Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Computational Advances in Cancer Informatics (A)

Overcome Support Vector Machine Diagnosis Overfitting

Henry Han[1,2] and Xiaoqian Jiang[3]

[1]Department of Computer and Information Science, Fordham University, New York, NY, USA. [2]Quantitative Proteomics Center, Columbia University, New York, NY, USA. [3]Division of Biomedical Informatics, University of California, San Diego, CA, USA.

ABSTRACT: Support vector machines (SVMs) are widely employed in molecular diagnosis of disease for their efficiency and robustness. However, there is no previous research to analyze their overfitting in high-dimensional omics data based disease diagnosis, which is essential to avoid deceptive diagnostic results and enhance clinical decision making. In this work, we comprehensively investigate this problem from both theoretical and practical standpoints to unveil the special characteristics of SVM overfitting. We found that disease diagnosis under an SVM classifier would inevitably encounter overfitting under a Gaussian kernel because of the large data variations generated from high-throughput profiling technologies. Furthermore, we propose a novel sparse-coding kernel approach to overcome SVM overfitting in disease diagnosis. Unlike traditional ad-hoc parametric tuning approaches, it not only robustly conquers the overfitting problem, but also achieves good diagnostic accuracy. To our knowledge, it is the first rigorous method proposed to overcome SVM overfitting. Finally, we propose a novel biomarker discovery algorithm: Gene-Switch-Marker (GSM) to capture meaningful biomarkers by taking advantage of SVM overfitting on single genes.

KEYWORDS: SVM, overfitting, biomarker discovery

SUPPLEMENT: Computational Advances in Cancer Informatics (A)

CITATION: Han and Jiang. Overcoming Support Vector Machine Overfitting in Diagnosis. *Cancer Informatics* 2014:13(S1) 145–158 doi: 10.4137/CIN.S13875.

RECEIVED: May 4, 2014. **RESUBMITTED:** September 13, 2014. **ACCEPTED FOR PUBLICATION:** September 16, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Review

FUNDING: XJ is partially supported by iDASH (grant no: U54HL108460), NIH R00 (grant no: R00 LM011392). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: xhan9@fordham.edu, x1jiang@ucsd.edu

Introduction

With the surge of bioinformatics, complex disease diagnosis and prognosis rely more and more on biomedical insights discovered from its molecular signatures.[1,2] However, a key challenge is to detect molecular signatures of disease from high-dimensional omics data, which are usually characterized with a large number of variables and a small number of observations, in an accurate and reproducible manner. Various classification algorithms have been proposed or adopted in molecular diagnosis of disease in recent studies for this purpose. These algorithms include logistic regression, ensemble learning methods, Bayesian methods, neural networks, and kernel learning methods such as support vector machines (SVMs).[3–8]

As a state-of-the-art machine learning algorithm, the standard SVM probably is one of the mostly employed methods in molecular diagnosis of disease for its good scalability for high-dimensional data.[6,9] However, omics data's special characteristic: ie, small number of samples and large number of variables, theoretically will increase the likelihood of an SVM classifier's overfitting in classification and lead to deceptive diagnostic results.[9–11] Overfitting means that a learning machine (classifier) loses its learning generalization capability and produces deceptive diagnostic results. It may achieve some good diagnostic results on some training data, but it has no way to generalize the good diagnostic ability to new test data. In other words, diagnostic results are only limited to few specific data instead of general data. In fact, it can even be trapped in one or a few diagnostic patterns for input data and produce totally wrong diagnostic results, because of inappropriate parameter setting (eg, kernel choice) in classification.

In fact, SVM overfitting on omics data is of significance in bioinformatics research and clinical applications when considering the popularity of SVM in molecular diagnosis of disease. On the other hand, it is also essential for kernel-based learning theory itself to develop new knowledge and technologies for omics data. However, there is almost no previous research available on this important topic. In fact, it remains unknown about the following important questions, ie, "why does overfitting happen, and how to conquer it effectively?" As such, a serious investigation on SVM overfitting is definitely of priority for the sake of robust disease diagnosis. In this work, we have investigated SVM overfitting on molecular diagnosis of disease using benchmark omics data and presented the following novel findings.

First, contrary to the general assumption that a nonlinear decision boundary is more effective in SVM classification than a linear one, we have found that SVM encounters overfitting on nonlinear kernels through rigorous kernel analysis. In particular, it demonstrates a major-phenotype favor diagnostic mechanism on Gaussian kernels under different model selections. That is, an SVM classifier can only recognize those samples with majority counts in training data. When the training data have an equal number of phenotypes, the SVM classifier will produce all false diagnostic results under a leave-one-out cross validation (LOOCV) because of the major-phenotype favor diagnostic mechanism.

Second, we have demonstrated that an SVM classifier under a linear kernel shows some advantages in diagnosis over the nonlinear kernels, and it has less likelihood of encountering overfitting also. Moreover, we have found that large pairwise distances between training samples, which are actually caused by molecular signal amplification mechanism in omics profiling systems, are responsible for the SVM overfitting on Gaussian kernels. We further have illustrated that general feature selection algorithms actually cannot overcome overfitting and contribute to improving diagnostic accuracy effectively.

Third, we have proposed a novel sparse-coding kernel approach to conquer SVM overfitting by imposing sparseness on training data, by seeking each training sample's nearest non-negative sparse approximation in L_1 and L_2 norms, before a kernel evaluation. Unlike traditional ad-hoc parametric tuning approaches, it not only robustly conquers overfitting in SVM diagnosis, but also achieves better diagnostic results in comparison with other kernels.

On the other hand, we demonstrate that sparse coding would be an effective way to optimize the kernel matrix structure to enhance a classifier's learning capability. To the best of our knowledge, it is the first rigorous method to overcome SVM overfitting and would inspire other following methods in data mining and bioinformatics fields. Finally, we have proposed a novel biomarker discovery by taking advantage of the special "gene switch" mechanism demonstrated by SVM overfitting on single genes to seek meaningful biomarkers.

This paper is structured as follows. Section 2 presents SVM disease diagnostics and benchmark datasets used in our overfitting analysis. Section 3 presents SVM overfitting results and rigorous kernel analysis results in disease diagnostics. Section 4 proposes our sparse-coding kernel approach to conquer SVM overfitting under the Gaussian kernel. Section 5 proposes our Gene-Switch-Marker (GSM) algorithm by taking advantage of SVM overfitting on single genes. Finally, we discuss the ongoing and future related work and conclude our paper in Section 6.

Support Vector Machine Diagnosis

Support Vector Machine diagnosis starts a set of samples drawn from omics data with known class labels, usually control vs disease, to build a linear decision function to determine an unknown sample's type by constructing an optimal separating hyperplane geometrically. Given training omics data $\{(x_i, y_i)\}_{i=1}^{m}, x_i \in R^n$ is a sample with n features, a feature refers to a gene, peptide, or protein in our context, and its label $y_i \in \{-1, +1\}$, where $y_i = -1$ if x_i is an observation from a control (negative) class; otherwise, $y_i = +1$ if it is from a disease (positive) class.

An SVM classifier computes an optimal separating hyperplane: $(w \cdot x) + b = 0$ in R^n, to attain the maximum margin between the two types of training samples. The hyperplane normal w and offset vector b are solutions of the following optimization problem, provided we assume the training data are linearly separable,

$$\begin{gathered} \min J(w,b) = \frac{1}{2}\|w\|^2, s.t. \\ y_i(w \cdot x_i + b) - 1 > 0,\ i = 1, 2 \ldots m \end{gathered} \tag{1}$$

It is equivalent to solving the following optimization problem:

$$\max L_d(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{2}$$

where $\alpha_i \geq 0$, $i = 1, 2, \ldots m$, the two parameters of the optimal separating hyperplane w and b can be calculated by solving equation (1)'s stationary conditions $w = \sum_{i=1}^{m} \alpha_i y_i x_i$ and KKT condition: $y_i(w \cdot x_i + b) - 1 = 0$. Finally, the decision function determining the class type of an unknown sample x' is formulated as, $f(x') = sgn((w \cdot x') + b)) = sgn((\sum_{i=1}^{m} \alpha_i (x_i \cdot x') + b)$. That is, x' will be diagnosed as a disease sample if $w \cdot x' + b > 0$ and a control sample otherwise. The training samples x_i corresponding to $\alpha_i > 0$ are called support vectors, and SVM disease diagnosis is totally dependent on the support vectors.

The standard SVM algorithm can be further generalized to handle the corresponding nonlinear problems by mapping training samples into a higher or infinite-dimensional feature space F using a mapping function ϕ: $X \to F$, and constructing an optimal nonlinear decision boundary in F to achieve more separation capabilities. Correspondingly, the decision function for an unknown sample x' is updated as $f(x') = \mathrm{sgn}((\sum_{i=1}^{m} \alpha_i(\phi(x_i)\cdot\phi(x')) + b)$. Since the inner product $(\phi(x_i)\cdot\phi(x_j))$ in F can be evaluated by any kernel $(\phi(x_i)\cdot\phi(x_j)) = k(x_i, x_j)$ implicitly in the input space R^n, provided its kernel matrix is positive definite, the decision function can be evaluated in the input space as $f(x') = sgn((\sum_{i=1}^{m} \alpha_i k(x', x_i) + b)$. In addition to quadratic $k(x, x') = (1 + (x_i \cdot x'))^2$, polynomial: $k(x, x') = (1 + (x_i \cdot x'))^3$, and multilayer perceptron: $k(x, x') = \tanh((x_i \cdot x') - 1)$ kernels, we mainly employ a Gaussian radial basis function ("*rbf*") $k(x, x') = \exp(||x - x'||^2/2\sigma^2)$, and a linear kernel $k(x, x') = (x_i \cdot x')$ in our experiments.

