Boluki *et al. BMC Bioinformatics* 2017, **18**(Suppl 14):552
DOI 10.1186/s12859-017-1893-4

BMC Bioinformatics

RESEARCH Open Access

Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors

Shahin Boluki[1*], Mohammad Shahrokh Esfahani[2], Xiaoning Qian[1] and Edward R Dougherty[1]

From The 14th Annual MCBIOS Conference
Little Rock, AR, USA. 23-25 March 2017

Abstract

Background: Phenotypic classification is problematic because small samples are ubiquitous; and, for these, use of prior knowledge is critical. If knowledge concerning the feature-label distribution – for instance, genetic pathways – is available, then it can be used in learning. Optimal Bayesian classification provides optimal classification under model uncertainty. It differs from classical Bayesian methods in which a classification model is assumed and prior distributions are placed on model parameters. With optimal Bayesian classification, uncertainty is treated directly on the feature-label distribution, which assures full utilization of prior knowledge and is guaranteed to outperform classical methods.

Results: The salient problem confronting optimal Bayesian classification is prior construction. In this paper, we propose a new prior construction methodology based on a general framework of constraints in the form of conditional probability statements. We call this prior the *maximal knowledge-driven information prior* (MKDIP). The new constraint framework is more flexible than our previous methods as it naturally handles the potential inconsistency in archived regulatory relationships and conditioning can be augmented by other knowledge, such as population statistics. We also extend the application of prior construction to a multinomial mixture model when labels are unknown, which often occurs in practice. The performance of the proposed methods is examined on two important pathway families, the mammalian cell-cycle and a set of p53-related pathways, and also on a publicly available gene expression dataset of non-small cell lung cancer when combined with the existing prior knowledge on relevant signaling pathways.

Conclusion: The new proposed general prior construction framework extends the prior construction methodology to a more flexible framework that results in better inference when proper prior knowledge exists. Moreover, the extension of optimal Bayesian classification to multinomial mixtures where data sets are both small and unlabeled, enables superior classifier design using small, unstructured data sets. We have demonstrated the effectiveness of our approach using pathway information and available knowledge of gene regulating functions; however, the underlying theory can be applied to a wide variety of knowledge types, and other applications when there are small samples.

Keywords: Optimal Bayesian classification, Prior construction, Biological pathways, Probabilistic Boolean networks

*Correspondence: s.boluki@tamu.edu
[1]Department of Electrical and Computer Engineering, Texas A&M University, MS3128 TAMU, 77843 College Station, TX, USA
Full list of author information is available at the end of the article

Background

Small samples are commonplace in phenotypic classification and, for these, prior knowledge is critical [1, 2]. If knowledge concerning the feature-label distribution is available, say, genetic pathways, then it can be used to design an optimal Bayesian classifier (OBC) for which uncertainty is treated directly on the feature-label distribution. As typical with Bayesian methods, the salient obstacle confronting OBC is prior construction. In this paper, we propose a new prior construction framework to incorporate gene regulatory knowledge via general types of constraints in the form of probability statements quantifying the probabilities of gene up- and down-regulation conditioned on the regulatory status of other genes. We extend the application of prior construction to a multinomial mixture model when labels are unknown, a key issue confronting the use of data arising from unplanned experiments in practice.

Regarding prior construction, E. T. Jaynes has remarked [3], "…there must exist a general formal theory of determination of priors by logical analysis of prior information – and that to develop it is today the top priority research problem of Bayesian theory". It is precisely this kind of formal structure that is presented in this paper. The formal structure involves a constrained optimization in which the constraints incorporate existing scientific knowledge augmented by slackness variables. The constraints tighten the prior distribution in accordance with prior knowledge, while at the same time avoiding inadvertent over restriction of the prior, an important consideration with small samples.

Subsequent to the introduction of Jeffreys' non-informative prior [4], there was a series of information-theoretic and statistical methods: Maximal data information priors (MDIP) [5], non-informative priors for integers [6], entropic priors [7], reference (non-informative) priors obtained through maximization of the missing information [8], and least-informative priors [9] (see also [10–12] and the references therein). The principle of maximum entropy can be seen as a method of constructing least-informative priors [13, 14], though it was first introduced in statistical mechanics for assigning probabilities. Except in the Jeffreys' prior, almost all the methods are based on optimization: max- or min-imizing an objective function, usually an information theoretic one. The least-informative prior in [9] is found among a restricted set of distributions, where the feasible region is a set of convex combinations of certain types of distributions. In [15], several non-informative and informative priors for different problems are found. All of these methods emphasize the separation of prior knowledge and observed sample data.

Although the methods above are appropriate tools for generating prior probabilities, they are quite general methodologies without targeting any specific type of prior information. In that regard, the problem of prior selection, in any Bayesian paradigm, is usually treated conventionally (even "subjectively") and independent of the real available prior knowledge and sample data.

Figure 1 shows a schematic view of the proposed mechanism for Bayesian operator design.

The *a priori* knowledge in the form of graphical models (e.g., Markov random fields) has been widely utilized in covariance matrix estimation in Gaussian graphical models. In these studies, using a given graphical model illustrating the interactions between variables, different problems have been addressed: e.g., constraints on the matrix structure [16, 17] or known independencies between variables [18, 19]. Nonetheless, these studies rely on a fundamental assumption: the given prior knowledge is complete and hence provides one single solution. However, in many applications including genomics, the given prior knowledge is uncertain, incomplete, and may be inconsistent. Therefore, instead of interpreting the prior knowledge as a single solution, e.g., a single deterministic covariance matrix, we aim at constructing a prior distribution on an uncertainty class.

In a different approach to prior knowledge, gene-gene relationships (pathway-based or protein-protein interaction (PPI) networks) are used to improve classification accuracy [20–26], consistency of biomarker discovery [27, 28], accuracy of identifying differentially expressed genes and regulatory target genes of a transcription factor [29–31], and targeted therapeutic strategies [32, 33]. The majority of these studies utilize gene expressions corresponding to sub-networks in PPI networks, for instance: mean or median of gene expression values in gene ontology network modules [20], probabilistic inference of pathway activity [24], and producing candidate sub-networks via a Markov clustering algorithm applied to high quality PPI networks [26, 34]. None of these methods incorporate the regulating mechanisms (activating or suppressing) into classification or feature-selection to the best of our knowledge.

The fundamental difference of the work presented in this paper is that we develop machinery to transform knowledge contained in biological signaling pathways to prior probabilities. We propose a general framework capable of incorporating any source of prior information by extending our previous prior construction methods [35–37]. We call the final prior distribution constructed via this framework, a *maximal knowledge-driven information prior* (MKDIP). The new MKDIP construction constitutes two steps: (1) Pairwise and functional information quantification: information in the biological pathways is quantified by an information theoretic formulation. (2) Objective-based Prior Selection: combining sample data and prior knowledge, we build an objective function, in which the expected mean log-likelihood is regularized by

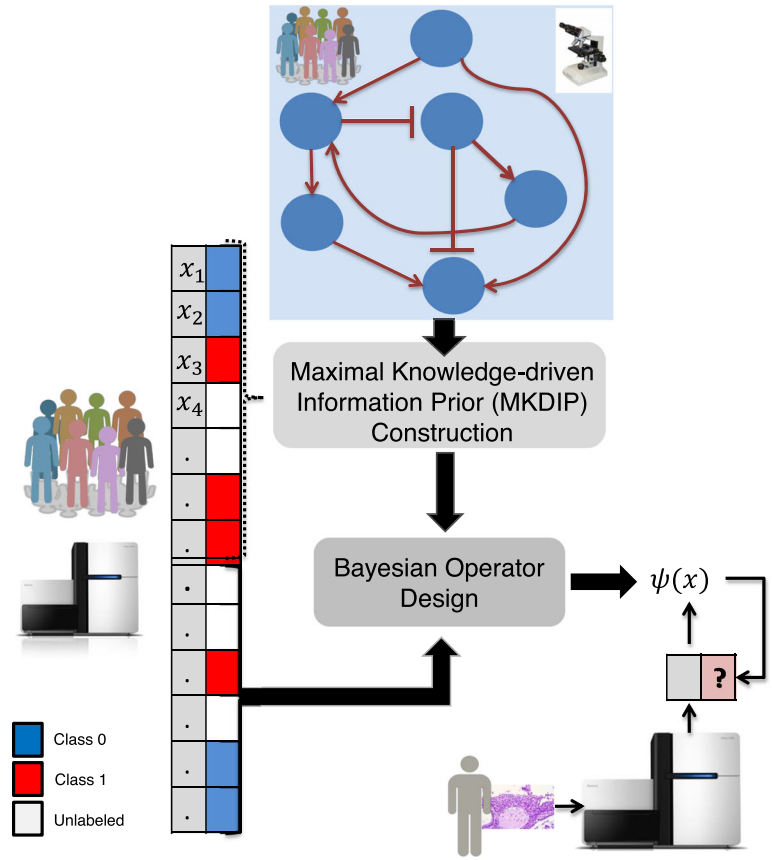

Fig. 1 A schematic illustration of the proposed Bayesian prior construction approach for a binary-classification problem. Information contained in the biological signaling pathways and their corresponding regulating functions is transformed to prior probabilities by MKDIP. Previously observed sample points (labeled or unlabeled) are used along with the constructed priors to design a Bayesian classifier to classify a new sample point (patient)

the quantified information in step 1. As a special case, where we do not have any sample data, or there is only one data point available for constructing the prior probability, the proposed framework is reduced to a regularized extension of the maximum entropy principle (MaxEnt) [38].

Owing to population heterogeneity we often face a *mixture model*, for example, when considering tumor sample heterogeneity where the assignment of a sample to any subtype or stage is not necessarily given. Thus, we derive the MKDIP construction and OBC for a mixture model. In this paper, we assume that data are categorical, e.g. binary or ternary gene-expression representations. Such categorical representations have many potential applications, including those wherein we only have access to a coarse set of measurements, e.g. epifluorescent imaging [39], rather than fine-resolution measurements such as microarray or RNA-Seq data. Finally, we emphasize that, in our framework, no single model is selected; instead, we consider all possible models as the uncertainty class that can be representative of the available prior information and assign probabilities to each model via the constructed prior.

