**BRIEFINGS IN BIOINFORMATICS.** VOL 9. NO 2. 102–118
Advance Access publication February 29, 2008
doi:10.1093/bib/bbn005

# Approaches to dimensionality reduction in proteomic biomarker studies

*Melanie Hilario and Alexandros Kalousis*

Submitted: 25th August 2007; Received (in revised form): 18th January 2008

## Abstract

Mass-spectra based proteomic profiles have received widespread attention as potential tools for biomarker discovery and early disease diagnosis. A major data-analytical problem involved is the extremely high dimensionality (i.e. number of features or variables) of proteomic data, in particular when the sample size is small. This article reviews dimensionality reduction methods that have been used in proteomic biomarker studies. It then focuses on the problem of selecting the most appropriate method for a specific task or dataset, and proposes method combination as a potential alternative to single-method selection. Finally, it points out the potential of novel dimension reduction techniques, in particular those that incorporate domain knowledge through the use of informative priors or causal inference.

***Keywords:*** *proteomics; mass spectra; biomarkers; dimensionality reduction; feature transformation; feature selection*

## INTRODUCTION

### Motivation

There is a growing body of research on biomarker discovery from high-throughput biological data. The search for biomarkers is typically cast as the task of finding the most discriminating variables (or features) for classification (e.g. discriminating healthy versus diseased, or different tumor stages). A major problem is the huge number of dimensions or features—genes or proteins/peptides—that represent potential biomarkers in a microarray or mass spectrum. This high dimensionality is compounded by data sparsity: in controlled experiments, the number of available cases/controls rarely exceeds a few dozens, so that the number of variables or features $p$ is usually much higher than the number of samples $n$. This so-called high-dimensional small-sample problem raises significant data-analytical issues [1, 2]. When $p > n$, certain techniques (e.g. standard linear discriminants) will fail; others will build classifiers on noisy as well as relevant features, thus degrading generalization; still others will only find a solution at extremely high computational cost. Dimensionality reduction (DR) is therefore an indispensable preliminary step to model building. More importantly, in biomarker studies, DR is not just a computational necessity, it is an intrinsic part of the knowledge discovery task. Building an accurate model for phenotype classification is inseparable from the task of distilling a handful of biologically meaningful biomarkers from the massive set of initial variables.

### The knowledge discovery framework

Figure 1 shows the place of DR in the knowledge discovery pipeline. First, the raw data are explored and preprocessed in preparation for the learning or modeling task. In proteomic mass-spectra based diagnosis or prognosis, which is the focus of this article, preprocessing includes baseline subtraction, smoothing or denoising, intensity normalization and peak detection and alignment. Preprocessing typically reduces tens of thousands of raw variables

Corresponding author. Melanie Hilario, Computer Science Department, University of Geneva, Battelle Bât. A, 7 route de Drize, CH-1227 Carouge, Switzerland. Tel: +41-22-379 0222; Fax: +41-22-379 0250; E-mail: Melanie.Hilario@cui.unige.ch

**Melanie Hilario** holds a PhD in Computer Science from the University of Paris VI and is Assistant Professor at the University of Geneva's Artificial Intelligence Laboratory. Her research interests include biological data and text mining.

**Alexandros Kalousis** received his PhD in Computer Science from the University of Geneva and currently works as senior researcher in the University's Artificial Intelligence Laboratory. His research focuses on machine learning in proteomics-related applications.

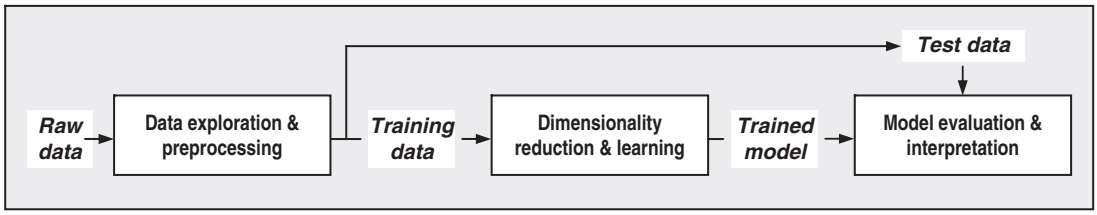


**Figure 1:** Dimensionality reduction in the knowledge discovery process.

($m/z$ points) to several hundreds of peaks (proteins or peptides). The preprocessed data are then split into training set and test set. The training set is used to build a classification model while further reducing data dimensionality; the test set is held aside for the evaluation of the trained classifier. Note that the diagram does not separate dimension reduction from model construction, for these two major facets of the data mining task can be coupled in diverse ways that will be detailed below.

## Definitions and distinctions

Formally DR is defined as follows: Given a set of $n$ vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathbb{R}^p$, find a set of lower-dimensional vectors $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \in \mathbb{R}^{p'}, p' < p$, that maximally preserve the information in the original data according to some criterion. For example, in a classification task involving $k$ classes $C = \{C_j | j = 1 \ldots k\}$, and where a class label $\gamma_i \in C$, $i = 1 \ldots n$, is associated with each data point $\mathbf{x}_i$ and its reduced version $\mathbf{z}_i$, the criterion could be some measure of how well the $p$ or $p'$ features discriminate between the different classes.

DR methods can be divided into two broad categories: feature (variable) transformation and feature selection. Feature transformation (FT) methods create a (possibly smaller) set of new features by transforming or combining the old. Feature selection methods reduce the size of the original feature set by eliminating irrelevant or redundant features. A complementary classification scheme is based on how DR is coupled to the learning (or model building) process. Filter methods perform DR as a preprocessing step, independently of the learning method. Wrapper methods wrap DR around the learning process and use the estimated performance of the learned model as the selection criterion; the effectiveness of the selected features depends strongly on the specific learning method used. In embedded methods, DR is programmed as an integral part of the learning algorithm. The filter-wrapper-embedded distinction first appeared in 1997 [3, 4] and remains widely used

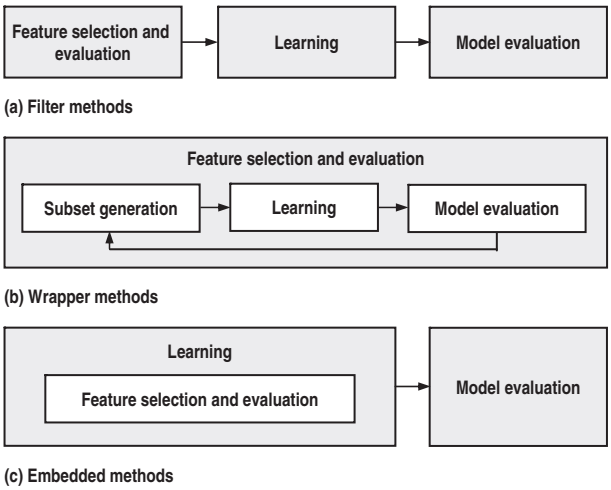


**Figure 2:** Filter, wrapper and embedded feature selection scheme.

in machine learning research [5]. Figure 2 visualizes the difference between these three schemes.

The goal of this article is to review current approaches to DR in mass-spectra based proteomic biomarker identification; Table 1 gives a synoptic view of these different approaches. The principal FT and selection methods are discussed successively in the next two sections. The problem of selecting the most effective dimension reduction method is then addressed, and method combination proposed as a potential alternative solution. The conclusion summarizes outstanding open issues and directions for future research.

## FEATURE TRANSFORMATION

FT methods construct new features as functions that express relationships between the initial features; hence FT methods have greater potential for conveying discriminatory content than feature selection methods. However, this advantage comes at the cost of comprehensibility, as complex functions of base-level features often blur the straightforward interpretation associated with the original

**Table 1:** Overview of major approaches to dimensionality reduction and examples of methods used in mass-spectra based biomarker discovery

| | Feature Transformation | | Feature Selection | | | |
|---|---|---|---|---|---|---|
| | | | Univariate | | Multivariate | |
| Filter | Unsupervised | PCA [6, 7] | Parametric | *t*-test [10, 6] | Heuristic | FS-MD [8] |
| | | Wavelets [8] | | $\chi^2$-test [19] | | CFS [39] |
| | Supervised | LDA [7] | | *F*-ratio [18] | | |
| | | PLS [16] | Non-parametric | Kolmogorov–Smirnov [22, 23] | Stochastic | GA-LDA [40] |
| | | | | Wilcoxon test [20, 21] | | Relief [42] |
| | | | | Mutual Info [26] | | |
| Wrapper | | | | | Heuristic | DTE [69] |
| | | | | | Stochastic | GA-SOM [52] |
| | | | | | | GA-SVM [54] |
| | | | | | | GA-KNN [55] |
| | | | | | | PSO-SVM [56, 57] |
| | | | | | | ACO-SVM [58] |
| Embedded | Supervised | LDA [10] | | CCM [29] | | DTs [46, 47, 48] |
| | | | | | | Boosting [50, 51, 23] |
| | | | | | | RFE [60] |
| | | PLS [15] | | Shrunken centroids [31] | | LASSO [65] |
| | | | | | | SMLR [68] |
| | | | | | | LIKNON [66] |

observations. FT methods typically learn a mapping from $\mathbb{R}^p$ to $\mathbb{R}^{p'}$, where $p'$ is not necessarily smaller than $p$. In other words, FT *per se* does not necessarily reduce dimensionality; however most FT methods provide principled ways of doing so. FT is either supervised or unsupervised, depending on whether or not class information is taken into account in designing the transformation function.

