Graph Constrained Discriminant Analysis: A New Method for the Integration of a Graph into a Classification Process

Vincent Guillemot[1,2]*, Arthur Tenenhaus[1,2], Laurent Le Brusquet[2], Vincent Frouin[1]*

[1] Laboratory of Functional Genomics – CEA, DSV, IRCM, Evry, France, [2] Department of Signals and Electronics Systems – Supélec, Gif-sur-Yvette, France

Abstract

Integrating gene regulatory networks (GRNs) into the classification process of DNA microarrays is an important issue in bioinformatics, both because this information has a true biological interest and because it helps in the interpretation of the final classifier. We present a method called graph-constrained discriminant analysis (gCDA), which aims to integrate the information contained in one or several GRNs into a classification procedure. We show that when the integrated graph includes erroneous information, gCDA's performance is only slightly worse, thus showing robustness to misspecifications in the given GRNs. The gCDA framework also allows the classification process to take into account as many *a priori* graphs as there are classes in the dataset. The gCDA procedure was applied to simulated data and to three publicly available microarray datasets. gCDA shows very interesting performance when compared to state-of-the-art classification methods. The software package gcda, along with the real datasets that were used in this study, are available online: http://biodev.cea.fr/gcda/.

Citation: Guillemot V, Tenenhaus A, Le Brusquet L, Frouin V (2011) Graph Constrained Discriminant Analysis: A New Method for the Integration of a Graph into a Classification Process. PLoS ONE 6(10): e26146. doi:10.1371/journal.pone.0026146

Editor: Arkady B. Khodursky, University of Minnesota, United States of America

Received February 11, 2011; **Accepted** September 20, 2011; **Published** October 14, 2011

Funding: This work was funded by a CEA - Région Ile de France grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Vincent.Guillemot@ibe.med.uni-muenchen.de (VG); vincent.frouin@cea.fr (VF)

Introduction

Very often, biologists and bioinformaticians have prior knowledge about the relationships that exist between genes under specific biological conditions. These structured priors are usually represented by a graph, called a gene regulation network (GRN) throughout this paper, in which the nodes are the genes and the edges represent interactions between genes. Integrating such a structured prior knowledge into the classification of microarray data is an important bioinformatics research field and has recently been addressed in the literature (for example, [1–4]). The properties of the Laplacian's graph eigen values are used by Rapaport *et al.* [2] and Li *et al.*[1] to compute a classifier intended to be "smooth" across the graph. Zhu *et al.* [3] encodes the graph by means of additional specific constraints in the support vector machines (SVM) [5] optimization problem (in a way that is also suggested by Rapaport *et al.* [6]). Binder *et al.* [4] proposed that the graph be incorporated in a boosting framework. All of these methods pursue the same general objective: two variables connected in the GRN must have close weights in the classification function. This type of constraint yields a better interpretability of the resulting classifier, but not necessarily better performance.

In this paper, we propose a new method - called graph constrained discriminant analysis (gCDA) - , which is a constrained version of the discriminant analysis [7], with constraint depending on information that is represented by one or more graphs. Here, we present a fully operational and validated method that has resulted from preliminary works reported in [8]. In the discriminant analysis (DA), the decision function involves the inverse of the within-class covariance matrix. In the high-dimensional setting ($n \ll p$) considered here, the usual maximum likelihood covariance estimator is singular. As a result, the use of shrinkage estimators for the covariance matrix is needed, as described in the regularized discriminant analysis (RDA) [9]. Our approach is two-fold: first, the within-class covariance estimation is shrunk by integrating the information contained in GRNs. Then, the new estimator is entered into a DA framework. The underlying motivation for this approach is to improve the accuracy of the predictions, at least when compared to RDA.

The present work is structured as follows: the first section is dedicated to the presentation of gCDA and of the state-of-the-art methods to which it is compared. The second part is devoted to the validation of gCDA on a simulated dataset and three publicly available gene expression microarray datasets [10–12].

Methods

The integration of a graph into the classification process of microarray data requires that GRNs describing the dependencies between genes and a given microarray dataset are considered. The structure of these two objects is radically different, and the challenging task we attempt to overcome in this paper is to combine these two sources of information.

Notations

Let $\mathbf{x}$ be a $n \times p$ matrix containing the expression profiles of the n individuals distributed in two classes. Each individual is associated with a response variable. In this paper, we evaluate only binary classification problems: an individual from class 1 will have a response variable equal to $y_1 = -1$ and equal to $y_2 = +1$ otherwise. The set of the individuals belonging to class k is denoted by $\mathcal{C}_k$. Let $X_k, k = 1,2$ be the two p-variate random variables that model the expression of the p genes in each class. The means and covariance matrices of these variables will be denoted as $\boldsymbol{\mu}_k$ and Σ_k, respectively. Moreover, the variables $X_k, k = 1,2$ are supposed to be multivariate and following Gaussian distributions. Therefore, an individual from the dataset, $\mathbf{x}$, is a realization of a multivariate Gaussian mixture variable of density $f = \sum_k \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$, with π_k representing the probability that an individual belongs to class k.