Methods

Notation

Boldface lower case letters represent column vectors. Occasionally, concatenation of several vectors is also shown by boldface lower case letters. For a vector $\boldsymbol{a}$, a_0 represents the summation of all the elements and a_i denotes its i-th element. Probability sample spaces are shown by calligraphic uppercase letters. Uppercase letters are for sets and random variables (vectors). Probability measure over the random variable (vector) X is denoted by $P(X)$, whether it be a probability density function or a probability mass function. $E_X[f(X)]$ represents the expectation of $f(X)$ with respect to X. $P(\boldsymbol{x}|\boldsymbol{y})$ denotes the conditional probability $P(X = \boldsymbol{x}|Y = \boldsymbol{y})$. $\boldsymbol{\theta}$ represents generic parameters of a probability measure, for instance $P(X|Y;\boldsymbol{\theta})$ (or $P_{\boldsymbol{\theta}}(X|Y)$) is the conditional probability parameterized by $\boldsymbol{\theta}$. $\boldsymbol{\gamma}$ represents generic hyperparameter vectors. $\pi(\boldsymbol{\theta};\boldsymbol{\gamma})$ is the probability measure over the

parameters $\boldsymbol{\theta}$ governed by hyperparameters $\boldsymbol{\gamma}$, the parameters themselves governing another probability measure over some random variables. Throughout the paper, the terms "pathway" and "network" are used interchangeably. Also, the terms "feature"" and "variable" are used interchangeably. $\mathcal{M}ult(\boldsymbol{p}; n)$ and $\mathcal{D}(\boldsymbol{\alpha})$ represent a multinomial distribution with vector parameter $\boldsymbol{p}$ and n samples, and a Dirichlet distribution with vector $\boldsymbol{\alpha}$, respectively.

Review of optimal Bayesian classification

Binary classification involves a feature vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T \in \Re^d$ composed of random variables (features), a binary random variable (label) Y and a *classifier* $\psi(\mathbf{X})$ to predict Y. The error is $\varepsilon[\psi] = P(\psi(\mathbf{X}) \neq Y)$. An optimal classifier, ψ_{bay}, called a *Bayes classifier*, has minimal error, called the *Bayes error*, among all possible classifiers. The underlying probability model for classification is the joint feature-label distribution. It determines the class prior probabilities $c_0 = c = P(Y = 0)$ and $c_1 = 1 - c = P(Y = 1)$, and the class-conditional densities $f_0(\mathbf{x}) = P(\mathbf{x}|Y = 0)$ and $f_1(\mathbf{x}) = P(\mathbf{x}|Y = 1)$. A Bayes classifier is given by

$$\psi_{\text{bay}}(\mathbf{x}) = \begin{cases} 1, & c_1 f_1(\mathbf{x}) \geq c_0 f_0(\mathbf{x}), \\ 0, & \text{otherwise.} \end{cases} . \tag{1}$$

If the feature-label distribution is unknown but belongs to an uncertainty class of feature-label distributions parameterized by the vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, then, given a random sample S_n, an *optimal Bayeisan classifier* (OBC) minimizes the expected error over $\boldsymbol{\Theta}$:

$$\psi_{\text{OBC}} = \arg\min_{\psi \in \mathcal{C}} E_{\pi^*(\theta)}[\varepsilon_{\boldsymbol{\theta}}[\psi]], \tag{2}$$

where the expectation is relative to the posterior distribution $\pi^*(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$, which is derived from the prior distribution $\pi(\boldsymbol{\theta})$ using Bayes' rule [40, 41]. If we let $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ denote the class 0 and class 1 parameters, then we can write $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = [c, \boldsymbol{\theta_0}, \boldsymbol{\theta_1}]$. If we assume that $c, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$ are independent prior to observing the data, i.e. $\pi(\boldsymbol{\theta}) = \pi(c)\pi(\boldsymbol{\theta}_0)\pi(\boldsymbol{\theta}_1)$, then the independence is preserved in the posterior distribution $\pi^*(\boldsymbol{\theta}) = \pi^*(c)\pi^*(\boldsymbol{\theta}_0)\pi^*(\boldsymbol{\theta}_1)$ and the posteriors are given by $\pi^*(\boldsymbol{\theta}_y) \propto \pi(\boldsymbol{\theta}_y)\prod_{i=1}^{n_y} f_{\boldsymbol{\theta}_y}(\mathbf{x}_i^y|y)$ for $y = 0, 1$, where $f_{\boldsymbol{\theta}_y}(\mathbf{x}_i^y|y)$ and n_y are the class-conditional density and number of sample points for class y, respectively [42].

Given a classifier ψ_n designed from random sample S_n, from the perspective of mean-square error, the best error estimate minimizes the MSE between its true error (a function of $\boldsymbol{\theta}$ and ψ_n) and an error estimate (a function of S_n and ψ_n). This Bayesian minimum-mean-square-error (MMSE) estimate is given by the expected true error, $\widehat{\varepsilon}(\psi_n, S_n) = \mathrm{E}_{\boldsymbol{\theta}}[\varepsilon(\psi_n, \boldsymbol{\theta})|S_n]$, where $\varepsilon(\psi_n, \boldsymbol{\theta})$ is the error of ψ_n on the feature-label distribution parameterized by $\boldsymbol{\theta}$ and the expectation is taken relative to the prior distribution $\pi(\boldsymbol{\theta})$ [42]. The expectation given the sample is over the posterior probability. Thus, $\widehat{\varepsilon}(\psi_n, S_n) = \mathrm{E}_{\pi^*}[\varepsilon]$.

The *effective class-conditional density* for class y is defined by

$$f_{\boldsymbol{\Theta}}(\mathbf{x}|y) = \int_{\boldsymbol{\Theta}_y} f_{\boldsymbol{\theta}_y}(\mathbf{x}|y)\, \pi^*(\boldsymbol{\theta}_y)\, d\boldsymbol{\theta}_y, \tag{3}$$

$\boldsymbol{\Theta}_y$ being the space for $\boldsymbol{\theta}_y$, and an OBC is given pointwise by [40]

$$\psi_{\text{OBC}}(\mathbf{x}) = \begin{cases} 0 \text{ if } \mathrm{E}_{\pi^*}[c] f_{\boldsymbol{\Theta}}(\mathbf{x}|0) \geq (1 - \mathrm{E}_{\pi^*}[c]) f_{\boldsymbol{\Theta}}(\mathbf{x}|1), \\ 1 \text{ otherwise.} \end{cases} . \tag{4}$$

For discrete classification there is no loss in generality in assuming a single feature X taking values in the set $\{1, \ldots, b\}$ of "bins". Classification is determined by the class 0 prior probability c and the class-conditional probability mass functions $p_i = P(X = i|Y = 0)$ and $q_i = P(X = i|Y = 1)$, for $i = 1, \ldots, b$. With uncertainty, we assume beta class priors and define the parameters $\boldsymbol{\theta}_0 = \{p_1, p_2, \ldots, p_{b-1}\}$ and $\boldsymbol{\theta}_1 = \{q_1, q_2, \ldots, q_{b-1}\}$. The bin probabilities must be valid. Thus, $\{p_1, p_2, \ldots, p_{b-1}\} \in \boldsymbol{\Theta}_0$ if and only if $0 \leq p_i \leq 1$ for $i = 1, \ldots, b-1$ and $\sum_{i=1}^{b-1} p_i \leq 1$, in which case, $p_b = 1 - \sum_{i=1}^{b-1} p_i$. We use the Dirichlet priors

$$\pi(\boldsymbol{\theta}_0) \propto \prod_{i=1}^{b} p_i^{\alpha_i^0 - 1} \text{ and } \pi(\theta_1) \propto \prod_{i=1}^{b} q_i^{\alpha_i^1 - 1}, \tag{5}$$

where $\alpha_i^y > 0$. These are conjugate priors, leading to the posteriors of the same form. The effective class-conditional densities are

$$f_{\boldsymbol{\Theta}}(j|y) = \frac{U_j^y + \alpha_j^y}{n_y + \sum_{i=1}^{b} \alpha_i^y}, \tag{6}$$

for $y = 0, 1$, and the OBC is given by

$$\psi_{\text{OBC}}(j) = \begin{cases} 0, & \text{if } E_{\pi^*}[c] f_{\boldsymbol{\Theta}}(j|0) \geq (1 - E_{\pi^*}[c]) f_{\boldsymbol{\Theta}}(j|1); \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

where U_j^y denotes the observed count for class y in bin j [40]. Hereafter, $\sum_{i=1}^{b} \alpha_i^y$ is represented by α_0^y, i.e. $\alpha_0^y = \sum_{i=1}^{b} \alpha_i^y$, and is called the precision factor. In the sequel, the sub(super)-script relating to dependency on class y may be dropped; nonetheless, availability of prior knowledge for both classes is assumed.

Multinomial mixture model

In practice, data may not be labeled, due to potential tumor-tissue sample or stage heterogeneity, but still we want to classify a new sample point. A mixture model is a natural model for this scenario, assuming each sample

point $\mathbf{x}_i$ arises from a mixture of multinomial distributions:

$$P_{\boldsymbol{\theta}}(\mathbf{x}_i) = \sum_{j=0}^{M-1} c_j P_{\boldsymbol{\theta}_j}(\mathbf{x}_i), \tag{8}$$

where M is the number of components. When there exists two components, similar to binary classification, $M = 2$. The conjugate prior distribution family for component probabilities (if unknown) is the Dirichlet distribution. In the mixture model, no closed-form analytical posterior distribution for the parameters exists, but Markov chain Monte Carlo (MCMC) methods [43] can be employed to numerically calculate the posterior distributions. Since the conditional distributions can be calculated analytically in the multinomial mixture model, Gibbs sampling [44, 45] can be employed for the Bayesian inference. If the prior probability distribution over the component probability vector ($\boldsymbol{c} = [c_0, c_1, \ldots, c_M]$) is a Dirichlet distribution $\mathcal{D}(\boldsymbol{\phi})$ with parameter vector $\boldsymbol{\phi}$, the component-conditional probabilities are $\boldsymbol{\theta}_j = [p_1^j, p_2^j, \ldots, p_b^j]$, and the prior probability distribution over them is Dirichlet $\mathcal{D}(\boldsymbol{\alpha}^j)$ with parameter vector $\boldsymbol{\alpha}^j$ (as in the classification problem), for $j = 1, \ldots, M$, the Gibbs updates are

$$\begin{aligned}
y_i^{(t)} &\sim P\left(y_i = j | \boldsymbol{c}^{(t-1)}, \boldsymbol{\theta}^{(t-1)}, \mathbf{x}_i\right) \propto c_j^{(t-1)} p_{\mathbf{x}_i}^{j,(t-1)} \\
\boldsymbol{c}^{(t)} &\sim P\left(\boldsymbol{c} | \boldsymbol{\phi}, \boldsymbol{y}^{(t)}\right) = \mathcal{D}\left(\boldsymbol{\phi} + \sum\nolimits_{i=1}^{n} \left[I_{y_i^{(t)}=1}, \ldots, I_{y_i^{(t)}=M}\right]\right) \\
\boldsymbol{\theta}_{\boldsymbol{j}}^{(t)} &\sim P\left(\boldsymbol{\theta}_j | \mathbf{x}, \boldsymbol{y}^{(t)}, \boldsymbol{\alpha}_j\right) \\
&= \mathcal{D}\left(\boldsymbol{\alpha}_j + \sum\nolimits_{i=1:y_i^{(t)}=j}^{n} \left[I_{\mathbf{x}_i=1}, \ldots, I_{\mathbf{x}_i=b}\right]\right),
\end{aligned}$$

where the super-script in parentheses denotes the chain iteration number, I_w is one if w is true, and otherwise I_w is zero. In this framework, if the inference chain runs for Is iterations, then the numerical approximation of the OBC classification rule is

$$\psi_{\mathrm{OBC}}(k) \approx \arg\max_{y \in \{1,\ldots,M\}} \sum_{t=1}^{Is} c_y^{(t)} p_k^{y,(t)}. \tag{9}$$

Without loss of generality the summation above can be over the iterations of the chain considering burn-in and thinning.

