The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 16):0
DOI 10.1186/s12859-016-1311-3

BMC Bioinformatics

RESEARCH Open Access

Integrating gene set analysis and nonlinear predictive modeling of disease phenotypes using a Bayesian multitask formulation

Mehmet Gönen

From The 10th International Workshop on Machine Learing in Systems Biology (MLSB)
Den Haag, The Netherlands. 3-4 September 2016

Abstract

Background: Identifying molecular signatures of disease phenotypes is studied using two mainstream approaches: (i) Predictive modeling methods such as linear classification and regression algorithms are used to find signatures predictive of phenotypes from genomic data, which may not be robust due to limited sample size or highly correlated nature of genomic data. (ii) Gene set analysis methods are used to find gene sets on which phenotypes are linearly dependent by bringing prior biological knowledge into the analysis, which may not capture more complex nonlinear dependencies. Thus, formulating an integrated model of gene set analysis and nonlinear predictive modeling is of great practical importance.

Results: In this study, we propose a Bayesian binary classification framework to integrate gene set analysis and nonlinear predictive modeling. We then generalize this formulation to multitask learning setting to model multiple related datasets conjointly. Our main novelty is the probabilistic nonlinear formulation that enables us to robustly capture nonlinear dependencies between genomic data and phenotype even with small sample sizes. We demonstrate the performance of our algorithms using repeated random subsampling validation experiments on two cancer and two tuberculosis datasets by predicting important disease phenotypes from genome-wide gene expression data.

Conclusions: We are able to obtain comparable or even better predictive performance than a baseline Bayesian nonlinear algorithm and to identify sparse sets of relevant genes and gene sets on all datasets. We also show that our multitask learning formulation enables us to further improve the generalization performance and to better understand biological processes behind disease phenotypes.

Keywords: Gene set analysis, Nonlinear predictive modeling, Disease phenotypes, Multiple kernel learning, Cancer, Tuberculosis

Background

Predictive modeling is frequently used to find molecular signatures of disease phenotypes from genomic data, which helps us better understand underlying biological processes behind phenotypes and reduce data acquisition cost for future clinical samples by doing targeted profiling instead of genome-wide screens. To this aim, supervised machine learning methods such as linear classification and regression algorithms are trained to predict disease phenotypes, and features with relatively higher importance values (e.g. features with larger magnitude weights) in these parametric models are included into the signature. However, as illustrated by existing studies [1, 2], molecular signatures identified by such algorithms may

Correspondence: mehmetgonen@ku.edu.tr
Department of Industrial Engineering, Koç University, 34450 İstanbul, Turkey

not be robust due to small sample size or highly correlated nature of genomic data.

Gene set analysis methods try to identify gene sets on which disease phenotypes are dependent by calculating an enrichment score for each gene and transforming these scores into gene set scores using a summarization procedure [3]. The main advantage of these approaches is the ability to bring prior biological knowledge into the analysis in the form of biological pathways or sets of genes with similar biological functions [4], leading to more robust and clinically interpretable results than predictive modeling approaches. However, they usually assume linear dependencies between genomic data and phenotype, which may not reflect the underlying biology of disease, and have difficulties in using very small or large gene sets in the analysis.

To benefit from the best of both worlds, integrating gene set analysis and predictive modeling is already considered in many existing studies [5–7], which modify linear classification and regression algorithms to include gene set information while doing feature selection for molecular signature extraction. Even though this family of methods capture dependencies between genes, they still fail to capture nonlinear dependencies between genomic data and phenotype.

We suggest to integrate these two components using a nonlinear framework by extending our earlier Bayesian formulation [8]. Here, we develop a novel Bayesian multiple kernel learning algorithm, which trains a binary classifier with a sparse set of active gene sets using a sparsity-inducing prior, i.e. the spike and slab prior [9]. Using gene sets within a probabilistic formulation helps us identify more robust signatures even with small sample sizes. Using a kernel-based formulation enables us to capture nonlinear dependencies between genomic data and phenotype, and to use overlapping gene sets and gene sets with different sizes without any major concern. We also generalize our proposed formulation to multitask learning setting to model multiple related datasets (e.g. different patient cohorts profiled against the same phenotype) conjointly, leading to better predictive performance and more robust molecular signatures. To the best of our knowledge, [10] provides the first joint formulation of gene set analysis and nonlinear predictive modeling, which performs a survival analysis on breast cancer patients using both clinical and genomic data, using an existing discriminative multiple kernel learning algorithm. However, our approach has important advantages over their method: (i) more robustness on clinical datasets with small sample size due to its probabilistic nature, (ii) its ability to perform automatic model selection (e.g. determining the sparsity level of kernel weights) due to its fully Bayesian inference mechanism and (iii) its ability to model multiple related datasets conjointly due to its multitask learning variant.

We perform repeated random subsampling validation experiments on two cancer and two tuberculosis datasets to demonstrate the better predictive performance of our two algorithms over a baseline Bayesian nonlinear algorithm and to show the biological relevance of the genes and gene sets selected to disease phenotypes modeled.

Materials

In this study, we use two cancer and two tuberculosis datasets, where we solve binary classification problems to predict phenotype values from genomic data and to extract molecular signatures of disease phenotypes.

Diagnosis of micro-satellite instability in colorectal and endometrial carcinomas

Micro-satellite instability is a hypermutable phenotype caused by the loss of DNA mismatch repair activity. It is frequently observed in several tumor types such as colorectal, endometrial, gastric, ovarian and sebaceous carcinomas [11]. Tumors with micro-satellite instability do not respond to chemotherapeutic strategies developed for micro-satellite stable tumors, leading to its clinical importance. That is why we address the problem of predicting micro-satellite instability status of cancer patients from their gene expression data. We use two publicly available datasets provided by 'the Cancer Genome Atlas' (TCGA) consortium: (i) 'colon and rectum adenocarcinoma' (COADREAD) patients [12] and (ii) 'uterine corpus endometrial carcinoma' (UCEC) patients [13].

The phenotype values of cancer patients for both datasets are downloaded from the TCGA website (https://tcga-data.nci.nih.gov), which groups the patients into three categories: (i) 'micro-satellite instability high' (MSI-H), (ii) 'micro-satellite instability low' (MSI-L) and (iii) 'micro-satellite stable' (MSS). The preprocessed genomic characterizations of primary tumors from the patients (i.e. mRNA gene expression) are downloaded from https://www.synapse.org/#!Synapse:syn300013, where 20,530 normalized gene expression intensities are provided for each profiled primary tumor. We remove the patients with missing phenotype value or genomic data from further analysis. At the end, there are 261 and 330 patients with available phenotype value and genomic data for COADREAD and UCEC datasets, respectively. Table 1 summarizes the final datasets by listing the numbers of patients in each category together with the total number of patients.

Diagnosis of tuberculosis in adult and pediatric individuals

Tuberculosis is responsible for 1.5 million deaths in 2013 according to the World Health Organization, which makes it the second greatest killer due to a single infectious agent after HIV. It is also the leading cause of death for HIV-infected people. Its diagnosis is currently based on

Table 1 Summary of two cancer datasets

	Number of patients			
Dataset	MSI-H	MSI-L	MSS	Total
COADREAD	37	43	181	261
UCEC	108	27	195	330

MSI-H Micro-satellite instability high, *MSI-L* Micro-satellite instability low, *MSS* Micro-satellite stable

clinical and radiological features, sputum microscopy and tuberculin skin testing, which usually give false results in HIV-infected individuals [14]. New clinical diagnostic tests, especially for resource poor settings such as low-income countries with high rates of HIV, are needed to identify tuberculosis cases correctly for proper treatment. That is why we address the problem of predicting tuberculosis status of individuals from genome-wide RNA expression in host blood. We use two publicly available datasets of HIV-infected and -uninfected individuals from South Africa and Malawi: (i) adult individuals (ADULT) [14] and (ii) pediatric individuals (PEDIATRIC) [15].

The phenotype values and the genomic data for ADULT and PEDIATRIC datasets are downloaded from NCBI's Gene Expression Omnibus using GEO Series accession numbers GSE37250 and GSE39940, respectively, where the individuals are grouped into three categories: (i) 'active tuberculosis' (ATB), (ii) 'latent tuberculosis infection' (LTBI) and (iii) 'other disease' (OD). These repositories contain background subtracted and quantile normalized intensities of 47 323 probes for each individual. There are 537 and 334 individuals with available phenotype and genomic data for ADULT and PEDIATRIC datasets, respectively. Table 2 summarizes the datasets by listing the numbers of individuals in each category together with the total number of individuals.

