Published in final edited form as:
J Proteome Res. 2008 September ; 7(9): 4013–4021. doi:10.1021/pr8002886.

The Knowledge-Integrated Network Biomarkers Discovery for Major Adverse Cardiac Events

Guangxu Jin[†,‡], **Xiaobo Zhou**[*,†], **Honghui Wang**[§], **Hong Zhao**[†], **Kemi Cui**[†], **Xiang-Sun Zhang**[‡], **Luonan Chen**[‖], **Stanley L. Hazen**[⊥], **King Li**[#], and **Stephen T. C. Wong**[†]

[†] Center for Biotechnology and Informatics, The Methodist Hospital Research Institute & Cornell University, Houston, Texas 77030

[‡] Academy of Mathematics and Systems Science, Chinese Academy of Science, Beijing, 100080, China

[#] Department of Radiology, The Methodist Hospital, Houston, Texas 77030

[§] Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Bethesda, Maryland 20892

[‖] Department of Electrical Engineering, Osaka Sangyo University, Osaka 574-8530, Japan

[⊥] Center for Cardiovascular Diagnostics and Prevention, CCF, Cleveland, Ohio 44195

Abstract

The mass spectrometry (MS) technology in clinical proteomics is very promising for discovery of new biomarkers for diseases management. To overcome the obstacles of data noises in MS analysis, we proposed a new approach of knowledge-integrated biomarker discovery using data from Major Adverse Cardiac Events (MACE) patients. We first built up a cardiovascular-related network based on protein information coming from protein annotations in Uniprot, protein–protein interaction (PPI), and signal transduction database. Distinct from the previous machine learning methods in MS data processing, we then used statistical methods to discover biomarkers in cardiovascular-related network. Through the tradeoff between known protein information and data noises in mass spectrometry data, we finally could firmly identify those high-confident biomarkers. Most importantly, aided by protein–protein interaction network, that is, cardiovascular-related network, we proposed a new type of biomarkers, that is, network biomarkers, composed of a set of proteins and the interactions among them. The candidate network biomarkers can classify the two groups of patients more accurately than current single ones without consideration of biological molecular interaction.

Keywords

Proteomics; Mass Spectrometry; MACE; Network biomarker; Cross validation; Systems biology

* To whom correspondence should be addressed. Tel, 1.713.441.8692; fax, 1.713.441.8696; XZhou@thms.org.

Supporting Information Available: Supplementary Figure 1 shows the analysis on the pair biomarker P10600-P61812 and Supplementary Figure 2 shows that MPO is also a candidate biomarker in our analysis. This material is available free of charge via the Internet at http://pubs.acs.org.

Introduction

Systematic proteomic studies to discover biomarkers are imperative since proteins perform the main cellular functions essential to signal transduction that lead to cell growth, differentiation, proliferation and death. Protein biomarkers have proven to extremely useful in providing valuable information that can be used during establishing a diagnosis or prognosis for a disease and developing targeted therapeutics.[1–9] Classic examples are Her2 protein for breast cancers diagnosis and treatment[10–12] and myeloperoxidase (MPO)[13] for predicting the risk of cardiovascular events.

Many diseases with a high incidence in the population, such as cardiac-cerebral vascular disease, cancer and diabetes, have a multifactorial basis. Though biomarker discovery resulted from intensive study of individual proteins, it is becoming increasingly clear that the predictive utility of individual biomarker proteins may be limited.[6,8,9,14] As an alternative, panels of proteins may be required to accurately gauge the level of perturbation of a biological system.[15,16]

Protein–protein interactions (PPIs) play a central role in many biological functions. For instance, signal cascades were mediated by PPIs of the signaling molecules from the exterior to interior of a cell.[17] This process, called signal transduction, plays a fundamental role in many biological processes and in many diseases. If interacted proteins maintain stable over time, they were called protein complexes, which are essential to biological processes.[18–20] Most works on biomarker discovery mainly focused on only single ones instead of interacting ones. Our work in this paper was desired to discover a new type of biomarkers with protein–protein interactions (e.g., network biomarker).

The principal enabling technology of proteomic discovery is mass spectrometry (MS).[21] However, the major obstacle to discover biomarkers from MS data is the data noises caused by instrument calibration. Although peak alignment and denoising processes can reduce the data noises greatly,[22,23] the data preprocessing will miss some candidate biomarkers only due to their bad performances in peak alignment. To avoid that, we used the established protein knowledge, such as protein annotations, PPI, and signaling pathway, to first filter out a cardiovascular-related protein network. A tradeoff was made between the protein knowledge and data noises in MS. By applying cardiovascular-related protein network without considerations of MS data, we can first identify some proteins really related to cardiovascular disease. Then, denoising processes and local peak alignments were applied to MS data for the identified proteins in cardiovascular-related network. Thus, the differently expressed proteins in MS data were identified by statistical methods. In this manner, we can select high-confident single biomarkers based on not only MS data, but also protein knowledge.

Here, Expression Difference Mapping using Ciphergen's SELDI ProteinChip technology was used to produce the MS data for cardiovascular disease. Plasma samples of two groups of patients, 60 MACEs (Major Adverse Cardiac Events) and 60 controls were used in this experiment (Materials and Methods). We proposed a new biomarker discovery method based on protein knowledge to discover biomarkers on the SELDI-TOF-MS data and derived a new type of biomarkers with protein–protein interactions (e.g., network biomarkers) that perform better performances than single biomarkers without any protein–protein interaction in patient classification, whose classification accuracy in 5-fold cross validation of SVM is nearly 80%.

Materials and Methods

Mass Spectrometry Data of Cardiovascular Disease

The plasma samples used in this study are the same as those used in Brenna's original work.[13] We use two groups of plasma samples: (1) MACE group of 60 patient samples, patients with chest pain and consistently negative Troponin T, but suffered MACE during the next 30-day or 6-month period, and (2) control group of 60 patient samples, patients with chest pain and consistently negative Troponin T and lived in next 5 years without any major cardiac events or death. To increase the coverage of proteins in SELDI protein profiles, the blood samples were fractionated with HyperD Q (anion ion exchange) into six fractions. The protein profiles of fractions 1, 3, 4, 5, and 6 were acquired with two SELDI Chips: IMAC and CM10. A total of 120 plasma samples, 24 reference samples, and 6 blanks were randomly divided into two groups, Group A and Group B, and were fractionated into six fractions using two 96-well plates containing anion exchange resin (Ciphergen, CA). Group A was processed in Day 1, while Group B was processed in Day 2. Two 96-well anion exchange resin plates were used to fractionate samples into six discrete fractions (pH 9 + flow through, pH 7, pH 5, pH 4, pH 3, and organic wash) as previously described.[37] Fractionation has been shown to greatly increase the number of proteins that can be resolved.

