

# Identification of Contaminants in Proteomics Mass Spectrometry Data

M. Duncan<sup>1</sup>, K. Fung<sup>1</sup>, H. Wang<sup>2</sup>, C. Yen<sup>2</sup> and K. Cios<sup>2</sup>,  
University of Colorado Health Sciences Center<sup>1</sup>  
University of Colorado at Denver<sup>2</sup>  
Denver, Colorado, U.S.A.

## Abstract

Mass spectrometry (MS) is a widely used method for protein identification. Peptide mass fingerprinting is the protein identification technique in which MS is employed to determine the masses of peptide fragments generated following enzymatic digestion of proteins. The masses of peptides are then submitted to a recognition program, e.g., MASCOT or MSFIT, for identification of a protein. The strategy is hampered, however, because not only are the peptide masses determined, but also the masses of multiple contaminants that are also present in the sample. Although the masses of some common and known contaminants are removed (e.g., peptides generated by trypsin autolysis), many others are inadvertently incorporated into the analysis. In this paper we present an approach for automatic identification of contaminant masses so that they can be removed prior to the identification process. For this purpose we have developed an algorithm that clusters mass values. We calculate the frequencies of all masses and then identify contaminants. We propose that masses with frequency higher than a given value are contaminants. In our analysis of 3,029 digested proteins, yielding 78,384 masses, we identified 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.

## Key words

Protein identification, mass spectrometry, contaminants, molecular weights, peptide mass fingerprinting, peptide mass mapping, clustering.

## Authors' mailing and email addresses

Mark Duncan	UC HSC	Mark.Duncan@uchsc.edu
Kim Fung	UC HSC	Kim.Fung@uchsc.edu
Heng Wang	UC Denver	Paul.Wang@cudenver.edu
Chia-Yu Yen	UC Denver	cyen@ouray.cudenver.edu
Krzysztof Cios	UC Denver	Krys.Cios@cudenver.edu

# 1 Introduction

Proteins are the end products of genes and serve as critical structural and functional components in cells and tissues. Changes in protein expression can result in disease. By determining differences in protein expression – an exercise now termed comparative proteomics - can lead to a better understanding of health and disease and yield improvements in accurate diagnosis and treatment.

Protein identification based on the methods of proteomics employ a small subset of the available data – the masses of a subset of the tryptic peptides – in the interpretation process. Under these conditions it is common for critical information to be lost leading investigators to make erroneous conclusions. Our goal is to identify contaminant masses because if these can be identified and removed from consideration, the probability of obtaining an accurate match markedly improves. Contaminant data are submitted with valid data for protein analysis, thereby skewing results.

This paper is organized as follows. First, we provide the theoretical foundation for the study. Second, we describe the method used in conducting the study and third, we present and discuss the results. We end with conclusions pertinent to our findings.

## 1.1 Proteins and amino acids

Proteins are comprised of 20 naturally occurring amino acids joined together by peptide bonds. A three-letter or single-letter code is used to represent each amino acid: e.g., “Lys” or “K” represents lysine, and “Arg” or “R” represents arginine, etc. Currently, there are over one million known proteins in the protein databases ranging in size from those comprised of just a few amino acids (usually referred to as peptides) to those composed of over 30,000 amino acids. Enzymes, such as trypsin, can break down other proteins into fragments called peptides. In the case of trypsin, cleavage is sequence specific and occurs after (i.e., on the C-terminal side) every K or R residue.

## 1.2 Protein identification process

Protein identification is generally based on the following process:

- Complex protein mixtures are separated based on physical techniques such as two-dimensional gel electrophoresis.
- The protein is removed from the gel and digested using an enzyme, usually trypsin.
- The masses of the tryptic fragments are experimentally determined by mass spectrometry.
- For each known protein (i.e., all entries in the protein database) the computer performs an *in silico* digestion to yield the set of predicted masses for each entry.
- The experimentally determined molecular weights for each protein are compared to the predicted molecular weights of every protein in the database to determine the best match.
- Possible protein IDs are sorted based on a number of criteria including how well the predicted molecular weights match the measured molecular weights, average mass errors and number of peptides matching.

### 1.3 Contaminants

There are two complicating factors which need to be considered in protein identification: first, all measured molecular weights have associated errors, both random and systematic; second, contaminants are invariably present and their molecular weights are measured and added to the list of protein-derived peptide masses against which the match will be made.

The presence of contaminant masses during database searches increases the probability of both false positive and false negative results. Possible contaminant sources are:

- Chemicals and contaminants present in the sample (or plastic ware) used in protein preparation
- Keratin (i.e., a protein found in skin and hair) and other protein contaminants
- Chemicals used to visualize proteins before they are excised from the 2D gels.