We consider finite, undirected graphs to model GRNs. A graph, $\mathcal{G}$, is an object defined by the set of its edges, $\mathcal{E}$, and the set of its vertices, $\mathcal{V}$. A vertex represents a gene. Hence, $\mathcal{V}$ contains p vertices. Let w be the function $w : \mathcal{V} \times \mathcal{V} \to \{0,1\}$ such that $w(j_1,j_2)$ is 1 if there is an edge between vertices j_1 and j_2 and 0 otherwise.

The Laplacian of a graph $\mathcal{G}$, denoted by $\mathcal{L}_{\mathcal{G}}$, is a semi-definite, positive, $p \times p$ matrix whose coefficients are:

$$[\mathcal{L}_{\mathcal{G}}]_{(j_1,j_2)} = \begin{cases} -w(j_1,j_2) & ,\text{if } j_1 \neq j_2 \\ d_j & ,\text{if } j_1 = j_2 = j \end{cases},$$

with d_j representing the connectivity degree of vertex j. Thus, each null term in $\mathcal{L}_{\mathcal{G}}$ corresponds to an absence of an edge in $\mathcal{G}$.

Related work

Rapaport *et al.* [2] proposed that a spectral transformation be applied to the Laplacian $\mathcal{L}_{\mathcal{G}}$. This gives a semi-definite, positive matrix, which is then used as a kernel matrix that is loaded into SVM. The authors of this work do not report any improvement of the performance of classification, but they suggest that this approach results in better interpretability of the classification model.

In a more recent study [6], Rapaport *et al.* integrate the given graph by adding constraints to the classical SVM optimization problem. These additional constraints encode the fact that two adjacent variables must have close weights in the final model. This idea is further developed by Zhu *et al.* [3], who proposed a method called network-based (NB)-SVM. This approach aims to solve the following optimization problem:

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} \mathbf{x}_i + \lambda \sum_{j_1 \sim j_2} M(j_1,j_2)$$

$$\forall i = 1,\ldots,N, y_i(\beta_0 + \mathbf{x}_i^T \beta) \geq 1 - \mathbf{x}_i$$

$$\forall j_1 \sim j_2, \left|\frac{\beta_{j_1}}{w_{j_1}}\right| \leq M(j_1,j_2) \text{and} \left|\frac{\beta_{j_2}}{w_{j_2}}\right| \leq M(j_1,j_2),$$

with $M(j_1,j_2) = \max\left(\left|\frac{\beta_{j_1}}{w_{j_1}}\right|,\left|\frac{\beta_{j_2}}{w_{j_2}}\right|\right)$, where $\mathbf{x}_i$ the expression profile of the ith individual and w_j a weight that is dependent on the degree d_j of gene j in the graph $\mathcal{G}$. The values proposed by [3] are $w_j = 1$, d_j or $\sqrt{d_j}$. In the comparison presented in our paper, we considered the case $w_j = d_j$.

The two methods described above are intended to solve the issue that variables connected in a given graph must have close coefficients in the decision function. This type of constraint clearly helps the interpretation of the resulting classifier, but it is not specifically designed to improve the performance of the classification, even if Zhu *et al.* [3] show results on simulated data that support such an improvement. By contrast, the method proposed here is explicitly designed to improve the classification accuracy.

In a nutshell, we propose to regularize the estimation of the covariance matrix by integrating information contained in the GRN(s). The resulting estimator can then simply be used in the context of DA. As described by [2], the key element of our integration procedure is the $p \times p$ Laplacian matrix of $\mathcal{G}$. As we will see in the section describing the Gaussian graphical model, the Laplacian matrix can be considered to be homogeneous to the inverse of a covariance matrix and will be used in our shrinkage target.

Discriminant Analysis

DA is a simple, yet very popular, classification method [7,13]. To implement gCDA, we focused more particularly on the Fisher's DA. This analysis aims to first determine a linear transformation, defined as V, of the dataset that is able to maximize the between versus within-class covariance ratio:

$$\mathrm{V} = \arg\max_{v \in \mathbb{R}^p} \frac{v^T \Sigma_b v}{v^T \Sigma_w v},$$

with $\Sigma_w = \pi_1 \Sigma_1 + \pi_2 \Sigma_2$, the within-class covariance matrix and $\Sigma_b = \pi_1 \pi_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$, and the between-class covariance matrix.

Considering that there are only two classes, the transformation V defines a 1-dimensional space: the discriminant axis. Once an individual $z \in \mathbb{R}^p$ is projected onto this discriminant axis, one can predict its class based on the following Bayesian decision function:

$$\delta : z \mapsto \ln \frac{P(z \in \mathcal{C}_1 | \mathrm{V}^T z)}{P(z \in \mathcal{C}_2 | \mathrm{V}^T z)}, \tag{1}$$

if $\delta(z) > 0$, it is decided that z belongs to class 1. Otherwise, z is attributed to class 2. The Gaussian assumption helps substantially to simplify the expression of δ because $V^T z$ is the realization of either a Gaussian variable $\mathcal{N}(\mu_1, \sigma_1^2)$ or $\mathcal{N}(\mu_2, \sigma_2^2)$, with probabilities equal to π_1 and π_2, respectively. This formula (1) can be rewritten as:

$$\delta(z) = \ln\left(\frac{\pi_1 \sigma_2}{\pi_2 \sigma_1}\right) + \frac{(\mathrm{V}^T z - \mu_2)^2}{2\sigma_2^2} - \frac{(\mathrm{V}^T z - \mu_1)^2}{2\sigma_1^2}. \tag{2}$$

