RESEARCH ARTICLE

Identification of biomarkers for risk stratification of cardiovascular events using genetic algorithm with recursive local floating search

Xiaobo Zhou[1, 2], Honghui Wang[3]*, Jun Wang[1, 4]*, Yuan Wang[1], Gerard Hoehn[3], Joseph Azok[3], Marie-Luise Brennan[5], Stanley L. Hazen[5], King Li[2], Shih-Fu Chang[4] and Stephen T. C. Wong[1, 2]*

[1] Center for Biotechnology and Informatics, The Methodist Hospital Research Institute & Cornell University, Houston, TX, USA
[2] Department of Radiology, The Methodist Hospital, Weill Cornell College of Medicine, Houston, TX, USA
[3] Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Bethesda, MD, USA
[4] Department of Electrical Engineering, Columbia University, New York, NY, USA
[5] Center for Cardiovascular Diagnostics and Prevention, CCF, Cleveland, OH, USA

Received: September 12, 2007
Revised: December 10, 2008
Accepted: December 15, 2008

Conventional biomarker discovery focuses mostly on the identification of single markers and thus often has limited success in disease diagnosis and prognosis. This study proposes a method to identify an optimized protein biomarker panel based on MS studies for predicting the risk of major adverse cardiac events (MACE) in patients. Since the simplicity and concision requirement for the development of immunoassays can only tolerate the complexity of the prediction model with a very few selected discriminative biomarkers, established optimization methods, such as conventional genetic algorithm (GA), thus fails in the high-dimensional space. In this paper, we present a novel variant of GA that embeds the recursive local floating enhancement technique to discover a panel of protein biomarkers with far better prognostic value for prediction of MACE than existing methods, including the one approved recently by FDA (Food and Drug Administration). The new pragmatic method applies the constraints of MACE relevance and biomarker redundancy to shrink the local searching space in order to avoid heavy computation penalty resulted from the local floating optimization. The proposed method is compared with standard GA and other variable selection approaches based on the MACE prediction experiments. Two powerful classification techniques, partial least squares logistic regression (PLS-LR) and support vector machine classifier (SVMC), are deployed as the MACE predictors owing to their ability in dealing with small scale and binary response data. New preprocessing algorithms, such as low-level signal processing, duplicated spectra elimination, and outliner patient's samples removal, are also included in the proposed method. The experimental results show that an optimized panel of seven selected biomarkers can provide more than 77.1% MACE prediction accuracy using SVMC. The experimental results empirically demonstrate that the new GA algorithm with local floating enhancement (GA-LFE) can achieve the better MACE prediction performance comparing with the existing techniques. The method has been applied to SELDI/MALDI MS datasets to discover an optimized panel of protein biomarkers to distinguish disease from control.

Keywords:
Biomarker panel / Genetic algorithm / Major adverse cardiac events (MACE) / Mass spectrometry / Recursive local floating search

Correspondence: Professor Xiaobo Zhou, Center for Biotechnology and Informatics, The Methodist Hospital Research Institute & Cornell University, Houston, TX 77030, USA
E-mail: xzhou@tmhs.org
Fax: +1-7134418696

Abbreviations: FDA, food and drug administration; **GA**, genetic algorithm; **MACE**, major adverse cardiac event; **MPO**, myeloperoxidase; **PCA**, ProteinChip arrays; **PLS-LR**, partial least squares logistic regression; **ROC**, receiver operating characteristic

* These authors have contributed equally to this work.

1 Introduction

Using the protein level of myeloperoxidase (MPO) and other known cardiovascular biomarkers, which are measured from the blood samples of patients to predict the early risk of cardiovascular disease, has recently been studied by Brennan *et al.* [1]. In their work, the researchers investigated 604 patients who presented in emergency room with chest pain and showed that the MPO to be a new biomarker for the prediction of the risk of major adverse cardiac event (MACE) in the ensuing 30-day and 6-month period with an accuracy better than 60% for these patients with consistent negative Troponin T.

The biomarkers generated from MS were investigated for disease diagnosis and prognosis, such as ovarian cancer identification [2]. In our study for MACE prediction, the high-throughput SELDI MS generates more than one thousand proteins or protein fragment peaks with multiple ProteinChips. The simplicity and concision requirement for the development of immunoassay cannot adapt to the complexity of the prediction model with all the generated biomarkers [3]. Therefore, the emergent task is to discover a panel of optimal biomarkers, which should have the strong relevance to MACE prediction and less redundancy among biomarkers themselves.

The study of biomarker discovery is intrinsically linked to the variable selection methodology in the fields of machine learning and pattern recognition. The mathematical definition of a variable selection problem is to select a subset $\{Y_1, Y_2,\ldots Y_M\}$ from a feature set $\{X_1, X_2,\ldots X_N\}$ to optimize a predefined fitness function $J(y_1, y_2,\ldots y_M)$. The values of N and M represent the number of elements in the original and target variable sets, and usually $M<<N$. For a classification or prediction problem, the fitness function is typically selected as the accuracy of the predictor. The exhausted search method requires going through all the possible $\mathrm{C}_{\mathrm{N}}^{\mathrm{M}}$ combinations, which could achieve exponential computation complexity $O(D^M)$, a NP-hard problem. Therefore the full search is often not feasible owing to extremely high computation cost incurred. Generally speaking, there are two categories of variable selection techniques, *filter* and *wrapper.* The filter based techniques use the data statistic characteristics as the criteria to find a subset of features which can keep most class relevance while reducing variable redundancy [4]. The wrapper techniques use the accuracy of the predictor as the criteria and then apply certain optimization techniques to obtain the global or local optima of the criteria function. Genetic algorithm (GA), originally proposed by Holland, is a conventional wrapper based method that mimics the evolutionary process of the survival of the fittest [5]. Particularly, in this biomarker selection problem, a population of abstract representations (usually called chromosomes) of candidate solutions (called individuals, representing a subset of biomarkers) evolves to achieve better prediction performance. Commonly, the chromosomes are encoded in binary strings of 0s and 1s, where 1s represent selected biomarkers and 0s represent omitted biomarkers (see Supporting Information material). In the following description, we use the term chromosome to denote a single candidate solution without specific declaration.

