Journal of Proteomics & Bioinformatics - Open Access

JPB/Vol.3 Issue 2

Research Article OPEN ACCESS Freely available online doi:10.4172/jpb.1000120

Clustering Mass Spectral Peaks Increases Recognition Accuracy and Stability of SVM-based Feature Selection

Mikhail Pyatnitskiy, Maria Karpova*, Sergei Moshkovskii, Andrey Lisitsa, and Alexander Archakov

Institute of Biomedical Chemistry, 119121, Pogodinskaya str., 10, Moscow, Russia

Abstract

Mass spectral profiling of serum or plasma is one of the tools widely used to make experimental diagnostic systems for different cancer types. In this approach, a set of discriminatory peaks serves as a multiplex cancer biomarker. Hence, adequate selection of peaks is a crucial stage in the development of diagnostic rule. In the present paper we propose using sequential filter and wrapper feature selection in a complete cross-validation scheme with feature selection performed at each run of cross-validation separately. Filter feature selection is represented by hierarchical cluster analysis; recursive feature elimination coupled with support vector machine is utilized as a wrapper feature selection method. The method performance is demonstrated on previously obtained dataset with ovarian cancer and non-cancer sera. Application of our approach led to a slight but statistically significant increase in accuracy. Peak clustering favoured more stable results of feature selection and provided a biological meaning to selected m/z values. We recommend clustering of peaks as a filter dimensionality reduction for further use in mass spectral studies.

Keywords: Mass spectrometry; SELDI; Biomarker discovery; Support Vector Machine; Recursive feature elimination; Clustering

Abbreviations: SVM: Support Vector Machine**;** RFE: Recursive Feature Elimination; MALDI: Matrix-assisted Laser Desorption/ionization; SELDI: Surface-enhanced Laser Desorption/ionization

Introduction

Matrix-assisted laser desorption/ionization (MALDI-TOF) mass-spectrometry as well as its protein-chip based modification surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) mass-spectrometry can rapidly provide information about tens of proteins and therefore is a promising instrument of proteome investigation. During the last decade a lot of effort was directed to utilizing mass-spectral profiles for discriminating cancer from non-cancer conditions. Although classification accuracy in many cases was excellent, no reliable biomarkers were found. Besides, poor reproducibility of the results between the laboratories caused a lot of controversy (Whelan et al., 2008).

Due to wide dynamic range of protein concentrations in serum it seems impossible to detect specific cancer products present in extremely low concentrations using direct profiling. Almost all proteins differentially expressed in cancer and detected on mass-spectra are related to inflammation and none of them can be used individually as a cancer biomarker (Hortin, 2006). Nevertheless, numerous reports of successful discrimination of cancer by MALDI profiling prove its diagnostic potential. Differences in MS profiles may be explained by differential modification pattern of major serum proteins as well as their truncated forms arising as a consequence of altered protease activity. These changes may be multiple and subtle, hence, development of reliable diagnostic algorithms based on complex mass-spectral data strongly depends on the correct usage of bioinformatical tools (Lumbreras et al., 2009).

There is quite a lot of experience in utilizing machine learning algorithms for development of diagnostic rules. One of the most widely used methods in genomic and proteomic research is Support Vector Machine (SVM). It is naturally combined with wrapper method of feature selection called Recursive Feature Elimination (RFE). On each step of RFE a SVM classifier is used to assign a relevance weight to each feature and then the feature with the lowest weight is eliminated. For the next iteration all weights are re-evaluated and dynamically adapted, while the process continues recursively (Gyuon et al., 2002). Eventually, the smallest set of top-ranked features achieving the highest classification accuracy is selected as a set of potential biomarkers.

However, SVM-RFE is characterized by high instability leading to different rankings of potential discriminative variables. In their pioneer work (Gyuon et al., 2002) Guyon and coworkers wrote: “We observed in real experiments that a slight change in the feature set often results in a completely different RFE ordering”.

Analysis of high-throughput data, including mass-spectra, often raises two important interrelated problems. First, the high dimensionality of MS data often leads to overfitting, so that the diagnostic model perfectly fits the training dataset but performs poorly on the independent dataset. This phenomenon, caused by so-called “the curse of dimensionality” can be partially avoided by minimizing number of features as an input to classifier. Hence, development of adequate methods for dimensionality reduction is crucial. The second common problem is a limited number of samples available for the analysis. Machine learning methods need a sufficiently large training dataset for feature selection and discrimination rule development as well as an independent test

***Corresponding author:** Maria Karpova, Institute of Biomedical Chemistry, 119121, Pogodinskaya str., 10, Moscow, Russia, Tel: +7-499-2461641, E-mail: karpova@bioinformatics.ru

Received January 13, 2010; **Accepted** February 12, 2010; **Published** February 12, 2010

Citation: Pyatnitskiy M, Karpova M, Moshkovskii S, Lisitsa A, Archakov A (2010) Clustering Mass Spectral Peaks Increases Recognition Accuracy and Stability of SVM-based Feature Selection. J Proteomics Bioinform 3: 048-054. doi:10.4172/jpb.1000120

set for calculation of diagnostic accuracy. A common way of testing diagnostic model in the case of small dataset is cross-validation. Cross-validation is a partitioning of data when the whole sample set is split in two parts, whereupon one part is used for classifier training, another is used to test the obtained model and the whole procedure is repeated many times.