It was shown (for example, [14]) that the unknown parameters of δ, defined in the equation (2), can be re-expressed as a function of μ_k, Σ_k:

$$\mathrm{V} = \Sigma_w^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = (\pi_1 \Sigma_1 + \pi_2 \Sigma_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$\sigma_k = \mathrm{V}^T \Sigma_k \mathrm{V} \text{ and } \mu_k = \mathrm{V}^T \boldsymbol{\mu}_k, k = 1,2.$$

Moreover, we can consider the linear and quadratic cases in the DA framework:

- in the linear case, Σ_1 is supposed to be equal to Σ_2. If this is the case, then δ is a linear function of the components of z.
- in the quadratic case, Σ_1 and Σ_2 are supposed to be different from each other, and, as a result, δ is a quadratic function of the components of z. Moreover, the quadratic case allows us to consider situations in which the GRNs from the two classes are different. In that case, we can integrate one GRN per class. Such an interesting property cannot be found in the methods presented in the literature [2],[1],[3], although it could be of interest to the biologist. Indeed, for example, new approaches have been developed with the purpose of estimating differences between GRNs that exist between two classes of patients [15]. Plus, the fact that our method is able to integrate one GRN per class is of interest in the cases where unexpected differences in phenotype are observed (e.g. some types of cancer with similar excision histology but different survival times). In those cases, differences in the connectivity of the GRN might be expected and may help in building a two-class predictor. Conversely, a better classification rate obtained when using two GRN variants could bring a validation of their biological relevance.

Due to the $n \ll p$ setting, the estimation of the covariance matrix used in the discriminant analysis has to be regularized. In gCDA, the GRNs are integrated into the covariance matrix estimator using Gaussian graphical models, hence realizing at the same time the needed regularization.

Gaussian Graphical Models

The theory of Gaussian graphical models [16] (GGM) allows for the description of the dependencies between variables by a graph and the formulation of correspondences between the graph and the covariance matrix of the considered Gaussian variables. Let X be a random, multivariate, Gaussian variable with mean μ and covariance matrix Σ. According to GGM, two variables, X_{j_1} and X_{j_2}, are independent conditionally to the remaining variables if $\left[\Sigma^{-1}\right]_{j_1,j_2}=0$. If the graph $\mathcal{G}$ describes the conditional independence between variables, then Σ has to respect the constraint:

$$j_1 \not\sim j_2 \Leftrightarrow \left[\Sigma^{-1}\right]_{j_1,j_2}=0. \quad (3)$$

With this property in mind, we propose the following shrinkage target, which integrates the *a priori* information encoded in $\mathcal{G}$:

$$\Sigma_{\mathcal{G}}^{-1}=\mathcal{L}_{\mathcal{G}}+I_p, \quad (4)$$

where I_p is the $p \times p$ identity matrix and the Laplacian matrix $\mathcal{L}_{\mathcal{G}}$ is a semi-positive matrix respecting (3).

Integrating the GRN

In the $n \ll p$ case, the empirical covariance matrix, S, is an unbiased estimator of the covariance matrix, but it shows poor performance with regard to its variance.

Guo *et al.* [17] propose to regularize this estimator in the following way: $\widehat{\Sigma}=\alpha S+(1-\alpha)I_p$. Schäfer *et al.* [18], propose to replace I_p with the so-called "target matrix" and provide a closed-form expression for the parameter α. Our method is inspired by those ideas: in gCDA, we use the model (4) to build our own target matrix $\Sigma_{\mathcal{G}}=\left(\mathcal{L}_{\mathcal{G}}+I_p\right)^{-1}$, which in turn is used to regularize the estimation of the covariance matrix

$$\widehat{\Sigma}=\alpha S+(1-\alpha)\Sigma_{\mathcal{G}}.$$

The value of the parameter α is determined with a cross-validation procedure. Let us note thereafter $\tilde{S}_w$ the estimation of the within-class covariance matrix we propose.

In the Linear gCDA, each class is supposed to have the same covariance, and there is only one GRN:

$$\tilde{S}_w(\alpha)=\alpha S_w+(1-\alpha)\Sigma_{\mathcal{G}},$$

with S_w representing the empirical, within-class covariance matrix.

In the Quadratic gCDA, each class is characterized by a different GRN:

$$\tilde{S}_w(\alpha_1,\alpha_2)=\frac{n_1}{n}\left(\alpha_1 S_1+(1-\alpha_1)\Sigma_{\mathcal{G},1}\right)+\frac{n_2}{n}\left(\alpha_2 S_2+(1-\alpha_2)\Sigma_{\mathcal{G},2}\right),$$

with $\Sigma_{\mathcal{G},k}$ and S_k representing the target matrix and the empirical covariance matrix for class $k=1,2$, respectively. The quadratic gCDA allows for the integration of two graphs, corresponding to two biological situations, into a classification process.