Protein spectra were obtained on immobilized metal affinity capture ProteinChip arrays coupled with copper (IMAC30-Cu2+, Ciphergen Biosystems, Inc., Fremont, CA) and weak cation exchange (CM10, Ciphergen Biosystems, Inc.) ProteinChip arrays. Fractions were subsequently profiled on both IMAC30-Cu2+ and CM10 protein arrays. Fraction 2 was not analyzed since experiments have shown that it contains little protein (data not shown). Samples from MACE and Control, as well as pooled samples from both groups and blank cases, were randomly distributed to the spots of ProteinChip arrays in Group A or Group B. All spectra were acquired in duplicate using two Bioprocessors, Bioprocessor 1 and Bioprocessor 2, which were processed at the same time using the same aliquot sample plate. The remaining portions of the samples were stored at −80 °C and were never reused for other ProteinChip arrays. ProteinChip arrays were analyzed utilizing a ProteinChip Reader, model PBSIIc (Ciphergen Biosystems, Inc.). Protein spectra were externally calibrated using the All-in-One Protein Standard II (Ciphergen Biosystems, Inc.) consisting of seven calibrants between 7 and 147 kDa. Data was collected between 0 and 200 kDa with the region between 2 and 20 kDa optimized. Spectra were generated by averaging 130 laser shots with a laser intensity (215–220) and a detector sensitivity (5–8) optimized for each fraction. MPO levels were measured with FDA approved assay (the assay name is CardioMPO), provided by Cleveland Clinic Foundation.

Protein Information

The protein–protein interaction data were downloaded from HPRD database (*Human Protein Reference Database* http://www.hprd.org/) in January, 2008. HPRD is composed of 18 796 proteins and 37 056 interactions (not including self-interaction). KEGG is a signal pathway database (*Kyoto Encyclopedia of Genes and Genomes* http://www.genome.jp/kegg/), which includes ‘Metabolism’, ‘Genetic Information Processing’, ‘Environmental Information Processing’, ‘Cellular Process’, ‘Human Disease’, and ‘Drug Development’ pathways. The signal pathways data were derived from KEGG in December, 2007. Uniprot (*Universal Protein Resource* http://www.pir.uniprot.org/) is the most comprehensive catalog of information of proteins. It is a central repository of protein sequence and function created by joining the information contained in UniProtKB/*Swiss-Prot* (http://www.ebi.ac.uk/swissprot/), *TrEMBL* (http://www.ebi.ac.uk/trembl/), and *PIR* (http://pir.georgetown.edu/). The knowledge data on proteins were drawn from Uniprot in January, 2008.

Cardiovascular-Related Network Construction

The cardiovascular-related network construction was completed by the following three steps. The first is to identify the cardiovascular-related proteins based on the knowledge of proteins, Uniprot. For most proteins, the important knowledge, such as related references and related diseases, can be found in Uniprot database. By searching the keyword 'cardiovascular' in the annotations of proteins of Uniprot, we got 76 proteins revealed to be closely related to cardiovascular disease. The next step is to build up the protein–protein interactions among these cardiovascular-related proteins. By checking the protein–protein interactions of these proteins in HPRD, we identified 17 proteins with at least one protein–protein interaction. The last step is to expand these 17 proteins to get a larger PPI network for cardiovascular disease using KEGG and HPRD. Until now, none of signal pathways for cardiovascular disease was available for systems biology study and its signal proteins or metabolisms were also hard to identify. Because of the identified important roles of these proteins in cardiovascular disease, it is reasonable to assume that the signal partners of these proteins in KEGG should also have their great contributions to the pathology of cardiovascular disease from the signal transduction viewpoint. In all signal proteins appearing in the signal pathways in KEGG, the interacting partners of identified 17 proteins have been expanded into the cardiovascular-related network. Thus, the cardiovascular-related network composed of 55 proteins with 122 protein–protein interactions was constructed based on the knowledge coming from Uniprot, HPRD, and KEGG databases (Figure 2).

Preprocessing and Local Peak Alignment in Mass Spectrometry

The data denoising and normalizing processes were applied to MS data got from SELDI-TOF-MS.[24–29] Comparing the mass of a protein with ones for every spectrometry data, we can find the mass location of the protein in the spectrometry data. However, the intensity in this mass location cannot just be simply considered as the expression for protein if the noises in the mass spectrometry are taken into account. Because of to the location of MS data may be moved in a small range by experiment noises, moving some peaks near to some location is very necessary for adjusting the accuracy of the data. The nearest peak in a window of −10 Da and +10 Da has been chosen as the peak of some location. If there is no peak in this window, the average of the intensities in this window will be considered as the intensity for the mass value.

P-Value Vector

For measuring distinct expressions for a protein in distinct fractions of mass spectrometry experiment, we proposed a *P*-value vector composed of *P*-values for distinct fractions in SELDI-TOF-MS (Figure 3). Instead of computing such a *P*-value for all fractions, the noise of mass spectrometry and distinct proteins remaining in different fractions of mass spectrometry experiments have been taken into account. If $\mathbf{V}_1 = (I_{c1}, I_{c2},\ldots, I_{c60})$ is the intensity vector of a protein for 60 control patients and $\mathbf{V}_2 = (I_{d1}, I_{d2},\ldots, I_{d60})$ is the intensity vector of a protein for 60 MACE disease patients in one fraction of SELDI-TOF-MS data, the *P*-value for $\mathbf{V}_1$ and $\mathbf{V}_2$ of the protein can be derived by using statistical methods (Student's *t* test, significant level: 0.05). Thus, all *P*-values for all fractions of mass spectrometry experiment can produce a *P*-value vector.

A 5-fold Cross Validation in SVM

A 5-fold cross-validation procedure in SVM was used to classify patients in MACE and controls. All intensity values in mass spectrum were normalized on [0, 1] interval. The training set for each split included 4/5 of the cases, while 1/5 of the samples were used as the test set and were not involved in training. In other words, the training set for each split includes 48 MACE patients and 48 control patients and the test set contains for each split 12 MACE patients and 12 control patients.[30–34]

Parameters in SVM

The SVM classifier used in this study is C-SVM where the kernel is Radial Basis Function kenel ($\exp(-\gamma \cdot |u - v|^2)$, $\gamma = 1/k$, k is the number of samples), and the parameter C is 1. In the cross validation of SVM, we chose the fold as 5.

Biomarker Identification

1. Candidate Single Biomarker Identification—The single biomarker discovery is based on the significantly different expressions of a protein in control and disease patients, or a significant low *P*-value for the protein's expressions. In our analysis, the *P*-value vector for every protein in cardiovascular disease has been used to single biomarker discovery. We searched all the *P*-values through the *P*-value vector to identify candidate biomarkers. If no significant low *P*-value was found, the protein would not be chosen as a candidate single biomarker. That means the protein does not represent significantly different expressions for control and disease patients in every fragment of MS data. In contrast, if at least one significantly low *P*-value in *P*-value vector can be found, it indicates that the protein should be a candidate for biomarkers.

2. Single Biomarker Identification—The biomarker identification in our analysis is based on not only its *P*-value, but also its performance on the 5-fold cross validation in SVM. For the identified candidate single biomarkers without any consideration on the protein–protein interaction, different number of them, 1, 2, and 3, was given into SVM to determine their performances in classification between control and disease patients. By this means, the best single biomarkers with not only best performance in SVM, but also significantly low *P*-value were chosen from candidate biomarkers. The intensities of single biomarkers used in SVM are just the original intensities in mass spectrometry data.