However, the standard GA cannot generate the satisfactory panel of biomarker selection results because its randomized search ability severely degenerates when encountering the high-dimensionality of the biomarker candidate set and the dramatically increased dimensionality of the target subset. Therefore, we proposed to use the recursively local floating search technique to enhance the individuals for each generation of evolution. Linear discriminate analysis is selected as the fitness function because of its efficient computation. With the discovered panel of biomarkers, the Partial Least Square Logistic Regression (PLS-LS) [6] and support vector machine classifier (SVMC) [7, 8] are applied as the predictor for validation of MACE prediction. The prediction accuracy is approximated using leave-one-out cross validation estimation [9].

The comparison studies on the accuracy of MACE prediction are conducted as follows: (a) comparing with the single MPO value, which is measured by Hazen's group (Cleveland Clinic Foundation) with an FDA approved assay called CardioMPO (tm); (b) comparing with partial datasets of 377 biomarkers, including MPO value; (c) comparing with the standard GA [10, 11] to provide the proof of the performance improvement; (d) comparing with sequential floating forward search (SFFS), which is reviewed as the best sequential search method [12], and, finally; (e) comparing with biomarker ranking based methods, like a *t*-test for example. The experimental results empirically demonstrated that the proposed method of GA algorithm with local floating searching embedding (GA-LFE) achieves better MACE prediction performance than the existing techniques. With a selected panel of only seven biomarkers, the prediction accuracy estimated by leave-one-out cross validation strategy is greater than 77%.

2 Materials and methods

2.1 Materials

The plasma samples used in this study are the same as those used in the original work of Brennan *et al.* [1]. We use two groups of plasma samples: (i) 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 (ii) 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 fraction 1, 3, 4, 5, and 6 were acquired with two SELDI Chips: IMAC and CM10. Total 120 plasma samples, 24 reference samples, and 6 blanks were randomly divided into two groups, Group A and 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 on 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 (Koopmann 2004). Fractionation has been shown to greatly increase the number of proteins that can be resolved.

Protein spectra were obtained on IMAC ProteinChip arrays (PCA) coupled to copper (IMAC30-Cu^{2+}, Ciphergen Biosystems, Inc., Fremont, CA) and weak cation exchange (CM10, Ciphergen Biosystems, Inc.) PCAs. Fractions were subsequently profiled on both IMAC30-Cu^{2+} 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 pooled samples from both groups and blank cases were randomly distributed to the spots of PCA in Group A or B. All spectra were acquired in duplicate using two Bioprocessors, Bioprocessor 1 and 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 re-used for other PCAs. PCAs were analyzed using 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(tm)), provided by Cleveland Clinic Foundation.

All SELDI MS data were processed with CiphergenExpress 3.0 to generate peak maps. All spectra were preprocessed with the baseline subtraction and followed by normalization based on TIC with CiphergenExpress 3.0. For MACE/control sub dataset, all spectra were kept in the dataset. To generate peak maps, all the peaks, except a few manually picked peaks were deleted before the clustering. The peaks from 2000–200 000 Da were auto-detected using the algorithms during the clustering based on the signal/noise ratios specified in Table 1. The number of peaks in each peak definition and the range of normalization factor are also listed in Table 1. This study aims to identify the MACE and control patients using the peaks that were defined as S/N = 3, Valley = 2 from fractions 1, 3, 4, 5, and 6, and MPO value. To ensure that all the peaks used in this study are well defined, 70 out of 444 peaks were manually removed from the peak list. Finally, a total of 377 peaks or biomarkers are used for following classification.

The reproducibility of the mass spectra was monitored with a pooled sample (12 samples were combined together to form a pooled sample) and total 24 spectra with the pooled sample were acquired at the same time from all samples. The intensities of the top 20 to 30 peaks in MS data were compared and statistically analyzed. The estimated measurement error on the peak intensity is about 20–30%. The peak intensity is in the relative scale with the highest value of 100%. Hence it is necessary to perform normalization

Table 1. Peak definition, peak number, and normalization factors in each fraction

