Identifying Phenotype-Relevant Protein Structural Variations by Mass Spectrometry Using Statistical Learning Approaches

Zheng Li [1,2], Catherine E. Costello [2], Mark E. McComb [2]

1. Biomedical Engineering, Boston University, Boston, MA
2. Cardiovascular Proteomics Center, Boston University School of Medicine, Boston, MA.

Introduction: Mass spectrometry is widely used for characterization of protein primary structure and structural changes, such as sequence variants and post-translational modifications. MALDI-TOF-MS is effective for peptide mass mapping, and LC-MS/MS is a powerful technique in protein/peptide sequencing and in localizing the site of structural changes. For expressed proteins, many variants and post- or co-translational modifications have been found associated to changes in biological function and may result in a disease phenotype. Here, we explore feature-based identification of peptide structural changes in MS peptide mapping using hemoglobin as a target protein. These feature changes can be intensity or isotopic peak distribution, or both.

Methods: Sample preparation: Whole blood was diluted and cleaned by centrifugal filtration. Trypsin digestion of intact globin chains was performed for peptide mass mapping and tandem mass spectrometric measurement. The digests were analyzed by MALDI-TOF MS (Bruker Reflex IV) and online nanoLC-MS/MS (QTOF-API-US, Waters Corporation). MALDI-TOF-MS data, LC-MS and LC-MS/MS data were processed and searched against SwissProt and custom programmed Hemoglobin/PTM databases using Mascot Server 2.2, Matrix Science and BUPID (Boston University Protein Identifier).

Statistical analysis: Statistical approaches including clustering, principal component analysis, ensemble learning based upon decision tree and support vector machine, were applied for phenotype-relevant feature selection. Here we applied an approach of ensemble learning combining decision forest (Figure 1) and support vector machine (SVM) to select features that contribute the most to the separation of phenotypes. Decision forest was applied to select the first few significant features. Decision forest is an ensemble of decision trees with the idea of combining multiple individual prediction models to reach a final decision based upon majority voting. The scheme of the methodology is shown in Figure 1. Each decision tree uses a subset of MS features to predict the phenotype by constructing a series of IF-THEN rules. Each individual decision tree was developed using a distinct set of evidence that was excluded from other models so that each individual model makes a unique contribution to the final decision.

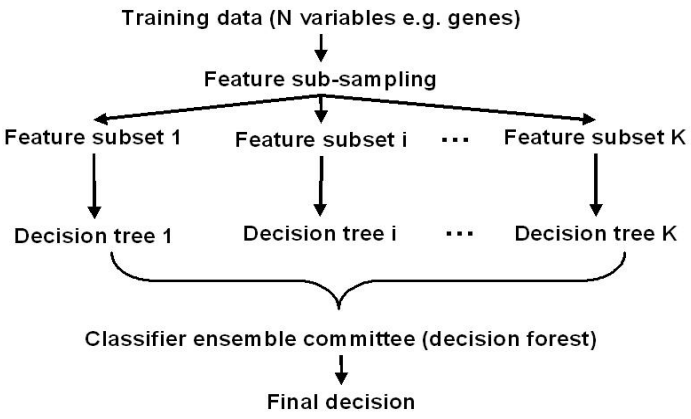

Fig1: Scheme of decision forest.

Recursive feature elimination is backward procedure consisting of three steps: (1) classifier training, (2) ranking score computation for each feature, (3) eliminating features with smallest ranking scores. We use a linear SVM as the classifier to discriminate blood samples of different phenotypes. SVM defines the classifier with $f(x) = \sum_{i=1}^{N} a_i y_i K(x_i, x) + b$, where a, b are parameters obtained with training procedure.

It is simplified into the following equation in linear case, $f(x) =< w, x > +b, w = (wi) = \sum_{i=1}^{N} a_i y_i x_i$. We rank the features with wi^2 and only the top M (e.g. 100) features with the highest scores are selected for classification. Combining decision forest with SVM_RFE enables the identification of significant features as well as subtle features for phenotype separation. A Matlab program was developed in house to implement the ensemble learning described above.

Results: MALDI-TOF-MS was performed for the tryptic digests of the samples. High sequence coverage of up to 95% for both alpha and beta globin chains was routinely achieved using PMF or LC-MS/MS with database searching, assuring the identification of any variants and possible post- or co-translational modifications in the globin chains. However, results often failed to discriminate different samples due to high false positive rates. Thus we used statistical approaches to discriminate samples based upon intensity data. Different levels of features including raw mzXML data, binned intensity data with different bin widths and peak distributions were extracted from the MALDI-TOF-MS data. For each level of features, different statistical learning approaches including clustering, PCA and ensemble learning based feature selection were applied to identify phenotype relevant features.

Raw data: It was found that clustering performed poorly on raw data with 186,000 data points per mass spectrum. Normal/sickle samples can not be clustered separately using all the features in the dataset. In contrast, PCA analysis performed reasonably well with the first three PCs to separate sickle vs normal samples (Figure 2). While feature selection approaches performed poorly due to the collinearity within the data set with multiple data points collected from the same peak. The first dozens of features were selected dominantly from the same peak (*m/z*~922.5).

Binned data: We performed the same statistical analysis on binned intensity data consisting of 1993 data points per spectrum. Similar to raw data analysis results, clustering cannot discriminate different phenotypes while PCA can separate the samples in a 3D space composed of the first three PCs. Ensemble learning based feature selection performed significantly better on binned data than raw data. All samples were predicted correctly with 100% accuracy of leave-one-out cross validation. The features selected were indicative of the protein structural variations between sickle and normal cells.

Peak data: Lastly, we compared to the statistical analysis results on peak-picked data which included 693 peaks identified using peak probability contrast algorithm (Stanford University). Clustering and PCA performed similarly to the raw and binned data sets. Clustering was unable to separate samples while PCA can separate samples in 3D-PCs space. Ensemble learning based upon peak list data identified the important peaks such as *m/z* 926 and 952.

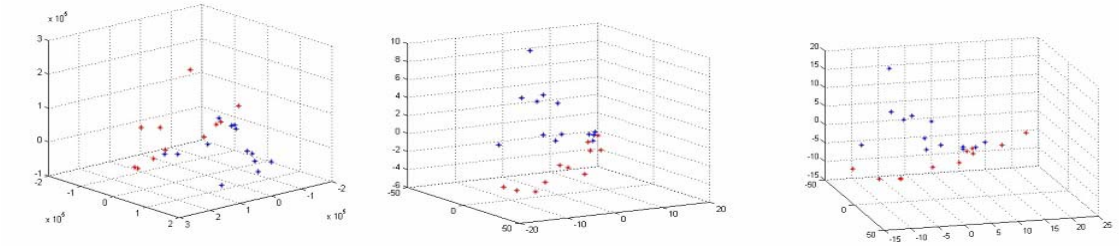

Fig2: PCA analysis of raw, binned and peak picking data.

Summary: Appropriate data reduction and model applications yield identification of structural variants in peptides and proteins, which allows for fast label-free sample classification with mass spectrometry data. Clustering performs poorly comparing to other approaches. However, with the preselected discriminative features, clustering can separate the phenotypes correctly. PCA is more robust to the data reductions in identifying underlying phenotype separation. Ensemble learning at the appropriate data level is efficient for biomarker identification.

Acknowledgement: NIH-NCRR grants P41 RR10888, S10 RR15942 and NIH-NHLBI contract N01 HV28178