Multiple Peak Alignment in Sequential Data Analysis: A Scale-Space-Based Approach

Weichuan Yu, Xiaoye Li, Junfeng Liu, Baolin Wu, Kenneth R. Williams, and Hongyu Zhao

Abstract—In this paper, we address the multiple peak alignment problem in sequential data analysis with an approach based on the Gaussian scale-space theory. We assume that multiple sets of detected peaks are the observed samples of a set of common peaks. We also assume that the locations of the observed peaks follow unimodal distributions (e.g., normal distribution) with their means equal to the corresponding locations of the common peaks and variances reflecting the extension of their variations. Under these assumptions, we convert the problem of estimating locations of the unknown number of common peaks from multiple sets of detected peaks into a much simpler problem of searching for local maxima in the scale-space representation. The optimization of the scale parameter is achieved using an energy minimization approach. We compare our approach with a hierarchical clustering method using both simulated data and real mass spectrometry data. We also demonstrate the merit of extending the binary peak detection method (i.e., a candidate is considered either as a peak or as a nonpeak) with a quantitative scoring measure-based approach (i.e., we assign to each candidate a possibility of being a peak).

Index Terms—Biomarker discovery, peak identification, multiple peak alignment, scale-space, prior information, energy minimization, parameter optimization.

1 INTRODUCTION

IN the postgenome era, proteomics has attracted more and more attention due to its potential in understanding biological functions and structures at the protein level. Although recent years have witnessed great advancement in the collection and analysis of gene expression microarray data, proteins are, in fact, the functional units that are of biological relevance. The rapid advances in mass spectrometry (MS) instrumentation and attendant protein profiling methodologies have substantially increased interest in using MS approaches to identifying peptide and protein biomarkers of disease. This great level of interest arises from the high potential of biomarkers to provide earlier diagnosis, more accurate prognosis and disease classification, to guide treatment, and to increase our understanding at the molecular level of a wide range of human diseases.

The MS data consists of paired mass-charge-ratio (m/z value) versus intensity data points. Generally, the total number of measured data points is extremely large (about

10^5 for a conventional MALDI-TOF instrument as compared to perhaps 10^6 for a MALDI-FTICR instrument covering the range from 700-28,000 Da), while the sample size is usually in the hundreds. This very high data size to sample size ratio poses significant challenges in MS data analysis. The disease biomarker discovery approach is directed entirely at the most important question: "Which relatively few (mass, intensity) data points in MS spectra obtained on sera and other biological samples can best differentiate all normal from disease patients?" In 2002, Petricoin et al. [17] first published an MS-based study of analyzing serum samples from 50 control and 50 case samples from patients with ovarian cancer. They identified five peptide biomarkers ranging in size from 534 to 2,465 Da and were able to correctly classify all 50 ovarian cancer samples in a masked set of serum samples from 116 patients that included 50 ovarian cancer patients and 66 women unaffected or with nonmalignant disorders. Since then, MS-based approaches have been used in the study of different types of cancer (ovarian, prostate, breast, bladder, renal, head and neck cancer [22]) and other diseases (e.g., [15], and sleeping sickness [16]). The biological fluids analyzed include serum, urine, cerebrospinal fluid, nipple aspirate fluid, and others.

In spite of these promising results, however, some concerns still need to be addressed before we can apply an MS-based approach in real clinical diagnosis of diseases. The major concern is that closer examination of the identified biomarkers (such as in [17]) suggests strongly that some of the biomarkers identified in some studies may arise from statistical artifacts rather than the ionization of biologically significant peptides. In addition, there are systematic differences among these spectra [5], which makes the direct comparison of multiple spectra unstable.

In order to address these concerns, we need appropriate preprocessing steps to extract biologically meaningful features from spectra and to build a mutual correspondence among them so that the systematic differences among

- W. Yu is with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Sai Kung, Kowloon, Hong Kong. E-mail: eeyu@ust.hk.
- X. Li is with the Department of Applied Mathematics, Yale University, 300 George Street, Suite 503, New Haven, CT 06520. E-mail: xiaoye.li@yale.edu.
- J. Liu is with Department of Statistics, West Virginia University, Hodges Hall 406, PO Box 6330, Morgantown, WV 26506. E-mail: jfliu@stat.wvu.edu.
- B. Wu is with the Division of Biostatistics, School of Public Health, University of Minnesota, A442 Mayo Building, MMC 303, 420 Delaware St. SE, Minneapolis, MN 55455. E-mail: baolin@biostat.umn.edu.
- K.R. Williams is with the Keck Laboratory, Yale University, 300 George Street, G005, New Haven, CT 06520. E-mail: kenneth.williams@yale.edu.
- H. Zhao is with the Department of Epidemiology and Public Health, Yale University, 60 College Street, LEPH 200, New Haven, CT 06520. E-mail: hongyu.zhao@yale.edu.

Manuscript received 18 Mar. 2005; revised 20 July 2005; accepted 30 Oct. 2005; published online 31 July 2006.

For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0016-0305.

different spectra do not affect the discovery of biomarkers. Generally, only the spectral peaks that result from ionization of biomolecules such as peptides and proteins are biologically meaningful and of use in applications. To detect and locate spectral peaks, different data preprocessing methods have been proposed. A commonly used protocol for MS data preprocessing consists of the following steps: spectrum calibration, baseline correction, smoothing, intensity normalization, peak identification, and peak alignment [23], [25], [7]. Among them, peak identification and alignment are the two most important preprocessing steps. Normally, spectral peaks are local maxima in MS data. But local maximum search alone does not work in peak detection/finding since we have to deal with chemical and electronic noise whose properties are unknown. To address this issue, Coombes et al. [7] defined noise as the median value of intensity. Satten et al. [19] used the negative part of the normalized MS data to estimate the variance of noise. Wavelets-based approaches [8], [18] have also been proposed to denoise the MS data before peak detection. Based on the observation that there were substantial measurement errors in the intensity value, Yasui et al. [25] argued that binary peak/nonpeak data was more useful than the absolute values of intensity, while they still used a local maximum search method to detect peaks. Clearly, the success of a noise-estimation-plus-threshold method largely depends on the correctness of the noise model, which remains to be developed and validated.