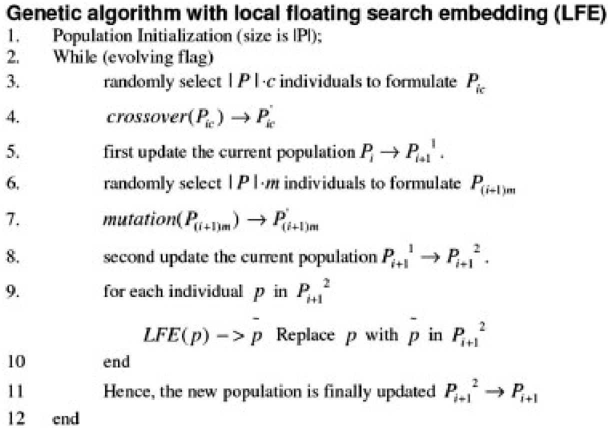
	Normalization factor	Peak no 1^{st} pass: sn3, v2, 10% 2^{nd} pass: 0.3% mass, sn2, v2	Peak no 1^{st} pass: sn3, v3, 10% 2^{nd} pass: 0.3% mass, sn1.5, v1.5	Peak no 1^{st} pass: sn2.5, v2.5, 10% 2^{nd} pass: 0.3% mass, sn1.5, v1.5	Peak no 1^{st} pass: sn2, v2, 10% 2^{nd} pass: 0.3% mass, sn1.0, v1.0	Peak no 1^{st} pass: sn2, v1.5, 10% 2^{nd} pass: 0.3% mass, sn1.5, v1.5	Peak no 1^{s} pass: sn1.5, v1.5, 10% 2^{nd} pass: 0.3% mass, sn1.0, v1.0
CM10 F1	0.21–4.96	57	65	81	132	210	265
CM10 F3 Group B recalibrated to Group A	0.5–1.6	31	40	48	68	97	151
CM10 F4	0.5–1.8	38	47	55	72	108	161
CM10 F5	0.4–2.4	34	37	50	105	156	232
CM10 F6	0.48–4.8	52	63	76	108	149	210
IMAC F1	0.23–13.0	54	66	92	151	195	253
IMAC F3	0.4–4.1	47	61	84	117	151	209
IMAC F4	0.36–2.7	46	62	72	98	145	200
IMAC F5	0.37–10.6	34	44	61	87	116	168
IMAC F6	0.43–4.73	51	52	70	118	160	217
	Total Peak no	444	537	689	1056	1487	2066

between the replicates. To normalize the mass spectra of the same sample in the two bio-processors, we first employ z-transform to each bio-processor, which makes the transformed data have zero in mean and one in variance. Then, we calculate the average of the two signals. In order to normalize the mass spectra in two chips of IMAC and CM10, we again apply z-transform to the mass spectra in each chip.

2.2 Genetic algorithm (GA) with recursively local floating searching embedding

The standard GA consists of two key operations, crossover and mutation. The crossover and mutation make GA explore a wide range of space while guaranteeing the entire population of each generation to move to an optimal stage. The main drawback is that GA cannot efficiently improve the individual to its local optimal. Hence, a hybrid GA embedded with local optimization methods is proposed by Oh *et al.* [12]. The hybrid GA is empirically proved to have better convergence than the standard GA and enjoys a slightly higher performance than standard GA.

In our pursuit of predicting MACE with just a few biomarkers, the searching burden is dramatically alleviated. Furthermore, the limited number valid bits on the chromosome encoding make the crossover procedure generate little diversity in offspring. Therefore we propose to use an enhanced GA with recursively local floating embedding to improve the biomarker selection (Fig. 1.). Compared to the standard GA, the technical merits of the proposed methods lie in the following aspects. First, after the crossover and mutation operation in each generation, a recursive local floating searching method is applied to hunt the local optimal solution around the current individuals. Second, because the forward floating procedure would exhaust all the possible biomarkers, the computation cost would be increased significantly. We thus try to limit the candidate sets by removing those biomarkers with low-prediction-relevance or high-biomarker redundancy. Third, the updating procedure for each generation has three corresponding steps to maintain a fixed size of population, *i.e.*, updating after crossover, updating after mutation, and updating after recursively local floating searching. Compared with the hybrid GA [12], the proposed method performs better because the recursively floating searching can cover much more biomarker combinations. The reduction of the candidate biomarkers with considering the MACE-relevance and biomarker-redundancy make the local searching procedure much more efficient.



Genetic algorithm with local floating search embedding (LFE)

```
1.  Population Initialization (size is |P|);
2.  While (evolving flag)
3.          randomly select $|P| \cdot c$ individuals to formulate $P_{ic}$
4.          $crossover(P_{ic}) \rightarrow P_{ic}'$
5.          first update the current population $P_i \rightarrow P_{i+1}^{1}$.
6.          randomly select $|P| \cdot m$ individuals to formulate $P_{(i+1)m}$
7.          $mutation(P_{(i+1)m}) \rightarrow P_{(i+1)m}'$
8.          second update the current population $P_{i+1}^{1} \rightarrow P_{i+1}^{2}$.
9.          for each individual $p$ in $P_{i+1}^{2}$
                $LFE(p) -> \bar{p}$  Replace $p$ with $\bar{p}$ in $P_{i+1}^{2}$
10          end
11          Hence, the new population is finally updated $P_{i+1}^{2} \rightarrow P_{i+1}$
12  end
```

Figure 1. The algorithmic steps of GA with local floating search embedding (GA-LFE).

The algorithm diagram is shown in Fig. 1. The parameters c and m are the crossover and mutation rates. The update rule in step (5) and (8) is based on the predictor accuracy of each individual. Besides the conventional crossover operator *crossover*(P_{ic}) and mutation operator *mutation*($P_{(i+1)m}$), the local floating searching embedding operator *LFE*(p) tries to find the local optimal solution around the current individual p. The evolving flag is initialized to be one and can be changed to zero under any the following conditions: the maximum generation is achieved, the evolving process achieves convergence or the maximum running time is used out. During the process of crossover and mutation, we add the strategy to make each individual has d selected biomarkers. The LFE operation, as described in Fig. 2, just tries to find an optimal solution around the current individual by two-direction recursively searching strategies, $add - r - remove - r$ and $remove - r - add - r$, where r is the step length of the bidirectional floating search. These two strategies are sequentially adding one feature and then removing one feature, or *vice versa*. For each step, the added or removed feature makes the fitness function maximum. Note that the $add - r$ procedure is an expensive step because all the possible features have to be exhaustedly evaluated. In order to reduce the computation time, the MACE-relevance and biomarker-redundancy are evaluated using information gain [13] as following:

