Identification of Contaminants in Proteomics Mass Spectrometry Data

M. Duncan, K. Fung
University of Colorado Health Sciences Center

H. Wang, C. Yen and K. Cios
University of Colorado at Denver

Krys.Cios@cudenver.edu

Abstract

This paper discusses the identification of potential contaminants in mass spectrometry data derived from proteomic studies. Contaminant masses are usually submitted with valid peptide masses to the protein identification algorithms which can potentially lead to false positive results. In this paper we present an approach for the automatic identification of contaminant masses so that they can be removed prior to the submission of the peak list for protein identification. For this purpose we have developed an algorithm that clusters mass values. We calculate the frequencies of all masses and then identify possible contaminant masses. We propose that masses that occur with high frequency are contaminants. In our analysis of 78,384 masses derived from 3,029 proteins, we identify 16 possible contaminants. Of these 16, four are known trypsin autolysis peptides. Removing these contaminant masses from the database search will lead to more accurate and reliable protein identification.

1. Introduction

Mass spectrometry (MS) is a widely used method for protein identification. Peptide mass fingerprinting is the protein identification technique where MS is employed to measure peptide masses generated by enzymatic digestion of proteins. The mass of each peptide detected is then submitted to a protein identification algorithm, e.g., Mascot or MS-Fit, which attempts to match the experimentally determined masses to those generated *in silico*. The strategy is hampered, however, as both the peptide masses and the masses of contaminants present in the sample are often submitted to the algorithm. There are many possible contaminant sources such as chemicals used during the sample preparation process, keratin (a ubiquitous protein found in skin and hair), and chemicals used to visualize proteins before they are excised from the 2D gel. In this paper we present an approach for identification of contaminant masses.

2. Methods

Data were collected at the Colorado Center for Innovative Proteomics at the University of Colorado Health Sciences Center. Pre-processing of the data included baseline correction, noise reduction, de-isotoping and mass calibration to known trypsin autolysis peptides. Peak lists were sorted by mass values and the data set on which we performed analysis included 3,029 proteins with 78,384 distinct mass values.

In order to identify possible contaminant masses, the mass values were clustered, the optimal clustering setting was determined and the frequency of each cluster was calculated. Clustering is the major unsupervised data mining tool that is used in this research. First, a similarity measure was designed. Second, a clustering algorithm was developed to cluster the data with different cluster settings that represent different radii. Third, the validity of clustering was verified because the number of clusters was unknown. Finally, the cluster centers were treated as target mass values for possible peptides allowing the frequencies of the peptide masses to be calculated. High frequencies of the peptide masses are good indicators that they are contaminants.

The similarity measure used to calculate the distance was the normalized distance formula:

$$D = \frac{|M_{\exp} - M_{known}|}{M_{known}} \times 10^6$$

where D was measured in terms of PPM (part per million), M_{known} were the real values of masses (such as the known trypsin masses), and M_{exp} were the values of masses obtained experimentally. Using the function shown above, a clustering algorithm was designed based on the "tentative add" idea. The preprocessed data, Q, is one-dimensional and sorted by mass value. The idea is to fetch the first mass value, M, from Q and try to add it into the current cluster, C. The radii of the new cluster, C', which contains the masses of C and M, is calculated from the cluster center to the left and right boundaries. If it is less than or equal to R, M is treated as a member of the current cluster. Otherwise, the leftmost mass, L_1 of C, is removed, based on the fitness evaluations of M and L_1. If the radius of a new cluster, whose boundary from the second leftmost mass L_2 to M is smaller than the radius of C, M would be added into C and L_1 would be removed from the cluster. This process is repeated until the new cluster is formed. Each mass removed from C becomes a new M and undergoes the same process as M. The process

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**

of evaluating each data point runs recursively. The other case is when M is less fit than L_1, and therefore belongs to another cluster (a new cluster) composed of only one mass, M. After re-clustering of all clusters, the next mass value is fetched from Q and the process is repeated for M until all mass values are processed.