Although the identity of many of the sample contaminants is unknown, they are observed in a large number of different samples. The presence of the same masses in many samples is a good indicator that those are contaminants. Once we know what the contaminant masses are, they can be filtered out and only the remaining masses submitted for database searching.

## 2 Methods

### 2.1 Preprocessing

Data were collected by the Proteomics Lab at the University of Colorado Health Sciences Center. They are in the form of a peak list of measured peptide masses. Pre-processing of the data included baseline correction, noise reduction, de-isotoping and mass calibration to known trypsin autolysis peptides. This results in the elimination of peaks that clearly represent noise. Data analysis was performed by the UCHSC Proteomics Facility. Peak lists were sorted by mass values and the resulting data set included 3,029 proteins with 78,384 distinct mass values.

### 2.2 Data Mining

In order to identify possible contaminant masses, the mass values were clustered, the optimal clustering setting determined and the frequency of each cluster calculated. Clustering was the major unsupervised data mining tool that was 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.

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

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

where  $D$  was measured in terms of PPM (part per million),  $M_{\text{known}}$  were the real values of masses (such as the known trypsin masses), and  $M_{\text{exp}}$  were the values of masses obtained

experimentally. An objective function  $D = R$  was used as a clustering condition where radius  $R$  was a parameter of this algorithm that controlled the size of clustering.

The reason for choosing the normalized similarity measure was that errors in mass assignment were directly proportional to the mass. This error is normalized by the respective PPM value. Using the measurement functions shown above, a clustering algorithm was designed based on the “try to add” idea: because the preprocessed data are one-dimensional and sorted by mass value, this set is treated as  $Q$ . The concept 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  $M$ , are calculated from the cluster center to the left and right boundaries. If they are less or equal to  $R$ ,  $M$  is treated as a member of the current cluster. Otherwise, the leftmost mass,  $L_1$  of  $C$ , is removed. This is 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 comparison is done until the new cluster is formed. For all masses removed from  $C$ , each becomes another  $M$  and undergoes the same process as  $M$ . The process of evaluating each data point runs recursively. The other case is that  $M$  is less fit than  $L_1$ , and therefore belongs to another cluster (i.e., a new cluster) composed of only one mass,  $M$ . After re-clustering of all existing 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  $M_c$  are the known masses of trypsin fragments and  $M_{mean}$  is the mean of each generated trypsin cluster. We calculate the mean of each cluster as:

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

where the frequency is either 1 for a not-weighted method, or the real frequency of the constituent mass value for a weighted method. The smaller the PPM error, the better the accuracy is. This validity measure is used as one criterion to determine the optimal setting of the clusters. The other measure is the ambiguity analysis.

### 2.3 Contaminant masses identification

After peptides are clustered, the frequency of each peptide in the data is calculated. The following formula is used:

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

Mass values with high frequencies are probable contaminant masses. Later we will discuss in more detail what frequency is considered to be high enough for the peptides to be considered contaminants, and how many peptide mass values are chosen based on this frequency.

### 2.4 Implementation

We have developed several programs in JAVA to perform functions like merging peak list files, removing redundant protein data, organizing peptide mass data as a one-

dimensional ranked list, sorting the data by the mass value, clustering the sorted data, and calculating the frequencies of the prototypes.

### 3 Results and discussion

As described, we cluster peptide mass values using the weighted method, where the frequency of each mass value is taken into account in calculating cluster centers. In contrast, in the not-weighted method the frequency is not used as a factor in calculating cluster centers. Another approach we used was to remove the exact values of known trypsins from the data and then cluster the mass values. The analysis of the results from the four methods shows that the outcomes are very similar (see Table 1).

	Without Trypsin Masses		With Trypsin Masses	
	Not-weighted	Weighted	Not-weighted	Weighted
<b>Number of Proteins</b>	3029	3029	3029	3029
<b>Number of Masses</b>	78380	78380	78384	78384
<b>Number of Clusters</b>	13946	14002	13946	14002

Table 1

The clustering results show no significant difference between the two settings of weighted vs. not-weighted, so we will only discuss the weighted results that include two settings: weighted with trypsin masses and weighted after removing trypsin masses.

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 just 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 factor and average error rate are discussed below.

#### 3.1 Ambiguity

The ambiguity is defined as the similarity between two or more clusters within a distance that is less than 50 PPM, because the maximum experimental mass error for these experiments was defined as 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 the two clusters. Figure 2 shows that with different cluster settings, from 1 PPM to 100 PPM, the minimum distance for each pair increases from 29 PPM to 30 PPM, and then jumps to another higher level from 30 PPM to 31 PPM. From this we conclude that the radius greater or equal to 30 PPM has no ambiguity, *i.e.*, 30 PPM is the minimum clustering setting with no ambiguity of clustering.