There exist two possible schemes of combining feature selection with cross-validation. First scheme employs feature selection utilizing the whole dataset with subsequent classifier training and accuracy estimation at each cycle of cross-validation. Second scheme uses only the training set (which is resampled at each run of cross-validation) for feature selection and classifier training. As a consequence, each step of cross validation produces different feature sets. The first method often gives extremely optimistic results in terms of accuracy, which are caused by overfitting. Simon and coworkers (Simon et al., 2003) clearly demonstrated the difference between these two schemes of cross-validation by random assignment of class labels to gene expression samples. They performed the analysis of 2000 simulated datasets consisting of 10 'class 1' samples and 10 'class 2' samples. There was no true underlying difference between two classes, therefore expected class prediction accuracy should be around 50%, like a random guess. They tested leave-one-out cross-validation scheme with feature selection on the whole dataset, which resulted in 90.2% of simulated datasets with no misclassifications. However, when gene selection was also subjected to cross-validation, the mean number of misclassified profiles was close to 50% (Simon et al., 2003).

Thus, although the first incomplete cross-validation scheme was widely used earlier in biomedical researches (Gevaert et al., 2008; Xue et al., 2008), including our previous experience (Moshkovskii et al., 2007), it became obvious that only the complete cross-validation scheme gives reliable results and should be further used (Zhang et al., 2006; Barla et al., 2008).

In the present paper we employed sequential use of filter and wrapper methods for dimensionality reduction. Clustering of mass-spectral peaks with Pearson correlation as a distance between variables was used as a filter feature selection, whereas SVM-RFE was used as a wrapper method for further dimensionality reduction. We utilized hierarchical clustering to find groups of highly correlated features and selected one variable with the highest area under the ROC-curve (AUC) from each cluster, while discarding the rest of them. A similar approach was utilized in paper (Shin et al., 2008); however, in their research no wrapper feature selection was used.

We employed a non-biased scheme of cross-validation with both steps of feature selection carried on training set only. Our result demonstrates that clustering of m/z values results in increased classification accuracy, improves RFE stability and provides more biological meaning to the features.

Materials and Methods

Specimen

For this work we used our previous SELDI-TOF mass-spectral data on ovarian cancer and control specimen (Moshkovskii et al., 2007). Samples included 34 sera from women with epithelial ovarian cancer, 14 sera from women with benign ovarian tumors, 17 sera from women with uterine myoma and 26 sera from healthy women. Cancer sera were compared to all the others taken together, so we addressed binary classification problem.

Spectra acquisition and preprocessing

Sample preprocessing was carried out using normal-phase chips NP20. Spectra were obtained by mass-spectrometer SELDI-TOF Protein Biology System II (PBS II) in m/z range 5,500 to 17,500 Da.

Spectra preprocessing including baseline subtraction, peak identification and alignment, was performed using Biomarker Wizard™ software (Ciphergen Biosystems) with the following settings: signal/noise (first pass) 10, signal/noise (second pass) 5, minimum peak threshold 0%, mass error 0.2%. Only 48 peaks were detected due to stringent criteria of peak selection applied to avoid artifactual peaks.

All data were normalized to have zero mean and unit variance for each variable. To discard outliers we performed Principal Component Analysis. First principal components revealed one gross outlier (patient #83), which was removed from subsequent analysis.

Filter feature selection

We applied hierarchical cluster analysis to find groups of closely correlating variables. The correlation matrix was calculated for all variable pairs $\boldsymbol{R} = (r_{ij})$, where r_{ij} – Pearson correlation between i-th and j-th variable. We used correlation matrix instead of covariance matrix because all variables were measured in the same units and had the same scale. Then we transformed correlation matrix to matrix of dissimilarities by subtracting absolute value of correlation coefficient from unity: $d_{ij}=1-|r_{ij}|$. We also tried to define $d_{ij}=1-r^2_{ij}$, which yielded approximately the same results.

The obtained dendrogram was cut at specified level of similarity and for each cluster we chose the only variable with the highest AUC value. Different cut-off values from 0 to 0.9 with step 0.02 were probed. Thus, the higher cut-off value was taken, the less input variables were left for further feature selection and subsequent discrimination between classes.

