Reducing the Cost of Protein Identifications from Mass Spectrometry Databases

B. Logan, L. Kontothanassis, D. Goddeau, P.J. Moreno, R. Hookway[1], D. Sarracino[2]
Cambridge Research Laboratory
HP Laboratories Cambridge
HPL-2004-139
August 9, 2004*

mass spectrometry, machine learning, workflow management, noise filtering

We present two techniques to improve the computational efficiency of protein discovery from mass spectrometry databases: noise filtering and hierarchical searching. Our approaches are orthogonal to existing algorithms and are based on the observation that typical mass spectrometry data contains a large amount of noise that can lead to wasteful computation. Our first improvement uses standard machine learning techniques with novel feature vectors derived from the mass spectra to identify and filter the noisy spectra. We demonstrate this approach results in computational gains of around 38% with less than 10% loss of peptides. Additionally we present a hierarchical searching scheme in which most samples are matched against a small database at low computational cost, leaving only a small number of samples to be searched against larger databases. Combining this scheme with the machine learning filters leads to a further performance improvement of 3%.

* Internal Accession Date Only
[1]Hewlett-Packard High Performance Technical Computing, Marlboro, MA, USA
[2] Harvard Partners Center for Genetics and Genomics, Cambridge MA, USA
Approved for External Publication

I. Introduction

Mass Spectrometry (MS) is a well-studied technique to determine the protein content of a sample. The sample is first cleaved using enzymes into smaller chains called peptides. The peptides are then fed into a mass spectrometer which applies an electrical charge to them, breaking them into smaller pieces and creating charged sub-peptides. A small electrical field then guides these sub-peptides to a plate where they arrive at different times based on their mass. Using the time of arrival at the plate, the masses of the sub-peptides can be derived and the collection of masses of all sub-peptides belonging to a single peptide results in the mass spectrum of that peptide. These observed spectra can then be compared against those derived from a database of known proteins in order to identify the proteins present.

A commonly used algorithm for such comparisons is the SEQUEST algorithm [1]. This operates as follows. First a set of peptides is computed from the database of known proteins by emulating computationally the chemical process used to split the proteins. Second the mass spectrum of each peptide is computed by emulating computationally the process of electrical charge application on the peptides. These two steps can be pre-computed for a particular database of proteins creating an indexed database. Finally the observed mass spectra are compared with the computed spectra and close matches are returned as likely peptides in the sample.

Currently this process is far from perfect. During ionization, peptides can undergo a chemical change or *modification* which affects their mass and hence their spectral signature. Since this process is non-deterministic, the search process must anticipate both modified and unmodified peptides and prepare the comparison database accordingly. For each type of modification, the database grows substantially in size which quickly becomes problematic since there are a large number of potential modifications. For example the database size of the human proteome can vary greatly from a few tens of megabytes when no modifications are considered to over a terabyte when all modifications are taken into account. Computational overhead is directly related to database size thus searching against all possible modifications all the time is inadvisable.

Additionally, due to machine malfunctions, poorly fragmenting samples and contaminants, many mass spectra are invalid; they will not match any protein, yet both researcher and computational time is still wasted searching for them in the databases. Finally a sample may not match a protein in any database because it contains a protein which is currently unknown.

We have developed two approaches to improve the search time for identifying proteins from a series of mass spectra. First, we use machine learning algorithms to reject invalid mass spectra which are unlikely to be found in any database. Second, we have developed a workflow system to automate searching against multiple databases which only pays the high cost associated with large databases when necessary.

Previous approaches have addressed the problem of improving the speed of the search algorithms for a particular database e.g. [2]. However we are unaware of any work that attempts to eliminate contaminated spectra or identifies the appropriate database for searching. Our work is likely to benefit from any improvements in the searching algorithms (e.g. [3], [4]) since we use those algorithms as black boxes, only invoking them when necessary and reducing the space that they search.

II. Methodology

A. Filtering Invalid Spectra

Most invalid mass spectra arise from non-peptide contaminants and poor chemical fragmentation of peptides. We have found that it is common for 60-70% (and sometimes up to 90%) of spectra to be invalid, thus there is a strong motivation to automatically reject these. We therefore explore using classifiers to distinguish between valid and invalid spectra.

i. Feature Vectors

A typical mass spectrum contains around 2000 data points. We therefore represent each spectrum by its lower 2000 points and learn the statistics of this. Many of these points are zero or close to zero however. Since it is difficult to model such sparse feature vectors, we also consider several transformations of the spectra to reduce the dimensionality. An obvious choice is Principal Components Analysis (PCA) [5] which transforms a set of vectors to a lower dimension while preserving as much variance information as possible.