Methods

We consider the problem of predicting phenotype values from genomic data using classification algorithms. Instead of training classifiers that use all available features, we want to develop classifiers that use very few but biologically relevant input features to identify a molecular signature of the phenotype and to reduce the data acquisition cost for test samples. However, the molecular signatures identified from, for example, gene expression data are not robust when we have limited training data [1, 2]. In such cases, we obtain different molecular signatures from different subsets of the same training set due to highly correlated nature of data, which makes knowledge extraction quite difficult. Instead, we can use our prior biological knowledge to group the input features and pick the relevant groups that are predictive of the phenotype while training the classification algorithm. We first discuss our proposed method that can learn a classifier and identify predictive gene sets simultaneously on a single dataset. We then explain how we extend our method to model multiple related datasets by identifying a common set of predictive gene sets across them.

Table 2 Summary of two tuberculosis datasets

	Number of individuals			
Dataset	ATB	LTBI	OD	Total
ADULT	195	167	175	537
PEDIATRIC	111	54	169	334

ATB Active tuberculosis, *LTBI* Latent tuberculosis infection, *OD* Other disease

Sparse Bayesian multiple kernel learning

We formulate the prediction task as a binary classification problem defined on the genomic data, denoted as domain $\mathcal{X}$, and the phenotype, denoted as domain $\mathcal{Y}$. We are given an independent and identically distributed sample $\{\boldsymbol{x}_i \in \mathcal{X}\}_{i=1}^N$ and a class label vector $\boldsymbol{y} = \{y_i \in \mathcal{Y}\}_{i=1}^N$, where N is the number of data points, and $\mathcal{Y} = \{-1, +1\}$. We are also given a list of gene sets $\{\mathcal{I}_m\}_{m=1}^P$, which encode our prior biological knowledge in terms of gene names, where $\mathcal{I}_m$ list the names of genes in the gene set m, which may be a set of genes from a biological pathway or a set of genes with similar biological functions, and P is the number of gene sets.

We choose to develop a nonlinear classifier to predict phenotype from genomic data using a kernel-based formulation due to its three main advantages [16]: (i) We can learn robust classifiers for tasks with very high dimensional representations such as genomic data and small sample size (i.e. large p, small n). (ii) We can learn better classifiers using nonlinear kernels such as the Gaussian kernel (i.e. kernel trick). (iii) We can use domain-specific kernels (e.g. graph and tree kernels for structured objects) to better capture the underlying biological processes [17]. To calculate similarities between the data points, we have multiple kernel functions defined over gene sets, namely, $\{k_m \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}\}_{m=1}^P$, which are used to calculate the kernel matrices $\{\mathbf{K}_m\}_{m=1}^P$. For each gene set, the corresponding kernel $k_m(\boldsymbol{x}_i, \boldsymbol{x}_j | \mathcal{I}_m)$ considers only the features extracted from or related to the genes in $\mathcal{I}_m$. We choose to learn a weighted combination of the input kernels $\{\mathbf{K}_m\}_{m=1}^P$ while training a binary classifier, which is known as multiple kernel learning [18], by extending our earlier Bayesian formulation [8] with a sparsity-inducing prior on the kernel weights. Figure 1 gives a schematic description of the proposed model.

Probabilistic model

Our proposed probabilistic model, called 'sparse Bayesian multiple kernel learning' (SBMKL), has three main parts: (i) finding kernel-specific latent variables using the same

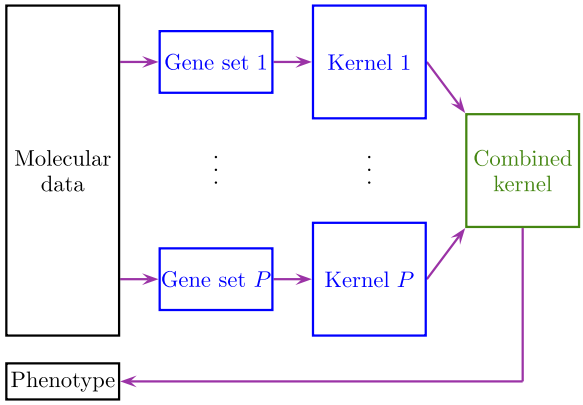

Fig. 1 Schematic description of sparse Bayesian multiple kernel learning. For each gene set, the corresponding kernel considers only the features extracted from or related to the genes in this gene set. We then learn a weighted sparse combination of these kernels while training a binary classifier to predict the phenotype values

set of sample weights over the input kernels, (ii) assigning sparse weights to these latent variables using the spike and slab prior [9] and (iii) generating predicted outputs using the latent variables and these sparse weights together with a bias parameter.

The first part has the following distributional assumptions:

$$\begin{aligned} \lambda_i &\sim \text{Gamma}(\lambda_i; \alpha_\lambda, \beta_\lambda) && \forall i \\ a_i|\lambda_i &\sim \text{Normal}\left(a_i; 0, \lambda_i^{-1}\right) && \forall i \\ g_i^m|\boldsymbol{a}, \boldsymbol{k}_{m,i} &\sim \text{Normal}\left(g_i^m; \boldsymbol{a}^\top \boldsymbol{k}_{m,i}, \sigma_g^2\right) && \forall (m,i), \end{aligned}$$

where the superscript indexes the rows, the subscript indexes the columns, $\text{Normal}(\cdot; \boldsymbol{\mu}, \Sigma)$ represents the normal distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ, and $\text{Gamma}(\cdot; \alpha, \beta)$ denotes the gamma distribution with the shape parameter α and the scale parameter β. We generate the latent variables $\boldsymbol{g}^m$ for each input kernel $\mathbf{K}_m$ using the same set of sample weights $\boldsymbol{a}$. Note that we need to use a small noise parameter σ_g while generating the latent variables to better generalize to test data points.

The second part has the following distributional assumptions:

$$\begin{aligned} \kappa &\sim \text{Beta}(\kappa; \zeta_\kappa, \eta_\kappa) && \\ s_m|\kappa &\sim \text{Bernoulli}(s_m; \kappa) && \forall m \\ \omega &\sim \text{Gamma}(\omega; \alpha_\omega, \beta_\omega) && \\ e_m|\omega &\sim \text{Normal}\left(e_m; 0, \omega^{-1}\right) && \forall m, \end{aligned}$$

where $\text{Beta}(\cdot; \zeta, \eta)$ denotes the beta distribution with the shape parameters ζ and η, and $\text{Bernoulli}(\cdot; \pi)$ represents the Bernoulli distribution with the success probability parameter π. We generate a binary indicator variable s_m and a normally distributed weight e_m for each input kernel. The product of these two variables $s_m e_m$ is a simple parameterization of the spike and slab prior, which is more amenable to approximate inference.

The third part has the following distributional assumptions:

$$\begin{aligned} \gamma &\sim \text{Gamma}(\gamma; \alpha_\gamma, \beta_\gamma) && \\ b|\gamma &\sim \text{Normal}(b; 0, \gamma^{-1}) && \\ f_i|b, \boldsymbol{e}, \boldsymbol{s}, \boldsymbol{g}_i &\sim \text{Normal}\left(f_i; (\boldsymbol{s} \circ \boldsymbol{e})^\top \boldsymbol{g}_i + b, 1\right) && \forall i \\ y_i|f_i &\sim \text{Kronecker}(f_i y_i > \nu) && \forall i, \end{aligned}$$

where $\circ$ represents the Hadamard product, and Kronecker$(\cdot)$ denotes the Kronecker delta function that returns 1 if its argument is true and 0 otherwise. The predicted outputs $\boldsymbol{f}$, similar to the discriminant outputs in support vector machines, are introduced to make the inference procedures efficient [19]. The nonnegative margin parameter ν is introduced to resolve the scaling ambiguity and to place a low-density region between two classes, similar to the margin idea in support vector machines, which is generally used for semi-supervised learning [20].

Figure 2 illustrates the proposed probabilistic model for binary classification with a graphical model.