Prior construction: general framework

In this section, we propose a general framework for prior construction. We begin with introducing a knowledge-driven prior probability:

Definition 1 *(Maximal Knowledge-driven Information Prior) If* Π *is a family of proper priors, then a maximal knowledge-driven information prior (MKDIP) is a solution to the following optimization problem:*

$$\arg\min_{\pi \in \Pi} E_\pi[\, C_{\boldsymbol{\theta}}(\xi, D)], \tag{10}$$

where $C_{\boldsymbol{\theta}}(\xi, D)$ *is a cost function that depends on* (1) $\boldsymbol{\theta}$*: the random vector parameterizing the underlying probability distribution,* (2) ξ*: state of (prior) knowledge, and* (3) D*: partial observation (part of the sample data).*

Alternatively, by parameterizing the prior probability as $\pi(\boldsymbol{\theta}; \boldsymbol{\gamma})$*, with* $\boldsymbol{\gamma} \in \Gamma$ *denoting the hyperparameters, an MKDIP can be found by solving*

$$\arg\min_{\boldsymbol{\gamma} \in \Gamma} E_{\pi(\boldsymbol{\theta};\boldsymbol{\gamma})}[\, C_{\boldsymbol{\theta}}(\xi, D, \boldsymbol{\gamma})]. \tag{11}$$

In contrast to non-informative priors, the MKDIP incorporates available prior knowledge and even *part* of the data to construct an informative prior.

The MKDIP definition is very general because we want a general framework for prior construction. The next definition specializes it to cost functions of a specific form in a constrained optimization.

Definition 2 *(MKDIP: Constrained Optimization with Additive Costs) As a special case in which* $C_{\boldsymbol{\theta}}$ *can be decomposed into additive terms, the cost function is of the form:*

$$C_{\boldsymbol{\theta}}(\xi, D, \boldsymbol{\gamma}) = (1 - \beta) g_{\boldsymbol{\theta}}^{(1)}(\xi, \boldsymbol{\gamma}) + \beta g_{\boldsymbol{\theta}}^{(2)}(\xi, D),$$

where β *is a non-negative regularization parameter. In this case, the MKDIP construction with additive costs and constraints involves solving the following optimization problem:*

$$\begin{aligned}
&\arg\min_{\boldsymbol{\gamma} \in \Gamma} E_{\pi(\boldsymbol{\theta};\boldsymbol{\gamma})}\left[(1 - \beta) g_{\boldsymbol{\theta}}^{(1)}(\xi, \boldsymbol{\gamma}) + \beta g_{\boldsymbol{\theta}}^{(2)}(\xi, D)\right] \\
&\textit{Subject to:} \quad E_{\pi(\boldsymbol{\theta};\boldsymbol{\gamma})}[g_{\boldsymbol{\theta},i}^{(3)}(\xi)] = 0;\ i \in \{1, \ldots, n_c\},
\end{aligned} \tag{12}$$

where $g_{\boldsymbol{\theta},i}^{(3)}$, $\forall i \in \{1, \ldots, n_c\}$*, are constraints resulting from the state of knowledge* ξ *via a mapping:*

$$\mathcal{T} : \xi \rightarrow E_{\pi(\boldsymbol{\theta};\boldsymbol{\gamma})}\left[g_{\boldsymbol{\theta},i}^{(3)}(\xi)\right], \forall i \in \{1, \ldots, n_c\}.$$

In the sequel, we will refer to $g^{(1)}(\cdot)$ and $g^{(2)}(\cdot)$ as the cost functions, and $g_i^{(3)}(\cdot)$'s as the knowledge-driven constraints. We begin with introducing information-theoretic cost functions, and then we propose a general set of mapping rules, denoted by $\mathcal{T}$ in Definition 2, to convert biological pathway knowledge into mathematical forms. We then consider special cases with information-theoretic cost functions.

Information-theoretic cost functions

Instead of having least squares (or mean-squared error) as the standard cost functions in classical statistical inference

problems, there is no universal cost function in the prior construction literature. That being said, in this paper, we utilize several widely used cost functions in the field:

1. (Maximum Entropy) The principle of maximum-entropy (MaxEnt) for probability construction [38] leads to the least informative prior given the constraints in order to prevent adding spurious information. Under our general framework MaxEnt can be formulated by setting:

$$\beta = 0,\ g_{\boldsymbol{\theta}}^{(1)} = -H[\,\boldsymbol{\theta}\,],$$

where $H[\,.]$ denotes the Shannon entropy.
2. (Maximal Data Information) The maximal data information prior (MDIP) introduced by Zellner [46] as a choice of the objective function is a criterion for the constructed probability distribution to remain maximally committed to the data [47]. To achieve MDIP, we can set our general framework with:

$$\begin{aligned}\beta = 0,\ g_{\boldsymbol{\theta}}^{(1)} &= \ln \pi(\boldsymbol{\theta};\boldsymbol{\gamma}) + H[\,P(x|\boldsymbol{\theta})]\\ &= \ln \pi(\boldsymbol{\theta};\boldsymbol{\gamma}) - E_{x|\boldsymbol{\theta}}[\,\ln P(x|\boldsymbol{\theta})]\,.\end{aligned}$$

3. (Expected Mean Log-likelihood) The cost function introduced in [35] is the first one that utilizes part of the observed data for prior construction. In that, we have

$$\beta = 1,\ g_{\boldsymbol{\theta}}^{(2)} = -\ell(\boldsymbol{\theta};D),$$

where $\ell(\boldsymbol{\theta};D) = \frac{1}{n_D}\sum_{i=1}^{n_D} \log f(\boldsymbol{x}_i|\boldsymbol{\theta})$ is the mean log-likelihood function of the sample points used for prior construction (D), and n_D denotes the number of sample points in D. In [35], it is shown that this cost function is equivalent to the average Kullback-Leibler distance between the *unknown* distribution (empirically estimated by some part of the samples) and the uncertainty class of distributions.

As originally proposed, the preceding approaches did not involve expectation over the uncertainty class. They were extended to the general prior construction form in Definition 1, including the expectation, in [36] to produce the regularized maximum entropy prior (RMEP), the regularized maximal data information prior (RMDIP), and the regularized expected mean log-likelihood prior (REMLP). In all cases, optimization was subject to specialized constraints.

For MKDIP, we employ the same information-theoretic cost functions in the prior construction optimization framework. MKDIP-E, MKDIP-D, and MKDIP-R correspond to using the same cost functions as REMP, RMDIP, and REMLP, respectively, but with the new general types of constraints. To wit, we employ *functional information* from the signaling pathways and show that by adding these new constraints that can be readily derived from prior knowledge, we can improve both supervised (classification problem with labelled data) and unsupervised (mixture problem without labels) learning of Bayesian operators.

From prior knowledge to mathematical constraints

In this part, we present a general formulation for mapping the existing knowledge into a set of *constraints*. In most scientific problems, the prior knowledge is in the form of conditional probabilities. In the following, we consider a hypothetical gene network and show how each component in a given network can be converted into the corresponding inequalities as general constraints in MKDIP optimization.

Before proceeding we would like to say something about contextual effects on regulation. Because a regulatory model is not independent of cellular activity outside the model, complete control relations such as $A \rightarrow B$ in the model, meaning that gene B is up-regulated if and only if gene A is up-regulated (after some time delay), do not necessarily translate into conditional probability statements of the form $P(X_B = 1|X_A = 1) = 1$, where X_A and X_B represent the binary gene values corresponding to genes A and B, respectively. Rather, what may be observed is $P(X_B = 1|X_A = 1) = 1 - \delta$, where $\delta > 0$. The pathway $A \rightarrow B$ need not imply $P(X_B = 1|X_A = 1) = 1$ because $A \rightarrow B$ is conditioned on the *context* of the cell, where by context we mean the overall state of the cell, not simply the activity being modeled. δ is called a *conditioning* parameter. In an analogous fashion, rather than $P(X_B = 1|X_A = 0) = 0$, what may be observed is $P(X_B = 1|X_A = 0) = \eta$, where $\eta > 0$, because there may be regulatory relations outside the model that up-regulate B. Such activity is referred to as cross-talk and η is called a *crosstalk* parameter. Conditioning and cross-talk effects can involve multiple genes and can be characterized analytically via context-dependent conditional probabilities [48].

Consider binary gene values $X_1, X_2, \ldots, X_m$ corresponding to genes $g_1, g_2, \ldots, g_m$. There are $m2^{m-1}$ conditional probabilities of the form

$$\begin{aligned}&P(X_i = k_i|X_1 = k_1, \ldots, X_{i-1} = k_{i-1}, X_{i+1} =\\ &\quad k_{i+1}, \ldots, X_m = k_m)\\ &= a_i^{k_i}(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m)\end{aligned} \tag{13}$$

to serve as constraints, the chosen constraints to be the conditional probabilities whose values are known (approximately). For instance, if g_2 and g_3 regulate g_1, with $X_1 = 1$ when $X_2 = 1$ and $X_3 = 0$, then, ignoring context effects,

$$a_1^1(1, 0, k_4, \ldots, k_m) = 1$$

for all $k_4, \ldots, k_m$. If, however, we take context conditioning into effect, then

$$a_1^1(1, 0, k_4, \ldots, k_m) = 1 - \delta_1(1, 0, k_4, \ldots, k_m),$$

where $\delta_1(1, 0, k_4, \ldots, k_m)$ is a conditioning parameter.

Moreover, ignoring context effects,

$$\begin{aligned} a_1^1(1, 1, k_4, \ldots, k_m) &= a_1^1(0, 0, k_4, \ldots, k_m) \\ &= a_1^1(0, 1, k_4, \ldots, k_m) = 0 \end{aligned}$$

for all $k_4, \ldots, k_m$. If, however, we take crosstalk into effect, then

$$\begin{aligned} a_1^1(1, 1, k_4, \ldots, k_m) &= \eta_1(1, 1, k_4, \ldots, k_m) \\ a_1^1(0, 0, k_4, \ldots, k_m) &= \eta_1(0, 0, k_4, \ldots, k_m) \\ a_1^1(0, 1, k_4, \ldots, k_m) &= \eta_1(0, 1, k_4, \ldots, k_m), \end{aligned}$$

where $\eta_1(1, 1, k_4, \ldots, k_m)$, $\eta_1(0, 0, k_4, \ldots, k_m)$, and $\eta_1(0, 0, k_4, \ldots, k_m)$ are crosstalk parameters. In practice it is unlikely that we would know the conditioning and crosstalk parameters for all combinations of $k_4, \ldots, k_m$; rather, we might just know the average, in which case, $\delta_1(1, 0, k_4, \ldots, k_m)$ reduces to $\delta_1(1, 0)$, $\eta_1(1, 1, k_4, \ldots, k_m)$ reduces to $\eta_1(1, 1)$, etc.