3. Pair Biomarker Identification—Distinguishing from discovering single biomarkers, pair biomarkers were identified based on not only the 5-fold cross validation in SVM but also the PPI network. Every pair biomarker is composed of two candidate single biomarkers and one protein–protein interaction between them. Then, distinct number of pair biomarkers, 1, 2, and 3, were put into SVM to show their performances for classification between control and disease patients. Thus, the best pair biomarkers with not only best performance in SVM, but also significantly low *P*-values were found using SVM. The intensity vectors of pair biomarkers used in SVM are the combined ones computed from original intensities of mass spectrometry as following,

Let P_1 and P_2 be the two interacted proteins involved in a pair biomarker. Denote p_1 and p_2 as the *P*-values of P_1 and P_2, respectively. And also denote $\mathbf{I}_1 = (I_{1,c_1}, I_{1,c_2},\ldots, I_{1,c_{60}}, I_{1,d_1}, I_{1,d_2},\ldots, I_{1,d_{60}})$ as the intensity vector of protein P_1 for not only 60 control patients, but also 60 MACE disease patients and $\mathbf{I}_2 = (I_{2,c_1}, I_{2,c_2},\ldots, I_{2,c_{60}}, I_{2,d_1}, I_{2,d_2},\ldots, I_{2,d_{60}})$ as the intensity vector of protein P_2 for both 60 control and 60 MACE disease patients, then the combined intensity vector $\mathbf{I}_{pair}$ for the pair biomarker is

$$\mathbf{I}_{pair} = \frac{\frac{1}{p_1}}{\frac{1}{p_1}+\frac{1}{p_2}}\mathbf{I}_1 + \frac{\frac{1}{p_2}}{\frac{1}{p_1}+\frac{1}{p_2}}\mathbf{I}_2 \quad (1)$$

Thus, the intensity vector of the most significant protein (with lowest *P*-value) can achieve the highest weight in the computing for $\mathbf{I}_{pair}$ due to the fact that the protein contributes to the pair-biomarker more than another relatively less significant protein.

4. Triple Biomarker Identification—Similar to pair biomarker, every triple biomarker is composed of three candidate single biomarkers and three protein interaction between every pair of them. Distinct number of triple biomarkers, 1, 2, and 3, were given into SVM to show their performances for classification between control and disease patients. Thus, the best triple biomarkers with not only best performance of 5-fold cross validation in SVM, but also significantly low *P*-values were found using SVM. The intensity vectors of triple biomarkers used in SVM are the combined ones computed from original intensities of mass spectrometry as following,

Let P_1, P_2 and P_3 be the three interacted proteins involved in a triple biomarker. Denotep_1, p_2 and p_3 as the *P*-values of P_1, P_2 and P_3, respectively. And also denote $\mathbf{I}_i = (I_{i,c1}, I_{i,c2},\ldots, I_{i,c60}, I_{i,d1}, I_{i,d2},\ldots, I_{i,d60})$ ($i = 1, 2, 3$) as the intensity vector of protein P_i for 60 control patients and 60 MACE disease patients, then the combined intensity vector $\mathbf{I}_{triple}$ for the triple biomarker is

$$\mathbf{I}_{triple}=\frac{\frac{1}{p_1}}{\frac{1}{p_1}+\frac{1}{p_2}+\frac{1}{p_3}}\mathbf{I}_1+\frac{\frac{1}{p_2}}{\frac{1}{p_1}+\frac{1}{p_2}+\frac{1}{p_3}}\mathbf{I}_2+\frac{\frac{1}{p_3}}{\frac{1}{p_1}+\frac{1}{p_2}+\frac{1}{p_3}}\mathbf{I}_3 \quad (2)$$

Thus, the intensity vector of the most significant protein (with lowest *P*-value) can achieve the highest weight in the computing for $\mathbf{I}_{triple}$ due to the fact that the protein contributes to the triple-biomarker more than other two relatively less significant proteins.

5. Multitype Biomarker Identification—Regardless of single, pair or triple biomarkers, all were given into SVM to train the best multitype biomarkers. Multitype biomarker is composed of different combinations of single ones, pair ones and triple ones. The best multitype biomarkers with both best performance of 5-fold cross validation in SVM and low *P*-values were found using SVM. The intensity vectors of multitype biomarkers input into SVM are the corresponding ones of single, pair and triple ones.

Results

The Scheme of Knowledge-Integrated Biomarker Discovery

The biomarker discoveries on distinct molecular levels, either mRNA[35,36] or protein, suffer from the data noises coming from expression instruments (microarray or mass spectrometry devices) or experimental design methods. Here, we proposed a novel biomarker discovery method based on protein knowledge to overcome the data noises in MS. Another extra advantage of such a biomarker discovery method can identify not only single biomarkers without any consideration of protein interactions, but also network biomarkers, a set of proteins with protein–protein interactions.

The knowledge-integrated biomarker discovery involves the integration of protein information from Uniprot, HPRD and KEGG, identification of candidate single biomarkers from MS data based on statistical methods, and identification of network biomarkers from protein–protein interaction network based on their performance in classification, as illustrated in Figure 1, and Materials and Methods.

Checking whether a protein is related to cardiovascular disease from the publications and disease annotations in Uniprot, the cardiovascular-related proteins were first identified. Through protein–protein interactions in HPRD and signal proteins in KEGG, the cardiovascular-related subnetwork was then constructed. With the use of a cardiovascular-

related subnetwork instead of whole protein–protein interaction network to discover biomarkers, it is ensured that more reliable proteins closely related to cardiovascular disease can enter into the process of biomarker identification so that the disturbance of noises coming from MS data can be easily avoided. Next, comparing with most previous works for discovering biomarkers from the peaks of MS data by machine learning methods, the present candidate single biomarkers were identified by statistical methods. Feature selection from aligned peaks is an indispensable step for most previous machine learning methods used in biomarker discovery. In contrast, feature selection is not a necessary step for us to identify biomarkers. Actually, for the limited cardiovascular-related proteins, the differently expressed proteins for control and disease patients can be easily identified by statistical method after the data preprocessing and local peak alignment in MS (Materials and Methods). Moreover, such a method can also easily avoid data noises by computing the *P*-values for different expressions of protein in control and disease patients. If too many noises instead of peaks occurring in the expressions of a protein, its *P*-value will not be significant low and thereby the protein will not be chosen as a candidate biomarker. Lastly, after the identification of candidate single biomarkers, network biomarkers were identified by their classification performance in SVM.

Cardiovascular-Related Protein Network Construction

The cardiovascular-related network integrated most protein information coming from Uniprot, HPRD, and KEGG databases (Materials and Methods), as illustrated in Figure 2. First, by checking publication and protein annotations in Uniprot, 76 cardiovascular-related proteins have been identified. Then, to derive the protein–protein interactions among these proteins, HPRD has been taken into consideration. Seventeen proteins of the 76 identified cardiovascular-related proteins appear in the HPRD database, which means that these 17 proteins take part in the protein–protein interactions. In consideration of the important roles of these proteins in the pathology of this disease, they should also be essential to the signal transduction for cardiovascular system, and thus, the protein interaction partners in signal proteins of KEGG were expanded into the cardiovascular-related network. At last, the cardiovascular-related network was constructed with 17 proteins identified from Uniprot and HPRD and their 38 signal partners expanded from KEGG signal proteins (Figure 2).

MS-Based Biomarker Discovery (Candidate Single Biomarkers)

The aim of MS-based biomarker discovery is to identify proteins differentially expressed in the serum or plasma of cardiovascular disease patients. A new and emerging technology, proteomics, has the potential to identify protein molecules in a high-throughput discovery approach in patient's serum. Electrospray ionization mass spectrometry, surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF-MS) technology can identify patterns or changes in thousands of proteins and can globally analyze almost all small molecular weight proteins in complex solutions such as serum or plasma.