After peaks have been detected, we must align them together before we compare peaks in different data sets. Previous studies have shown that the variation of peak location in different data sets is nonlinear [21], [10]. The example in [28] shows that this variation still exists even when we use technical replicates. The reasons that underlie data variation are extremely complicated. Taking the MALDI-MS data as an example, differences in sample preparation, chemical noise, cocrystallization and deposition of the matrix-sample onto the target, laser position on the target, etc. can all cause data variation. Our focus is not to identify these reasons. Instead, we focus on developing a framework to reduce the variation and align these peaks together.

Toward this direction, some methods have been proposed. Yasui et al. [25] believed that the m/z axis shift of peaks were approximately $\pm 0.1\%$ to $\pm 0.2\%$ of the m/z value. Thus, they expanded each peak to its local neighborhood with the width equal to 0.4 percent of the m/z value of the middle point. This method may oversimplify the problem.

Randolph and Yasui [18] used wavelets to represent the MS data in a multiscale framework. They first aligned peaks at a dominant coarser scale from multiple samples and then aligned the remaining peaks at a finer scale. From the signal representation point of view, this approach is very interesting. But, it remains to study if representing peaks at multiscales is biologically reasonable, i.e., if peaks at coarse scales really correspond to true peptides.

Johnson et al. [12] assumed that the peak variation is less than the typical distance between peaks and they used a closest point matching method in peak alignment. This method is limited by the data quality and it cannot handle large peak variation or false positive peak detection results.

Dynamic programming (DP) based approaches [14], [21] have also been proposed. DP has been used in gene expression analysis to warp one gene expression time series to a similar series obtained from a different biological replicate [1], where the correspondence between the two gene expression time series is guaranteed. In MS data analysis, however, the situation is more complicated since the one-to-one correspondence between two data sets does not always exist. Although it is still possible to apply DP to deal with the short-of-correspondence problem, some modifications are necessary (such as adding an additional distance penalty term in the estimation of the correspondence matrix). It also remains unclear how DP can identify and ignore outliers during the matching.

Eilers [10] proposed a parametric warping model with polynomial functions or spline functions to align chromatograms. In order to fix warping parameters, he added calibration example sequences into chromatograms. While the idea of using parametric model is interesting, it is difficult to repeat the same method in MS data since we cannot add many calibrator compounds into the MS samples. Also, it is unclear if a polynomial function with the second order would be precise enough to describe the nonlinear shift of MS peaks.

Tibshirani et al. [20] applied a hierarchical clustering method to construct a dendrogram of all peaks from multiple samples. They cut off the dendrogram using a predefined parameter and clustered the remaining branches into different groups. They then considered the centers of these groups as common peaks and aligned every peak set with respect to the common peaks. The implicit assumption behind their approach is that around each of the common peaks, the observed peaks from multiple samples obey a certain kind of distribution with the mean equal to the location of the common peak. The assumption agrees well with the motivation of peak alignment. However, the cut-off parameter and the final clustering results can be influenced by changing a few nodes in the dendrogram, while some noisy points or outliers (e.g., caused by false positive peak detection results) often cause such changes, as we will show later in this paper. Consequently, the clustering method may not be stable against noise.

In addition, the distance used in the clustering method is only based on the m/z information. It is not described how intensity or other kinds of related information can be used in the construction of the dendrogram.

Recently, a Hidden Markov Model-based approach was proposed by Listgarten et al. [13] to align multiple time series data. This approach considered the intensity information in the alignment. The implicit assumption behind this approach is that there exists a one-to-one correspondence among multiple time series, as shown by the replicate data examples in [13]. However, this method cannot be directly applied to the MS data, which does not always guarantee a one-to-one correspondence among peaks. Another point worth mentioning is that Listgarten et al. used EM-based training method to estimate the locations of *latent trace* and then used these latent trace points to align original time series. While the idea of training is interesting, they have to manually determine the number of parameters before applying the EM algorithm, which is normally unknown in MS data.

Our contribution in this paper is that we address the peak alignment problem from a scale-space point of view. In our approach, we do not have to manually determine any parameters thanks to the parameter optimization scheme. Our simulation results as well as real data analysis also show that this new approach is more robust against noise than the hierarchical clustering method. In addition, it is possible in our approach to use both m/z information and intensity values (or similar measure) in the alignment process, while most previous approaches (including the DP method and the hierarchical clustering method, but excepting [13] mentioned above) used only the m/z location information during the alignment of peaks.

Our approach can also be used in the analysis of other forms of sequential data. By sequential data we mean that the data points are ordered with a quantitative index. The order, however, does not necessarily mean that the data points with lower index values will influence data points with higher index values or every data point has to be considered, as in the case of the time series data [13]. Typical examples of sequential data that can be analyzed with our approach include the nuclear magnetic resonance (NMR) data (whose format is almost the same as the MS data except that only the interpretation of index is slightly different) [11] and optical mapping data [3]. From now on, we will focus on the MS data analysis to simplify the explanation. But, keep in mind that the same method can be applied to the analysis of other kinds of sequential data as well.

We construct the rest of the paper as follows: In Section 2, we propose a Gaussian scale-space approach to align multiple peak sets with the possibility of including both m/z values and intensity into the alignment process. Comparisons and experimental results are shown in Section 3. In Section 4, we describe how the binary peak identification can be extended and the benefit of the extension. We finally conclude the paper in Section 5 with discussion on various related issues and future work.

2 MULTIPLE PEAK ALIGNMENT

The goal of multiple peak alignment is to identify their mutual correspondence. We formulate the “true” locations of peaks as unknown parameters and regard peak detection results as the observed samples. Then, the problem of aligning multiple peak sets is converted to the problem of estimating mean values (we call them common peaks from now on) with the assumption that the majority of peaks locate close to the “true” locations with only a few outliers. After estimating the common peaks, we can straightforwardly align peaks with respect to the common peaks using a closest point matching method, i.e., for every common peak, its counterpart in a sample has the smallest distance among all peaks in the same sample. Because the set of common peaks is estimated using all information we have, it is more trustworthy than any individual peak set. In this sense, the set of common peaks serves as a template and false positive or false negative peak detection results in any individual peak set will be checked by the template and will NOT be automatically included in the alignment process, thus making the alignment result more robust against noise or feature absence.

