5-2018

Wisdom of artificial crowds feature selection in untargeted metabolomics: An application to the development of a blood-based diagnostic test for thrombotic myocardial infarction

Patrick J. Trainor
University of Louisville

Roman V. Yampolskiy
University of Louisville, roman.yampolskiy@louisville.edu

Andrew P. DeFilippis
University of Louisville

Original Publication Information
Patrick J. Trainor, Roman V. Yampolskiy, Andrew P. DeFilippis, Wisdom of artificial crowds feature selection in untargeted metabolomics: An application to the development of a blood-based diagnostic test for thrombotic myocardial infarction, Journal of Biomedical Informatics, Volume 81, 2018, Pages 53-60, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2018.03.007. (https://www.sciencedirect.com/science/article/pii/S1532046418300492)

Journal of Biomedical Informatics 81 (2018) 53–60

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Wisdom of artificial crowds feature selection in untargeted metabolomics: An application to the development of a blood-based diagnostic test for thrombotic myocardial infarction

Patrick J. Trainor[a,*], Roman V. Yampolskiy[b,1], Andrew P. DeFilippis[a,1]

[a] Department of Medicine, Division of Cardiovascular Medicine, University of Louisville, United States
[b] Department of Computer Science and Engineering, University of Louisville, United States

ARTICLE INFO

Keywords:
Evolutionary computation
Wisdom of artificial crowds
Metabolomics
Feature selection
Classification
Diagnostic test
Myocardial infarction

ABSTRACT

Introduction: Heart disease remains a leading cause of global mortality. While acute myocardial infarction (colloquially: heart attack), has multiple proximate causes, proximate etiology cannot be determined by a blood-based diagnostic test. We enrolled a suitable patient cohort and conducted a non-targeted quantification of plasma metabolites by mass spectrometry for developing a test that can differentiate between thrombotic MI, non-thrombotic MI, and stable disease. A significant challenge in developing such a diagnostic test is solving the NP-hard problem of feature selection for constructing an optimal statistical classifier.

Objective: We employed a Wisdom of Artificial Crowds (WoAC) strategy for solving the feature selection problem and evaluated the accuracy and parsimony of downstream classifiers in comparison with traditional feature selection techniques including the Lasso and selection using Random Forest variable importance criteria.

Materials and methods: Artificial Crowd Wisdom was generated via aggregation of the best solutions from independent and diverse genetic algorithm populations that were initialized with bootstrapping and a random subspaces constraint.

Results/Conclusions: Strong evidence was observed that a statistical classifier utilizing WoAC feature selection can discriminate between human subjects presenting with thrombotic MI, non-thrombotic MI, and stable Coronary Artery Disease given abundances of selected plasma metabolites. Utilizing the abundances of twenty selected metabolites, a leave-one-out cross-validation estimated misclassification rate of 2.6% was observed. However, the WoAC feature selection strategy did not perform better than the Lasso over the current study.

1. Introduction

Heart disease remains the most prevalent cause of death worldwide despite dramatic reductions in the incidence of heart disease associated mortality in developed countries [1]. Acute Myocardial Infarction (AMI), an acute manifestation of heart disease, is characterized by myocardial ischemia (oxygen starvation in heart muscle) and necrosis (a form of cell death) secondary to atherosclerotic plaque disruption or other cause. While ischemia and necrosis are the common pathological characteristics of AMI there are multiple underlying causes that can lead to ischemia and necrosis [2]. An important etiological distinction can be made between thrombotic and non-thrombotic MI. Thrombotic MI results from spontaneous atherosclerotic plaque disruption (e.g. rupture or erosion) that results in the formation of an occluding coronary thrombus [2]. In contrast, MI secondary to oxygen supply/demand mismatch resulting from conditions not associated with plaque rupture such as vasospasm or stress cardiomyopathy are categorized as non-thrombotic MI. These etiological distinctions are of critical importance as the course of treatment depends on the underlying cause and diagnostic misclassification may result in negative outcomes such as iatrogenic bleeding [3].

Blood-based tests that measure the release of troponin from the myocardium can provide evidence of myocardial necrosis which may be used to substantiate a diagnosis of AMI [4]. However, a non-invasive test that enables the discrimination of thrombotic from non-thrombotic MI has yet to be developed. In response, we set out to develop a blood-based test that could differentiate thrombotic MI from non-thrombotic MI and stable coronary artery disease (CAD). In developing a blood-based diagnostic test we chose a plasma medium—plasma contains enzymes, lipoproteins, hormones, metabolic intermediates, and other

* Corresponding author.
E-mail address: patrick.trainor@louisville.edu (P.J. Trainor).
[1] Equally contributing authors.

https://doi.org/10.1016/j.jbi.2018.03.007
Received 5 August 2017; Received in revised form 15 December 2017; Accepted 12 March 2018
Available online 22 March 2018
1532-0464/

small molecules dissolved in suspension. Plasma provides a suitable medium as plasma is a repository of biochemicals that reflects the state of the entire organism at the time of sampling [5]. We focused on metabolites—or more precisely small molecules—as metabolite concentrations are dynamic and reflect the "end result" of genetic factors, environmental exposures, and gene-environment interactions at the time of sampling [6]. A patient cohort was recruited that allowed for the discrimination of thrombotic MI from multiple control populations [7–9]. This cohort consisted of three study groups: thrombotic MI, non-thrombotic MI, and stable coronary artery disease subjects. The non-thrombotic MI group controlled for metabolic changes associated with myocardial ischemia and necrosis, while the stable coronary artery disease group was used to control for the underlying disease process of atherosclerosis. Both control groups served as procedural controls as all three study groups underwent a cardiac catheterization procedure following study enrollment.