In the analysis of MS data, researchers usually use a common protocol that consists of preprocessing, peak detection, and peak alignment, especially for those using classification to select biomarkers, because MS data, that is, spectra, may be affected by errors and noise as a result of sample preparation and instrument approximation or the mass/charge axis shift. Previous works paid more attention on the peak alignment and peak detection to ensure the good performance of classification algorithm. Their hypothesis is the peaks are different from noises on MS. One obvious disadvantage of these methods is that proteins may be missed merely due to the bad peak alignment or no detected peaks on some data. Alternative method for dealing with peaks in MS has been proposed in our analysis. For a protein, its mass in the mass/charge axis was first identified and then its nearest peak or mean of the masses in the window of −10 Da and +10 Da of its mass was identified as one of the expression intensities

for the protein. Thus, the intensity vectors for different conditions can be derived from 60 controls and 60 disease patients.

Distinguishing from the machine learning methods based on the peaks of MS data, our method does not merely focus on the proteins chosen from peak alignment, whose mass/charge is exactly located at certain location of the mass/charge axis where the intensities are just the peaks of the MS data sets, but focus on those proteins discovered from their protein knowledge, whose intensities may be composed of not only peaks but also some nonpeak values. Whether nonpeak intensities are data noises is not essential to our biomarker discovery. To elucidate the significantly distinct expressions of a protein between control and cardiovascular disease patients, we adopt a statistical method instead of machine learning to discover biomarkers. If the intensity vectors for a protein are affected by the data noises significantly, the *P*-value to evaluate the different expressions of the protein will not be significantly low and the peptide will not provide evidence for its protein to prove that it is a candidate biomarker discovered from the MS data. In other words, the protein with relatively low *P*-value implies that its intensities should not be disturbed by the data noises greatly as well as they are differently expressed in control and disease patients, and thereby be considered as a candidate biomarker.

The five fraction profiles acquired from SELDI protein chip were washed using distinct washing chemicals (Materials and Methods). Considering the different proteins remaining in the different fractions of MS data, we introduced a *P*-value vector to evaluate the different expressions of a protein in all fractions of MS (Figure 3). If a protein does not lie in some fraction of MS, the majority of its intensities on MS will be composed of noises and then the protein's *P*-value for control and disease patients will be not significantly low. On the other hand, if a protein does lie in some fraction and is a candidate biomarker displaying significantly distinct expressions between control and disease patients, its *P*-value should be relatively low in this fraction of MS data. By searching through the *P*-value vector, if no significantly low *P*-value can be found, we can firmly say that the protein is not a candidate biomarker. Otherwise, at least one low *P*-value in *P*-value vector can be found, which implies that the protein should be a candidate biomarker. Thus, for every protein in cardiovascular-related network, we can easily identify whether it is really a candidate biomarker by its *P*-value vector. Totally, 31 proteins were found with significant *P*-value vectors in cardiovascular-related network.

Network Biomarker Identification in Cardiovascular-Related Network

The protein–protein interaction information in cardiovascular-related network was not considered in the identification process of candidate single biomarkers for MS data. The interactions between proteins are important for many biological functions. Because of the essential roles of protein interactions in biological processes, we integrated the protein–protein interaction information into the biomarker discovery process. We revealed a new type of biomarkers, called network biomarkers, composed of a set of proteins and the protein interactions among them.

Network biomarkers considered in our analysis can be divided into three types, single biomarker without any protein–protein interaction, pair-biomarker with two proteins and one protein–protein interaction, triple-biomarker with three proteins and three protein–protein interactions. After the identification of candidate single biomarkers using *P*-value vector, the intensity vectors of control and disease patients, respectively, for a protein can be identified by the lowest *P*-value. For a single biomarker, its intensities are just the original ones from the MS data; however, the intensities for a pair-biomarker and a triple-biomarker are *P*-value weighted summation of the intensities of their composed single proteins (Materials and Methods).

Classification based on 5-fold cross validation of SVM was applied to identify network biomarkers based on their classification performances. First, different number, 1 or 2 or 3, of same type of network biomarkers was put into SVM. By their performance, we can easily identify the best ones for patient classification. We found that the best performance for single biomarkers, P06858, P35555, and Q07954, is 71.67%, while the best performances for pair biomarkers, P04180-P01023, P10600-P61812, and P11802-P36897, and triple biomarkers, Q04771-O14920-P36897, P36897-P61812-P10600, and P35555-P15502-P07585, are 77.50% and 72.50%, respectively. In Table 1, the results show that the performances for network biomarkers considering protein–protein interaction information, that is, pair-biomarkers and triple-biomarkers, are higher than the single ones without any protein–protein interaction information. Next, different number, 1 or 2 or 3, of combinations of multiple types of network biomarkers was given into SVM (Table 2). By the same means, we found that the best classification performance, 78.33%, occurred in the combination of network biomarkers, Q07954-Q01023, P63151-P36897, and P35555-P15502-P07585 (Figure 4), which can be considered as the best network biomarkers for cardiovascular diseases.

To analyze and explain the performances of different type of biomarkers in cross validation, we compared the ROC curves of three types of biomarkers, that is, the best single biomarker (P06858, P35555, Q07954), pair biomarker (P04180-P01023, P10600-P61812, P11802-P36897), and multitype biomarker (Q07954-P01023, P63151-P36897, P35555-P15502-P07585) (Figure 5). We found that the AUCs (Area Under ROC Curve) of these three types of biomarkers, that is, single, pair, and multitype, are 71.26, 79.68, and 80.58, respectively. By comparing the AUCs of these three types of biomarkers, we found the biomarkers with protein interactions (pair biomarker and multitype biomarker) are better than the single biomarker without consideration on protein interaction information.

Discussion

The knowledge-integrated biomarker discovery method integrated most protein information of their publications, signal transductions and protein–protein interactions into the biomarker discovery process through cardiovascular disease related network. We used a statistical method to avoid the disturbing of data noises (not peak data) and to select the candidate single biomarker from MS data. By the combination of protein–protein interactions among these candidate single biomarkers, we defined a novel type of biomarkers with protein–protein interactions, called network biomarker. According to the performance of network biomarkers in the 5-fold cross validation of SVM, we found that network biomarkers can classify the cardiovascular patients from control patients more accurately. Therefore, the advantages of the knowledge-integrated biomarker discovery include not only easily avoiding data noises by cardiovascular-related network, but also deriving high-confident network biomarkers.

Our method started from the cardiovascular-related network identified by protein information. This step is to ensure that most known protein knowledge of the cardiovascular disease can be integrated into our biomarker discovery so that the biomarker discovery process can be less disturbed by the errors existing in MS data. We made a tradeoff between protein known knowledge and data noises of MS data in the biomarker discovery process. Aided by protein information, the biomarkers discovered from MS data may suffer from some data noises. The the high performances of discovered network biomarkers in classification implies that the integration of protein knowledge into biomarker discovery is a very important strategy for the discovery of high confident biomarkers from MS data with noises.

The identification of single candidate biomarkers from cardiovascular related network is based on statistical method. Actually, the biomarkers are defined as a small subset of differentially expressed proteins from a large volume of profiling data and used as targets for further

development in molecular diagnostics and therapeutics. The statistical method used in our approach has itself superiority in discovering biomarkers. One advantage is that the low *P*-values computed from the intensities of proteins in both control and disease patients have the ability to identify the differently expressed proteins. Another is that statistical methods can easily avoid the disturbance from the data noises. If the intensities of a protein are mainly composed of data noises instead of peaks, its *P*-value will not be significantly low and it will not be chosen as a candidate single biomarkers. Most importantly, such a method can pick up some proteins with significantly low data noises in spite of their bad peak alignments caused by instrument calibration in MS.