Wrapper feature selection

We used SVM-RFE as a wrapper method for feature selection to rank the variables, selected on previous step. RFE is a sequential backward feature selection algorithm based on SVM (Gyuon et al., 2002). Initially, RFE started with all the features. Coefficients w_i of obtained decision function $D(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x}+b$ were used as feature weights. At each iteration, one feature with minimal weight was removed and SVM was trained with the remaining features. This procedure continued until all features were ranked according to the order of their removal. In this work we used SVM with linear kernel. For optimal tuning of SVM cost parameter C we employed svmpath package (Hastie et al., 2004).

The complete scheme of 10-fold cross-validation was implemented in the present research. The entire data set was randomly split into 10 non-overlapping parts. Both feature selection and classifier training were carried out on the training set (nine parts), whereas the remaining part was used for accuracy estimation. The whole procedure was repeated 50 times, each time for a

new split of a dataset. Thus, there were total 500 runs of feature selection, classifier training and testing. We also computed 95% confidence intervals for means of classification accuracy, sensitivity and specificity using the well-known formula $[\mu_1;\mu_2] = m \pm t_{\alpha/2,N-1}\, s/\sqrt{N}$, where m – sample mean, N – sample volume, $t_{\alpha/2,N-1}$, – value of t-distribution at given significance level α, s – sample standard deviation.

All computations were done with a set of in-house scripts for R statistical language. Sources are freely available from authors upon request.

Results

Dependence of diagnostic accuracy on clustering

The overall computational scheme is illustrated by the Figure 1. We outline that for each run of cross-validation the whole feature selection procedure was performed separately. Thus, at each iteration we obtained new ranking of features which yielded the best classification accuracy.

To investigate the influence of the first stage of feature selection, the dendrogram of feature correlation was cut at different levels from 0 to 0.9 with step 0.02. For each cut-off value we obtained clusters of variables. From each cluster we picked the only variable with the highest AUC and SVM-RFE was carried out to rank selected variables. For each partial subset of top-ranked variables we trained SVM classifier and assessed classification accuracy, sensitivity and specificity on test data.

Figure 2 illustrates the dependence of the highest achieved classification accuracy on cut-off value. There is a tendency for increasing accuracy with cut-off value up to 0.7. When running the whole scheme without the first step of feature selection (cut-off value equal to zero, i.e. no variable clustering was performed) we report the accuracy about 78.0±1.3% with the following discriminative m/z peaks: 11681, 6454, 10265, 6575. The highest diagnostic accuracy 81.1±1.1 % was achieved with cut-off value of 0.7 using peaks 11681, 6454, 10265, 13769, 8829. Although

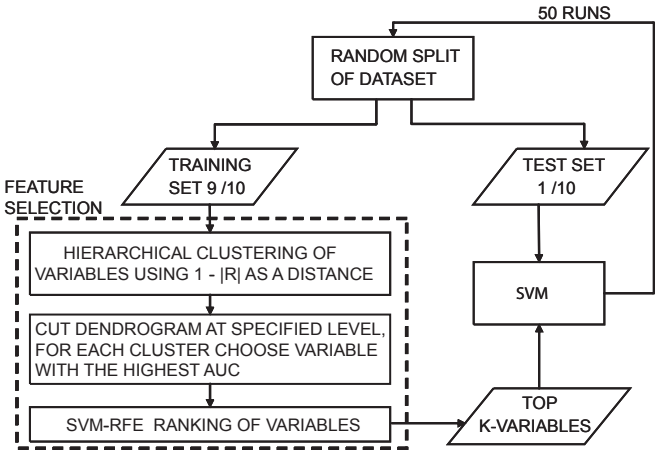

Figure 1: Overall scheme of computational analysis.