## Unsupervised methods

**Principal Component Analysis** (PCA) is the most prominent method in this category. PCA seeks linear transformations that best explain the variance in the data. Formally, it obtains the transformed data set $\mathbf{Z} = \mathbf{XW}$ by solving the optimization problem arg $\max_{\mathrm{w}}$ var($\mathbf{XW}$), where $\mathbf{X}$ is the $n \times p$ data matrix and the linear transformation is given by the orthogonal $p \times p'$ matrix $\mathbf{W}$. It does this by computing the eigenvalue decomposition of the covariance matrix $\mathbf{S} = 1/N\left(\sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T\right)$ and projecting the training instances onto the basis defined by the resulting eigenvectors (also called principal components). It can be shown that the projection directions $\mathbf{W}$ that maximize the variance are given by the solutions $\mathbf{SW}_{.i} = \lambda_i \mathbf{W}_{.i}$, where the eigenvector $\mathbf{W}_{.i}$ (the i$^{\text{th}}$ column of $\mathbf{W}$) corresponds to the eigenvalue $\lambda_i$. PCA orders the eigenvectors in decreasing order of their corresponding eigenvalues, which measure their importance in terms of the amount of variance they account for. The number of non-null eigenvalues, and therefore the dimensionality of the projection space, is upper-bounded by min ($p$, $n$) where $n$ is the number of training instances. Consequently, in cases where $n < p$, PCA will reduce dimensionality to at most $n$ without any information loss. Further reduction can be obtained by selecting the first $p' < n$ eigenvectors.

In developing a diagnostic test for African trypanosomiasis, Papadopoulos *et al.* [6] used PCA to reduce a set of 206 peaks extracted from serum mass spectra; 41 principal components accounting for 90% of the variance were selected to build and test candidate diagnostic models. PCA has also been used to reduce dimensionality of raw mass spectra prior to further reduction by Fisher's linear discriminant [7]. Raw mass spectra collected in ovarian and prostate cancer studies contained over 15 000 features; to meet constraints imposed by the use of Fisher's linear discriminant (see 'Supervised Methods' section), PCA was applied to obtain a feature set smaller than the set of training examples.

Other popular FTs include the **Fourier and the wavelet transforms**. Both of these depict a signal, in this case a mass spectrum, as a linear combination of prespecified basis functions. Discrete wavelet transforms were used to reduce the dimensionality of a prostate cancer dataset that contained 48 538 features and 248 training instances. Wavelet coefficients with

a value lower than a given threshold were discarded, yielding a new feature set consisting of 1271 wavelet coefficients [8].

## Supervised methods

In a supervised setting such as mass-spectra based classification, one of the drawbacks of unsupervised FT methods is that they do not exploit the information provided by class labels. As a result, the transformations they compute may not reflect the underlying class structure; in PCA for instance, the maximum variance directions do not necessarily guarantee maximum discrimination.

The best-known supervised FT method is **Fisher's Linear Discriminant Analysis** (LDA) [9]. Though often presented as a classification method, it is in fact a FT method that projects the initial data onto a linear subspace of dimensionality $k-1$, where $k$ is the number of classes. Like PCA, LDA effects a linear transformation of the form $\mathbf{Z}=\mathbf{XW}$ where the projection dimensions, i.e. the linear discriminants, simultaneously maximize between-class distance and minimize within-class variance. They result from the solution of a generalized eigenvalues problem $\mathbf{S}_B\mathbf{W}_{.i} = \lambda_i\mathbf{S}_W\mathbf{W}_{.i}$ where $\mathbf{S}_B$ is the between-class scatter matrix, $\mathbf{S}_W$ the within-class scatter matrix and the $i$th column, $\mathbf{W}_{.i}$, of $\mathbf{W}$ is given by the generalized eigenvector that corresponds to the $i$th largest eigenvalue $\lambda_i$ (scatter matrices are essentially unscaled covariance matrices). The solution requires the inversion of $\mathbf{S}_W$; however, when $p > n-k$, as is typical with mass-spectral data, the matrix is not invertible. One way to solve this problem is to reduce the feature set size to less than $n-k$ prior to LDA, using feature selection [10] or another FT method such as PCA [7]. An alternative solution is to use the pseudo-inverse instead of the inverse, as was done in a different application area using spectroscopic data [11].

**Partial Least Squares** (PLS) is a regression method that incorporates FT [12], but it is equally applicable to classification problems [13]. Contrary to LDA, it is not bound by any $p < n$ constraint and is therefore better adapted to high-dimensional small samples; in addition it can handle highly correlated features. Like PCA, PLS finds linear combinations of the input features that maximize variance; unlike PCA, it does this while simultaneously maximizing correlation with the target variable. For this reason, PLS usually performs better than PCA in prediction problems. Furthermore, PLS is considerably more efficient than PCA: its computational cost is $O(np)$, i.e. linear in the number of cases $n$ and the number of original predictors $p$, whereas that of PCA is on the order of min $(np^2+p^3, pn^2+n^3)$, i.e. cubic in $n$ or $p$, whichever is smaller [14]. Two studies on a lung cancer dataset illustrate two different uses of PLS in a classification setting. The first [15] directly utilized the PLS model as a classifier. A wavelet transform was used to reduce the initial set of $\sim$60 000 $m/z$ values to 545 wavelet coefficients. Based on the wavelet representation PLS built a model to discriminate lung cancer cases from controls. Potential biomarkers were identified by selecting wavelet coefficients with high regression weights in the PLS model; these wavelet coefficients were then inverse-transformed to the original $m/z$ values. The second study [16] simply employed PLS as a dimension reduction method. After binning of raw $m/z$ data points, the preprocessed dataset contained 545 features, which were transformed by PLS into a set of new features called latent components. The number of latent components to retain was based on the residual sum of squares estimated via split-sample validation. The selected PLS components were then used as predictors in classifiers built using logistic regression and discriminant analysis.

# FEATURE SELECTION

Feature selection methods output a subset of the original input features without transforming them. They are either univariate or multivariate, based on whether they evaluate individual features or feature subsets. Both univariate and multivariate methods can be used as filters prior to learning or embedded in the learning algorithm; the wrapper schema is specific to multivariate methods. A survey of feature selection techniques in a variety of bioinformatics applications can be found in [17].

## Univariate methods

Univariate methods assume mutual independence of the predictive variables. Each feature is scored or ranked based on its individual relevance, in isolation from all other features. The final feature subset is determined by a user-defined threshold on the computed scores or ranks. Most often, a feature (representing a peptide or protein) is selected when it is shown to be differentially expressed at a statistically significant level in the classes of interest (e.g. diseased versus controls). Thus standard **statistical tests**

have been widely used for this purpose. These tests rely on the same basic procedure to evaluate each variable: partition the sample according to classes (e.g. healthy versus diseased), compute a test statistic of the variable for each class, and then check for significant differences in the values of this statistic. Parametric tests, which assume a specific probability distribution of the data, as well as nonparametric tests have been used, whether in a filter or an embedded setup.

***Univariate filters***

Univariate statistical tests have been widely used as filters due to their flexibility: the selected feature set can be fed into any learning algorithm to build a predictive model. Examples of parametric tests that have been used in proteomic analyses are the *t*-test [10, 6], the *F*-ratio [18], the $\chi^2$-test [19]; nonparametric tests include the Kolmogorov–Smirnov test [22, 23] and the Wilcoxon rank test, which is equivalent to the Mann–Whitney or AUC (area under the ROC curve) test [20, 21]. Alternative criteria include *information-theoretic measures* such as mutual information [25] (known as information gain in the machine learning community), which quantifies the reduction in class entropy brought about by a given feature. Mutual information proved to be an effective variable ranking criterion in MS-based lung cancer prediction [26].