Most previous researches mainly focused on the single biomarker discovery while our work considered the network biomarkers based on protein–protein interactions. A complex pathology of a disease could not be easily explained by single proteins or single biomarkers. From systems biology's viewpoint, we should resort to the network biomarkers, which may correspond to some protein complexes or signal pathways essential to discover the underlying mechanism of some diseases. Our work was desired to make a great attempt in this direction. Definitely, from the classification results of network biomarkers on 5-fold cross validation of SVM, we can firmly say that network biomarkers are more reliable for predicting the risk of cardiovascular events.

We not only set up the classification experiments on single, pair, and triple biomarkers, but also do the same numerical experiments on the subnetworks with four, five, and six proteins, illustrated in Figure 6. Comparing the accuracies in Figure 6, we found that the subnetworks with more than 3 proteins have relatively low classification accuracies than single, pair, and triple biomarkers. Therefore, it is reasonable to choose the single, pair, and triple biomarkers for our analysis and thereby their combination can consistently provide the best performance.

The results for same type and multitype network biomarkers in classification have been shown. One may notice that a triple biomarker for same type classification analysis in Table 1, such as P35555-P15502-P07585, also appears in Table 2 for multitype classification analysis. However, the pair biomarker, P10600-P61812, for same type classification analysis in Table 1 cannot be found in Table 2 for multitype classification analysis. To identify the roles of pair biomarker, P10600-P61812, in classification, we recomputed the results of involvement of it into SVM and found that the performance in multitype classification analysis is also as high as 75% (Supplementary Figure 1). Thus, undoubtedly, the network biomarkers, that is, pair ones and triple ones, can derive high performance regardless of same type or multitype classification analysis.

To indicate the roles of protein information in the biomarker discovery, we also compared our method to the general biomarker discovery method without assistance of protein knowledge. Here, we adopted a general biomarker discovery method: peak detection from mass spectrums, peak alignment, and doing classification on the found peaks on the treated MS data (baseline removal, denoised, normalized). We found that the best classification accuracy of this method is only 75.00% and it is not better than that of multiple biomarkers, nearly 80%.

Additionally, MPO has been identified as a biomarker for cardiovascular disease.[13] By our statistical methods, the significant difference of MPO peaks was found between controls and MACE patients, showing that MPO is a candidate single biomarker for cardiovascular events without consideration of protein interactions (*P*-value less than 0.01) (Supplementary Figure 2). Because it has no interaction partner, we did not put it into our network biomarker discovery process.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is funded by the Bioinformatics Core Research Grant at The Methodist Research Institute, Cornell University. Dr. Zhou is partially funded by The Methodist Hospital Scholarship Award. He and Dr. Wong are also partially funded by NIH grants R01LM08696, R01LM009161, and R01AG028928.

References

1. Malik G, Rojahn E, Ward MD, Gretzer MB, Partin AW, Semmes OJ, Veltri RW. SELDI protein profiling of dunning R-3327 derived cell lines: identification of molecular markers of prostate cancer progression. Prostate 2007;67(14):1565–75. [PubMed: 17705230]
2. Oh JH, Nandi A, Gurnani P, Knowles L, Schorge J, Rosenblatt KP, Gao JX. Proteomic biomarker identification for diagnosis of early relapse in ovarian cancer. J Bioinf Comput Biol 2006;4(6):1159–79.
3. Paweletz CP, Liotta LA, Petricoin EF 3rd. New technologies for biomarker analysis of prostate cancer progression: Laser capture microdissection and tissue proteomics. Urology 2001;57(4 Suppl. 1):160–3. [PubMed: 11295617]
4. Vlahou A, Schellhammer PF, Mendrinos S, Patel K, Kondylis FI, Gong L, Nasim S, Wright GL Jr. Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. Am J Pathol 2001;158(4):1491–502. [PubMed: 11290567]
5. Wulfkuhle JD, McLean KC, Paweletz CP, Sgroi DC, Trock BJ, Steeg PS, Petricoin EF. New approaches to proteomic analysis of breast cancer. Proteomics 2001;1(10):1205–15. [PubMed: 11721633]
6. Cox J, Mann M. Is proteomics the new genomics. Cell 2007;130(3):395–8. [PubMed: 17693247]
7. Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. Nature 2008;452(7187):571–9. [PubMed: 18385731]
8. Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nat Biotechnol 2006;24(8):971–83. [PubMed: 16900146]
9. Simpson RJ, Bernhard OK, Greening DW, Moritz RL. Proteomics-driven cancer biomarker discovery: looking to the future. Curr Opin Chem Biol 2008;12(1):72–7. [PubMed: 18295612]
10. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, Baselga J, Norton L. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 2001;344(11):783–92. [PubMed: 11248153]
11. Asgeirsson KS, Agrawal A, Allen C, Hitch A, Ellis IO, Chapman C, Cheung KL, Robertson JF. Serum epidermal growth factor receptor and HER2 expression in primary and metastatic breast cancer patients. Breast Cancer Res 2007;9(6):R75. [PubMed: 17976236]
12. Pal SK, Pegram M. HER2 targeted therapy in breast cancer…beyond Herceptin. Rev Endocr Metab Disord 2007;8(3):269–77. [PubMed: 17899385]
13. Brennan ML, Penn MS, Van Lente F, Nambi V, Shishehbor MH, Aviles RJ, Goormastic M, Pepoy ML, McErlean ES, Topol EJ, Nissen SE, Hazen SL. Prognostic value of myeloperoxidase in patients with chest pain. N Engl J Med 2003;349(17):1595–604. [PubMed: 14573731]
14. McGuire JN, Overgaard J, Pociot F. Mass spectrometry is only one piece of the puzzle in clinical proteomics. Briefings Funct Genomics Proteomics 2008;7(1):74–83.
15. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5(2):101–13. [PubMed: 14735121]
16. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature 1999;402(6761 Suppl):C47–52. [PubMed: 10591225]
17. Pawson T, Nash P. Protein-protein interactions define specificity in signal transduction. Genes Dev 2000;14(9):1027–47. [PubMed: 10809663]