Inference using variational Bayes

We need to infer the posterior distribution over the model parameters and the latent variables, which we denote as $\Theta = \{\boldsymbol{\lambda}, \boldsymbol{a}, \mathbf{G}, \kappa, \boldsymbol{s}, \omega, \boldsymbol{e}, \gamma, b, \boldsymbol{f}\}$, given the input kernel matrices $\{\mathbf{K}_m\}_{m=1}^P$ and the class labels $\boldsymbol{y}$ to find the predictive distribution for test data points. Unfortunately, exact inference for our proposed probabilistic model is intractable. Instead of using a computationally expensive

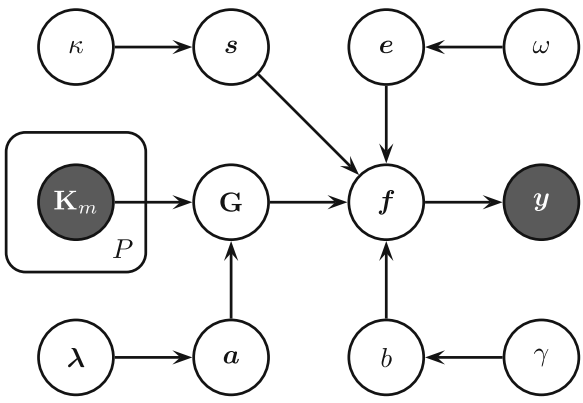

Fig. 2 Graphical model of sparse Bayesian multiple kernel learning. Random variables are shown as *empty circles*, whereas observed variables are shown as *filled circles*. Hyper-parameters are ignored for simplicity

Gibbs sampling approach [21], we choose to perform variational inference, which maximizes a lower bound on the marginal likelihood using an ensemble of factored posteriors to infer the joint parameter distribution [22].

We approximate the posterior distribution over the model parameters and the latent variables by a variational distribution:

$$p(\Theta|\{\mathbf{K}_m\}_{m=1}^P, \boldsymbol{y}) \approx q(\Theta),$$

where we assume that the variational distribution has a simpler form than the posterior distribution to make inference tractable. The inference problem can be defined as finding the nearest variational distribution to the posterior distribution with respect to a distance function. We perform mean-field variational Bayes, which measures the distance between distributions q and p using 'the Kullback–Leibler divergence' denoted as $\mathcal{KL}(q||p)$. We can decompose the log evidence as

$$\log p(\boldsymbol{y}|\{\mathbf{K}_m\}_{m=1}^P) = \underbrace{\int q(\Theta) \log \frac{p\left(\Theta, \boldsymbol{y}|\{\mathbf{K}_m\}_{m=1}^P\right)}{q(\Theta)} \mathrm{d}\Theta}_{\mathcal{L}(q)} + \underbrace{\int -q(\Theta) \log \frac{p\left(\Theta|\{\mathbf{K}_m\}_{m=1}^P, \boldsymbol{y}\right)}{q(\Theta)} \mathrm{d}\Theta}_{\mathcal{KL}(q||p)},$$

where we assume without loss of generality that all model parameters and latent variables are continuous variables, and see that minimizing $\mathcal{KL}(q||p)$ amounts to maximizing the lower bound $\mathcal{L}(q)$.

We start by writing $q(\Theta)$ as a factorized approximation:

$$q(\Theta) = q(\boldsymbol{\lambda})q(\boldsymbol{a})q(\mathbf{G})q(\kappa)q(\boldsymbol{s})q(\omega)q(\boldsymbol{e}|\boldsymbol{s})q(\gamma)q(b)q(\boldsymbol{f}),$$

where we couple the weights $\boldsymbol{e}$ with the binary indicator variables $\boldsymbol{s}$ due to their strong correlation. Note that we choose not to have the factorization as $q(\boldsymbol{e})q(\boldsymbol{s})$ because it gives a unimodal distribution, but the true posterior distribution may have exponentially many modes. To capture this multimodal structure, we choose to formulate the factorization as $q(\boldsymbol{e}|\boldsymbol{s})q(\boldsymbol{s})$, which can be approximated efficiently [23]. We then write $\mathcal{L}(q)$ in the form of expectations:

$$\mathcal{L}(q) = \mathrm{E}_{q(\Theta)}[\log p(\Theta, \boldsymbol{y}|\{\mathbf{K}_m\}_{m=1}^P)] - \mathrm{E}_{q(\Theta)}[\log q(\Theta)],$$

where we iteratively maximize $\mathcal{L}(q)$ with respect to each factor until convergence. The approximate posterior distribution of a specific factor τ can be found as

$$q(\tau) \propto \exp\left(\mathrm{E}_{q(\Theta\setminus\tau)}\left[\log p\left(\Theta, \boldsymbol{y}|\{\mathbf{K}_m\}_{m=1}^P\right)\right]\right).$$

Inference details

We define the factors for the first part of our probabilistic model as

$$\begin{aligned} q(\boldsymbol{\lambda}) &= \prod_{i=1}^N \mathrm{Gamma}(\lambda_i; \alpha(\lambda_i), \beta(\lambda_i)) \\ q(\boldsymbol{a}) &= \mathrm{Normal}(\boldsymbol{a}; \mu(\boldsymbol{a}), \Sigma(\boldsymbol{a})) \\ q(\mathbf{G}) &= \prod_{i=1}^N \mathrm{Normal}(\boldsymbol{g}_i; \mu(\boldsymbol{g}_i), \Sigma(\boldsymbol{g}_i)), \end{aligned}$$

where $\alpha(\cdot), \beta(\cdot), \mu(\cdot)$, and $\Sigma(\cdot)$ denote the shape parameter, the scale parameter, the mean vector and the covariance matrix of their arguments, respectively. The approximate posterior distributions can be updated as

$$\begin{aligned} \alpha(\lambda_i) &= \alpha_\lambda + 1/2 \\ \beta(\lambda_i) &= \left(1/\beta_\lambda + \langle a_i^2 \rangle/2\right)^{-1} \\ \Sigma(\boldsymbol{a}) &= \left(\mathrm{diag}(\langle \boldsymbol{\lambda} \rangle) + \sigma_g^{-2} \sum_{m=1}^P \mathbf{K}_m \mathbf{K}_m^\top\right)^{-1} \\ \mu(\boldsymbol{a}) &= \Sigma(\boldsymbol{a}) \left(\sigma_g^{-2} \sum_{m=1}^P \mathbf{K}_m \langle (\boldsymbol{g}^m)^\top \rangle\right) \\ \Sigma(\boldsymbol{g}_i) &= \left(\sigma_g^{-2}\mathbf{I} + \langle (\boldsymbol{s} \circ \boldsymbol{e})(\boldsymbol{s} \circ \boldsymbol{e})^\top \rangle\right)^{-1} \\ \mu(\boldsymbol{g}_i) &= \Sigma(\boldsymbol{g}_i) \left(\sigma_g^{-2} [\, \boldsymbol{k}_{1,i} \ \ldots \ \boldsymbol{k}_{P,i}]^\top \langle \boldsymbol{a} \rangle \right. \\ &\quad \left. + \langle f_i \rangle \langle \boldsymbol{s} \circ \boldsymbol{e} \rangle - \langle b \rangle \langle \boldsymbol{s} \circ \boldsymbol{e} \rangle\right), \end{aligned}$$

where $\langle h(\cdot) \rangle$ denotes the posterior expectation as usual, i.e. $\mathrm{E}_{q(\cdot)}[\, h(\cdot)]$.