Non-separable cases. If the training data are not separable, an SVM classifier can separate many but not all samples by using a soft margin that permits misclassification.[6,10] Mathematically, it is equivalent to adding slack variables ξ_i and a penalty parameter C to equation (1) under L_1 or L_2 norms. For example, the corresponding L_1 norm problem minimizes $\frac{1}{2}||w||^2 + c\sum_{i=1}^{m}\xi_i$ under the conditions $y_i(w \cdot x_i + b) - 1 > \xi_i$ and $\xi_i \geq 0$, where the penalty parameters C imposes weights on the slack variables to achieve a strict separation between two types of samples.

Model selection. There are quite a few model selection methods for SVM diagnosis to minimize the expectation of diagnostic errors.[6] In this work, we mainly employ different cross-validation methods for model selection that include LOOCV and k-fold cross validation (k-fold CV), because they are widely employed in disease diagnostics. The LOOCV removes one sample from the training data and constructs the decision function to infer the class type for the removed one. The k-fold CV randomly partitions the training data to form k disjoint subsets with approximately equal size, removes the i^{th} subset from the training data, and employs the remaining $k - 1$ subsets to construct the decision function to infer the class types of the samples in the removed subset.

Benchmark omics data. Table 1 includes benchmark genomics and proteomics data used in our experiment.[12–17] We have three criteria to choose benchmark data. First, they are generated from different omics profiling technologies for well-known cancer disease studies. For example, The *Ovarian-qaqc* data are generated from surface enhanced laser desorption and ionization time-of-flight (SELDI-TOF) profiling technologies and the *Cirrhosis* and *Colorectal* are mass spectral from matrix-assisted laser desorption time-of-flight (MALDI-TOF) technologies.[14–17] Second, they contain some widely used omics data in the literature. For example, the *Medulloblastoma* and *Breast* datasets are widely used gene expression data in secondary data analysis.[12,18] Third, the omics data should be processed by different normalization/preprocessing methods.

It is noted that these omics data are not raw data. Instead, they are normalized data by using different normalization and preprocessing methods.[9,14] For example, robust multiarray average (RMA)[43] method is employed to normalize the *Stroma* data, which is quite different from the normalization methods employed in the *Medulloblastoma* and *Breast* data.[12,18] Moreover, the *Ovarian-qaqc* and *Cirrhosis* data have been preprocessed by using "lowess" and "least-square polynomial" smoothing methods, respectively, in addition to the same baseline correction, AUC normalization, and peak alignment processing.[9] We have not conducted preprocessing for the *Colorectal* data, and the details about its preprocessing can be found in the work of Alexandrov et al.[16]

Analyze SVM Overfitting in Disease Diagnosis

Given an omics dataset with binary labels $\{(x_i, y_i)\}_{i=1}^{m}$, $y_i \in \{-1, +1\}$, $x_i \in R^n$, we define following measures for the convenience of SVM overfitting analysis.

1. The majority (minority) type in omics data is the class type with more (less) counts among all samples. Let $m^- = |\{y_i|y_i| = -1\}|$, $m^+ = |\{y_i|y_i| = +1\}|$, we define the majority/minority type ratios as $\gamma_{\max} = \max(m^-, m^+)/m$ and $\gamma_{\min} = \min(m^-, m^+)/m$, respectively.
2. The pairwise distance between two omics samples x_i and x_j is an Euclidean distance defined as: $d_{ij} = ||x_i - x_j|| = (\sum_{k=1}^{m}(x_{ki} - x_{kj})^2)^{1/2}$, and data total variation is defined as $\rho = \sum_{i=1}^{m}\sum_{j=1}^{m}||x_i - x_j||^2$.
3. The absolute difference between x_i and x_j at the k^{th} feature is a Manhattan distance defined as: $\delta_{ijk} = |x_{ki} - x_{kj}|$, $k = 1, 2\cdots n$. Correspondingly, the maximum absolute difference (MAD) between x_i and x_j is defined as $\delta_{ij} = \max_k \Delta_{ijk}$, which measures the maximum expression difference across all features for two samples.

There is a strong need to examine omics data and find their latent characteristics for the sake of better understanding

Table 1. Benchmark omics data.

DATA	#FEATURE	THE NUMBER OF SAMPLES	PLATFORM
Stroma	18,995	13 *inflammatory breast cancer* + 34 *non-inflammatory breast cancer*	oligonucleotide
Medulloblastoma	5,893	25 *classic* + 9 *desmoplastic*	oligonucleotide
Breast	24,188	46 *patients with* 5-*year metastasis* + 51 *without* 5-*year metastasis*	oligonucleotide
Ovarian-qaqc	15,000	95 *controls* + 121 *cancer*	SELDI-TOF
Cirrhosis	23,846	78 *HCC* + 51 *cirrhosis*	MALDI-TOF
Colorectal	16,331	48 *controls* + 64 *cancer*	MALDI-TOF

SVM overfitting. As such, we have checked the ratio between pairwise sample distance and MAD for each omics dataset, which answers the following query: "compared with the pairwise distance between omics samples x_i and x_j, how large will their MAD be?" Interestingly, we have found that the ratio for any omics data is always between 1 and $\sqrt{n}$, where n is the total feature number for the omics data. Such a ratio indicates that the pairwise distance can be much larger than corresponding MAD because of the large number of variables for a given omics data. The following theorem states the result about the ratio estimation.

The sample distance and MAD ratio theorem. *Given an omics data with m observations across n features, ie, $X \in R^{n\times m}$, the ratio between the distance of samples x_i and x_j and their MAD satisfies the following inequality when $i \neq j$,*

$$1 \leq \frac{d_{ij}}{\delta_{ij}} \leq \sqrt{n} \tag{3}$$

Proof. Suppose MAD is achieved at the k_*^{th} feature (eg, gene) ie, $\delta_{ij} = |x_{k_*i} - x_{k_*j}|$, it is clear that we have $\delta_{ij} \leq (|x_{k_*i} - x_{k_*j}|^2)^{1/2} + \sum_{k\neq k_*}(x_{ki} - x_{kj})^2)^{1/2} = d_{ij}$. On the other hand, $d_{ij}^2 = \sum_{k=1}^{n}(x_{ki} - x_{kj})^2 \leq n|x_{k_*i} - x_{k_*j}|^2 = n\delta_{ij}^2$. Combining the two previous equations, we have $1 \leq \frac{d_{ij}}{\delta_{ij}} \leq \sqrt{n}$.

The distance between training samples x_i and x_j in the feature space of a *rbf*-SVM, which is an SVM classifier under the '*rbf*' kernel, is the entry K_{ij} in the learning machine's kernel matrix K, which can be calculated by plugging x_i and x_j into the '*rbf*' kernel $k(x, y) = e^{-||x-y||^2/2\sigma^2}$, ie, $k_{ij} = \exp(-d_{ij}^2/2\sigma^2)$. We have the following estimation about K_{ij} by using Theorem 1 result.

Corollary 1. *Given an omics data with m observations across n features, ie, $X \in R^{n\times m}$, then each entry K_{ij} in the SVM kernel matrix K under the Gaussian 'rbf' kernel: $k(x, y) = e^{-||x-y||^2/2\sigma^2}$ satisfies* $\exp(-n\delta_{ij}^2/2\sigma^2) \leq K_{ij} \leq \exp(-d_{\min}^2/2\sigma^2)$.

According to $\delta_{ij} \leq d_{ij} \leq \sqrt{n}\delta_{ij}$ and $\delta_{ij}^2 \leq \min_{i\neq j} d_{ij}^2 \leq d_{ij}^2$, it is easy to have $d_{\min}^2 \leq d_{ij}^2 \leq n\delta_{ij}^2$. Substituting it into $K_{ij} = \exp(-d_{ij}^2/2\sigma^2)$, we have $\exp(-n\delta_{ij}^2/2\sigma^2) \leq K_{ij} \leq \exp(-d_{\min}^2/2\sigma^2)$.

It is clear that the upper bound of K_{ij} is determined by d_{min}, the minimum pairwise distance among all training samples, and the bandwidth parameter σ, according to Corollary 1. Considering the popularity of setting $\sigma = 1$ by default in most *rbf*-SVM diagnosis, we treat this case as an important '*rbf*' kernel scenario in our overfitting analysis.

Identity or isometric identity kernel matrices. Although choosing the '*rbf*' kernel with bandwidth $\sigma = 1$ is generally recommended in the literature,[19,20] we have found that the *rbf*-SVM classifier would inevitably encounter overfitting, because of an identity or isometric identity kernel matrix. This is because the pairwise sample distances are quite large, which will cause their distances in the feature space of the '*rbf*' kernel to be zero or approximately zero, that is, $K_{ij} = exp(\ d_{ij}^2/2) \sim 0$ for $i \neq j$.

Figure 1 shows the minimum ($d_{\min}^2$), first percentile ($d_{0.01}^2$), median (d_{median}^2), and maximum ($d_{\max}^2$) values of the pairwise distance squares ($d_{ij}^2, i \neq j$) for all samples in each omics data. It is interesting to see that $\log_{10} d_{0.01}^2 > 2$, for all data, which indicates that the upper bound of K_{ij} will be zero or approximately zero for $i \neq j$, because of the fact that $K_{ij} \leq \exp(-10^2/2) = 9.287 \times 10^{-22}$.

