Douglas J. Slotta[1]
Lenwood S. Heath[1]
Naren Ramakrishnan[1]
Rich Helm[2]
Malcolm Potts[3]

[1]Department of Computer Science
[2]Department of Wood Science and Forest Products
[3]Department of Biochemistry
Virginia Polytechnic Institute and State University,
Blacksburg, VA, USA

Clustering mass spectrometry data using order statistics

Mass spectrometry data is inherently uncertain. Rather than compare peak heights across samples, a comparison can be made of the relative ordering of the peak height across samples. Order statistics are used to provide a distance metric between each ordered list of peak heights from the samples. A principal component analysis is performed on the set of distance vectors to highlight to important components.

Keywords: Mass spectrometry / Order statistics / Singular value decomposition PRO 0517

1 Introduction

Mass spectrometry data is inherently uncertain since the accuracy of any mass to charge ratio (*m/z*) determination is approximately 0.1%. The problem lies in determining which *m/z* denote the same peptide, and which *m/z* do not. It is possible that three samples exist, A, B and C, where A is within 0.1% of B, and B is within 0.1% of C, and yet A is not within 0.1% of C. In this case, which peptide is B equivalent to: A or C? The problem is further complicated in that two peptides could have the same *m/z* and still be different. In this case, the determination of equivalence cannot be made without further data.

Furthermore, once the equivalent peptides have been identified, peak values must be compared to determine the differences between the subjects. However, the magnitude of peak height from one sample to another may not be directly comparable. Rather than compare peak height to peak height, a comparison can be made between the relative orders of the peak heights. If samples A and B have *m/z* M and N with ($A_M<A_N$) and ($B_M>B_N$) (where A_M is the peak height for sample A at *m/z* ratio M and A_N is the peak height for sample A at *m/z* N and so forth) then M and N form an inversion pair. The relative orders of two samples can be measured by the total number of inversion pairs between the lists of *m/z*, ordered by the peak heights.

2 Materials and methods

Figure 1 shows a plot of all of the peaks found with a *m/z* between 4100 and 4600 for each subject. A cursory visual inspection reveals apparent vertical columns where the peptides are most likely the same. The first step in our datamining process is to link all of the samples together that could possibly be the same. To begin, all of the samples are sorted in order by their *m/z*. Then, each adjacent pair is checked to see if they are within the margin of error (0.1%) of each other. In this manner, chains of linked samples are formed where each item in the set is within the margin of error. The overall size of the chain may exceed the error margin. An example of this is shown in Fig. 2, which has the same data as Fig. 1, with the addition of a line drawn through each sample belonging to the same chain. The widest chain in the picture begins at *m/z* 418 4.472 and is 3.39% wide.

Given this list of all possible chains, the next step is to break up those chains that are larger than the margin of error into smaller, more reasonable chains. There exist vertical regions of the chart that contain a greater density of samples than adjacent areas. Statistically, these regions are more likely to be the same peptide, and those areas that are sparser contain either outliers or different peptides. If a count is taken at every sample that is within a certain margin of error, then the location of the vertical regions with the highest density is easily determined. The sliding window size chosen for Fig. 3 was 0.2%, since it is assumed that a *m/z* measurement will be accurate within ±0.1%. Figure 3 shows the count for the number of samples within the sliding window that begins at the current *m/z* for each chain. Peaks indicate those regions that contain the most samples and are at least 0.1% away from each other. The peaks are used to determine the final chains, or sets of like peptides. The results are shown in Fig. 4. For those experimental subjects that have the same peptide within different fractions, the *m/z* with the highest peak value was chosen as the representative peptide for that subject. A different selection method could be used based on more advanced knowledge of the fractionalization method used.

