level for pp and pp2 values*

The confidence level can also be determined for both pp and pp2. Suppose there are $r$ theoretical spectra within the protein sequence database. If we assume that all theoretical spectra are uncorrelated, eqn. 37 gives $\phi$, the number of random matches that have a pp value greater than or equal to $\beta$ under $H_0$ for any given experimental spectrum.

$$\phi = r\int_\beta^{+\infty} f_\beta(x)dx = r\int_\beta^{+\infty} p_1 \ln(10) 10^{-x} dx = r p_1 10^{-\beta}, \quad (37)$$

The confidence level, $\psi$, is defined as

$$\psi = -\log(\phi) = -\log(r\, p_1\, 10^{-\beta}) = \beta - \log(r) - \log(p_1) \quad (38)$$

where $\beta$ is either the pp or pp2 value, $r$ is the number of theoretical spectra within the protein sequence database, and $p_1$ is given in eqn. 4. Confidence levels calculated from pp value and pp2 value are referred as confidence level and confidence level2 respectively.

The confidence level is the negative common logarithm of the expected number of random matches with a pp value bigger than or equal to the one we observe for the corresponding experimental spectrum. Therefore, if the confidence level is below 0, more than one random match for the spectrum is expected and the corresponding match is highly suspect. From eqn. 38, the pp value is directly related to the confidence value. The confidence level is dependent upon the size of the database and degrades as the number of peptide created from the database increases.

### *The protein pp score*

The pp model is also used to calculate pp values for protein matches. Let $r_{\text{protein}}$ denote the total number of theoretical spectra created from a protein sequence and $n_{\text{spectra}}$ denote the total number of experimental spectra in the data set. The cross match of all experimental spectra with theoretical peptides for the protein sequence generates $n_{\text{match\_protein}} = r_{\text{protein}} \times n_{\text{spctra}}$ potential matches. The sum of reported pp values of all matches for the protein is calculated from eqn. 39.

$$B = \sum_{i=1}^{n_{\text{match\_protein}}} \beta_i \quad (39)$$

The statistical model is used to test the following hypotheses: $H_0$ – All peptide matches for a given protein are random and $H_A$ – At least one peptide match for a given protein is not random. We assume that $r_{\text{spectra}}$ theoretical spectra created from the protein sequence are uncorrelated to each other and that $n_{\text{spectra}}$ experimental spectra from the data set are uncorrelated to each other. Since $n_{\text{match\_protein}}$ is normally very large, $B$ is approximately a normal distribution with a mean of $\mu_B = n_{\text{match\_protein}} \times \mu_\beta$ and a variance of $\sigma_B^2 = n_{\text{match\_protein}} \times \sigma_\beta^2$ according to the central limit theorem.

According to the distribution of the pp value for random matches described above, the mean and variances of a random match are given by the following equations:

$$\mu_\beta = \int_0^{+\infty} x f_\beta(x)dx = \int_{\beta_c}^{+\infty} x p_1 \ln(10) 10^{-x} dx = \left(\frac{p_1}{\ln(10)} + p_1\beta_c\right) 10^{-\beta_c} \quad (40)$$

and

$$\sigma_\beta^2 = E(\beta^2) - \mu_\beta^2 = \int_0^{+\infty} x^2 f_\beta(x)dx - \mu_\beta^2 = \int_{\beta_c}^{+\infty} x^2 p_1 \ln(10) 10^{-x} dx - \mu_\beta^2$$
$$= \left(\frac{2p_1}{[\ln(10)]^2} + \frac{2p_1\beta_c}{\ln(10)} + p_1\beta_c^2\right) 10^{-\beta_c} - \left[\left(\frac{p_1}{\ln(10)} + p_1\beta_c\right) 10^{-\beta_c}\right]^2 \quad (41)$$

where $p_1$ is given in eqn. 4 and $\beta_c$ is the pp value threshold. Likewise for the sum of pp values for the protein, $B$, the mean and variance for the distribution under $H_0$ are given in eqn. 42:

$$\begin{cases} \mu_B = n_{\text{match\_protein}} \left(\frac{p_1}{\ln(10)} + p_1\beta_c\right) 10^{-\beta_c} \\ \sigma_B^2 = n_{\text{match\_protein}} \left\{ \left[\frac{2p_1}{[\ln(10)]^2} + \frac{2p_1\beta_c}{\ln(10)} + p_1\beta_c^2\right] 10^{-\beta_c} - \left[\left(\frac{p_1}{\ln(10)} + p_1\beta_c\right) 10^{-\beta_c}\right]^2 \right\} \end{cases} \quad (42)$$