The factors for the second part of our probabilistic model are defined as

$$\begin{aligned} q(\kappa) &= \mathrm{Beta}(\kappa; \zeta(\kappa), \eta(\kappa)) \\ q(\boldsymbol{s}) &= \prod_{m=1}^P \mathrm{Bernoulli}(s_m; \pi(s_m)) \\ q(\omega) &= \mathrm{Gamma}(\omega; \alpha(\omega), \beta(\omega)) \\ q(\boldsymbol{e}|\boldsymbol{s}) &= \prod_{m=1}^P \mathrm{Normal}(e_m|s_m; \mu(e_m|s_m), \Sigma(e_m|s_m)), \end{aligned}$$

where $\zeta(\cdot), \eta(\cdot)$ and $\pi(\cdot)$ denote the shape parameters and the success probability parameter of their arguments. We can update the approximate posterior distributions as

$$\zeta(\kappa) = \zeta_\kappa + \sum_{m=1}^{P} \langle s_m \rangle$$
$$\eta(\kappa) = \eta_\kappa + P - \sum_{m=1}^{P} \langle s_m \rangle$$
$$\pi(s_m) = 1/(1+\exp(-r_m))$$
$$\alpha(\omega) = \alpha_\omega + P/2$$
$$\beta(\omega) = \left(1/\beta_\omega + \sum_{m=1}^{P} (\langle 1-s_m \rangle \langle e_m^2|0\rangle + \langle s_m \rangle \langle e_m^2|1\rangle)/2\right)^{-1}$$
$$\Sigma(e_m|0) = 1/\langle \omega \rangle$$
$$\mu(e_m|0) = 0$$
$$\Sigma(e_m|1) = 1/(\langle \omega \rangle + \langle \boldsymbol{g}^m (\boldsymbol{g}^m)^\top \rangle)$$
$$\mu(e_m|1) = \Sigma(e_m|1) \sum_{i=1}^{N} \left[(\langle f_i \rangle - \langle b \rangle) \langle g_i^m \rangle - \sum_{l \neq m} \langle s_l \rangle \langle e_l|1 \rangle \langle g_i^l g_i^m \rangle \right],$$

where the auxiliary variable r_m is defined as

$$r_m = \left\langle \log \frac{\kappa}{1-\kappa} \right\rangle - \frac{1}{2} \langle e_m^2|1 \rangle \langle \boldsymbol{g}^m (\boldsymbol{g}^m)^\top \rangle + \langle e_m|1 \rangle \sum_{i=1}^{N} \left[\left(\langle f_i \rangle - \langle b \rangle \right) \langle g_i^m \rangle - \sum_{l \neq m} \langle s_l \rangle \langle e_l|1 \rangle \langle g_i^l g_i^m \rangle \right].$$

We define the factors for the third part of our probabilistic model as

$$q(\gamma) = \text{Gamma}(\gamma; \alpha(\gamma), \beta(\gamma))$$
$$q(b) = \text{Normal}(b; \mu(b), \Sigma(b))$$
$$q(\boldsymbol{f}) = \prod_{i=1}^{N} \text{TruncatedNormal}(f_i; \mu(f_i), \Sigma(f_i), \rho(f_i)),$$

where $\text{TruncatedNormal}(\cdot; \boldsymbol{\mu}, \Sigma, \rho(\cdot))$ denotes the truncated normal distribution with the mean vector $\boldsymbol{\mu}$, the covariance matrix Σ and the truncation rule $\rho(\cdot)$ such that $\text{TruncatedNormal}(\cdot; \boldsymbol{\mu}, \Sigma, \rho(\cdot)) \propto \text{Normal}(\cdot; \boldsymbol{\mu}, \Sigma)$ if $\rho(\cdot)$ is true, and $\text{TruncatedNormal}(\cdot; \boldsymbol{\mu}, \Sigma, \rho(\cdot)) = 0$ otherwise. The approximate posterior distributions can be updated as

$$\alpha(\gamma) = \alpha_\gamma + 1/2$$
$$\beta(\gamma) = (1/\beta_\gamma + \langle b^2 \rangle/2)^{-1}$$
$$\Sigma(b) = (\langle \gamma \rangle + N)^{-1}$$
$$\mu(b) = \Sigma(b) \left(\sum_{i=1}^{N} \langle f_i \rangle - \langle (\boldsymbol{s} \circ \boldsymbol{e})^\top \rangle \langle \boldsymbol{g}_i \rangle \right)$$
$$\Sigma(f_i) = 1$$
$$\mu(f_i) = \langle (\boldsymbol{s} \circ \boldsymbol{e})^\top \rangle \langle \boldsymbol{g}_i \rangle + \langle b \rangle$$
$$\rho(f_i) \triangleq f_i y_i > \nu,$$

where we can fortunately calculate the expectation of the truncated normal distribution in closed-form.

Prediction scenario

We can replace $p\left(\boldsymbol{a}|\{\mathbf{K}_m\}_{m=1}^P, \boldsymbol{y}\right)$ with its approximate posterior distribution $q(\boldsymbol{a})$ and obtain the posterior predictive mean of the latent variables $\boldsymbol{g}_\star$ for a new data point $\boldsymbol{x}_\star$ as

$$\langle \boldsymbol{g}_\star \rangle = [\, \boldsymbol{k}_{1,\star} \ \ldots \ \boldsymbol{k}_{P,\star}]^\top \langle \boldsymbol{a} \rangle.$$

The posterior predictive mean of the predicted output $f_\star$ can also be found by replacing $p(b, \boldsymbol{e}, \boldsymbol{s}|\{\mathbf{K}_m\}_{m=1}^P, \boldsymbol{y})$ with its approximate posterior distribution $q(b)q(\boldsymbol{e}|\boldsymbol{s})q(\boldsymbol{s})$:

$$\langle f_\star \rangle = \langle (\boldsymbol{s} \circ \boldsymbol{e})^\top \rangle \langle \boldsymbol{g}_\star \rangle + \langle b \rangle,$$

where we use $\langle f_\star \rangle$ to predict the class label by looking at its sign.

Sparse Bayesian multitask multiple kernel learning

We formulate the joint modeling of prediction tasks on multiple datasets using a multitask learning approach, which models distinct but related tasks conjointly to improve overall generalization performance. We are given T datasets, and, for each dataset, we have an independent and identically distributed sample $\{\boldsymbol{x}_{t,i} \in \mathcal{X}\}_{i=1}^{N_t}$ and a class label vector $\boldsymbol{y}_t = \{y_{t,i} \in \mathcal{Y}\}_{i=1}^{N_t}$, where N_t is the number of data points in the dataset t. We also have a list of gene sets $\{\mathcal{I}_m\}_{m=1}^P$, which are shared across the tasks, and the corresponding kernel functions $\{k_{t,m}(\cdot,\cdot|\mathcal{I}_m)\}_{m=1}^P$ for each task.

Probabilistic model

Our single-task learning model SBMKL is extended towards multitask learning to obtain 'sparse Bayesian multitask multiple kernel learning' (SBMTMKL).

The distributional assumptions of the first part can be modified as

$$\lambda_{t,i} \sim \text{Gamma}(\lambda_{t,i}; \alpha_\lambda, \beta_\lambda) \qquad \forall (t,i)$$
$$a_{t,i}|\lambda_i \sim \text{Normal}\left(a_{t,i}; 0, \lambda_{t,i}^{-1}\right) \qquad \forall (t,i)$$
$$g_{t,i}^m|\boldsymbol{a}_t, \boldsymbol{k}_{t,m,i} \sim \text{Normal}\left(g_{t,i}^m; \boldsymbol{a}_t^\top \boldsymbol{k}_{t,m,i}, \sigma_g^2\right) \qquad \forall (t,m,i),$$

where we have task-specific model variables and latent variables.

The distributional assumptions of the second part are written as

$$\kappa \sim \text{Beta}(\kappa; \zeta_\kappa, \eta_\kappa)$$
$$s_m|\kappa \sim \text{Bernoulli}(s_m; \kappa) \qquad \forall m$$
$$\omega_t \sim \text{Gamma}(\omega_t; \alpha_\omega, \beta_\omega) \qquad \forall t$$
$$e_{t,m}|\omega_t \sim \text{Normal}(e_{t,m}; 0, \omega_t^{-1}) \qquad \forall (t,m),$$

where the binary indicator variables are shared across the tasks, which helps us transfer information between them.

The distributional assumptions of the third part can be modified as

$$\begin{aligned}
\gamma_t &\sim \text{Gamma}\left(\gamma_t;\alpha_\gamma,\beta_\gamma\right) && \forall t\\
b_t|\gamma_t &\sim \text{Normal}\left(b_t;0,\gamma_t^{-1}\right) && \forall t\\
f_{t,i}|b_t,\boldsymbol{e}_t,\boldsymbol{s},\boldsymbol{g}_{t,i} &\sim \text{Normal}\left(f_{t,i};(\boldsymbol{s}\circ\boldsymbol{e}_t)^\top\boldsymbol{g}_{t,i}+b_t,1\right) && \forall(t,i)\\
y_{t,i}|f_{t,i} &\sim \text{Kronecker}\left(f_{t,i}y_{t,i}>\nu\right) && \forall(t,i),
\end{aligned}$$

where we have task-specific bias parameters and predicted outputs.

