Published in final edited form as:
Proceedings (IEEE Int Conf Bioinformatics Biomed). 2012 October 4; : 648–653. doi:10.1109/BIBMW.2012.6470214.

A statistical approach to peptide identification from clustered tandem mass spectrometry data

Soyoung Ryu[*], **David R. Goodlett**[†], **William S. Noble**[‡], and **Vladimir N. Minin**[§]

[*]Department of Statistics, University of Washington, Seattle, WA, USA, claireryu@gmail.com

[†]Department of Medicinal Chemistry, University of Washington, Seattle, WA, USA, goodlett@u.washington.edu

[‡]Department of Genome Sciences and Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA, william-noble@uw.edu

[§]Department of Statistics, University of Washington, Seattle, WA, USA

Abstract

Tandem mass spectrometry experiments generate from thousands to millions of spectra. These spectra can be used to identify the presence of proteins in biological samples. In this work, we propose a new method to identify peptides, substrings of proteins, based on clustered tandem mass spectrometry data. In contrast to previously proposed approaches, which identify one representative spectrum for each cluster using traditional database searching algorithms, our method uses all available information to score all the spectra in a cluster against candidate peptides using Bayesian model selection. We illustrate the performance of our method by applying it to seven-standard-protein mixture data.

Keywords

Bayesian analysis; Bioinformatics; Clustered tandem mass spectra; False discovery rate; Peptide identification; Proteomics

I. Introduction

The field of proteomics attempts to study all expressed proteins in an organism (i.e. the proteome) rather than single proteins as was traditionally practiced [1]. Although understanding the roles and interactions of all proteins in an organism simultaneously is very challenging, the development of high performance liquid chromatography-tandem mass spectrometry (HPLC-MS/MS) methods, commonly referred to as shotgun proteomics, creates the possibility of success. Additionally, the development of bioinformatics tools to analyze the large MS/MS data sets created by these shotgun proteomic experiments, makes the following four goals achievable [1]: 1) identifying all (or many) of the proteins in a sample, 2) profiling differences in protein expression between varying conditions, 3) determining how proteins interact with each other in the living systems, and 4) identifying how and where proteins are modified. Among these four goals, we are interested in the first goal, which can be accomplished by identifying as many peptides (i.e. short stretches of amino acid polymers) in a sample as possible. Recently, several research groups proposed

Corresponding author. vminin@uw.edu.

methods to save computational database searching time in the sequence matching step by first clustering similar tandem mass spectra and then using the clustered spectra, rather than individual spectra, for peptide identification [2]–[4]. Assuming that tandem mass spectra within each cluster are repeated observations of the same peptide, one can treat the tandem mass spectra within each cluster as repeated experimental observations. Two common approaches to clustered tandem mass spectral analysis proceed by either taking a single “highest quality” representative spectrum from each cluster [3], [4] or by averaging all spectra in the cluster and finding the best matching peptide for this representative pseudo-spectrum [2], [3]. The rationale behind both of these approaches is to increase the database searching speed by eliminating redundant queries and to improve the signal-to-noise ratio. Although current methods for clustered spectral analysis can reduce the database searching time, these methods do not fully explore all information offered by clustered spectra. We propose a formal statistical approach that we refer to as $BaMS^2$ (Bayesian Model Selection for tandem Mass Spectra) that uses repeated experimental observations, afforded by clustered tandem mass spectra, more efficiently.

Taking a statistical point of view, we treat peptide identification as a model selection problem assuming that individual tandem mass spectra within a cluster represent repeated fragmentations of the same peptide. As demonstrated in Figure 1, a tandem mass spectrum, obtained together with an associated precursor ion peptide mass-to-charge (m/z) value, contains both m/z and intensity values of the selected precursor ion’s fragment ions. Given clustered observed tandem mass spectra, we may have hundreds to thousands of peptide candidates selected based on the precursor ion peptide mass. Assuming each candidate peptide defines a model for generating multiple observed tandem mass spectra in a cluster, we formulate a probabilistic data generating model. Next, we follow the Bayesian paradigm and measure relative model fit by the probability of each candidate model/peptide given the observed clustered tandem mass spectra. This procedure allows us to capture more data features than merely the mean intensities of a cluster. The focus of the paper is not on clustering algorithms for tandem mass spectra, but rather on novel peptide identification methods based on already clustered spectra. We first cluster tandem mass spectra using the algorithm of Frank *et al.* [3]. We then use our clustered tandem mass spectra-based database searching algorithm to assign the top-ranked peptide sequence for each cluster. Then, we differentiate correct peptide identifications from the incorrect ones using a false discovery rate (FDR) procedure coupled with a decoy database analysis [5], [6]. The FDR procedure is needed because the top-ranked peptide sequence for a cluster may not be a correct assignment. To validate our approach, we apply our method to a standard mixture of seven proteins. We show that $BaMS^2$ identifies more peptides than the competing method [3] combined with traditional peptide identification algorithms, such as Sequest [7]. The advantage of our method becomes even more apparent as the cluster size increases. We believe that our model-based approach shows great promise, because in contrast to model-free approaches, our model can be further improved more easily and refined to yield better performance.