In this paradigm, the constraints resulting from our state of knowledge are of the following form:

$$\begin{aligned} &g_{\boldsymbol{\theta},i}^{(3)}(\xi) = \\ &P(X_i = k_i | X_1 = k_1, \ldots, X_{i-1} = k_{i-1}, X_{i+1} = k_{i+1}, \\ &\quad \ldots, X_m = k_m) - a_i^{k_i}(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m). \end{aligned} \tag{14}$$

The basic setting is very general and the conditional probabilities are what they are, whether or not they can be expressed in the regulatory form of conditioning or crosstalk parameters. The general scheme includes previous constraints and approaches proposed in [35] and [36] for the Gaussian and discrete setups. Moreover, in those we can drop the regulatory-set entropy because it is replaced by the set of conditional probabilities based on the regulatory set, whether forward (master predicting slaves) or backwards (slaves predicting masters) [48].

In this paradigm, the optimization constraints take the form

$$\begin{aligned} &a_i^{k_i}(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m) - \\ &\quad \varepsilon_i(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m) \\ &\leq E_{\pi(\boldsymbol{\theta};\boldsymbol{\gamma})}[\, P(X_i = k_i | X_1 = k_1, \ldots, X_{i-1} = k_{i-1}, \\ &\qquad X_{i+1} = k_{i+1}, \ldots, X_m = k_m)] \\ &\leq a_i^{k_i}(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m) + \\ &\quad \varepsilon_i(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m), \end{aligned} \tag{15}$$

where the expectation is with respect to the uncertainty in the model parameters, that is, the distribution of the model parameter $\boldsymbol{\theta}$, and ε_i is a slackness variable. Not all will be used, depending on our prior knowledge. In fact, the general conditional probabilities will not likely be used because they will likely not be known when there are too many conditioning variables. For instance, we may not know the probability in Eq. (13), but may know the conditioning on part of the variables which can be extracted from some interaction network (e.g. biological pathways). A slackness variable can be considered for each constraint to make the constraint framework more flexible, thereby allowing potential error or uncertainty in prior knowledge (allowing potential inconsistencies in prior knowledge). When using slackness variables, these variables also become optimization parameters, and a linear function (summation of all slackness variables) times a regulatory coefficient is added to the cost function of the optimization in Eq. (12). In other words, when having slackness variables, the optimization in Eq. (12) can be written as

$$\begin{aligned} &\arg \min_{\boldsymbol{\gamma} \in \Gamma, \boldsymbol{\varepsilon} \in \mathcal{E}} E_{\pi(\boldsymbol{\theta};\boldsymbol{\gamma})}\Big[\lambda_1[\,(1-\beta) g_{\boldsymbol{\theta}}^{(1)}(\xi, \boldsymbol{\gamma}) + \beta g_{\boldsymbol{\theta}}^{(2)}(\xi, D)] \\ &\qquad + \lambda_2 \sum_{i=1}^{n_c} \varepsilon_i\Big] \\ &\text{Subject to: } -\varepsilon_i \leq E_{\pi(\boldsymbol{\theta};\boldsymbol{\gamma})}[\, g_{\boldsymbol{\theta},i}^{(3)}(\xi)] \leq \varepsilon_i;\ i \in \{1, \ldots, n_c\}, \end{aligned} \tag{16}$$

where λ_1 and λ_2 are non-negative regularization parameters, and $\boldsymbol{\varepsilon}$ and $\mathcal{E}$ represent the vector of all slackness variables and the feasible region for slackness variables, respectively. For each slackness variable, a possible range can be defined (note that all slackness variables are non-negative). The higher the uncertainty is about a constraint stemming from prior knowledge, the greater the possible range for the corresponding slackness variable can be (more on this in the "Results and discussion" section).

The new general type of constraints discussed here introduces a formal procedure for incorporating prior knowledge. It allows the incorporation of knowledge of the functional regulations in the signaling pathways, any constraints on the conditional probabilities, and also knowledge of the cross-talk and conditioning parameters (if present), unlike the previous work in [36] where only partial information contained in the edges of the pathways is used in an ad hoc way.

An illustrative example and connection with conditional entropy

Now, consider a hypothetical network depicted in Fig. 2. For instance, suppose we know that the expression of gene g_1 is regulated by g_2, g_3, and g_5. Then we have

$$P(X_1 = 1 | X_2 = k_2, X_3 = k_3, X_5 = k_5) = a_1^1(k_2, k_3, k_5).$$

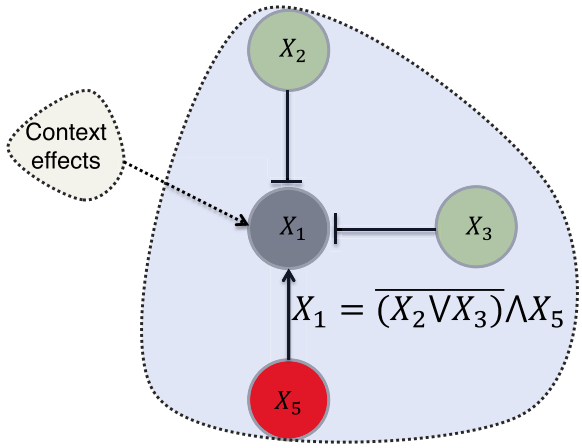

Fig. 2 An illustrative example showing the components directly connected to gene 1. In the Boolean function {AND, OR, NOT} = {∧, ∨, −}. Based on the regulating function of gene 1, it is up-regulated if gene 5 is up-regulated and genes 2 and 3 are down-regulated

As an example,

$$P(X_1 = 1|X_2 = 1, X_3 = 1, X_5 = 0) = a_1^1(1_2, 1_3, 0_5),$$

where the notation 1_2 denotes 1 for the second gene. Further, we might not know $a_1(k_2, k_3, k_5)$ for all combinations of k_2, k_3, k_5. Then we use the ones that we know. In the case of conditioning with g_2, g_3, and g_5 regulating g_1, with g_1 on if the others are on,

$$a_1^1(1_2, 1_3, 1_5) = 1 - \delta_1(1_2, 1_3, 1_5).$$

If limiting to 3-gene predictors, g_3, and g_5 regulate g_1, with g_1 on if the other two are on, then

$$a_1^1(k_2, 1_3, 1_5) = 1 - \delta_1(k_2, 1_3, 1_5),$$

meaning that the conditioning parameter depends on whether $X_2 = 0$ or 1.

Now, considering the conditional entropy, assuming that $\delta_1 = \max_{(k_2,k_3,k_5)} \delta_1(k_2, k_3, k_5)$ and $\delta_1 < 0.5$, we may write

$$\begin{aligned} &H[X_1|X_2, X_3, X_5] = \\ &-\Bigg[\sum_{\mathcal{X}_2,\mathcal{X}_3,\mathcal{X}_5} [P(X_1 = 0|X_2 = x_2, X_3 = x_3, X_5 = x_5) \\ &\quad \times P(X_2 = x_2, X_3 = x_3, X_5 = x_5) \\ &\quad \log[P(X_1 = 0|X_2 = x_2, X_3 = x_3, X_5 = x_5)] \\ &\quad + P(X_1 = 1|X_2 = x_2, X_3 = x_3, X_5 = x_5) \\ &\quad \times P(X_2 = x_2, X_3 = x_3, X_5 = x_5) \\ &\quad \log[P(X_1 = 1|X_2 = x_2, X_3 = x_3, X_5 = x_5)]\Bigg] \\ &\quad \leq h(\delta_1), \end{aligned}$$

where $h(\delta) = -[\delta \log(\delta) + (1 - \delta) \log(1 - \delta)]$. Hence, bounding the conditional probabilities, the conditional entropy is in turn bounded by $h(\delta_1)$:

$$\lim_{\delta_1 \to 0^+} H[X_1|X_2, X_3, X_5] = 0.$$

It should be noted that constraining $H[X_1|X_2, X_3, X_5]$ would not necessarily constrain the conditional probabilities, and may be considered as a more relaxed type of constraints. But, for example, in cases where there is no knowledge about the status of a gene given its regulator genes, constraining entropy is the only possible approach.

In our illustrative example, if we assume that the Boolean regulating function of X_1 is known as shown in Fig. 2 and context effects exist, then the following knowledge constraints can be extracted from the pathway and regulating function:

$$\begin{aligned} a_1^0(k_2, k_3, 0_5) &= 1 - \delta_1(k_2, k_3, 0_5) \\ a_1^0(k_2, 1_3, k_5) &= 1 - \delta_1(k_2, 1_3, k_5) \\ a_1^0(1_2, k_3, k_5) &= 1 - \delta_1(1_2, k_3, k_5) \\ a_1^1(0_2, 0_3, 1_5) &= 1 - \delta_1(0_2, 0_3, 1_5). \end{aligned}$$

Now if we assume that the context does not affect the value of X_1, i.e. the value of X_1 can be fully determined by knowing the values of X_2, X_3, and X_5, then we have the following equations:

$$a_1^0(k_2, k_3, 0_5) = P(X_1 = 0|X_5 = 0) = 1 \quad (17a)$$

$$a_1^0(k_2, 1_3, k_5) = P(X_1 = 0|X_3 = 1) = 1 \quad (17b)$$

$$a_1^0(1_2, k_3, k_5) = P(X_1 = 0|X_2 = 1) = 1 \quad (17c)$$

$$a_1^1(0_2, 0_3, 1_5) = P(X_1 = 1|X_2 = 0, X_3 = 0, X_5 = 1) = 1. \quad (17d)$$

It can be seen from the equations above that for some setups of the regulator values, only a subset of them determines the value of X_1, regardless of the other regulator values. If we assume that the value of X_5 cannot be observed, for example X_5 is an extracellular signal that cannot be measured in gene expression data and thereafter X_5 is not in the features of our data, the only constraints relevant to the feature-label distribution that can be extracted from the regulating function knowledge will be

$$\begin{aligned} a_1^0(k_2, 1_3, k_5) &= P(X_1 = 0|X_3 = 1) = 1 \\ a_1^0(1_2, k_3, k_5) &= P(X_1 = 0|X_2 = 1) = 1. \end{aligned} \quad (18)$$