Correspondence: Douglas J. Slotta, Department of Computer Science (0106), Virginia Polytechnic Institute and State University, Blacksburg VA 24061, USA
E-mail: slotta@csgrad.cs.vt.edu
Fax: +1-540-231-6075

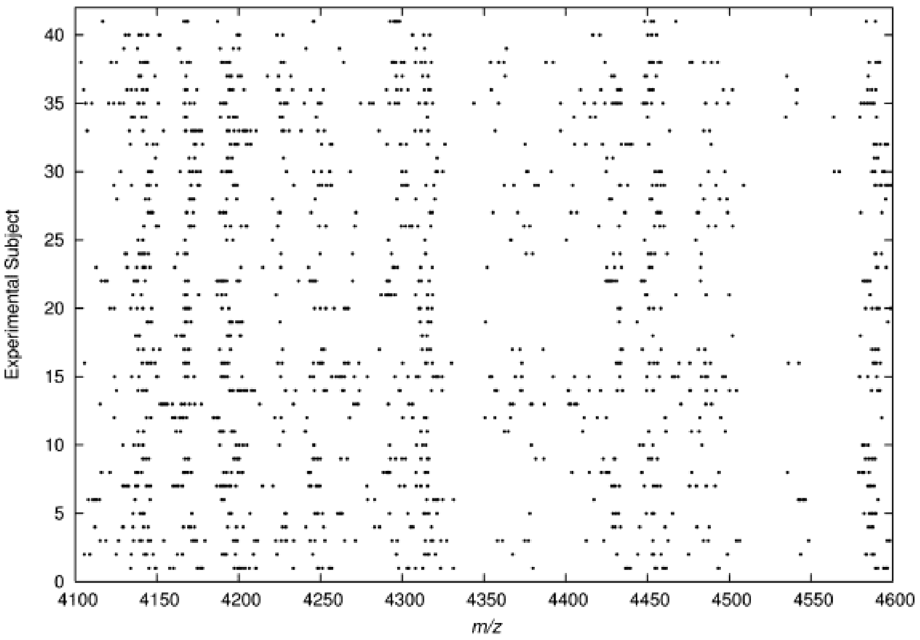

Figure 1. Initial data

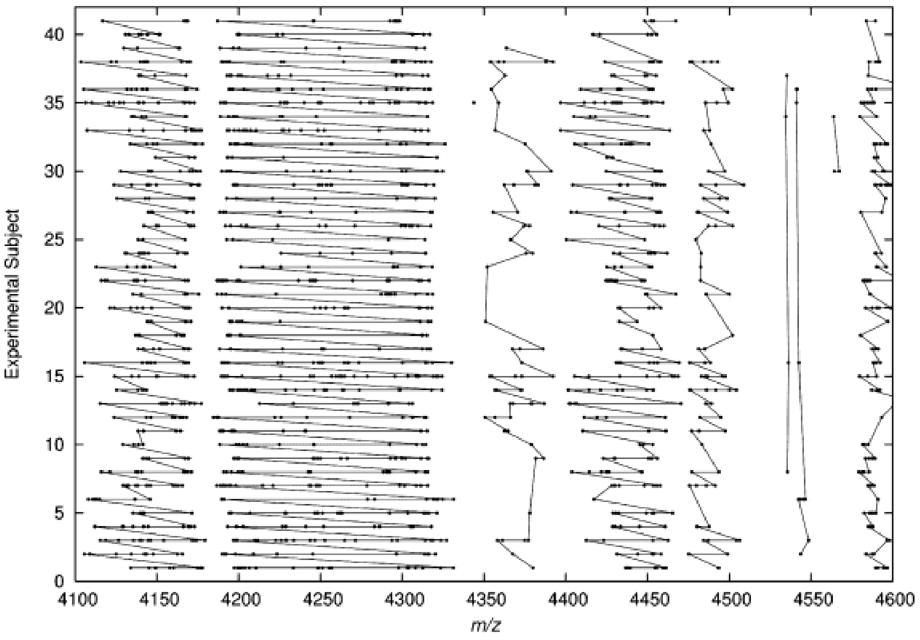

Figure 2. Preliminary chains

The first stage found 367 possible chains of more than two samples, 205 of which were greater than 0.2% wide. The final stage of the process broke these into 358 distinct chains, or peptides. The peptides of each experimental subject were formed into lists ordered by their peak height, normalized by adjusting for the difference of the medians of each fraction from the raw data. Pairwise comparisons were conducted by counting the number of inversions between ordered pairs amongst the two lists. The possible inversion counts for all permutations of a list follow a normal distribution [1]. By comparing the resulting inversion count to the expected inversion count of a random ordering of the same list, the likelihood of correspondence can be computed [2].