The p-value for a protein, $\alpha_{\text{protein}}$, is defined to be the probability that the protein hit can have a sum of pp values from all its peptide matches greater than or equal to $B$ under $H_0$. Thus $\alpha_{\text{protein}}$ is given by

$$\alpha_{\text{protein}} = \int_B^{+\infty} f_B(x)dx = \int_B^{+\infty} \frac{e^{-(x-\mu_B)^2/(2\sigma_B^2)}}{\sqrt{2\pi}\sigma_B} dx. \quad (43)$$

and the protein pp value becomes

$$\text{protein pp value} = -\log(\alpha_{\text{protein}}) = -\log(\int_{B}^{+\infty} \frac{e^{-(x-\mu_B)^2/(2\sigma_B^2)}}{\sqrt{2\pi}\sigma_B} dx) \quad (44)$$

## Discussion

### *Effect of various spectral characteristics on scoring*

Five example spectra, shown in Figure 1, are used to illustrate the effect of various spectral characteristics on scoring. All spectra were collected on an LTQ-Orbitrap mass spectrometer (ThermoElectron Finnigan, San Jose, CA, USA) [29]. Precursor and product ions were mass analyzed by the Orbitrap to achieve a mass accuracy of < 5.0 ppm. The pp and pp2 values at different mass accuracies (0.01 Da, 0.1 Da and 1.0 Da) were calculated and listed in Table 1. Mass accuracy tolerances were specified as either relative or absolute for precursor ions but only as absolute tolerances for product ions. Absolute mass accuracy tolerances for product ions are computationally cheaper and yield a reasonable compromise between computational expense and accuracy. Good quality spectra (Figure 1a,1b) yielded high empirical and statistical scores as expected. These sequences in Figure 1a &1b illustrate that peptide length has little effect on the scoring. This observation is consistent with the statistical model lack of peptide sequence bias and the peptide length penalty included in the empirical score (eqn. 1). Low quality data (i.e. low signal to noise ratio) can still yield good scores if the most abundant ions are dominated by signal (Figure 1c). The most challenging spectra are those with few dominant signal peaks. Examples are shown in 1d &1e. These figures show the spectra with one single dominant ion due to N-terminal fragmentation of an internal Pro residue and the neutral loss of $H_2O$ at an N-terminal Glu residue [30]. The empirical scores were poorer for these cases since only a single ion mainly contributes to score (eqn. 1). However, the pp and pp2 values were not as severely affected and able to accurately discriminate these matches from false positives.

### *Comparison between pp and pp2 values*

The pp value is the primary discriminator for quality of matches. The pp2 value can provide a complementary assessment of quality when pp values are suspect. Although the pp and pp2 value have the same statistical basis, there are several differences between them: The pp value is based on the number of matched product ions and the pp2 value is based on the total abundance of matched product ions. The pp value can be underestimated when noise is present in the experimental spectrum especially at low mass accuracy. Because noise normally has lower abundance than product ions, the pp2 value, on the other hand, is generally unaffected. As shown in Table 1, pp value for the spectrum with low signal to noise ratio and majority of noise peaks (Figure 1c) were affected negatively by the noise peaks and relative low compared with those for normal spectra at mass accuracy of 1.0 Da. However, pp2 value was not affected by the noise peaks.

While pp value can be precisely calculated, there are three assumptions needed to estimate the pp2 values. Assumption 1 and 3 for the pp2 test are not plausible when the number of product ions in the experimental spectrum, $n$ is small. Therefore, pp2 value estimated by the central limit theorem cannot evaluate the quality of matches with a small number of product ions. Furthermore the normal distribution under the central limit theorem is less tailed than the true distribution of $Y$, pp2 value is normally overestimated when it is large (> 16) as shown in Table 1.