Intuitively, the relative distance between peak samples and common peaks may be further used as a quality measure in alignment because a good alignment center should attract more sample peaks with relatively smaller distances than a poor alignment center does. Suppose we have N samples with each sample containing some detected peaks. We use K_j to indicate the peak number in the j th sample and

$$\sum_{j=1}^N K_j = M.$$

Please note that the number of peaks in these samples may not necessarily be the same (i.e., K_j is not necessarily equal to K_m if $j \neq m$). By denoting $d_{i,j}$ as the distance between the i th peak in the j th sample and its closest common peak, we define an average distance as:

$$D_{av} = \frac{\sum_{j=1}^N \sum_{i=1}^{K_j} \|d_{i,j}\|^2}{M}. \quad (1)$$

Now, the key issue is to estimate the locations of the common peaks. Our goal is to estimate the locations of k common peaks $\vec{\mu} = (\mu_1, \mu_2, \dots, \mu_k)$ from the observed peaks across multiple spectra. The challenge now is that k is usually unknown and we face a parameter estimation problem in the feature space with unknown dimension. In the following, we use a scale-space-based approach to address this challenge.

The scale-space representation was first introduced by Babaud et al. [4]. In this seminal work, they proved that Gaussian is the only kernel whose filtering response has monotonically increasing maxima when the bandwidth of the filter is increased [4]. Since then the Gaussian has been widely used in the scale-space filtering approaches as it provides a complete representation of features in a continuous space.

In this paper, we use a scale-space representation to alleviate the identification of common peaks among multiple MS spectra. Concretely, we represent the peak identification results as a set of Dirac components with different locations and weighting coefficients:

$$P(x) = \sum_{i=1}^M a_i \delta(x - x_i) \quad (2)$$

with x_i denoting the peak location (m/z value) and a_i denoting the weighting coefficient ($0 \leq a_i \leq 1$). Here, the number M denotes the total number of peaks in all N samples and we do not care about which peak is from which sample. Then, we define a Gaussian function with zero mean and σ^2 variance as

$$G(x) = e^{-\frac{x^2}{2\sigma^2}}. \quad (3)$$

For simplicity, we ignore the term $\frac{1}{\sqrt{2\pi}\sigma}$ in the Gaussian function since we are only interested in the relative value in the scale-space. After convolving $P(x)$ with $G(x)$, we obtain a scale-space representation of all peaks:

$$S(x; \sigma) = P(x) * G(x) = \sum_{i=1}^M a_i e^{-\frac{(x-x_i)^2}{2\sigma^2}}.$$

This representation is just a diffused version of peak distribution, with diffusion controlled by the scale parameter σ . By changing σ , we can easily change the scale. Correspondingly, searching for the location means (or cluster centers) is simplified as searching for the local maxima in $S(x; \sigma)$.

The scale-space representation can simplify the problem of model identification: We do not have to fix the number k during the estimation of $\bar{\mu}$ since k changes naturally when we change σ . As long as the range of σ covers all possible scales, the representation of the peak distribution is complete thanks to the scale-space theory.

Another nice point about this representation is that the relative importance of different peaks is naturally weighted by the coefficients a_i . This makes it possible to align peaks based on their relative importance. For example, suppose we know from prior knowledge that some peaks close to a certain m/z location correspond to a certain peptide and they should definitely be aligned together. We can simply set the a_i values of these peaks much larger than those values of neighboring peaks, thus ensuring that the alignment of the chosen candidates has a higher priority over the alignment of neighboring peaks. Note that, it is still possible to treat all peaks uniformly by setting all a_i values equal to one, making m/z value based alignment a special case of our approach.

Certainly, we still need to determine at which scale level we estimate $\bar{\mu}$. We propose to use an energy minimization method to find out the best scale. The energy function consists of a data-fitting term and a regularization term. The data-fitting term measures the differences between the sample peaks and the scale-space-based representation. We use the common sum-of-square-difference here:

$$D(\sigma) = \sum_{i=1}^M (P(x_i) - S(x_i; \sigma))^2. \quad (4)$$

Note that each peak point is used in the data fitting once and only once.

The regularization term prevents overfitting (we can easily find out the best fit by assuming the scale level is infinitesimal and let the number of peak locations equal the number of peaks in the data) and oversmoothing (a too large σ blurs or even destroys the gaps between neighboring common peaks). Here, we assume that the best scale follows a Gaussian distribution $N(\mu_s, \sigma_s)$ in the scale-space (i.e., there is no negative scale parameter), where μ_s and σ_s are determined using data. Concretely, we denote the m/z location of the i th peak in the j th sample as $m_{i,j}$ and use the m/z distances between neighboring peaks in all samples to estimate μ_s and σ_s :

$$\mu_s = \frac{\text{median}\{m_{i+1,j} - m_{i,j}\}}{6}, i = 1, \dots, K_j - 1; j = 1, \dots, N,$$

and

$$\sigma_s = \frac{\text{std}\{m_{i+1,j} - m_{i,j}\}}{6}, i = 1, \dots, K_j - 1; j = 1, \dots, N.$$

with i denoting the peak index in each sample, j denoting the sample index, and

$$\sum_{j=1}^N K_j = M.$$

We made such a choice based on the understanding that, if two groups of Gaussian diffused peaks are well-separated, the distance between their centers should be at least 6σ (3σ for each side of the Gaussian distribution function). Thus, the median of this distance can be used to reflect our belief on the best scale (i.e., μ_s) and the standard variation of these distances can be used to reflect how widely the best scale can spread (i.e., σ_s). In the next section, we will show how these parameters influence the extraction of the locations of the common peaks.

The regularization term then reads

$$R(\sigma) = 1 - e^{-\frac{(\sigma - \mu_s)^2}{2\sigma_s^2}}. \quad (5)$$

Finally, we construct the energy function as

$$E(\sigma) = D(\sigma) + \lambda R(\sigma), \quad (6)$$

where λ is a parameter adjusting the relative importance between $D(\sigma)$ and $R(\sigma)$.

So far, we have described a scale-space framework to align multiple peaks. In the next section, we will test its robustness against noise and show its superiority over a hierarchical clustering method proposed in the literature [20].

3 EXPERIMENT

3.1 Synthetic Examples

In this section, we test the scale-space approach and compare it with a hierarchical clustering method [20]. Both methods are distance-based exact methods in the sense that no training is required. To make a fair comparison, we only use the m/z information in our approach by putting the weighting coefficients of all peaks equal to one. Here, we do not include the EM-based continuous profile model [13] into the comparison because we do not want to predefine the number of common peaks, which is necessary for [13] to work.

