Tsinghua Science and Technology

Volume 20 | Issue 5 Article 3

2015

Accurate Identification of Mass Peaks for Tandem Mass Spectra Using MCMC Model

Hui Li
the Department of Computer Science, Howard University, Washington, DC 20059, USA.

Chunmei Liu
the Department of Computer Science, Howard University, Washington, DC 20059, USA.

Mugizi Robert Rwebangira
the Department of Computer Science, Howard University, Washington, DC 20059, USA.

Legand Burge
the Department of Computer Science, Howard University, Washington, DC 20059, USA.

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 03/10 pp453-459
Volume 20, Number 5, October 2015

Accurate Identification of Mass Peaks for Tandem Mass Spectra Using MCMC Model

Hui Li, Chunmei Liu*, Mugizi Robert Rwebangira, and Legand Burge

Abstract: In proteomics, many methods for the identification of proteins have been developed. However, because of limited known genome sequences, noisy data, incomplete ion sequences, and the accuracy of protein identification, it is challenging to identify peptides using tandem mass spectral data. Noise filtering and removing thus play a key role in accurate peptide identification from tandem mass spectra. In this paper, we employ a Bayesian model to identify proteins based on the prior information of bond cleavages. A Markov Chain Monte Carlo (MCMC) algorithm is used to simulate candidate peptides from the posterior distribution and to estimate the parameters for the Bayesian model. Our simulation and computational experimental results show that the model can identify peptide with a higher accuracy.

Key words: mass spectrometry; Fourier transform; noise filtering; Markov Chain Monte Carlo (MCMC)

1 Introduction

Currently, peptide identification by tandem Mass Spectra (MS/MS) plays an indispensable role in proteomics research, disease diagnosis, and biomarker discovery as well as drug development. True mass peak detection is usually the first step in preprocessing tandem mass spectra, which aims to detect true mass peaks and remove noise peaks from raw MS/MS spectra. The performance of the preprocessing algorithm directly affects the accuracy of subsequent analysis such as protein identification, quantification, biomarker discovery, and classification of different samples.

Many researchers are making efforts to increase the accuracy and efficiency of mass spectral data analysis through efficiently and correctly identifying all the peaks and assigning them to the right peptides. Challenged by huge data size and rich information contained, the following problems should be addressed in peptide detection: (1) a peptide species may register several peaks in the spectrum which are commonly referred to as isotopic peaks. The isotopic distribution of a known peptide is usually determined by the elemental formula of the peptide and the natural abundance of heavy isotopes[1]; (2) the dynamic ranges of peptide signals from MS/MS data caused by the complexity of the samples have a wide range of dynamic concentration. If the resolution of mass spectra is too low to be resolved by baseline, and when these low resolution spectra are of more biological importance, they are likely to be buried under noise or interfering signals which would give rise to too many false positives[2]. For high-resolution peptides, they are comparably easier to be identified; (3) the signal density is probably high and overlapping peptide peaks are commonly observed[3]. A variety of algorithms have been proposed for peptide identification. For example, SEQUEST and Mascot[4] are popular software focusing on database searching for peptide identification. But only a small portion of peptides can be correctly identified due to spectra matching ambiguity. Other

• Hui Li, Chunmei Liu, Mugizi Robert Rwebangira, and Legand Burge are with the Department of Computer Science, Howard University, Washington, DC 20059, USA. E-mail: hli3302@gmail.com; chunmei@scs.howard.edu; rweba@scs.howard.edu; blegand@scs.howard.edu.
∗ To whom correspondence should be addressed.
Manuscript received: 2015-01-22; revised: 2015-08-07; accepted: 2015-08-10

algorithms such as PepList[5], msInspect[6], Noy's method[7], Decon2LS[8], and OpenMS[9] are template matching based methods. These methods are not perfect due to the errors which result from many sources. For instance, the problem of this kind of matching method is that it may be ineffective in detecting overlapping peptides. If the observed signal matches the proposed template, it will be reported as a feature and then will be subtracted from the spectrum. But if the peak cluster of a peptide is not correctly matched and subtracted, the rest of the peptides detected are not correct. The consequence of this matching will give error propagation. To overcome the drawbacks of these template matching based algorithms, variable selection algorithms are developed by Du and Angeletti[10]. They selected the least number of candidate isotope series to explain the spectrum and identified the corresponding peptides. Zhang et al.[11] proposed a Bayesian model to estimate the model parameters based on the observed spectrum, and calculated the existence probability of a peptide ion peak at each charge state and isotope position. Sun et al.[12] proposed a Bayesian method to detect and interpret the peptide identified from a mass spectrum. They reported that their method has better performance than other state-of-art methods.