From the above discussion, the pp value is more reliable and accurate than the pp2 value under most circumstances, but it can be affected by noise. Under these circumstances, the number of product ions in the spectrum is normally large and pp2 value can be well estimated and complementary to pp value. Thus the combination of the two scores provides an excellent means to ascertain the quality of matches under conditions where one might fail.

### *Effect of mass accuracy on pp values*

In the pp model, the two most important parameters ($p_1$, the probability that a theoretical precursor randomly matches the experimental and $p_2$, the probability that a theoretical product ion randomly matches any product ions in the MS/MS spectrum) are set in accordance with the predetermined mass accuracy of mass spectrometer. These parameters' values decrease as mass accuracy increases. This effect is shown in Table 1. A more thorough list of all parameters used in calculating the empirical and statistical scores is provided as supplementary material [see Additional file 1]. The statistical model specifically takes each parameter into account when calculating the statistical scores. Therefore, these two parameters have a substantial effect on the pp values for both random matches and true matches. Consequently, pp and pp2 values are very sensitive to the accuracy of mass spectrometer.

As is shown in the Figure 2, the probability of random matches having high pp values is substantially reduced as we increase mass accuracy. Increasing mass accuracy resulted in a shift of the pp value distribution for random matches to lower values. At the same time the pp value distribution for true matches moves to higher pp values. This effect is evident from the pp values in Table 1. As mass accuracy improved, the pp values improved for all peptide matches in Figure 1. Thus higher mass accuracy improves sensitivity and selectivity for a search and help discriminate true matches from random matches.