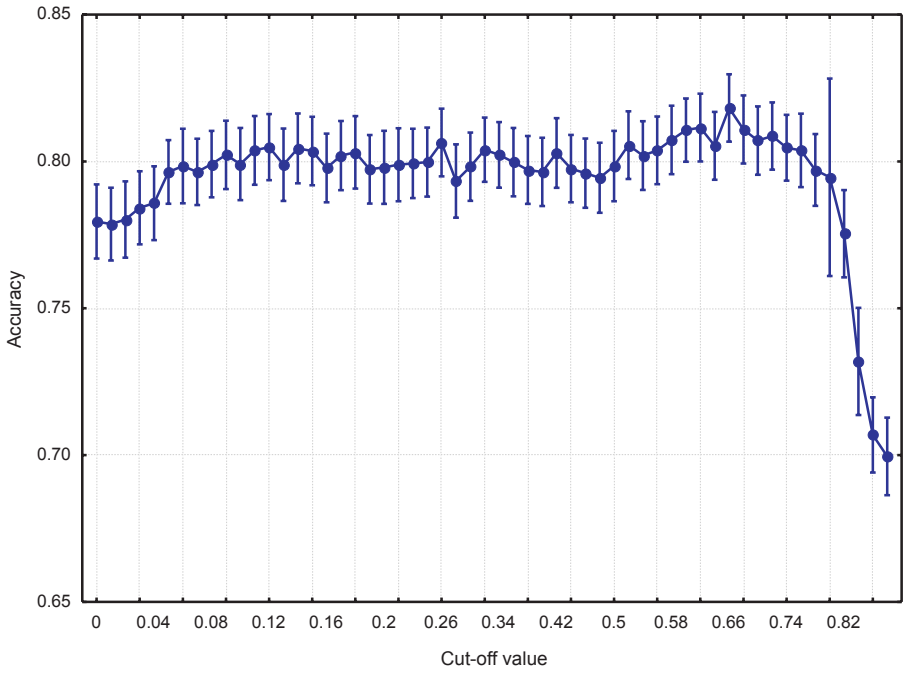

Figure 2: Dependence of the highest average classification accuracy on the cut-off value of correlation dendrogram of m/z values. Whiskers indicate 95% confidence intervals for mean accuracy.

the difference in diagnostic accuracy does not seem to be dramatic, it is indeed statistically significant. We can conclude that application of variable clustering improves the overall classification accuracy.

Increase in RFE stability at high cut-off values

We utilized SVM-RFE to rank the features selected on the clustering step. However, because of using complete cross-validation scheme, we obtain different variable rankings at each iteration. Also SVM-RFE is known for being rather sensitive to subtle changes in the data (Gyuon et al., 2002). All this leads to instability of selected discriminative variables. As a result, instead of fixed set of discriminative m/z values we can only operate with frequencies of selected variables having the specified rank. Figure 3 illustrates frequency of assigning the first rank to different features when no clustering was used (cut-off 0) and at cut-off 0.7. Cut-off value 0.7 was chosen because of the highest accuracy (81.1±1.1 %) achieved. The variable with m/z 11681 Da had the first rank in most runs of cross-validation both when clustering was used and when this stage was omitted. However, with variable clustering it was top-ranked two fold more frequently: in 421 runs of cross validation from total 500 (84%) compared to 220 runs (44%) without clustering. Increasing cut-off values resulted in larger cluster sizes and, therefore, fewer uncorrelated non-redundant variables were used for SVM-RFE procedure. This illustrates that variable clustering as filter feature selection allows more robust variable ranking obtained from wrapper feature selection.

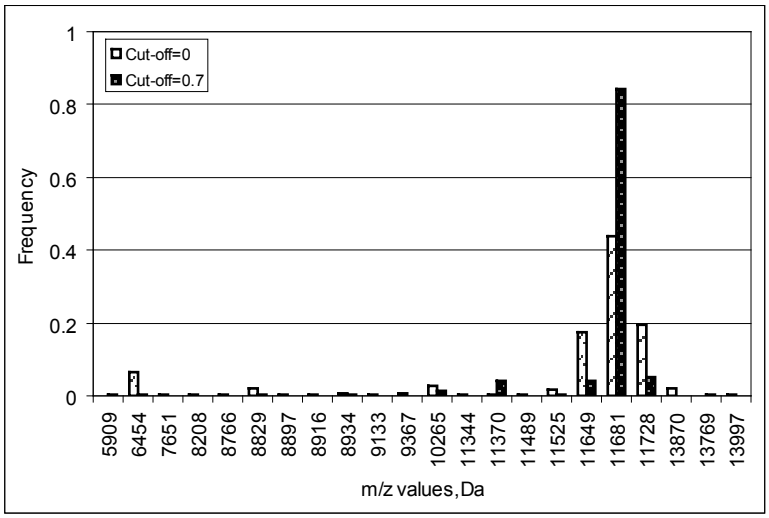

Figure 3: Frequency of assigning each feature the first rank with cut-off 0 (no clustering as feature selection) and at cut-off level 0.7.