II. Method

A. Data

To test our method, we used thirty HPLC-MS/MS data sets acquired on mixtures of known proteins [8]. There were six samples that contained seven proteins with varying concentrations. From these experiments, 7,478 clusters with a size greater than two (spectra) were used for our analysis. About 15% of the clusters (Standard dataset A) were used for the model tuning, while the rest of clusters (Standard dataset B) were used for the model evaluation and model comparison with a Frank *et al.*’s approach [3]. For the protein database, we attached a shuffled yeast database and the contaminant database (human keratins, porcine trypsin, bonvine trypsin) to the seven-standard-protein database and treated this combined database as a target database.

B. Preprocessing

Before modeling tandem mass spectrometry data, performing adequate preprocessing steps is essential [9]. This is necessary because inadequate preprocessing may cause mild to serious model violations as well as low performance in assigning the correct peptide sequence to a cluster. Our preprocessing steps included clustering tandem mass spectra, binning m/z values of peaks, and rescaling peak intensities in each spectrum. Recently developed clustering algorithms for tandem mass spectra have only a few differences among them [2]–[4]. Encouraged by results of [3], who synthesized two previously published methods, we used their software, MS-Clustering (available from http://peptide.ucsd.edu) to cluster tandem mass spectra. MS-Clustering starts by removing low-quality spectra, which are not likely to yield peptide identifications. For convenience in model building, we considered only a certain range of m/z values (200–2,000Da) [9] and removed peaks within a 10Da window around the precursor mass-to-charge value to eliminate the possibility of matching a predicted fragment ion to the mass-to-charge ratio of the precursor ion [7]. Then, we discretized spectrum i by dividing the m/z values into B bins and collecting the total intensities per bin into a vector $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{iB})$. The bin width is related to the resolution of the instrument. Using the fact that fragment mass tolerance of the LTQ is about 0.8 [10], we set 1 Da as a bin width. In our model formulation, we assumed that log-intensities of bin j in spectra within the same cluster were generated from a normal distribution with a common mean intensity value. However, peak intensities of different spectra were not on the same scale due to the difference in their precursor ion intensities. We therefore normalized each spectrum i by subtracting an arithmetic mean of log-transformed non-zero intensities in spectrum i from all log-transformed non-zero intensities in spectrum i.

C. Bayesian Peptide Identification

1) Probabilistic Data Generating Model—We formulated a probabilistic data generating model given a theoretical tandem mass spectrum. In our theoretical spectrum, we included the first and second isotope peaks of predominant ions, which are called b- and y-ions. Our method can be easily extended to include more types of ions (i.e. z-, c-, and a-ions [11]) by simply adding more ions in the theoretical spectrum. Since intensity values of the theoretical spectrum were unknown, our theoretical spectrum was represented by a binary vector $\mathbf{T} = (T_1, \dots, T_B)$, where $T_j = 0$ indicates the absence of a signal peak in bin j and $T_j = 1$ indicates that we expect to see a signal peak in this bin. Given these indicators of signal peak absence/presence, we modeled the observed intensities of signal and noise peaks. When $T_j = 1$, we observe a signal peak in bin j with a probability p and log-intensities of signal peaks are normally distributed with mean μ_j and variance σ_j^2. When $T_j = 0$, we observe a noise peak with a probability q and log-intensities of noise peaks are normally distributed with a mean ν and a variance γ^2. Note that the signal model has bin-specific means and variances while the noise model has a common mean and a common variance. We also point out that we transformed observed intensities to the log scale. This choice of transformation was dictated by mathematical convenience rather than by physical modeling. The logarithmic transformation itself allowed us to use normal distribution for noise peak intensities or bin-specific signal peak intensities, and we empirically confirmed that a normal distribution is appropriate for the log-transformed clustered mass spectrometry data observed in practice. In order to obtain the likelihood of the observed clustered spectra, we made two assumptions. The first assumption was that observations corresponding to the same bin are independent and identically distributed (iid) across spectra within a cluster. In this assumption, we considered the intensities of common peaks from the same peptide (with the same charge state) were generated from the same distribution. This assumption is reasonable when our rescaling preprocessing step is performed properly. Furthermore, we assumed that intensities across bins are independently distributed. Thus, the peak intensity in one bin is not associated with the peak intensity in another bin. This assumption is disputable due to dependences among isotopic peaks, between b- and y-ions, among the same ions with different charge states [12], [13], and due to rescaling in our preprocessing step. However, we prefer to build a simple and computationally efficient model with the hope that our crude approximation of reality will be sufficient for peptide identification. Under our two assumptions, the likelihood of observed clustered spectra can be written as