Whole blood was collected immediately prior to cardiac catheterization from subjects in each of the study groups and plasma metabolite relative abundances were determined by a non-targeted mass spectrometry approach. While non-targeted mass spectrometry is well suited for determining a metabolic signature or biomarkers of a disease state or phenotype, a diagnostic test requires a targeted strategy such as mass spectrometry based multiple reaction monitoring with stable isotopically labeled standards (MRM) [10] or an enzyme-linked immunosorbent assay (ELISA) [11]. In both cases, a small set of discriminatory biochemicals is highly desirable, as the marginal effort requirement of quantifying additional biochemicals is significant. Two challenges arise in developing a practical diagnostic classifier from non-targeted mass spectrometry data given the constraint that the final classifier should use only a small number of metabolites. The first is that the dimension of the feature space (1032 detected metabolites) may be significantly larger than the number of samples (11 thrombotic MI, 12 non-thrombotic MI, and 15 stable CAD in the case of the current cohort). To determine the best subset of metabolites to be included in a classifier is thus a search problem for which a brute-force approach is not advisable. For example, to determine the optimal classifier with five metabolites, 9.7×10^{12} combinations are possible. Second, the parameter estimates of a classification model may be highly unstable given the small sample size. Consequently, an algorithm that searches for the best possible subset of metabolites for inclusion in a classifier should converge in reasonable time and should minimize a measure of expected prediction error subject to the constraint of few metabolites included. In the current study, we evaluated the use of a Wisdom of Artificial Crowds (WoAC) [12] approach to feature selection for developing a blood-based diagnostic test for thrombotic myocardial infarction. A WoAC approach to problem solving is predicated on the Wisdom of Crowds concept that holds that under certain conditions the aggregation of independent "expert" solutions will outperform individual solutions. Improved performance using Wisdom of Crowds over individual solutions has been shown for classical search problems such as the traveling salesman problem given both human crowds [13] and artificial crowds [12] and in problems in the domain of molecular biology such as inferring gene networks by combining the solutions of multiple experiments and models [14]. Wisdom of Crowds aggregation is fundamentally related to bootstrap aggregation or "bagging" [15], however we use the term Wisdom of Crowds to emphasize that consensus wisdom is applied to feature selection as a search problem as opposed to the aggregation of predictions. We loosely follow the process utilized by Yampolskiy and Barkouky [12] in that we first generate individual solutions using a genetic algorithm and then aggregate a proportion of the solutions as "experts" to determine a consensus solution to feature selection. In the current work we evaluated the performance of this strategy for feature selection against the Least absolute shrinkage and selection operator (Lasso) [16] and selection using the measure variable importance determined by the permutation of observed values in the generation of Random Forest ensembles [17]. Following the selection of fixed numbers (3, 5, 7, 10, 15, 20 and 25) of metabolites by each method, multinomial logit generalized linear model (GLM) and Random Forest classifiers were constructed. In the case of multinomial logit GLM, Elastic Net regularization [18] was also applied to yield a separate set of classifiers for evaluation.

2. Materials and methods

2.1. Study cohort

Patients presenting to two hospitals in Louisville, Kentucky with suspected acute myocardial infarction were enrolled in the study upon providing written informed consent. Additionally, patients presenting for an elective outpatient procedure for the treatment of stable coronary artery disease were enrolled as stable disease controls upon providing written informed consent. Novel criteria was developed by our group for differentiating thrombotic MI and non-thrombotic MI and has been previously discussed [7–9] and is presented in Supplementary Table 1. Briefly, this criteria required clinical presentation consistent with the universal definition of AMI and a finding of positive Troponin for inclusion in either AMI study group. For the thrombotic MI group, recovery of a histologically confirmed coronary thrombus as well as $\geqslant$ 50% stenosis in the same vessel was an inclusion criteria. For inclusion in the non-thrombotic MI group, subjects must not have had significant stenosis or evidence of flow-limiting stenotic lesions evaluated by angiography and must not have had a thrombus recovered. The strict inclusion criteria was designed to limit misclassification of thrombotic MI and non-thrombotic MI subjects, and many borderline cases were eliminated (Supplemental Fig. 2). The cohort described in this study included 11 thrombotic MI, 12 non-thrombotic MI, and 15 stable CAD subjects.

2.2. Plasma samples and analytical measurement

Whole blood was collected from study subjects immediately prior to cardiac catheterization and plasma was extracted via centrifugation. The detection and quantification of metabolite relative abundances was conducted by Metabolon, Inc. Samples were analyzed by UPLC-MS/MS with positive ion mode electrospray ionization, negative ion mode electrospray ionization, and a negative ionization optimized for polar molecule detection and GC–MS. The identity of biochemicals detected was based on retention index matching, mass to charge ratio matching, and spectral data from libraries of known standards. All metabolites were identified at MSI level 1 unless denoted "unknown". After identification, relative abundances were quantified by determining the area-under-the curve. After quantification, minimum values were imputed for missing abundances predicated on the assumption that these compounds were either not present or below the limit of detection. The resulting data was then median scaled and log-transformed (base 2).

2.3. Genetic algorithm

We begin our description of the approach used to develop a classifier by introducing our notation. The phenotype (thrombotic MI, non-thrombotic MI, or stable CAD) of the ith plasma sample for $i = 1,2,...,N$, is represented as:

$$Y_{i,j} = \begin{cases} 1 & \text{if } c(Y_i) = j \\ 0 & \text{else} \end{cases}$$

where $c(Y_i)$ returns a phenotype index and $j \in \{1,2,...,J\}$ is the index of phenotypes. The vector $\mathbf{x}_i$ represents the metabolite abundances from the i th sample but the notation X_m is used to emphasize that the abundance of the m th metabolite is a random variable. The probability the ith sample has phenotype index is then:

$$\pi_{ij} = P(Y_i = j).$$

A multinomial logit model was assumed for determining the phenotype probabilities of each sample. This model is a generalized linear model with the following form [19]:

$$\eta_{ij} = \log\frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + \mathbf{x}_i^T\boldsymbol{\beta}_j.$$

The estimated phenotype probabilities are then:

$$\hat{\pi}_{ij} = \frac{\exp\hat{\eta}_{ij}}{\sum_{k=1}^{J}\exp\hat{\eta}_{ik}}.$$

The subset of metabolites included as predictors in the multinomial model was denoted $\mathscr{M}$, with the complement subset (metabolites not included) denoted $\mathscr{M}^C$.