The point we like to emphasize is that the setting of a cut-off point in a hierarchical clustering method [20] is very sensitive to noise. We start with simulation examples. We simulated 50 samples with each sample having at most 20 detected peaks. The designed locations of these 20 peaks were arranged using the following nonlinear equation:

$$\mu_i = 9.5 \cdot i \left(1 + \frac{i}{20} \right), \quad i = 1, \dots, 20. \quad (7)$$

Around each of these 20 peak, we used a normal distribution $N(0, \sigma_i)$ to simulate the locations of the detected peaks. In order to simulate the linear diffusion effect of the mass spectrometer, we set σ_i to be linearly increasing:

$$\sigma_i = 1 + \frac{i}{25}, i = 1, \dots, 20. \quad (8)$$

To simulate biological variation, we randomly deleted 20 percent of the peak points. To simulate noise, we further added 50 points around the seventh common peak. We also tested different noise distributions here: These 50 points

Fig. 1. (a) Simulated peak distribution. Each circle represents an observed peak. The ground truths of the 20 peak locations are drawn as vertical lines. The horizontal axis denotes the m/z location and the vertical axis denotes the sample index. The noisy points are added between the sixth and the ninth true peak location. (b) This plot shows how noisy points were added. We observe 20 naturally grouped clusters on the left serving as the ground truth.

Fig. 2. Scale space representation of all peaks whose intensities are all equal to one. (b) Total energy versus scale parameter σ . We display several plots with different λ values. (c) Relative estimation error in percentage by using $\lambda = 1$.

followed a Gaussian distribution $N(0, 10)$ in the first test and followed a uniform distribution $U(-10, 10)$ in the second test. Fig. 1 shows one peak distribution example which is corrupted by Gaussian-distributed noise. Note that we only added noisy peaks around one common peak (the seventh peak) for simplicity. Generally, noisy peaks are distributed across the whole m/z range. We shall discuss the more general case in the next section (cf. Fig. 6).

If we do not add noise, the dendrogram in the hierarchical clustering method shows that we can cluster all peaks into 20 groups by choosing a cut-off height between 9.1 and 16.4. Here, we arbitrarily cut off the dendrogram at a height of 15. The centers of these groups are very close to the locations of the common peaks, indicating a satisfactory performance of the clustering method in this example, except for the fact that the cut-off point has to be determined manually.

Then, we tested the scale-space approach. Fig. 2 shows the resulting scale-space representation. Fig. 2 also shows the total energy term. In order to test the sensitivity of the total energy with respect to the weighting coefficient λ , we used different λ values and show the corresponding plots accordingly in Fig. 2 as well. Clearly, the regularization term plays a very important role to stabilize the total energy

term, while the value of the relative importance coefficient λ does not have much influence on the total energy. In this example, we simply set $\lambda = 1$.

After knowing the scale parameter that minimizes the energy function, it is then straightforward to obtain the number and locations of the common peaks at the estimated scale level by local maximum search. A comparison shows that the scale-space approach provides similar performance as the clustering method except for the first peak, where the chosen scale is coarser than the real one and the estimation error is consequently slightly larger. The average distance measure in Table 1 also confirms that both methods have similar performances.

However, the situation will be different if noise is added. A direct result of adding noise is that there is no more appropriate range of cut-off height that can group peaks into the desired 20 clusters: When we used the height of 15 as before, we obtained 21 clusters; when we tried to keep the cluster number as 20, for example, by choosing the cut-off height as 17 (in the middle of the range that keeps the number of clusters equal to 20), the mutual correspondence was not as expected, either, as shown in Fig. 3. In contrast, the scale-space approach was more robust against noise and still obtained reasonable results. Also, the quantitative

TABLE 1
Estimated Average Distances in the Synthesized Examples

Simulation	Hierarchical Clustering	Scale Space without Weighting Information
no noise added	1.1193	1.1181
Gaussian-distributed noise added	1.4654	1.1368
Uniform-distributed noise added	1.2028	1.1309

The first row shows the result without any noise added to the data. The cut-off height is set to be 15 in the clustering method. The second row shows the result when Gaussian-distributed noise ($N(0, 10)$) is added. The cut-off height is manually set to be 17 in the clustering method. The third row shows the result when uniformly-distributed noise ($U(-10, 10)$) is added. The cut-off height is manually set to be 16.4 in the clustering method.

Fig. 3. (a) Estimated cluster centers (circles) versus ground truth of peak locations (stars) when we cut off the clustering tree at the height of 15. There is a total of 21 clusters. The two clusters within the black window are corresponding to the same peak. Clearly, the added noisy points cause the number increase and shift of cluster centers. (b) Estimated cluster centers (circles) versus ground truth of peak locations (stars) when we cut off the clustering tree at the height of 17. Notice the incorrect correspondences within two black windows. The noisy points in this example follow Gaussian distribution $N(0, 10)$.

Fig. 4. (a) Peak detection result. Each point represents an observed peak. The horizontal axis denotes the $\log(m/z)$ location and the vertical axis denotes the sample index. (b) Sorted peak locations. There is no distinctively grouped cluster of peaks.

average distances in Table 1 indicate a similar conclusion: The distance using the clustering method is significantly higher than that using the scale-space approach.

We repeated the same comparison using uniformly-distributed noise $U(-10, 10)$. A similar conclusion holds with the quantitative average distances shown in Table 1.

3.2 Real Examples

Here, we use replicate data from a real study to ensure that the sample variation is minimal. From the National Ovarian Cancer Early Detection Program at Northwestern University Hospital, we have obtained eight different samples, each with six technical replicates. Among these replicates, four replicates on aliquots of the same sample were used by the Micromass M@LDI-L/R instrument and two replicates

were used manually to obtain MS spectra. As one of the manual mass spectra was missing, we had, therefore, a total of 47 data sets for the test. The data are available at <http://bioinformatics.med.yale.edu/MSDATA>. They are all obtained in the Reflectron mode with the range 800-3,500 Da. Fig. 4 shows the peak detection results before alignment and a sorted plot of all peaks.

As the ground truth of the peak alignment results is unknown, the only criterion we could use is the average distance. A smaller average distance indicates a better alignment. Fig. 5 shows the variation of the common peak numbers with respect to the variation of the cut-off/scale parameters in both the clustering method and the scale-space approach. The plot on the right also shows the average distance values at different common peak numbers.