#### 3.2 Average error rate

The quality of clustering is also evaluated using error rates of the four known trypsin masses. The average error rate for is calculated by the sum of the cluster centers minus the ideal values of the four trypsin masses, and then divided by 4. Figure 3 shows the average error rate from 1 PPM to 100 PPM. From this graph we see that the average error rate increases from 1 PPM to approximately 40 PPM. Since cluster settings greater than 50 PPM were not used in this study, the settings from 50 PPM to 100 PPM are shown

only for reference. We have therefore focused on settings in the range 0 – 50 PPM. In this range we found that at 30 PPM the ambiguity of clustering reduces to 0, while the average error rate is lowest between 30 – 50 PPM.

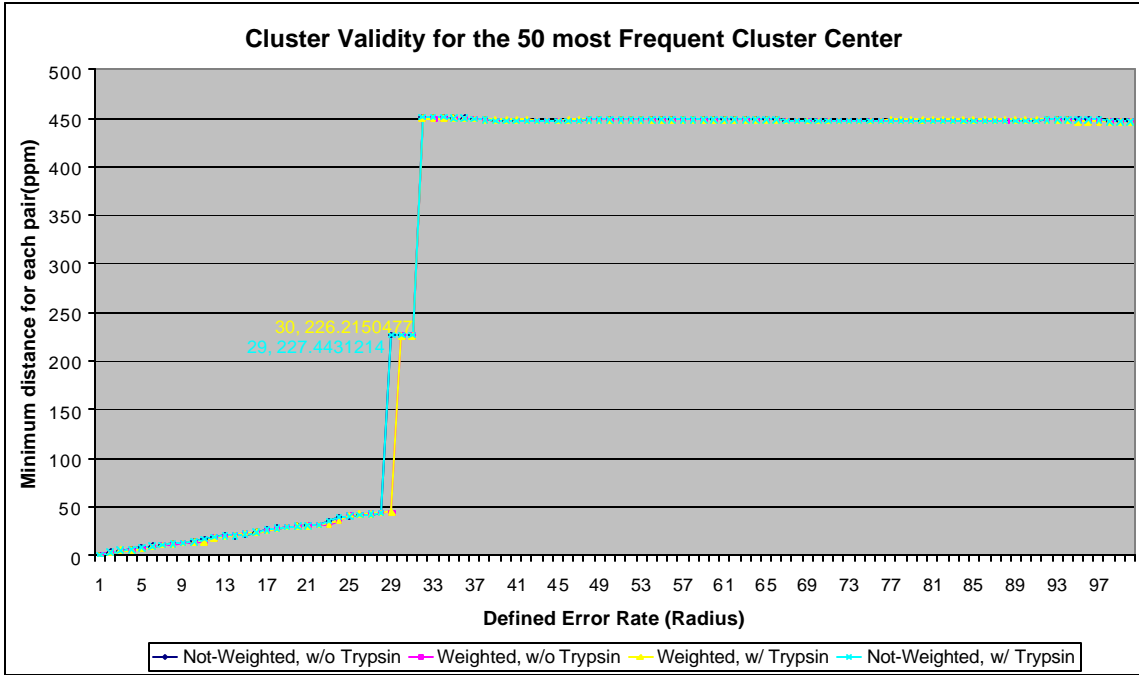


Figure 2

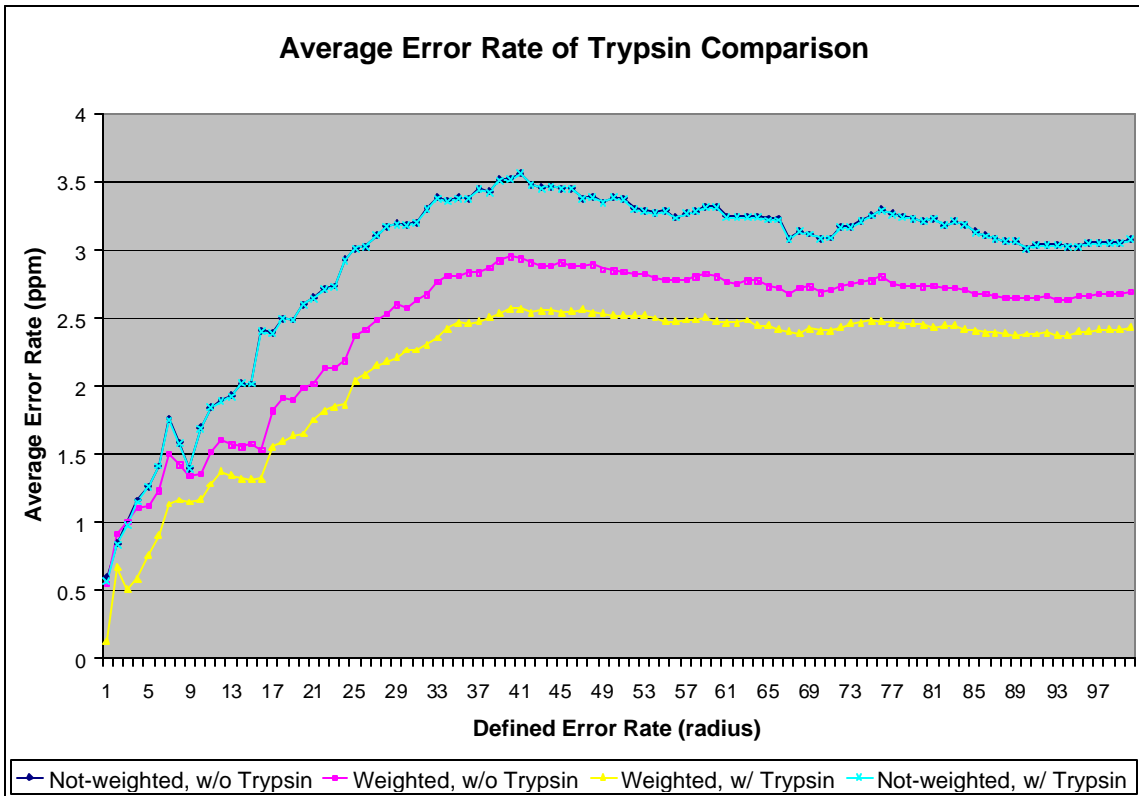


Figure 3

Figure 4 shows the top 50 clusters with the highest frequencies. The cluster number is ordered by the mass values of the cluster centers. The highest frequency is 85.8%, which occurred at  $m/z$  842.5099; the lowest frequency is 6.5%, which occurred at  $m/z$  843.0798.

Based on the frequencies of the mass values shown in Figure 4 we propose that the masses with high frequencies can be treated as possible contaminants. The threshold for identifying potential contaminants was defined at 20%.