Visual inspection of valid and invalid mass spectra also led us to hypothesize several novel feature transformations. Valid spectra tend to have relatively few peaks which are well separated from the noise floor. Invalid spectra have a high noise floor which makes it difficult to distinguish the valid peaks.

We therefore form feature vectors as histograms of data points in different regions of the spectra since the characteristics of such histograms should differ markedly for valid and invalid spectra. Figure 1 illustrates the construction of histogram features from a typical mass spectrum. The image is divided into N horizontal regions as shown and the number of points in each region counted to form an N dimensional vector. We use an uneven division of the space to capture more detail for the lower intensities. We denote this type of feature vector `1D Hist'. We also consider a two dimensional division of the image, resulting in a N*N dimensional vector which we denote `2D Hist'.

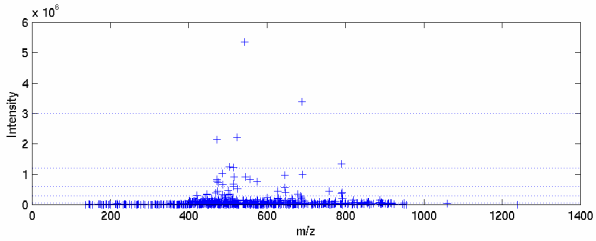

Fig. 1: Feature extraction from a typical mass spectrum. The image is divided into N horizontal regions as shown and the number of points in each region counted to form an N dimensional vector.

ii. Modeling Techniques

We use two standard machine learning techniques to model the features extracted from valid and invalid mass spectra: Gaussian mixture models (GMMs) and boosting [5]. GMMs model the probability distribution of the feature vectors from data classes. In our case, the data classes are valid and invalid spectra. Boosting learns the boundary between two classes of data. The classifier is formed as a weighted sum of so-called *weak learners*. Each round of the algorithm learns a weak learner, focusing on more and more difficult regions of the feature space.

B. Workflow Improvements

In addition to filtering invalid spectra, we improve the mass spectrometry process by searching the various peptide databases in an intelligent manner. Specifically, we search the available databases in a hierarchical fashion, scanning the smaller databases first in the hope that the sample will be identified cheaply. Constructing a hierarchy of databases requires some domain knowledge since the number of database modifications that can be considered is quite large.

Table 1 summarizes the three different workflows we evaluated and the databases comprising each workflow. Our first workflow is a degenerate case, consisting of a full database with all contaminants, all human proteins, and all possible modifications as discussed below. The second workflow consists of the following series of databases: a) a database of known contaminants (Inc1), b) a database of human proteins with no modifications (Inc2), c) a database of human proteins where Methionine residues may be oxidized (Inc3), d) a database of human proteins where Cysteine residues may be carboxyamidomethylated (Inc4), and e) a database of human proteins where Serine, Threonine, and Tyrosine residues may be phosporylated (Inc5). In all of the above cases a database contains both the modified and unmodified versions of the peptides. Such databases are termed inclusive.

We also considered a third workflow where modifications in databases *must* be present as opposed to being optional. Those databases are termed exclusive and represented by the identifiers Exc1 to Exc5 in Table 1. Exclusive databases are smaller and hence faster to search but can sometimes result in fewer hits, especially when more than one modification is considered simultaneously as is the case for database (e). Peptides that have just one modification will not hit an exclusive database that requires all of the modifications to be present. However it makes sense to consider exclusive databases since some modifications occur simultaneously in practice and thus searching a smaller exclusive database can be just as effective and computationally cheaper.

For each step in the workflow process we use the SEQUEST analysis tool to search the corresponding database. At the end of the step, we score the results of the search. Peptides found to have a high enough score with respect to the searched database are removed from the search set. The search continues until either the search steps are exhausted or the search set becomes empty.

TABLE I

Name and contents of peptide databases used. The databases contain contaminants, clean (unmodified) and modified peptides. A • denotes whether the data is included. Databases are either inclusive or exclusive.

Name	Contam (a)	Clean (b)	Modifications		
			M+Oxid. (c)	C+Acet. (d)	STY+Phosp. (e)
Full	•	•	•	•	•
Inc1	•				
Inc2		•			
Inc3		•	•		
Inc4		•		•	
Inc5		•			•
Exc1	•				
Exc2		•			
Exc3			•		
Exc4				•	
Exc5					•

III. RESULTS

A. Initial Filtering Experiments

We first conduct a series of experiments to validate the performance of the filtering algorithms. We use a set of mass spectra from 20 patients. 10 patients are healthy and 10 have tumors. The total number of spectra is 101360 and is roughly divided equally between the patients. This data is further evenly divided into training and test sets with the healthy and unhealthy patients evenly distributed between the sets.