Thus, the SVM kernel matrices of the omics data under the '*rbf*' kernel with the bandwidth $\sigma = 1$ are identity or isometric identity. The zero or approximately zero pairwise sample distances in the classifier *rbf*-SVM's feature space actually force the classifier to lose the diagnostic capability to distinguish

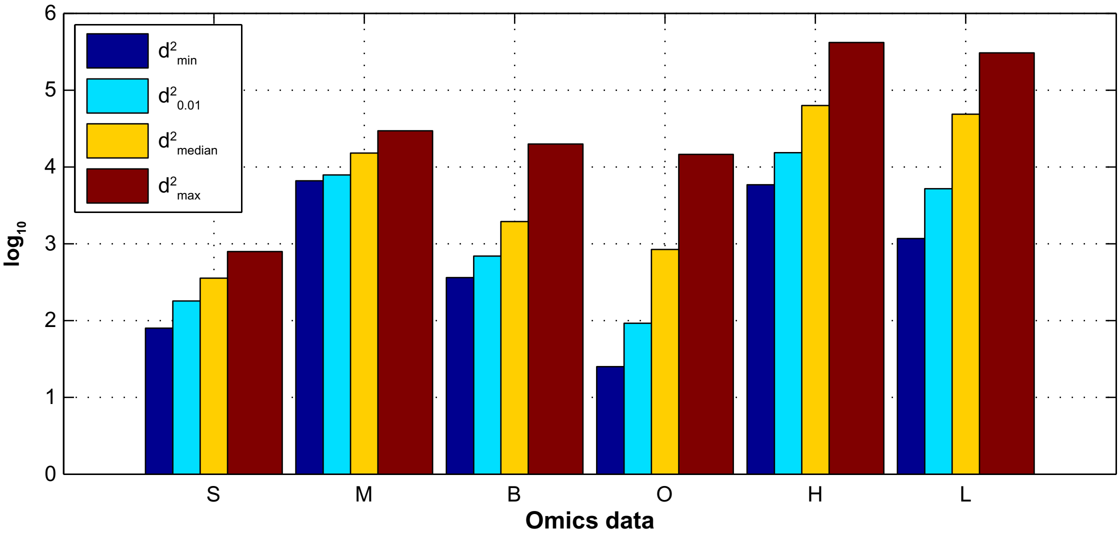

Figure 1. The minimum, first percentile, median, and maximum of d_{ij}^2 across different omics data. The minimum pair-sample distances are approximately 10^2. Each dataset is represented by its first letter except the *Colorectal* dataset, which is represented by "L".

any test samples, not to mention its generalization. That is, the *rbf*-SVM classifier loses diagnostic capability because it encounters overfitting for input omics data.

Major-phenotype favor diagnosis. According to that the *rbf*-SVM's kernel matrix is the identity or isometric identity matrix, the following SVM overfitting theorem demonstrates that overfitting will lead to a major-phenotype favor diagnosis, in which the *rbf*-SVM classifier always diagnoses an unknown omics sample as the type of sample with the majority count in the training data. If there is no major type in the training data, the classifier will fail to conduct any diagnosis.

SVM overfitting theorem. *Given an omics training data-set with binary labels* $\{(x_i, y_i)\}_{i=1}^{m}, y_i \in \{-1,+1\}, x_i \in R^n$, *let* $m^- = |\{y_i|y_i| = -1\}|$, $m^+ = |\{y_i|y_i| = +1\}|$, *an SVM classifier under the Gaussian 'rbf' kernel (*$\sigma = 1$*) always predicts x' as the majority type in the training data. That is, it has the following decision rule about a test omics sample* $x' \in R^n$.

$$f(x') = sgn(m^+ - m^-) \tag{4}$$

Proof. Let $f(x') = sgn((\sum_{i=1}^{m} \alpha_i k(x', x_i) + b)$ be the decision function for an unknown type sample *x*', it is clear that it will be totally dependent on the offset vector *b* because of $k(x' \cdot x_i) = \exp(-||x' - x_i||^2/2) \sim 0$, according to our previous results. In fact, the offset term *b* is determined by the weight vector $w = \sum_{i=1}^{m} \alpha_i y_i \phi(x_i)$, ie, $b = -\frac{1}{2}(w^T\phi(x_p) + w^T\phi(x_n))$, where x_p and x_n are two support vectors with positive and negative labels, respectively, namely,

$$b = -\frac{1}{2}\sum_{i=1}^{m}(\alpha_i y_i k(x_i \cdot x_p) + \alpha_i y_i k(x_i \cdot x_n)) \tag{5}$$

Since $k(x_i \cdot x_p) \sim 0$ and $k(x_i \cdot x_n) \sim 0$ when $i \neq p$ and $i \neq n$, we have $b = -\frac{1}{2}(\alpha_p - \alpha_n)$, where α_p and α_n are corresponding alpha values. In fact, all α values can be solved by the following equation with conditions $\sum_{i-1}^{m} \alpha_i y_i = 0$, and $0 \leq \alpha_i < C$, $i = 1, 2, \cdots m$,

$$\max L_d(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j k(x_i \cdot x_j) \tag{6}$$

The equation is further reduced as

$$\max L_d(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\alpha_i^2 \tag{7}$$

under the same conditions because of $k(x_i \cdot x_i) = 1$ and $k(x_i \cdot x_j) = 0$ for $i \neq j$.

It is easy to have $\alpha_1 = \alpha_2 = \cdots \alpha_{m^-} = \frac{m^+}{m}$ and $\alpha_1 = \alpha_2 = \cdots \alpha_{m_+} = \frac{m^-}{m}$. That is, there are two different alpha values for positive and negative samples: $\alpha_p = \frac{m^-}{m}$ and $\alpha_n = \frac{m^+}{m}$. As such,

$$b = -\frac{1}{2}(\alpha_p - \alpha_n) = \frac{1}{2}(\alpha_n - \alpha_p) = \frac{m^+ - m^-}{2m} \tag{8}$$

Thus, the decision function $f(x') = sgn(b)$ for an unknown sample *x*' is reduced as $f(x') = sgn(\frac{m^+ - m^-}{2m})$. According to the sign function's definition, the decision function is further simplified as $f(x') = sgn(m^+ - m^-)$. Obviously, the class type of the sample will be totally determined by the majority type in training data. If there is no majority type, ie, $m^- = m^+$, the learning machine cannot determine the class type of the input sample, ie, $f(x') = 0$. In this scenario, the SVM classifier cannot determine the omics sample type any more.

Diagnostic measures. It is obvious that the identity or isometric identity kernel matrix under the '*rbf*' kernel causes the SVM classifier to lose diagnostic capabilities by only diagnosing a test sample as the majority type in the training data. Before presenting further SVM diagnosis overfitting results under LOOCV and *k*-fold CV model selection, we introduce important diagnostic measures: diagnostic accuracy, sensitivity, and specificity as follows.

Given a classifier, diagnostic accuracy is ratio $r_c = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$, where TP(TN) is the number of positive (negative) samples correctly diagnosed, and FP (FN) is the number of negative (positive) samples incorrectly diagnosed. The sensitivity and specificity are defined as $r_{\text{sen}} = \frac{\text{TP}}{\text{TP+FN}}$ and $r_{\text{spe}} = \frac{\text{TN}}{\text{TN+FP}}$, respectively.

Overfitting under model selection. It is noted that the SVM overfitting under the Gaussian '*rbf*' kernel ($\sigma = 1$) would always happen in diagnosis under LOOCV and *k*-fold CV according to the SVM overfitting theorem. An extreme case will happen under LOOCV when input data have the same number of positive and negative samples. The following balanced data overfitting theorem states the extreme case in detail.

Balanced data overfitting theorem. Given a balanced omics dataset with binary labels $\{(x_i, y_i\})_{i=1}^{m}$, $y_i \in \{-1,+1\}, x_i \in R^n$, where $m^- = |\{y_i|y_i| = -1\}| = m/2$, $m^+ = |\{y_i|y_i| =. +1\}| = m/2$, an SVM classifier with the Gaussian '*rbf*' kernel ($\sigma = 1$) under LOOCV has a zero diagnostic accuracy.

Proof. It is clear that there are totally *m* trials of diagnoses in LOOCV, each of which has a training dataset consisting of *m* – 1 samples and a test dataset has only one sample.

Suppose the test sample *x*' in *i*th trial ($1 \leq i \leq m$) is positive ("+1") and the majority type is "–1" for the training data, in which $m^- = m/2$ and $\text{m}^+ = (m - 1)/2$. According to the SVM overfitting theorem, the decision function outputs "–1": $f(x') = sgn(m^+ - m^-) = -1$, ie, it is misdiagnosed as a negative sample. Similarly, the test sample will be diagnosed as positive if it is a negative sample. Finally, all test samples will be misdiagnosed as its opposite types and the classifier has a zero diagnostic accuracy.

A real extreme case example. To further demonstrate this balanced data overfitting theorem, we include a *CNS* (central nervous system) dataset with 10 *Medulloblastoma* and 10 *Malignant glioma* samples, which are labeled "–1" and "+1", respectively, in our experiment, and employ the *rbf*-SVM classifier ($\sigma = 1$) to conduct diagnosis for this data under LOOCV.

We have found that all samples in this data are misdiagnosed as their "opposite" types each time consistently. We have the following interesting results.

First, the bias term b in the decision function $f(x') = sgn(b)$ only takes two values: $b = 0.0526$ (1/19) in the first 10 trials and $b = -0.0526$ (−1/19) in the last 10 trials, respectively. They indicate that each test sample of the first and last 10 trials is diagnosed as "+1" and "−1", respectively, though they are actually labeled as "−1" and "+1" correspondingly.

Second, all the training samples are support vectors in each trial instead of a few of them as we usually expect for an SVM classification. That is, there are 10 positive samples and 9 negative samples in the training data, we have corresponding α values solved from $\alpha_p = \alpha_1 = \alpha_2 = \cdots \alpha_{10} = \frac{9}{19} = 0.4737$, $\alpha_n = \alpha_{11} = \alpha_{12} = \cdots \alpha_{19} = \frac{10}{19} = 0.5263$. Thus, $b = 0.0526 = \frac{10}{19} - \frac{9}{19}$ and the decision function for each test sample is $f(x') = sgn\,(10 - 9) = +1$, that is all the negative samples are misdiagnosed as positive samples because the majority type is "+1" in the training data. Similarly $b = -0.0526 = \frac{9}{19} - \frac{10}{19}$, $f(x') = sgn(b) = sgn\,(9 - 10) = -1$; where $\alpha_n = \frac{9}{19}$ and $\alpha_p = \frac{10}{19}$, and all the positive samples are misdiagnosed as negative samples in the last 10 trials.

Figure 2 shows the values of b and the α values in the first 10 trials and last 10 trials from left to right, respectively. Furthermore, we have the following more general results about SVM diagnosis under model selections. We skip their proof for the convenience of concise description.