Fig. 5. (a) The number of common peaks versus the cut-off parameter. (b) The number of common peaks versus the scale parameter σ . (c) A comparison of average distances. The circles denote the distances using the clustering method for peak alignment, while the stars denote the distances using the scale-space approach.

The scale-space approach always provided a lower average distance value.

3.3 Impact of Different Alignment Methods on Classification

Now, let us check how significantly the alignment method would influence downstream statistical analysis, e.g., classification analysis. For this purpose, we preprocessed the data with the same method under the same parameter setting until we came to the stage of peak alignment. Then, we used both the hierarchical clustering method and the scale-space approach to align the detected peaks. After that, we fed the alignment results of both methods separately into the same classification method and checked the classification accuracy. Based on the analysis in [24], we chose the Random Forests algorithm [6] in our test.

For the sake of comparing classification performance, we used serum samples from 48 control individuals and 47 ovarian cancer patients (also from the National Ovarian Cancer Early Detection Program at Northwestern University Hospital). We then subjected these samples to automated desalting and MALDI-MS on a M@LDI-LR instrument in the Keck Laboratory at Yale University. Here, we study the MS spectra obtained in the Reflectron mode with the range 800-3,500 Da. The data are available at <http://bioinformatics.med.yale.edu/MSDATA>.

We, in total, detected 73,706 peaks across 95 samples. As we did not know the ground truth of the common peak number, we used different parameters in the clustering method as well as the scale-space approach to align the detected peaks and identified the best parameter by the classification results. For the reason that we will explain in the next section, in the preprocessing step we kept as many weak peak candidates as possible. Consequently, we could not guarantee that the “best” scale parameter estimated using the energy minimization method is not corrupted by noisy peaks. Intuitively, it would be more appropriate to test multiple parameters under such a noisy condition. In the test, we used 60 samples for training (30 from the control group and 30 from the patient group) in the Random Forests method and used 35 samples for testing.

After running the same Random Forests method 50 times, we calculated the median of the misclassification rates. The

lowest misclassification rate in the scale-space approach was 40 percent, while the corresponding rate in the clustering method was about 44 percent. Here we only used about $\frac{2}{3}$ of total samples to train our classifier. If we used the 10-fold cross validation or .632+ method as suggested in [9], we might have a lower and more stable error estimation result. A lower error rate on the same data set was reported in [24], [20]. However, the cross validation in these two papers was not external to the feature selection or parameter setting. This would cause the underestimation of the error rate, as pointed out in [2]. Note that it is a very active area of research to develop statistical methods to reduce the misclassification rate. Our focus here is to evaluate the impact of different peak alignment methods on sample classification by fixing everything except the alignment method. Thus, our results indirectly suggest that our scale-space approach may lead to better classification performance than the hierarchical clustering method, although the difference is not statistically significant. More data sets are needed to fully explore the relative performance of different peak alignment methods on sample classification.

4 PEAK IDENTIFICATION USING A SCORING MECHANISM

Our review in Section 1 clearly shows that current peak detection methods belong to the binary identification category (i.e., a candidate is either a peak point or a nonpeak point). From the data analysis point of view, this simplification is prone to the influence of noise (noise may also produce some local maximal values). In addition, the use of a uniform threshold value may exclude some weak peaks in the MS data, while the existence/nonexistence of some weak peaks may be the most informative biomarkers.

In order to preserve the flexibility of including weak peaks in the peak identification process and have a quantitative measure of how strongly we believe that a candidate point is a peak, we propose to assign to each candidate point a score as its possibility of being a peak. The benefits of introducing such a possibility measure are twofold: 1) The dependence of the binary peak identification method on the threshold value is removed. 2) This method makes it easier to embed other information (such as peak width, area, or even shape information) into the

quantitative measure, thus making it possible to extract useful signals from noisy data. In the following, we will first describe some possible methods of using the scoring mechanism during peak identification and then show their advantages using a simulation example.

4.1 Scoring Mechanisms

Given a raw MS data (x, I) with x denoting the m/z value and I the intensity, we first correct the baseline and smooth the data. Then, we detect local maxima I_i ($i = 1, \dots$) at locations x_i . We only consider these points (x_i, I_i) as peak candidates and assign them scores based on certain pre-defined criteria. The criteria can be very flexible. Here, we illustrate the scoring principle using intensity and width criteria. The width of the peak candidate (x_i, I_i) is defined as

$$W_i = x_i^l - x_i^r$$

with x_i^l and x_i^r denoting the m/z values of two points in the smoothed data that have the local intensity minima and are close to the candidate (x_i, I_i) from the left and from the right, respectively. In this way, the candidates are extended to the form (x_i, I_i, W_i) . Certainly, we could replace the intensity with the area of the region between x_i^l and x_i^r . Also, we could define the width as the Full Width of the peak at Half its Maximum height (i.e., the well-known FWHM definition). Here, we stick to the above simple definition and it is beyond the scope of this paper to study which definition is the most appropriate in MS data analysis.

The score may be constructed in two alternative ways: In the first method, we can use a distribution model with parameters Θ_p to describe the typical distribution of the peak width and peak intensity in a data set. Similarly, we can use another distribution with parameters Θ_n to describe the distribution of the nonpeak width and nonpeak intensity. The parameters Θ_p and Θ_n of the distribution model can be fixed using the training samples. For a candidate point (x_i, I_i, W_i) , we can carry out a likelihood ratio test

$$L(x_i; I_i, W_i) = \frac{P(I_i, W_i | \Theta_p)}{P(I_i, W_i | \Theta_n)}$$

to see how likely this point can be a peak given the prior distributions of the peak width and peak intensity. After obtaining all the ratio values, we can further obtain a normalized measure for each candidate. If the value is close to one, then the candidate has a stronger possibility as a peak. If the value is close to zero, then the candidate has a weaker possibility as a peak.