Figure 5 shows the results of the pairwise comparisons between the ordered lists. The lighter the value, the more alike the order of the list, the darker, the less alike the two

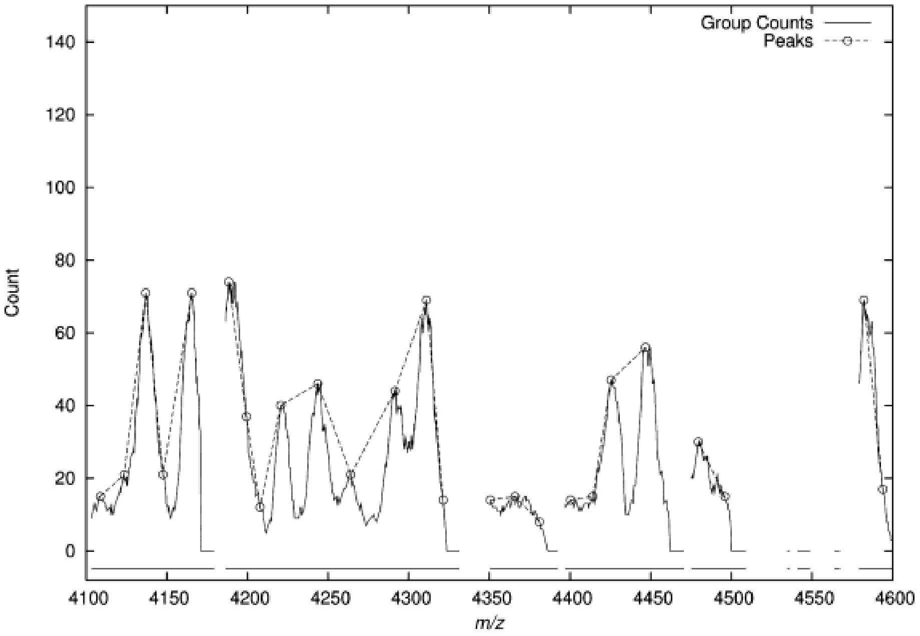

Figure 3. Sliding window density

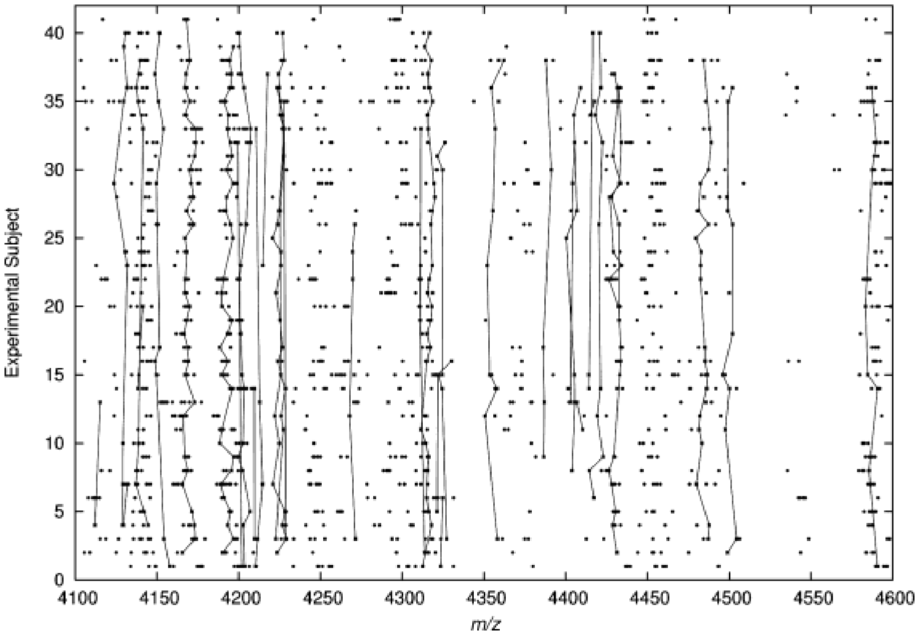

Figure 4. Final chains

lists are. Therefore, the boxes on the main diagonal where each list is compared to itself, are all white. Figure 6 shows the results of performing a principal components analysis on the data of Fig. 5 and subsequent reconstruction, where all but the highest two values in the singular values from Singular Value Decomposition (SVD) [3, 4] were set to zero.