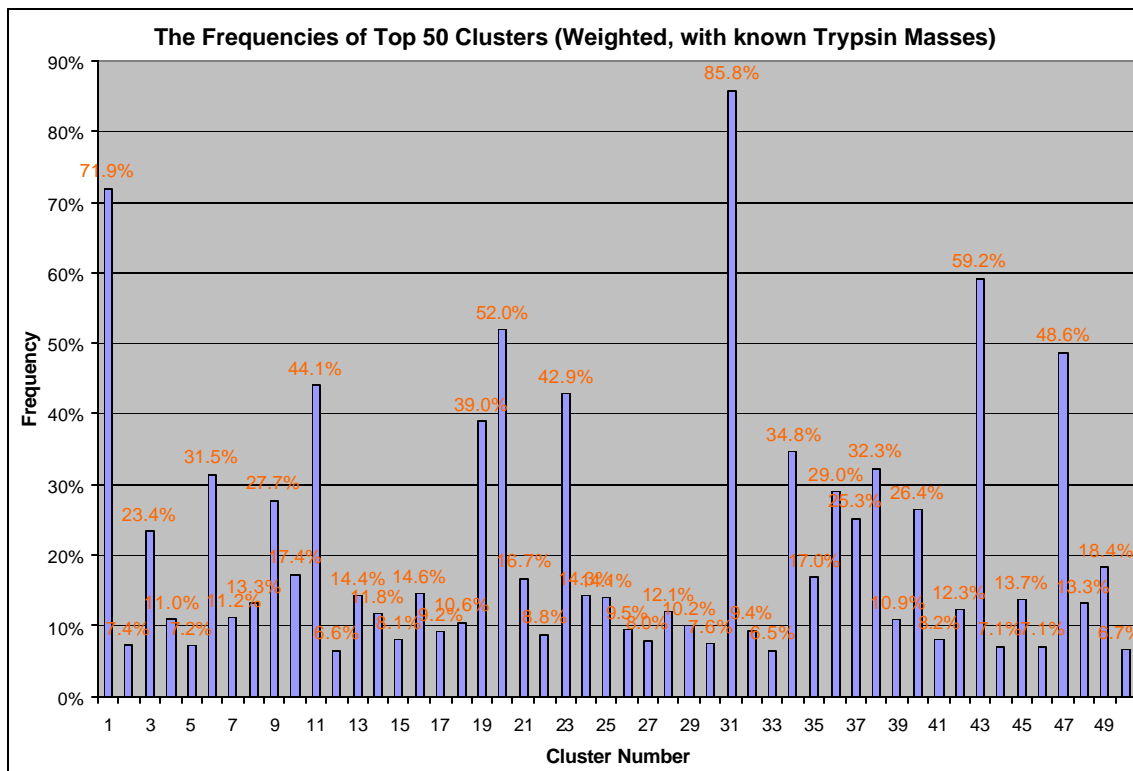


Figure 4

Figure 5 and 6 show 16 masses that occur with a frequency greater than 20%. Figure 5 is based on the weighted setting without known trypsin masses, while Figure 6 is based on weighted setting with known trypsin masses.