18. Azpiazu I, Gautam N. Role of G protein beta gamma complex in receptor-G protein interaction. Methods Enzymol 2002;344:112–25. [PubMed: 11771376]
19. Banci L, Bertini I, Cantini F, Felli IC, Gonnelli L, Hadjiliadis N, Pierattelli R, Rosato A, Voulgaris P. The Atx1-Ccc2 complex is a metal-mediated protein-protein interaction. Nat Chem Biol 2006;2 (7):367–8. [PubMed: 16732294]
20. Das D, Scovell WM. The binding interaction of HMG-1 with the TATA-binding protein/TATA complex. J Biol Chem 2001;276(35):32597–605. [PubMed: 11390376]
21. Issaq HJ, Veenstra TD, Conrads TP, Felschow D. The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification. Biochem Biophys Res Commun 2002;292(3):587–92. [PubMed: 11922607]
22. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 2004;20(5):777–85. [PubMed: 14751995]
23. Karpievitch YV, Hill EG, Smolka AJ, Morris JS, Coombes KR, Baggerly KA, Almeida JS. PrepMS: TOF MS data graphical preprocessing tool. Bioinformatics 2007;23(2):264–5. [PubMed: 17121773]
24. Malyarenko DI, Cooke WE, Adam BL, Malik G, Chen H, Tracy ER, Trosset MW, Sasinowski M, Semmes OJ, Manos DM. Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. Clin Chem 2005;51(1):65–74. [PubMed: 15550476]
25. Bensmail H, Golek J, Moody MM, Semmes JO, Haoudi A. A novel approach for clustering proteomics data using Bayesian fast Fourier transform. Bioinformatics 2005;21(10):2210–24. [PubMed: 15769836]
26. Perrin C, Walczak B, Massart DL. The use of wavelets for signal denoising in capillary electrophoresis. Anal Chem 2001;73(20):4903–17. [PubMed: 11681466]
27. Wang P, Tang H, Zhang H, Whiteaker J, Paulovich AG, McIntosh M. Normalization regarding non-random missing values in high-throughput mass spectrometry data. Pac Symp Biocomput 2006:315–26. [PubMed: 17094249]
28. Alfassi ZB. On the normalization of a mass spectrum for comparison of two spectra. J Am Soc Mass Spectrom 2004;15(3):385–7. [PubMed: 14998540]
29. Marcuson R, Burbeck SL, Emond RL, Latter GI, Aberth W. Normalization and reproducibility of mass profiles in the detection of individual differences from urine. Clin Chem 1982;28(6):1346–8. [PubMed: 7074944]
30. Bro R, Kjeldahl K, Smilde AK, Kiers HA. Cross-validation of component models: a critical look at current methods. Anal Bioanal Chem 2008;390(5):1241–51. [PubMed: 18214448]
31. Sundararajan S, Shevade S, Keerthi SS. Fast generalized cross-validation algorithm for sparse model learning. Neural Comput 2007;19(1):283–301. [PubMed: 17134326]
32. Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics 2005;21(9):1979–86. [PubMed: 15691862]
33. Mavroforakis ME, Theodoridis S. A geometric approach to support vector machine (SVM) classification. IEEE Trans Neural Networks 2006;17(3):671–82.
34. Byvatov E, Schneider G. SVM-based feature selection for characterization of focused compound collections. J Chem Inf Comput Sci 2004;44(3):993–9. [PubMed: 15154767]
35. Allantaz F, Chaussabel D, Banchereau J, Pascual V. Microarray-based identification of novel biomarkers in IL-1-mediated diseases. Curr Opin Immunol 2007;19(6):623–32. [PubMed: 18036805]
36. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol 2007;3:140. [PubMed: 17940530]
37. Koopmann J, Zhang Z, White N, Rosenzweig J, Fedarko N, Jagannath S, Canto MI, Yeo CJ, Chan DW, Goggins M. Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry. Clin Cancer Res 2004;10(3):860–8. [PubMed: 14871961]

Abbreviations

MS	mass spectrometry
MACE	Major Adverse Cardiac Events
SVM	support vector machine
SELDI-MS-TOF	surface-enhanced laser desorption/ionization time-of-flight mass spectrometry
PPI	protein–protein interaction

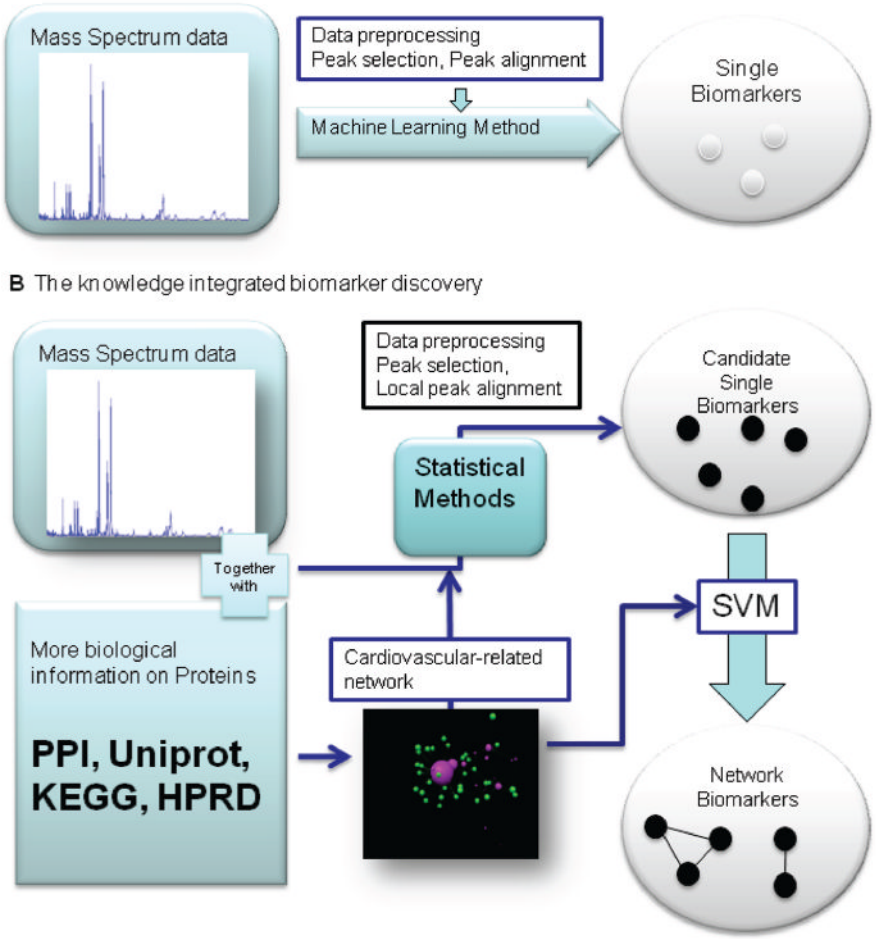

Figure 1.
The scheme for knowledge integrated biomarker discovery.