To employ a Wisdom of Crowds approach for metabolite selection, a synthetic crowd was generated. A crowd was generated by first determining optimal metabolite subsets using a non-standard genetic algorithm. This algorithm emulated four biologically-inspired processes (birth/external immigration, recombination, mutation, and death) mimicking the evolution of chromosomes as the material of trait inheritance [20]. Each iteration of the algorithm represented one temporal generation for the population of genetic material. The algorithm was initialized by generating an initial list of Boolean vectors $\mathscr{B}^{(1)} = \{\mathbf{b}_l^{(1)} | l \in 1,2,...,L\}$ with entries $b_{ml}^{(1)}$ defined as:

$$b_{ml}^{(1)} = \begin{cases} 0 & \text{if } X_m \in \mathscr{M}^c \\ 1 & \text{if } X_m \in M \end{cases}.$$

The initial Boolean vectors were generated by simulating Bernoulli random variables to generate a population with limiting distribution $\boldsymbol{\phi}^0 = (\phi_{\mathscr{M}^c}, \phi_{\mathscr{M}})^T$.

After initialization, the cost of each vector was estimated prediction error using repeated k-fold cross-validation. Each multinomial logit model was used to generate phenotype probability estimates $\hat{\pi}_{ij}$. From these estimates the cross-entropy loss [21] was determined as:

$$L(Y_i, \hat{\boldsymbol{\pi}}_i) = \sum_{j=1}^{J} Y_i \log\hat{\pi}_{ij},$$

with corresponding cross-entropy error of prediction:

$$\epsilon = \sum_{i=1}^{N}\sum_{j=1}^{J} Y_i \log\hat{\pi}_{ij}$$

Cross-validation was used to estimate the expected error of prediction [15]. The observed data $\{(\mathbf{x}_i, Y_i)\colon\ i = 1,2,...,N\}$ was randomly partitioned into k folds (we choose $k = 10$). Representing this partition as a mapping of samples to folds $\kappa\colon\ \{1,2,...,N\} \mapsto \{1,2,...,k\}$, the cross-validation estimated error of prediction was then:

$$\hat{\epsilon} = \sum_{i=1}^{N}\sum_{j=1}^{J} Y_i \log\hat{\pi}_{ij}^{-\kappa(i)},$$

where $\hat{Y}_i^{-\kappa(i)}$ denotes the predicted phenotype of the ith sample with the $\kappa(i)$ fold removed in the estimation of the multinomial logit model. Given the potential variability of cross-validation estimates of expected prediction error over small samples [22], repeated cross-validation was employed in which the random partitioning and estimation were conducted repeatedly (10 times) and the mean of the estimates was taken as the cost of the vector $\mathbf{b}_l$. The fitness of $\mathbf{b}_l$ was inversely proportional to the cost.

After determining the fitness of each inclusion vector, a two-point crossover recombination operator was used to generate child chromosomes [23]. This operator selected two chiasmata (location of crossover) to mimic a double crossover of homologous chromosomes [20]. As a concrete example consider $\mathbf{b}_1 = (0,1,0,1,0,1,0)$ and $\mathbf{b}_2 = (1,0,1,0,1,0,1)$ for crossover with chiasmata at positions 2 and 5. Possible progeny vectors are then: $\mathbf{b}_3 = (0,1|1,0,1|1,0)$ and $\mathbf{b}_4 = (1,0|0,1,0|0,1)$. If a single endpoint was selected as chiasma, then a single-point crossover could be produced. The fitness of each $\mathbf{b}_l$ determined the probability that that vector would be a parent in recombination. The probability of recombination participation was determined by mapping the empirical cumulative distribution function (ECDF) of the fitness of each $\mathbf{b}_l^{(1)} \in \mathscr{B}^{(1)}$ to the cumulative distribution function of a Beta distribution, $Beta(1,3)$, to generate Bernoulli priors, p_l, where p_l was the probability of recombination participation. In addition to generating new genetic material via a recombination operator, new genetic material was generated by an "immigration" which created a small proportion of new inclusion vectors using the same generation function that generated the initial population. A death operator was defined for maintaining a stable population size and to retire inclusion vectors of low fitness. The death operator proceeded as follows: a cost function was defined as a weighted average of inclusion vector fitness (the inverse of repeated cross-validation estimated misclassification error) and age (the number of generations over which the vector had existed). As the recombination operator generated new genetic material, low fitness inclusion vectors were removed from the population by the death operator. A mutation operator was defined by generating a transition matrix $\mathbf{A}$ for each vector $\mathbf{b}_l^{(1)} \in \mathscr{B}^{(1)}$ to maintain the steady state distribution $\boldsymbol{\pi}^0 = (\pi_{\mathscr{M}^c}, \pi_{\mathscr{M}})^T$, that is $\mathbf{A}\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^0$ (although the mutation operator was applied only once per generation). The entries of the transition matrix were determined by varying the proportion of metabolites switching state inversely with fitness, that is the mutation rate $\varphi_l = a_{\mathscr{M},\mathscr{M}^c} + a_{\mathscr{M}^c,\mathscr{M}}$ varied linearly with inverse fitness. Each binary metabolite indicator random variable $b_{ml}^{(t)}$ thus was a discrete nonhomogeneous Markov chain [24].

2.4. Artificial crowds aggregation

Following the method introduced in Yampolskiy and Barkouky [12], we generate artificial crowd wisdom by aggregating intermediate genetic algorithm solutions. For crowd wisdom to generate "wise" solutions a crowd must be diverse (solutions determined from private information) and independent [25]. Similarly the aggregation of classifiers benefits from the introduction of diversity, such as with Random Forest ensembles [26]. As with Random Forest ensembles, we introduce diversity via bootstrapping and a random subspaces constraint. Specifically, 1000 bootstrapped samples, $(\mathbf{X}^*, \mathbf{Y}^*)_k$, were drawn with replacement from the original data. The columns of $\mathbf{X}_k^*$ were then sampled without replacement to generate a new matrix $\mathbf{X}_k^{*'}$ such that $\dim(\mathbf{X}_k^{*'}) = n \times \text{floor}(p/3)$. Each sample $(\mathbf{X}^{*'}, \mathbf{Y}^*)_k$ was then used to initialize a genetic algorithm population. A single iteration of the algorithm consisted of the application of the following operators to the population: recombination, immigration, death, and mutation. Each initial population of inclusion vectors $\mathscr{B}^{(1)}$ had a population size of 250 and underwent 200 iterations, representing 200 generations of evolution. From each population of 250 inclusion vectors, the best solution of the inclusion vectors was retained as an "expert". The proportion of times a metabolite was included in a multinomial model was then computed over the independent populations as: $p_m = \frac{1}{L}\sum_{l=1}^{L} b_{ml}^{(T)}$ where $l = 1,2,...,L$ indexed the independent GA populations.

