Aligning Peaks Across Multiple Mass Spectrometry Data Sets Using A Scale-Space Based Approach

Weichuan Yu, Xiaoye Li, and Hongyu Zhao

Center for Statistical Genomics and Proteomics, Yale University

Introduction. In mass spectrometry (MS) based proteomics research, only the spectral peaks are biologically relevant in analysis applications such as biomarker discovery and disease classification. The number and locations of peaks in different spectra are very likely to be different given the noisy nature of biological samples and variations of MS data acquisition conditions. Aligning peaks across multiple spectra and determining their locations is therefore an important task as all comparisons and biologically significant conclusions are based on the alignment results. Due to the fact that the shift curves of different peaks are non-linear, however, peak alignment still remains challenging even after instrument calibration with internal markers. To address this problem, different methods have been proposed. When we have only two peak sets, dynamic programming [1] and parametric warping [2] based approaches are suitable. But when we have multiple peak sets, we cannot use the pair-wise peak alignment algorithm repeatedly to build a unique mutual correspondence among multiple peak sets since we do not know the standard peak set. If we arbitrarily choose one peak set as the standard, the final correspondence may vary when we use a different peak set as the standard, causing an ambiguity problem. In other words, the major difficulty in multiple peak alignment is to estimate the unknown number and locations of standard peaks. Recently, the hierarchical clustering method [3] has achieved certain degree of success in aligning multiple peak sets. But this method still has a few limitations: 1) We have to manually determine a cut-off parameter during the construction of a standard set. 2) The clustering result is sensitive to the existence of outliers. 3) The clustering method does not use intensity values, which is a valuable source of information. These limitations motivated us to propose a scale-space based approach [4] to tackle the multiple peak alignment problem.

Method. We considered detected peaks as observed values of the unknown standard peaks and assumed locations of detected peaks follow a Gaussian distribution with mean equal to the locations of standard peaks (This assumption is based on the central limit theorem, which is valid for relatively large number of spectra (such as 50)). We represented detected peaks as a set of Dirac components with different locations and weighting coefficients. Then, we convolved these peaks with a set of Gaussian functions (with zero mean and varying standard deviation σ) to form a two-dimensional scale-space representation (figure 2 left). The problem of estimating unknown number of sample means was then converted into a simpler problem of searching for local maxima in the scale-space representation. Once the scale parameter was estimated, we obtained a standard peak set by simply searching for the local maxima at the corresponding scale level and counting the number of these local maxima. A nice point of the scale-space approach is that we can use a parameter optimization scheme to avoid the manual determination of the scale parameter σ. Concretely, we minimized a distance-based energy function, which consists of a data-fitting term (which is the sum of squared distance between standard peak set components and detected peaks) and a regularization term (which reflects our belief on the most suitable value of σ and penalizes deviations from this value). Different forms of the regularization term can be used. For simplicity, here we assumed that the best scale follow a Gaussian distribution in the scale-space (i.e. there is no negative scale parameter) with its mean and variance determined by neighboring peak distances from the data. After estimating the standard peaks, the remaining problem of building a standard peak set-based mutual correspondence among multiple spectra can be solved using either dynamic programming, parametric warping, or even a simple closest point matching approach. Here we used a closest point matching approach.

Preliminary Result. We used a simple example with known ground-truth to compare the performance of our scale-space approach with a current hierarchical clustering method. Figure 1 shows the distribution of peaks from 50 peak sets with some noise points added (marked as red circles). Here we would check if the number and locations of 20 true peaks can be correctly estimated using both methods. For quantitative comparison, we also defined a measure called average distance as:

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

Here $d_{i,j}$ denotes the distance between the i-th peak in the j-th sample and its closest peak in the standard peak set, K_j indicates the peak number in the j-th sample, and M represents the total number of peaks in N samples. Intuitively, a smaller D_{av} indicates a better standard peak set. When we used the hierarchical clustering method, we had to manually determine the cut-off parameter and the result was sensitive to noise (figure1). In contrast, the scale-space approach provided reasonable results (figure 2). The average distance values in table 1 also gave an quantitative comparison.

Figure 1: Multiple peak alignment using the hierarchical clustering method. **Top Left:** Simulated peak distribution. Circles denote noisy points. **Top Right:** Cluster number vs. the cut-off height in the hierarchical clustering method. **Bottom Left:** Estimated cluster centers (circles) vs. ground truth (stars) when the cut-off height equals 15. A mistake is shown in the black box. **Bottom Right:** Estimated cluster centers (circles) vs. ground truth (stars) when the cut-off height equals 17. Two mistakes are shown in black boxes.

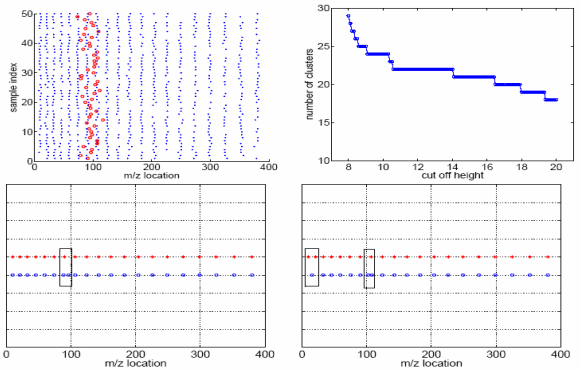

Figure 2: Left: Two-dimensional scale-space representation of all peaks. **Middle:** Energy function value vs. σ value. **Right:** Relative estimation error of 20 peak locations (normalized by the true locations).

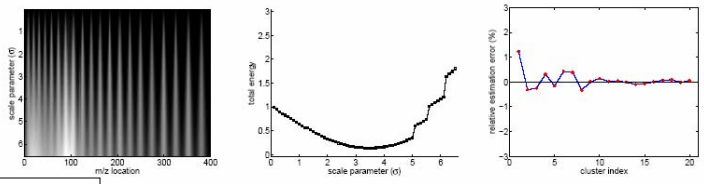

Table 1: D_{av} values for peaks shown in figure 1. While both the hierarchical clustering method and the scale-space approach provide similar D_{av} values for the peak distribution without noisy points (without circles), the scale-space provides a smaller D_{av} value if noisy points (circles) are added (row 2).

Simulation	Hierarchical Clustering	Scale Space
no noise added	1.1193	1.1181
noise added	1.4654	1.1368

Conclusion. We proposed a scale-space approach to automatically align multiple MS peak sets without manual parameter determination. It is more robust against noise than the hierarchical clustering method. In addition, it is possible to embed intensity information into the alignment framework, thus generalizing current approaches that use only the m/z information during the alignment of peaks. Our tests showed that this generalization brought some extra advantages for peak alignment, although we did not show concrete examples here due to the space limitation.

Acknowledgment

This work was supported with Federal funds from NHLBI/NIH contract N01-HV-28186, NIDA/NIH grant 1 P30 DA018343-01, NIGMS R01-59507, and NSF grant DMS-0241160.

Reference

[1] J. Aach and G.M. Church. Aligning gene expression time series with time warping algorithms. Bioinformatics, 17(6): 495-508, 2001.

[2] P.H.C. Eilers. Parametric time warping. Analytical Chemistry, 76(2): 404--411, 2004.

[3] 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, 20(17):3034--3044, 2004.

[4] T. Lindeberg. Scale-space theory in computer vision. Kluwer, Netherlands, 1994.