$$\Pr(\mathbf{X}|\mathbf{T},p,q,\mu,\sigma,\nu,\gamma)=\prod_{j=1}^{B}\prod_{i=1}^{N}\left[(1-p)^{1_{[x_{ij}=0]}}\left\{\frac{p}{\sigma_j}\phi\left(\frac{\log(x_{ij})-\mu_j}{\sigma_j}\right)\right\}^{1_{[x_{ij}>0]}}\right]^{T_j}\left[(1-q)^{1_{[x_{ij}=0]}}\left\{\frac{q}{\gamma}\phi\left(\frac{\log(x_{ij})-\nu}{\gamma}\right)\right\}^{1_{[x_{ij}>0]}}\right]^{1-T_j}, \quad (1)$$

where N is the number of spectra in a cluster, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N)$ is a matrix of observed peak intensities for a given cluster, $(p, q, \boldsymbol{\mu}, \boldsymbol{\sigma}, \nu, \gamma)$ are model parameters, and $\phi(x)$ is the standard normal density. The clusters with a minimum cluster size of 3 ($N \geq 3$) were used for this model.

2) Bayesian Model Selection Method—Now, suppose we want to compare K peptides/models, $\mathbf{T}_1, \dots, \mathbf{T}_K$, using the same observed spectral cluster $\mathbf{X}$. Taking a statistical point of view, we need to measure relative fit of these models to the observed data. Notice that such a comparison is complicated by the fact that model parameter vectors, $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$, may have different numbers of components, where $\boldsymbol{\theta} = (p, q, \boldsymbol{\mu}, \boldsymbol{\sigma}, \nu, \gamma)$. There are several approaches to this model selection problem. We took one of them, which is based on integrated likelihood [14], [15]. First, we assigned conjugate prior distributions to model parameters, $\Pr(\boldsymbol{\theta}_m \mid \mathbf{T}_m)$, which is a choice based on mathematical convenience. The prior distributions for signal bin parameters are:

$$\begin{aligned}&p \sim \mathrm{Beta}(a_p, b_p),\\&\mu_j|\sigma_j^2 \sim \mathrm{Normal}(\mu_{j0}, \sigma_j^2/\kappa_{j1}),\\&\sigma_j^2 \sim \mathrm{Inverse-Gamma}(\alpha_{j1}, \beta_{j1}).\end{aligned} \quad (2)$$

We can interpret prior parameters in terms of prior observations if we look at the posterior expectation as a weighted average of the prior and empirical expectations [16]. We interpret a_p as a prior observed number of non-zero signal bins, where $a_p + b_p$ is a prior sample size of all signal bins. We think of μ_{j0} as the prior sample mean of κ_{j1} non-zero signal bins with variance σ_j^2. The interpretation of β_{j1}/α_{j1} is a prior sample variance of non-zero signal intensities in bin j, where the prior sample size is $2\alpha_{j1}$. Parameters of prior noise bin distributions listed below in (3.3), can be interpreted similarly to prior parameters for signal bins except that ν_0 and γ^2/κ_0 are not bin-specific:

$$\begin{aligned}&q \sim \mathrm{Beta}(a_q, b_q),\\&\nu|\gamma^2 \sim \mathrm{Normal}(\nu_0, \gamma^2/\kappa_0),\\&\gamma^2 \sim \mathrm{Inverse-Gamma}(\alpha_0, \beta_0).\end{aligned} \quad (3)$$

After the priors were determined, we proceeded by calculating the marginal likelihood for each model T_m:

$$\Pr(\mathbf{X}|\mathbf{T}_m)=\int \Pr(\mathbf{X}|\mathbf{T}_m,\theta_m)\Pr(\theta_m|\mathbf{T}_m)d\theta_m. \quad (4)$$