SVM overfitting under model selection theorem. Given an omics dataset with binary labels $\{(x_i, y_i)\}_{i=1}^{m}$, $y_i \in \{-1, +1\}$, $x_i \in R^n$, let $m^- = |\{y_i|y_i| = -1\}|$, $m^+ = |\{y_i|y_i| = +1\}|$, an SVM classifier with the Gaussian '*rbf*' kernel ($\sigma = 1$) under LOOCV and k-fold CV has the following diagnostic results.

1. The expected diagnostic accuracy $E(r_c)$ will be exactly the majority type ratio $\gamma_{\max}$ in the input data, where the expected sensitivity $E(r_{\text{sen}})$ and specificity $E(r_{\text{spe}})$ are 100% and 0%, respectively, or vice versa under LOOCV.
2. The expected diagnostic accuracy $E(r_c)$ will approximate or equal the majority type ratio $\gamma_{\max}$ in the input data, where the expected sensitivity $E(r_{\text{sen}})$ and specificity $E(r_{\text{spe}})$ are approximately 100% and 0%, respectively, or vice versa under k-fold CV.

SVM overfitting under 50% holdout cross-validation (HOCV). Although we only focus on k-fold CV and LOOCV in our model selection, it does not mean that our overfitting results would not apply to other model selection cases. Here we illustrate overfitting under a new model selection method: 50% holdout cross-validation (HOCV), where 1,000 trials of training and test data are randomly generated for each dataset. The final diagnostic performance is evaluated by using the expectation of the three statistics in the 1,000 trials of diagnoses.

Table 2 shows the expected diagnostic accuracies $E(r_c)$, sensitivities $E(r_{\text{sen}})$, specificities $E(r_{\text{spe}})$, and their standard deviations: std(r_c), std(r_{sen}), and std(r_{spe}) of an SVM learning machine with a standard Gaussian kernel ('*rbf*') for *Medulloblastoma and Ovarian-qaqc* data. As a representative among our six omics datasets, the former is a gene expression dataset with 34 samples across 5,893 genes and the latter is a proteomics dataset with 216 samples across 15,000 *m/z* ratios.

Because, for each trial, the SVM classifier can only recognize the majority type and the data partition is based on 50% HOCV, the expected diagnostic accuracy $E(r_c)$ will approximate the majority type ratio $\gamma_{\max}$, in addition to the fact that $E(r_{\text{sen}})$ and $E(r_{\text{spe}})$ will be complementary to each other in diagnosis. For example, $E(r_c) = 0.72988$ and $E(r_c) = 0.550880$ approximates the majority type ratio: 25/34 (0.7353) of the *Medulloblastoma* data and the majority type ratio: 121/216 (0.5602) of the *Ovarian-qaqc* data, respectively. Clearly, the overfitting can be detected by the complementary of average sensitivities and specificities easily. It is noted that similar results can be observed for other datasets also.

The biological root of SVM overfitting. The mathematical reason for the SVM overfitting under the Gaussian '*rbf*' kernel ($\sigma = 1$) lies in the fact that the pairwise distances between omics samples are large or even huge. The '*rbf*' kernel $k(x, y) = \exp(-||x - y||^2/2)$ maps it to zero or a tiny value approximate to zero in the feature space. Finally, a corresponding

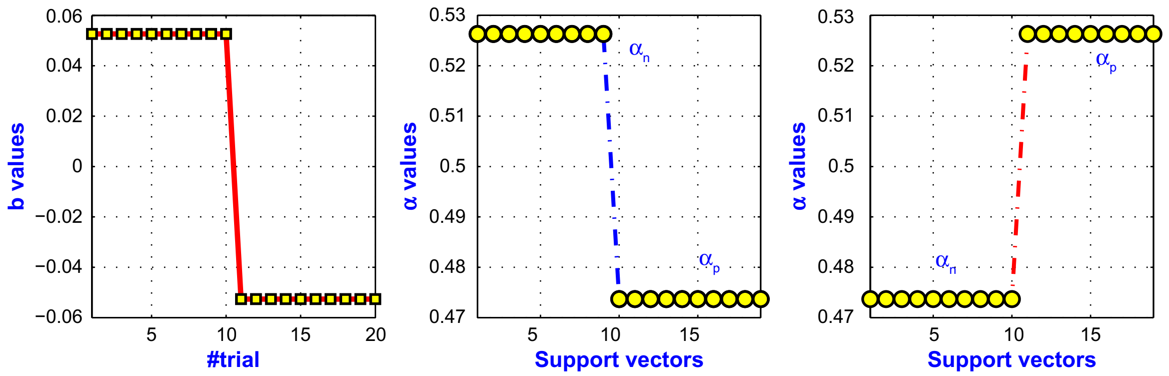

Figure 2. The offset b values (1/19 and −1/19) in all the 20 trials of SVM diagnoses and corresponding alpha values (α_p and α_n) in the first and last 10 trials. All test samples are misdiagnosed as their "opposite types" in each trial, where all training samples are support vectors and diagnostic results only rely on the majority type in the training data.

Table 2. SVM diagnostics with an "*rbf*" kernel (1,000 trials of 50% HOCV).

DATASET	$E(r_c) \pm std(r_c)$	$E(r_{sen}) \pm std(r_{sen})$	$E(r_{spe}) \pm std(r_{spe})$
Medulloblastoma	0.729882 ± 0.084460	0.002000 ± 0.044699	0.998000 ± 0.044699
Ovarian-qaqc	0.550880 ± 0.046736	0.036000 ± 0.186383	0.964000 ± 0.186383

identity or isometric identity kernel matrix will be generated, which forces the linear machine to lose its diagnostic capability by demonstrating the major-phenotype favor mechanism in such a situation.

In fact, the large or even huge pairwise sample distances in each omics dataset are actually rooted in the molecular signal amplification mechanism in omics profiling, where different techniques are employed in various profiling platforms to amplify expression signals for the sake of phenotype or genotype identification at a molecular level.[21,22] For example, like RNA-Seq, gene expression profiling technologies usually employ quantitative real-time PCR or similar approaches to amplify the expression signals over each probe to increase the sensitivity in the phenotype or genotype identification.[23,24] The PCR amplification makes it possible to distinguish disease signatures at a molecular level, but it directly contributes to increasing the pairwise distances between two samples also. Similarly, there is the amplification of the ionized molecule signals in mass spectral proteomic profiling to get high-resolution protein expression values at a molecular level.[25] As such, the SVM overfitting under the Gaussian '*rbf*' kernel ($\sigma = 1$) would be inevitable to some degree if no special action is taken to overcome it.

Can feature selection avoid such overfitting?. A traditional misconception believes that such an SVM overfitting in disease diagnosis could be avoided by conducting feature selection, because it would produce a "more meaningful" low-dimensional omics dataset than the original one. However, we have found that the low-dimensional dataset is actually unable to avoid the SVM overfitting. This is mainly because the pairwise sample distances after feature selection are still quite large or even huge, which causes the corresponding pairwise distances in the *rbf*-kernel space to be zero or approximately zero.

To demonstrate this, we employ Bayesian *t*-test to simulate a generic feature selection algorithm to obtain low-dimensional datasets for each omics data before a *rbf*-SVM classifier diagnosis ($\sigma = 1$).[26] In fact, we select the top 100, 200, 500, 1,000, and 2,000 differentially expressed features (genes/peptides) ranked by the Bayesian *t*-test for each omics data. Then, we conduct the *rbf*-SVM classifier diagnosis for each low-dimensional dataset under the LOOCV and five-fold CV. Interestingly, we have found that they all encounter overfitting and the SVM classifier diagnoses all test samples as the majority type of the training data.

Table 3 illustrates the SVM diagnosis results obtained by using the top 200 features selected by Bayesian *t*-test for each data under the five-fold CV. It is clear that the SVM classifier demonstrates overfitting by achieving a deceptive diagnostic accuracy for each data, because the accuracy is actually the majority type ratio of the original data. For example, the average diagnostic accuracies for *Stroma* and *Colorectal* data are 72.56% and 57.15%, respectively, which reach or approximate their corresponding majority type ratios $0.7234 = \frac{34}{47}$ and $0.5714 = \frac{64}{112}$.

In fact, we have found that even single genes encounter overfitting under LOOCV, that is, an SVM classifier can only recognize the majority type in diagnosis when input data consist of only a single gene. For example, 17 genes among 200 top-ranked genes by Bayesian *t*-test from the *Stroma* data encounter overfitting under the *rbf*-SVM classifier under LOOCV. The single gene overfitting case actually demonstrates that general feature selection may not avoid such an overfitting because they may not be able to decrease the built-in large pairwise sample distance though they can lower the input data dimensionality.

Conquer SVM Overfitting

A traditional way to overcome overfitting is to tune the bandwidth parameter σ in the '*rbf*' kernel directly to avoid an identity or isometric identity kernel matrix. However, there is no robust rule available to guide how to choose the bandwidth value appropriately.[6,10,11] In fact, a small σ value cannot avoid

Table 3. SVM overfitting under Bayesian *t*-test under five-fold CV.

DATA	ACCURACY (%)	SENSITIVITY (%)	SPECIFICITY (%)
Stroma	72.56 ± 3.63	100.0 ± 0.0	00.00 ± 0.0
Medulloblastoma	73.81 ± 5.32	00.00 ± 0.0	100.0 ± 0.0
Breast	52.58 ± 1.77	100.0 ± 0.0	00.00 ± 0.0
Ovarian-qaqc	56.01 ± 0.45	00.00 ± 0.0	100.0 ± 0.0
Cirrhosis	52.00 ± 0.75	00.00 ± 0.0	100.0 ± 0.0
Colorectal	57.15 ± 1.94	00.00 ± 0.0	100.0 ± 0.0

the identity or isometric identity kernel matrix issue. a large σ value will cause the kernel matrix to be a uniform matrix with only entries 1, which leads to an under-fitting problem where an SVM classifier has a low detection capability.