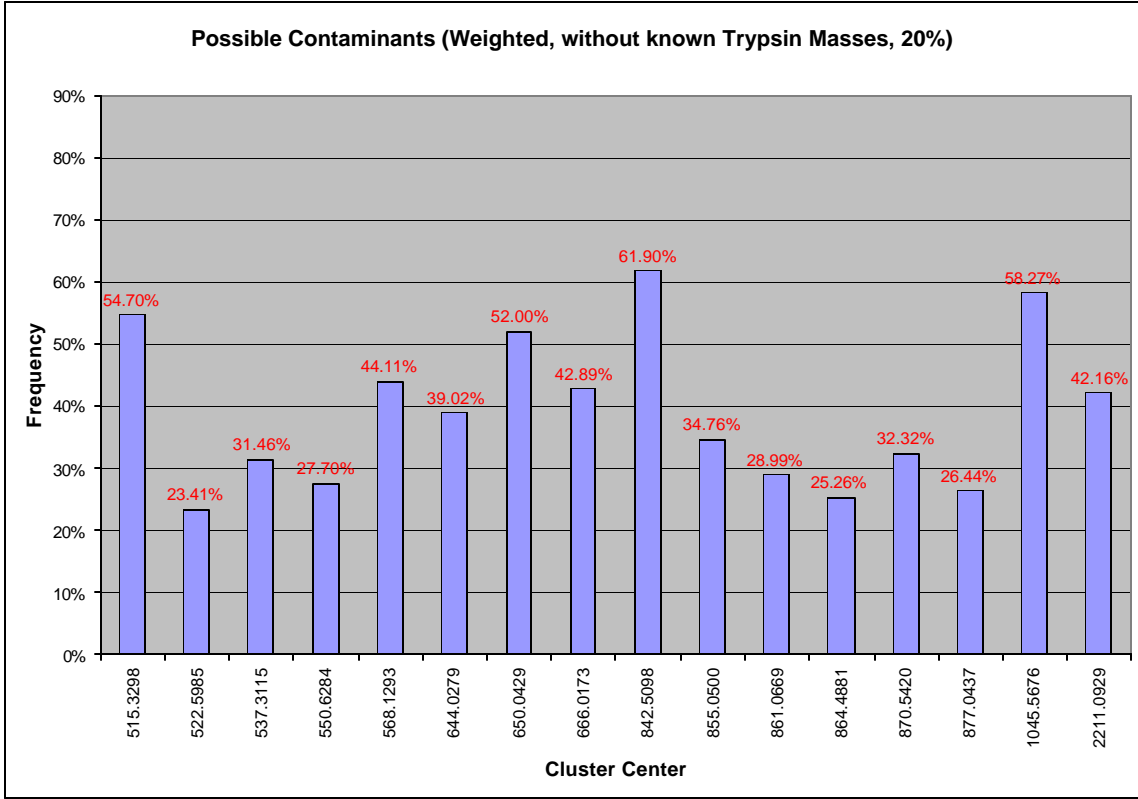


Figure 5

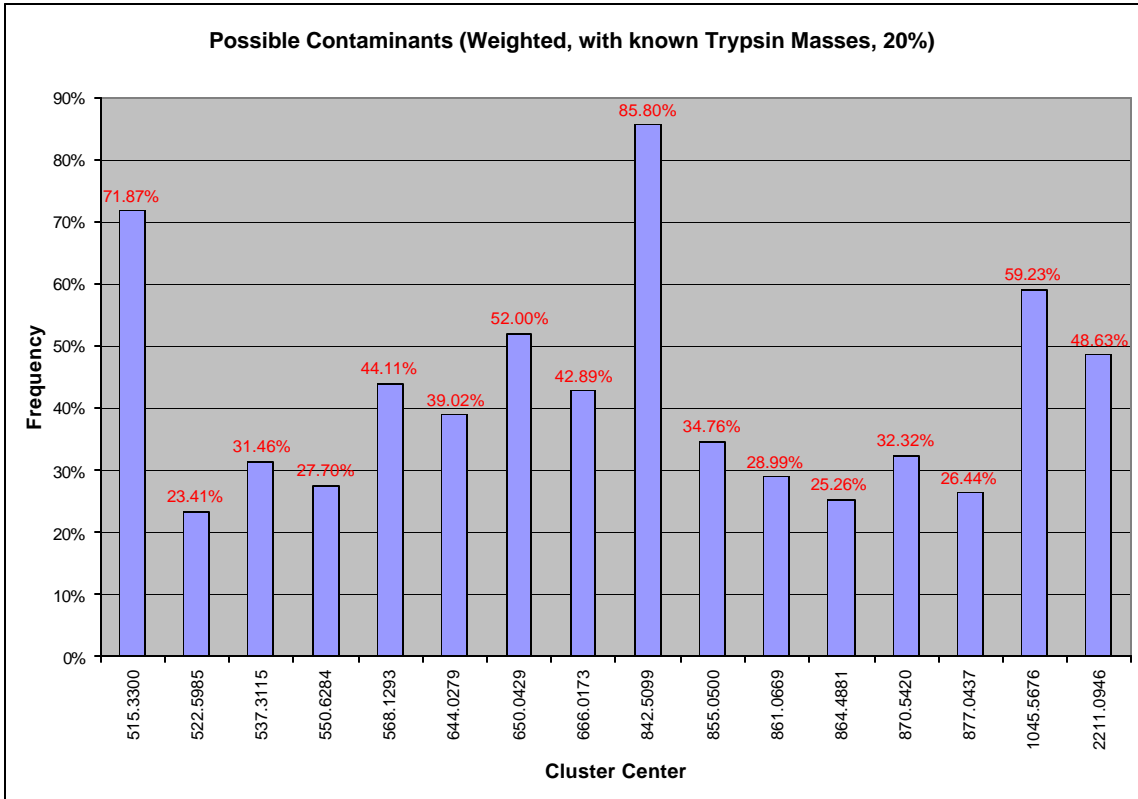


Figure 6

Table 2 and Figure 7 shows that the suspected contaminant masses are the same regardless of which method is used.

No	Without Trypsin Masses		With Trypsin Masses		StdDev
	Not-weighted	Weighted	Weighted	Not-weighted	
1	515.329343	515.329817	515.330004	515.329345	0.000335
2	522.597728	522.598473	522.598473	522.597728	0.000430
3	537.311673	537.311515	537.311515	537.311673	0.000091
4	550.628080	550.628417	550.628417	550.628080	0.000195
5	568.129540	568.129322	568.129322	568.129540	0.000126
6	644.027929	644.027945	644.027945	644.027929	0.000012
7	650.042975	650.042859	650.042859	650.042975	0.000067
8	666.017431	666.017338	666.017338	666.017431	0.000053
9	842.509444	842.509794	842.509851	842.509445	0.000219
10	855.050019	855.049958	855.049958	855.050019	0.000035
11	861.067085	861.066872	861.066872	861.067085	0.000123
12	864.487895	864.488094	864.488094	864.487895	0.000116
13	870.541746	870.542039	870.542039	870.541746	0.000169
14	877.043596	877.043721	877.043721	877.043596	0.000073
15	1045.567934	1045.567608	1045.567553	1045.567930	0.000204
16	2211.091197	2211.092931	2211.094642	2211.091213	0.001646