3 Results

Figure 6 shows four clusters of light valued cells along the main diagonal. X01 and X02 form one cluster, X03 through X26 form the second cluster, X27 through X32 form the third, and X33 through X41 form the final cluster. The peptides responsible for the differences between the

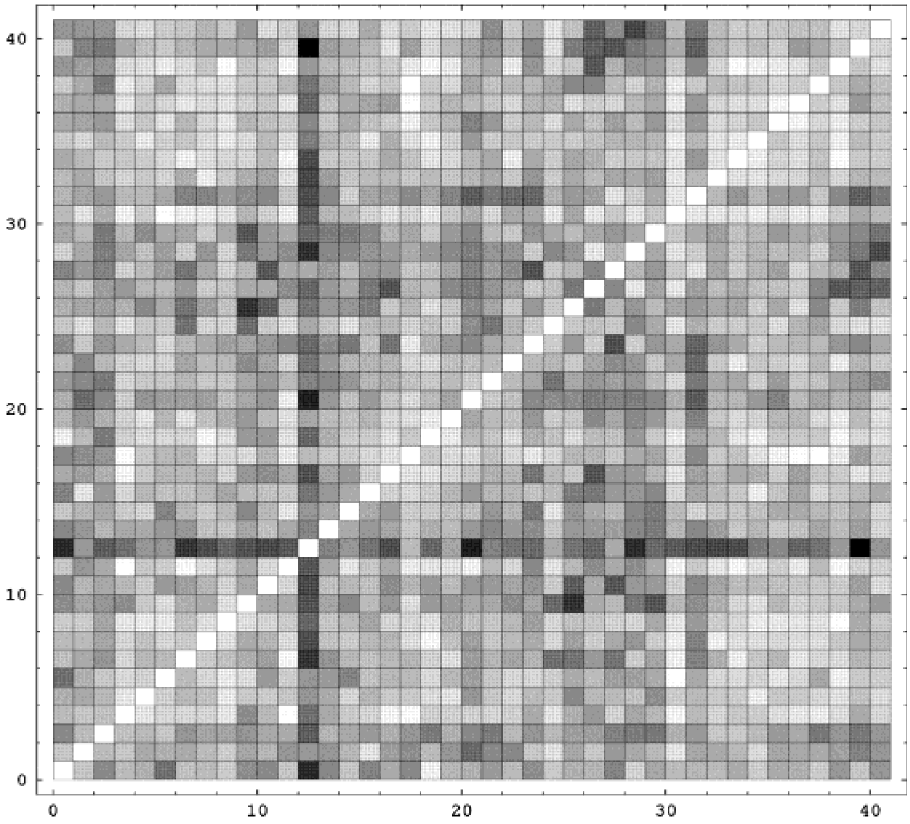

Figure 5. Pair wise list comparisons

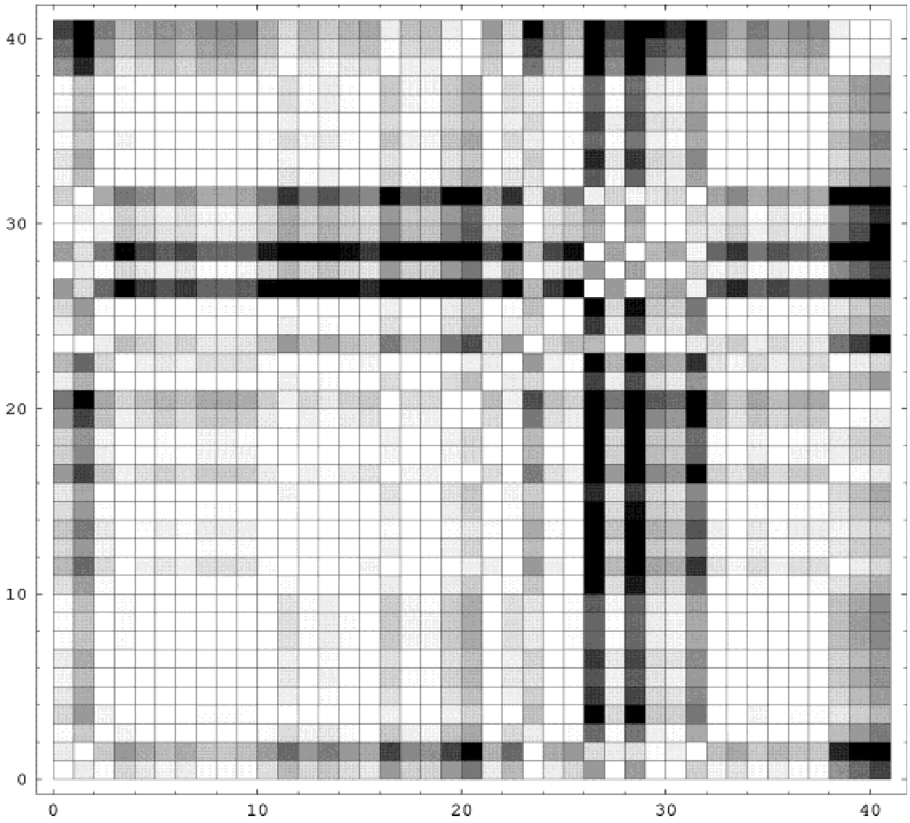

Figure 6. Post SVD reconstruction

experimental subjects can be determined by looking for those peptides with the highest individual inversion count. The *m/z*'s with the ten highest inversion counts in this study are: 66433, 15127, 15873, 6642, 8135, 7577, 28058, 5386, 9656 and 33293.

4 Concluding remarks

The methods outlined in this extended preliminary study show promise for future investigation. The results are dependent on precisely matching up the peptides according to their *m/z*, but not on comparing peak height across the subjects; only the relative order of the peak heights within each subject is needed. One immediate prospect is to use order-theoretic notions to do feature selection (*i.e.