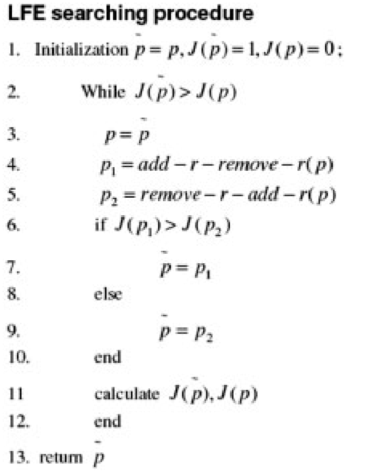
$$IG(X,\ Y) = H(X) - H(X|Y) =$$

$$-\sum_{\mathrm{i}} p(x_{\mathrm{i}}) \log_2 p(x_{\mathrm{i}}) + \sum_{\mathrm{j}} p\left(y_{\mathrm{j}}\right) \sum_{\mathrm{i}} p\left(x_{\mathrm{i}}|y_{\mathrm{j}}\right) \log_2 p\left(x_{\mathrm{i}}|y_{\mathrm{j}}\right)$$

where $H(X)$ is the entropy of a variable X and $H(X|Y)$ denotes the entropy of X after observing the values of another variable Y. The probability density function (pdf) of variable X is $p(x_{\mathrm{i}})$ and $p(x_{\mathrm{i}}|y_{\mathrm{j}})$ represents the conditional probability density of variable X, given Y. Hence the information gain reflects the additional information about X provided by the observations of Y. In order to compute the above measurement, Parzen window method is applied to estimate the pdf $p(x_{\mathrm{i}})$, as described in ref. [14].

$p(x) = \dfrac{1}{n} \sum_{\mathrm{i}=1}^{\mathrm{n}} \phi(x - x_{\mathrm{i}}, h)$ where n is the number of samples located in the window region and the Gaussian window function is commonly used with the kernel as:

$$\phi(x, h) = \frac{1}{(2\pi)^{\mathrm{d}/2} h^{\mathrm{d}} |\sum|^{1/2}} \exp\left(-\frac{x^{\mathrm{T}} \sum^{-1} x}{2h^2} \right);$$

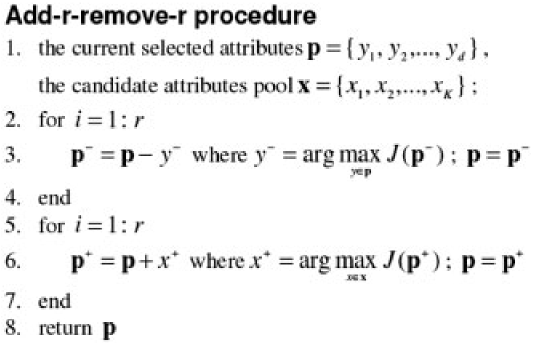


```
LFE searching procedure
Initialization $\tilde{p} = p, J(\tilde{p}) = 1, J(p) = 0$;
    While $J(\tilde{p}) > J(p)$
        $p = \tilde{p}$
        $p_1 = add - r - remove - r(p)$
        $p_2 = remove - r - add - r(p)$
        if $J(p_1) > J(p_2)$
            $\tilde{p} = p_1$
        else
            $\tilde{p} = p_2$
        end
        calculate $J(\tilde{p}), J(p)$
    end
return $\tilde{p}$
```

Figure 2. The algorithmic steps of the recursive local floating search embedding procedure.

Σ is the covariance matrix of the variable x and d is the dimensionality of variable x, h is the kernel size of Gaussian window.

Considering the reducing procedure of the biomarker candidates, the bi-directional floating search approach with step length r can be described as the algorithm chart shown in Fig. 3. In our experiments, in order to reduce the computation cost, we use the simplest version of bi-directional floating search with fixed step length 1. Assume the selected biomarkers in the chromosome p are $\{y_1, y_2, \ldots y_M\}$ and the remainder biomarker set is $\{x_1, x_2, \ldots, x_{N-M}\}$. The procedure of local optimization of the add-one-remove-one strategy consists of three basic steps. First, considering the current selected biomarkers and the output class c, the reduced candidates biomarker set $\{\tilde{x}_1, \tilde{x}_2, \ldots \tilde{x}_k\}$ are constructed by selecting the biomarkers from $\{x_1, x_2, \ldots, x_{N-M}\}$ with high MACE relevance IG $(x_i, c) > \zeta$, $i = 1, 2, \ldots, N - M$ and low biomarker redundancy IG $(x_i, y_j) < \xi$, $i = 1, 2, \ldots, N - M$, $j = 1, 2, \ldots, M$. The thresholds ζ, ξ of MACE relevance and biomarker redundancy are used to filter the original candidate. In practical, we set dynamic values of ζ and ξ to guarantee around one-tenth of the biomarkers are remained for generating the new generation of subsets, *i.e.*, $k \approx (N - M)/10$. Second, one biomarker is selected to generate a chromosome $p^+ = \{y_1, y_2, \ldots, y_M, x^+\}$, which can maximize the fitness function $\mathrm{x}^+ = \arg\max_{\mathrm{x}} J(y_1, y_2, \ldots, y_\mathrm{M}, x)$. Third, remove one biomarker y^- from p to obtain the new chromosome, which has the exact number of biomarkers. The removed y^- also maximizes the fitness function as $y^- = \arg\max_{y} J(p^-)$. Note here we use the signs of addition or subtraction to represent set operations. For example, $p + x^+$ means add one biomarker x^+ to the current biomarker set p. The remove-one-add-one strategy has the similar flow chart with the only difference in the order of operations of removing and adding biomarkers. Because the floating search is recursively executed, the local optimal set of biomarkers can be quickly acquired in several iterations.



```
Add-r-remove-r procedure
the current selected attributes $\mathbf{p} = \{y_1, y_2, \ldots, y_d\}$,
the candidate attributes pool $\mathbf{x} = \{x_1, x_2, \ldots, x_K\}$;
for $i = 1 : r$
    $\mathbf{p}^- = \mathbf{p} - y^-$ where $y^- = \arg\max_{y \in \mathbf{p}} J(\mathbf{p}^-)$; $\mathbf{p} = \mathbf{p}^-$
end
for $i = 1 : r$
    $\mathbf{p}^+ = \mathbf{p} + x^+$ where $x^+ = \arg\max_{x \in \mathbf{x}} J(\mathbf{p}^+)$; $\mathbf{p} = \mathbf{p}^+$
end
return $\mathbf{p}$
```