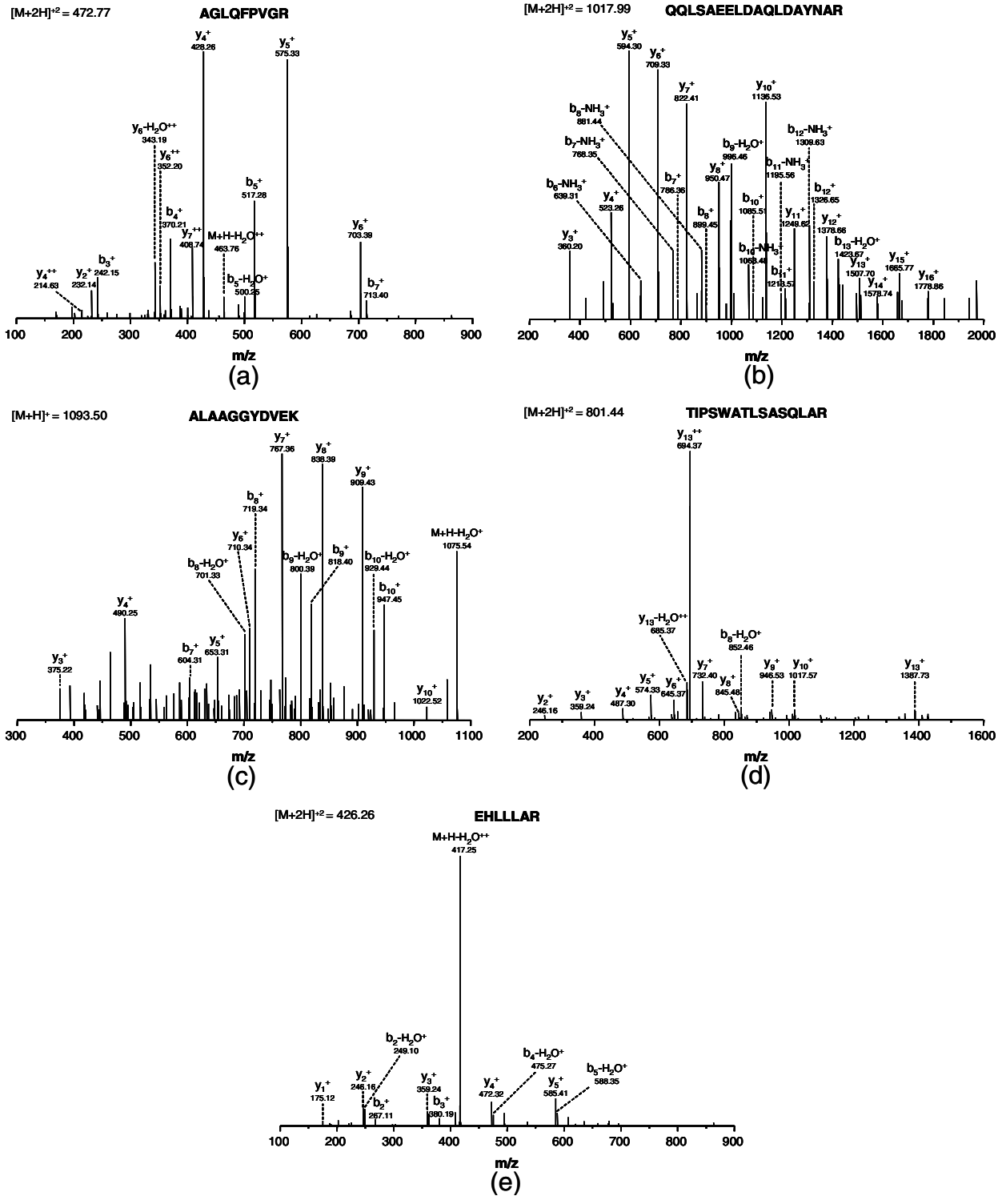


**Figure 1**

Examples of spectral influences on scoring. High quality data for peptides of short (a) and long (b) lengths. Poorer quality data from a low signal-to-noise spectrum (c), a spectrum from a peptide with a single dominant product ion due to fragmentation at the N-terminal side of Pro (d) and a spectrum of a peptide with a single dominant product ion due to a extensive neutral loss of water at the N-terminal Glu (e) [30].

**Table 1: Empirical and statistical scores along with associated parameters (*p*1 and *p*2) for each spectrum shown in Figure 1. The data were obtained for mass accuracies of 1.0 Da, 0.1 Da and 0.01 Da. Confidence levels were calculated based on a search space of 2726345 theoretical peptides. The confidence levels for the pp and pp2 scores are denotes as confidence level and confidence level2**

| Mass accuracy | Spectrum | *p*1 | *p*2 | score | pp | pp2 | Confidence level | Confidence level 2 |
|---|---|---|---|---|---|---|---|---|
| 0.01 Da | a | $2.1. \times 10^{-5}$ | $7.9. \times 10^{-4}$ | 69 | 53.8 | 307.6 | 52.0 | 305.8 |
| | b | $2.1. \times 10^{-5}$ | $2.1. \times 10^{-3}$ | 75 | 69.0 | 307.6 | 67.2 | 305.8 |
| | c | $2.1. \times 10^{-5}$ | $9.2. \times 10^{-4}$ | 59 | 43.0 | 307.6 | 41.2 | 305.8 |
| | d | $2.1. \times 10^{-5}$ | $1.5. \times 10^{-3}$ | 19 | 75.2 | 307.6 | 73.4 | 305.8 |
| | e | $2.1. \times 10^{-5}$ | $6.5. \times 10^{-4}$ | 27 | 36.6 | 307.6 | 34.8 | 305.8 |
| 0.1 Da | a | $2.1. \times 10^{-4}$ | $7.8. \times 10^{-3}$ | 77 | 38.6 | 203.9 | 35.8 | 201.1 |
| | b | $2.1. \times 10^{-4}$ | $2.1. \times 10^{-2}$ | 95 | 46.7 | 157.5 | 43.9 | 154.7 |
| | c | $2.1. \times 10^{-4}$ | $9.1. \times 10^{-3}$ | 68 | 26.9 | 274.7 | 24.1 | 271.9 |
| | d | $2.1. \times 10^{-4}$ | $1.5. \times 10^{-2}$ | 20 | 45.2 | 37.3 | 42.4 | 34.5 |
| | e | $2.1. \times 10^{-4}$ | $6.5. \times 10^{-3}$ | 28 | 23.7 | 82.8 | 20.9 | 80.0 |
| 1.0 Da | a | $2.1. \times 10^{-3}$ | $7.8. \times 10^{-2}$ | 100 | 22.1 | 21.3 | 18.3 | 17.5 |
| | b | $2.1. \times 10^{-3}$ | $2.1. \times 10^{-1}$ | 120 | 16.0 | 12.1 | 12.2 | 8.3 |
| | c | $2.1. \times 10^{-3}$ | $9.1. \times 10^{-2}$ | 74 | 7.8 | 23.1 | 4.0 | 19.3 |
| | d | $2.1. \times 10^{-3}$ | $1.5. \times 10^{-1}$ | 55 | 15.5 | 6.1 | 11.7 | 2.3 |
| | e | $2.1. \times 10^{-3}$ | $6.4. \times 10^{-2}$ | 36 | 14.7 | 9.0 | 10.9 | 5.2 |