After obtaining integrated likelihoods for all candidate models, using Bayes' rule, we arrived at the posterior model probabilities:

$$\Pr(\mathbf{T}_m|\mathbf{X})=\frac{\Pr(\mathbf{X}|\mathbf{T}_m)\Pr(\mathbf{T}_m)}{\sum_{l=1}^{K}\Pr(\mathbf{X}|\mathbf{T}_l)\Pr(\mathbf{T}_l)}. \quad (5)$$

Our algorithm selects a model/peptide with the highest posterior probability. In this analysis, we assumed that $\Pr(\mathbf{T}_i) = 1/K$ for all $i = 1, \ldots, K$, which means that we select a model/peptide with the highest integrated likelihood. The detectibility score [17] is another reasonable choice for $\Pr(\mathbf{T}_i)$ since it is the probability of observing a peptide T_i in a sample.

D. Measuring Uncertainty of Peptide Identification

The predictive ability of scores used for peptide identification determines the performance in measuring uncertainty of peptide/protein identification [18], [19], thus we need good predictor scores that can distinguish correct from incorrect identifications. Since the cluster size varies by cluster, the integrated likelihoods are not comparable across clusters, preventing us from using them as a predictive score. The potential scores we considered were the posterior probability that the observed cluster is generated from a certain peptide/model, the difference between the top two log integrated likelihoods normalized by the standard deviation of log integrated likelihoods of all candidate peptides (DeltaLIL), the noise emission probability $\hat{q}$, the signal emission probability $\hat{p}$, the average intensity of noise peaks $\hat{\nu}$, and the average of $\hat{\mu}_j$'s which is the average intensity of signal peaks. We compared the performance of these scores in differentiating the correct identifications from incorrect identifications using precision-recall [20], [21] and receiver operating characteristic (ROC) curves [22], shown in Figure 2. Seven standard protein mixture data (Standard dataset A) were used for this comparison since the correct and incorrect identifications can be approximated as described in the section II.C. In Figure 2, the ROC curve closer to the upper-left comer indicates better performance while the precision-recall curve closer to the upper-right comer indicates better performance. Since our data contain a much smaller number of positives (correct identification) compared to the number of negatives (incorrect identifications), the precision-recall curves highlight the important differences among scores better than the ROC curves [20]. Based on this analysis, we decided to use DeltaLIL as the predictive score, which performed the best among the potential scores. Using DeltaLIL, we computed a p-value to test the null hypothesis that a match between the observed cluster and the top-ranked peptide occurred by chance. Since such a null distribution is unknown for real biological samples, we approximated the score distribution of incorrect peptide identifications using a decoy database approach (e.g. a database that contains reversed sequences of proteins or a shuffled database) [5], [18], [23]. The derived p-value was the proportion of decoy identifications equal to or larger than the score of a given identification. Our computed p-value distribution was most dense near zero and flat elsewhere. This shape displayed no violations of the assumptions of methods operating on p-values and suggested that several clusters were correctly identified [24]. Finally, we corrected the multiple testing problem by computing the q-value, which is the minimal positive false discovery rate (PFDR) for each identification [6]. The database search time for our Bayesian approach is O(BKC) where B is the number of bins, K is the number of candidate peptides, and C is the number of clusters in a dataset. Since the number of clusters, C is much less than S, the number of spectra in a particular dataset, the time complexity of this algorithm is much less than the time complexity of traditional database searching algorithms, O(BKS). Comparing to traditional database searching algorithms, we preprocess the spectra (i.e. clustering, normalizing, extracting sufficient statisticis), however, the time complexity for this step is only O(BS).

III. RESULTS AND DISCUSSION

On the seven-protein standard mixture data (Standard dataset B), our *BaMS*^{2} method outperformed the model-free approach in terms of the number of clusters identified with high confidence. The number of clusters identified by our approach was consistently larger than the number of clusters identified by the competing approach for q-value cutoffs ∈ (0, 0.1) as shown in Figure 3(a). For q-value< 0.01, *BaMS*^{2} identified 20.80% more clusters than the model-free approach (Table I). Moreover, our *BaMS*^{2} approach identified 18.07% more peptides and 22.06% more spectra than the model-free method with the same q-value cutoff. The number of identified spectra was computed by multiplying the number of identified clusters by their cluster size under the assumption that all spectra from the same cluster are from the same peptide. Better *BaMS*^{2} performance in identifying spectra rather than clusters (compared to the model-free approach) suggests that our approach performs better for larger clusters. We examined the performance of both methods for various cluster sizes in Figure 3(b). Both methods did better as the number of spectra per cluster increased. More importantly, as the cluster size increased, the advantage of our method over the model-free approach became more pronounced. The latter pattern emerged because more spectra in a cluster enabled the *BaMS*^{2} algorithm to use information more efficiently. Surprisingly, our *BaMS*^{2} did well even for small clusters in the standard mixture data. Good performance of the *BaMS*^{2} algorithm on modestly sized clusters (<30 spectra per cluster) may be due to our choice of prior parameters which happened to work well for this particular data set.