Inference using variational Bayes

We approximate the posterior distribution over the model parameters and the latent variables by a variational distribution:

$$p\left(\Theta|\{\{\mathbf{K}_{t,m}\}_{m=1}^P,\boldsymbol{y}_t\}_{t=1}^T\right)\approx q(\Theta),$$

where we start inference by writing $q(\Theta)$ as a factorized approximation:

$$\begin{aligned}
q(\Theta)=&\prod_{t=1}^T\Big[q(\boldsymbol{\lambda}_t)q(\boldsymbol{a}_t)q(\mathbf{G}_t)\Big]q(\kappa)q(\boldsymbol{s})\prod_{t=1}^T\Big[q(\omega_t)q(\boldsymbol{e}_t|\boldsymbol{s})\Big]\\
&\times\prod_{t=1}^T\Big[q(\gamma_t)q(b_t)q(\boldsymbol{f}_t)\Big].
\end{aligned}$$

Inference details

The update equations of the approximate posterior distributions for all model parameters and latent variables are very similar to those of SBMKL except for the binary indicator variables. We can update the approximate posterior distribution of them as

$$\pi(s_m)=1/(1+\exp(-r_m))$$

where the auxiliary variable r_m is defined as

$$\begin{aligned}
r_m=&\left\langle\log\frac{\kappa}{1-\kappa}\right\rangle-\frac{1}{2}\sum_{t=1}^T\langle e_{t,m}^2|1\rangle\langle\boldsymbol{g}_t^m(\boldsymbol{g}_t^m)^\top\rangle\\
&+\sum_{t=1}^T\langle e_{t,m}|1\rangle\sum_{i=1}^{N_t}\left[\left(\langle f_{t,i}\rangle-\langle b_t\rangle\right)\langle g_{t,i}^m\rangle-\sum_{l\neq m}\langle s_l\rangle\langle e_{t,l}|1\rangle\langle g_{t,i}^l g_{t,i}^m\rangle\right].
\end{aligned}$$

Prediction scenario

We can use the approximate posterior distribution $q(\boldsymbol{a}_t)$ instead of $p(\boldsymbol{a}_t|\{\{\mathbf{K}_{t,m}\}_{m=1}^P,\boldsymbol{y}_t\}_{t=1}^T)$ and obtain the posterior predictive mean of the latent variables $\boldsymbol{g}_{t,\star}$ for a new data point $\boldsymbol{x}_{t,\star}$ in the task t as

$$\langle\boldsymbol{g}_{t,\star}\rangle=[\,\boldsymbol{k}_{t,1,\star}\;\ldots\;\boldsymbol{k}_{t,P,\star}]^\top\langle\boldsymbol{a}_t\rangle.$$

The posterior predictive mean of the predicted output $f_{t,\star}$ can also be found by replacing $p\left(b_t,\boldsymbol{e}_t,\boldsymbol{s}|\left\{\{\mathbf{K}_{t,m}\}_{m=1}^P,\boldsymbol{y}_t\right\}_{t=1}^T\right)$ with its approximate posterior distribution $q(b_t)q(\boldsymbol{e}_t|\boldsymbol{s})q(\boldsymbol{s})$:

$$\langle f_{t,\star}\rangle=\langle(\boldsymbol{s}\circ\boldsymbol{e}_t)^\top\rangle\langle\boldsymbol{g}_{t,\star}\rangle+\langle b_t\rangle,$$

where we use $\langle f_{t,\star}\rangle$ to predict the class label by looking at its sign.

Baseline algorithm

We use a kernelized Bayesian classification algorithm, which is known as relevance vector machine [24], as the baseline algorithm. Its distributional assumptions are defined as

$$\begin{aligned}
\lambda_i &\sim \text{Gamma}\left(\lambda_i;\alpha_\lambda,\beta_\lambda\right) && \forall i\\
a_i|\lambda_i &\sim \text{Normal}\left(a_i;0,\lambda_i^{-1}\right) && \forall i\\
\gamma &\sim \text{Gamma}\left(\gamma;\alpha_\gamma,\beta_\gamma\right)\\
b|\gamma &\sim \text{Normal}\left(b;0,\gamma^{-1}\right)\\
f_i|\boldsymbol{a},b,\boldsymbol{k}_i &\sim \text{Normal}\left(f_i;\boldsymbol{a}^\top\boldsymbol{k}_i+b,1\right) && \forall i\\
y_i|f_i &\sim \text{Kronecker}\left(f_i y_i>\nu\right) && \forall i,
\end{aligned}$$

where the predicted outputs of data points are modeled as a linear function of their kernel representations (i.e. $\boldsymbol{a}^\top\boldsymbol{k}_i+b$). We again learn the posterior distribution over the model parameters and the latent variables using a deterministic variational approximation as we do for our methods. We call this algorithm 'Bayesian relevance vector machine' (BRVM). We have three main reasons for choosing this particular baseline algorithm: (i) BRVM can make use of kernel functions to obtain nonlinear models like our methods. (ii) We can see the effect of using gene set information by comparing our methods to BRVM. (iii) BRVM uses the same type of inference mechanism with our methods.

Results and discussion

To illustrate the effectiveness of our proposed methods SBMKL and SBMTMKL, we report their results on four datasets (i.e. two cancer and two tuberculosis datasets) and compare them to the baseline algorithm BRVM, which does not make use of gene set information, using repeated random subsampling validation experiments.

Experimental settings

For each dataset, we create 100 random train/test splits to obtain robust results. For each replication, the training set is defined by randomly selecting 75 % of the data points with stratification on the phenotype, and the remaining 25 % of the samples are used as the test set. The training set is normalized to have zero mean and unit standard deviation, and the test set is then normalized using the mean and the standard deviation of the original training set.

We extract gene sets from 'the Molecular Signatures Database' (MSigDB) [3], which contains curated pathway gene sets from online databases such as 'the Kyoto Encyclopedia of Genes and Genomes' (KEGG) [25] and 'the

Pathway Interaction Database' (PID) [26]. In our experiments, we use 196 PID pathways reported in MSigDB as our gene set collection.

To calculate similarity between data points for all methods, we use the Gaussian kernel:

$$k_{\text{Gaussian}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2/(2s^2)\right),$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm, and we set the kernel width s to the mean of pairwise Euclidean distances between the data points:

$$s = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2.$$

For BRVM, we calculate a single kernel over all input features. For SBMKL and SBMTMKL, we calculate a separate kernel function for each gene set over the corresponding features. Note that the Gaussian kernels calculated on the gene sets take values between 0 and 1 by definition, and there is no need for eliminating small/large gene sets or performing additional normalization steps to remove the effect of gene set size.

The hyper-parameter values of BRVM are selected as $(\alpha_\lambda, \beta_\lambda) = (1,1)$, $(\alpha_\gamma, \beta_\gamma) = (1,1)$ and $\nu = 1$. The hyper-parameter values of SBMKL and SBMTMKL are selected as $(\alpha_\lambda, \beta_\lambda) = (1,1), \sigma_g = 0.1, (\zeta_\kappa, \eta_\kappa) = (1,999), (\alpha_\omega, \beta_\omega) = (1,1), (\alpha_\gamma, \beta_\gamma) = (1,1)$ and $\nu = 1$. Note that $(\zeta_\kappa, \eta_\kappa)$ are set to these particular values to produce very sparse binary indicator variables, leading to classifiers with very few gene sets used for prediction. For BRVM, we perform 200 iterations during variational inference, whereas we perform 50 iterations for SBMKL and SBMTMKL.

We use 'area under the receiver operating characteristic curve' (AUROC) to compare classification results. AUROC is used to summarize the receiver operating characteristic curve, which is a curve of true positives as a function of false positives while the threshold to predict labels changes. Larger AUROC values correspond to better performance.

Classification results on the cancer datasets

On the cancer datasets, we run binary classification experiments to separate MSI-H patients from others (i.e. MSI-L and MSS), which is in agreement with the earlier studies that combine MSI-L and MSS tumors into the same group [11]. For BRVM and SBMKL methods, we train a separate classification model on each dataset, whereas, for SBMTMKL, we train a joint model on both datasets. Figure 3 compares the performance of BRVM, SBMKL and SBMTMKL on both datasets in terms of AUROC over 100 replications using box-and-whisker plots, and also reports the average AUROC value for each experiment. We clearly see that our methods with sparse gene set weights, leading to classifiers with very few active features, obtain results comparable to or even better than BRVM. Note that BRVM uses all available input features of the genomic data for classification. For example, SBMKL falls behind BRVM just by 0.1 % on COADREAD dataset, but obtains 2.9 % higher average AUROC on UCEC dataset. The average AUROC values become even higher if we model both datasets together using our multitask learning method SBMTMKL, which outperforms BRVM by 1.0 and 3.4 % on COADREAD and UCEC, respectively. Our sparse classifiers obtain these results using very few active features (i.e. features related to the genes in the gene sets with nonzero binary indicator variables); SBMKL uses 154.19 (3.40) and 403.03 (8.27) out of 20 530 (196) input features (gene sets) on the average, whereas SBMTMKL uses 484.03 (9.96) features (gene sets) on the average (i.e. less than 2.5 % of the input features) and obtains better classification results than BRVM and SBMKL on both datasets.