2.5. Feature selection

Over the expert solutions from the genetic algorithm populations, the proportion p_m that each metabolite was included was then used as a measure of relative variable importance. As references for performance evaluation, both the Lasso [16] and variable importance determined from a Random Forest ensemble [17,26] were used.

In this context, the Lasso corresponded to maximizing the multinomial logit likelihood with added L_1 norm penalization, that is optimization of the following log-likelihood [27]:

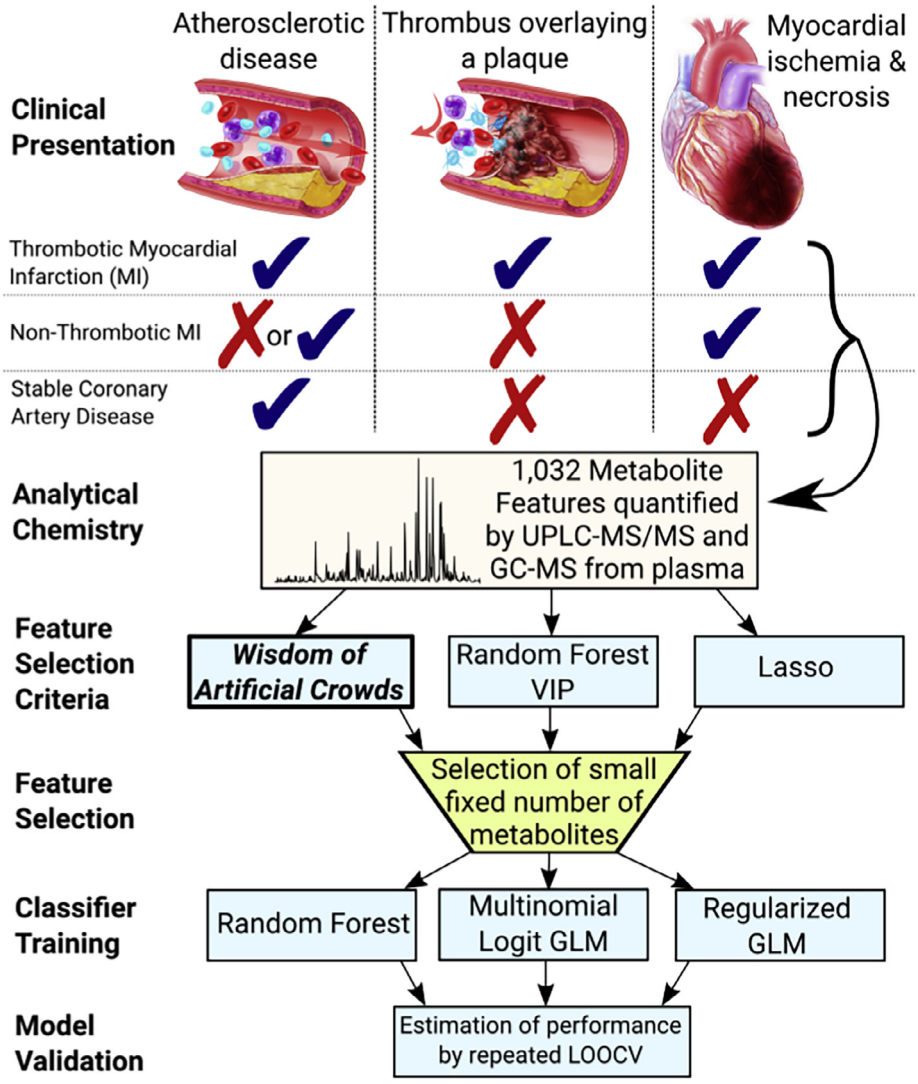

Fig. 1. Flow chart diagram of the process used to determine a diagnostic classifier for Acute Myocardial Infarction (AMI) from the abundance of circulating metabolites in plasma. Blood samples were drawn from human subjects presenting with Thrombotic MI, Non-thrombotic MI, and Stable Coronary Artery Disease (CAD). Abundances of plasma metabolites were quantified via a non-targeted approach using ultra performance liquid chromatography-tandem mass spectrometry (UPLC-MS/MS) and gas chromatography mass spectrometry (GC–MS). Feature selection was conducted by Wisdom of Artificial Crowds with other approaches (the Lasso and Random Forest Variable Importance) employed for comparison. A small fixed number of metabolites was selected, and classifiers were trained and evaluated.

$$\ell(\{\beta_{0k},\beta_k\}_1^K) = -\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k=1}^{K} y_{il}(\beta_{0k} + x_i^T\beta_k) - \log\left(\sum_{k=1}^{K}\exp(\beta_{0k} + x_i^T\beta_k)\right)\right] + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$\ell(\{\beta_{0k},\beta_k\}_1^K) = -\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k=1}^{K} y_{il}(\beta_{0k} + x_i^T\beta_k) - \log\left(\sum_{k=1}^{K}\exp(\beta_{0k} + x_i^T\beta_k)\right)\right] + \lambda\sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right].$$

In this notation, K represents the number of phenotypes and p represents the number of metabolites in the model. The value of λ was selected that minimized cross-entropy error estimated by 10-fold cross-validation. A measure of feature importance was determined using a Random Forest ensemble [17,26] of size 10,000 trees by determining the increase in misclassification error over the ensemble as the observed abundance values of each metabolite are permuted within "out-of-bag" samples.