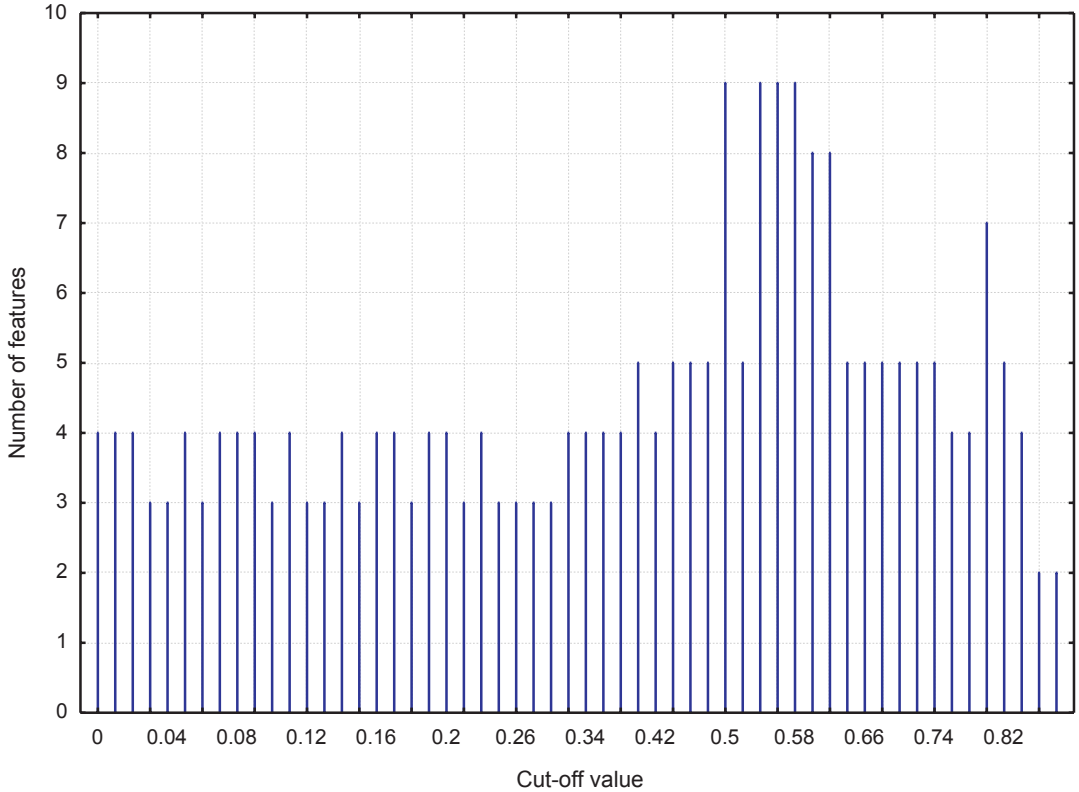

Figure 4: Number of features included in the model to obtain the highest diagnostic accuracy at different cut-off values of correlation dendrogram.

Discriminative m/z peaks and their biological sense

Figure 4 illustrates the dependence between number of features included in the model to obtain the highest diagnostic accuracy and cut-off value. Interestingly, the number of features increased up to 9 at cut-off levels of 0.5-0.6. Accuracies at these cut-off levels also reached the highest values of approximately 80±1%. Thus, number of features included in the model was comparable to the total number of features at these cut-off-levels (16 at cut-off value 0.5, 10-11 at cut-off level 0.7). Of course, all these peaks are not "true" cancer biomarkers; most of them correspond to major serum proteins or their modifications. Relatively high diagnostic accuracy obtained using these peaks let us draw a conclusion that in our case cancer versus non-cancer discrimination was due to differences in modification patterns of major serum proteins.

As it was already mentioned, the identity of features with selected rank differs between runs of cross-validation. Therefore, we report discriminative features which most frequently obtained the highest ranking for cut-off values 0 and 0.7. We also ran the 10-fold cross-validation without any feature selection for the same cut-off values. This experiment corresponded to "unfair" scheme, with feature selection carried out for the whole data prior to cross-validation. As expected, we obtained higher accuracy values (84.1±1.0% for cut-off 0 and 84.8±0.9% for cut-off 0.7) compared to "fair" scheme with feature selection step embedded in cross-validation cycle (78.01.3% for cut-off 0 and 81.1±1.1% for cut-off 0.7). Observed difference in accuracy can be explained by 'inadequate' runs of cross-validation, when irrelevant features are selected because of non-representative training chunk of data.

Although identification of peaks in this research was not carried out, the identity of some peaks is obvious because of previous knowledge about serum MALDI-TOF spectra structure. Peaks with m/z 11681 Da and 11525 Da were previously identified as serum amyloid A1 alpha and truncated form of serum amyloid A1 alpha respectively (Moshkovskii et al., 2005). Peak with m/z 13769 Da corresponded to transthyretin, 13870 Da - to cysteinylated form of transthyretin (Zhang et al., 2004). Men-