In this paper, we propose a Bayesian model with the aim of identifying true peptides based on the observed spectrum. To obtain the true peptides, a Markov Chain Monte Carlo (MCMC) algorithm is employed to simulate candidate peptide sequences from the posterior distribution. The peptide with the largest posterior probability is estimated as the true peptide. The results of our model demonstrate the effectiveness of our approach.

2 Methods

A Bayesian approach uses prior knowledge of the peptides to help us get a good estimate of the true peptide. Due to the complexity of the posterior density in mass spectrum, we use MCMC algorithm to estimate the posterior probabilities. Furthermore, MS/MS spectral data is quite noisy, the observed spectrum thus first needs to be preprocessed. The commonly used preprocessing method including baseline subtraction, sinusoidal noise removal, and normalization is described below.

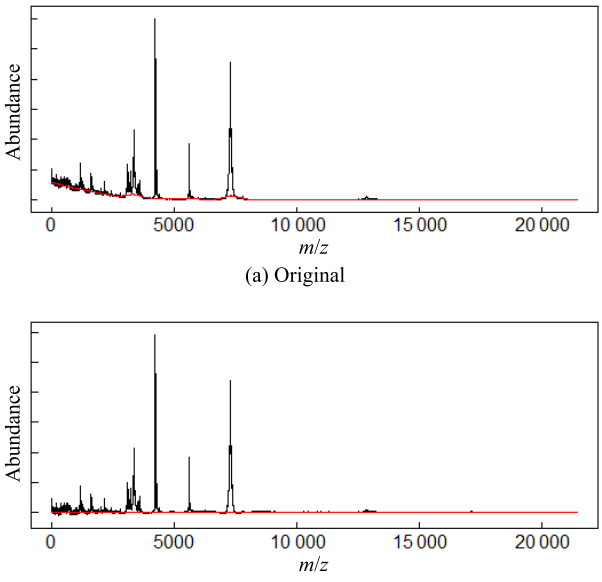
2.1 Processing

To guarantee an unbiased comparison of ion intensities, preprocessing steps including peak detection, RT alignment, peak matching, and normalization need to be handled. In order to decrease the noisy peaks in the spectrum for different peptides, we use baseline subtraction and Haar wavelet transform to differentiate noises from signal peaks. The baseline subtraction component uses an iterative algorithm to discover the best fitted curve and remove the baseline by calculating a set of estimated baseline points. The classical approach for MS/MS data processing consists of decomposing the acquired mass spectrum into three components: the true signal s, baseline drift d, and the noise σ. Consequently, each mass spectrum can be schematically modeled by Eq. (1)

$$r = s + d + \sigma \tag{1}$$

The true signal consists of a number of peaks at different m/z values. The intensities of the signal peaks have the same order of magnitude as the background noises in same cases. The effect of the baseline subtraction is shown in Fig. 1. The top panel shows an original spectrum and the bottom is the spectrum after the baseline subtraction.

The data is decomposed into multiple levels by the multi-scale wavelet decomposition. The detailed coefficient is based on the wavelet variance threshold. Then the MS/MS spectrum is reconstructed from such decomposed peaks with selected coefficient. The time domain signal contains all the characteristic frequencies of the measured



(a) Original

(b) Baseline corrected spectrum

Fig. 1 The effect of the baseline subtraction data.

ions at given intensities. Figure 2 shows the peaks are decomposed into different peaks using the Haar wavelet transformation.

We include peak information generated by biological sample, baseline from matrix background noise, and random white noise caused by MS instrument system in a normalized spectrum. The signal generated by the peptide candidate is thus modeled by the following equation:

$$y_i = \sum_{k=1}^{K} f_k(x_i) + \sum_{k=1}^{K} g_k(x_i),\ i = 1, 2, \cdots, N \quad (2)$$

where x_i is the i-th m/z in the spectrum and y_i is its corresponding output intensity. N is the length of spectrum, and K is a total number of peaks. $f(x_i)$ and $g(x_i)$ represent the peak information and baseline information of the spectrum, respectively. $g_k(x_i) \sim \mathrm{N}(0, \sigma^2)$ is the Gaussian random noise with zero-mean and standard deviation σ. $f_k(x_i)$ is the k-th peak signal. Radial basis function is used to model $f(x_i)$ and polynomial function is used for modeling $g(x_i)$.