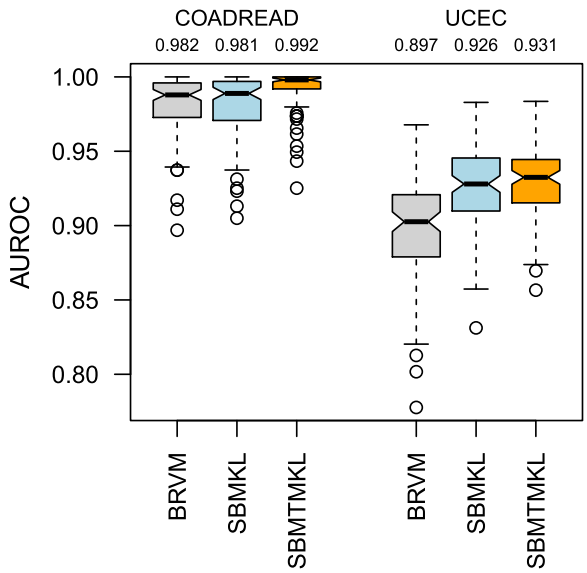

Fig. 3 AUROC values on the cancer datasets for MSI-H versus others classification. The *box-and-whisker plot* shows the results over 100 replications in repeated random subsampling validation experiments of BRVM, SBMKL and SBMTMKL on both datasets. The numbers just below the dataset names give the average AUROC value for each experiment

Classification results on the tuberculosis datasets

On the tuberculosis datasets, we perform binary classification experiments to separate individuals with ATB from others (i.e. individuals with LTBI or OD), which is critical in clinical settings because we should correctly identify individuals who need tuberculosis treatment [14]. Figure 4 compares the performance of BRVM, SBMKL

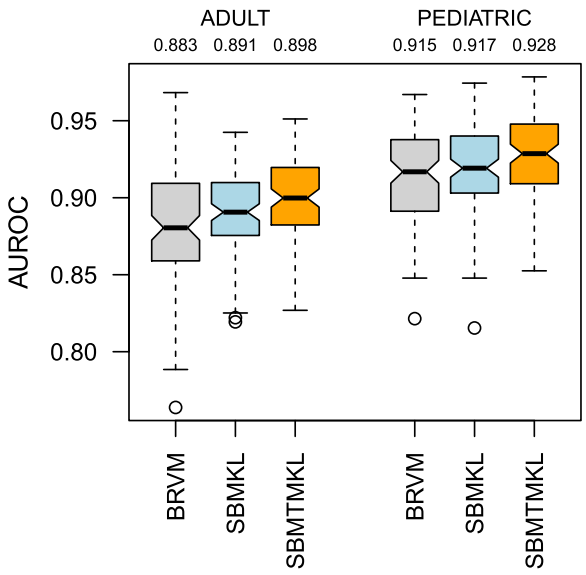

Fig. 4 AUROC values on the tuberculosis datasets for ATB versus others classification. The *box-and-whisker plots* show the results over 100 replications in repeated random subsampling validation experiments of BRVM, SBMKL and SBMTMKL on both datasets. The numbers just below the dataset names give the average AUROC value for each experiment

and SBMTMKL on both datasets. We see that our methods obtain results better than BRVM. On ADULT and PEDIATRIC datasets, SBMKL outperforms BRVM by 0.8 and 0.2 % using 782.21 (11.41) and 569.51 (7.88) out of 47, 323 (196) input features (gene sets) on the average, respectively. Our multitask learning method SBMTMKL again has the highest AUROC values on both datasets and outperforms BRVM by 1.5 % on ADULT and 1.3 % on PEDIATRIC using 1 102.65 (16.07) features (gene sets) on the average.

Biological results on the cancer datasets

To illustrate the biological relevance of our methods, we analyze their abilities to identify relevant gene sets based on the binary indicator variables inferred during training. For each gene set, we count the number of replications in which the corresponding binary indicator variable is nonzero. Table 3 lists the top 10 most frequently selected gene sets together with their selection frequencies for three scenarios: (i) SBMKL on COADREAD, (ii) SBMKL on UCEC and (iii) SBMTMKL on COADREAD and UCEC. We see that SBMKL is able to identify `WNT_NONCANONICAL_PATHWAY` and `TGFBRPATHWAY` as the top two gene sets in the first scenario, which are reported to be involved in the initiation and progression of colorectal cancer [12]. However, their selection frequencies are quite low (i.e. less than or equal to 0.10). Similarly, for UCEC, it is able to identify two apoptosis-related gene sets, namely, `P53DOWNSTREAMPATHWAY` and `NOTCH_PATHWAY`, as the top gene sets with more than 0.80 frequencies, which are known to be associated with endometrial cancer [13]. When we jointly model both datasets using our multitask learning method SBMTMKL, we are able to identify `P53DOWNSTREAMPATHWAY`, `NOTCH_PATHWAY` and `WNT_NONCANONICAL_PATHWAY` as the top gene sets with increased frequencies compared to those of SBMKL. We see that multitask learning decreases the effect of random subsampling by picking relevant gene sets more frequently, leading to more robust knowledge extraction for both datasets.

We also count the number of replications for each gene in which it is included in the final classifier. Figure 5 displays the top 50 most frequently selected genes together with their selection frequencies for three scenarios. CREBBP, EP300, JUN and MDM2 are among the top 50 genes for all scenarios, which is reasonable

Table 3 Gene set selection results on the cancer datasets for MSI-H versus others classification

SBMKL on COADREAD		SBMKL on UCEC		SBMTMKL on COADREAD and UCEC	
Gene set name	Frequency	Gene set name	Frequency	Gene set name	Frequency
`WNT_NONCANONICAL_PATHWAY`	0.10	`P53DOWNSTREAMPATHWAY`	0.92	`P53DOWNSTREAMPATHWAY`	0.99
`TGFBRPATHWAY`	0.09	`NOTCH_PATHWAY`	0.83	`NOTCH_PATHWAY`	0.92
`DELTANP63PATHWAY`	0.07	`NFAT_TFPATHWAY`	0.26	`WNT_NONCANONICAL_PATHWAY`	0.61
`TAP63PATHWAY`	0.07	`IL5_PATHWAY`	0.24	`NFAT_TFPATHWAY`	0.41
`RB_1PATHWAY`	0.07	`P53REGULATIONPATHWAY`	0.24	`AR_PATHWAY`	0.34
`NFAT_3PATHWAY`	0.06	`CDC42_REG_PATHWAY`	0.20	`RHOA_PATHWAY`	0.21
`ATF2_PATHWAY`	0.06	`AVB3_OPN_PATHWAY`	0.15	`REG_GR_PATHWAY`	0.21
`SMAD2_3NUCLEARPATHWAY`	0.05	`WNT_NONCANONICAL_PATHWAY`	0.14	`UPA_UPAR_PATHWAY`	0.17
`P73PATHWAY`	0.05	`REG_GR_PATHWAY`	0.13	`BMPPATHWAY`	0.17
`MYC_ACTIVPATHWAY`	0.05	`UPA_UPAR_PATHWAY`	0.11	`RAC1_PATHWAY`	0.14

The table displays the top 10 most frequently selected gene sets together with their selection frequencies for three scenarios