After all mass values are clustered we use this validity measure:

$$ErrorPPM = \frac{|M_c - M_{mean}|}{M_c} \times 10^6$$

where M_c is the known mass of trypsin fragments and M_{mean} is the mean of each generated trypsin cluster. We calculate the mean of each cluster as:

$$M_{mean} = \frac{\sum(Mass \quad Value \quad \times \quad Frequency)}{\sum Freqeuncy \quad of \quad Mass \quad Value}$$

where the frequency is either 1 for a not-weighted method, or the real frequency of the constituent mass value for a weighted method. In a weighted method, we take into account the frequency with which a given mass appears in the data. The smaller the error (PPM), the better the accuracy. This validity measure was used as one criterion to determine the optimal setting of the clusters. The other measure was the ambiguity analysis. After peptides were clustered, the frequency of each peptide in the data was calculated. The following formula was used:

$$Frequency(\%) = \frac{Frequency \quad of \quad Prototype}{Number \quad of \quad Proteins} \times 100\%$$

3. Results and discussion

The optimal cluster setting is defined as the radius of clusters for which the masses are clustered without ambiguity (see the definition below). The optimal cluster radius using the defined optimal cluster setting was found to be 30 PPM. This was determined by combining the criteria of the lowest possible clustering ambiguity and the smallest average error rate, using as a reference the four known trypsin masses.

The ambiguity is defined as the similarity between two or more clusters within a distance that is less than 50 PPM. We sorted the top 50 clusters with the highest frequencies by mass values from lowest to the highest, and then calculated the distance between each of two clusters. We found that the radius greater than or equal to 30 PPM had no ambiguity. The quality of clustering was also evaluated using error rates of the four known trypsin masses. The average error rate was calculated by the sum of the cluster centers minus the theoretical values of the four trypsin masses, and then divided by 4. We found that at 30 PPM, the ambiguity of clustering reduces to 0, while the average error rate is lowest in the range of 30 – 50 PPM.

4. Conclusions

We have presented a clustering method that shows that 30 PPM is the optimal cluster setting. Based on this setting, sixteen masses were identified as probable contaminants, occurring with frequencies greater than 20% above the threshold. Four of those 16 masses are known trypsin masses, which leaves 12 possible contaminant masses that should be eliminated before the data are submitted to a search engine for identification. This would improve matching accuracy. Second, we presented a hypothesis for determining contaminant masses by using the frequency of those mass values. Although this study was based on one experimental protocol (MALDI-TOF) it can be extended and applied to determine possible contaminant masses when alternative MS experiments are conducted.

5. References

[1] D.C. Liebler. *Introduction to Proteomics*. Humana Press, Totowa, NJ, 2002.

[2] D. Perkins, D. Pappin, D. Creasy, J. Cottrell. Probability-based Protein Identification By Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* 1999, 20, 3551-3567.

[3] D.F. Hochstrasser, J. Sanchez, R.D. Appel. Proteomics and its trends facing nature's complexity. *Proteomics* 2002, 2, 807-812.

[4] K. Cios, W. Pedrycz, R.W. Swiniarski. *Data Mining Methods for Knowledge Discovery.* Kluwer Academic Publishers, Norwell, MA, 1998.

[5] K. Fung, D. Friedman, M.W. Duncan. Identification of the peptide and protein constituents of human seminal fluid. *Proceedings 49th ASMS Conference on Mass Spectrometry and Allied Topics*. Chicago, Illinois; 2001.

[6] T. Rabilloud. Detecting Proteins Separated By 2D Gel Electrophoresis. *Analytical Chemistry*. 72, 48A-55A. 2000.

[7] K.J. Cios, (ed.). 2001. Medical Data Mining and Knowledge Discovery. Springer Physica-Verlag.

[8] J.R. Yates. Mass Spectrometry and the Age of the Proteome. J. *Mass. Spectrometry.* 33, 1-19. 1998.

[9] A. Cerpa-Poljak, M.W. Duncan. 1998. Amino acid analysis of peptides and proteins on the Femtomole scale by gas chromatography/mass spectrometry. *Analytical Chemistry*. 70:890-896.