***Univariate embeddings***

Individual variable ranking or scoring can be also embedded in the learning algorithm. One example is an extension of Tukey's **compound covariate method** (CCM) [27]. In the original paper, as in Hedenfalk's [28] application of the method to gene expression analysis for breast cancer diagnosis, features were selected individually through a standard *t*-test. A linear classifier was then built using the selected features, each weighted by its computed *t*-statistic. Yanasigawa [29] adapted this method to the search for proteomic biomarkers of non-small-cell lung cancer. To reduce variance and enhance robustness, the standard *t*-test was replaced by a battery of six statistical tests (e.g. Kruskal–Wallis test, random permutation *t*-test, information score); a feature was selected if it passed three of the six tests. This method resulted in the selection of 82 peaks that were differentially expressed between tumor and normal lung tissue. The final compound covariate model, a linear combination of the selected features, perfectly classified a blinded test set as either diseased or normal. However, the method did less well in discriminating histological subgroups such as adenocarcinoma versus large-cell (94% test set accuracy) or mediastinal nodal involvement (75% test set accuracy).

A related approach is **centroid shrinkage**, a feature selection method embedded in the nearest-centroid classification algorithm. In this simple learning scheme, the training samples are used to compute the class centroids; a test sample is assigned to the class with the closest centroid. Class centroid computation is strictly univariate: the *i*th component of the centroid of a given class $k$ is $\bar{x}_{ik} = \sum_{j=1}^{n_k} x_{ij}/n_k$, where $x_{ij}$ is the value of the *i*th variable in case $j \in k$ and $n_k$ is the number of cases in class $k$. Similarly, the *i*th component of the overall centroid is $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$, where $n$ is the total number of cases. To reduce the number of features, the distance between the class centroids and the overall centroid is shrunk by an amount determined by $\triangle$, a user-tuned parameter; the higher the shrinkage parameter, the more rapidly the class centroids move to the overall mean. Shrinkage can reduce the distance between the class mean and the overall mean to zero for noisy or non-discriminatory features, which are in effect eliminated. The method was first applied to gene expression analysis for diagnosis of small round blue cell tumors [30]. It was later integrated into a comprehensive classification procedure for mass spectra called 'peak probability contrasts' and applied to ovarian cancer diagnosis. Preprocessing of the raw mass spectra led to the extraction of 192 peaks, which were reduced by centroid shrinkage to seven peaks. In a comparative study of seven methods, the resulting classifier achieved a cross-validation error of 23 (out of 89) and was outperformed only by a support vector machine (SVM), which scored two errors less by using all initial 91 360 peak sites [31].

The main advantage of the univariate approach is its efficiency; it requires computing no more than $p$ scores. However, it has a number of drawbacks [32]: it cannot detect correlated or redundant features, or interacting features (i.e. features which are irrelevant by themselves but highly discriminatory in combination with others). Multivariate approaches are meant to overcome these limitations.

## Multivariate methods

Multivariate methods [5] assess the predictive power of feature subsets rather than individual features, hence they take feature dependencies into account in the feature subset selection process. The major

difficulty is that the number of possible subsets increases exponentially with the number of features. This precludes exhaustive search (i.e. the strategy of generating and evaluating all $2^p - 1$ possible subsets of $p$ features [33]) for all but trivial datasets; heuristic search strategies are needed. Forward selection and backward elimination [33] are classical examples of heuristic search. Forward selection starts with an empty variable subset $S$ and selects the variable that maximizes a predefined scoring function. Thereafter it selects from the remaining variables the one which, added to $S$, maximizes the score of the resulting subset. The process continues until a predefined criterion is met, e.g. until the score of $S$ ceases to improve. Backward elimination proceeds in the reverse direction; it starts with the full variable set and at each step removes the variable whose elimination yields the highest score for the remaining subset. Both are greedy search strategies that are not guaranteed to achieve optimal results. As a partial remedy to the myopia of greedy search, so-called floating strategies [34] allow forward (backward) selection to eliminate (add) previously selected (eliminated) features. Alternatively, stochastic search methods use randomization to overcome the main pitfall of greedy methods, that of being trapped in local optima. Among these, biologically inspired techniques—so called because they mimic mechanisms underlying the behavior or evolution of living populations—have proved to be effective strategies for finding discriminatory feature subsets. Examples are Genetic Algorithms [35], Ant Colony Optimization [36] and Particle Swarm Optimization [37], whose application to protein biomarker selection from mass spectra are discussed in the following subsections.

### *Multivariate filters*

A number of variable subset selection strategies have been used as filters prior to the learning process. Forward selection has been used with different scoring functions in two mass-spectral applications. In one experimental study on prostate cancer detection [8], a discrete wavelet transform reduced the initial mass set to 1271 derived features. Forward search was then applied to find a subset that maximized the Mahalanobis distance (MD) between the cancer cases and controls. Informally, the MD between two groups is computed as the Euclidean distance between their centers (group means), normalized by their covariance. This method (henceforth FS-MD) resulted in a subset of 11 variables, which were then used to build a linear discriminant model. In another forward selection scheme called **Correlation-Based Feature Selection** (CFS), the evaluation criterion is based on the idea that useful features are highly correlated with the class yet uncorrelated with each other [38]. Correlation is measured in terms of symmetrical uncertainty [25], a normalized form of mutual information that captures dependencies other than linear correlation. In a comparative study where CFS and five univariate filters were coupled with five different learning algorithms, CFS was found to yield best performance for mass-spectra based ovarian cancer diagnosis [39].

To discover lung cancer biomarkers based on a dataset of 41 serum MALDI-TOF mass spectra, Baggerly *et al.* [40] focused their search on very small feature subsets. Low-level preprocessing of the raw mass spectra reduced the 60 831 initial $m/z$ values to 506; however, given the available data, a much smaller feature set was needed to build a linear discriminant. Exhaustive search was used to find promising subsets of one or two features, and **Genetic Algorithms** (GAs) subsets of 3–5. A feature set was considered optimal if it maximized the MD between the lung cancer and control groups. Fifty GA runs were done for each subset size $p = 1$ to 5; for each run, an initial population of 200 $p$-feature sets was generated and allowed to evolve for 250 generations using the standard genetic operations: selection, mutation and crossover. For each feature set, Fisher's LDA was used to build a separating hyperplane between the two classes, and the classification error estimated using leave-one-out cross-validation. A single misclassification error (2.4%) was reported for the 5-feature linear classifier.

**Relief** [41] computes the relevance of each predictive variable using a method based on K-nearest neighbors. In a binary classification problem, it repeatedly picks a case at random and identifies the case's nearest neighbor from the same class and its nearest neighbor from the other class. It then adjusts feature weights by rewarding features that discriminate neighbors from different classes while penalizing those with different values for neighbors of the same class. Although feature weights are updated separately, Relief is basically multivariate since distance computation underlying nearest neighbor identification takes joint account of all features. Relief can be used as a feature selection

filter for any learning algorithm; different variants of Relief were shown to outperform a number of state-of-the-art feature selection methods as a filter to SVMs [42] and other machine learning algorithms [26]. In [43] Relief is viewed as an online algorithm that solves a convex optimization problem with a margin-based objective function; this recent interpretation sheds new light on Relief's strengths and weaknesses as well as suggests ways of mitigating the latter.

***Multivariate embeddings***

**Decision trees** (DTs) like CART [44] and C4.5 [45] are classical examples of learning algorithms that embed heuristic feature selection. DT construction is driven by a sequential forward search in the space of candidate feature subsets, much like FS-MD and CFS (see previous subsection). At each leaf node of the partially built tree, the algorithm selects the feature that maximally reduces the class impurity (or entropy) of the examples associated with that node. One measure of class entropy reduction is C4.5's information gain, defined as $I(X; C) = H(C) - H(C|X)$, where $C$ is the class variable, $X$ a predictor variable and $H(\cdot)$ is the entropy. In words, information gain is the decrease in class entropy brought about by the predictor variable $X$. However, DTs are multivariate rather than univariate; in reality, what is measured is the *cumulative* reduction in entropy brought about by the feature subset consisting of all features along the path from the root to the current node. DTs have been applied directly to SELDI-MS peaks extracted by Cipherghen's built-in software: five biomarkers of renal cell carcinoma were identified using C4.5 [46], while three biomarkers of prostate cancer were selected by CART [47]. DT-based biomarker identification can also be preceded by other feature selection methods such as univariate ranking filters [48].

Another embedded multivariate technique consists in building **ensembles** or committees of univariate classifiers, which are then combined to yield a single prediction. A widely used ensemble learning method is **boosting** [49], which builds a sequence of classifiers from adaptively generated data. At each iteration, a classifier is built and its accuracy on the training data is estimated. The weights of misclassified cases are increased and those of correctly classified cases decreased, so that the learning process progressively focuses on the most difficult cases. After training, a test sample $\mathbf{x}$ is classified using the rule $H(\mathbf{x}) = \mathrm{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})\right)$, where $t$ indexes each of the $T$ base classifiers, $h_t$ is classifier $t$'s prediction and $a_t$ its weight, determined by its accuracy on the training set. The result is a multivariate feature selection scheme embedded in the boosting algorithm. As any base learner can be used provided it performs at least slightly better than chance, simple schemes suffice to build univariate classifiers: single-node decisions trees [50], linear discriminants [51] and nearest centroids [23] have been used to build boosted ensembles for cancer diagnosis.