2.6. Classifier construction

Multinomial logit GLMs and Random Forest ensembles were constructed following feature selection using the best 3, 5, 7, 10, 15, 20, and 25 metabolites. Given the number of parameters to be estimated in the multinomial logit GLMs relative to sample size, a separate set of multinomial logit classifiers was fit using a regularized likelihood. The elastic-net penalty was used for regularization of the likelihood which includes L_1 and L_2 norm penalties [18]. The regularized log- likelihood then has the following form [27]:

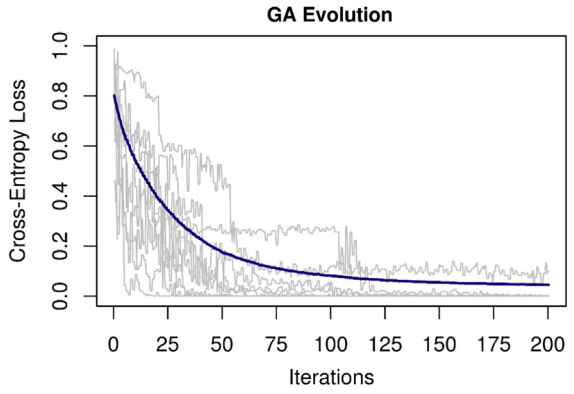

Fig. 2. Cost-paths of the GA solution with the minimum final cost (cross-entropy loss) from 10 randomly selected populations. The average cost over all populations is also shown (blue line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

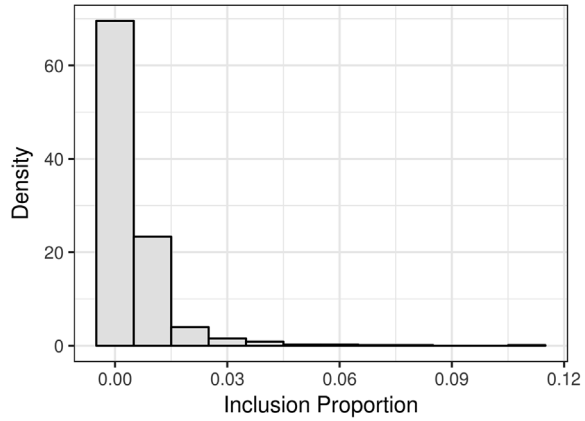

Fig. 3. Empirical distribution of inclusion proportion for metabolites with inclusion proportion greater than zero.

The parameter α controls the tradeoff between the Lasso penalty and the ridge regression penalty (see Fig. 1).

2.7. Performance evaluation

The performance of the WoAC approach to feature selection towards the development of a classifier for the discrimination of acute myocardial infarction in comparison to selection via the Lasso and Random Forest variable importance was estimated via leave-one-out cross-validation.

The WoAC feature selection methods were developed in the R software environment [28] and are available at the repository https://github.com/trainorp/WoAC. Functions were utilized from the following R packages: *randomForest* [29], *nnet* [30], *cvTools* [31], *doParallel* [32], and *glmnet* [27].

3. Results

The cost paths of the genetic algorithm (GA) solution exhibiting minimal final cost (cross-entropy error) for 10 randomly selected populations are shown in Fig. 2. These solutions are representative members of the Artificial Crowds. Diminishing returns in cross-entropy loss reduction are observed with increasing evolutionary time. The empirical distribution of p_m for the metabolites with a non-zero p_m value is shown in Fig. 3. Similarly, the course of the Lasso over varying values of λ is shown in Fig. 4. In this figure, the estimated cross-entropy error is shown as a function of λ [sub-figure (A)], the number of metabolites selected (non-zero coefficients) as a function of error is shown in sub-figure (B), and the L_1 norm of the coefficients and the number of metabolites selected is shown as a function of λ [sub-figure (C)]. The distribution of p_m was extremely skewed. Of the 1032 metabolites, only 29 had $p_m > 0.025$. These metabolites included: 5 lysophospholipids (LysoPE and LysoPC species), 5 steroid metabolites (pregnenolone and corticosteroid metabolites), 4 monoacylglycerols, 3 amino acids (His, Lys, Ser), 3-aminoisobutyrate, 3-hydroxypyridine sulfate, 3-methyl catechol sulfate, 4-allylphenol sulfate, methyl-4-hydroxybenzoate, nicotinamide, phosphate, and 5 unknowns. The pairwise correlations between these metabolites are presented in Fig. 5 and the abundance distributions of the selected metabolites are shown in Supplemental Fig. 2. Significant correlations were observed in the steroid hormone abundances and in the monoacylglycerols abundances. The abundance of steroid hormones was negatively correlated with the abundance of Ser, Lys, His, nicotinamide, and phosphate.

The results of the evaluation of WoAC feature selection performance are presented in Table 1. The best misclassification rate observed was for WoAC selection with 25 metabolites and Lasso selection with 20 or 25 metabolites with a 2.6% misclassification rate estimated by LOO-CV. Each of these classifiers achieving this 2.6% misclassification rate was a regularized multinomial logit classifier. With respect to cross-entropy error, the lowest observed error was for the Lasso solution, with an