## Conclusion

A new statistically derived scoring algorithm was developed for characterization of peptides, proteins and their posttranslational modifications from tandem MS data. The probability based algorithm implicitly incorporates mass accuracy into scoring the potential peptide and protein matches. This approach is separate and distinct from algorithms that filter precursor and product ion matches based on mass accuracy. The statistical model involves no empirical parameters and its scores correlate to the probability that a match is a random occurrence. A novel statistically derived algorithm to rigorously calculate protein scores from the probability based peptide scores was also developed. Thus the protein scores reflect the significance of protein matches and can be used to differentiate true protein matches from random matches. The algorithm is incorporated in an automated database search program MassMatrix.

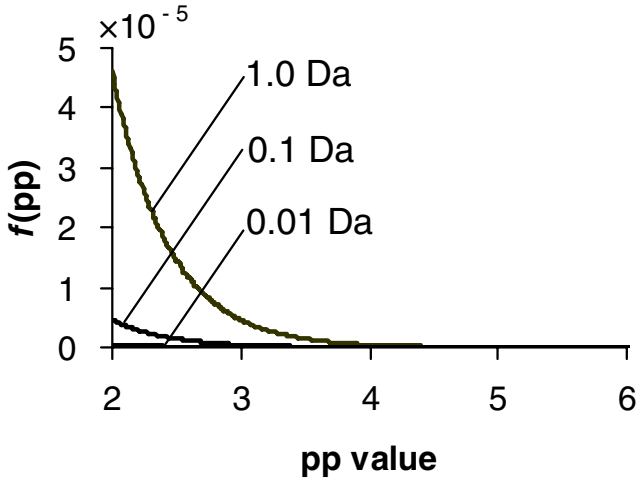


**Figure 2**
Effect of mass accuracy on the theoretical probability density function (*f*(pp)) of pp and pp2 values for random matches ($\beta \geq \beta_c$).

## Authors' contributions

HX designed and mathematically proved the statistical model and drafted the manuscript. MAF was the principle investigator and provided overall guidance of the project, and also revised the manuscript critically. Both authors read and approved the final manuscript.

## Acknowledgements

The study was funded by the Ohio State University, the National Institutes of Health CA107106, the V Foundation/American Association for Cancer Research Translational Cancer Research Grant and the Leukemia & Lymphoma Society.

## References

1. Sadygov RG, Cociorva DC, Yates JR: **Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book.** *Nature Methods* 2004, **1(3):**195-202.
2. Hunt DF, Yates JR, Shabanowitz J, Winston S, Hauer CR: **Protein sequencing by tandem mass-spectrometry.** *Proc Natl Acad Sci U S A* 1986, **57:**6233-6237.
3. Bakhtiar R, Guan ZQ: **Electron capture dissociation mass spectrometry in characterization of peptides and proteins.** *Biotechnol Lett* 2006, **28(14):**1047-1059.
4. Nikolaev EN, Somogyi A, Smith DL, Gu CG, Wysocki VH, Martin CD, Samuelson GL: **Implementation of low-energy surface-induced**