Furthermore, we have found that such an ad-hoc parameter tuning may avoid overfitting sometimes at the cost of low diagnostic accuracies. For example, we set σ^2 as the total variation and average total variation of the training data, respectively, and find that such an approach may not lead to a real improvement in disease diagnosis by overcoming overfitting generically, though they may contribute to some slight improvements for some individual data.

As such, we propose a sparse-coding kernel technique to conquer the SVM overfitting in order to achieve a good diagnostic accuracy for each data. A sparse-coding kernel aims at lowering both pairwise distances and data variations of training data in a kernel function by using sparse-coding techniques. In particular, the sparse coding in our context refers to representing each omics sample by "coding" it in a sparse way, where most of its entries take values zero or close to zero, whereas only a few entries take non-zero values. Obviously, such a sparse-coding technique imposes a data localization mechanism on input omics data such that each sample is only represented by a few non-zero components. Thus, it is quite clear that pairwise sample distances of the training data will decrease significantly under such a sparse representation, and the corresponding kernel matrix will no longer be the identity or isometric identity matrix. Instead, they will be more sensitive to distinguish the signatures of diseases represented by those omics samples under the sparse representation mechanism.

Sparse kernels. A sparse kernel $k(x, y) = k(f_s(x), f_s(y))$ first employs a sparse-coding function $f_s(.)$ to map an input sample to its nearest non-negative vector under a sparse-coding measure δ_s, such that they have the same L_1 and L_2 norms. That is, $f_s(x)$: $x \rightarrow x_s \geq 0$, and $f_y(y)$: $y \rightarrow y_s \geq 0$, where $||x||_1 = ||x_s||_1$, and $||x||_2 = ||x_s||_2$; $||y||_1 = ||y_s||_1$, and $||y||_2 = ||y_s||_2$ respectively. Then, the corresponding nearest non-negative vectors x_s and y_s are evaluated by a kernel function, which can be any kernel functions theoretically. In our experiment, however, we only focus on the '*rbf*' kernel with $\sigma = 1$ for the sake of overcoming overfitting.

Sparseness and sparse coding. The sparseness (measure) of an omics sample (vector) u is defined as a ratio between 0 and 1 as follows.

$$\delta_s(u) = \frac{\sqrt{n} - (\sum_{i=1}^{n} |u_i|)/(\sum_{i=1}^{n} u_i^2)^{1/2})}{\sqrt{n} - 1} \tag{9}$$

A large sparseness indicates that the vector has a few number of positive entries. The two extreme cases $\delta_s(u) = 1$ and $\delta_s(u) = 0$ refer to that there is only one entry and all entries are equal in the vector, respectively. The sparse coding of the omics sample u seeks the closest non-negative vector $\upsilon \geq 0$ in the same dimensional space on behalf of L_1 and L_2 norms such that the υ has a specified sparseness value.

In fact, the omics sample u is normalized by its L_2 norm such that it has a unit L_2 norm: $\sum_{i=1}^{n} u_i^2 = 1$ for the convenience of implementations. It is equivalent to calculating the non-negative intersection point between a hyperplane $\pi_1 : \sum_{i=1}^{n} |u_i| = \sum_{i=1}^{n} s_i$ and a hypersphere $\pi_2 : \sum_{i=1}^{n} s_i^2 = 1$ such that the non-negative vector s has the specified sparseness: $\delta_s(s) = \frac{\sqrt{n} - \sum_{i=1}^{n} s_i)}{\sqrt{n} - 1}$. This optimization problem can be solved in real-time by traditional approaches[27] or by a simple but efficient method presented in Ref. [28].

Figure 3 illustrates the sparse coding for an inflammatory breast cancer (IBC) sample and a non-inflammatory breast cancer (non-IBC) sample in the *Stroma* data with sparseness 0.3 and 0.5. It is clear that relatively large amounts of zeros are introduced in the vector with an increase in sparseness, and the pairwise distances of the nearest non-negative vectors will be much smaller than those of their corresponding original samples.

We formulate a sparse kernel ("*sparse-kernel*") by applying our sparse-coding techniques to the "*rbf*" kernel and compare its performance in SVM diagnosis with the other kernels such as linear ("*linear*"), quadratic ("*quad*"), polynomial ("*poly*"), multilayer perceptron kernels ("*mlp*"), and an *rbf* kernel with adjusted sigma ("*rbf-sigma*"). It is noted that the bandwidth of the "*rbf-sigma*" kernel is set as the total variation of all training data: $\sigma = \sum_{i=1}^{m} \sum_{j=1}^{m} || x_i - x_j ||^2$ each time.

Figure 4 illustrates the average SVM diagnosis accuracy, sensitivity, specificity, and positive prediction ratio for the "*sparse-kernel*" and other five kernels. It is interesting to find that SVM diagnosis with a sparse kernel not only successfully overcomes overfitting, but also achieves almost best performance among all kernels stably, though the linear kernel achieves the same level of performance on the *Cirrhosis* and *Medulloblastoma* data. Moreover, it seems that such a "*sparse-kernel*" SVM brings a lower standard deviation value than that with the *linear-SVM* under the same-level performance scenarios. For example, both of them achieve 94.02% diagnostic accuracy, but the sparse kernel has only 1.43% standard deviation compared with the 2.69% of the linear kernel.

In particular, we have found that the linear kernel actually encounters or at least moves closer to overfitting on the *Stroma* data, where it achieves 91.43% sensitivity but only 40% specificity, but the sparse kernel overcomes overfitting completely with 87.67% sensitivity and 78.24% specificity. As such, we say that the sparse kernel demonstrates an obvious advantage than the linear kernel in overcoming overfitting, in addition to its prediction capability.

In addition, we have found that the sparseness value demonstrates interesting impacts on the SVM diagnosis. It seems that any too low (eg, <0.2) or too high sparseness values (eg, 0.8) will not stably enhance SVM diagnosis for all data sets but will only avoid overfitting. We uniformly set sparseness $\delta_s(u) = 0.35$ for all datasets except the *Stroma* data ($\delta_s(u) = 0.42$)

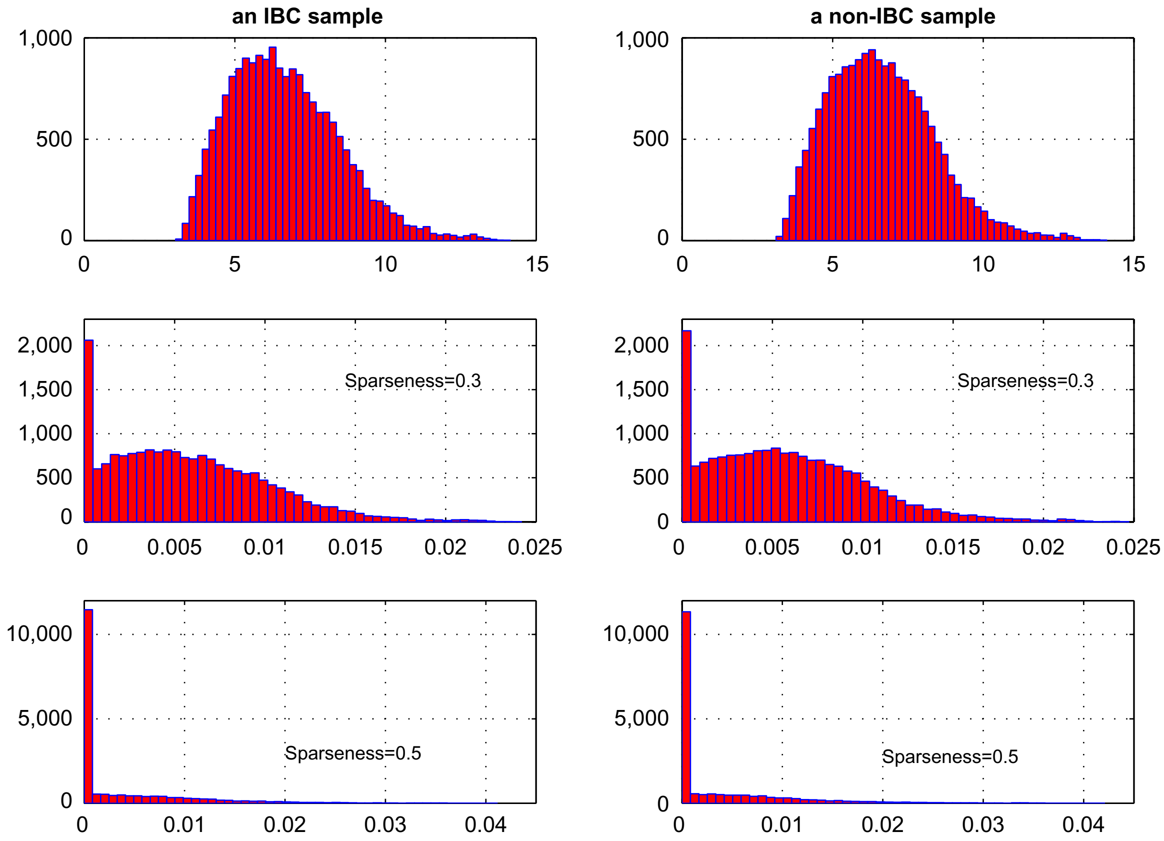

Figure 3. The sparse coding of an inflammatory breast cancer (IBC) and a non-inflammatory breast cancer (non-IBC) sample in the *Stroma* data, each of which has 18,895 genes, under 0.3 and 0.5 sparseness.