We run the SEQUEST algorithm on each spectrum using a complete database with all modifications taken into account in order to identify whether it is valid. SEQUEST returns a number of parameters such as correlations and scores for the most likely peptides which match the spectra. We use a combination of expert-specified thresholds on these parameters to identify valid and invalid spectra. According to these criteria, around 10% of the spectra are labeled as valid.

We then transform the data into feature vectors and learn classifiers on the training set as described in Section II. We use these classifiers to label the data in the test set. Figure 2 shows the ROC curves for Gaussian classifiers learnt on the various feature vectors considered. The area under these curves describes the efficacy of each feature transformation and is indicated on the plot. From these curves we can see that the raw spectral features and 2D histogram features have comparable performance. However, in practice the cost of false negatives is higher than false positives since peptides which are missed could lead to missing a protein altogether. Therefore, we wish to set the operating point of our classifier to the upper right hand part of the plot. Therefore, we choose to use the 2D histogram features since these have better performance in this area.

We also investigated using mixture of Gaussian classifiers but the performance degraded. For example, for the 2D histogram features, the ROC area decreased from 0.80 to 0.74 when the number of mixtures increased from 1 to 2. Most likely this is because we do not have sufficient examples to train more than one mixture component.

We now compare the Gaussian classifier to the boosting classifier for the 2D histogram features. The ROC curves and corresponding areas are shown in Figure 3. We see that the boosting classifier has better performance than the Gaussian classifier, although their performance is comparable in the operating region of interest.

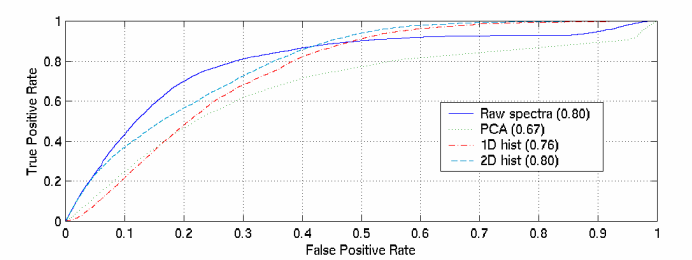

Fig. 2: ROC curves comparing a Gaussian classifier for four different feature vectors: raw mass spectra, PCA, 1D histograms and 2D histograms. Areas under the curves are shown in parentheses.

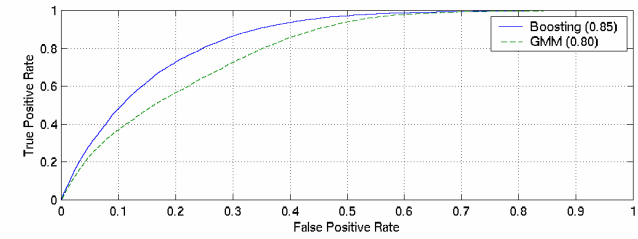

Fig. 3: ROC curves comparing a Gaussian to a Boosting classifier for 2D histogram features. Areas under the curves are shown in parentheses.

B. Initial Workflow Experiments

We next investigate the performance of a system incorporating the workflow improvements described in Section II. We arbitrarily choose one healthy and one tumor patient for these experiments. We compare the time taken to analyze the mass spectra from the two patients using our three different workflow systems. The first system, denoted `Full', analyses the spectra using SEQUEST with the full database. The second system, denoted `Inclusive', successively runs SEQUEST using the series of inclusive databases Inc1 to Inc5 as described earlier. The final system, denoted `Exclusive', successively runs SEQUEST using the series of exclusive databases Exc1 to Exc5.

Figure 4 shows the results of this experiment. We see that breaking the database search into sequential steps is detrimental. This is due to the fact that around 70% of the spectra produced by the spectrometer cannot be found in any of the databases. Therefore, these spectra traverse the whole workflow, incurring the full cost of searching all the databases in the hierarchy. This is slower than simply searching the single `Full' database.