Results

In this section, we apply gCDA to simulated and real datasets. The performance is evaluated in a Monte Carlo cross validation (MCCV) framework: The dataset is randomly split into a training dataset (two thirds), and the rest of the dataset is used as a test dataset. The whole procedure is iterated 100 times. The tuning parameters (e.g. α or (α_1,α_2) for gCDA) of the considered

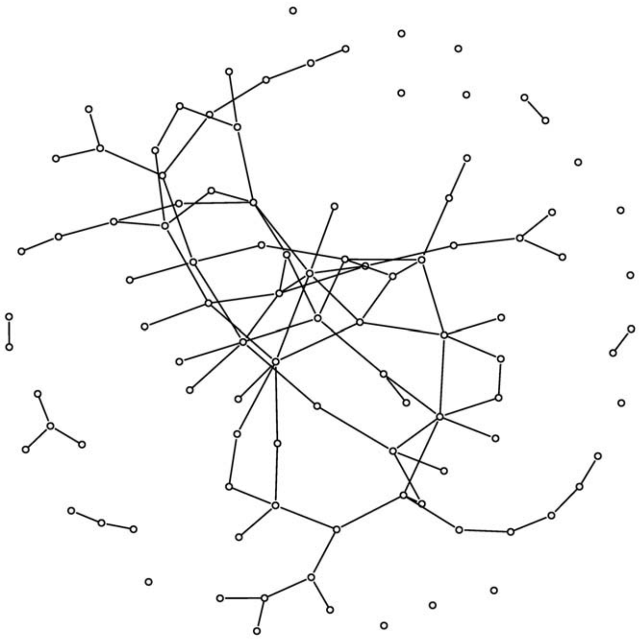

Figure 1. Graph used to generate simulated data: an Erdös-Rényi graph.
doi:10.1371/journal.pone.0026146.g001

Table 1. Results using simulated datasets.

Setting	p	RDA	SVM	LP-SVM	NB-SVM	gCDA
$\Sigma_1 = \Sigma_2$	$p = 50$	66.12 (13.79)	80.32 (6.55)	69.97 (10.04)	70.24 (10.54)	88.74 (5.07)
	$p = 100$	76.00 (21.37)	92.59 (3.58)	70.91 (11.90)	74.76 (9.70)	96.56 (2.81)
	$p = 200$	65.26 (19.36)	81.24 (7.21)	70.56 (13.10)	67.06 (8.79)	93.38 (4.13)
$\Sigma_1 \neq \Sigma_2$	$p = 50$	71.44 (12.90)	77.50 (6.43)	71.97 (9.09)	70.94 (9.06)	80.29 (6.24)
	$p = 100$	70.59 (18.73)	84.47 (5.76)	71.59 (9.97)	70.47 (9.79)	86.65 (5.92)
	$p = 200$	72.35 (21.70)	87.50 (5.44)	73.65 (12.57)	73.74 (11.77)	92.56 (4.66)

Mean of the good classification percentage (and standard deviation) over 100 MCCV iterations. Results obtained using simulated datasets. p is the number of variables. The number of individuals is set to $n = 100$. We used the linear version of gCDA when $\Sigma_1 = \Sigma_2$ and the quadratic version when $\Sigma_1 \neq \Sigma_2$.

doi:10.1371/journal.pone.0026146.t001

classification methods are computed with an internal, 10-fold cross validation. We compared gCDA to the network-based support vector machines (NB-SVM) presented by Zhu *et al.* and to the reference method they considered [3], namely linear programming (LP)-SVM [19]. Rapaport's method is not considered in the comparison because the authors stated that it does not perform better than a regular SVM classification. We also computed the performance of the regular SVM method (as implemented in the R package e1071 [20]) and the RDA (implemented in the package rda [21]).

The results presented here were obtained from simulated and microarray data. In the latter case, the performance assessment and comparisons were performed while varying the method to get the GRN and the number of GRN nodes.

Results obtained on simulated datasets

To demonstrate the performance of the presented methods, we generated simulated data. We used Erdös-Rényi's graphs (see Figure 1) to model the interactions between genes, which allows loops, hubs, and multiple connected components. We used the following algorithm:

(i) compute one Erdös-Rényi graph $\mathcal{G}$ for both classes or two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ (one per class),

(ii) for each graph, use the model given in equation (4) to build a covariance matrix,

(iii) and model the two classes by random multivariate Gaussian variables $X_1 \sim \mathcal{N}(0, \Sigma_1)$ and $X_2 \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_2)$. $\boldsymbol{\mu}$ represents the mean difference between the two datasets.

A comparison of the results obtained from simulated data for gCDA, NB-SVM, RDA, and SVM is presented in Table 1. The key result of these simulations is that the integration of the known model for the covariance matrix greatly improves the classification performance (see the performance of RDA compared to the results of gCDA). Finally, the performance is always better for gCDA than NB-SVM, the other method that integrates the known graph.

To explore the limits of gCDA, we also run the linear version of the method on a simulated dataset containing $p = 1000$ variables and $n = 100$ individuals split into two classes. The computation times of a single MCCV iteration lasted 667.82 s for gCDA against 12.65 s for SVM (on a personal computer with a processor Pentium(R) Dual core CPU E5800 3,20GHz×2 and 3.42 GB RAM). It has to be stated that the methods LP-SVM and NB-SVM could not be used on this dataset due to limited computer memory. The results are quite interesting, since SVM (86% of mean good classification rate) performs as well as gCDA (87%), whereas RDA performs as bad as a random assignment of the classes (47%). It shows that the regularization of the estimation of covariance matrices we apply in gCDA is more efficient than the one in RDA.