After the noises are removed, we introduce a Bayesian model which aims at the identification of the correct peptides without depending on the database of peptides. Instead of database dependent, we make use of prior information to find the true peptide from mass spectrum data. Regarding the parameters in Bayesian model, MCMC method is used to estimate the parameters. The details of the model and Bayesian model are described in the following subsections.

2.2 Bayesian model of peptide identification from mass spectrum data

Regarding the prior information in Bayesian model,

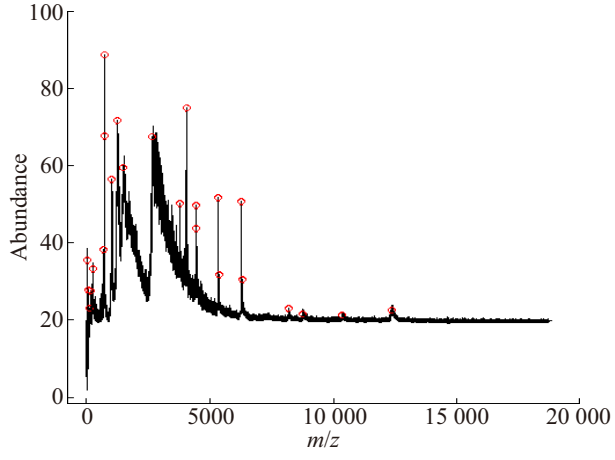


Fig. 2 Plot of Haar wavelet transformation.

Huang et al.[13] estimated the average bond cleavage abundance for each amino acid pair for both the b and y ions for gas-phase dissociation spectra. They inspired us about cleavage pair abundance if we want to know cleavages in the pairs of amino acid residues. Therefore, this information will be used in our Bayesian model as prior information to identify the true peptide. For the Bayesian model used in this paper, we use $p(D_k)$ to estimate a paired peptide j and spectrum k. D_k is the collection of spectrum of the peptides, Y_j denotes the protein of peptides. Among the estimated posterior probability of the matched peptide, only the highest likelihood is kept and selected as the best candidate peptide. By estimating the prior probability $P(Y_j = y_j)$, the $p(D_k)$ is calculated as below:

$$p(D_k) = \frac{p(D_k|Y_j)p(Y_j)}{\sum_{y_j} p(D_k|Y_j = y_j)p(Y_j = y_j)} \quad (3)$$

The likelihood measures how well the observed spectrum matches the theoretical spectrum. Due to the complexity of the posterior density in an MS/MS spectrum, we use MCMC to estimate it.

2.3 MCMC

MCMC is a statistical method for sampling from an invariant distribution based upon the construction of a Markov chain. The strong mathematics background makes it popularly used in a variety of the fields. The Metropolis-Hastings algorithm[14–17] is one kind of MCMC method which gets a sequence of random samples with a complex invariant distribution. The Gibbs sampling is a special case of the Metropolis-Hastings in which the random value is always accepted and only univariate full conditional distributions are considered[18, 19]. Gibbs sampling is an iterative scheme by sampling a subset of parameters at a time while fixing the rest at the sample values from the previous iteration. To produce a sample from the full joint distribution, Gibbs sampling simulates n random variables sequentially from the n univariate conditionals[20]. The process of Gibbs Sampling for the k-th peptide candidate and the derivations of the conditional posterior distributions of important model parameters are briefly summarized below.

2.4 Gibbs sampling

Our final goal is to locate the latent variables and sample all complete data of the mass spectrum. How do we get $p(p_i|P, I, O, \alpha)$? We first construct Markov chains

based on each peptide sequence as the following. In order to construct an MCMC that can be reversible, we should fix the length of the mass spectrum. In this paper, we use the possible maximum length of the mass spectrum. Each time we can randomly pick up the position of the mass spectrum and the probability is $1/n$, where n represents the length of the mass spectrum. We get the maximum length of peptide sequence as in Eq. (4).

$$\text{Maxlen} = \frac{\text{parent\_value}}{\text{minimum\_acid}} \tag{4}$$

According to Markov blanket, the latent variable, its corresponding observed peak value, and its associated intensity, we get the joint distribution as in Eq. (5).