The success of the above method largely depends on the accuracy of the training process and the appropriate choice of the distribution model, which is highly biological expertise dependent. Alternatively, we can use an approach which does not require an accurate distribution model. Suppose our prior knowledge indicates that a candidate is definitely a peak point if its width W is broader than W_1 and is definitely not a peak if its width is narrower than W_2 ($W_2 < W_1$) no matter what the intensity value is (imagine that a narrow spike with a high intensity value often results in a false positive peak detection). If its width is between W_2 and W_1 , the prior indicates that the higher the intensity, the

more likely it will be a peak. Based on this prior, we can formulate a scoring function as

$$Y(x_i; I_i, W_i) = \begin{cases} 1, & \text{if } W_i > W_1 \\ \frac{I_i - I_{\min}}{I_{\max} - I_{\min}}, & \text{if } W_2 \leq W_i \leq W_1 \\ 0, & \text{if } W_i < W_2, \end{cases} \quad (9)$$

where $I_{\max}$ and $I_{\min}$ denote the maximal and minimal intensity values in an MS spectrum, respectively.

The advantages and limitations of the above scoring functions are certainly worth further studying. There are probably better choices in constructing the scoring function. But, the point here is that we can replace a purely intensity- and/or width-based **hard** threshold value with a **soft** criterion in peak identification. The benefits are twofold: First, we are able to keep *weak* candidates by assigning smaller scoring values to these points, while in traditional peak identification processes we exclude these candidates from the peak list. Second, this formulation gives us the flexibility to incorporate other information of candidates into the peak identification process. With our better understanding of MS peaks, it is straightforward to incorporate more parameters in peak identification. In the following, we will show how the introduction of the scoring mechanism may improve the robustness of the method against noise.

4.2 Robustness Test

We use the same peak set as in Fig. 1, but assign to each peak a weighting coefficient between zero and one. The weighting coefficients were first designed to obey a uniform distribution. To simulate some strong peaks, we randomly chose 12 common peaks and made the corresponding weighting coefficients of peak points around these common peaks to be at least 0.75. To simulate noise, we added about 1,000 points with their weighting coefficients set between zero and 0.2. Fig. 6 shows a 3D view and a 2D view of the example.

For such a noisy example, the hierarchical clustering method no longer provides a stable range that can be used to cut off the dendrogram (cf. Fig. 7). If we really want to have 20 clusters, we have to manually find a cut-off height (which equals 27.3 in this example). But, the corresponding cluster centers do not agree well with the true locations of the common peaks, as shown in Fig. 7.

Now, let us check the performance of the scale-space approach. If we do not use the weighting coefficient information, the distribution of the true peaks is largely blurred, as shown in Fig. 8a. Consequently, the scale parameter μ_s is much smaller than expected and we obtain more local maxima than expected, as shown in Table 2. In contrast, if we use the weighting coefficient information, the scale-space representation is much cleaner and the estimated μ_s (we estimate it using weighted distance) is much closer to the expected value. Consequently, we still can obtain reasonable estimation results.

5 DISCUSSION AND FUTURE WORK

In this paper, we have proposed a scale-space approach to aligning multiple MS peak sets. Compared to the current hierarchical clustering method, our method is more robust against noise. More importantly, while the hierarchical

Fig. 6. (a) A 3D view of the peaks and noisy points with different weighting coefficients. We use the same peak set as in Fig. 1, but assign to each peak a weighting coefficient between zero and one. In addition, we add about 1,000 points with their weighting coefficients set between zero and 0.2. Peaks are represented with dots and noisy points are represented with circles. (b) A 2D view only shows the locations of the peaks and noisy points.

Fig. 7. (a) The number of clusters versus the cut-off height. (b) If we manually cut off the dendrogram at a height of 27.3 to produce 20 clusters, this plot shows the positions of cluster centers relative to the true locations of common peaks. Here, the first row shows true common peaks and the second row shows cluster centers. (c) The relative distances of cluster centers to common peaks normalized by the true locations of common peaks.

clustering method only uses the m/z location information, our scale-space approach makes it possible to embed intensity or other information into the same alignment process. This extension enables us to keep those weak candidates during peak identification, which may be informative biomarkers. With the traditional threshold framework, these weak peaks may be excluded at the right beginning.

Our method is similar to a recent work [13] in the sense that both methods consider the intensity information during the alignment. Their difference is that the authors in [13] had to make an assumption on the one-to-one correspondence among multiple data and they used an EM-based training method, while in our method both are not necessary.

In our current approach, we favor strong peaks as they weight more during the alignment. If we know that some weak peaks in certain regions should definitely be aligned together, we can adjust their weighting coefficients and the current framework still works. Certainly, we need the help of prior biological knowledge in the adjustment of the weights.

Our energy minimization method globally searches one optimal scale for the entire spectrum. One may argue that two close peaks may produce only one local maximum at a coarser scale level in the scale-space and thus yield only one common peak in the alignment. This is true under the current setting, although we also like to point out that the closeness between two neighboring peaks in the same MS spectrum usually has a lower bound that can be determined by the resolution of a given mass spectrometer, charge state

of test ions (e.g., single charge or double charge), and the fact that the minimal difference of chemical components is roughly 1 Da. For example, we usually expect that the minimal distance among isotopically resolved peaks will be around $\frac{1}{Z}$ Da, where Z denotes the charge state.

To improve the robustness of our algorithm with respect to very close peaks, we may implement the following approach in future work: We sequentially divide the mass spectra into different segments and assume that, within each segment the peaks, are roughly uniformly distributed, i.e., can be characterized with one scale. This assumption can be supported by the argument that the observed peak width in MS spectra is roughly monotonically increasing with the increase of the m/z value. It should be noted that the peak width is a measure of peak location variation. During the MS data acquisition process, we usually acquire about 20 to 40 spectra from one sample and then export their average as the final spectrum. If there is no variation in peak location, the peaks should look like very sharp Dirac functions. It is the peak location variation and the averaging computation that form a blurred shape for peaks. Consequently, peak width serves as a measure of peak location variation. The monotonically increasing property of peak width, therefore, supports the assumption that the scale is roughly constant within a relatively short segment.

After the division, we can use the energy minimization method to find the best scale segment by segment (The scales for different segments may be different) and align

Fig. 8. (a) The scale-space representation of peaks shown in Fig. 6. Here, the weighting coefficients are all set to be one. (b) The scale-space representation after considering the real weighting coefficients. We can observe 12 common peaks with larger weighting coefficients. (c) The first row shows locations of common peaks. The second row shows estimated locations using the weighting coefficient information. The third row shows estimated locations without using the weighting coefficient information. (d) This plot shows the distances of estimated group centers to common peaks normalized by the true locations of common peaks. The results are much improved comparing to the clustering method (cf. Fig. 7c).