Figure 3(a) plots the numbers of clusters identified against q-value cutoffs. Here, we used two types of q-values, the predicted q-values [6] and estimated q-values. The latter q-values were approximated based on the information of proteins present in the samples and the shuffled yeast database. More specifically, for each predictive score (DeltaLIL for *BaMS*^{2} and DeltaCn for the model-free approach) threshold, we estimated pFDR by dividing the number of identified clusters as a peptide neither from one of seven standard proteins in the sample nor from one of contaminants by the number of clusters with the predictive score exceeding the specified threshold. Then, we computed the q-values which are the minimum pFDRs at which a given predictive score is accepted [6]. Figure 3(a) shows that both [6]'s and estimated q-values are very similar except for very small q-value thresholds (q-value< 0.01).

IV. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a novel database searching algorithm, *BaMS*^{2}, for clustered tandem mass spectra. By eliminating redundant entries for database search, this clustering based algorithm has lower time complexity than the traditional algorithms and can be implemented to reduce the actual database searching time. Its performance is good for larger clusters in terms of identifying more peptides with high confidence. For small clusters, our approach performs similarly to the model-free approach, but the performance of our *BaMS*^{2} method seems to increase faster than the model-free method as the cluster size increases. We also think that our *BaMS*^{2} has the potential to improve its performance since our proposed method is flexible. For example, we can incorporate better prior knowledge into our model. In our analysis, we assigned the same weak informative prior for all the signal bins. However, in reality, the intensity of a signal peak depends on the fragment behavior of peptides which can be predicted by the amino acid composition of the peptide. Thus, it is more appropriate to assign a prior mean intensity separately for each signal bin. Combining the efficiency of our method with a better choice of priors can improve the *BaMS*^{2} performance, especially for small clusters. We can also improve our method by making our data generating model more realistic so that it accounts for dependencies among signal bins (i.e. b- and y-ions, the first and second isotope peaks).

Our *BaMS*^{2} algorithm takes advantage of repeated spectra from the same peptide and works well with large clusters (30 or more). Our approach is useful in the situation where multiple tandem mass spectra from the same peptides are produced. For example, researchers can take advantage of our method for precursor acquisition independent from ion count (PAcIFIC) [25] which produces many more tandem mass spectra from the same precursor for one run than regular mass spectrometry experiments. In general, our method is timely, especially in light of ongoing developments of the mass spectrometry experiments that produce a large amount of often redundant information. We believe that model-based statistics methods provide a natural framework for leveraging this redundancy.

Acknowledgments

This research was supported by the National Institues of Health grant [5R33CA099139-03 to D.R.G., 1RC4AI092828-01 to V.N.M., P41 RR0011823 to W.S.N].