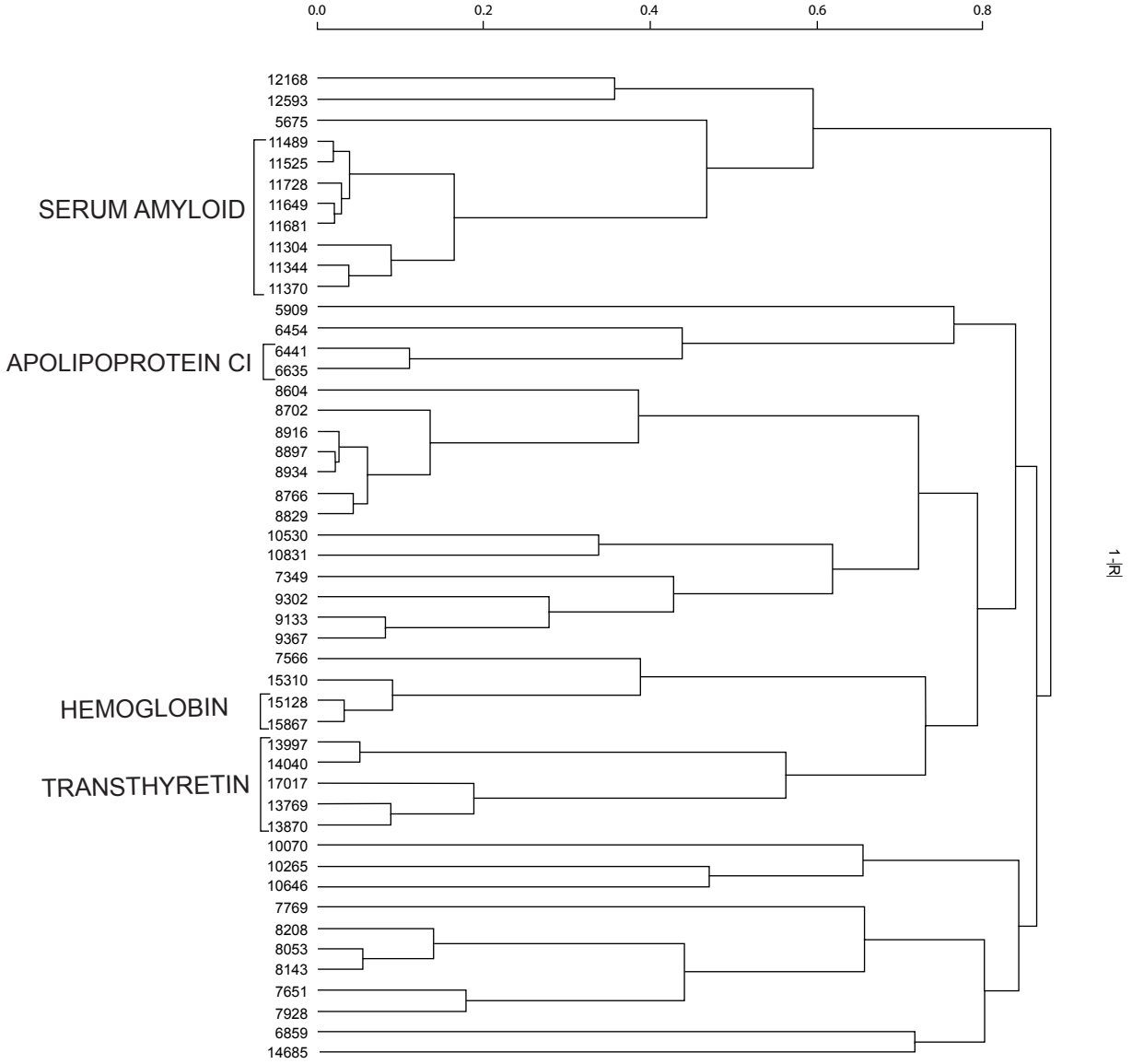

Figure 5: Correlation dendrogram of m/z values built for the whole dataset. Average linkage clustering was used, distance defined as 1– |r|, where *r* is Pearson correlation coefficient.

tioned proteins were reported as differentially expressed in a number of cancer types (Koomen et al., 2005; Kozak et al., 2005; Ehmann et al., 2007). Peak with m/z 6454 Da most likely corresponded to oxidated form of apolipoprotein CI which was reported to be visualized on mass-spectra of healthy people (Nedelkov et al., 2006).

Clustering of variables also provided valuable information about mass-spectral structure because highly correlated features combined in clusters were likely to be modified forms of the same protein. Dendrogram built on the whole dataset is shown on Figure 5. We note that the actual dendrograms used for feature selection were built using only the training set and hence were a bit different at each run of cross-validation.

Clusters of serum amyloid, transthyretin, apolipoprotein C1 and hemoglobins alpha and beta can be clearly distinguished on the dendrogram. Each cluster encompasses covalent modifications and truncated forms of distinct protein. Close correlation of selected MS peaks with identified protein may contribute to speculations about their identity. Other sources of highly-correlated m/z peaks are multiple-charged peaks. Clustering of MS peaks removes unnecessary details in spectra and can simplify biological interpretations of obtained discriminative features.