Special case of Dirichlet distribution

Fixing the value of a single gene, being ON or OFF (i.e. $X_i = 0$ or $X_i = 1$, respectively), corresponds to a partition of the states, $\mathcal{X} = \{1, \ldots, b\}$. Here, the portions of $\mathcal{X}$ for which $(X_i = k_1, X_j = k_2)$ and $(X_i \neq k_1, X_j = k_2)$, for any $k_1, k_2 \in \{0, 1\}$, are denoted by $\mathcal{X}^{i,j}(k_1, k_2)$ and $\mathcal{X}^{i,j}(k_1^c, k_2)$, respectively. For the Dirichlet distribution, where $\boldsymbol{\theta} = \boldsymbol{p}$ and $\boldsymbol{\gamma} = \boldsymbol{\alpha}$, the constraints on the expectation over the conditional probability in (15) can be explicitly written as functions of the prior probability parameters (hyperparameters). For the parameter of the Dirichlet distribution, a vector $\boldsymbol{\alpha}$ indexed by $\mathcal{X}$, we denote the variable indicating the summation of its entities in $\mathcal{X}^{i,j}(k_1, k_2)$ by $\overline{\alpha}^{i,j}(k_1, k_2) = \sum_{k \in \mathcal{X}^{i,j}(k_1,k_2)} \alpha_k$. The notation can be easily extended for the cases having more than two fixed genes. In this setup, if the set of random variables corresponding to genes other than g_i and the vector of their corresponding values are shown by $\tilde{X}_i$ and $\tilde{x}_i$, respectively, the expectation over the conditional probability in (15) is [36]:

$$
\begin{aligned}
&E_{\boldsymbol{p}}\left[P\left(X_i = k_i | X_1 = k_1, \ldots, X_{i-1} = k_{i-1},\right.\right.\\
&\quad \left.\left. X_{i+1} = k_{i+1}, \ldots, X_m = k_m\right)\right]\\
&= E_{\boldsymbol{p}}\left[\frac{\sum_{k \in \mathcal{X}^{i,\tilde{X}_i}(k_i,\tilde{x}_i)} p_k}{\sum_{k \in \mathcal{X}^{i,\tilde{X}_i}(k_i,\tilde{x}_i)} p_k + \sum_{k \in \mathcal{X}^{i,\tilde{X}_i}(k_i^c,\tilde{x}_i)} p_k}\right]\\
&= \frac{\overline{\alpha}^{i,\tilde{X}_i}\left(k_i, \tilde{x}_i\right)}{\overline{\alpha}^{i,\tilde{X}_i}\left(k_i, \tilde{x}_i\right) + \overline{\alpha}^{i,\tilde{X}_i}\left(k_i^c, \tilde{x}_i\right)},
\end{aligned}
\tag{19}
$$

where the summation in the numerator and the first summation in the denominator of the second equality is over the states (bins) for which $(X_i = k_i, \tilde{X}_i = \tilde{x}_i)$, and the second summation in the denominator is over the states (bins) for which $(X_i = k_i^c, \tilde{X}_i = \tilde{x}_i)$.

If there exists a set of genes that completely determines the value of gene g_i (or only a specific setup of their values that determines the value, as we had in our illustrative example in Eq. (17)), then the constraints on the conditional probability conditioned on all the genes other than g_i can be changed to be conditioned on that set only. Specifically, let $\boldsymbol{R}_i$ denote the set of random variables corresponding to such a set of genes/proteins and suppose there exists a specific setup of their values $\boldsymbol{r}_i$ that completely determines the value of gene g_i. If the set of all random variables corresponding to the genes/proteins other than X_i and $\boldsymbol{R}_i$ is denoted by $\boldsymbol{B}_i = \tilde{X}_{(i,\boldsymbol{R}_i)}$, and their corresponding values by $\boldsymbol{b}_i$, then the constraints on the conditional probability can be written as

$$
\begin{aligned}
&E_{\boldsymbol{p}}\left[P\left(X_i = k_i | \boldsymbol{R}_i = \boldsymbol{r}_i\right)\right]\\
&= E_{\boldsymbol{p}}\left[\frac{\sum_{\boldsymbol{b}_i \in O_{\boldsymbol{B}_i}} \sum_{k \in \mathcal{X}^{i,\boldsymbol{R}_i,\boldsymbol{B}_i}(k_i,\boldsymbol{r}_i,\boldsymbol{b}_i)} p_k}{\sum_{\boldsymbol{b}_i \in O_{\boldsymbol{B}_i}} \sum_{k \in \mathcal{X}^{i,\boldsymbol{R}_i,\boldsymbol{B}_i}(k_i,\boldsymbol{r}_i,\boldsymbol{b}_i)} p_k}\right.\\
&\qquad \left.\frac{}{+ \sum_{\boldsymbol{b}_i \in O_{\boldsymbol{B}_i}} \sum_{k \in \mathcal{X}^{i,\boldsymbol{R}_i,\boldsymbol{B}_i}(k_i^c,\boldsymbol{r}_i,\boldsymbol{b}_i)} p_k}\right]\\
&= \frac{\sum_{\boldsymbol{b}_i \in O_{\boldsymbol{B}_i}} \overline{\alpha}^{i,\boldsymbol{R}_i,\boldsymbol{B}_i}(k_i, \boldsymbol{r}_i, \boldsymbol{b}_i)}{\sum_{\boldsymbol{b}_i \in O_{\boldsymbol{B}_i}} \overline{\alpha}^{i,\boldsymbol{R}_i,\boldsymbol{B}_i}(k_i, \boldsymbol{r}_i, \boldsymbol{b}_i)}\\
&\qquad \frac{}{+ \sum_{\boldsymbol{b}_i \in O_{\boldsymbol{B}_i}} \overline{\alpha}^{i,\boldsymbol{R}_i,\boldsymbol{B}_i}(k_i^c, \boldsymbol{r}_i, \boldsymbol{b}_i)},
\end{aligned}
\tag{20}
$$

where $O_{\boldsymbol{B}_i}$ is the set of all possible vectors of values for $\boldsymbol{B}_i$.

For a multinomial model with a Dirichlet prior distribution, a constraint on the conditional probabilities translates into a constraint on the above expectation over the conditional probabilities (as in Eq. (15)). In our illustrative example and from the equations in Eq. (17), there are four of these constraints on the conditional probability for gene g_1. For example, in the second constraint from the second line of Eq. (17) (Eq. 17b), $X_i = X_1$, $k_i = 0$, $\boldsymbol{R}_i = \{X_3\}$, $\boldsymbol{r}_i = [0]$, and $\boldsymbol{B}_i = \{X_2, X_5\}$. One might have several constraints for each gene extracted from its regulatory function (more on extracting general constraints from regulating functions in the "Results and discussion" section).

Results and discussion

The performance of the proposed general prior construction framework with different types of objective functions and constraints is examined and compared with other methods on two pathways, a mammalian cell-cycle pathway and a pathway involving the gene TP53. Here we employ Boolean network modeling of genes/proteins (hereafter referred to as entities or nodes) [49] with perturbation (BNp). A Boolean Network with p nodes (genes/proteins) is defined as $B = (V, F)$, where V represents the set of entities (genes/proteins) $\{v_1, \ldots, v_p\}$, and F is the set of Boolean predictor functions $\{f_1, \ldots, f_p\}$. At each step in a BNp, a decision is made by a Bernoulli random variable with the success probability equal to the perturbation probability, p_{pert}, as to whether a node value is determined by perturbation of randomly flipping its value or by the logic model imposed from the interactions in the signaling pathways. A BNp with a positive perturbation probability can be modeled by an ergodic Markov chain, and possesses a steady-state distribution (SSD) [50]. The performance of different prior construction methods can be compared based on the expected true error of the optimal Bayesian classifiers designed with those priors, and also by comparing these errors with some other well

known classification methods. Another comparison metric of prior construction methods is the expected norm of the difference between the true parameters and the posterior mean of these parameters inferred using the constructed prior distributions. Here, the true parameters are the vectors of the true class-conditional SSDs, i.e. the vectors of the true class-conditional bin probabilities of the BNp.

Moreover, the performance of the proposed framework is compared with other methods on a publicly available gene expression dataset of non-small cell lung cancer when combined with the existing prior knowledge on relevant signaling pathways.

Mammalian cell cycle classification

A Boolean logic regulatory network for the dynamical behavior of the cell cycle of normal mammalian cells is proposed in [51]. Figure 3(a) shows the corresponding pathways. In normal cells, cell division is coordinated via extracellular signals controlling the activation of CycD. Rb is a tumor suppressor gene and is expressed when the inhibitor cyclins are not present. Expression of p27 blocks the action of CycE or CycA, and lets the tumor-suppressor gene Rb be expressed even in the presence of CycE and CycA, and results in a stop in the cell cycle. Therefore, in the wild-type cell-cycle network, expressing p27 lets the cell cycle stop. But following the proposed mutation in [51], for the mutated case, p27 is always inactive (i.e. can never be activated), thereby creating a situation where both CycD and Rb might be inactive and the cell can cycle in the absence of any growth factor.

The full functional regulations in the cell-cycle Boolean network are shown in Table 1.

Following [36], for the binary classification problem, $y = 0$ corresponds to the normal system functioning based on Table 1, and $y = 1$ corresponds to the mutated (cancerous) system where CycD, p27, and Rb are permanently down-regulated (are stuck at zero), which creates a situation where the cell cycles even in the absence of any growth factor. The perturbation probability is set to 0.01 and 0.05 for the normal and mutated system, respectively. A BNp has a transition probability matrix (TPM), and as mentioned earlier, with positive perturbation probability can be modeled by an ergodic Markov chain, and possesses a SSD [50]. Here, each class has a vector of steady-state bin probabilities, resulting from the regulating functions of its corresponding BNp and the perturbation probability. The constructed SSDs are further marginalized to a subset of seven genes to prevent trivial classification scenarios. The final feature vector is $\mathbf{x} = [\text{E2F}, \text{CycE}, \text{CycA}, \text{Cdc20}, \text{Cdh1}, \text{UbcH10}, \text{CycB}]$, and the state space size is $2^7 = 128$. The true parameters for each

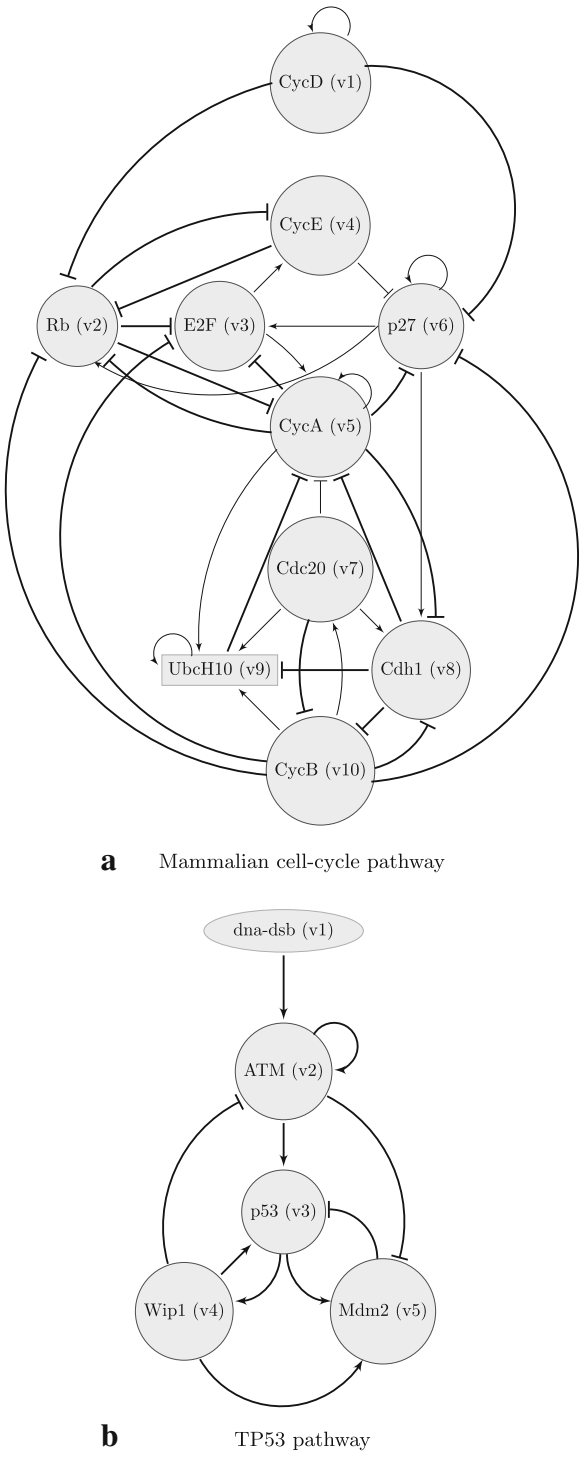

Fig. 3 Signaling pathways corresponding to Tables 1 and 2. Signaling pathways for: 3(a) the normal mammalian cell cycle (corresponding to Table 1) and 3(b) a simplified pathway involving TP53 (corresponding to Table 2)