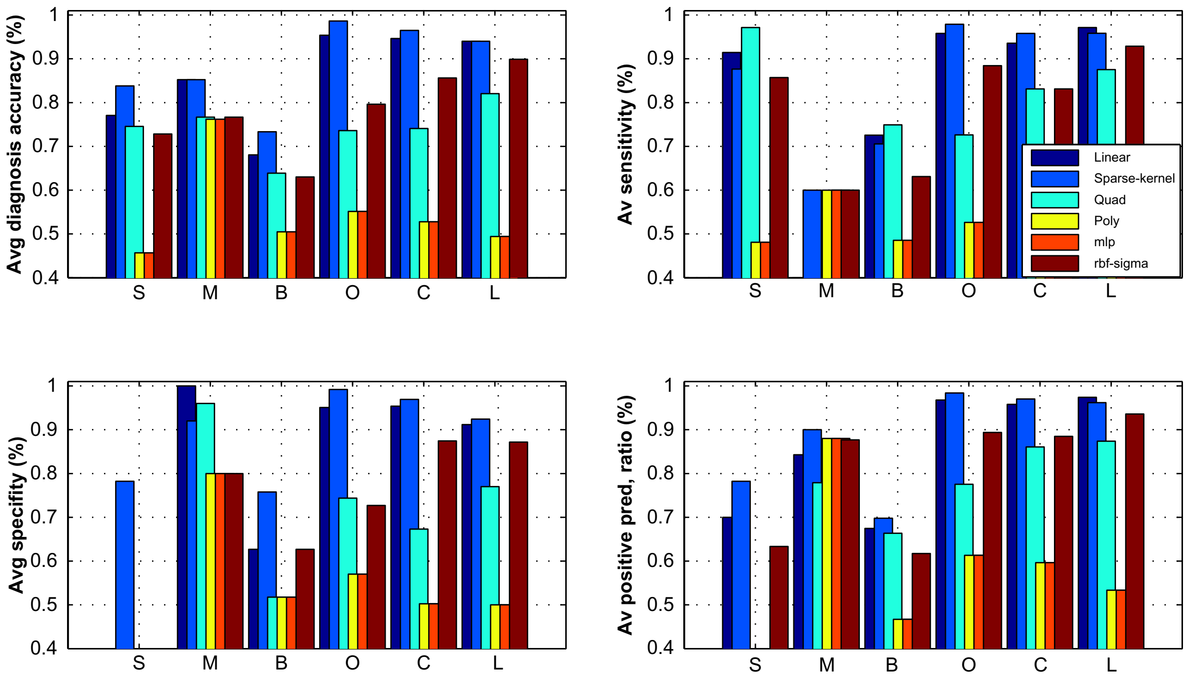

Figure 4. The comparison of the SVM diagnosis for "sparse-kernel", "linear", "quadratic", "polynomial", multilayer perceptron kernel ("mlp"), and an "*rbf*" kernel with adjusted sigma value on six omics datasets on average accuracy, sensitivity, specificity, and positive prediction ratios. The sparse kernel conquers overfitting with the best diagnosis performance compared with other kernels. Each dataset is represented by its first letter expect the Colorectal dataset, which is represented by "L."

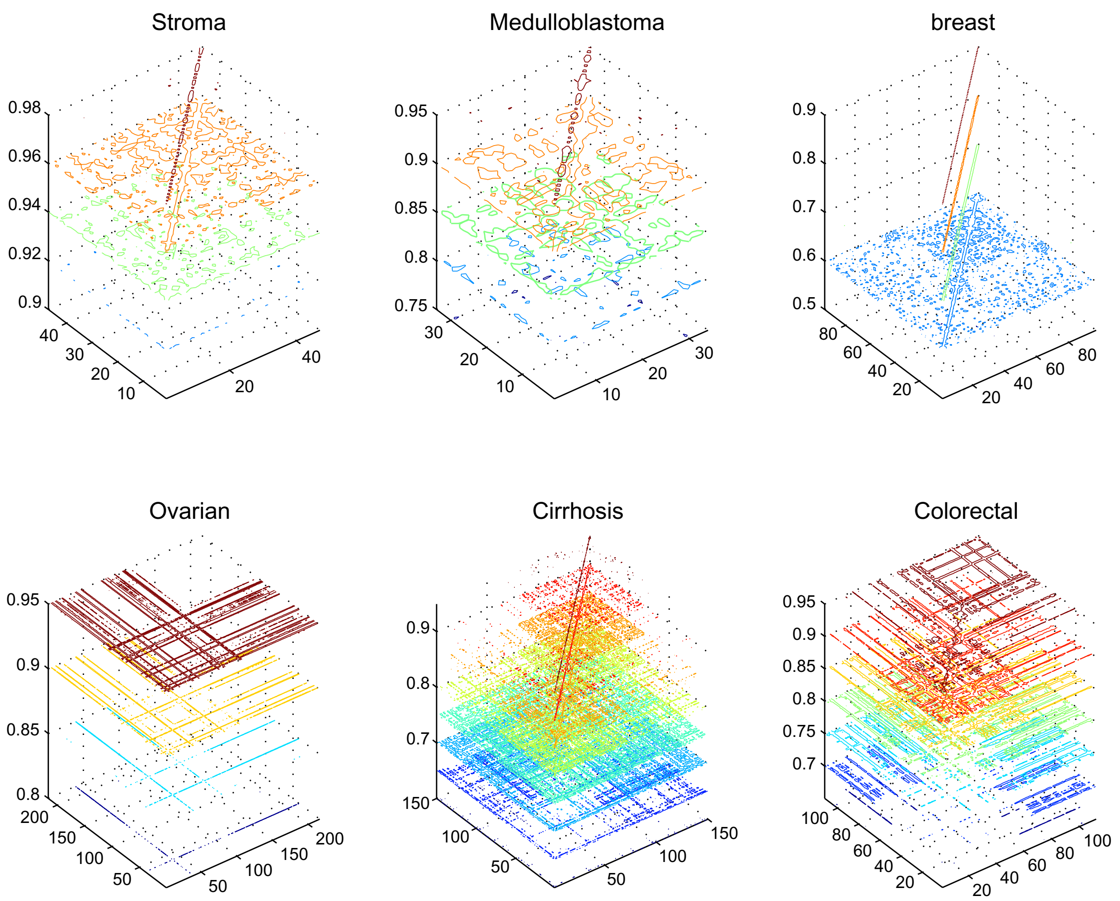

Figure 5. The contour plots of the kernel matrices of all six omics datasets under the sparse kernel. Most of the kernel matrices have entry values spanning more layers, which contributes to enhancing the SVM classifier's diagnostic power because of the optimized kernel structures.

to demonstrate the effectiveness of sparse kernels, though adaptively selecting sparseness values can result in better performance for each data.

Figure 5 illustrates the kernel matrices' contour plots under the sparse kernel for the six omics datasets, where all samples in each dataset are viewed as the training data in the SVM diagnosis for the convenience of analysis. It is obvious that our sparse coding successfully avoids the original identity or isometric identity kernel matrices associated with the '*rbf*' kernel with bandwidth $\sigma = 1$, and causes each kernel matrix to be a meaningful kernel matrix. Moreover, it is interesting to see that most of the kernel matrices have entry values spanning more layers in the contour plot, which contributes to enhancing the SVM classifier's diagnostic power. Instead, the kernel matrices, whose entry values have relatively small ranges, may lead to a low diagnostic performance. For example, the kernel matrix of the *Breast* data has most entries on or close to the surface $z = 0.6$, which corresponds to the lowest diagnostic accuracies among the six datasets.

The reason why the SVM overfitting is conquered by the sparse kernel lies in the fact that our sparse coding decreases the pairwise distances in each kernel matrix and optimizes it to be a more meaningful representative structure because of the data localization mechanism brought by the sparse-coding kernels. Figure 6 illustrates the minimum ($d^2_{\min}$), first percentile ($d^2_{0.01}$), median (d^2_{median}), and maximum ($d^2_{\max}$) values of the pairwise distance squares d^2_{ij} in the kernel matrices of the "*sparse-kernel*" SVM classifier for all samples in each omics dataset. Compared with the fact that the original pairwise distance square minimum values $d^2_{\min}$ are in the order of 10^2 under the original '*rbf*' kernel, the values are in a much smaller interval under the sparse kernel for all data, ie, $10^{-3.079699} \leq d^2_{\min} \leq 10^{-0.157466}$. It means corresponding minimum non-diagonal entries will be between $\exp(-10^{-0.157466}/2) = 0.7061$ and $\exp(-10^{-3.079699}/2) = 0.9996$, for $i \neq j$. In other words, the kernel matrices under the sparse kernel are representative and meaningful instead of the original identity or isometric identity matrices.

Moreover, we have examined the eigenvalues of the "*sparse-kernel*" matrices and found that all their eigenvalues are not only different, but also demonstrate quite large sensitivity degrees from tiny to large values (please see the lower plot in Fig. 6), by comparison with the original all "1" eigenvalues from the '*rbf*' kernel matrices. Although some eigenvalues can be relatively small for each kernel matrix, all the kernel matrices are positive definite full-rank matrices. Interestingly, the