Additionally, Figure 2 depicts the values of the parameter α that was selected by cross validation: the selected values are close to 0, which reveals that the graph was taken into account in the estimation of the covariance matrix.

The classification performance of gCDA also depends on the quality of the integrated graph.

We empirically observed the evolution of gCDA's performance as a function of the quality of the integrated graph on simulated data. Starting with the graph that was used to simulate the dataset, we generated a set of gradually different graphs by randomly reassigning some of its edges to different vertices. The difference between two graphs is calculated as the number of different vertices between the union of the two graphs and their intersection, which corresponds to the structural Hamming distance [22] because the considered graphs are undirected. The results are shown on Figure 3. Although the best results are obtained with the real graph, our method performs robustly in spite of misspecified edges in the integrated graph. Moreover, we see that gCDA maintains its performance at least at the same level as the SVM's, even when the graph is incorrectly specified.

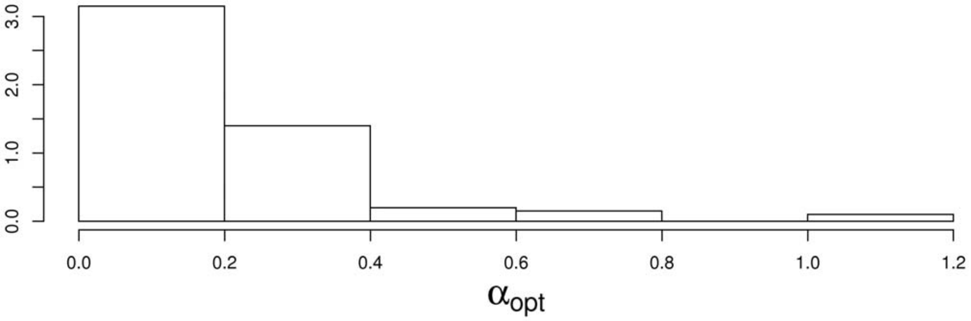

Figure 2. Histogram of the optimal values of α. These values were selected by 10-fold cross validation obtained on simulated data (linear setting, $p = 200$ and $n = 100$).

doi:10.1371/journal.pone.0026146.g002

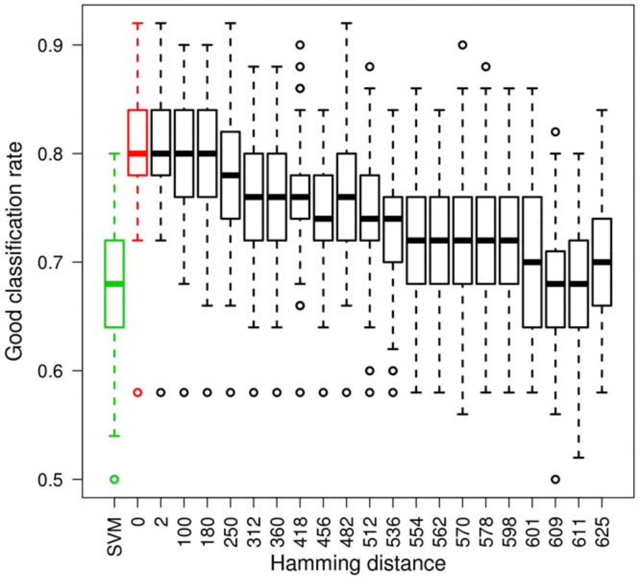

Figure 3. Plot of the classification performance as a function of the Hamming distance between the real graph and the graph integrated in gCDA. For this part of the simulation study, the number of variables is set to $p=200$ and the number of individuals to $n=100$.
doi:10.1371/journal.pone.0026146.g003

Results obtained from gene expression microarray datasets

To evaluate the performance of gCDA with real data, we chose three gene expression microarray datasets. The characteristics of the three datasets are summarized in Table 2. These datasets are available from the Gene Expression Omnibus (GEO) public database [23] and pertain to colon [10], prostate [11] and lung [12] cancers.

When dealing with gene expression microarray datasets obtained from specific tissues and under particular experimental conditions with the gCDA method, two major issues must be pointed out: 1) the graph describing the various interactions between genes is not known and has to be inferred, and 2) differences between the covariance matrices of the two classes should be evaluated. In the results reported for the real data, we investigated pragmatic choices. We selected two recognized approaches to infer the GRNs to be integrated in the classification and we also built GRNs based on reported interactions gathered in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [24] database. Namely, the three sources we used to get GRNs are

- ARACNE [25], a method based on the computation of mutual information, implemented in the package minet [26],
- ridge.net [27], based on the estimation of the partial correlation matrix, implemented in the package parcor [28],
- and the KEGG pathway hsa05200, that pertains to the biological samples we considered. The network was extracted thanks to the Bioconductor library KEGGgraph [29].

For the two first kinds of GRN mentioned above, we inferred the graphs based on a dataset independent from the one used for the classification process.