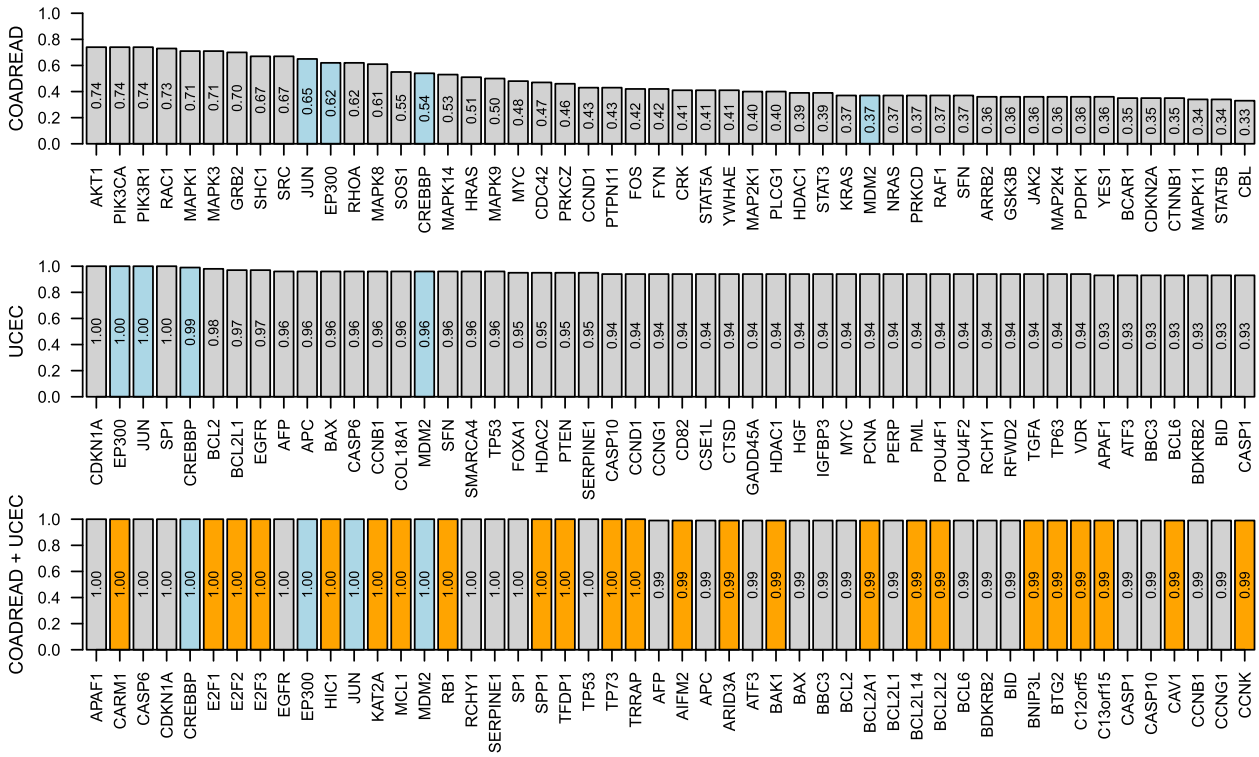

Fig. 5 Gene selection results on the cancer datasets for MSI-H versus others classification. The *bar plots* display the top 50 most frequently selected genes together with their selection frequencies for three scenarios. *Blue bars* show the genes that are in the top 50 for all scenarios, and *orange bars* show the genes that are in the top 50 only for multitask learning scenario

considering their functions in cell cycle. We see that the selection frequencies of the first two scenarios are lower than those of the third scenario, which is consistent with our gene set selection results. Our multitask learning method SBMTMKL includes several genes in the top 50 that are not selected by SBMKL in two other scenarios, which may lead to interesting findings. For example, E2F1, E2F2 and E2F3 are used in the final classifier in all replications, which are reported to be related to cellular proliferation [27].

Biological results on the tuberculosis datasets

We also evaluate the gene set selection results of our methods on the tuberculosis datasets. Table 4 lists the top

Table 4 Gene set selection results on the tuberculosis datasets for ATB versus others classification

SBMKL on ADULT		SBMKL on PEDIATRIC		SBMTMKL on ADULT and PEDIATRIC	
Pathway name	Frequency	Pathway name	Frequency	Pathway name	Frequency
ERBB_NETWORK_PATHWAY	0.73	A6B1_A6B4_INTEGRIN_PATHWAY	0.31	RHODOPSIN_PATHWAY	0.67
AP1_PATHWAY	0.55	RAS_PATHWAY	0.27	ERBB_NETWORK_PATHWAY	0.63
CONE_PATHWAY	0.44	INTEGRIN1_PATHWAY	0.24	AP1_PATHWAY	0.60
AR_TF_PATHWAY	0.42	RAC1_PATHWAY	0.21	SYNDECAN_1_PATHWAY	0.51
CERAMIDE_PATHWAY	0.31	RHODOPSIN_PATHWAY	0.20	PLK1_PATHWAY	0.50
RHODOPSIN_PATHWAY	0.31	SYNDECAN_1_PATHWAY	0.17	CERAMIDE_PATHWAY	0.42
SYNDECAN_1_PATHWAY	0.29	ATM_PATHWAY	0.16	ATM_PATHWAY	0.41
FANCONI_PATHWAY	0.25	ATF2_PATHWAY	0.15	AR_TF_PATHWAY	0.40
RXR_VDR_PATHWAY	0.24	THROMBIN_PAR1_PATHWAY	0.15	ATF2_PATHWAY	0.37
HNF3BPATHWAY	0.23	IL12_2PATHWAY	0.13	HNF3APATHWAY	0.35

The table displays the top 10 most frequently selected gene sets together with their selection frequencies for three scenarios

10 most frequently selected gene sets together with their selection frequencies for three scenarios: (i) SBMKL on ADULT, (ii) SBMKL on PEDIATRIC and (iii) SBMTMKL on ADULT and PEDIATRIC. We see that the gene set selection frequencies of SBMKL on PEDIATRIC dataset are quite low (i.e. between 0.13 and 0.31) compared to those on ADULT dataset. However, when we model both datasets using our multitask learning method SBMTMKL, the selection frequencies of the top 10 gene sets become significantly higher (i.e. between 0.35 and 0.67), leading to more robust gene set signatures.

Figure 6 displays the top 50 most frequently selected genes together with their selection frequencies for three scenarios. We see that the genes that are part of signaling mechanisms such as MAPK1, MAPK3, MAPK8, PIK3CA, PIK3R1 and RAC1 are selected in the top 50 genes for all scenarios. Similar to the results on the cancer datasets, the selection frequencies of the first two scenarios are lower than those of the third scenario, which shows the robustness of multitask learning approach. As an interesting finding, SBMTMKL includes three genes from interleukin family, namely, IL8, IL2 and IL6, in the top 50, which are shown to be diagnostically associated with tuberculosis [28–30], whereas they are not picked in the top 50 by SBMKL in single dataset experiments.

Conclusions

Integrating gene set analysis and predictive modeling is already considered by many existing studies, which fail either to capture nonlinear dependencies between genomic data and phenotype or to model multiple related datasets conjointly.

In this study, we integrate gene set analysis and nonlinear predictive modeling of disease phenotypes by casting this problem into a binary classification framework defined on the gene sets with a sparsity-inducing prior on their weights. To this aim, we propose a Bayesian multiple kernel learning algorithm, which produces a classifier with sparse gene set weights, by extending our earlier Bayesian formulation [8]. We then generalize this new algorithm to multitask learning to be able to model multiple related datasets conjointly, leading to better generalization performance and to more robust molecular signatures. The main novelty of our methods is the integration of gene set analysis and nonlinear predictive modeling using a probabilistic formulation, which enables us to robustly capture nonlinear dependencies between genomic data and phenotype even with small sample sizes, and to use overlapping gene sets and gene sets with different sizes without any major concern. Our approach brings us two side advantages: (i) We can identify very few gene sets