References

1. Liebler, DC. Introduction to Proteomics: Tools for the New Biology. Humana Press; 2002.
2. Beer I, Barnea E, Ziv T, Admon A. Improving large-scale proteomics by clustering of mass spectrometry data. Proteomics. 2004; vol. 4:950–960. [PubMed: 15048977]
3. Frank AM, Bandeira N, Shen Z, Tanner S, Briggs SP, Smith RD, Pevzner PA. Clustering millions of tandem mass spectra. Journal of Proteome Research. 2007; vol. 7(no. 1):113–122. [PubMed: 18067247]
4. Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR. Similarity among tandem mass spectra from proteomic experiments: Detection, significance, and utility. Analytical Chemistry. 2003; vol. 75(no. 10):2470–2477. [PubMed: 12918992]
5. Käll L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. Journal of Proteome Research. 2007; vol. 7(no. 1):29–34. [PubMed: 18067246]
6. Storey JD. A direct approach to false discovery rates. Journal Of The Royal Statistical Society, Series B. 2002; vol. 64(no. 3):479–498.
7. Eng J, McCormack A, Yates JI. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry. 1994; vol. 5:976–989.
8. Ryu S, Gallis B, Goo YA, Shaffer SA, Radulovic D, Goodlett DR. Comparison of a label-free quantitative proteomic method based on peptide ion current area to the isotope coded affinity tag method. Cancer Informatics. 2008
9. Li, Q. Ph.D. Dissertation. University of Washington; 2008. Statistical methods for peptide and protein identification using mass spectrometry.
10. Xu H. MassMatrix database search engine: search form help. 2009 Nov. http://www.xumatrix.comlcgi-bin/mm-cgi/home.py.
11. Roepstorff P, Fohlman J. Letter to the editors. Biological Mass Spectrometry. 1984; vol. 11(no. 11):601–601.
12. Wan Y, Yang A, Chen T. PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. Analytical Chemistry. 2005; vol. 78(no. 2):432–437.
13. Klammer AA, Reynolds SM, Bilmes JA, MacCoss MJ, Noble WS. Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. Bioinformatics. 2008; vol. 24(no. 13):i348–i356. [PubMed: 18586734]
14. Leonard, T.; Hsu, JS. Bayesian methods: an analysis for statisticians and interdisciplinary researchers. Cambridge University Press; 1999.
15. Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association. 1995; vol. 90:773–795.
16. Hoff, PD. A First Course in Bayesian Statistical Methods. Springer; 2009.

17. Tang H, Arnold RJ, Alves P, Xun Z, Clemmer DE, Novotny MV, Reilly JP, Radivojac P. A computational approach toward label-free protein quantification using predicted peptide detectability. Bio informatics. 2006; vol. 22(no. 14):e481–e488.

18. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nature Methods. 2007; vol. 4(no. 11):923–925. [PubMed: 17952086]

19. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Analytical Chemistry. 2002; vol. 74:5383–5392. [PubMed: 12403597]

20. Davis, J.; Goadrich, M. The relationship between precision-recall and ROC curves; ICML ’06: Proceedings of the 23rd international conference on Machine learning; 2006. p. 233-240.Association for Computing Machinery

21. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; vol. 21(no. 20):3940–3941. [PubMed: 16096348]

22. Provost, FJ.; Fawcett, T.; Kohavi, R. The case against accuracy estimation for comparing induction algorithms; ICML ’98: Proceedings of the Fifteenth International Conference on Machine Learning; 1998. p. 445-453.

23. Choi H, Nesvizhskii AI. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. Journal of Proteome Research. 2007; vol. 7(no. 1):254–265. [PubMed: 18159924]

24. Pounds SB. Estimation and control of multiple testing error rates for microarray studies. Briefings in Bioinformatics. 2006; vol. 7(no. 1):25–36. [PubMed: 16761362]

25. Panchaud A, Schcrl A, Shaffer SA, von Haller PD, Kulasekara HD, Miller SI, Goodlett DR. Precursor acquisition independent from ion count: How to dive deeper into the proteomics ocean. Analytical Chemistry. 2009; vol. 81(no. 15):6481–6488. pMID: 19572557. [PubMed: 19572557]