Table 1 Boolean regulating functions of normal mammalian cell cycle [51]. In the Boolean functions {AND, OR, NOT} = {∧, ∨, −}

Gene	Node name	Boolean regulating function
CycD	v_1	Extracellular signal
Rb	v_2	$(\overline{v_1} \wedge \overline{v_4} \wedge \overline{v_5} \wedge \overline{v_{10}}) \vee (v_6 \wedge \overline{v_1} \wedge \overline{v_{10}})$
E2F	v_3	$(\overline{v_2} \wedge \overline{v_5} \wedge \overline{v_{10}}) \vee (v_6 \wedge \overline{v_2} \wedge \overline{v_{10}})$
CycE	v_4	$(v_3 \wedge \overline{v_2})$
CycA	v_5	$(v_3 \wedge \overline{v_2} \wedge \overline{v_7} \wedge \overline{(\overline{v_8} \wedge v_9)}) \vee (v_5 \wedge \overline{v_2} \wedge \overline{v_7} \wedge \overline{(\overline{v_8} \wedge v_9)})$
p27	v_6	$(\overline{v_1} \wedge \overline{v_4} \wedge \overline{v_5} \wedge \overline{v_{10}}) \vee (v_6 \wedge \overline{(v_4 \wedge v_5)} \wedge \overline{v_{10}} \wedge \overline{v_1})$
Cdc20	v_7	v_{10}
Cdh1	v_8	$(\overline{v_5} \wedge \overline{v_{10}}) \vee (v_7) \vee (v_6 \wedge \overline{v_{10}})$
UbcH10	v_9	$(\overline{v_8}) \vee (v_8 \wedge v_9 \wedge (v_7 \vee v_5 \vee v_{10}))$
CycB	v_{10}	$(\overline{v_7} \wedge \overline{v_8})$

class are the final class-conditional steady-state bin probabilities, $\boldsymbol{p}^0$ and $\boldsymbol{p}^1$ for the normal and mutated systems, respectively, which are utilized for taking samples.

Classification problem corresponding to TP53

TP53 is a tumor suppressor gene involved in various biological pathways [36]. Mutated p53 has been observed in almost half of the common human cancers [52], and in more than 90% of patients with severe ovarian cancer [53]. A simplified pathway involving TP53, based on logic in [54], is shown in Fig. 3(b). DNA double-strand break affects the operation of these pathways, and the Boolean network modeling of these pathways under this uncertainty has been studied in [53, 54]. The full functional regulations are shown in Table 2.

Following [36], two scenarios, dna-dsb=0 and dna-dsb=1, weighted by 0.95 and 0.05, are considered and the SSD of the normal system is constructed based on the ergodic Markov chain model of the BNp with the regulating functions in Table 2 by assuming the perturbation probability 0.01. The SSD for the mutated (cancerous) case is constructed by assuming a permanent down regulation of TP53 in the BNp, and perturbation probability 0.05. Knowing that dna-dsb is not measurable, and to avoid trivial classification situations, the SSDs are marginalized to a subset of three entities $\mathbf{x} =$ [ATM, Wip1, Mdm2]. The state space size in this case is $2^3 = 8$. The true parameters for each class are the final class-conditional steady-state bin probabilities, $\boldsymbol{p}^0$ and $\boldsymbol{p}^1$ for the normal and mutated systems, respectively, which are used for data generation.

Table 2 Boolean regulating functions corresponding to the pathway in Fig. 3(b) [54]. In the Boolean functions {AND, OR, NOT} = {∧, ∨, −}

Gene	Node name	Boolean regulating function
dna − dsb	v_1	Extracellular signal
ATM	v_2	$\overline{v_4} \wedge (v_2 \vee v_1)$
P53	v_3	$\overline{v_5} \wedge (v_2 \vee v_4)$
Wip1	v_4	v_3
Mdm2	v_5	$\overline{v_2} \wedge (v_3 \vee v_4)$

Extracting general constraints from regulating functions

If knowledge of the regulating functions exists, it can be used in the general constraint framework of the MKDIP, i.e. it can be used to constrain the conditional probabilities. In other words, the knowledge about the regulating function of gene i can be used to set $\varepsilon_i(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m)$, and $a_i^{k_i}(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m)$ in the general form of constraints in (15). If the true regulating function of gene i is known, and it is not context sensitive, then the conditional probability of its status, $a_i^{k_i}(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m)$, is known for sure, and $\delta_i(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m) = 0$. But in reality, the true regulating functions are not known, and are also context sensitive. The dependence on the context translates into $\delta_i(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m)$ being greater than zero. The greater the context effect on the gene status, the larger δ_i is. Moreover, the uncertainty over the regulating function is captured by the slackness variables $\varepsilon_i(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m)$ in Eq. (15). In other words, the uncertainty is translated to the possible range of the slackness variable values in the prior construction optimization framework. The higher the uncertainty is, the greater the range should be in the optimization framework. In fact, slackness variables make the whole constraint framework consistent, even for cases where the conditional probability constraints imposed by prior knowledge are not completely in line with each other, and guarantee the existence of a solution.

As an example, for the classification problems of the mammalian cell-cycle network and the TP53 network, assuming the regulating functions in Tables 1 and 2 are the true regulating functions, the context effect can be observed in the dependence of the output of the Boolean regulating functions in the tables on the extracellular signals, non-measurable entities, and the genes that have been marginalized out in our setup. In the absence of quantitative knowledge about the context effect, i.e. $a_i^{k_i}(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m)$ for all possible setups of the regulator values, one can impose only those with such knowledge. For example, in the mammalian cell-cycle network, CycB's regulating function only depends on the values included in the observed feature set; therefore the conditional probabilities are known for all regulator value setups. But for CycE the regulating function depends on Rb, which is marginalized out in our feature set, and also itself depends on an extracellular signal. Hence, the conditional probability constraints for CycE are known only

for the setup of the features that determine the output of the Boolean regulating function independent of the other regulator values.

In our comparison analysis, $a_i^{k_i}(k_1, \ldots, k_{i-1}, k_{i+1}, \ldots, k_m)$ for each gene/protein in Eq. (15) is set to one for the feature value setups that determine the Boolean regulating output regardless of the context. But since the observed data are not fully described by these functions, and the system has uncertainty, we let the possible range for the slackness variables in Eq. (15) be $[0, 1)$.

We now continue the examples on two of the mammalian cell-cycle network nodes, CycB and CycE. For CycB the following constraints on the prior distribution are extracted from its regulating function:

$$E_{\boldsymbol{p}}[P(\nu_{10} = 0|\nu_8 = 1)] \geq 1 - \epsilon_1$$
$$E_{\boldsymbol{p}}[P(\nu_{10} = 0|\nu_7 = 1)] \geq 1 - \epsilon_2$$
$$E_{\boldsymbol{p}}[P(\nu_{10} = 1|\nu_7 = 0, \nu_8 = 0)] \geq 1 - \epsilon_3.$$

For CycE, one of its regulators is Rb (ν_2), which is not included in the feature set, i.e. not observed, but is known to be down-regulated in the mutated (cancerous) case. Thus, the set of constraints extracted from the regulating function of CycE for the normal case includes only

$$E_{\boldsymbol{p}}[P(\nu_4 = 0|\nu_3 = 0)] \geq 1 - \epsilon_1$$

and for the mutated case consists of

$$E_{\boldsymbol{p}}[P(\nu_4 = 0|\nu_3 = 0)] \geq 1 - \epsilon_1$$
$$E_{\boldsymbol{p}}[P(\nu_4 = 1|\nu_3 = 1)] \geq 1 - \epsilon_2.$$

As another example, for the TP53 network, the set of constraints extracted from the regulating functions in Table 2 for the normal case are shown in the left panel of Table 3.

The first and second constraints for MKDIP in the left panel of Table 3 come from the regulating function of ν_2 in Table 2. Although ν_1 is an extracellular signal, the value of ν_4 imposes two constraints on the value of ν_2. But the regulating function of ν_4 in Table 2 only depends on ν_3, which is not included in our feature set, so we have no imposed constraints on the conditional probability from its regulating function. The other two constraints for MKDIP in the left panel of Table 3 are extracted from the regulating function of ν_5 in Table 2. Although ν_3 is not included in the observed features, for two setups of its regulators, ($\nu_2 = 1$) and ($\nu_2 = 0, \nu_4 = 1$), the value of ν_5 can be determined, so the constraint is imposed on the prior distribution from the regulating function. For comparison, the constraints extracted from the pathway in Fig. 3(b) based on the method of [36] are provided in the right panel of Table 3.