Table 2

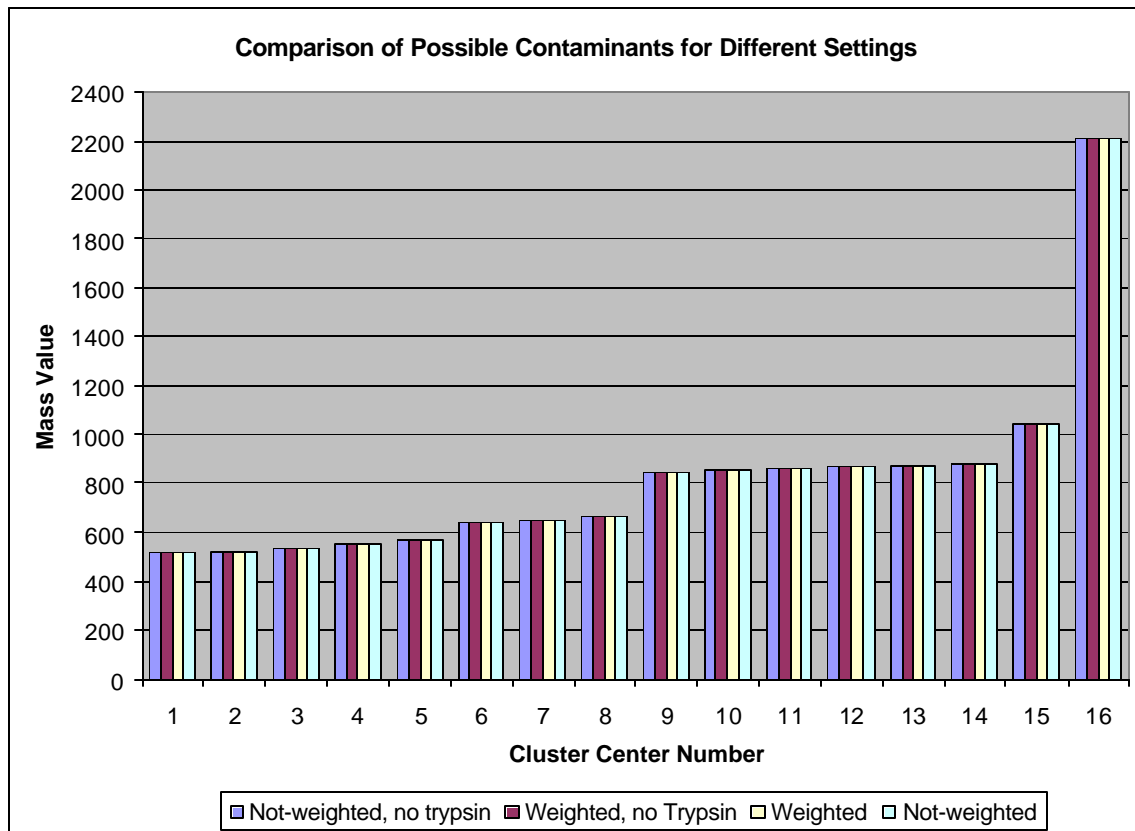


Figure 7

## 4 Conclusions

Our aim was to identify masses that may be contaminants rather than those derived from the protein following digestion. By doing so we believe that we will be able to make more accurate and reliable identification of proteins.

We have found that 30 PPM is the optimal cluster setting for clustering of mass values leading to identification of possible contaminants. This conclusion is based on minimum ambiguity and minimum average error rate, when compared to four known trypsin masses. We identified 16 masses as probable contaminants, with frequencies greater than 20% of 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.

We have presented a new clustering method to find probable contaminant masses. The method can be adopted for analysis of much larger protein data sets. Second, we presented a hypothesis for determining contaminant masses based on the frequency of those mass values. Although this study was based on one experimental protocol, it can be extended and applied to determine possible contaminant masses when alternative experiments are conducted.

## 5 Acknowledgements

The authors would like to thank Dr. Srdjan Askovic and Allison Gehrke for their help with this work.

## 6 Bibliography

- [1] S.J. Cordwell, V.C. Wasinger, A. Cerpa-Poljak, M.W. Duncan, I. Humphery-Smith. 1997. Conserved motifs as the basis for recognition of homologous proteins across species boundaries using peptide-mass fingerprinting. *Journal of Mass Spectrometry*. 32(4):370-8.
- [2] D.C. Liebler. *Introduction to Proteomics*. Humana Press, Totowa, NJ, 2002.
- [3] 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.
- [4] J.R. Yates. *Electrophoresis* 1998, 19, 893-900.
- [5] N.E. Sherman, N.A. Yates, J. Shabanowitz, D.F. Hunt, W. Jeffery, M. Bartlet-Jones, D.J.C. Pappin, *Proceedings of the 43<sup>rd</sup> ASMS Conference on Mass Spectrometry and Allied Topics*, Atlanta, GA, May 21-26, 1995, p.626.
- [6] M.W. Duncan, K. Fung, J.A. Zirrolli, F. Basile, L.M. Glode, Y. Miller. Qualitative and quantitative proteomics applied to the identification of cancer biomarkers in biological fluids. *The Second Principal Investigator's (PI) Meeting of the Innovative Molecular Analysis Technologies (IMAT) Program*. Washington, DC; 2001:W52.
- [7] D. Pappin, D. Rahman, H.F. Hansen, M. Bartlet-Jones, W. Jeffery, A.J. Bleasby. In A.L. Burlingame, S.A. Carr, editors, *Chemistry Mass Spectrometry and Peptide-Mass Databases: Evolution of Methods for the Rapid Identification and Mapping of Cellular Proteins*, Humana, Totowa, NJ 1996, 135-150.
- [8] B. Kuster, P. Mortensen, M. Mann. *Proceedings of the 47<sup>th</sup> ASMS Conference on Mass Spectrometry and Allied Topics*, Dallas, TX, June 13-17, 1999, 1897-1898.