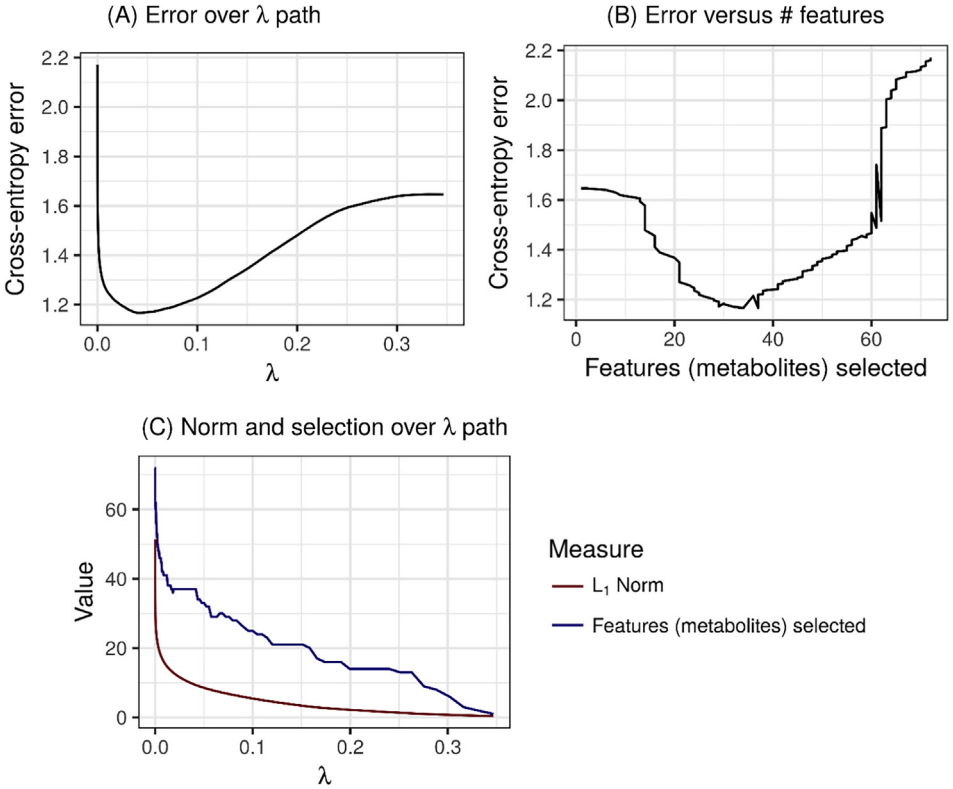

Fig. 4. Lasso path over varying values of λ. (A) Estimated cross-entropy error as a function of λ. (B) Number of metabolites selected (with non-zero coefficients) as a function of error. (C) The L_1 norm of the coefficients and the number of metabolites as a function of λ.

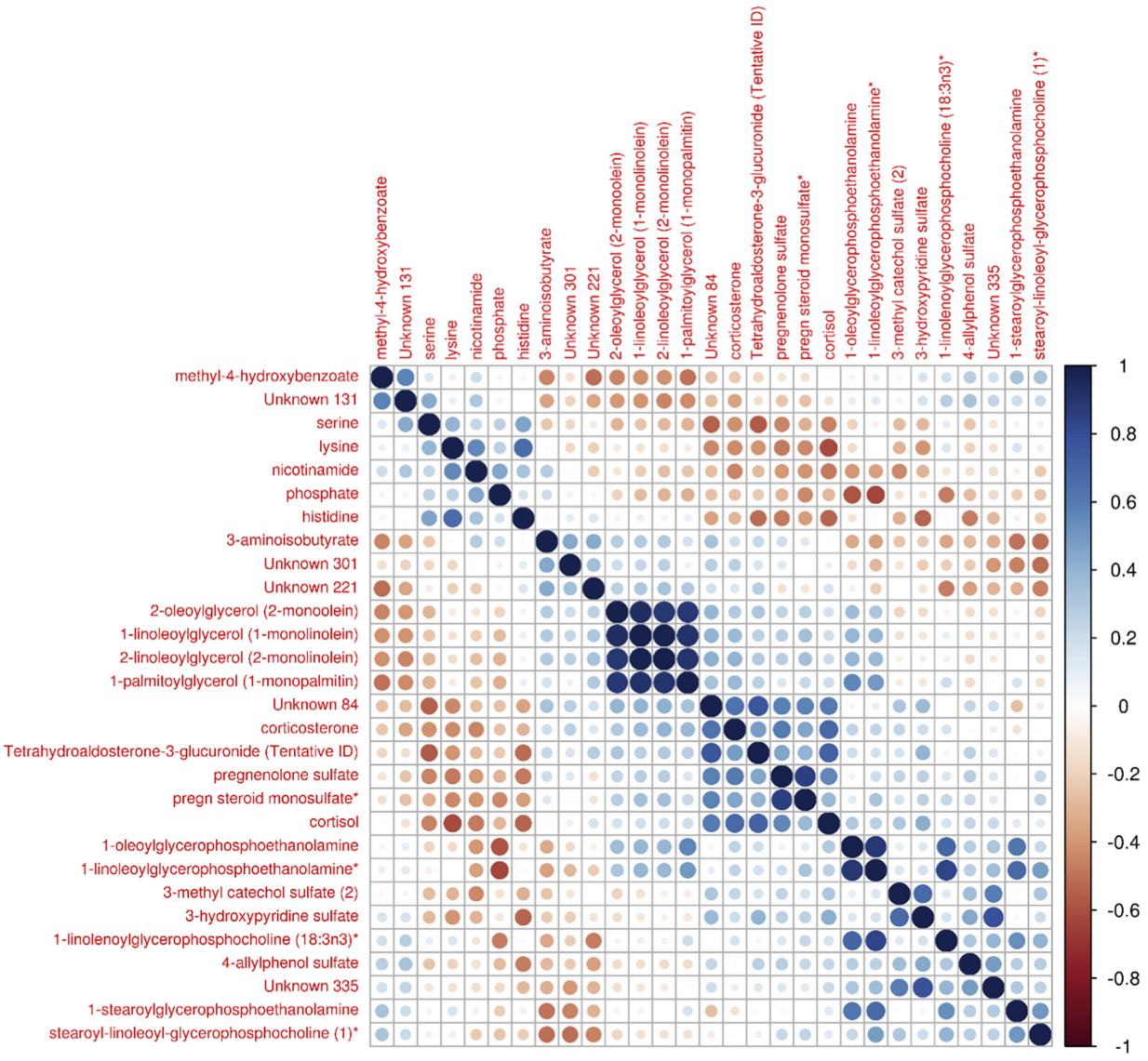

Fig. 5. Correlation plot of plasma metabolites with inclusion proportion greater than 0.025 using the WoAC approach.

error of 0.057. A trend was observed of lower cross-entropy error using the Lasso relative to WoAC selection or Random Forest variable importance selection. Over the entire evaluation, substantial performance improvements were observed by increasing the number of metabolites included in the classifier from 3 to at least 20. For example, the estimated misclassification error rate dropped from 18.4% to 2.6% using WoAC selection and a regularized multinomial logit classifier.

4. Discussion

In this paper we detailed an evaluation of Wisdom of Artificial Crowds using intermediate genetic algorithm solutions for feature selection. The successful application of genetic algorithms to feature selection in metabolomics is well established with early papers pairing GA feature selection with partial least squares regression (PLSR) for determining important spectral features from mass spectrometry [33] and nuclear magnetic resonance [34] data. However, unlike these works, we do not directly incorporate a genetic algorithm solution or set of solutions directly via embedding. Instead, our application first generates independently evolved and diverse (via bootstrapping and the random subspaces constraint) populations of genetic algorithm solutions as a crowd and extracts the consensus wisdom of that crowd. This wisdom was then applied to the feature selection problem, to enable classifier construction.