Feature subset selection is typically a yes/no decision (features are either selected or eliminated). It is therefore prone to high variance: small changes in the data can lead to very different models. To attenuate this problem, so-called soft feature selection methods assign continuous weights to features and use these weights to determine the final subset. In Recursive Feature Elimination (RFE) [60], the candidate feature set (initially the set of all variables) is used to train a linear SVM; the features are ranked in decreasing order of their (squared) weights in the SVM hyperplane, and the lowest ranked features are eliminated. The process is reiterated until a predefined feature subset size is reached. When the cost function is a quadratic function of the model weights, this is equivalent to selecting the feature subset that minimizes the cost function. RFE was first used to select genes for cancer classification [60]; it has since been extended in diverse ways and applied to proteomics-based diagnosis of ovarian [61, 62] and breast cancer [63].

While RFE reduces the feature set by recursively applying the SVM learning algorithm, other feature weighting schemes directly build sparse models (regressors or classifiers), i.e. models in which most of the features have zero weights. Two such methods build linear models while minimizing the $L_1$ norm or $\|\mathbf{w}\|_1 = \sum_{j=1}^{p} |w_j|$ (in words, the sum of the absolute values of the weights). The **Lasso** method [64] solves the problem $\min_{\mathbf{w},b} \sum_{i=1}^{n} (\mathbf{w} \cdot \mathbf{x}_i - \gamma_i)^2$ subject to $\sum_{j=1}^{p} |w_j| \leq t$, an optimization problem that is solved via quadratic programming. Making the threshold $t$ sufficiently small will cause some of the coefficients to be exactly zero, thus eliminating the corresponding features. The Lasso was used to build a linear discriminant between cases and controls in a study on head and neck squamous cell carcinoma. Input data consisted of 32 055 $m/z$ values per SELDI-TOF mass spectrum; of these, only 65 retained non zero weights in the final model, which achieved 68% sensitivity and 73% specificity on a masked test

set [65]. **LIKNON** [66] embeds the optimization problem in SVM classification and uses linear programming to minimize the $L_1$ norm of the weight vector. Applied to a 200-sample ovarian cancer dataset, it produced a classifier with non-zero weights for only 52 of the 15 154 initial features. Finally, **Sparse Multinomial Logistic Regression** (SMLR) [67] adopts a Bayesian perspective on the same problem; to build sparse models, the likelihood of the training data is regularized by a sparsity-promoting prior belief on the weights. SMLR uses a Laplacian prior, which has been shown to be equivalent to minimizing the $L_1$ norm [64]: formally, $p(\mathbf{w}) \sim e^{-\lambda||\mathbf{w}||_1}$. In a lung cancer study involving 68 mass spectra, low-level preprocessing yielded an initial set of 308 peaks from raw mass spectra of lung tissue samples [68]. SMLR used these peaks to build a classification model with four candidate biomarkers.

***Multivariate wrappers***

Multivariate wrappers were motivated by the insight that the quality of a feature set is best revealed by the performance of the model built on it. Hence DR is implemented as an outer loop around the learning process, and the feature set that produces a model with the highest test accuracy is selected. Geurts *et al.* [69] followed a more complex procedure based on binary DTs together with the entropy reduction measure

$$I(\text{node}) = |S|H(S) - |S_T|H(S_T) - |S_F|H(S_F),$$

where $S$ is the sample that reaches the node and $S_T$ ($S_F$) is the sample subset for which the test is true (false); given a sample $s$, $|s|$ is its size and $H(s)$ its class impurity. Contrary to information gain (see 'Multivariate Embeddings' subsection), $I$(node) assigns higher scores to splits near the node. After DT construction, the feature selection process can be summarized as follows:

(1) Assign a score to each feature as the sum of $I$ (node) for all nodes that were split using the feature.
(2) Rank all features in decreasing order of their total scores, producing a nested series of candidate feature subsets.
(3) Rerun the learning algorithm on progressively larger feature subsets, starting from the top-ranked features. The accuracy estimates on these nested feature subsets typically increase to a maximum, then decrease.
(4) Select the smallest feature subset that produced a tree whose error rate is within 1 SE of the smallest observed error.

Biomarkers for rheumatoid arthritis and inflammatory bowel disease were identified by wrapping this procedure around ensemble learners such as boosted DTs.

**Genetic Algorithms** have been used as feature selection wrappers in several studies on ovarian cancer diagnosis. In Petricoin *et al.*'s [52] work, the input to the data mining process contained 15 154 $m/z$ values or features. Evolutionary computation started with an initial population of 1500 feature sets, each containing between 5 and 20 $m/z$ values. Each feature set was used to train a Self-Organizing Map (SOM), a neural network that clusters the input samples while preserving the topology of the input space [53]. Feature sets that produced a map with homogeneous cancer and control clusters were selected to spawn a new generation of feature sets through crossover and mutation. The learning process halted after 250 generations or when a map was found that perfectly separated the cancer and control cases. Two other studies on the same datasets wrapped GAs around SVMs [54] and K nearest neighbors (KNN) [55]. GA-SVMs consistently outperformed a univariate method based on a standard statistical test, and ROC curves confirmed this dominance. In the GA-KNN setup, the 10 features that appeared most frequently in the final population were used to train a 5-NN which achieved 97% accuracy, averaged over 50 runs.

Like GAs, swarm intelligence methods such as **Particle Swarm Optimization** (PSO) [37] and **Ant Colony Optimization** (ACO) [36] are population-based, stochastic optimization methods. While GAs are based on the evolutionary metaphor, the latter methods draw inspiration from the collective intelligence manifested by bird flocks (PSO) or ant colonies (ACO). Like GAs, they are initialized with a population of random solutions and iteratively update generations to find optimal solutions; candidate solutions are assessed using a chosen fitness function. In PSO, each particle (or solution) moves with randomly increasing velocity towards a location determined by its own previous best position and the known best collective position—either the best global solution of the entire swarm or the best local solution found within the particle's neighborhood. PSO has been used to reduce dimensionality

and identify potential biomarkers of hepatocellular carcinoma (HCC) [56, 57]. Low-level preprocessing of SELDI-QqTOF serum mass spectra reduced the ∼340 000 raw $m/z$ values to 368 features ($m/z$ windows). From these features, PSO generated an initial population of $N$ particles, each composed of $p$ randomly selected features. Each particle or feature set was used to train a linear SVM, whose performance determined whether the particle was fit enough to contribute to the next generation of particles. Experiments with different particle sizes $p = 5 \ldots 10$ and population sizes $N = 50$ and 100 led to the selection of 7–9 potential biomarkers. These achieved up to 91% sensitivity and 92% specificity in discriminating HCC cases from healthy controls on an independent test set.

In a subsequent study [58], the same core team used ACO to identify MALDI-TOF based biomarkers that distinguish HCC from cirrhosis. In ACO, artificial ants build a solution much like real ants build pheromone trails to optimize their search for food. The solution space is modeled as a fully connected construction graph $G_C = (V, E)$ where $V$ is a set of vertices and $E$ a set of edges. At a given time point $t$, each ant $k$ moves from vertex $i$ to a vertex $j$ within its neighborhood according to a transition probability function:

$$p_{ij}^{k}(t) = \frac{\tau_{ij}(t)^{\alpha}\eta_{ij}^{\beta}}{\sum_{l \in N_i^k} \tau_{il}(t)^{\alpha}\eta_{il}^{\beta}},$$

where $\tau_{ij}$ and $\eta_{ij}$ represent, respectively the pheromone intensity and heuristic value associated with the edge $(i, j)$, and $\alpha$ and $\beta$ are parameters that determine their relative importance. The above function shows clearly that an ant is more likely to select an edge with a higher pheromone level and/or heuristic value. In the HCC/cirrhosis study, the heuristic value was computed using Golub *et al.*'s [59] signal-to-noise ratio $\eta_i = |\mu_{1i} - \mu_{2i}|/\sigma_{1i} + \sigma_{2i}$ where $\mu_{1i}$ and $\mu_{2i}$ are the mean intensities of peak $i$ in the HCC and cirrhosis groups, respectively, and $\sigma_{1i}$ and $\sigma_{2i}$ the corresponding SDs. $\alpha$ and $\beta$ were both set to 1, so that an ant is more likely to select an edge that maximizes the product of pheromone intensity and heuristic information.

After each move of all ants at time $t$, the pheromone intensity of each arc is updated as follows:

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \rho \sum_{s \in S_{\text{good}}} F(s),$$

where $\rho$ is the so-called evaporation factor and $F(s)$ is a fitness function that quantifies the quality of solution $s$. In words, pheromone update blends a form of forgetting with reinforcement of edges traversed by good solutions.