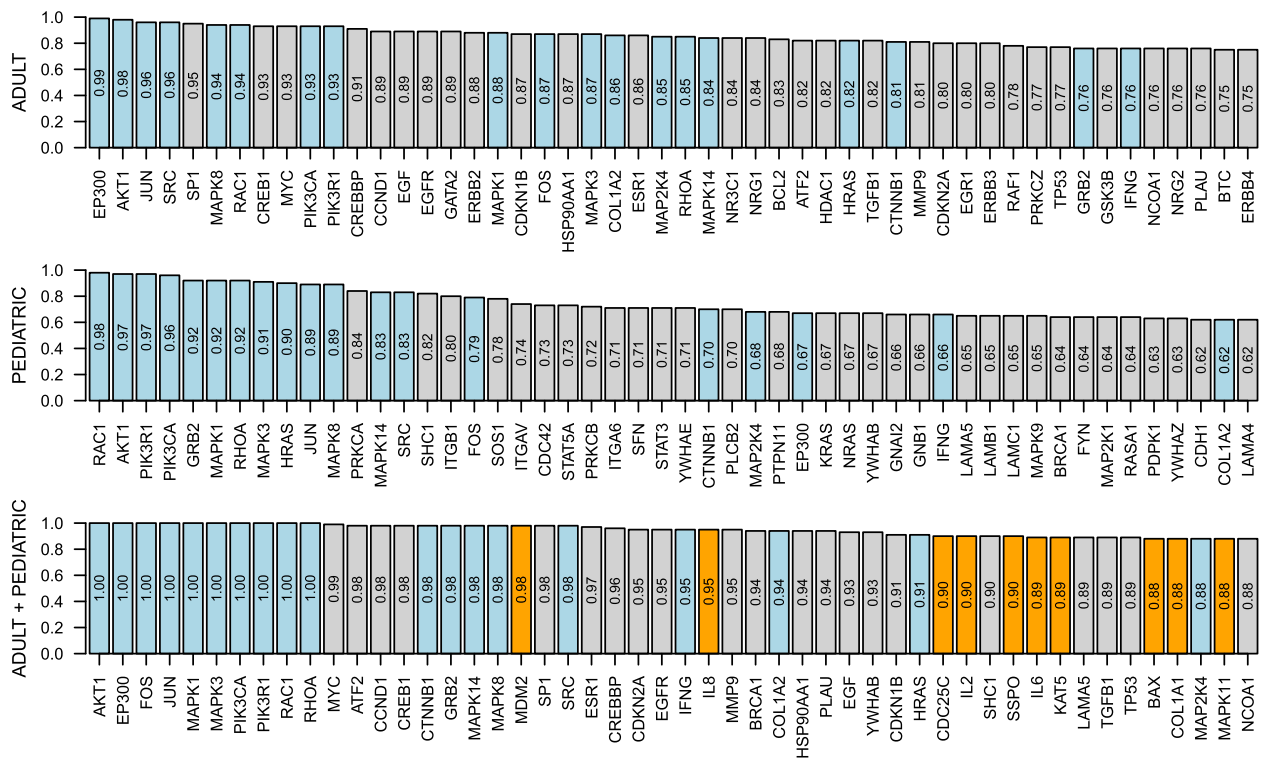

Fig. 6 Gene selection results on the tuberculosis datasets for ATB versus others classification. The *bar plots* display the top 50 most frequently selected genes together with their selection frequencies for three scenarios. *Blue bars* show the genes that are in the top 50 for all scenarios, and *orange bars* show the genes that are in the top 50 only for multitask learning scenario

predictive of the phenotype, which may shed light on underlying biological processes. (ii) We can reduce the data acquisition cost for test samples in clinical settings by collecting only the features used in our classifier.

To demonstrate the performance of our algorithms SBMKL and SBMTMKL, we perform repeated random subsampling validation experiments on four datasets of two major human diseases, namely, cancer and tuberculosis. On the two cancer datasets [12, 13], we decide whether a colorectal or endometrial tumor displays micro-satellite instability using its mRNA gene expression data. On the two tuberculosis datasets [14, 15], we diagnose whether an adult or pediatric individual has an active tuberculosis infection using his/her whole blood RNA expression data. We compare our two methods to a baseline Bayesian nonlinear algorithm that is trained on all available genomic data without using gene set information. Our methods obtain comparable or even better predictive performance using very few features (i.e. less than 2.5 % of the input features) on all datasets. We also show that we are able to identify biologically relevant genes and gene sets for cancer and tuberculosis phenotypes, which are validated by the existing studies from the literature. The results of our multitask learning algorithm show that modeling multiple related datasets conjointly enables us to further improve the generalization performance and to better understand biological processes behind disease phenotypes.

In the experiments reported, we use real-valued gene expression measurements as genomic data. Our methods can also be applied to discrete data such as mutation profiles of tumors, which are hard to use in classical gene set analysis methods due to their very sparse nature. As a possible extension, we plan to use our kernel-based formulations on cancer datasets to identify driver mutations using kernels for discrete data such as the Jaccard similarity coefficient.

Declarations
This article has been published as part of *BMC Bioinformatics* Volume 17 Supplement 16, 2016: Proceedings of the Tenth International Workshop on Machine Learning in Systems Biology (MLSB 2016). The full contents of the supplement are available online at http://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-16.

Funding
Publication of this article was funded by Koç University.

Availability of data and materials
All datasets used in this study were made publicly available previously by the corresponding data providers as mentioned in the manuscript. Matlab and R implementations of our two methods are available at https://github.com/mehmetgonen/sbmkl.

Authors' contributions
MG designed the study, implemented the algorithms, carried out the computational experiments, analyzed the results, and wrote the manuscript.

Competing interests
The author declares that he has no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Published: 13 December 2016

References

1. Ein-Dor L., et al. Outcome signature genes in breast cancer: Is there a unique set Bioinformatics. 2005;21:171–8.
2. Ein-Dor L., et al. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci USA. 2006;103:5923–8.
3. Subramanian A., et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102:15545–50.
4. Khatri P, et al. Ten years of pathway analysis: Current approaches and outstanding challenges. PLoS Comput Biol. 2012;8:e1002375.
5. Tai F, Pan W. Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. Bioinformatics. 2007a;23:1775–82.
6. Tai F, Pan W. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. Bioinformatics. 2007b;23:3170–7.
7. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008;24:1175–82.
8. Gönen M. Bayesian efficient multiple kernel learning. In: Proceedings of the 29th International Conference on Machine Learning. Edinburgh: Omnipress; 2012.
9. Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. J Amer Statist Assoc. 1988;83:1023–32.
10. Seoane JA, et al. A pathway-based data integration framework for prediction of disease progression. Bioinformatics. 2014;30:838–45.
11. Vilar E, Gruber SB. Microsatellite instability in colorectal cancer—the stable evidence. Nat Rev Clin Oncol. 2010;7:153–62.
12. The Cancer Genome Atlas Network Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7.
13. The Cancer Genome Atlas Research Network Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497:67–73.
14. Kaforou M, et al. Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: A case-control study. PLoS Med. 2013;10:e1001538.
15. Anderson ST, et al. Diagnosis of childhood tuberculosis and host RNA expression in Africa. N Engl J Med. 2014;370:1712–23.
16. Schölkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. 2002.
17. Schölkopf B, et al. Kernel Methods in Computational Biology. 2004.
18. Gönen M, Alpaydın E. Multiple kernel learning algorithms. J Mach Learn Res. 2011;12:2211–68.
19. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. J Amer Statist Assoc. 1993;88:669–79.
20. Lawrence ND, Jordan MI. Semi-supervised learning via Gaussian processes. Adv Neural Inf Process Syst. 2005;17:753–60.
21. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. J Amer Statist Assoc. 1990;85:398–409.
22. Jordan MI, et al. An introduction to variational methods for graphical models. Mach Learn. 1999;37:183–233.
23. Titsias MK, Lázaro-Gredilla M. Spike and slab variational inference for multi-task and multiple kernel learning. Adv Neural Inf Process Syst. 2011;24:2339–47.
24. Tipping ME. Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res. 2001;1:211–44.
25. Ogata H., et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 1999;27:29–34.
26. Schaefer CF, et al. PID: The Pathway Interaction Database. Nucleic Acids Res. 2009;37:D674–D9.
27. Timmers C, et al. E2f1, E2f2, and E2f3 control E2F target expression and cellular proliferation via a p53-dependent negative feedback loop. Mol Cell Biol. 2007;27:65–78.

28. Boggaram V, et al. Early secreted antigenic target of 6 kDa (ESAT-6) protein of Mycobacterium tuberculosis induces interleukin-8 (IL-8) expression in lung epithelial cells via protein kinase signaling and reactive oxygen species. J Biol Chem. 2013;288:25500–11.
29. Mamishi S, et al. Diagnostic accuracy of IL-2 for the diagnosis of latent tuberculosis: A systematic review and meta-analysis. Eur J Clin Microbiol Infect Dis. 2014;33:2111–9.
30. Martinez AN, et al. Role of interleukin 6 in innate immunity to Mycobacterium tuberculosis infection. J Infect Dis. 2013;207:1253–61.