Discussion

Most papers devoted to discrimination between samples using mass spectra employ different approaches to feature selection (Oh et al., 2005; Zhang et al., 2006; Shin et al., 2008). Discovery and subsequent identification of discriminative peaks is of fundamental and practical interest.

The overall idea of proposed algorithm was simple – we wanted to obtain minimal non-redundant set of peaks, which would be uncorrelated but still highly discriminating between classes. We adopted two-step approach, where filter feature selection was used at the first stage, and wrapper – at the second stage. We also used a complete scheme of cross-validation to minimize possible overfitting. We stress, that all feature selection was performed on training data only, thus estimated classification accuracy values can be treated with confidence.

Filter feature selection was implemented as clustering of variables with Pearson correlation as a distance. Variable clustering is a partitioning a set of variables into hierarchical groups of classes, which often leads to helpful insights into the data structure. Choosing one representative variable from each cluster reduces the redundancy among the variables.

Another possible way to reduce the redundancy among the variables is factor analysis, where variables are represented as linear combinations of hidden "factors". However, existence and interpretation of these hypothetical factors is unclear and should be carefully investigated in each case. On the contrary, cluster analysis deals with principal dimensionality in the data, rather than abstract factors and results of clustering are easily interpreted.

One of the promising methods for feature selection and ranking is recursive feature elimination (RFE). RFE is coupled with classification method, most often with SVM. SVM-RFE is a wrapper feature selection method and operates in a recursive manner where features are reordered between runs based on weights received from the classification rule (Gyuon et al., 2002). RFE is superior to naïve ranking methods (such as ranking features according to AUC values or p-values in t-test (Yu et al., 2004) because it concerns the data as a whole, including all their complex interrelations. It has been applied to mass spectral data and showed high prediction accuracy (Duan et al., 2005). However, SVM-RFE is known to be rather unstable, and subtle differences in the input data may lead to dramatic altering of variable ranking. Instability of feature selection in machine-learning methods can be one of the reasons for low reproducibility of discriminative features reported by different research groups. Using an additional step of filter dimensionality reduction prior to SVM-RFE removes redundant features from the dataset and makes results of SVM-RFE more stable.

At high cut-off values high accuracy was obtained with many m/z values used in classification model. For example, at cut-off value 0.6 the highest accuracy of 81.0±1.2% was reached with eight features. It says for the whole spectra performance as a discriminator rather than presence of any true cancer biomarkers seen on MS spectra. Crucial feature selection is limited to removing highly correlated variables at the filtering stage; after that using almost all the left features leads to high diagnostic accuracy. Thus, it is likely that no cancer-specific products are present in selected panels. Experimental conditions utilized in this research were extremely simple and lacked stages of any fractionation or depletion. Therefore, all the peaks surely correspond to some abundant serum proteins.

For example, peaks corresponding to serum amyloid A (SAA), transthyretin (TTR) and oxidated apolipoprotein CI are often included in diagnostic models. SAA and TTR are well-known inflammatory proteins and thus lack specificity (Hortin, 2006). However, even absence of any cancer-specific proteins detected on mass-spectra does not exclude the possibility of using mass-spectra as discriminators. Cancer is often accompanied by altered protease activity, which may lead to arising of cancer-specific truncated forms of major serum proteins. Differential modification pattern of serum proteins in cancer has been demonstrated earlier (Fung et al., 2005; Miguet et al., 2006).

Using complete scheme of cross-validation decreased the accuracy to 78.0±1.3% (without clustering) compared to 89.5±0.7% reported in our previous work (Moshkovskii et al., 2007). Utilizing our approach we achieved classification accuracy 81.1±1.2%. However, the present results are much more reliable, whereas previous estimates were too optimistic and probably caused by overfitting.

Conclusion

In the present paper we proposed clustering of MS peaks as the first step of feature selection. We tested our algorithm of sequential filter and wrapper feature selection on ovarian cancer sample set. Proposed approach allowed us to achieve higher accuracy compared to wrapper feature selection only. Clustering of mass-spectral peaks also simplifies biological interpretation.

Acknowledgements

This work was supported by the Program "Proteomics for Medicine and Biotechnology" of Russian Academy of Medical Sciences.