Performance comparison in classification setup

For both the mammalian cell cycle and TP53 problems, the performance of 11 methods are compared for classification performance. OBC with the Jeffreys' prior, OBC with our previous prior construction methods in [36] (RMEP, RMDIP, REMLP), OBC with our proposed general framework of constraints (MKDIP-E, MKDIP-D, MKDIP-R), and also well known methods including Histogram rule (Hist), CART [55], Random Forest (RF)[56], and Support Vector Machine classification (SVM) [57, 58]. Also, for all the Bayesian methods using OBC, the posterior mean of the parameters' distance from the true parameters is calculated and compared. The samples from the true distributions are stratified fixing two different class prior probabilities. Following [36], we assume that $\max_i p_i^{y,true}$, *for* $y \in \{0, 1\}$, is known within a $+/-$ 5% interval (can come from existing population statistics in practice). Two simulation scenarios are performed: one assuming the complete knowledge of the optimal precision factors [36] $\alpha_0^y = \sum_{i=1}^b \alpha_i^y, y \in \{0, 1\}$ for prior construction methods (oracle precision factor); and the other estimating the optimal precision factor from the observed data itself. Two class prior probabilities, $c = 0.6$ and $c = 0.5$, are considered. Along with the true class-conditional SSDs of the two classes, the corresponding Bayes error corresponds to the best performance that any classification rule for that classification problem (feature-label distribution) can yield. Fixing c and the true class-conditional bin probabilities, n sample

Table 3 The set of constraints extracted from the regulating functions and pathways for the TP53 network. Constraints extracted from the Boolean regulating functions in Table 2 corresponding to the pathway in Fig. 3(b) used in MKDIP-E, MKDIP-D, MKDIP-R (left). Constraints extracted based on [36] from the pathway in Fig. 3(b) used in RMEP, RMDIP, REMLP (right)

(a) MKDIP Constraints		(b) Constraints in Methods of [36]	
Node	Constraint	Node	Constraint
ν_2	$E_{\boldsymbol{p}}[P(\nu_2 = 0\|\nu_4 = 1)] \geq 1 - \epsilon_1$	ν_2	$E_{\boldsymbol{p}}[P(\nu_2 = 0\|\nu_4 = 1)] \geq 1 - \epsilon_1$
ν_2	$E_{\boldsymbol{p}}[P(\nu_2 = 1\|\nu_4 = 0)] \geq 1 - \epsilon_2$	ν_5	$E_{\boldsymbol{p}}[P(\nu_5 = 1\|\nu_2 = 0, \nu_4 = 1)] \geq 1 - \epsilon_2$
ν_5	$E_{\boldsymbol{p}}[P(\nu_5 = 0\|\nu_2 = 1)] \geq 1 - \epsilon_3$		
ν_5	$E_{\boldsymbol{p}}[P(\nu_5 = 1\|\nu_2 = 0, \nu_4 = 1)] \geq 1 - \epsilon_4$		

points by stratified sampling ($n_0 = \lceil cn \rceil$ sample points from class 0 and $n_1 = n - n_0$ sample points from class 1) are taken for prior construction (if used by the method), classifier training, and posterior distribution calculations. Then the designed classifier's true classification error is calculated for all classification methods. The posterior mean of parameter distance from the true parameter (true steady-state bin probabilities vector) is calculated based on $\sum_{y=0}^{1} ||\boldsymbol{\alpha}^{y*}/\alpha_0^{y*} - \boldsymbol{p}^y||^2$, where $\boldsymbol{\alpha}^{y*}$ and $\boldsymbol{p}^y$ represent the parameters of the posterior distribution and true bin probabilities vector for class y, respectively. For each fixed c and n, 800 Monte Carlo repetitions are done to calculate the expected classification errors and posterior distances from the true parameters for each parameter setup. For REMLP and MKDIP-R, which use a fraction of data in their prior construction procedure, 10 data points from each class are used for prior construction, and all for the inference and posterior calculation (here the number of data points used for prior construction is not fine-tuned, but a small number is chosen to avoid overfitting). The overall procedure taken for a fixed classification problem and a fixed sample size (fixed n) in each Monte Carlo repetition is as follows:

- The true bin probabilities $\boldsymbol{p}^0$ and $\boldsymbol{p}^1$ are fixed.
- n_0 and n_1 are determined using c as $n_0 = \lceil cn \rceil$ and $n - n_0$.
- Observations (training data) are randomly sampled from the multinomial distribution for each class, i.e. $(U_1^y, \ldots, U_b^y) \sim \mathcal{M}ult(\boldsymbol{p}^y; n_y)$, for $y \in \{0, 1\}$.
- 10 data points are randomly taken from the training data points of each class to be used in the prior construction methods that utilize partial data (REMLP and MKDIP-R)
- All the classification rules are trained based on their constructed prior (if applicable to that classification rule) and the training data.
- The classification errors associated with the classifiers are computed using $\boldsymbol{p}^0$ and $\boldsymbol{p}^1$. Also for the Bayesian methods, the posterior probability mass (mean) distance from the true parameters (true bin probabilities, $\boldsymbol{p}^0$ and $\boldsymbol{p}^1$) is calculated.

The regularization parameter λ_1 is set to 0.5, and λ_2 is set to 0.25 and 0.5 for the mammalian cell cycle classification problem and the TP53 classification problem, respectively. The results of expected classification error and posterior mean distance from the true parameters for the mammalian cell-cycle network are shown in Tables 4 and 6, respectively. Tables 5 and 7 contain the results of expected classification error and posterior mean distance from the true parameters for the TP53 network.

The best performance (with the lowest error in Tables 4 and 5, and lowest distance in Tables 6 and 7) for each sample size, are written in bold. For the mammalian cell-cycle network, MKDIP methods show the best (or as good as the best) performance in all the scenarios in terms of both the expected classification error and posterior parameter estimates. For the TP53 network, MKDIP methods show the best performances in posterior parameter estimates, and are competitive with the previous knowledge-driven prior construction methods in terms of the expected classification error.

Performance comparison in mixture setup

The performance of the OBC with different prior construction methods, including OBC with the Jeffreys' prior, OBC with prior constructions methods of [36] (RMEP, RMDIP, REMLP), and OBC with the general framework of constraints (MKDIP-E, MKDIP-D, MKDIP-R), are further compared in the mixture setup with missing labels, for both the mammalian cell-cycle and the TP53 systems. Also, the OBC with prior distribution centered on the true parameters with a relatively low variance (hereinafter abbreviated as PDCOTP method in Tables 8 and 9) is considered as the comparison baseline, though it is not a practical method. Similar to the classification problems, we assume that only two components (classes) exist, normal and mutated (cancerous). Here, c_0 is fixed at 0.6 ($c_1 = 1 - c_0 = 0.4$), but the sampling is not stratified. The component-conditional SSDs (bin probabilities) for the two components are as before in the classification problem, i.e. the same as the class-conditional SSDs in the classification problem.

For each sample point, first the label (y) is generated from a Bernoulli distribution with success probability c_1, and then the bin observation is generated given the label, from the corresponding class-conditional SSD (class conditional bin probabilities vector, $\boldsymbol{p}^y$), i.e. the bin observation is a sample from a categorical distribution with parameter vector $\boldsymbol{p}^y$ but the label is hidden for the inference chain and classifier training. n sample points are generated and fed into the Gibbs inference chain with different priors from the different prior construction methods. Then the OBC is calculated based on Eq. 9. For each sample size, 400 Monte Carlo repetitions are done to calculate the expected true error and the error of classifying the unlabeled observed data used for the inference itself.

To have a fair comparison of different methods' class-conditional prior probability construction, we assume that we have a rough idea of the mixture weights (class probabilities). In practice this can come from existing population statistics. That is, the Dirichlet prior distribution over the mixture weights (class probabilities) parameters, $\boldsymbol{\phi}$ in $\mathcal{D}(\boldsymbol{\phi})$, are sampled in each iteration from a uniform distribution that is centered on the true mixture weights

Table 4 Expected true error of different classification rules for the mammalian cell-cycle network. The constructed priors are considered using two precision factors: optimal precision factor (left) and estimated precision factor (right), with $c = 0.5$, and $c = 0.6$, where the minimum achievable error (Bayes error) is denoted by Err_{Bayes}

(a) $c = 0.5$, optimal precision factor,$Err_{Bayes} = 0.2648$

Method/ n	30	60	90	120	150
Hist	0.3710	0.3423	0.3255	0.3155	0.3081
CART	0.3326	0.3195	0.3057	0.3031	0.2975
RF	0.3359	0.3160	0.3015	0.2991	0.2933
SVM	0.3359	0.3112	**0.2977**	0.2959	0.2940
Jeffreys'	0.3710	0.3423	0.3255	0.3155	0.3081
RMEP	0.3236	0.3070	0.3010	0.2946	0.2910
RMDIP	0.3236	0.3070	0.3010	0.2946	0.2910
REMLP	0.3425	0.3264	0.3146	0.3067	0.3011
MKDIP-E	0.3221	0.3070	0.3010	0.2949	0.2910
MKDIP-D	0.3232	0.3070	0.3010	0.2952	0.2910
MKDIP-R	**0.3149**	**0.3028**	0.2985	**0.2943**	**0.2907**

(b) $c = 0.5$, estimated precision factor, $Err_{Bayes} = 0.2648$

Method/ n	30	60	90	120	150
Hist	0.3710	0.3423	0.3255	0.3155	0.3081
CART	0.3326	0.3195	0.3057	0.3031	0.2975
RF	0.3359	0.3160	0.3015	0.2991	0.2933
SVM	0.3359	0.3112	0.2977	0.2959	0.2940
Jeffreys'	0.3710	0.3423	0.3255	0.3155	0.3081
RMEP	0.3315	0.3059	0.2985	0.2963	0.2930
RMDIP	0.3314	0.3060	0.2986	0.2965	0.2931
REMLP	0.3488	0.3352	0.3202	0.3101	0.3048
MKDIP-E	0.3313	0.3056	0.2982	0.2962	0.2929
MKDIP-D	0.3315	0.3061	0.2986	0.2965	0.2931
MKDIP-R	**0.3205**	**0.3041**	**0.2969**	**0.2947**	**0.2919**

(c) $c = 0.6$, optimal precision factor, $Err_{Bayes} = 0.31$

Method/ n	30	60	90	120	150
Hist	0.3622	0.3608	0.3624	0.3641	0.3652
CART	0.3554	0.3556	0.3507	0.3510	0.3447
RF	0.3524	0.3514	0.3467	0.3476	0.3420
SVM	0.3735	0.3684	0.3615	0.3602	0.3544
Jeffreys'	0.3620	0.3559	0.3519	0.3502	0.3472
RMEP	**0.3415**	0.3385	**0.3394**	**0.3390**	**0.3386**
RMDIP	**0.3415**	**0.3383**	**0.3394**	**0.3390**	**0.3386**
REMLP	0.3666	0.3625	0.3587	0.3558	0.3530
MKDIP-E	**0.3415**	0.3384	**0.3394**	**0.3390**	**0.3386**
MKDIP-D	**0.3415**	0.3386	**0.3394**	**0.3390**	**0.3386**
MKDIP-R	0.3437	0.3409	0.3404	0.3401	0.3389

(d) $c = 0.6$, estimated precision factor, $Err_{Bayes} = 0.31$

Method/ n	30	60	90	120	150
Hist	0.3622	0.3608	0.3624	0.3641	0.3652
CART	0.3554	0.3556	0.3507	0.3510	0.3447
RF	0.3524	0.3514	0.3467	0.3476	0.3420
SVM	0.3735	0.3684	0.3615	0.3602	0.3544
Jeffreys'	0.3620	0.3559	0.3519	0.3502	0.3472
RMEP	0.3528	0.3415	0.3407	0.3388	0.3378
RMDIP	0.3529	0.3415	0.3408	0.3388	0.3378
REMLP	0.3700	0.3650	0.3603	0.3578	0.3546
MKDIP-E	0.3525	**0.3413**	**0.3405**	**0.3387**	**0.3377**
MKDIP-D	0.3532	0.3418	0.3409	0.3389	0.3379
MKDIP-R	**0.3486**	0.3416	0.3416	0.3402	0.3387

The lowest error for each sample size is written in bold

vector $+/-$ 10% interval, and fixed for all the methods in that repetition. For the REMLP and MKDIP-R that need labeled data in their prior construction procedure, the predicted labels from using the Jeffreys' prior are used and one fourth of the data points are used in prior construction for these two methods, and all for inference. The reason for using a larger number of data points in prior construction within the mixture setup compared to the classification setup is that in the mixture setup, data points are missing their true class labels, and the initial label estimates may be inaccurate. One can use a relatively larger number of data points in prior construction, which still avoids overfitting. The regularization parameters λ_1 and λ_2 are set as in the classification problem. Optimal precision factors are used for all prior construction methods. The results are shown in Tables 8 and 9 for the mammalian cell-cycle and TP53 models, respectively. The best performance (lowest error) for each sample size and the best performance among practical methods (all other than PDCOTP), if different, is written in bold. As can be seen from the tables, in most cases the MKDIP methods have the best performance among the practical methods. With larger sample sizes, MKDIP-R even outperforms PDCOTP in the mammalian cell-cycle system.