TABLE 2
Performances of Different Versions of the Scale-Space Approach

Parameter	Expected Value	Without Weighting Information	With Weighting Information
μ_s	3.4764	2.2556	3.8678
σ_s	28.0380	46.7990	25.1673
peak number	20	65	20

peaks together. The spectrum alignment is then accomplished by combining alignment results from different segments. One remaining issue is that some peaks belonging to the same common peak near the segment boundary may be mistakenly separated into two neighboring segments. To solve this problem, we can use another set of segments with each segment centered at the boundary region (we may call these new segments as boundary segments) and use the alignment result in the middle of each boundary segment to replace the previous alignment result in the boundary region. This is similar to carrying out a Digital Fourier Transform (DFT) on a very long signal sequence, where we have to calculate the DFT piece by piece and where the above approach is commonly used. Certainly, we still need to study how to set the length of the segments and how to combine different segments.

Within the current framework, all peaks are treated individually. It might be possible that the

appearance/absence of combinations of some weak peaks are the most informative biomarkers. Feature extraction in higher-dimensional space, therefore, should be considered to explore this possibility, as described by some recent publications [27], [26].

A biomarker discovered in the MS data may correspond to many possible biological sources (i.e., a spectral peak may come from different proteins). Therefore, it is necessary to identify peptides and their parent proteins in order to fully understand the relation between protein structure and disease development. This understanding can also be very useful in drug design and development. Currently, tandem MS technique (MS/MS), coupled with database searching, has become the method of choice for the rapid and high-throughput identification, characterization, and quantification of proteins. It will be interesting to study how our peak alignment method can be applied to tandem MS data under such a setting.

ACKNOWLEDGMENTS

This work was supported with federal funds from NHLBI/NIH contract N01-HV-28186, NIDA/NIH grant P30 DA018343-01, NIGMS grant R01-59507, and US National Science Foundation grant DMS-0241160. The authors would like to thank Dr. Gusfield and Dr. Skiena and the anonymous reviewers for the very detailed comments, which greatly helped to improve the manuscript. The replicate data, ovarian cancer data, and the demonstration Matlab code are accessible at <http://bioinformatics.med.yale.edu/MSDATA>.

REFERENCES

- [1] J. Aach and G.M. Church, "Aligning Gene Expression Time Series with Time Warping Algorithms," *Bioinformatics*, vol. 17, no. 6, pp. 495-508, 2001.
- [2] C. Ambrose and G.J. McLachlan, "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data," *Proc. Nat'l Academy of Sciences of the United States of Am.*, vol. 99, no. 10, pp. 6562-6566, 2002.
- [3] T.S. Anantharaman, B. Mishra, and D.C. Schwartz, "Genomics via Optical Mapping II: Ordered Restriction Maps," *J. Computational Biology*, vol. 4, pp. 91-118, 1997.
- [4] J. Babaud, A.P. Witkin, M. Baudin, and R.O. Duda, "Uniqueness of the Gaussian Kernel for Scale-Space Filtering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 8, pp. 26-33, 1986.
- [5] K.A. Baggerly, J.S. Morris, J. Wang, D. Gold, L.C. Xiao, and K.R. Coombes, "A Comprehensive Approach to the Analysis of Matrix-Assisted Laser Desorption/Ionization-Time of Flight Proteomics Spectra from Serum Samples," *Proteomics*, vol. 3, pp. 1667-1672, 2003.
- [6] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [7] K.R. Coombes, H.A. Fritsche Jr., C. Clarke, J. Chen, K.A. Baggerly, J.S. Morris, L. Xiao, M. Hung, and H.M. Kuerer, "Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid by Surface-Enhanced Laser Desorption and Ionization," *Clinical Chemistry*, vol. 49, no. 10, pp. 1615-1623, 2003.
- [8] K.R. Coombes, S. Tsavachidis, J.S. Morris, K.A. Baggerly, M. Hung, and H.M. Kuerer, "Improved Peak Detection and Quantification of Mass Spectrometry Data Acquired from Surface-Enhanced Laser Desorption and Ionization by Denoising Spectra with the Undecimated Discrete Wavelet Transform," technical report, Univ. of Texas M.D. Anderson Cancer Center, 2004.
- [9] B. Efron and R. Tibshirani, "Improvements on Cross-Validation: The .632+ Bootstrap Method," *J. Am. Statistical Assoc.*, vol. 92, no. 438, pp. 548-560, 1997.
- [10] P.H.C. Eilers, "Parametric Time Warping," *Analytical Chemistry*, vol. 76, no. 2, pp. 404-411, 2004.
- [11] A. Gryff-Keller and S. Molchanov, "Systematic Discrepancy of Theoretical Predictions of NMR Chemical Shifts for Chlorinated Aromatic Carbons Using the GIAO DFT Method," *Molecular Physics*, vol. 102, no. 18, pp. 1903-1908, 2004.
- [12] K.J. Johnson, B.W. Wright, K.H. Jarman, and R.E. Synovec, "High-Speed Peak Matching Algorithm for Retention Time Alignment of Gas Chromatographic Data for Chemometric Analysis," *J. Chromatography A*, no. 996, pp. 141-155, 2003.
- [13] J. Listgarten, R.M. Neely, S.T. Roweis, and A. Emili, "Multiple Alignment of Continuous Time Series," *Advances in Neural Information Processing Systems 17*, Cambridge, Mass.: MIT Press, 2004.
- [14] N.V. Nielsen, J.M. Carstensen, and J. Smedsgaard, "Aligning of Single and Multiple Wavelength Chromatographic Profiles for Chemometric Data Analysis Using Correlation Optimised Warping," *J. Chromatography A*, vol. 805, pp. 17-35, 1998.
- [15] F. Nomura, T. Tomonaga, K. Sogawa, T. Ohashi, M. Nezu, M. Sunaga, N. Kondo, M. Iyo, H. Shimada, and T. Ochiai, "Identification of Novel and Downregulated Biomarkers for Alcoholism by Surface Enhanced Laser Desorption/Ionization-Mass Spectrometry," *Proteomics*, vol. 4, no. 4, pp. 1187-1194, 2004.
- [16] M.C. Papadopoulos, P.M. Abel, D. Agranoff, A. Stich, E. Tarelli, B.A. Bell, T. Planche, A. Loosemore, S. Saadoun, P. Wilkins, and S. Krishna, "A Novel and Accurate Diagnostic Test for Human African Trypanosomiasis," *The Lancet*, vol. 363, no. 9418, pp. 1358-1363, 2004.
- [17] E.F. Petricoin III, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohne, and L.A. Liotta, "Use of Proteomic Patterns in Serum to Identify Ovarian Cancer," *The Lancet*, vol. 359, no. 9306, pp. 572-577, 2002.
- [18] T.W. Randolph and Y. Yasui, "Multiscale Processing of Mass Spectrometry Data," *Univ. of Washington Biostatistics Working Paper Series, Number 230*, 2004.
- [19] G.A. Satten, S. Datta, H. Moura, A.R. Woolfitt, G. Carvalho, R. Facklam, and J.R. Barr, "Standardization and Denoising Algorithms for Mass Spectra to Classify Whole-Organism Bacterial Specimens," *Bioinformatics*, vol. 20, no. 17, pp. 3128-3136, 2004.
- [20] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q. Le, "Sample Classification from Protein Mass Spectrometry, by 'Peak Probability Contrasts,'" *Bioinformatics*, vol. 20, no. 17, pp. 3034-3044, 2004.
- [21] R.J.O. Torgrip, M. Aberg, B. Karlberg, and S.P. Jacobsson, "Peak Alignment Using Reduced Set Mapping," *J. Chemometrics*, vol. 17, pp. 573-582, 2003.
- [22] J.T. Wadsworth, K.D. Somers, L.H. Cazares, G. Malik, B.L. Adam, B.C. Stack Jr., G.L. Wright Jr., and O.J. Semmes, "Serum Protein Profiles to Identify Head and Neck Cancer," *Clinical Cancer Research*, vol. 10, no. 5, pp. 1625-1632, 2004.
- [23] M. Wagner, D. Naik, and A. Pothen, "Protocols for Disease Classification from Mass Spectrometry Data," *Proteomics*, vol. 3, no. 9, pp. 1692-1698, 2003.
- [24] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of Statistical Methods for Classification of Ovarian Cancer Using Mass Spectrometry Data," *Bioinformatics*, vol. 19, pp. 1636-1643, 2003.
- [25] Y. Yasui, M. Pepe, M.L. Thompson, B. Adam, G.L. Wright Jr., Y. Qu, J.D. Potter, M. Winget, M. Thornquist, and Z. Feng, "A Data-Analytic Strategy for Protein Biomarker Discovery: Profiling of High-Dimensional Proteomic Data for Cancer Detection," *Biostatistics*, vol. 4, no. 3, pp. 449-463, 2003.
- [26] J.S. Yu and X.W. Chen, "Bayesian Neural Network Approaches to Ovarian Cancer Identification from High-Resolution Mass Spectrometry Data," *Bioinformatics*, vol. 21, supplement 1, pp. i487-i494, 2005.
- [27] J.S. Yu, S. Ongarello, R. Fiedler, X.W. Chen, G. Toffolo, C. Cobelli, and Z. Trajanoski, "Ovarian Cancer Identification Based on Dimensionality Reduction for High-Throughput Mass Spectrometry Data," *Bioinformatics*, vol. 21, no. 10, pp. 2200-2209, 2005.
- [28] W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, and H. Zhao, "Detecting and Aligning Peaks in Mass Spectrometry Data with Applications to MALDI," *Computational Biology and Chemistry*, vol. 30, no. 1, pp. 27-38, 2006.