Figure 3. The algorithmic steps of the bi-directional floating search procedure of *add* − *r* − *remove* − *r* operation.

2.3 Validation by MACE prediction using PLS-LR and SVC

Notice that in our study, the available patient set is very small (120 patients/samples) while the original biomarker set is relatively big for all the fractions and ProteinChips. Most traditional pattern classification methods are well established for the large dataset learning. Here we use two famous small-set and binary response based classification techniques, PLS-LR and SVMC, as the classifier for our MACE prediction study. PLS-LR is developed based on the partial least squares and ridge penalized logistic regression techniques [6, 15, 16]. The ridge penalty is integrated in the partial least squares step and the dimensionality reduction is incorporated in the classification procedure. Although the approach of PLS-LR remains valid for multi-categorical classification problem, the experimental results from [6] demonstrates that this extended PLS technique have better performance for the binary response data. SVC executes the classification task by finding a hypersurface in the space of possible inputs to split the positive examples from the negative examples. The split hyperplane will be chosen to have the largest distance from the plane to the nearest of the positive and negative examples, which is named as maximum-margin hyperplanes [7, 8]. Because the construction of the hyperplane only depends on the support vectors instead of the entire input samples, it fits well to the small set based binary response learning problem. The characteristics of PLS-LR and SVC match also suit our MACE prediction problem well because SELDI-MS data obtained are binary response biomarkers indicating the MACE group and control group as well as the small size of the sample set.

3 Results

In our experiments, we use the biomarkers generated from the fraction 1, 3, 4, 5, and 6 with two SELDI chips, IMAC and CM10. There are totally 377 biomarkers, including the MPO value. The objective is to select several biomarkers for MACE prediction. The classifier used for predicting the patient group is based on the PLS-LR and SVC approaches. In order to derive a stable and robust statistical estimation of the prediction error, the 10-folder cross validation procedure is applied. The accuracy is simply the mean of tests, which generally provides a good estimation because the trained classifiers are similar in all the tests. With different partition of the data, the cross validation was repeated 10 times. The mean and SD of the prediction accuracy are recorded in Table 2. The proposed biomarker selection method is compared with the single MPO value, *t*-test ranked biomarkers, standard GA selected biomarkers, and the SFFS technique, which is claimed as the best sequential search method [6]. All these compared approaches are executed using all the candidate biomarkers (a total of 377 biomarkers) for fair compassion.

For our comparison study, both the standard GA and GA-LFE have the same setting: the population size is 120, the maximum generation size is 20, the crossover rate is 0.5, and the mutation rate is 0.3. The experimental results show that the GA-LFE significantly improved the prediction accuracy comparing with the single MPO value. For example, with seven selected biomarkers, GA-LFE relatively improved 30.91% with PLS-LR classifier (from 57.17 to 74.83%) and 39.51% (from 55.25 to 77.08%), respectively. Additionally, the best subset of seven biomarkers selected by GA-LFE achieved the highest performance of 77.08% using SVC. The extensive comparison results using three, five, and seven selected biomarkers are listed in Table 2 (mean and SD accuracy), and Table 3 (sensitivity and specificity). In Fig. 4, we compared four biomarker selection approaches, GA-LFE, SFFS, standard GA, and *t*-test by calculating the receiver operating characteristic (ROC) curve and the corresponding area under ROC curve (AUC). The ROC and AUC comparison is based on SVC classifier with seven selected biomarkers, where GA-LFE achieved the best performance. Particularly, when the false positive rate ranges from around 0.15 to 0.3, GA-LFE obtains a significant performance gain of the true positive rate. In order to obtain a visualized validation, the optimal biomarkers are projected to 2-D space using orthogonal locality preserving projections (OLPP) [17]. Figure 5 shows that the two samples, MACE and Control, are roughly clustered into two groups.

Table 4 lists the corresponding selected biomarkers. We can see that the proteins CM10 F3 107433, IMAC C4212.3, and MPO always remain in the selected biomarkers. In Fig. 6, we show the mass spectra map of the selected biomarkers: they are CM10 F6 37089, IMAC F4 2758.6, IMAC F6 4212.3, CM10 F3 107433, CM10 F6 51404, and IMAC F6 56652. Clearly all of these the selected peaks are true peaks. The protein separation and identification for these six peaks are underway and will be reported separately.