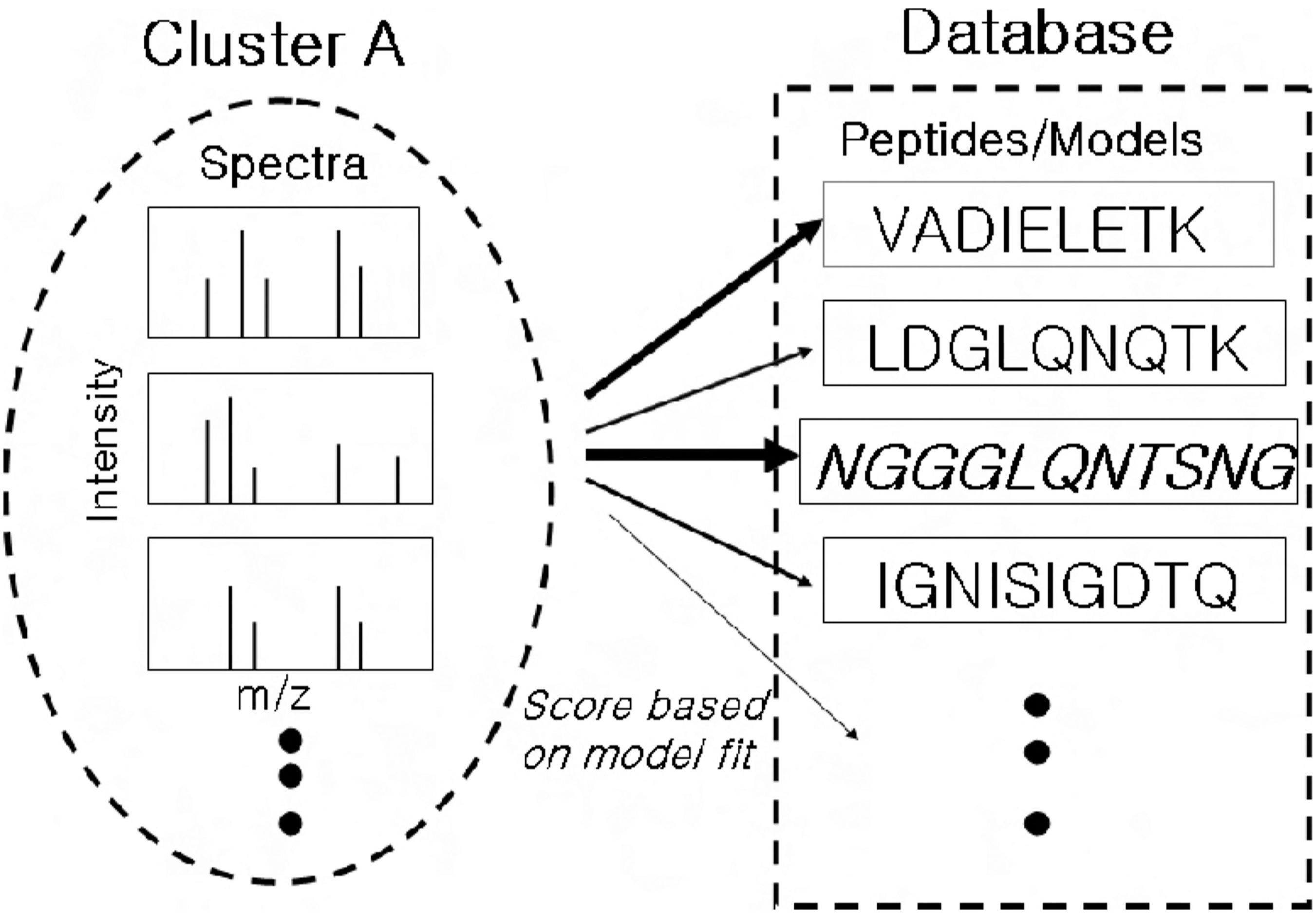
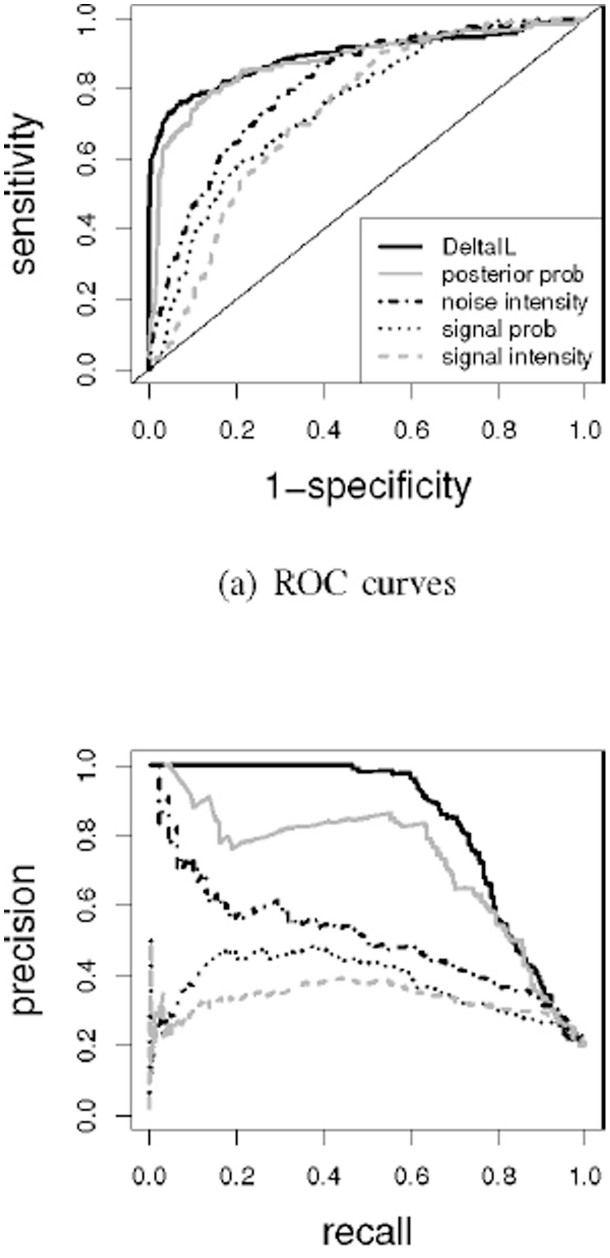