These methods impose limits on the set of variables to be considered down to several hundreds. Therefore, we selected a restricted set of genes corresponding to the KEGG pathway for human cancer (hsa05200). To avoid any bias in the classification process, we never used the same dataset to compute the GRNs and to measure the classification performance. We considered a couple of distinct datasets (see Table 2) corresponding to the same tissue and pathology: one dataset was used to infer the graphs and the other was used to build the classification models. To test the hypothesis that several covariance matrices are different, we used a statistical test adapted to high dimensional datasets presented in [30]. We used simulated data to ensure that this test is, indeed, able to distinguish between situations where the covariance matrices are equal or different. The results are not shown, but the reader is encouraged to run the example implemented in the package gcda. This test was applied to the three datasets; the obtained p-values are summarized in Table 3. It appears that the quadratic version of gCDA has to be applied only to the dataset on prostate cancer.

When necessary, we re-annotated the probe sets to associate them with a corresponding specific gene (essentially for Affymetrix chips). We used the UCSC database (http://genome.ucsc.edu/, March, 2006 (NCBI36/hg18)). For each gene, we chose the probe set whose position is the closest to the transcription initiation site. When several probe sets were selected, the mean value of the measurements was computed.

As shown in Table 4, gCDA's performance - when coupled with GRNs inferred with ARACNE - was always at least as good as the performance of SVM with a linear kernel. More importantly, gCDA always outperformed NB-SVM and RDA. The fact that gCDA outperforms RDA is indeed very interesting: it shows that the method we propose to regularize the estimation of the covariance matrix is efficient even on real datasets, when the real network is not known.

More importantly, we also assessed the way gCDA depends on the information in the GRNs by integrating three different types of GRNs: GRNs inferred with ridge.net or ARACNE and GRNs extracted from the KEGG database. The comparison between the obtained performance is presented on Table 5. This table shows that the nature of the network integrated into the classification thanks to gCDA has an interesting influence on the classification performance. For the three real datasets we analyzed, the

Table 2. Characteristics of the datasets.

outcome	n_1:n_2	p	Disease	Reference	Network inferred on
control/tumor	30:12	97	colon cancer	[10]	The rest of the original dataset
control/tumor	50:52	282	prostate cancer	[11]	Another dataset [31]
relapse/no relapse	69:69	325	lung cancer	[12]	Another dataset (GSE8332)

Summary of the characteristics of each of the datasets. $n_k, k=1,2$ represents the number of individuals in the class, k. The last column indicates whether the networks are inferred on an independent part of the dataset or on another dataset. In both cases, the dataset used to compute the networks is never used in the classification process.
doi:10.1371/journal.pone.0026146.t002

Table 3. Test on the covariance matrices.

	Colon	Lung	Prostate
p-value	0.26	0.65	$<$ 1e-3

We tested each dataset to determine whether the covariance matrices are statistically similar. The test we chose is robust enough to handle instances in which the number of variables is of the same order as the number of individuals. The null hypothesis is "$\Sigma_1 = \Sigma_2$". As a result, we rejected the null hypothesis when the p-value was lower than the threshold of 0.05.

doi:10.1371/journal.pone.0026146.t003

performance yielded with the KEGG network matches the performance obtained with one of the inferred networks for the colon cancer dataset and the lung cancer dataset, but the top performance is always achieved when using an inferred metwork. It is not such an unexpected result, since KEGG reports multiple types of gene interactions which are not necessarily relevant for transcriptomic data. On the contrary, when inferring a graph with ridge.net or ARACNE, the interactions are estimated directly on gene expression datasets, hence resulting into graphs that are much more in adequacy with the data.

Taken together, the results given in Table 4 and Table 5 show that gCDA integrates successfully GRNs into the building of the classification function. It appears to be robust enough to compensate for errors in the graph. Those results also illustrate the interest of choosing pertinent GRNs. Due to the integration of the KEGG pathways, some variables had to be removed from the analysis, which explains the differences between the two tables for gCDA (ARACNE). In addition to these results, Table 6 presents the edge differences between the three integrated GRNs for each dataset.

To conclude this section, we report the results obtained by applying gCDA to real microarray datasets while considering more than 1000 genes. The R Bioconductor package KEGGgraph was used to extract the extended version of the hsa05200 pathway from KEGG. This pathway was integrated into the classification process of lung and prostate cancer datasets. The colon cancer dataset was not considered in this part of the analysis, since only approximately 200 of its collection's genes belonged to the extended KEGG pathway. For the lung cancer dataset, there were 1252 genes in common between the collection probe sets and the extended KEGG pathway. For the prostate cancer dataset, there were 1033 genes in common between the collection probe sets and the extended KEGG pathway. We then applied linear gCDA on these two real datasets. Table 7 shows the results of these two experiments in terms of the mean of good classification rate over 100 MCCV iterations. The results obtained in this high dimensional setting raise two remarks. First, it is worth mentioning that the comparison of the mean performance is always favorable to gCDA. However, given the standard deviation, the difference between gCDA and SVM performances may not be significant. Second, the comparison between RDA and gCDA is remarkable. Indeed, the sole difference between these two methods is that for gCDA the within-class covariance estimation is shrunk by integrating KEGG prior information. From Table 7, we can observe a stable and significant improvement produced by the incorporation of the KEGG pathway information.