$$p(\text{peak\_value}, s) = f(P, O, I, m, \theta) \tag{5}$$

We can remove parameter θ and get the join distribution for the i-th prior probability $p(m_i)$ which is the probability of the observing expected fragment that has a mass in the data. The probability for all expected fragment ions is obtained from Ref. [21]. $p(m_i)$ is calculated as the following:

$$p(b - m_{\mathrm{H_2O}}) = 0.66 \tag{6}$$

$$p(y - m_{\mathrm{H_2O}}) = 0.21 \tag{7}$$

$$p(b - m_{\mathrm{H_2O}} - m_{\mathrm{H_2O}}) = 0.71 \tag{8}$$

$$p(y - m_{\mathrm{H_2O}} - m_{\mathrm{H_2O}}) = 0.09 \tag{9}$$

$$p(b - m_{\mathrm{NH_3}}) = 0.28 \tag{10}$$

$$p(y - m_{\mathrm{NH_3}}) = 0.19 \tag{11}$$

$$p(b - m_{\mathrm{H_2O}} - m_{\mathrm{NH_3}}) = 0.09 \tag{12}$$

$$p(y - m_{\mathrm{H_2O}} - m_{\mathrm{NH_3}}) = 0.01 \tag{13}$$

In the above equations, $m_{\mathrm{H_2O}}$ is the water's relative molecular mass (18). $m_{\mathrm{NH_3}}$ is $\mathrm{NH_3}$'s relative molecular mass (17). The conditional probability equals the scoring function given below:

$$p(s_1, s_2, \cdots, s_n) = M_1, M_2, \cdots, M_n \to a_1, a_2, \cdots, a_n \tag{14}$$

where M_i is the m/z of mass spectrum, a_i is the intensity of the mass spectrum.

As shown in Fig. 3, the processing of the analysis steps of the algorithm starts from walk 1. In each step, the different fragment of the peptide is recognized and kept till walk n to get the final sequence. Gibbs sampling from the ultimate goal is a multi-dimensional distribution of $f(X_1, X_2, \cdots, X_n)$ that generates a random number.

Given a set of observations $x_1, \cdots, x_n, y_1, \cdots, y_n$,

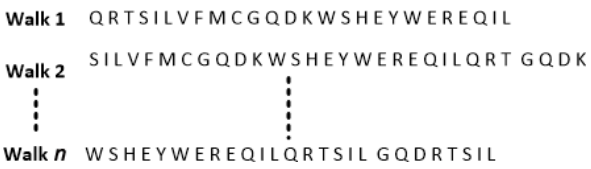


Fig. 3 The plot of the processing of the algorithm.

where x indicates the mass peak value and y is the intensity, we need to specify the number of peaks in the mass spectra and the location of the mass spectra. After the preprocessing of the MS/MS data, we first use the baseline subtraction and the wavelet transformation to remove the noises from the spectra according to the following steps:

Step 1 Initialization

Set length $k_{\max}$ of the peptide sequence based on the peptide sequence. The actual length of the experimental spectrum is m_l. $k_{\max} \geqslant m_l$. Set T be the number of iterations of MCMC. Build the array of the peptide sequence and put the observed experimental data into the array and mark the rest ones as empty. The process of initialization is shown in Fig. 4.

Step 2 Four operations for current position

Deletion of a specific mass is shown in Fig. 5 and the combination of the mass spectrum depicts in Fig. 6.

The processing of adding one H2O is shown in Fig. 7, for rare fragments (b–$\mathrm{H_2O}$), etc., let

$$M_i = M_i + m_{\mathrm{H_2O}} \tag{15}$$

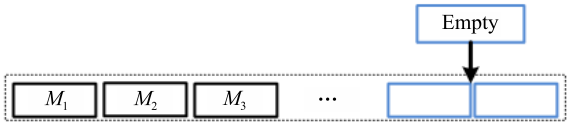


Fig. 4 The initial array of the peptide sequence.

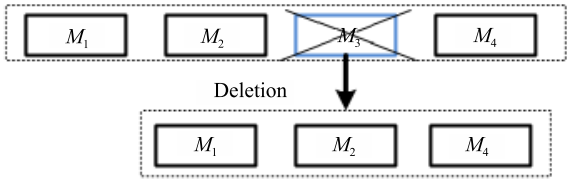


Fig. 5 The deletion of a specific mass.

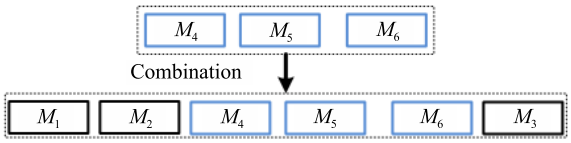


Fig. 6 The combination of specific mass peaks.