In the HCC versus cirrhosis experiments, the evaporation factor $\rho$ was set to 0.1, and the fitness function $F(s)$ used was the cross-validated accuracy of a linear SVM classifier built on each candidate feature subset $s$. A set of 50 ants (candidate solutions) was generated, each containing five features randomly selected from 228 peaks ($m/z$ windows) detected via low-level preprocessing. The ACO-SVM wrapper was run 100 times, and all peaks were ranked based on their frequency of occurrence in the final solutions of all 100 runs. The eight top-ranked peaks were used to build an SVM classifier, which scored 94% sensitivity and 100% specificity on a blinded test set.

## THE METHOD SELECTION PROBLEM

The above survey, though far from exhaustive, gives an idea of the profusion and diversity of techniques for DR. This raises a critical question: which is the best method for the given task and data? We call this the DR method selection problem to distinguish it from the related but distinct issue of model selection in machine learning. On both issues, there is no universally superior model or method; the most appropriate choice depends on multiple interacting factors relative to the domain task, the available data and the user's priorities.

However, the major contending DR approaches—FT versus feature selection, univariate versus multivariate feature selection, filter versus wrapper versus embedded setups—have known strengths and weaknesses that have been described above. Matching these with the characteristics of the available data would significantly reduce the combinatorics of choice. For instance, a number of rules of thumb can be applied, based on whether the learning process should tackle raw or preprocessed mass spectra. Raw spectra of biological samples contain thousands to hundreds of thousands of features, many of them correlated. This immediately precludes the use of multivariate wrappers, which would incur prohibitive computational costs. However, the number of remaining candidates that would both reduce and decorrelate the feature set remains large. When DR takes place downstream of a

**Table 2:** In binary classification problems, the confusion matrix shows the number of correctly and incorrectly classified cases

| Actual Predicted | Positive | Negative |
|---|---|---|
| Yes | TP | FP |
| No | FN | TN |

TP (true positives): correctly classified positive cases; TN (true negatives): correctly classified negative cases; FP (false positives): negative cases classified as positive; FN (false negatives): positive cases classified as negatives.

**Table 3:** The most widely used classification performance measures, defined on the basis of the quantities shown in confusion matrix A

| Performance measure | Formal definition |
|---|---|
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| Sensitivity (Recall) | $\frac{TP}{TP+FN}$ |
| Specificity | $\frac{TN}{TN+FP}$ |
| Balanced accuracy | $(\text{Sensitivity} + \text{Specificity})/2$ |
| Geometric mean accuracy | $\sqrt{\text{Sensitivity} \times \text{Specificity}}$ |

TP (true positives): correctly classified positive cases; TN (true negatives): correctly classified negative cases; FP (false positives): negative cases classified as positive; FN (false negatives): positive cases classified as negatives.

preprocessing pipeline, the range of options is even larger. DR method selection should take into account the number, type (proteins/peptides or derived features such as principal components or wavelets) and characteristics (e.g. correlated or not) of the preprocessed features.

## Evaluation criteria

In all cases, principled method selection relies on clear *quantifiable criteria*:

- The *classification performance* of the final model is the outstanding criterion in the data-analytic phase of biomarker discovery, pending definitive validation through, e.g. biological assays and clinical trials. The most commonly used performance measure is accuracy, or the proportion of correctly classified test cases. Its main disadvantage is that, in the case of imbalanced class distributions, the accuracy rate is dominated by the majority class and becomes misleading when correct prediction of the minority class is critical. Hence preference is sometimes given to class-specific performance measures such as sensitivity and specificity, defined, respectively as the proportion of correctly classified cases in the positive and the negative class. Less biased scores that measure performance over both classes while giving them equal importance are the geometric mean or the arithmetic mean of sensitivity and specificity. These performance measures are defined formally in Tables 2 and 3.
- A second criterion is the *complexity* of the feature set, quantified for our purposes as the number of selected features. (The complexity of the learned model depends partly on the complexity of the feature set and is a much more intricate issue that is beyond the scope of this article.) The size of the feature set has a significant impact on both the performance and the interpretability of the final model. Experiments reported in [70] suggest that the DR method and the number of selected features are more important than the learning algorithm in constructing a reliable predictive model. Should there be a conflict between feature set size and generalization performance, protein biomarker identification is one task where minimizing the number of features might be more important than classification performance. In the current state of technology, the number of candidate biomarkers that biology researchers can identify and validate for diagnostic purposes is severely limited. For this reason, most proteomic studies focus on biomarker panels reduced to a handful of features despite the possibility that much larger feature sets might yield a gain in accuracy.
- A third criterion is the *stability* of the selected feature set. Users instinctively have more confidence in DR methods that select reasonably similar feature subsets across slightly varying experimental conditions. For instance, one of the striking findings of Ressom *et al.*'s [56] study (see 'Multivariate Embeddings' subsection) was the fact that a set of seven features consistently appeared in the best 7–8 feature subsets produced by PSO-SVM using different population sizes as well as in a 128-feature set selected using a two-tailed *t*-test. Although feature stability cannot override classification performance when selecting between candidate feature subsets, it is a useful auxiliary criterion when performance rates of models built from them are not significantly different. The issue of feature stability has been relatively neglected to date; an in-depth analysis in the context of proteomic, genomic and biological text mining applications can be found in [71].

## Evaluation methodology

Similarly to model selection in learning, method selection in DR requires a rigorous evaluation methodology. The results of many proteomics experiments need to be taken with caution due to methodological flaws in data analysis experiments. Similarly, the performance measures given in the preceding sections for illustrative purposes should not be taken as indications of the relative merits of the methods used. These measures are simply incomparable across such different experimental settings.

The fundamental evaluation rule is that a model should never be tested on the same samples that were used to build the model in the first instance; otherwise performance estimates will be overly optimistic. The dataset should be split into two disjoint sets: the first is used to train the classifier and the second held aside for performance test, hence the name *holdout split*. To increase statistical reliability, performance measures are estimated by computing their means, SEs and/or confidence intervals over several holdout splits. With limited available data, resampling techniques such as bootstrap and cross-validation allow us to randomly generate the necessary number of training/test partitions from the original sample. *Bootstrap* resampling consists in randomly drawing cases with replacement to create multiple samples of the same size as the original data set. In *k-fold cross-validation*, the initial dataset is split into $k$ subsets. At each iteration $i$, the subset is reserved for testing and a model is built using the $k-1$ remaining subsets. The overall performance measure (e.g. misclassification error) is obtained by averaging over the $k$ iterations. Cross-validation with $k=n$, where $n$ is the number of examples, is known as *leave-one-out cross-validation*.

A major pitfall to avoid is what is now known as the *feature selection bias*. It is an unfortunate but common practice (e.g. [40, 10]) to perform DR on the whole dataset and use the low-dimensional version of the same dataset to train and evaluate a classifier. Since DR is an integral part of the learning process, this is tantamount to building and testing the model on the same data, and leads to optimistically biased performance estimates. Ambroise and McLachlan [70] have shown that with this method, one can get near perfect accuracy estimates on a randomly labeled training set, simply by increasing the number of features. However, such optimistic results are completely misleading; reliable measures of generalization performance can only be obtained on test instances that have been used neither for learning nor for DR. In the case of high-dimensional small samples for which cross-validation is used, this requires two nested cross-validation loops: an outer loop for training and testing the classifier, and at each step of this outer loop, an inner loop for selecting the feature set [2].

# METHOD COMBINATION

Given the difficulty of selecting the right method for reducing high-dimensional data, an attractive idea is to combine several of the existing methods. The rationale is that purposive combination of different methods could leverage their strengths to overcome their respective weaknesses, in the same way that ensemble learning [72] combines the predictions of several base-level classification models to improve predictive accuracy. However, DR method combination is distinct from ensemble learning: the former blends feature sets obtained via different selection or transformation techniques, not learned models. Similarly to ensemble learning, DR method combination techniques can follow a serial or a parallel scheme [73].

In the **serial** scheme, the different methods are applied successively, and the output of one method is input to the next. The goal is to distribute the DR task among methods with complementary strengths. In feature selection, for instance, univariate methods are extremely fast but ignore feature redundancy and interaction; multivariate methods capture what univariate methods ignore, but incur higher computational costs. By applying a univariate method to a very large feature set and a multivariate method to the resulting reduced set, we obtain the advantage of multivariate methods at a much lower cost. Kozak *et al.* [20] applied univariate statistical tests to an initially large peak set and identified 10 differentially expressed proteins between ovarian cancer cases and controls. Logistic regression with embedded backward elimination was then applied to remove redundant proteins, producing the final 5-protein biomarker panel in significantly less time. Lilien *et al.* [7] sequenced two FT methods, each at the processing endpoint where it was most effective: LDA cannot be applied to raw spectral data where $p \gg n$, but PCA can. On the other hand, LDA is a supervised method and can, contrary to PCA, exploit class information to reduce dimensionality drastically while ensuring maximal class separation. To exploit

these complementary advantages, PCA was applied to reduce the >15 000 raw spectral features to $n-k$ derived features; these were then used by LDA to produce a discriminant with $k-1$ features. All that was left to do after LDA was tune a threshold to optimize classification performance.