Table 4. Comparison of gCDA's performance with the performance of three other classification methods.

	gCDA	RDA	SVM	NB-SVM
Colon	79.36 (9.63)	69.50 (13.62)	75.07 (9.87)	54.57 (22.83)
Lung	55.93 (6.00)	49.13 (6.68)	55.02 (6.12)	50.41 (6.09)
Prostate	87.10 (5.59)	64.88 (12.1)	88.62 (5.38)	56.12 (13.2)

Comparison of the performance of gCDA with the performance obtained with RDA, SVM and NB-SVM. For NB-SVM and gCDA, we chose to integrate the GRNs inferred with ARACNE. In this table are presented the mean (standard deviation) of the good classification rate over 100 MCCV iterations.

doi:10.1371/journal.pone.0026146.t004

Table 5. Performance of the considered classification methods on three gene expression microarray datasets.

	ridge.net	ARACNE	KEGG
Colon	67.857 (11.77)	70.357 (11.37)	66.143 (12.17)
Lung	59.413 (5.88)	56.457 (6.31)	56.37 (5.83)
Prostate	87.441 (6.09)	87.029 (5.40)	84.353 (6.78)

The graphs integrated in the classification methods NB-SVM and gCDA were either inferred with two methods, ridge.net and ARACNE, or extracted from KEGG. In this table are presented the mean (standard deviation) of the good classification rate over 100 MCCV iterations.

doi:10.1371/journal.pone.0026146.t005

Discussion

Performance of gCDA

In this work, we propose a binary supervised classification algorithm of gene expression datasets that is able to integrate the information contained in gene regulation networks. The performance of gCDA is always equal to, or better than, classical SVM. When compared to state-of-the-art methods that integrate a graph, we show a significant improvement in classification performance. This result holds true whether the underlying graph is known, in the case of simulated data, or when the underlying graph of regulation is inferred, in the case of real microarray data. On real datasets, however, our method seems not to clearly outperform SVM. However, the increase in performance from RDA to gCDA, both methods based on discriminant analysis, shows that the regularization of the covariance matrix we propose is promising.

Choice of the graph integrated in gCDA

The pipeline proposed in this paper consists of two parts, the graph inference, based on classical methods, and a second step that relies on an original constrained classification algorithm. These two parts raise two major issues. First, it must be noticed

Table 6. Comparison of the integrated graphs.

# edges in:	$\mathcal{G}^{(R)}\cap\mathcal{G}^{(A)}$	$\mathcal{G}^{(R)}\cap\mathcal{G}_2^{(K)}$	$\mathcal{G}^{(A)}\cap\mathcal{G}^{(K)}$	$\mathcal{G}^{(R)}\cup\mathcal{G}^{(A)}$	$\mathcal{G}^{(R)}\cup\mathcal{G}^{(K)}$	$\mathcal{G}^{(A)}\cup\mathcal{G}^{(K)}$
Colon	35	2	6	315	158	344
Lung	263	62	18	3204	3311	1680
Prostate	69	4	19	1099	1300	1979

Comparison of the structure of the integrated graphs using ridge.net ($\mathcal{G}^{(R)}$), ARACNE ($\mathcal{G}^{(A)}$) or KEGG ($\mathcal{G}^{(K)}$). The table contains the number of edges in the intersection and the union. When two graphs were inferred, they were simply merged into a unique graph.

doi:10.1371/journal.pone.0026146.t006

Table 7. Linear gCDA applied on high dimensional microarray datasets.

	SVM	RDA	linear gCDA
Lung	59.74 (6.65)	49.30 (6.93)	60.44 (7.32)
Prostate	84.68 (5.69)	71.59 (10.35)	85.06 (5.81)

Application of gCDA to more than 1000 variables. Comparison of SVM, RDA and linear gCDA on the lung and prostate cancer datasets: mean (standard deviation) of good classification rate over 100 MCCV iterations.
doi:10.1371/journal.pone.0026146.t007

that the graph describing the various interactions between genes is not known. It has to be inferred from another dataset or from a graph that has been extracted from referenced interactions available on public databases. Second, it is usually not clear whether the covariance matrices of the two classes are different or the same.

These issues have not been addressed very often in the literature. In the procedure proposed by Rapaport *et al.*, Zhu *et al.*, Li *et al.* and Binder *et al.* [1–4] general GRNs are extracted from public knowledge databases, such as KEGG and subsequently integrated into the classification process. This kind of GRNs describes very general interactions between genes (like promoter-regulee or protein-protein interactions) and their adequacy to the biological process under study is difficult to assess without a thorough study by a specialist. We showed on real datasets that when a GRN extracted from a public database (KEGG) is used within gCDA, the resulting classification is worse than when inferred networks are used. We exemplified that one has to be very cautious when choosing a GRN to integrate into the classification process.

Linear *vs* Quadratic gCDA

To determine whether the covariance matrices of the two classes are different, we propose to use a statistical test adapted to high dimensional datasets presented in [30]. The result of this test allows choosing between the linear or the quadratic version of gCDA. The fact that it allows to integrate one GRN per class if needed is a unique feature of our method compared to other classifiers.