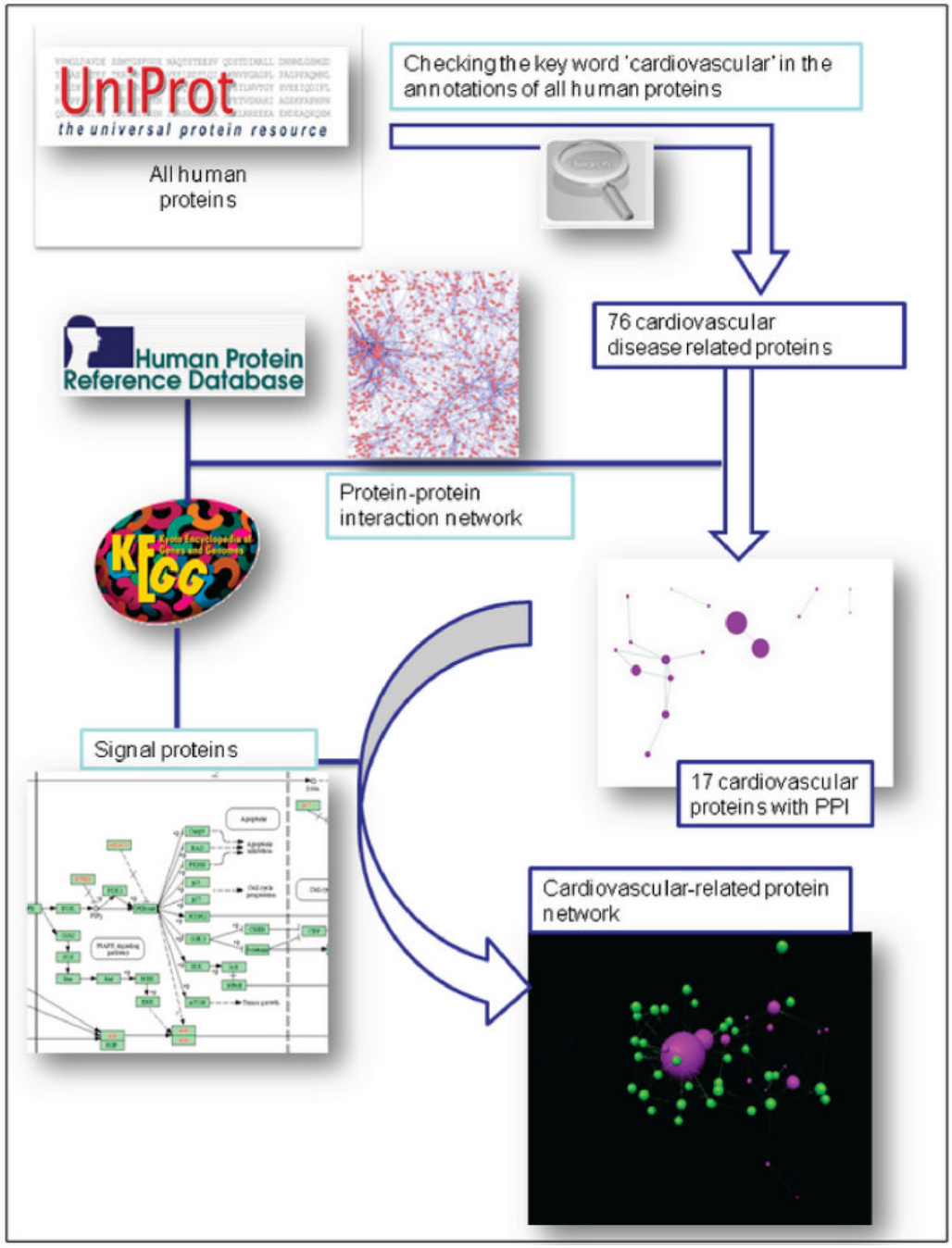

Figure 2.
Flowchart for cardiovascular-related protein network.

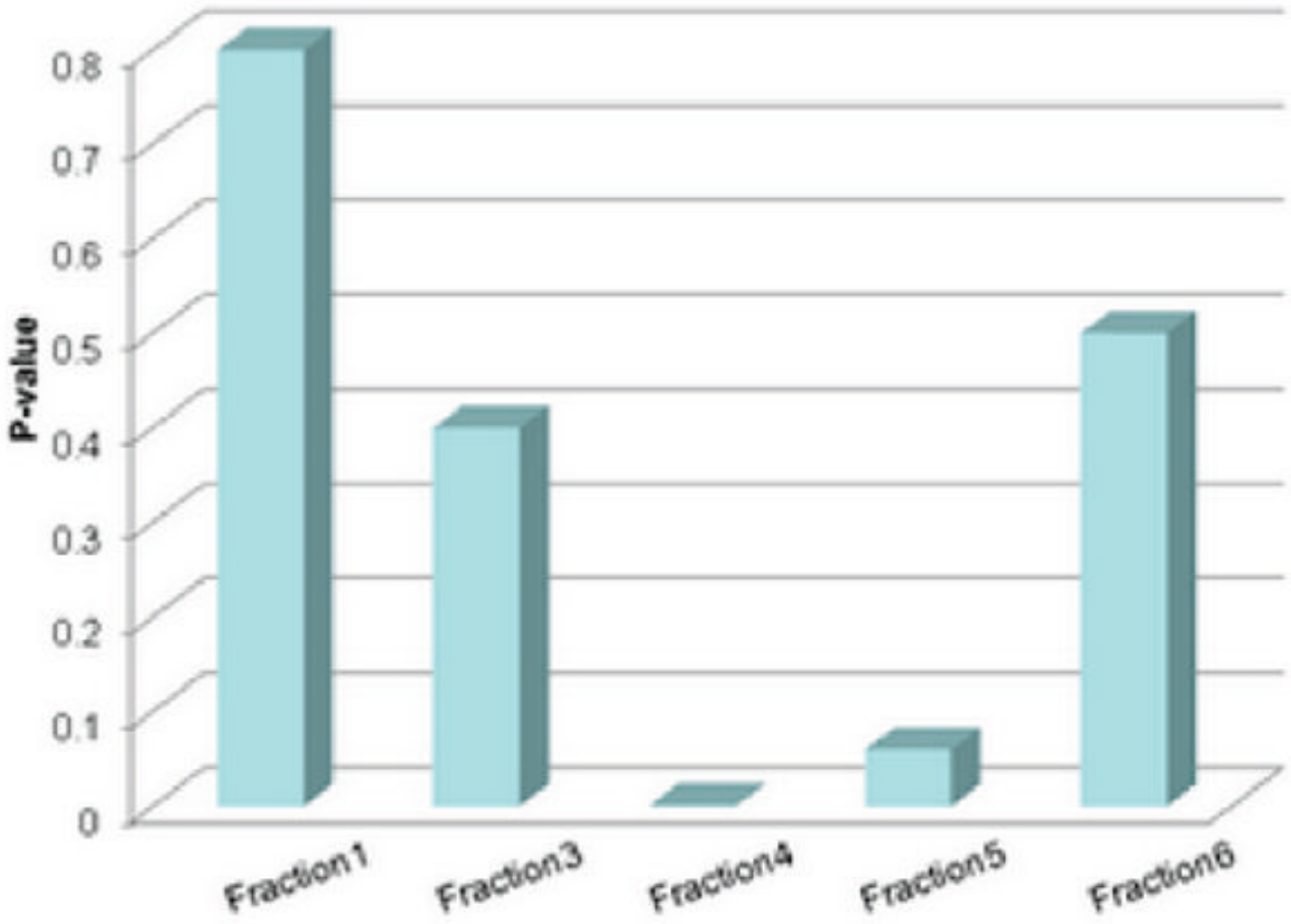

Figure 3.
The *P*-value vector for a protein in different factions of mass spectrometry data.