While we hypothesized that a WoAC feature selection strategy would perform better for the task of selecting sets of metabolites for classifier construction (given constraints on the number of metabolites) than traditional feature selection methods such as the Lasso, we did not observe evidence to support this hypothesis. Whether this is a general rule deserves further study. One specific aspect that may lead to a different performance is the choice of cost function utilized to determine each genetic algorithm solution. In this work we utilized the cross-entropy error from multinomial logit classifiers, however this general methodology would be amenable to the choice of other cost functions based on a different classifier model or different loss function.

A significant finding of this work is that using the abundances of specific metabolites in plasma, discrimination of thrombotic MI, non-thrombotic MI, and stable CAD subjects is possible with minimal error. Based on the current work, it does appear to be possible to accurately discriminate between these three phenotypes given the concentrations of a very small number metabolites in plasma, such as 3. To advance any of the classifiers constructed in the current work, further steps are required including measuring the selected metabolites utilizing a targeted and quantitative approach such as mass spectrometry based

Table 1
Estimates of classifier performance following feature selection utilizing WoAC, Random Forest variable importance, or the Lasso for classifiers with number of metabolites ranging from 3 to 25.

Method	# Of metabolites	Misclassification rate			Cross-entropy error			Minimum error	
		Multinomial logit GLM	Regularized GLM	Random forest	Multinomial logit GLM	Regularized GLM	Random forest	Misclassification	Cross-entropy
WoAC	3	0.184	0.184	0.211	0.729	0.584	0.509	0.184	0.509
	5	0.289	0.184	0.211	2.132	0.575	0.521	0.184	0.521
	7	0.211	0.132	0.132	5.364	0.509	0.455	0.132	0.455
	10	0.158	0.105	0.105	4.172	0.399	0.420	0.105	0.399
	15	0.237	0.079	0.079	14.136	0.515	0.430	0.079	0.430
	20	0.184	0.053	0.079	6.379	0.345	0.435	0.053	0.345
	25	0.237	0.026	0.105	6.487	0.137	0.445	0.026	0.137
RF	3	0.289	0.316	0.237	1.485	0.763	0.576	0.237	0.576
	5	0.263	0.158	0.132	3.352	0.784	0.455	0.132	0.455
	7	0.368	0.132	0.158	63.391	0.758	0.481	0.132	0.481
	10	0.289	0.184	0.158	20.956	0.688	0.493	0.158	0.493
	15	0.263	0.158	0.105	6.285	0.462	0.430	0.105	0.430
	20	0.263	0.158	0.105	13.640	0.417	0.454	0.105	0.417
	25	0.368	0.184	0.105	18.873	0.441	0.449	0.105	0.441
Lasso	3	0.237	0.237	0.211	1.023	0.535	0.547	0.211	0.535
	5	0.211	0.158	0.184	6.078	0.388	0.423	0.158	0.388
	7	0.184	0.184	0.158	4.316	0.415	0.424	0.158	0.415
	10	0.079	0.053	0.079	0.166	0.112	0.420	0.053	0.112
	15	0.158	0.053	0.079	0.166	0.145	0.426	0.053	0.145
	20	0.079	0.026	0.053	1.770	0.088	0.486	0.026	0.088
	25	0.211	0.026	0.053	2.901	0.057	0.473	0.026	0.057

multiple reaction monitoring with stable isotopically labeled standards, refitting classifiers, and validation in an independent cohort, however given the findings we feel that this effort would be well justified.

In addition to exhibiting low estimated misclassification rates when utilized for the construction of statistical classifiers, the metabolites selected by WoAC report some of the metabolic consequences of acute thrombotic MI. Specifically, an increased abundance of selected pregnenolone and corticosteroid metabolites was observed in the thrombotic MI group relative to the non-thrombotic and stable CAD groups. This may be indicative of stimulation of the hypothalamic-pituitaryadrenal axis following thrombotic MI. Evidence of activation of this axis following MI has been demonstrated in other studies that have shown increased levels of circulating adrenocorticotropic hormone [35] and copeptin [36] in the hours following acute MI. As this signal was stronger in the thrombotic MI group than the non-thrombotic MI group (especially for cortisol), an increase in these hormones in circulation may be associated with thrombosis directly in addition to acute stress. Others have demonstrated a mechanistic relationship between platelet activating factor, an important factor in stimulating platelet activation and aggregation, and glucocorticoids [37,38]. Increased abundance of selected monoacylglycerols was observed in both acute MI groups relative to the stable CAD group and these abundance distributions exhibited significant pairwise correlations. An increased abundance of monoacylglycerols may be indicative of increased hydrolysis of triacylglycerol molecules or impaired uptake of these molecules from plasma. Decreased plasma concentrations of selected amino acids in the thrombotic MI group relative to the non-thrombotic and stable CAD groups is an interesting finding and may indicate increased catabolism of amino acids to furnish ATP for the ischemic heart. All of the amino acids identified can be catabolized to produce ATP either via gluconeogenesis or ketogenesis [39]. Under ischemic conditions, the heart must utilize metabolic substrates that do not require oxygen [40]; hence, the inability to oxidize fatty acids may lead to increased utilization of amino acids. Alternatively, a decrease in plasma amino acid concentrations may be due to increased protein synthesis in activated platelets. Platelet activation results in signal dependent translation of factors involved in thrombosis such as tissue factor (TF) [41] and plasminogen activator inhibitor-1 (PAI-1) [42], which could result in diminished concentrations of amino acids following MI in thrombotic MI subjects.

Acknowledgements

The authors would like to thank the human participants who graciously agreed to take part in the study. The authors thank the members of the Atherosclerosis and Atherothrombosis Research laboratory at the University of Louisville for help with sample processing and Samantha Carlisle, M.S. for her assistance.

Funding sources

This work was supported in part by a grant from the American Heart Association (11CRP7300003) and the National Institute of General Medical Sciences, National Institutes of Health (GM103492).

Disclosures

Plasma metabolites were measured by Metabolon, Inc. (Research Triangle Park, NC).

Conflict of interest statement

None.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2018.03.007.

References

[1] D. Mozaffarian, et al., Heart disease and stroke statistics-2016 update: a report from the American Heart Association, Circulation 133 (4) (2016) e38–e360.
[2] K. Thygesen, et al., Third universal definition of myocardial infarction, J. Am. Coll. Cardiol. 60 (16) (2012) 1581–1598.
[3] E.A. Amsterdam, et al., 2014 AHA/ACC guideline for the management of patients with non-ST-elevation acute coronary syndromes: a report of the American College of Cardiology/American Heart Association task force on practice guidelines, J. Am.