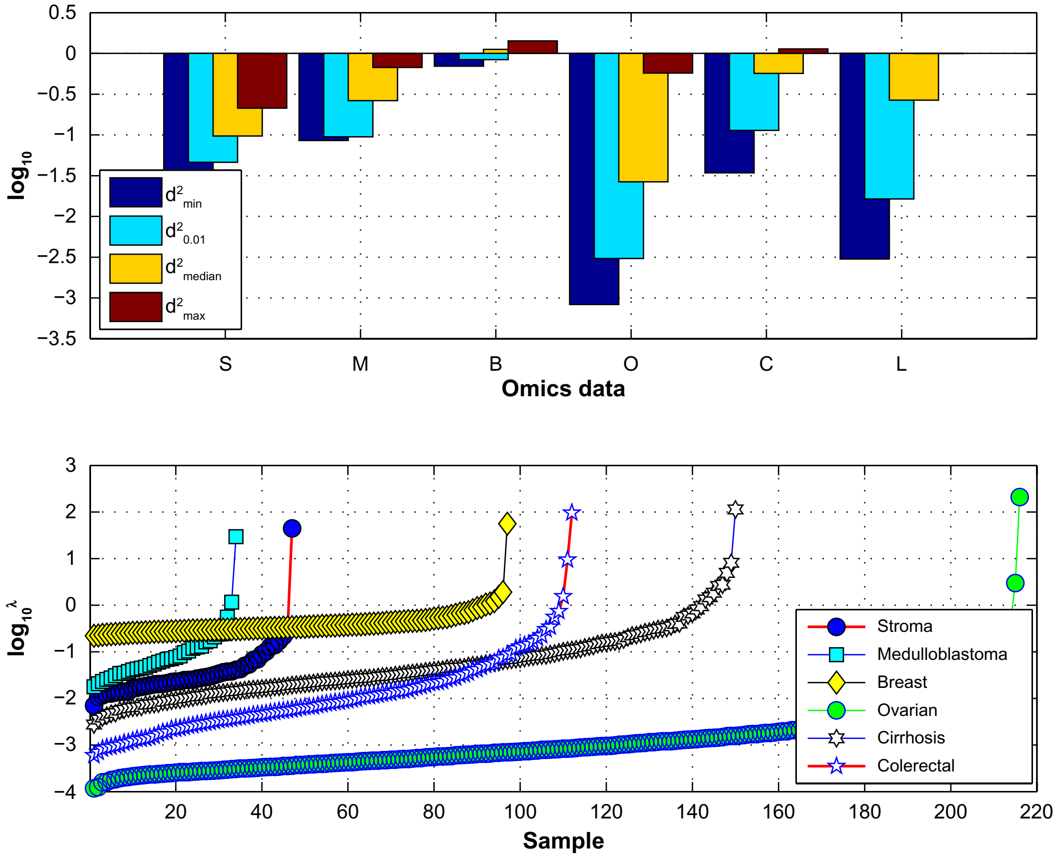

Figure 6. The minimum, median, first percentile, and maximum of the pairwise distance squares: d_{ij}^2 under the sparse kernel and the eigenvalues of the "*spare-kernel*" matrices across six omics datasets.

kernel matrices with eigenvalues spanning relatively a small value interval indicate low diagnosis accuracy. For example, the eigenvalues of the kernel matrix of the *Breast* data are in the interval [0.2202, 56.02], the smallest interval compared with those of the others. Its corresponding SVM diagnosis accuracy is the lowest among those of the six datasets, which is obviously consistent with our previous results obtained by their kernel matrix contour analysis.

Seeking Biomarkers Through Overfitting

As we have pointed out before, a single gene can still encounter overfitting by only recognizing its majority type samples in the SVM diagnostics under LOOCV. It will be more desirable to investigate such an interesting fitting by seeking its biological meaning and possible applications in cancer biomarker discovery. Such a single gene overfitting mechanism actually indicates a gene switch mechanism from a diagnostic viewpoint. That is, an individual gene loses its diagnostic ability when it encounters overfitting under a classifier (eg, SVM). In other words, as a switch, the gene turns off itself and fails to provide any useful diagnostic information. To some degree, such a gene switch can be viewed as a special case in the well-known silencing of gene expression in oligonucleotides,[29] where some genes encounter a gene silencing status and their expressions are meaningless for a classifier under a cross-validation.

However, such a gene switch mechanism provides us an option to seek meaningful biomarkers with better discriminative capabilities by only searching for those genes whose switches are "on" in diagnostics. In other words, biomarker search domains are limited to those genes whose switches turn on to conduct more targeted search and improve the biomarkers' detection power in diagnosis. As such, we propose a novel biomarker discovery algorithm: GSM by taking advantage of such a special gene switch property demonstrated by the single gene overfitting. The GSM algorithm is described as follows.

Algorithm: GSM biomarker discovery algorithm. Input:

1. An omics dataset with m samples across n genes: $X \in \mathrm{R}^{n \times m}$
2. N: the number of gene candidates to be selected in the *Bayesian t-test* filtering

Output:

A biomarker set $G = \mathrm{U}_{k=1}^{M} g_k$ with the largest diagnostic accuracy

1. Bayesian t-test filtering

a. *Score each gene in X by the Bayesian t–test*
b. *Select N genes with the smallest Bayesian factors to set S_b, such that $|S_b| = max(\lceil n \times 0.01 \rceil, N)$*

2. PCA-ranking for each gene

a. *Compute the principal component (PC) matrix for input data: $U \leftarrow pca(X)$*
b. *Project $X = X - \frac{1}{n}\vec{1}(\vec{1})^T X$ to U: $P \leftarrow X^*U$, where $\vec{1} \in \Re^n$ is a vector with all "1"s*
c. *Calculate the PCA–ranking score $\tau = \sum_{i=1}^{n} P_i^2$ for each gene (P_i is the i^{th} row of P).*

3. Biomarkers greedy capturing

1. *Initialize a biomarker set: $G \leftarrow \emptyset$*
2. *Conduct disease diagnosis with each gene in S_b with an rbf-SVM under LOOCV*
3. *Add all overfitting genes in set S_f and update $S_b \leftarrow S_b - S_f$*
4. *Add the gene g_1 with the highest accuracy and smallest τ value to G: $G \leftarrow \{g_1\}$, $S_b \leftarrow S_b - \{g_2\}$*
5. *Add the gene g_2 with the smallest τ value such that the rbf-SVM reaches its maximum accuracy under $G \cup \{g_1\}$, $G \leftarrow G \cup \{g_1\}$, $S_b \leftarrow S_b - \{g_2\}$*
6. *Proceeding like this until the rbf-SVM's accuracy stops increasing under G*
7. *Return G*

We have applied our GSM algorithm to the *Stroma* data, where $N = 200$ genes with the smallest Bayes factors are selected under the Bayesian filtering. We have found that 17 genes among them actually encounter overfitting, that is, their *rbf*-SVM diagnostic accuracy under LOOCV is always the majority ratio of this dataset: 0.7234 (34/47). Table 4 lists the PCA-ranking scores and Bayes factors of the overfitting genes, which turn off themselves in diagnosis under the *rbf*-SVM classifier.

Our GSM algorithm has identified four biomarkers and the final *rbf*-SVM diagnostic accuracy reaches 97.87% (sensitivity 92.31% and specificity 100%) under LOOCV. Table 5 illustrates its PCA-ranking scores,[44] Bayes factors, and individual SVM diagnostic accuracy. It is noted that the PCA-ranking score indicates a gene's dysregulation degree: the smaller, the more regulated. Our GSM algorithm always picks the gene with the smallest PCA-ranking score to the biomarker set *G* among several gene candidates, each of which has the same diagnostic accuracy by combining with the genes in the current biomarker set *G*.

Table 4. The 17 overfitting genes of the top 200 selected genes.

GENE-NAME	PCA-RANKING	BAYES FACTORS
CCNT2	0.4913	0.0904
RWDD3	8.4614	0.1134
USP2	7.0963	0.1380
APPBP2	33.4301	0.1475
CCPG1	5.9958	0.2206
NAB2	1.8507	0.2275
HUS1	32.8207	0.2517
SH3TC1	1.5249	0.2817
MOCS2	5.5084	0.2953
TM9SF3	10.7518	0.3062
UBE2J1	1.5724	0.4996
VPS53	8.2941	0.5198
KBTBD4	11.2837	0.5219
DLG1	0.5650	0.5310
GTF2H4	18.0115	0.5655
TRMT1	1.3772	0.5927

In addition to its excellent diagnostic accuracy, we have found that the biomarkers identified are quite meaningful and closely related to breast cancer. For example, the first biomarker is gene *USP46*, which is a broadly expressed gene reported to be a gene associated with breast cancer and glioblastomas.[30] The second biomarker is *FOSL2*, which is one of four members in the *FOSL* gene family. It is responsible for encoding leucine zipper proteins, which dimerize with proteins of the JUN family and form the transcription factor complex AP–1.[31] As a regulator in cell proliferation, differentiation, and transformation, recent studies have showed that it is one of the important genes associated with breast cancer, by being involved in the regulation of breast cancer invasion and metastasis.[32] The third biomarker is gene *RPL5*, which encodes a ribosomal protein that catalyzes protein synthesis. It was reported to be associated with biosynthesis and energy utilization, which is a cellular function associated with the pathogenesis of breast cancer.[33] In addition, it connects to breast cancer by lowering MDM2, a major regulator of p53 levels that prevents p53 ubiquitination and increases its transcriptional activity.[34] The fourth biomarker *KIF1C* is reported to be involved in podosome regulation and is associated with HPV-tumors.[35] It is interesting to see that such biomarker discovery result brings us a new biomarker *KIF1C* and three known biomarkers: *USP46*, *FOSL2*, and *RPL5* compared with our previous MICA-based biomarker discovery.[9] On the other hand, it demonstrates that the first three genes are repeatable biomarkers captured by different methods.

Table 5. Biomarkers identified by GSM for the *Stroma* data.

GENE	PCA-RANKING	BAYES FACTORS	SVM-ACCURACY
USP46	6.6276	0.0093	0.8936
FOSL2	0.3481	0.0418	0.8085
RPL5	2.6423	0.5056	0.5957
KIF1C	0.7895	0.0073	0.7872

Table 6. Biomarkers identified by GSM for the *Medulloblastoma* data.

GENE	PCA-RANKING	BAYES FACTORS	SVM-ACCURACY
NDP	100.1712	1.00E-03	0.9118
RPL21	410.9091	4.85E-03	0.8529

In addition, we have applied the proposed GSM method to the *Medulloblastoma* data and identified two important biomarkers, where 9 genes among the 100 top-ranked genes in the set S_b are overfitting genes (please see Table 6). The first biomarker is *NDP*, a gene related to Norrie disease that is reported to be a rare genetic disorder characterized by bilateral congenital blindness caused by a vascularized mass behind each lens caused by pseudoglioma. Such a finding not only strongly suggests that the medulloblastoma disease would have some very similar phenotypes to glioma, but also indicates that there are some genes related to both cancers. Interestingly, medulloblastoma was considered as a type of glioma disease in the past[36]; The second biomarker is *RPL21*, a gene encoding ribosomal proteins and has multiple processed pseudogenes dispersed through the genome. It was reported to be one of the biomarkers related to brain and other CNS cancer diseases.[36,37] In particular, the total *rbf*-SVM accuracy of the two biomarkers is 97.06% with 100.0% specificity and 88.89% sensitivity under LOOCV.