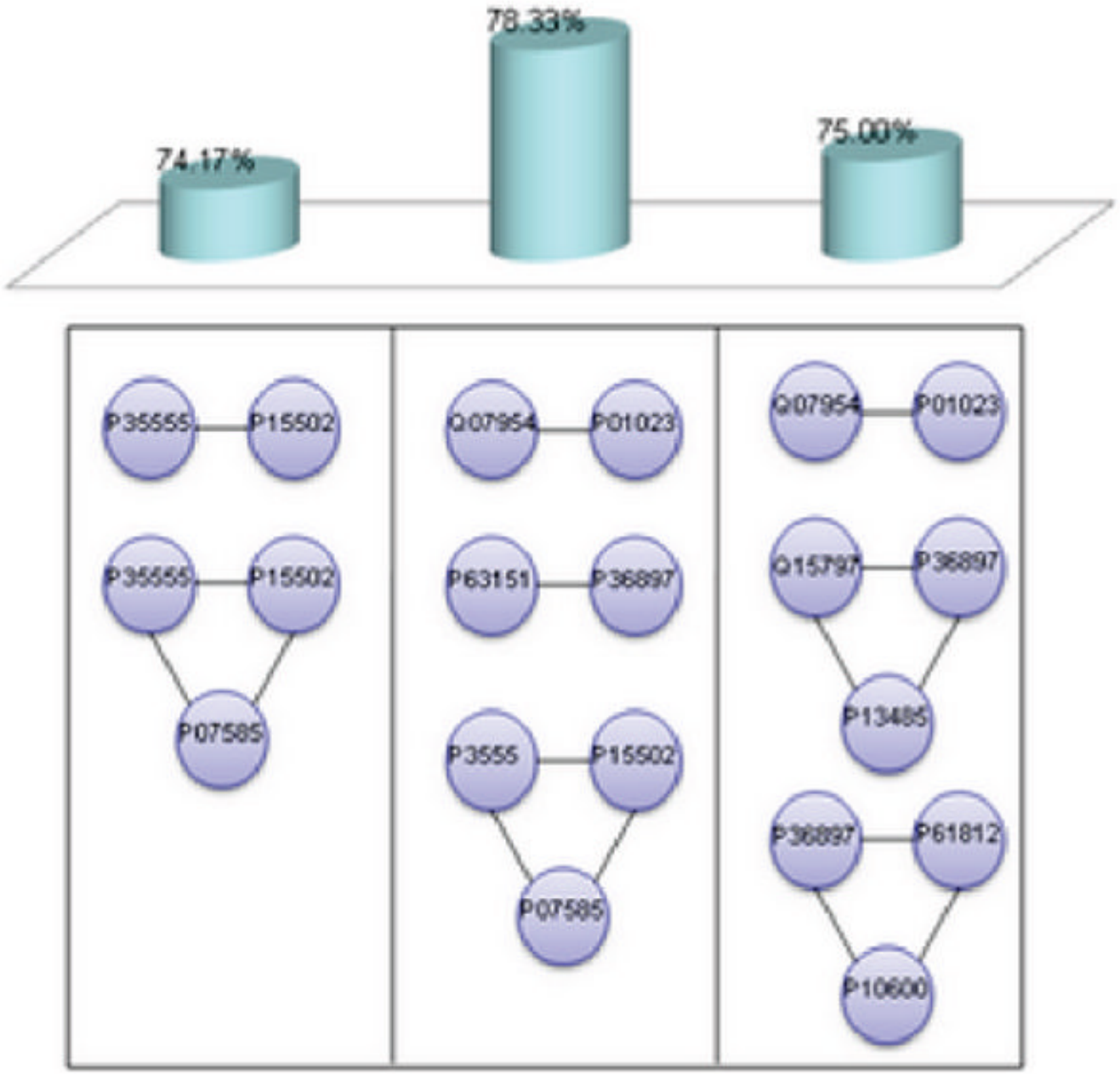

Figure 4.
The best classification accuracy is obtained by multitype biomarker.

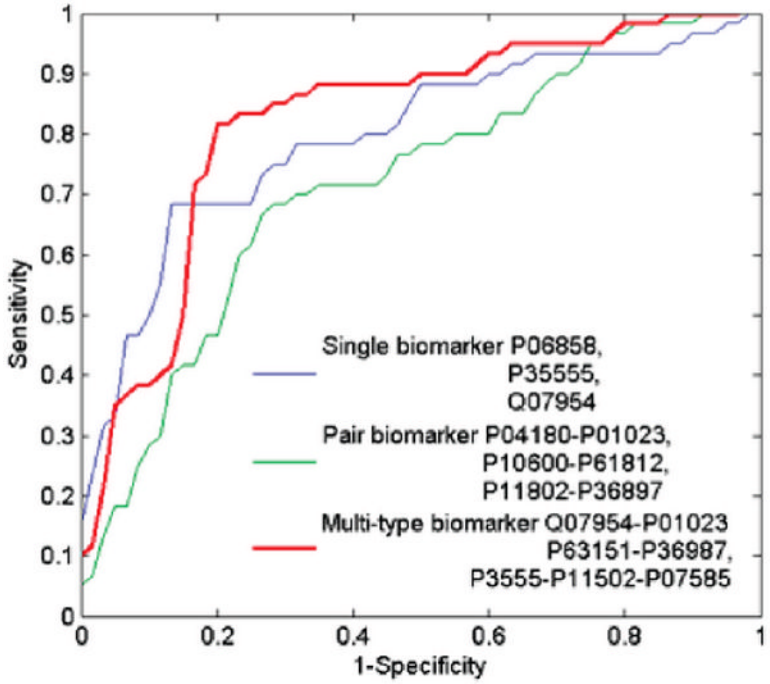

Figure 5.
The ROC curves for the chosen biomarkers.

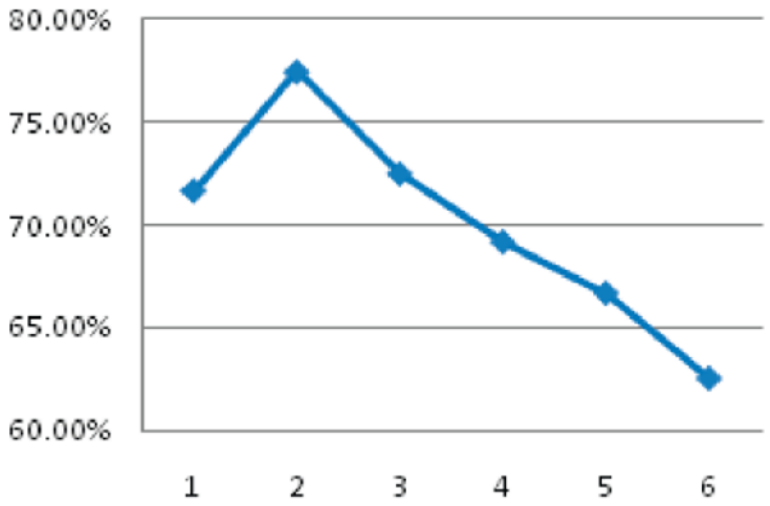

Figure 6.
The best classification performances for the subnetworks with different proteins.

Table 1

The Classification Accuracies for Single, Pair, Triple Biomarkers[a]

	numbers of biomarkers given in SVM		
biomarker types	**number = 1**	**number = 2**	**number = 3**
Accuracy	62.5%	66.67%	71.67%
Single-biomarker	P55058	P35222	P06858
		Q16671	P35555
			Q07954
Accuracy	64.17%	71.67%	77.50%
Pair-biomarker	O75052-Q07954	P10600-P61812	P04180-P01023
		P35222-Q95405	P10600-P61812
			P11802-P36897
Accuracy	65.83%	68.33%	72.50%
Triple-biomarker	P35555-P15502- P07585	P36897-P61812- P10600	Q04771-O14920- P36897
		P35555-P15502-P07585	P36897-P61812- P10600
			P35555-P15502-P07585

[a]The symbol ‘A-B’ means the pair biomarker and ‘A-B-C’ means the triple biomarker.

Table 2

The Classification Accuracies for Multitype Biomarkers

	numbers of biomarkers given in SVM		
biomarker types	**number = 1**	**number = 2**	**number = 3**
Accuracy	71.67%	73.33%	75.83%
Single and pair[a]	Q9HAU4	P61812	P61812
	P35555-P15502	Q63151	Q07954-P01023
		P06858-Q07954	P63151-P36897
Accuracy	70.00%	73.33%	75.83%
Single and triple[a]	P15502	P61812	P63151
	Q04771-O14920-P36897	P63151	P35555-P15502-P07585
		P35555-P15502-P07585	P36897-P61812-P10600
Accuracy	74.17%	78.33%	75.00%
Pair and triple[a]	P35555-P15502	Q07954-P01023	Q07954-P01023
	P35555-P15502-P07585	P63151-P36897	Q15797-P36897-Q13485
		P35555-P15502-P07585	P36897-P61812-P10600

[a]The symbol ‘A-B’ means the pair biomarker and ‘A-B-C’ means the triple biomarker.