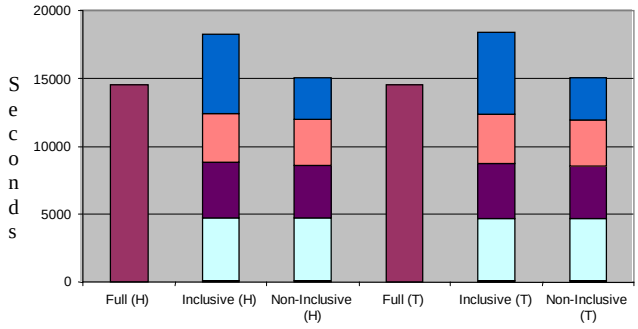

Fig. 4: Time in seconds to analyze data from a healthy (H) and tumor (T) sample for the different workflow systems. Different colors represent different sequential stages in the search.

C. *Combining Filtering and Workflow*

We now consider combining the filtering of valid and invalid spectra with the workflow in the hope that filtering will eliminate failed spectra and thus make the workflow system more efficient. We perform the same experiments as in Section

B but first filtering out invalid spectra using either the Gaussian or boosting classifier. The results for the healthy sample are shown in Figure 5. The results for the tumor sample are comparable.

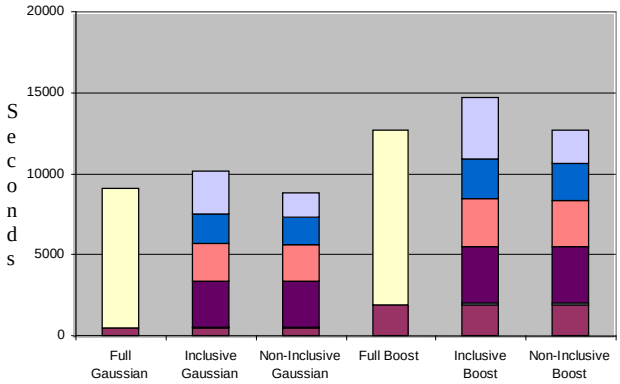

Fig. 5: Time in seconds to analyze data from a sample for different workflow systems with either a Gaussian or boosting classifier as a prefilter. Different colors represent different sequential stages in the search.

From these results we see that filtering speeds the process considerably. When searching the full database, filtering reduces the runtime by 38%. The hierarchical searches benefit even more, seeing reductions in runtime of 43.7% and 41.6% for the inclusive and exclusive workflows respectively. The lowest runtime is for the system which combines filtering with searching the series of exclusive databases. This is an improvement of approximately 3% over filtering and searching the full database.

TABLE 2
Number of correct peptides eliminated by various filters and total number of valid peptides for a healthy and tumor sample.

Filter	Healthy	Tumor
Gaussian	57/927	87/870
Boosting	15/927	14/870

However, the increase in speed when using the filters comes at a cost. As we saw earlier on the ROC curves, neither classifier is perfect. Although we correctly reject invalid peptide spectra, we sometimes also reject valid peptides. This could lead to errors in protein identification. Table 2 lists the number of incorrect peptides found when the filters are used. We see that at most 10% of the good peptides are eliminated. For some applications, this would justify the increased speed introduced by filtering.

IV. CONCLUSION

We have presented two main techniques which improve the mass spectrometry process. Our first approach filters invalid spectra before performing database searches and can lead to a runtime reduction of almost 40% with less than 10% loss in peptides. We also present a number of searching workflows. The performance of these workflows was highly dependent on the quality of filtering and the length of the workflow itself. Without filtering, the workflows were detrimental to performance. When combined with filtering however, a 3% improvement was seen.

Future work will concentrate on improved filtering techniques and differently configured workflows. We are also studying how protein identification is affected by the rejection of good peptides.

REFERENCES

[1] J. K. Eng, A. L. Mc Cormack and J. R. Yates, “An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database”, 1994, *J. Amer. Soc. Mass Spectrom.*, vol. 5, pp. 976-989.

[2] R. Carter, “Riptide: Fast protein identification from mass spectrometer data”, 2003, *Proc. 2003 IEEE Bioinformatics Conference*, Stanford, CA, pp. 398-399.

[3] T. Chen, J. Jaffe and G. Church, “Algorithms for identifying protein cross-links via tandem mass spectrometry”, 2001, *J. Comput. Bio*. vol. 8, no. 6, pp. 571-583.

[4] D. Anderson, W. Li, D. Payan and W. Noble, “A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores”, 2003, *Journal of Proteome Research*, vol. 2, no. 2, pp. 137-146.
[5] R. O. Duda, P. E. Hart and D. G. Stork, “Pattern Classification”, 2001, John Wiley & Son.