4 Discussion and conclusion

There exist different factors, such as the hardware status and environmental setting, could affect the reliability of the

Table 2. The performance comparison of accuracy and standard derivation (%) of the MACE prediction using three, five, and seven selected biomarkers by different approaches, *t*-test, standard GA, SFFS, and GA-LFE. The performance of single MPO value is also evaluated for comparison

Classifier	MPO value	377 Biomarkers	d	*t*-test	Standard GA	SFFS	GA-LFE
			3	58.83 (±1.58)	64.83 (±1.35)	66.91 (±1.42)	**68.00**
PLS-LR	57.17(±1.68)	51.33 (±1.43)	5	61.42 (±1.18)	71.42 (±1.36)	64.25 (±2.71)	**72.83 (±1.12)**
			7	62.83 (±1.37)	72.92 (±1.48)	68.58 (±2.36)	**74.83 (±1.46)**
			3	56.00 (±2.6)	63.50 (±3.47)	65.5 (±2.43)	**67.50 (±0.79)**
SVC	55.25 (±4.73)	56.83 (±4.39)	5	59.25 (±2.89)	70.33 (±1.68)	71.33 (±1.05)	**72.92 (±1.81)**
			7	61.42 (±2.08)	74.92 (±1.54)	73.5 (±1.10)	**77.08 (±1.43)**

Table 3. The performance comparison of sensitivity and specificity (%) of the MACE prediction using three, five, and seven selected biomarkers by different approaches, *t*-test, standard GA, SFFS, and GA-LFE; The performance of single MPO value is also evaluated for comparison

Classifier	MPO value	377 Biomarkers	d	*t*-test	Standard GA	SFFS	GA-LFE
			3	56.17/61.50	66.33/63.33	64.83/69.00	67.83/68.17
PLS-LR	58.33/55.50	46.50/56.17	5	54.50/68.33	72.17/70.67	61.50/67.00	73.33/72.33
			7	61.67/64.00	74.17/71.67	67.50/69.67	75.33/74.33
			3	54.50/57.50	69.33/57.67	65.50/65.50	65.17/69.83
SVC	44.50/66.00	52.00/61.67	5	60.00/58.50	70.33/70.33	70.17/72.50	73.33/72.50
			7	63.17/59.67	76.33/73.50	74.00/73.00	76.33/77.83

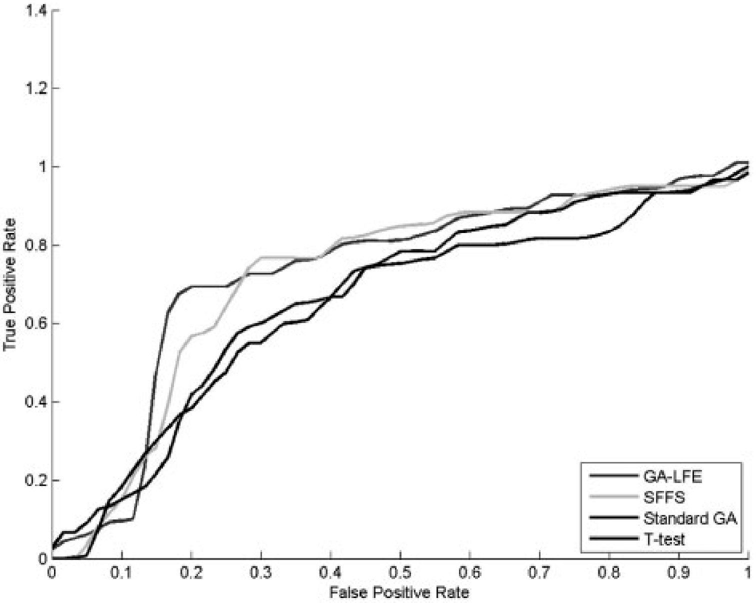


Figure 4. Comparing GA-LFE, SFFS, standard GA and *t*-test by calculating ROC curve of SVC classifier using seven selected biomarkers. The value of AUC of GA-LFE, SFFS, standard GA, and *t*-test are 0.7236, 0.7201, 0.6494, and 0.6614, respectively.