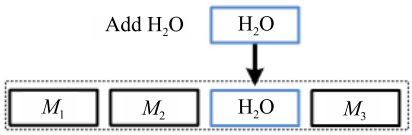


Fig. 7 The processing of adding water.

Add one NH_3, the process is shown in Fig. 8. If the fragments belong to rare fragments, $(b - NH_3)$ etc., let

$$M_i = M_i + m_{NH_3} \tag{16}$$

Figure 9 depicts the processing of adding both H_2O and NH_3.

Step 3 Sample random probability

The sample random probability $p = \mu(0, 1)$ is calculated as follows. If $p(s)p(x_i - 1) > p$, accept, else reject.

Loop until T meets the criteria, then update the peak value according to the acceptance rate. $T = T + 1$. End after several random walks and get all possible peptide candidates by sampling simultaneously. According to the Bayesian rule, the posterior probability is proportional to the likelihood times the prior probability, so that the posterior probability will update in each iteration.

3 Results

We used 100 000 runs to generate our spectra data and burned first 4000 runs. In order to test the performance of our peak detection method, we first use the MCMC to randomly generate peak location i. The MCMC model has several operations which include deletion, combination, acceptance, and adding H_2O. It not only can handle noise peaks but also can handle missing peaks.

In contrast, our peak detection method successfully detects all correct peaks. Based on the MCMC simulation results, we plot a Receiver Operating Characteristic (ROC) curve of sensitivity versus FDR (False Discovery Rate) as shown in Fig. 10. The right curve is the our method compared with the left curve for PepNovo[21]. It shows that our method is little better than PepNovo.

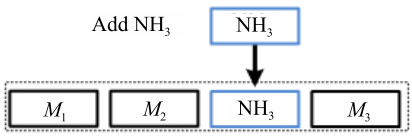


Fig. 8 The processing of NH_3.

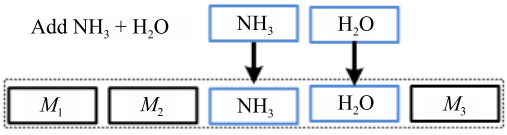


Fig. 9 The plot of adding both H_2O and NH_3.

4 Discussion

The common problems such as a limited number of known genome sequences, noisy data, and incomplete ion sequences lead to the limit accuracy of protein identification. Methods that have maximum sensitivity with minimum false discovery rate for the identification of peptide sequences are extremely in need in proteomics research area. In this paper, we describe an MCMC algorithm to improve the accuracy of the identification of proteins. With the MCMC algorithms, we approximate the target distribution in case of the posterior distribution of the unknown peptide sequence. The advantage of our method is that it is not dependent upon known peptides. We expect that our method will obtain more accurate estimations of the true peptides and help identifying correct proteins.

Acknowledgements

This work was supported by an NSF Science and Technology Center, under Grant Agreement CCF-0939370 and 2 G12 RR003048 from the RCMI program, Division of Research Infrastructure, National Center for Research Resources, NIH.

References

[1] P. Du and R. H. Angeletti, Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution, *Anal. Chem.*, vol. 78,

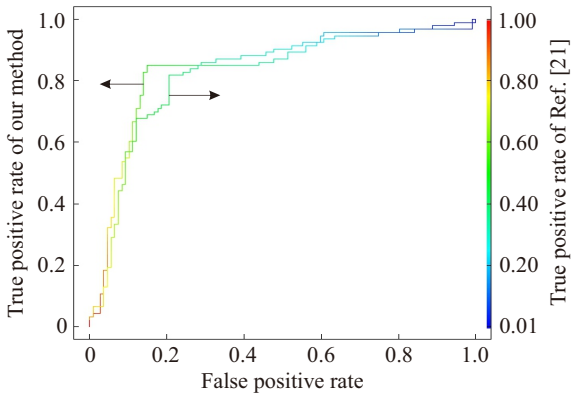


Fig. 10 ROC curve of the sensitivity on the *y*-axis and specificity on the *x*-axis.

pp. 3385–3392, 2006.

[2] X. J. Li, E. C. Yi, C. J. Kemp, H. Zhang, and R. Aebersold, A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry, *Molecular Cell Proteomics*, vol. 4, pp. 1328–1340, 2005.

[3] O. Schulz-Trieglaf, N. Pfeifer, C. Gröpl, O. Kohlbacher, and K. Reinert, Lc-MSsim a simulation software for liquid chromatography mass spectrometry data, *BMC Bioinformatics*, vol. 9, p. 423, 2008.