References

1. Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, et al. (2008) Machine learning methods for predictive proteomics. Brief Bioinform 9: 119-128. » CrossRef » PubMed » Google Scholar

2. Duan KB, Rajapakse JC, Wang H, Azuaje F (2005) Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans Nanobioscience 4: 228-234. » CrossRef » PubMed » Google Scholar

3. Ehmann M, Felix K, Hartmann D, Schnolzer M, Nees M, et al. (2007) Identification of potential markers for the detection of pancreatic cancer through comparative serum protein expression profiling. Pancreas 34: 205-214. » CrossRef » PubMed » Google Scholar

4. Fung ET, Yip TT, Lomas L, Wang Z, Yip C, et al. (2005) Classification of cancer types by measuring variants of host response proteins using SELDI serum assays. Int J Cancer 115: 783-789. » CrossRef » PubMed » Google Scholar

5. Gevaert O, De Smet F, Van Gorp T, Pochet N, Engelen K, et al. (2008) Expression profiling to predict the clinical behaviour of ovarian cancer fails independent evaluation. BMC cancer 8: 18. » CrossRef » PubMed » Google Scholar

6. Gyuon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Machine Learning 46: 389-422. » CrossRef » PubMed » Google Scholar

7. Hastie T, Rosset S, Tibshirani R, Zhu J, Cristianini N (2004) The entire regularization path for the support vector machine. J Mach Learn Res 5: 1391-1415. » CrossRef » PubMed » Google Scholar

8. Hortin GL (2006) The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome. Clin Chem 52: 1223-1237. » CrossRef » PubMed » Google Scholar

9. Koomen JM, Shih LN, Coombes KR, Li D, Xiao LC, et al. (2005) Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. Clin Cancer Res 11: 1110-1118. » CrossRef » PubMed » Google Scholar

10. Kozak KR, Su F, Whitelegge JP, Faull K, Reddy S, et al. (2005) Characterization of serum biomarkers for detection of early stage ovarian cancer. Proteomics 5: 4589-4596. » CrossRef » PubMed » Google Scholar

11. Lumbreras B, Porta M, Marques S, Pollan M, Parker LA, et al. (2009) Sources of error and its control in studies on the diagnostic accuracy of "-omics" technologies. Proteomics Clin Appl 3: 173-184. » CrossRef » PubMed » Google Scholar

12. Miguet L, Bogumil R, Decloquement P, Herbrecht R, Potier N, et al. (2006) Discovery and identification of potential biomarkers in a prospective study of chronic lymphoid malignancies using SELDI-TOF-MS. J Proteome Res 5: 2258-2269. » CrossRef » PubMed » Google Scholar

13. Moshkovskii SA, Serebryakova MV, Kuteykin-Teplyakov KB, Tikhonova OV, Goufman EI, et al. (2005) Ovarian cancer marker of 11.7 kDa detected by proteomics is a serum amyloid A1. Proteomics 5: 3790-3797. » CrossRef » PubMed » Google Scholar

14. Moshkovskii SA, Vlasova MA, Pyatnitskiy MA, Tikhonova OV, Safarova MR, et al. (2007) Acute phase serum amyloid A in ovarian cancer as an important component of proteome diagnostic profiling. Proteomics Clin Appl 1: 107-117. » CrossRef » PubMed » Google Scholar

15. Nedelkov D, Kiernan UA, Niederkofler EE, Tubbs KA, Nelson RW (2006) Population proteomics: the concept, attributes, and potential for cancer biomarker research. Mol Cell Proteomics 5: 1811-1818. » CrossRef » PubMed » Google Scholar

16. Oh JH, Gao J, Nandi A, Gurnani P, Knowles L, et al. (2005) Diagnosis of early relapse in ovarian cancer using serum proteomic profiling. Genome Inform 16: 195-204. » CrossRef » PubMed » Google Scholar

17. Shin H, Sheu B, Joseph M, Markey MK (2008) Guilt-by-association feature selection: identifying biomarkers from proteomic profiles. J Biomed Inform 41: 124-136. » CrossRef » PubMed » Google Scholar

18. Simon R, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 95: 14-18. » CrossRef » PubMed » Google Scholar

19. Whelan LC, Power KA, McDowell DT, Kennedy J, Gallagher WM (2008) Applications of SELDI-MS technology in oncology. J Cell Mol Med 12: 1535-1547. » CrossRef » PubMed » Google Scholar

20. Xue R, Lin Z, Deng C, Dong L, Liu T, et al. (2008) A serum metabolomic investigation on hepatocellular carcinoma patients by chemical derivatization followed by gas chromatography/mass spectrometry. Rapid Commun Mass Spectrom 22: 3061-3068. » CrossRef » PubMed » Google Scholar

21. Yu JK, Chen YD, Zheng S (2004) An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics. World J Gastroenterol 10: 3127-3131. » CrossRef » PubMed » Google Scholar

22. Zhang X, Lu X, Shi Q, Xu XQ, Leung HC, et al. (2006) Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC bioinformatics 7: 197. » CrossRef » PubMed » Google Scholar

23. Zhang Z, Bast RC Jr, Yu Y, Li J, Sokoll LJ, et al. (2004) Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res 64: 5882-5890. » CrossRef » PubMed » Google Scholar