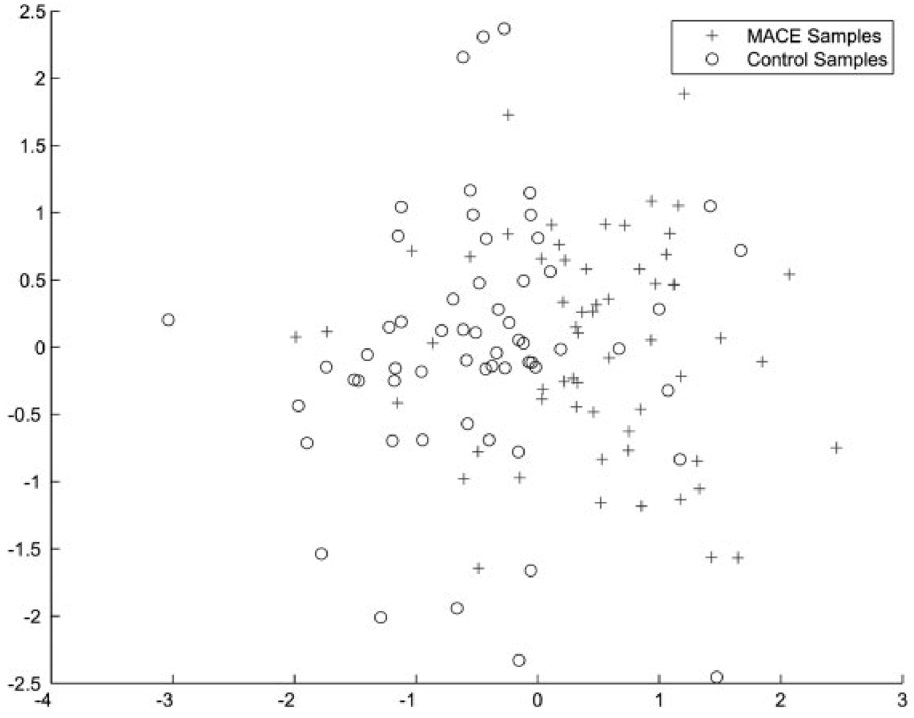


Figure 5. The 2-D projections of the samples with the five selected biomarkers using OLPP technique.

obtained SELDI-MS data. Therefore, it is necessary to remove the outlier or noisy data using the unsupervised information. High normalization factor means high variation in the sample sets. Hence the evaluation is tested by using the normalization factors. By setting a cutoff at three, we remove the samples whose normalization factors are higher than the cutoff. In our experiments, however, we found that the classification accuracy did not change significantly. The reason may be that we normalized the data between the replicates and the two different chips. Notice that each patient has 20 spectra (totally two chips, two processors *per* chip, and five fractions *per* processor). Therefore it makes sense that a couple of spectra with high normalization factors in the 20 spectra of some patients could not influence the classification accuracy of 120 patients in our experiments.

Table 4. The three, five, and seven selected biomarkers using GA-LFE

No of selected biomarkers	Position of biomarkers							
3	Fraction no.	3	6					
	SELDI Chips	CM10	IMAC					
	Name	C107433	C4212.3	MPO				
5	Fraction no.	3	6	6	6			
	SELDI Chips	CM10	CM10	CM10	IMAC			
	Name	C107433	C15193	C51404	C4212.3	MPO		
7	Fraction no.	3	4	6	6	6	6	
	SELDI Chips	CM10	IMAC	CM10	CM10	IMAC	IMAC	
	Name	C107433	C2758.6	C37089	C51404	C4212.3	C56652	MPO

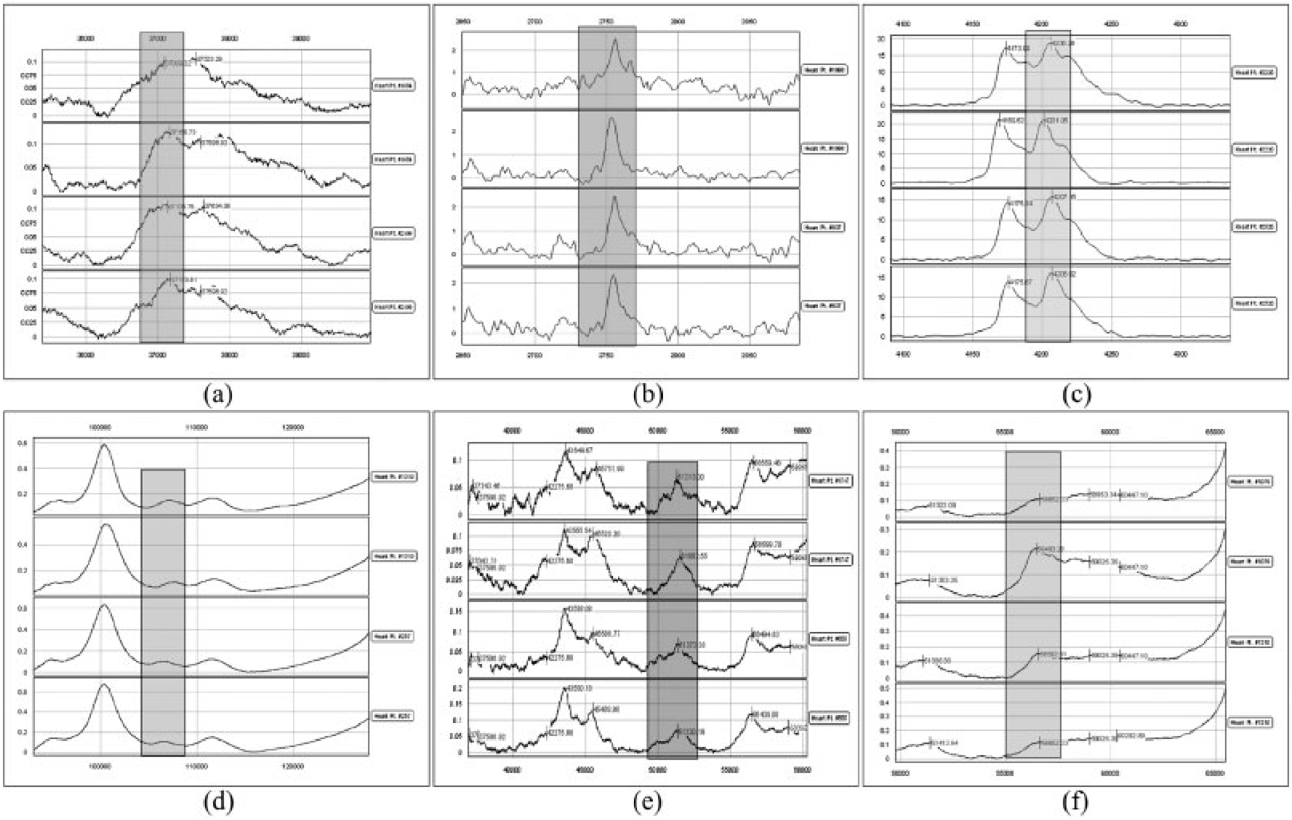


Figure 6. Mass spectra map of the selected biomarkers: (a) CM10 F6 37089; (b) IMAC F4 2758.6; (c) IMAC F6 4212.3; (d) CM10 F3 107433; (e) CM10 F6 51404; and (f) IMAC F6 56652.