In the *parallel* method combination scheme, several alternative methods are applied independently to achieve a given task, and their results integrated via a decision rule. The goal is to reduce the variance or instability of feature sets obtained by individual methods. A simple form of this approach is illustrated in univariate ranking systems where single statistical tests are replaced by up to half a dozen parametric and nonparametric tests; only features that pass all or the majority of these tests are retained [20, 24, 29]. Chan *et al*. [74] combine FT (e.g. PCA) with univariate (e.g. Wilcoxon) and multivariate (e.g. CFS) feature selection, the latter using different search strategies. The feature subset selected by each method was fed to two learning algorithms, neural networks and Naive Bayes. Each feature was then assigned a weighted score based on the accuracy rates of all classifiers built using that feature. Formally each feature's weighted score is defined as $WS(f_j) = \sum_{i=1}^{M} (1/F_i) e_{ij} a_i$, where $M$ is the number of classifiers, $e_{ij} = 1$ if the feature $f_j$ was used in classifier $i$, $a_i$ is the accuracy of classifier $i$ and $F_i$ is the number of features for classifier $i$. A classifier was built using the highest scoring features, successively adding a feature until performance ceased to improve. On one ovarian cancer dataset, neural nets achieved 100% and Naive Bayes 98.4% cross-validated accuracy using six features. The selected 6-feature set was then used to build a neural net classifier on a different ovarian cancer dataset and attained 83% accuracy, demonstrating that the selected features were reasonably robust. Consensus biomarker selection [75] uses four different scoring functions and unifies their individual rankings using two distinct methods. The first is a rank aggregation algorithm, which defines a Markov chain over the features and computes transition probabilities between them based on their position in the partial rankings. The final aggregated ranking is a list of features sorted by their stationary probabilities. In the second method, a consensus feature set is created by taking the union of the $k$ features top-ranked by the univariate rankers. Principal components analysis is then applied to the data matrix reduced to these $k$ consensus features; only the components that explain at least 0.1% of the variance are retained, producing an even smaller set of uncorrelated consensus features. In an experiment on prostate cancer, features derived from consensus PCA and standard PCA were fed into four learning algorithms. In 20 cross-validation runs, consensus PCA consistently outperformed standard PCA for values of $k$ varying between 2 and 30. However, its superiority was less clear in tests conducted on an ovarian dataset.

## FUTURE DIRECTIONS

Similarly to model selection for classification, method selection for DR remains a widely open issue. In the wake of impressive successes in model combination over the past decade, advances in DR could be expected by exploring fresh tracks in method combination. At the same time, the search for novel individual techniques remains indispensable, as successes in method combination will depend in part on the power of the base-level methods mobilized in the aggregation process.

Despite a few remarkable attempts to exploit techniques from machine learning and data mining, work in proteomic biomarker discovery has been dominated by the use of standard statistical tests aimed at identifying differentially expressed proteins among the different groups under study. Many state-of-the-art methods for dimension reduction remain untapped. For instance, Random Projection [76, 77] is a family of FT methods that map the original $p$-dimensional data onto a $k$-dimensional ($k << p$) hyperplane via a random $k \times p$ matrix $\mathbf{R}$ whose columns have unit length. Its key property arises from the Johnson–Lindenstrauss lemma [78]: for every set of $n$ points in a $p$-dimensional vector space, there exists a mapping onto a subspace whose dimensionality is lower-bounded by $\log n/\varepsilon^2$ (where $\varepsilon$ is some desired distortion level), such that the distances between the points are approximately preserved. It has been shown that this lemma holds for any random matrix whose elements follow a zero mean, unit variance distribution, ensuring bounded distortion in the move from a high to a lower-dimensional space. This guarantee of low distortion, coupled with significantly lower computational costs than, e.g. PCA, makes Random Projection an advantageous alternative to traditional projection-based DR methods. Since the dimensionality $p'$ of the projection space is determined only by the cardinality $n$ of the original data rather than its dimensionality $p$, Random Projection is particularly

appealing when $n << p$. It has been successfully applied to the selection of the number of clusters in gene expression clustering [79] and might be profitably exploited in proteomic biomarker studies, assuming that the original distance (which uses all the features) is correct.

Contrary to the linear FT techniques that have been applied in proteomics (e.g. standard PCA, LDA and PLS, see 'Feature Transformation' section), a number of recent FT techniques perform nonlinear DR. Kernel PCA [80] extends standard PCA by finding principal components that are nonlinear combinations of the original features. Conceptually, this is done by mapping the input data onto a high dimensional feature space and by performing PCA in that space. Kernel PCA therefore finds principal components that are nonlinear combinations of the original variables in the input space. In reality, the computational overhead of working under high dimensionality is circumvented through the use of kernels, i.e. functions that compute inner products of instances in high dimensional spaces without explicitly computing the mapping. One major difference is that the number of principal components is upper-bounded by $\min(n, p)$ in standard PCA and by $\min(n, p_{fs})$ in kernel PCA, where $p_{fs}$ is the dimensionality of the feature space. Whereas this could raise a problem in classical applications where $n > p$, in the small-sample setting typical of genomic and proteomic experiments, the initial number of principal components derived by kernel PCA remains far smaller than the dimensionality of the feature space. However, the loss of interpretability incurred by kernel PCA is a potential issue.

Among the feature selection methods, recent advances in $L_p$-based regularization could be fruitfully exploited in proteomic data analysis. $L_1$ (e.g. Lasso) and $L_2$ norm based methods (e.g. SVMs and ridge regression) control the variance of the feature weights and therefore improve predictive accuracy, especially when many features are correlated—a typical case in proteomic and microarray data. The difference between the two norms lies in the fact that, due to the nature of the $L_1$ norm, many coefficients are forced to become exactly zero ('Multivariate Embeddings' section), thus favoring feature sparsity. In addition, it has been shown [81] that the sample complexity of $L_1$-regularized logistic regression is logarithmic in the number of features, whereas that of $L_2$-regularized logistic regression is linear in the number of features. However $L_1$-norm based methods still suffer from a number of limitations. One of these is the difficulty of choosing $\lambda$, the regularization coefficient that controls the trade-off between accuracy and sparseness of the solution. This has been typically adjusted via cross-validation, but an algorithm has recently been proposed [82] that efficiently derives the complete solution path, i.e. all variable coefficients for all possible values of $\lambda$, thus greatly facilitating model selection in $L_1$-norm SVMs. Preliminary tests on gene biomarkers for leukemia diagnosis suggest the potential utility of the method in proteomics. Another limitation is that in the case of highly correlated variables (a common situation in mass-spectra based proteomics), $L_1$ methods typically select one of the correlated variables regardless of its merit relative to the others. To overcome this problem, the elastic net method [83, 84], based on a convex combination of $L_1$ and $L_2$, assigns similar weights to highly correlated features, in effect grouping them so that they are included in or excluded from the model together. The elastic net has been adapted to logistic regression using the $L_p$ norm, $p \leq 1$, and tested on a number of microarray datasets [85].

Other promising research directions can be gathered from work on metric learning. The goal is to learn the most appropriate distance metric for a given problem, typically by assigning weights to the different features, both in supervised [86–88] and unsupervised [89, 90] learning. The problem is cast as a mathematical optimization task, e.g. minimize the sum of distances between objects from the same class while maximizing the sum of distances between objects from different classes. The distance metric between two data points $x$ and $y$ is defined in terms of a transformation matrix A: $d^2_{\mathbf{A}}(x, y) = \|x - y\|^2_{\mathbf{A}} = (x - y)^T \mathbf{A}(x - y)$ where A should be positive semi-definite in order to ensure two defining characteristics of a metric, nonnegativity and the triangle inequality. Here too, one can impose the same type of sparsity constraint on the $L_p$ norms of the transformation matrix A.

A relatively neglected but crucial issue is the role of prior knowledge in proteomic data DR. The use of $L_p$ norms in feature weighting/selection methods can be viewed from a Bayesian perspective as equivalent to the introduction of a prior over the feature weight vectors [91]; the regularized solution is then equivalent to the maximum a posteriori solution. The $L_2$ norm corresponds to a Normal

prior, $p(\mathbf{w}) \sim N(\mathbf{0},\text{diag}(\sigma^2))$, while $L_1$ corresponds to a Laplacian prior, $p(\mathbf{w}) \sim e^{-\lambda L_1(\mathbf{w})}$. All these methods allow us to directly model relevant domain knowledge in the form of priors over variable weights when this information is available. A related method is the Relevance Vector Machine [92], which allows for the possibility of assigning a distinct parameterized prior $p(\mathbf{w}) \sim \prod_i N(0,\sigma_i^2)$ to each training point.