Discussion and Conclusion

In this study, we rigorously investigated the SVM overfitting problem in molecular diagnosis of disease and proposed a novel sparse kernel approach to conquer the overfitting. Our overfitting analysis unveils the special characteristics of the SVM overfitting on omics data through kernel analysis, which is essential to avoid deceptive diagnostic results and improve cancer molecular pattern discovery efficiency. As the first rigorous method proposed to conquer overfitting, our novel sparse kernel method is not only an alternative way to achieve a good disease diagnostic performance, but also a novel way to optimize kernel matrix structures in kernel-based learning. Thus, it will have a positive impact on data mining and bioinformatics.

It is noted that our sparse kernel method still needs to be further polished to improve its completeness and efficiency. For example, current sparseness degree selection is totally an empirical way instead of an optimal one. Although it is natural for us to choose a large sparseness degree for input data with a high dimensionality, it is still unknown how to adaptively select it for each omics training data in a data-driven approach. However, we are employing entropy theory to seek an optimal sparseness selection in our ongoing work.[38] Moreover, we are applying different feature-selection techniques to our sparse kernel method to filter redundant features so that the SVM classifier's kernel matrix structures can be further optimized in a low-dimensional input space.

Theoretically, the proposed sparse kernel method is a kernel optimization method to conquer an SVM classifier's overfitting on gene expression and proteomics data, in addition to enhancing the learning machine's prediction capability. We are interested in exploring its potential in multiple kernel learning and investigating its application on other omics data such as RNA-Seq and TCGA data.[39,46] Furthermore, our current overfitting analysis is only limited to binary-class diagnosis (disease vs control). However, multi-class diagnosis can be more general in determining different cancer subtypes from a clinical viewpoint. Thus, we are also interested in extending our current results to multi-class disease diagnosis by decomposing it to different binary-class cases through the "one-against-one" model.[40,45]

Although we analyze the SVM overfitting under the *k*-fold CV, LOOCV, and 50% HOCV, we have not conducted similar overfitting analysis for the widely used independent test set cross-validation. However, because of the lack of mature mathematical models and the ad-hoc training data selection, it would be hard to conduct a robust overfitting analysis. However, it does not mean that a similar overfitting problem would not happen in the situation. In fact, most investigators might neglect the occurrence of overfitting because of kernel parameter tuning and ad-hoc training/test data selection. In our future work, we plan to develop a novel mathematical model to investigate SVM diagnostic overfitting under the independent test set approach. In addition, we plan to examine systematically the relationships between the gene switch mechanism demonstrated by the SVM overfitting on individual genes and gene silencing, and seek their applications in reproducible biomarker discovery and consistent phenotype discrimination.[41,42]

Author Contributions

Conceived and designed the experiments: HH. Analyzed the data: HH. Contributed to the writing of the manuscript: HH. Agree with manuscript results and conclusions: HH, XJ. Jointly developed the structure and arguments for the paper: HH, XJ. Made critical revisions and approved final version: HH, XJ. Both authors reviewed and approved of the final manuscript. HH and XJ thank our scientific editors for their excellent work.

REFERENCES

1. Altman RB. Introduction to translational bioinformatics collection. *PLoS Comput Biol*. 2012;8(12):e1002796.
2. Shah NH, Tenenbaum JD. The coming age of data-driven medicine: translational bioinformatics' next frontier. *J Am Med Inform Assoc*. 2012;19:e2–4.
3. Fort G, Lambert-Lacroix S. Classification using partial least squares with penalized logistic regression. *Bioinformatics*. 2005;21(7):1104–11.
4. Nguyen D, Rocke D. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*. 2002;18:39–50.
5. Larrañaga P, Calvo B, Santana R, et al. Machine learning in bioinformatics. *Brief Bioinform*. 2005;7(1):86–112.
6. Vapnik V. *Statistical Learning Theory*. New York: John Wiley; 1998.
7. Oh S, Lee MS, Zhang BT. Ensemble learning with active example selection for imbalanced biomedical data classification. *IEEE/ACM Trans Comput Biol Bioinform*. 2011;8(2):316–25.

8. Han X. Nonnegative principal component analysis for cancer molecular pattern discovery. *IEEE/ACM Trans Comput Biol Bioinform*. 2010;7(3):537–49.
9. Han H, Li X. Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery. *BMC Bioinformatics*. 2011;12(S1):S7.
10. Shawe-Taylor J, Cristianini N. *Support Vector Machines and other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press; 2000.
11. Cucker F, Smale S. On the mathematical foundations of learning. *Bull. Am. Math. Soc*. 2002;39(1):1–49.
12. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*. 2004;101(12):4164–9.
13. Boersma BJ, Reimers M, Yi M, et al. A stromal gene signature associated with inflammatory breast cancer. *Int J Cancer*. 2008;122(6):1324–32.
14. Han H. Nonnegative principal component analysis for mass spectral serum profiles and biomarker discovery. *BMC Bioinformatics*. 2010;11(suppl 1):S1.
15. Conrads TP, Fusaro VA, Ross S, et al. High-resolution serum proteomic features for ovarian detection. *Endocr Relat Cancer*. 2004;11:163–78.
16. Alexandrov T, Decker J, Mertens B, et al. Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. *Bioinformatics*. 2009;25(5):643–9.
17. Ressom HW, Varghese RS, Drake SK, et al. Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics*. 2007; 23(5):619–26.
18. van ‘t Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530–6.
19. Hsu CW, et al. *A Practical Guide to Support Vector Classification (Technical Report)*. Taipei: Department of Computer Science and Information Engineering, National Taiwan University; 2003.
20. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: a library for large linear classification. *J Mach Learn Res*. 2008;9:1871–4.
21. Berger B, Peng J, Singh M. Computational solutions for omics data. *Nat Rev Genet*. 2013;14(5):333–46.
22. Weaver JM, Ross-Innes CS, Fitzgerald RC. The ’-omics’ revolution and oesophageal adenocarcinoma. *Nat Rev Gastroenterol Hepatol*. 2014;11(1):19–27.
23. Blomquist TM, Crawford EL, Lovett JL, et al. Targeted RNA-sequencing with competitive multiplex-PCR amplicon libraries. *PLoS One*. 2013;8(11):e79120.
24. Nagy ZB, Kelemen JZ, Fehér LZ, Zvara A, Juhász K, Puskás LG. Real-time polymerase chain reaction-based exponential sample amplification for microarray gene expression profiling. *Anal Biochem*. 2005;337(1):76–83.
25. López E, Wang X, Madero L, López-Pascual J, Latterich M. Functional phosphoproteomic mass spectrometry-based approaches. *Clin Transl Med*. 2012;1:20.
26. Gönen M. The Bayesian t-test and beyond. *Methods Mol Biol*. 2012;620:179–99.
27. Boyd S, Vandenberghe L. *Convex Optimization*. New York: Cambridge University Press; 2004.
28. Hoyer P. Non-negative matrix factorization with sparseness constraints. *J Mach Learn Res*. 2004;5:1457–69.
29. Deleavey GF, Damha MJ. Designing chemically modified oligonucleotides for targeted gene silencing. *Chem Biol*. 2012;19(8):937–54.
30. Holtkamp N, Ziegenhagen N, Malzer E, Hartmann C, Giese A, von Deimling A. Characterization of the amplicon on chromosomal segment 4q12 in glioblastoma multiforme. *Neuro Oncol*. 2007;9(3):291–7.
31. Milde-Langosch K, Janke S, Wagner I, et al. Role of Fra-2 in breast cancer: influence on tumor cell invasion and motility. *Breast Cancer Res Treat*. 2008;107(3):337–47.
32. Langer S, Singer CF, Hudelist G, et al. Jun and Fos family protein expression in human breast cancer: correlation of protein expression and clinicopathological parameters. *Eur J Gynaecol Oncol*. 2006;27(4):345–52.
33. Yu K, Lee CH, Tan PH, Tan P. Conservation of breast cancer molecular subtypes and transcriptional patterns of tumor progression across distinct ethnic populations. *Clin Cancer Res*. 2004;10:5508–17.
34. Lacroix M, Toillon R, Leclercq G. p53 and breast cancer, an update. *Endocr Relat Cancer*. 2006;13(2):293–325.
35. Buitrago-Pérez A, Garaulet G, Vázquez-Carballo A, Paramio JM, García-Escudero R. Molecular signature of HPV-induced carcinogenesis: pRb, p53 and gene expression profiling. *Curr Genomics*. 2009;10:26–34.
36. Smoll NR. Relative survival of childhood and adult medulloblastomas and primitive neuroectodermal tumors (PNETs). *Cancer*. 2012;118(5):1313–22.
37. Stein A, Litman T, Fojo T, Bates S. A Serial Analysis of Gene Expression (SAGE) database analysis of chemosensitivity: comparing solid tumors with cell lines and comparing solid tumors from different tissue origins. *Cancer Res*. 2014;64:2805–16.
38. Kapur JN, Kesevan HK. *Entropy Optimization Principles with Applications*. Toronto: Academic Press; 1992.
39. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509–17.
40. Hus C, Lin C. A comparison of methods for multi-class support vector machines. *IEEE Trans Neural Netw*. 2002;13(2):415–25.
41. McDermott JE, Wang J, Mitchell H, et al. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. *Expert Opin Med Diagn*. 2013;7(1):37–51.
42. Han H, Li XL, Ng SK, Ji Z. Multi-resolution-test for consistent phenotype discrimination and biomarker discovery in translational bioinformatics. *J Bioinform Comput Biol*. 2013;11(6):1343010.
43. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis. *Biostatistics*. 2010;11(2):242–53.
44. Jolliffe I. *Principal Component Analysis*. New York: Springer; 2002.
45. Gnen M, Alpaydin E. Multiple kernel learning algorithms. *J Mach Learn Res*. 2011;12:2211–68.
46. Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y. Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS One*. 2013;8(8):e71462.