**dissociation (eV SID) and high-energy collision-induced dissociation (keV CID) in a linear sector-TOF hybrid tandem mass spectrometer.** *Int J Mass Spectrom* 2001, **212:**535-551.
5. Syka JEP, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF: **Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.** *Natl Acad Sci USA* 2004, **101:**9528-9533.
6. Hunt DF, Buko AM, Ballard JM, Shabanowitz J, Giordani AB: **Sequence-analysis of polypeptides by collision activated dissociation on a triple quadrupole mass-spectrometer.** *Biomed Mass Spectrom* 1981, **53:**397-408.
7. Biemann K: **Contributions of mass-spectrometry to peptide and protein-structure.** *Biomed Environ Mass Spectrom* 1988, **16:**99-111.
8. Nesvizhskii AI, Aebersold R: **Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS.** *Drug Discov Today* 2004, **9(4):**173-181.
9. Dancík V, Addona TA, Clauser KR, Vath JE, Pevzner PA: **De novo peptide sequencing via tandem mass spectrometry.** *J Comput Biol* 1999, **6:**327-342.
10. Standing KG: **Peptide and protein de novo sequencing by mass spectrometry.** *Curr Opin Struc Biol* 2003, **13(5):**595-601.
11. Eng JK, McCormack AL, Yates JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.** *J Am Soc Mass Spectrom* 1994, **5:**976-989.
12. Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, Miller CA, Fenyo D, Eng JK, Adkins JN, Omenn GS, Simpson RJ: **An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis.** *Proteomics* 2005, **5(13):**3475-3490.
13. Keller A, Nesvizhskii A, Kolker E, Aebersold R: **Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.** *Anal Chem* 2002, **74:**5383-5392.
14. Qian W, Liu T, Monroe ME, Strittmatter EF, Jacobs JM, Kangas LJ, Petritis K, Camp DG, Smith RD: **Probability-Based Evaluation of Peptide and Protein Identifications from Tandem Mass Spectrometry and SEQUEST Analysis: The Human Proteome.** *J Proteome Res* 2005, **4:**53-62.
15. Zhang N, Aebersold R, Schwikowshi B: **ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data.** *Proteomics* 2002, **2:**1406-1412.
16. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS: **Probability-based protein identification by searching sequence database using mass spectrometry data.** *Electrophoresis* 1999, **20:**3551-3567.
17. Hansen BT, Jones JA, Mason DE, Liebler DC: **SALSA: a pattern recognition algorithm to detect electrophile-adducted peptides by automated evaluation of CID spectra in LC-MS-MS analyses.** *Anal Chem* 2001, **73:**1676-1683.
18. Mann M, Wilm M: **Error tolerant identification of peptides in sequence databases by peptide sequence tags.** *Anal Chem* 1994, **66:**4390-4399.
19. Bafna V, Edwards N: **SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database.** *Bioinformatics* 2001, **17:**S13-S21.
20. Colinge J, Masselot A, Giron M, Dessingy T, Magnin J: **OLAV: towards high-throughput tandem mass spectrometry data identification.** *Proteomics* 2003, **3:**1454-1463.
21. MacCoss MJ, Wu CC, Yates JR: **Probability-based validation of protein identifications using a modified SEQUEST algorithm.** *Anal Chem* 2002, **74:**5593-5599.
22. Havilio M, Haddad Y, Smilansky Z: **Intensity-based statistical scorer for tandem mass spectrometry.** *Anal Chem* 2003, **75:**435-444.
23. Sadygov RG, Yates JR: **A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases.** *Anal Chem* 2003, **75:**3792-3798.
24. Sadygov RG, Liu H, Yates JR: **Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases.** *Anal Chem* 2004, **76:**1664-1671.
25. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm.** *J Proteome Res* 2004, **3:**958-964.
26. Olsen JV, de Godoy LMF, Li G, Macek B, Mortensen P, Pesch R, Makarov A, Lange O, Horning S, Mann M: **Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap.** *Mol Cell Proteomics* 2005, **4(12):**2010-2021.
27. Clauser KR, Baker P, Burlingame AL: **Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching.** *Anal Chem* 1999, **71:**2871-2882.
28. Sullivan AG, Brancia FL, Tyldesley R, Bateman R, Sidhu K, Hubbard SJ, Oliver SG, Gaskell SJ: **The exploitation of selective cleavage of singly protonated peptide ions adjacent to aspartic acid residues using a quadrupole orthogonal time-of-flight mass spectrometer equipped with a matrix-assisted laser desorption/ionization source.** *Int J Mass Spectrom* 2001, **210/211:**665-676.
29. Xu H, Freitas AF: **MassMatrix: A Database Searching Program for Rapid Characterization of Proteins and Peptides from Tandem Mass Spectrometry Data.** *To be submitted* .
30. Paizs B, Suhai S: **Fragmentation pathways of protonated peptides.** *Mass Spectrom Rev* 2005, **24:**508-548.