The objective of this work is to develop a molecular diagnostic tool based on the biomarkers or models discovered by clinical proteomics to predict the risk of MACE in 6 months with improved accuracy compared with that of MPO alone, the latter has recently been approved by FDA to evaluate patients presenting with chest pain that are at risk for MACE, including myocardial infarction, need for revascularization, or death (PrognostiX CardioMPO Test, 510(k) summary, May 10, 2005).

The number of proteins in human blood is estimated to be in the order of ten thousands, and the application of proteomics approaches for protein profiling can generate large arrays of data for the development of optimized biomarker protein panels. The simplicity and concision requirement for the development of immunoassay can only tolerate the complexity of the prediction model with a small number of selected biomarkers. As described early, since we reduce the local search space to 1/10 of the original pool size, the computation cost is around 5–15 times of the standard GA approach, which is acceptable for this specific application.

The conventional optimization techniques for biomarker discovery, such as GA, are alleviated on the search ability because of the high-dimensional input biomarkers, small sample set, and the dramatically decreased of the number of target biomarkers. In this paper, we proposed an improved GA with local floating search embedding to identify an optimal panel of biomarkers. The standard GA has the extended search diversities through the crossover and mutation operation. It, however, does not have the local optimization strategy. Here we use the new bi-directional floating search approach, *i.e.*, $add - r - remove - r$ or $remove - r - add - r$ biomarkers, to find the local optimal solution around individuals of each generation during the evolving. In order to reduce the high computation cost, the MACE-relevance and biomarker redundancy are estimated using information gain. Therefore, the local optimization is only conducted on the biomarkers with high MACE-relevance and small biomarker redundancy.

Finally, we demonstrated the selected optimal panel biomarkers using the MACE prediction experiments. The well developed classifiers, PLS-LS and SVC, specially designed for binary variable response and small set learning, are constructed to predict the MACE. The comparison study with several classical variable selection techniques demonstrates that the proposed method has the technique merits on our biomarker discovery task for risk stratification of cardiovascular events. The average prediction accuracies with 3, 5, and 7 selected biomarkers are 67.50, 72.92, and 77.08% using SVC, respectively. Future work would include improvement of low level processing of the MS data and the design of more efficient and problem specified MACE predictor.

This research is partially funded by the HCNR Center for Bioinformatics Research Grant, Harvard Medical School (STCW), NIH Clinical Center (KL), and The Methodist Hospital Research Institute Scholarship Award (XZ).

5 References

[1] Brennan, M. L., Penn, M. S., Lente, F. V., Nambi, V. *et al.*, Prognostic value of myeloperoxidase in patients with chest pain. *N. Engl. J. Med.* 2003, *349*, 1595–1604.

[2] Kozak, K. R., Amneus, M. W., Pusey, S. M., Su, F. *et al.*, Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: Potential use in diagnosis and prognosis. *Proc. Natl. Acad. Sci.* 2003, *100*, 12343–12348.

[3] Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J., and Kovach, J., Detection of cancer-specific markers amid massive mass spectral data. *Proc. Natl. Acad. Sci.* 2003, *100*, 14666–14671.

[4] Peng, H., Long, F., Ding, C., Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 2005, *27*, 1226–1238.

[5] Holland, J. H.,*Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence* 2nd edn., MIT Press, Cambridge, MA 1996.

[6] Fort, G., Lambert-Lacroix, S., Classification using partial least squares with penalized logistic regression. *Bioinformatics* 2005, *21*, 1104–1111.

[7] Vapnik, V., *Statistical Learning Theory*, Wiley-Interscience, New York 1998.

[8] Cristianini, N., Shawe-Taylor, J., *An Introduction to Support Vector Machines*, Cambridge University Press, UK 2000.

[9] Duda, R. O., Hart, P. E., Stork, D. G. *Pattern Classification*, 2nd edn., Wiley-Interscience, New York 2001.

[10] Liu, J. J., Cutler, G., Li, W., Pan, Z. *et al.*, Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics* 2005, *21*, 2691–2697.

[11] Jarvis, R. M., Goodacre, R., Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. *Bioinformatics* 2005, *21*, 860–868.

[12] Oh, I., Lee, J. S., Moon, B. R., Hybrid Genetic Algorithms for Feature Selection. *IEEE Trans. Pattern Anal. Mach. Intell.* 2004, *26*, 1424–1437.

[13] Yu, L., Liu, H., Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.* 2004, *5*, 1205–1224.

[14] Kwak, N., Choi, C. H., Input feature selection by mutual information based on Parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.* 2002, *24*, 1667–1671.

[15] Le Cessie, S., Van Houwelingen, J., Ridge estimators in logistic regression. *J. R. Stat. Soc. Ser. C Appl. Stat.* 1992, *41*, 191–201.

[16] Helland, I., On the structure of partial least squares regression. *Comm. Statist. Simulation Comput.* 1988, *17*, 581–607.

[17] Cai, D., He, X., Han, J., Zhang, H., Orthogonal laplacianfaces for face recognition. *IEEE Trans. Image Process.* 2006, *15*, 3608–3614.