- [9] H.J. Issaq, T.P. Conrads, G.M. Janini, T.D. Veenstra. Methods for fractionation, separation and profiling of proteins and peptides. *Electrophoresis* 2002, **23**, 3048-3061. World Wide Web, [http://www.spectroscopynow.com/Spy/basehtml/SpyH/1,2466,10-1-2-0-0-news\\_detail-0-1502,00.html](http://www.spectroscopynow.com/Spy/basehtml/SpyH/1,2466,10-1-2-0-0-news_detail-0-1502,00.html).
- [10] D.F. Hochstrasser, J. Sanchez, R.D. Appel. Proteomics and its trends facing nature's complexity. *Proteomics* 2002, **2**, 807-812. World Wide Web, [http://www.spectroscopynow.com/Spy/basehtml/SpyH/1,2466,10-1-2-0-0-news\\_detail-0-1341,00.html](http://www.spectroscopynow.com/Spy/basehtml/SpyH/1,2466,10-1-2-0-0-news_detail-0-1341,00.html).
- [11] S. Karlin, B.E. Blaisdell, & P. Bucher. *Protein Engineering*. 5(8):729-738. 1992.
- [12] K. Cios, W. Pedrycz, R.W. Swiniarski. *Data Mining Methods for Knowledge Discovery*. Kluwer Academic Publishers, Norwell, MA, 1998.
- [13] 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.
- [14] K. Cios, A. Teresinska, S. Konieczna, J. Potocka, S. Sharma. 2000. Diagnosing Myocardial Perfusion SPECT Bull's-eye Maps - A Knowledge Discovery Approach. *IEEE Engineering in Medicine and Biology Magazine*, 19(4): 17-25
- [15] D.F. Hochstrasser, J.C. Sanchez, R.D. Appel. Proteomics and Its Trends Facing Nature's Complexity. *Proteomics*, 2002, **2**, 807-812.
- [16] J.R. Yates. *Mass Spectrometry*. 1998, **33**, 1-19.
- [17] S. Sechi. A Method to Identify and Simultaneously Determine the Relative Quantities of Proteins Isolated By Gel Electrophoresis. *Rapid Communications in Mass Spectrometry*. 2002; **16**:1416-1424.
- [18] T. Rabilloud. Detecting proteins Separated By 2D Gel Electrophoresis. *Analytical Chemistry*. **72**, 48A-55A. 2000.
- [19] M.P. Washburn, D. Wolters, J.R. Yates. Large-scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology. *Nature Biotechnology*. **19**, 242-247. 2001.
- [20] O.N. Jensen, M. Wilm, A. Shevchenko, M. Mann. Sample Preparation Methods for Mass Spectrometric Peptide Mapping Directly from 2-DE Gels. *Methods in Molecular Biology*. **112**, 513-530. 1999.
- [21] K.R. Jonscher, J.R. Yates, The Quadrupole Ion Trap Mass Spectrometer: A Small Solution to a Big Challenge. *Analytical Biochemistry*. **244**, 1-15. 1997.
- [22] K.J. Cios, (ed.). 2001. Medical Data Mining and Knowledge Discovery. Springer Physica-Verlag.
- [23] J.R. Yates. Mass Spectrometry and the Age of the Proteome. *J. Mass. Spectrometry*. **33**, 1-19. 1998.
- [24] K.J. Cios, G.W. Moore. 2002. Uniqueness of Medical Data Mining. *Artificial Intelligence in Medicine*, **26**(1-2): 1-24.
- [25] 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.
- [26] R. Gras, M. Muller, E. Gasteiger, S. Gay, P.A. Binz, W. Bienvenut, *et al.* Improving Protein Identification From Peptide Mass Fingerprinting Through a Parameterized Multi-level Scoring Algorithm and an Optimized Peak Detection. *Electrophoresis*. **20**, 3535-3550. 1999.

- [27] O.N. Jensen, A.V. Podtelejnikov, M. Mann. Identification of the Components of Simple Protein Mixtures By High-accuracy Peptide Mass Mapping and Database Searching. *Analytical Chemistry*. 69, 4741-4750. 1997.