Coll. Cardiol. 64 (24) (2014) e139–e228.
[4] L.K. Newby, et al., ACCF 2012 expert consensus document on practical clinical considerations in the interpretation of troponin elevations, J. Am. Coll. Cardiol. 60 (23) (2012) 2427–2463.
[5] N. Psychogios, et al., The human serum metabolome, PLoS One 6 (2) (2011) e16957.
[6] J.K. Nicholson, et al., Metabolic phenotyping in clinical and surgical environments, Nature 491 (7424) (2012) 384–392.
[7] A.P. DeFilippis, et al., Circulating levels of plasminogen and oxidized phospholipids bound to plasminogen distinguish between atherothrombotic and non-atherothrombotic myocardial infarction, J. Thromb. Thrombol. 42 (1) (2016) 61–76, http://dx.doi.org/10.1007/s11239-015-1292-5.
[8] A.P. DeFilippis, et al., Identification of a plasma metabolomic signature of thrombotic myocardial infarction that is distinct from non-thrombotic myocardial infarction and stable coronary artery disease, Plos One 12 (4) (2017) e0175591.
[9] P.J. Trainor, et al., Systems characterization of differential plasma metabolome perturbations following thrombotic and non-thrombotic myocardial infarction, J. Prot. 160 (2017) 38–46, http://dx.doi.org/10.1016/j.jprot.2017.03.014.
[10] Freue G.V. Cohen, C.H. Borchers, Multiple reaction monitoring (MRM): principles and application to coronary artery disease. Circulation: cardiovascular, Genetics 5 (3) (2012) 378.
[11] R.M. Lequin, Enzyme Immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA), Clin. Chem. 51 (12) (2005) 2415–2418.
[12] R.V. Yampolskiy, A.E.L. Barkouky, Wisdom of artificial crowds algorithm for solving NP-hard problems, Int. J. Bio-Insp. Comput. 3 (6) (2011) 358.
[13] S.K. Yi, et al., The wisdom of the crowd in combinatorial problems, Cogn. Sci. 36 (3) (2012) 452–470.
[14] D. Marbach, et al., Wisdom of crowds for robust gene network inference, Nat. Meth. 9 (8) (2012) 796–804.
[15] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics, second ed., Springer, New York, NY, 2009. xxii, 745 p.
[16] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, J. Roy. Statist. Soc. Ser. B 58 (1) (1996).
[17] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Variable selection using random forests, Patt. Recogn. Lett. 31 (14) (2010) 2225–2236.
[18] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Statist. Soc.: Ser. B (Statist. Methodol.) 67 (2) (2005) 301–320.
[19] A. Agresti, Categorical Data Analysis, Wiley Series in Probability and Statistics, 3rd ed., Wiley, Hoboken, NJ, 2013, xvi, 714 p.
[20] A.J.F. Griffiths, et al., Introduction to Genetic Analysis, eleventh ed., W.H. Freeman & Company, A Macmillan Education, New York, NY, 2015 imprint. xxiii, 868 pages.
[21] C.M. Bishop, Pattern Recognition and Machine Learning, Information Science and Statistics, Springer, New York, 2006. xx, 738 p.
[22] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of k-fold cross-validation, J. Mach. Learn. Res. 5 (Sep) (2004) 1089–1105.
[23] T. Nomura, An analysis on linear crossover for real number chromosomes in an infinite population size, IEEE Int. Conf. Evol. Comput. (1997) 111–114.
[24] G.F. Lawler, Introduction to Stochastic Processes, second ed., Chapman & Hall/CRC, Boca Raton, 2006. xiii, 234 p.
[25] J. Surowiecki, The Wisdom of Crowds: Why the Many are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations, first ed., Doubleday, New York, 2004. xxi, 296 p.
[26] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[27] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Statist. Softw. 33 (1) (2010).
[28] R.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016.
[29] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2 (3) (2002).
[30] W.N. Venables, B.D. Ripley, W.N. Venables, Modern Applied Statistics with S, Statistics and Computing, forth ed., Springer, New York, 2002, xi, 495 p.
[31] A. Alfons, cvTools: Cross-Validation Tools for Regression Models, R package version 0.3.0, 2012.
[32] S. Weston, M. Corporation, doParallel: Foreach Parallel Adaptor for the Parallel Package, R Package, 2016.
[33] D. Broadhurst, et al., Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry, Anal. Chim. Acta 348 (1–3) (1997) 71–86.
[34] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemomet. Intell. Lab. Syst. 41 (2) (1998) 195–207.
[35] F. Paganelli, et al., Hypothalamo-pituitary-adrenal axis in acute myocardial infarction treated by percutaneous transluminal coronary angioplasty: effect of time of presentation, J. Endocrinol. Invest. 26 (5) (2003) 407–413.
[36] A. Maisel, et al., Copeptin helps in the early detection of patients with acute myocardial infarction: primary results of the CHOPIN trial (Copeptin Helps in the early detection Of Patients with acute myocardial INfarction), J. Am. Coll. Cardiol. 62 (2) (2013) 150–160.
[37] T. Shimada, et al., Platelet-activating factor acts on cortisol secretion by perfused guinea-pig adrenals via calcium-/phospholipid-dependent mechanisms, J. Endocrinol. 184 (2) (2005) 381–391.
[38] T. Aikawa, et al., Effect of platelet-activating factor on cortisol and corticosterone secretion by perfused dog adrenal, Lipids 26 (12) (1991) 1108–1111.
[39] D. Voet, J.G. Voet, C.W. Pratt, Fundamentals of Biochemistry: Life at the Molecular Level, forth ed., Wiley, Hoboken, NJ, 2013.
[40] K.J. Drake, et al., Amino acids as metabolic substrates during cardiac ischemia, Exp. Biol. Med. (Maywood) 237 (12) (2012) 1369–1378.
[41] O. Panes, et al., Human platelets synthesize and express functional tissue factor, Blood 109 (12) (2007) 5242–5250.
[42] H. Brogren, Platelets synthesize large amounts of active plasminogen activator inhibitor 1, Blood 104 (13) (2004) 3943–3948.