*, identify the most relevant features of the dataset that capture the ordinalities in the original dataset) [5]. This will allow us to information-theoretically quantify the usefulness of various subsets of *m/z* for characterizing and clustering proteomics data. Our methods can also be used for defining application-specific distance measures for model-based clustering [6] and graph partitioning [7].

This work is supported in part by ONR MURI grant N00014-01-1-0852 and NSF grant EIA-0103660.

Received October 15, 2002

5 References

[1] Knuth, D. E., *The Art of Computer Programming*, Vol. 3, *Sorting and Searching*, 2nd Ed., Addison Wesley, NJ 1998, pp. 11–22.

[2] Slotta, D. J., Heath, L. S., Ramakrishnan, N., Helm, R., Potts, M., *Computational Approaches to Combining Predictive Biological Models*, Proc. High Performance Computing Symposium, Tenter, A. (Ed.); Advanced Simulation Technologies Conference, Soc. Computer Simulation Intl., San Diego, CA, 2002, pp. 75–80.

[3] Aeter, O., Brown, P. O., Botstein, D., *Proc. Nat. Acad. Sci. USA* 2000, *97*, 10101–10106.

[4] Golub, G. H., VanLoan, C. F., *Matrix Computations*, 3rd ed., John Hopkins University Press, Baltimore 1996.

[5] Koller, D., Sahami, M., *Proc. Thirteenth Int. Conf. Machine Learning*, Bari, Italy 1996, pp. 284–292.

[6] Joyner, I., Cook, D. J., Holder, L. B., *J. Machine Learning Res.* 2001, *2*, 19–43.

[7] Karypis, G., Han, E.-H., Kumar, V., *IEEE Computer* 1999, *32*, 68–75.