[4] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, vol. 20, pp. 3551–3567, 1999.

[5] X. Li, E. C. Yi, C. J. Kemp, H. Zhang, and R. Aebersold, A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry, *S. Mol. Cell Proteom*., vol. 4, pp. 1328–1340, 2005.

[6] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. Lin, et al., A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS, *Bioinformatics*, vol. 22, no. 15, pp. 1902–1909, 2006.

[7] K. Noy and D. Fasulo, Improved model-based, platformindependent feature extraction for mass spectrometry, *Bioinformatics*, vol. 23, pp. 2528–2535, 2007.

[8] N. Jaitly, A. Mayampurath, K. Littlefield, J. N. Adkins, G. A. Anderson, and R. D. Smith, Decon2LS: An opensource software package for automated processing and visualization of high resolution mass spectrometry data, *BMC Bioinformatics*, vol. 10, p. 87, 2009.

[9] M. Sturm, A. Bertsch, C. Grpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, and O. Kohlbacher, OpenMS — An open-source software framework for mass spectrometry, *BMC Bioinformatics*, vol. 9, p. 163, 2008.

[10] P. Du and R. H. Angeletti, Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution, *Anal. Chem*., vol. 78, pp. 3385–3392, 2006.

[11] J. Zhang, H. Wang, A. Suffredini, D. Gonzales, E. Gonzales, Y. Huang, and X. Zhou, Bayesian peak detection for pro-TOF MS MALDI data, in *Proc of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 661–664.

[12] Y. Sun, J. Zhang, U. Braga-Neto, and E. R. Dougherty, BPDA—A Bayesian peptide detection algorithm for mass spectrometry, *BMC Bioinformatics*, vol. 11, p. 490, 2010. doi: 10.1186/1471-2105-11-490.

[13] Y. Huang, J. M. Triscari, L. Pasa-Tolic, A. G. Anderson, M. S. Lipton, R. D. Smith, and V. H. Wysocki, Dissociation behavior of doubly-charged tryptic peptides: Correlation of gas-phase cleavage abundance with ramachandran plots, *Journal of American Chemical Society*, vol. 126, pp. 3034–3035, 2004.

[14] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, *Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[15] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[16] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan, An introduction to MCMC for machine learning, *Machine Learning*, vol. 50, no. 1, p. 543, 2003.

[17] D. Sorensen and D. Gianola, *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. Springer, 2002.

[18] S. Geman and D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

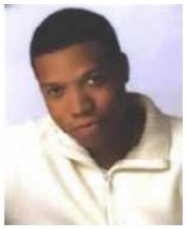
[19] A. Gelfand and A. Smith, Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 398–409, 1990.

[20] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer-Verlag, 1990.

[21] A. Frank and P. Pevzner, PepNovo: de novo peptide sequencing via probabilistic network modeling, *Anal. Chem*., vol. 77, no. 4, pp. 964–973, 2005.

Chunmei Liu received her BS and MS degrees in computer software from Anhui University in 1999 and 2002, respectively. She received her PhD degree in computer science from The University of Georgia in 2006. She has been a full professor since 2014 in the Department of Computer Science of Howard University. Her research interests include computational biology, graph algorithms, and theory of computation. Her recent research involves designing computationally efficient algorithms for protein identification, protein structure prediction, and protein-protein interactions.

Hui Li is currently a postdoctoral fellow at the Department of Computer Science in Howard University. He received his PhD degree in computer science from Beijing University of Technology in 2009. His research interests include computational biology, bioinformatics, pattern recognition, and algorithm.



Legand Burge received his BS degree in computer and information science from Langston University in 1992 and his PhD degree in computer science from Oklahoma State University in 1998. He has been a full professor at Howard University since 2009. His current research interests lie in the field of distributed computing. The primary thrust of his current research is in global resource management in large-scale distributed systems. In particular, he is interested in middleware technology to support scalable infrastructures for pervasive environments capable of servicing a very large number of small (possibly mobile) distributed and embedded devices efficiently. He is also interested in the application of distributed high performance computing to solve computational science problems in Biology, Physics, and Chemistry.

Mugizi Robert Rwebangira received his BS degree in systems and computer science from Howard University in 2002 and his PhD degree in computer science from Carnegie Mellon University in 2008. He has been an assistant professor at Howard University since 2010. He has received grant funding from the Army Research Lab and the National Science Foundation and published in the areas of semi-supervised learning algorithms, computational biology, and voting theory. His current research interests are in transfer learning and computational sociolinguistics.