In addition to informative priors, specifically biological knowledge should be brought to bear on the dimension reduction and modeling process. As shown in [1], one of the consequences of proteomic data sparsity is that several feature sets will yield identically high and even perfect test accuracy; this nonuniqueness casts doubt on the biological relevance of so-called optimal biomarker panels and argues for feature selection techniques that are not purely data-driven [93]. The search for meaningful biomarkers needs to be constrained and directed by current biological knowledge concerning, e.g. suspected genes/proteins or pathogenetic pathways, or simply the experimental environment that gave rise to the data under study. In this regard, promising tracks have been opened by groundbreaking work on the role of causality in feature selection [94]. Sorting out causal relationships from the correlations or statistical dependencies uncovered by relevance measures has many advantages. Causal inference from domain knowledge can help us understand the data structure and distinguish, for instance, features that are essential to the system under study (e.g. an *m/z* point representing a protein that is overexpressed due to the target disease) from simple experimental artifacts (e.g. an *m/z* point whose intensity corresponds to the baseline). Also, from a practical point of view, knowing which features are causes or consequences of the target (e.g. disease) under study is critical in medical decision-making: a diagnosed risk of disease can be averted by acting on a cause but not by acting on a consequence of the disease. A preliminary theoretical analysis of causality-based feature selection has refined the concept of feature relevance in the framework of causal Bayesian networks. These are computational models that are fully defined by their graphs (directed acyclic graphs in which nodes represent features and an edge from node $X_1$ to $X_2$ means $X_1$ directly causes $X_2$) and the conditional probabilities $P(X_i|\text{DirectCauses}(X_i))$ [94]. There have been a few early attempts to exploit causality in microarray data analysis [95, 96]; with recent successes in scaling up Bayesian networks to very high-dimensional data [97, 98], causal feature selection has taken place as a competitive technique for proteomic biomarker discovery.

**Key Points**

- DR is the critical step in biomarker discovery, which involves reducing tens of thousands of *m/z* points to a small set of predictive biomarkers. It has been observed that the choice of DR method is more important than that of the classification algorithm.
- The two broad approaches to DR have distinct advantages. FT has greater expressive power as its derived features represent relations between features. Feature selection is more interpretable because it retains the original mass spectral features, which are meaningful to biologists.
- Univariate feature selection methods are quick and easy to use, but slower multivariate methods can capture feature redundancy and interaction ignored by univariate methods.
- Filter approaches are quick and flexible, but they cannot directly optimize performance of the final classifier. Wrapper methods can, but they incur prohibitive costs for high-dimensional samples. Embedded approaches do what wrappers can at lower cost.
- DR method selection remains an open issue. An alternative path is DR method combination—blending different methods to exploit their synergies, overcome their individual drawbacks, and yield more stable selected feature sets.
- Future research in proteomic biomarker studies stands to gain from novel dimension reduction algorithms that exploit regularization or informative priors, or that incorporate biological knowledge through causal inference.

## *References*


1. Somorjai RL, Dolenko B, Baumgartner R. Class prediction and discovery using gene microarray and proteomics mass spectrometry data: curses, caveats, cautions. *Bioinformatics* 2003;**19**:1484–91.
2. Simon R. Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *ACM SIGKDD Explor Newslett* 2003;**5**:31–36.
3. Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell* 1997;**97**:273–324.
4. Blum A, Langley P. Selection of relevant feaures and examples in machine learning. *Artif Intell* 1997;**97**: 245–71.
5. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res* 2003;**3**:1157–82.
6. Papadopoulos M, Abel PM, Agranoff D, *et al*. A novel and accurate diagnostic test for human African trypanosomiasis. *Lancet* 2004;**363**:1358–63.
7. Lilien RH, Farid H, Donald BR. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *J Comput Biol* 2003;**10**: 925–946.
8. Qu Y, Adam BL, Thornquist M, *et al*. Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensional data. *Biometrics* 2003;**59**:143–51.

9. Duda R, Hart P, Stork D. *Pattern Classification*. Wiley, 2000; 117–24.
10. Wu B, Abbott T, Fishman D, *et al*. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 2003;**19**:1636–43.
11. Krzanowski W, Jonathan P, McCarthy W, *et al*. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Appl Stat* 1995;**44**:101–15.
12. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2001.
13. Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high dimensional genomic data. *Brief Bioinform* 2006;**8**:32–44.
14. Dai JJ, Lieu L, Rocke D. Dimension reduction for classification with gene expression microarray data. *Stat Appl Genet Mol Biol* 2006;**5**:6. http://www.bepress.com/sagmb/vol5/iss1/art6.
15. Lee KR, Lin X, Park DC, Eslava S. Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. *Proteomics* 2003;**3**:1680–6.
16. Purohit PV, Rocke DM. Discriminant models for high-throughput proteomics mass spectrometer data. *Proteomics* 2003;**3**:1699–703.
17. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**:2507–17.
18. Wagner M, Naik D, Pothen A. Protocols for disease classification from mass spectrometry data. *Proteomics* 2003;**3**: 1692–8.
19. Rogers MA, Clarke P, Noble J, *et al*. Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility. *Cancer Res* 2003;**63**:6971–83.
20. Kozak K, Amneus M, Pusey S, *et al*. Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: potential use in diagnosis and prognosis. *Proc Natl Acad Sci USA* 2003;**100**:12343–8.
21. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* 2003;**4**:24.
22. Yu J, Ongarello S, Fiedler R, *et al*. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics* 2005;**21**: 2200–9.
23. Levner I. Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics* 2005;**6**:68.
24. Bhanot G, Alexe G, Venkataraghavan B, *et al*. A robust meta-classification strategy for cancer detection from mass spectra data. *Proteomics* 2006;**6**:592–604.
25. Cover T, Thomas J. *Elements of Information Theory*. Wiley, 1991.
26. Hilario M, Kalousis A, Muller M, *et al*. Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics* 2003;**3**:1716–19.
27. Tukey JW. Tightening the clinical trial. *Controll Clin Trials* 1992;**14**:266–85.
28. Hedenfalk I, Duggan D, Chen Y, *et al*. Gene expression profiles in hereditary breast cancer. *N Engl J Med* 2001;**344**: 539–48.
29. Yanasigawa K, Shyr Y, Xu B, *et al*. Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet* 2003; **362**:433–9.
30. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 2002;**99**: 6567–72.
31. Tibshirani R, Hastie T, Narasimhan B, *et al*. Sample classification from protein mass spectrometry by ''peak probability contrasts''. *Bioinformatics* 2004;**20**:3034–44.
32. Guyon I, Elisseeff A. Introduction. In: Guyon I, Gunn S, Nikravesh M, *et al*. (eds). *Feature Extraction: Foundations, & Applications*. Springer, 2006;1–25.
33. Reunanen J. Search Strategies. Chapter 4. In: Guyon I, Gunn S, Nikravesh M, *et al*. (eds). *Feature Extraction: Foundations, & Applications*. Springer, 2006.
34. Somol P, Pudil P, Novovicová J, Paclík P. Adaptive floating search methods in feature selection. *Pattern Recogn Lett* 1999; **20**:1157–63.
35. Holland J. *Adaptation in Natural, & Artificial Systems*. MIT Press, 1992.
36. Dorigo M, Di Caro G, Gambardella L. Ant algorithms for discrete optimization. *Artif Life* 1999;**5**:137–72.
37. Kennedy J, Eberhart R. Particle swarm optimization. *Proc. IEEE Int Conf Neural Netw.* 1995;**IV**:1942–8.
38. Hall M, Holmes G. Benchmarking attribute selection techniques for discrete data class data mining. *IEEE T Knowl Dat En* 2003;**15**:1437–47.
39. Liu H, Li J, Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform* 2002;**13**: 51–60.
40. Baggerly KA, Morris JS, Wang J, *et al*. A comprehensive approach to the analysis of matrix-assisted laser desorption/ ionization time of flight proteomics spectra from serum samples. *Proteomics* 2003;**3**:1667–72.
41. Kira K, Rendell L. The feature selection problem: traditional methods and a new algorithm. *Proc Natl Conf Artif Intell (AAAI-92)* 1992;129–34.
42. Guyon I, Bitter HM, Ahmed Z, *et al*. Multivariate non-linear feature selection with kernel multiplicative updates and Gram-Schmidt Relief. In: *Proceedings of the BISC FLINT CIBI 2003 Workshop*. CA: Berkeley, 2003.
43. Sun Y, Li J. Iterative Relief for feature weighting. In: *Proceedings of the 23rd International Conference on Machine Learning*. PA: Pittsburgh, 2006;913–20.
44. Breiman L, Friedman JH, Olshen RA, *et al*. *Classification and Regression Trees*. Belmont: Wadsworth, 1984.
45. Quinlan JR. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
46. Won Y, Song HJ, Kang TW, *et al*. Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons. *Proteomics* 2003;**3**: 2310–6.
47. Bañez LL, Prasanna P, Sun L, *et al*. Diagnostic potential of serum proteomic patterns in prostate cancer. *J Urol* 2003; **170**:442–6.
48. Adam BL, Qu Y, Davis JW, *et al*. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res* 2002;**62**:3609–14.

49. Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: *Proceedings of the 13th International Conference on Machine Learning*. Bari, 1996;148–56.
50. Qu Y, Adam BL, Yasui Y, *et al*. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin Chem* 2002;**48**:1835–43.
51. Yasui Y, Pepe M, Thompson M, *et al*. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* 2003;**4**:449–63.
52. Petricoin EF, Ardekani AM, Hitt BA, *et al*. Use of proteomic patterns in serum of identify ovarian cancer. *Lancet* 2002;**359**:572–7.
53. Kohonen T. *Self-Organizing Maps*. Springer-Verlag, 1995.
54. Li L, Tang H, Wu Z, *et al*. Data mining techniques for cancer detection using serum proteomic profiling. *Artif Intell Med* 2004;**32**:71–83.
55. Li L, Umbach DM, Terry P, *et al*. Application of the GA/KNN method to SELDI proteomics data. *Bioinformatics* 2004;**20**:1638–40.
56. Ressom H, Varghese R, Sahia D, *et al*. Particle swarm optimization for analysis of mass spectral serum profiles. In: *Proceedings of the 2005 Conference on Genetic and Evolutionary Computation (GECCO-05)*. Washington, DC, 2005;431–8.
57. Ressom H, Varghese R, Sahia D, *et al*. Analysis of mass spectral serum profiles for biomarker selection. *Bioinformatics* 2005;**21**:4039–45.
58. Ressom H, Varghese R, Drake S, *et al*. Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* 2007;**23**:619–26.
59. Golub TR, Slonim DK, Tamayo P, *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;**286**:531–7.
60. Guyon I, Weston J, Barnhill S, *et al*. Gene selection for cancer classification using support vector machines. *Mach Learn* 2002;**46**:389–422.
61. Jong K, Marchiori E, Sebag M, *et al*. Feature selection in proteomic pattern data with support vector machines. In: *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. San Diego, CA, 2004;41–8.
62. Oh JH, Gao J, Nandi A, *et al*. Diagnosis of early relapse in ovariance cancer using serum proteomic profiling. *Genome Inform* 2005;**16**:195–204.
63. Zhang X, Lu X, Shi Q, *et al*. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* 2006;**7**:197.
64. Tibshirani R. Regression shrinkage and selection via the Lasso. *J. Roy Stat Soc B* 1996;**58**:267–88.
65. Soltys S, Le Q, Shi G, *et al*. The use of plasma surface-enhanced laser desorption/ionization time-of-flight mass spectrometry proteomic patterns for detection of head and neck squamous cell cancers. *Clin Cancer Res* 2004;**10**: 4806–12.
66. Bhattacharya C, Grate L, Rizki A, *et al*. Simultaneous classication and relevant feature identication in high-dimensional spaces: application to molecular profiling data. *Signal Process* 2003;**83**:729–43.
67. Krishnapuram B, Carin L, Figuereido M, *et al*. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2005;**27**:957–968.
68. Pratapa PN, Patz E, Hartemink AJ. Finding diagnostic biomarkers in proteomic spectra. *Pac Symp Biocomput* 2006; **11**:279–90.
69. Geurts P, Fillet M, de Seny D, *et al*. Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics* 2005;**21**:3138–45.
70. Ambroise C, McLachlan G. Selection bias in gene extraction on the basis of microarray gene-expression data. In: *Proceedings of the Natl Acad Sci USA* 2002;**99**:6562–6.
71. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl Inform Syst* 2007;**12**:95–116.
72. Dietterich TG. Ensemble methods in machine learning. In: Kittler J and Roli F (eds). *Multiple Classifier Systems*. Springer, 2000.
73. Tuv E. Ensemble Learning. In: Guyon I, Gunn S, Nikravesh M, *et al*. (eds). *Feature Extraction: Foundations and Applications* Springer, 2006;187–204.
74. Chan D, Bridges SM, Burgess S. An ensemble method for identifying robust features for biomarker identification. In: Liu H and Motoda H (eds). *Computational Methods of Feature Selection*. Chapman & Hall/CRC, 2007;377–392.
75. Dutkowski J, Gambin A. On consensus biomarker selection. *BMC Bioinformatics* 2007;**8**(Suppl 5):5.
76. Achlioptas D. Database-friendly random projections. In: *Proceedings of the ACM Symposium on the Principles of Database Systems*. Santa Barbara, CA, 2001;274–81.
77. Bingham E, Mannila H. Random projection in dimensionality reduction. Applications to image and text data. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2001;245–50.
78. Johnson WB, Lindenstrauss J. Extensions of Lipschitz mappings into a Hilbert space. In: *Proceedings of the Conference in Modern Analysis and Probability*. New Haven, CT, 1982;189–206.
79. Bertoni A, Valentini G. Model order selection for bio-molecular data clustering. *BMC Bioinformatics* 2007; **8**(Suppl 2):7.
80. Schölkopf B, Smola A, Müller KR. Kernel principal component analysis. In: *Advances in Kernel Methods: Support Vector Learning*. MIT Press, 1999;327–52.
81. Ng A. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: *Proceedings of the 21st International Conference on Machine Learning*. Canada: Banff, 2004;78.
82. Zhu J, Rosset S, Hastie T, *et al*. 1-norm support vector machines. In: *Advances in Neural Information Processing Systems* 2004;16.
83. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B* 2005;**67**:301–20.
84. Zou H, Hastie T. Model building and feature selection with genomic data. In: Liu and Motoda, (ed). *Computational Methods of Feature Selection*. Chapman & Hall/CRC, 2007; 393–411.
85. Liu Z, Jiang F, Tian G, *et al*. Sparse logistic regression with $L_p$ penalty for biomarker identification. *Stat Appl Genet Mol Biol* 2007;**6**:6.
86. Domeniconi C, Gunopoulos D. Adaptive nearest neighbor classification using support vector machines. In: *Advances in Neural Information Processing (NIPS 14)*MIT Press, 2002.

87. Goldberger A. Roweis S, Hinton G, *et al*. Neighbourhood component analysis. In: *Advances in Neural Information Processing (NIPS 17)*. Cambridge, MA: MIT Press, 2005.

88. Weinberger K, Blitzer J, Saul L. Distance metric learning for large-margin nearest neighbor classification. In: *Advances in Neural Information Processing (NIPS 18)*MIT Press, 2006.

89. Xing E, Ng AY, Jordan MI, *et al*. Distance metric learning with application to clustering with side-information. In: *Advances in Neural Information Processing (NIPS 15)*MIT Press, 2003;505–13.

90. Bar Hillel A, Hertz T, Shental N, *et al*. Learning distance functions using equivalence relations. In: *Proceedings of the International Conference on Machine Learning*. Washington, DC, 2003;11–18.

91. Bishop C. *Pattern Recognition and Machine Learning* Springer, 2006.

92. Tipping ME. Sparse Bayesian learning and the relevance vector machine. *J Mach Learn Res* 2001;**1**:211–44.

93. Sima C, Dougherty ER. What should be expected from feature selection in small-sample settings. *Bioinformatics* 2006;**22**:2430–6.

94. Guyon I, Aliferis C, Elisseeff A. Causal feature selection. In: Liu H and Motoda H (eds). *Computational Methods of Feature Selection*. Chapman & Hall/CRC, 2007;63–85.

95. Xing E, Jordan MI, Karp R. Feature selection for high-dimensional genomic microarray data. In: *Proceedings of the 18th International Conference on Machine Learning*. San Francisco: CA, 2001;601–8.

96. Friedman N, Linial M, Nachman I, *et al*. Using Bayesian networks to analyse expression data. In: *Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB)*. Tokyo, 2000; 127–35.

97. Tsamardinos I, Aliferis CF, Statnikov A. Algorithms for large-scale Markov blanket discovery. In: *Proceedings of the 16th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*. St Augustine: Florida, 2003; 376–80.

98. Aliferis CF, Tsamardinos I, Statnikov A. HITON, a novel Markov blanket algorithm for optimal variable selection. *American Medical Informatics Association (AMIA)* 2003;21–5.