Figure 1.
Database searching example. Cluster A contains several observed spectra (on the left). Candidate peptide sequences, stored in a database, are shown on the right. The cluster is scored via the Bayesian model selection method against all candidate peptides in the database. Higher scores are shown with thicker arrows and the highest scoring peptide is depicted in bold italic.

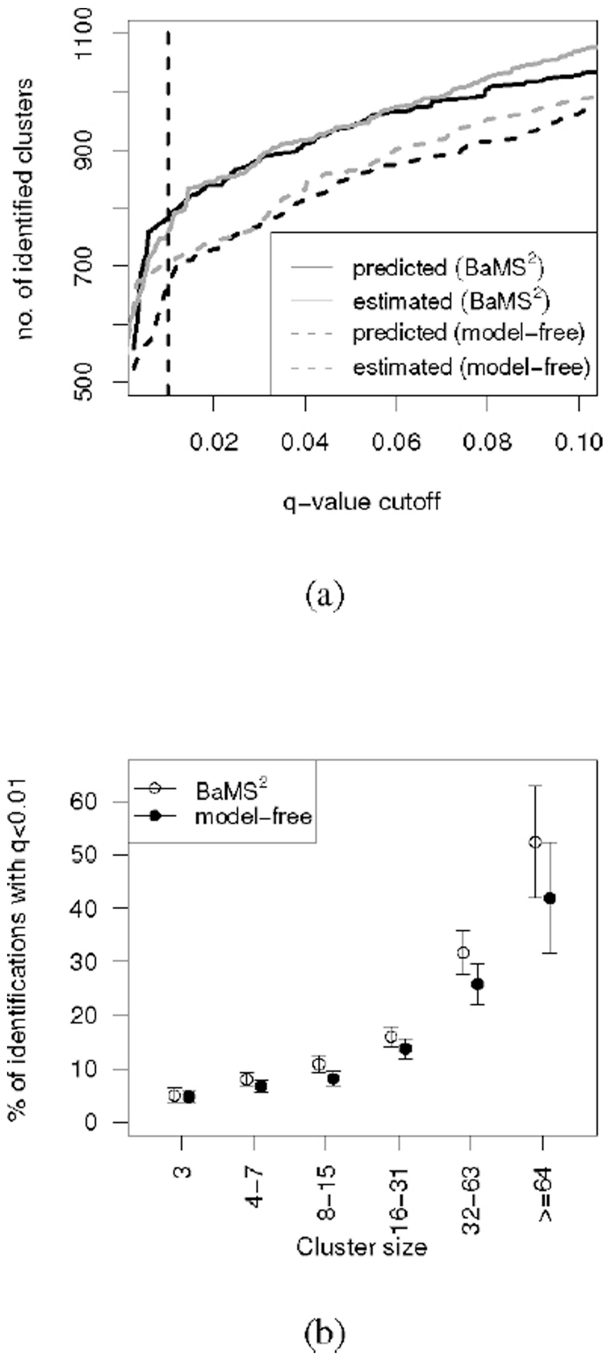


(a) ROC curves

(b) Precision-reall curves

Figure 2.

The ROC and precision-recall curves. The curves are estimated based on the correctly identified clusters of the seven-protein mixture data. The line descriptions in (b) are same as (a).



(a)

(b)

Figure 3.

Seven-protein standard mixture analysis. (a) The number of clusters identified with the corresponding q-value cutoffs. The black curves in this figure are the numbers of identified clusters given predicted q-value thresholds while the gray curves are the numbers of clusters given estimated q-value thresholds. The vertical dotted line marks the q-value threshold (= 0.01) that we used for our analysis. (b) The plot shows percentages of identifications with high confidence ($q < 0.01$) with respect to their cluster size. Open circles represent results from our *BaMS*^{2} method while filled circles show results of the model-free approach. The 95% confidence intervals are shown in the plot as vertical bars with whiskers.

Table I

PERFORMANCE IN PEPTIDE IDENTIFICATION (Q-VALUE< 0.01).

	no. of clusters	no. of spectra	no. of peptides
BaMS^2	784	18,880	647
model-free	649	15,468	548