Impact of GRN integration

In our comparison, we restricted the reference methods to those with a direct connection to the NB-SVM method (LP-SVM and SVM) and to gCDA (RDA) to focus on how much the integrated graph can improve the classification performance. Both the interpretability and the performance of our classifier is clearly not necessarily improved compared to the approach of Rapaport *et al.*, for example, probably because of the complexity of the automatically inferred network. Apart from the fact that there may be incorrect edges, another important feature of real networks is that the weight associated to each edge is also unknown. gCDA copes with this issue by assuming an arbitrary model between the network structure and the weights. This characteristic may explain why there is no definitely significant improvement over SVM in our applications on real datasets. Future work will be dedicated to the estimation of these weights. Nevertheless, our method still shows promising classification performance on both simulated and real datasets with various complexities.

Acknowledgments

We thank Sylvain Baulande (PartnerChip) for his help in analyzing the datasets and for his useful comments. We also thank Anne-Laure Boulesteix for her helpful comments.

Author Contributions

Conceived and designed the experiments: VF LLB AT. Performed the experiments: VG. Analyzed the data: VG. Wrote the paper: VG AT ALB VF.

References

1. Li C, Li H (2008) Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24: 1175–1182.
2. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert J (2007) Classification of microarray data using gene networks. BMC Bioinformatics 8: 35.
3. Zhu Y, Shen X, Pan W (2009) Network-based support vector machine for classification of microarray samples. BMC Bioinformatics 10(Suppl 1): S21.
4. Binder H, Schumacher M (2009) Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. BMC Bioinformatics 10: 18.
5. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20: 273–297.
6. Rapaport F, Barillot E, Vert JP (2008) Classification of arraycgh data using fused svm. Bioinformatics 24: i375–i382.
7. Fisher RA (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics 7: 179–188.
8. Guillemot V, Le Brusquet L, Tenenhaus A, Frouin V (2008) Graph-constrained discriminant analysis of functional genomics data. In: Proc IEEE International Conference on Bioinformatics and BiomeidcineWorkshops BIBMW 2008.. pp 207–210. doi:10.1109/BIBMW.2008.4686237.
9. Friedman JH (1989) Regularized discriminant analysis. Journal of the American Statistical Association 84: 165–175.
10. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A 96: 6745–6750.
11. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, et al. (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1: 203–209.
12. Lee ES, Son DS, Kim SH, Lee J, Jo J, et al. (2008) Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. Clin Cancer Res 14: 7397–7404.
13. Hand DJ (2006) Classifier technology and the illusion of progress. Statistical Science 21: 1–15.
14. McLachlan GJ, Do KA, Ambroise C (2004) Analyzing Microarray Gene Expression Data. John Wiley & Sons Inc.
15. Mansmann U, Schmidberger M, Strobl R, Jurinovic V (2010) Indirect comparison of interaction graphs. In: Kneib T, Tutz G, eds. Statistical Modelling and Regression Structures, Physica- Verlag HD. pp 249–265.
16. Whittaker J (1990) Graphical Models in Applied Multivariate Statistics. New York, Wiley.
17. Guo Y, Hastie T, Tibshirani R (2007) Regularized linear discriminant analysis and its application in microarrays. Biostatistics 8: 86–100.
18. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4: Article32.
19. Bradley PS, Mangasarian OL (1998) Feature selection via concave minimization and support vector machines. In: Machine Learning Proceedings of the Fifteenth International Conference (ICML'98). pp 82–90.
20. Dimitriadou E, Hornik K, Leisch F, Meyer D, et al. (2010) e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. URL http://CRAN.R-project.org/package = e1071. R package version 1.5-24.
21. Hastie YGT, Tibshirani R (2009) rda: Shrunken Centroids Regularized Discriminant Analysis. URL http://CRAN.R-project.org/package = rda. R package version 1.0.2.
22. Kalisch M, Bühlmann P (2007) Estimating high-dimensional directed acyclic graphs with the pcalgorithm. J Mach Learn Res 8: 613–636.
23. Edgar R, Domrachev M, Lash AE (2002) Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210.
24. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The kegg resource for deciphering the genome. Nucleic Acids Res 32: D277–D280.
25. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7(Suppl 1): S7.

26. Meyer PE, Lafitte F, Bontempi G (2010) minet: Mutual Information Network Inference. URL http://CRAN.R-project.org/package = minet. R package version 2.4.0.
27. Krämer N, Schäfer J, Boulesteix AL (2009) Regularized estimation of large scale gene association networks using gaussian graphical models. BMC Bioinformatics 10: 384.
28. Krämer N, Schäfer J (2010) parcor: Regularized estimation of partial correlation matrices. URL http://CRAN.R-project.org/package = parcor. R package version 0.2-2.
29. Zhang JD, Wiemann S (2010) KEGGgraph: KEGGgraph: A graph approach to KEGG PATHWAY in R and Bioconductor. URL http://www.dkfz.de/en/mga/index.html. R package version 1.8.0.
30. Schott JR (2007) A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. Computational Statistics & Data Analysis Volume 51: 6535–6542.
31. Chandran UR, Ma C, Dhir R, Bisceglia M, Lyons-Weiler M, et al. (2007) Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. BMC Cancer 7: 64.