bioinformatics, computational biology, biomedical imaging, signal processing, and computer vision. He is a member of the IEEE Computer Society.

analyze the popular classification algorithm random forest and to improve it. He has served as referee for the *Journal of Bioinformatics and Computational Biology* and *Expert Review of Proteomics*.

Weichuan Yu received the PhD degree in computer vision and image analysis from University Kiel, Germany, in 2001. He was a postdoctoral associate at Yale University from 2001 to 2004. Currently, he is a research faculty member in the Center for Statistical Genomics and Proteomics at Yale University. He is interested in computational analysis problems with biological and medical applications. He has published papers on a variety of topics including

Xiaoye Li received the bachelor's degree from the University of Science and Technology of China in 2001 and the master of philosophy degree from Yale University in 2004. Currently, he is a PhD candidate in the Applied Mathematics Department at Yale University. He is interested in machine learning and statistical learning theory and the application of machine learning technique to various data mining problems, especially those arising from genomics and proteomics studies. His dissertation takes an initiative step to

tions in bioinformatics. He is a two-time recipient (2003, 2005) of the Institute of Mathematical Statistics Laha Award, which is granted to new researchers within two years after obtaining the PhD degree.

Baolin Wu received the Bsc degree from Peking University in 1999 and the PhD degree from Yale University in 2004. In 2004, he joined the Division of Biostatistics at the University of Minnesota as an assistant professor. His main research areas include computational biology and statistical learning.

Kenneth R. Williams received the PhD degree in biochemistry from the University of Vermont in 1976. Since then, he has been a postdoctoral associate, associate research scientist, research scientist, senior research scientist, and professor (adjunct) research in the Department of Molecular Biophysics and Biochemistry at Yale University. In 1980, Dr. Williams founded the Keck Laboratory (<http://info.med.yale.edu/wmkeck/>) and, in 1986, he was one of the six founding members of the Association of Biomolecular Resource Facilities (ABRF), an international organization that represents biotechnology resource laboratories (<http://www.abrf.org/>). He has 155 publications and he directs the Yale/NHLBI Proteomics Center (<http://info.med.yale.edu/nhlbi-proteomics/>), the Yale/NIDA Neuroproteomics Center (<http://info.med.yale.edu/nida-proteomics/>), and the Proteomics Core of the Northeast Biodefense Center.

Hongyu Zhao received the Bachelor's degree from Peking University in 1990 and the PhD degree from the University of California at Berkeley in 1995. He was an adjunct assistant professor and assistant professor in residence at University of California at Los Angeles from 1995 to 1996. At the Yale University School of Medicine, he was an assistant professor from 1996 to 2000 and is currently the Ira V. Hiscock Associate Professor of Public Health and Genetics. He is interested in developing statistical and computational methods to address scientific problems in molecular biology and genetics. He has published more than 120 articles in peer-reviewed journals and books, and is an associate editor for *Biometrics*, *Journal of Agricultural, Biological, and Environmental Statistics*, *Statistical Applications in Genetics and Molecular Biology*, *Pharmacogenomics*, and *Statistica Sinica*.

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.