Performance comparison on a real data set

In this section the performance of the proposed methods are examined on a publicly available gene expression dataset. Here, we have considered the classification of two subtypes of non-small cell lung cancer (NSCLC), lung adenocarcinoma (LUA) versus lung squamous cell carcinoma (LUS). Lung cancer is the second most

Table 5 Expected true error of different classification rules for the TP53 network. The constructed priors are considered using two precision factors: optimal precision factor (left) and estimated precision factor (right), with $c = 0.5$, and $c = 0.6$, where the minimum achievable error (Bayes error) is denoted by Err_{Bayes}

(a) $c = 0.5$, optimal precision factor,$Err_{Bayes} = 0.3146$

Method/ n	15	30	45	60	75
Hist	0.3586	0.3439	0.3337	0.3321	0.3296
CART	0.3633	0.3492	0.3350	0.3314	0.3295
RF	0.3791	0.3574	0.3461	0.3400	0.3362
SVM	0.3902	0.3481	0.3433	0.3324	0.3322
Jeffreys'	0.3809	0.3439	0.3457	0.3321	0.3334
RMEP	0.3399	0.3392	0.3360	0.3315	0.3328
RMDIP	0.3399	0.3392	0.3360	0.3315	0.3328
REMLP	0.3405	**0.3340**	**0.3320**	**0.3292**	0.3287
MKDIP-E	**0.3397**	0.3398	0.3351	0.3306	0.3297
MKDIP-D	**0.3397**	0.3398	0.3347	0.3306	0.3297
MKDIP-R	0.3435	0.3354	0.3321	0.3295	**0.3283**

(b) $c = 0.5$, estimated precision factor, $Err_{Bayes} = 0.3146$

Method/ n	15	30	45	60	75
Hist	0.3586	0.3439	**0.3337**	0.3321	0.3296
CART	0.3633	0.3492	0.3350	**0.3314**	0.3295
RF	0.3791	0.3574	0.3461	0.3400	0.3362
SVM	0.3902	0.3481	0.3433	0.3324	0.3322
Jeffreys'	0.3809	0.3439	0.3457	0.3321	0.3334
RMEP	0.3791	0.3489	0.3377	0.3329	0.3302
RMDIP	0.3789	0.3490	0.3378	0.3329	0.3302
REMLP	**0.3417**	**0.3372**	0.3350	0.3318	0.3292
MKDIP-E	0.3675	0.3470	0.3373	0.3326	0.3298
MKDIP-D	0.3668	0.3472	0.3374	0.3327	0.3298
MKDIP-R	0.3471	0.3402	0.3349	0.3316	**0.3287**

(c) $c = 0.6$, optimal precision factor, $Err_{Bayes} = 0.2691$

Method/ n	15	30	45	60	75
Hist	0.3081	0.2965	0.2906	0.2883	0.2846
CART	0.3173	0.2988	0.2882	0.2846	**0.2796**
RF	0.3333	0.3035	0.2946	0.2850	0.2842
SVM	0.3322	0.3091	0.2991	0.2926	0.2857
Jeffreys'	0.3105	0.2936	0.2860	**0.2828**	0.2819
RMEP	**0.2924**	0.2922	0.2847	0.2843	0.2835
RMDIP	**0.2924**	0.2922	0.2847	0.2843	0.2835
REMLP	0.3003	**0.2908**	0.2869	0.2839	0.2832
MKDIP-E	**0.2924**	0.2909	**0.2837**	0.2851	0.2837
MKDIP-D	**0.2924**	0.2909	**0.2837**	0.2851	0.2837
MKDIP-R	0.3032	0.2917	0.2868	0.2843	0.2825

(d) $c = 0.6$, estimated precision factor, $Err_{Bayes} = 0.2691$

Method/ n	15	30	45	60	75
Hist	0.3081	0.2965	0.2906	0.2883	0.2846
CART	0.3173	0.2988	0.2882	0.2846	**0.2796**
RF	0.3333	0.3035	0.2946	0.2850	0.2842
SVM	0.3322	0.3091	0.2991	0.2926	0.2857
Jeffreys'	0.3105	0.2936	**0.2860**	**0.2828**	0.2819
RMEP	0.3346	0.3024	0.2894	0.2860	0.2823
RMDIP	0.3344	0.3023	0.2895	0.2858	0.2823
REMLP	**0.3054**	**0.2930**	0.2910	0.2870	0.2850
MKDIP-E	0.3341	0.3025	0.2898	0.2864	0.2822
MKDIP-D	0.3347	0.3024	0.2898	0.2862	0.2822
MKDIP-R	0.3096	0.2981	0.2910	0.2869	0.2849

The lowest error for each sample size is written in bold

commonly diagnosed cancer and the leading cause of cancer death in both men and women in the United States [59]. About 84% of lung cancers are NSCLC [59] and LUA and LUS combined account for about 70% of lung cancers based on the American Cancer Society statistics for NSCLC. We have downloaded LUA and LUS datasets (both labeled as TCGA provisional) in the form of mRNA expression z-scores (based on RNA-Seq profiling) from the public database cBioPortal [60, 61] for the patient sets tagged as "All Complete Tumors", denoting the set of all tumor samples that have mRNA and sequencing data. The two datasets for LUA and LUS consist of 230 and 177 sample points, respectively. We have quantized the data into binary levels based on the following preprocessing steps. First, to remove the bias for each patient, each patient's data are normalized by the mean of the z-scores of a randomly selected subset from the list of the recurrently mutated genes (half the size of the list) from the MutSig [62] (directly provided by cBioPortal). Then, a two component Gaussian mixture model is fit to each gene in each data set, and the normalized data are quantized by being assigned to one component, namely 0 or 1 (1 being the component with higher mean). We confine the feature set to {EGFR,PIK3CA,AKT,KRAS,RAF1,BAD,P53,BCL2} which are among the genes in the most relevant signaling pathways to the NSCLC [63]. These genes are altered, in different forms, in 86% and 89% of the sequenced LUA and LUS tumor samples on the cBioPortal, respectively. There are 256 bins in this classification setting, since the feature set consists of 8 genes. The pathways relevant to the NSCLC classification problem considered here are collected from KEGG [64, 65] Pathways for NSCLC

Table 6 Expected difference between the true model (for mammalian cell-cycle network) and estimated posterior probability masses. Optimal precision factor (left) and estimated precision factor (right), with $c = 0.5$, and $c = 0.6$

(a) $c = 0.5$, optimal precision factor

Method/ n	30	60	90	120	150
Jeffreys'	0.2155	0.1578	0.1300	0.1134	0.1010
RMEP	0.1591	0.1293	0.1126	0.1020	0.0912
RMDIP	0.1591	0.1294	0.1126	0.1020	0.0912
REMLP	0.1863	0.1436	0.1225	0.1088	0.0970
MKDIP-E	0.1589	0.1293	0.1126	0.1019	0.0911
MKDIP-D	0.1591	0.1293	0.1126	0.1020	0.0912
MKDIP-R	**0.1563**	**0.1283**	**0.1118**	**0.1012**	**0.0907**

(b) $c = 0.5$, estimated precision factor

Method/ n	30	60	90	120	150
Jeffreys'	0.2155	0.1578	0.1300	0.1134	0.1010
RMEP	0.1761	**0.1381**	**0.1177**	0.1032	**0.0943**
RMDIP	0.1761	**0.1381**	**0.1177**	0.1032	**0.0943**
REMLP	0.2060	0.1607	0.1315	0.1120	0.1019
MKDIP-E	0.1760	**0.1381**	**0.1177**	**0.1031**	**0.0943**
MKDIP-D	0.1761	**0.1381**	**0.1177**	0.1032	**0.0943**
MKDIP-R	**0.1742**	0.1392	0.1184	0.1036	0.0949

(c) $c = 0.6$, optimal precision factor

Method/ n	30	60	90	120	150
Jeffreys'	0.2183	0.1595	0.1322	0.1146	0.1027
RMEP	0.1628	0.1332	0.1154	0.1039	0.0946
RMDIP	0.1628	0.1333	0.1154	0.1039	0.0947
REMLP	0.1867	0.1471	0.1247	0.1101	0.0990
MKDIP-E	0.1627	0.1332	0.1154	0.1038	0.0946
MKDIP-D	0.1628	0.1332	0.1154	0.1039	0.0946
MKDIP-R	**0.1598**	**0.1317**	**0.1144**	**0.1032**	**0.0940**

(d) $c = 0.6$, estimated precision factor

Method/ n	30	60	90	120	150
Jeffreys'	0.2183	0.1595	0.1322	0.1146	0.1027
RMEP	0.1805	**0.1408**	0.1201	**0.1061**	**0.0961**
RMDIP	0.1805	**0.1408**	0.1201	**0.1061**	**0.0961**
REMLP	0.2065	0.1635	0.1346	0.1166	0.1036
MKDIP-E	**0.1804**	**0.1408**	**0.1200**	**0.1061**	**0.0961**
MKDIP-D	0.1805	**0.1408**	0.1201	**0.1061**	**0.0961**
MKDIP-R	0.1814	0.1421	0.1207	0.1065	0.0965

The lowest distance for each sample size is written in bold

Table 7 Expected difference between the true model (for TP53 network) and estimated posterior probability masses. Optimal precision factor (left) and estimated precision factor (right), with $c = 0.5$, and $c = 0.6$

(a) $c = 0.5$, optimal precision factor

Method/ n	15	30	45	60	75
Jeffreys'	0.2285	0.1716	0.1429	0.1242	0.1114
RMEP	0.1427	0.1165	0.1051	0.0934	0.0880
RMDIP	0.1424	0.1163	0.1048	0.0932	**0.0878**
REMLP	0.1698	0.1337	0.1199	0.1091	0.0985
MKDIP-E	0.1412	0.1161	0.1050	0.0933	0.0880
MKDIP-D	**0.1407**	**0.1158**	**0.1047**	**0.0931**	**0.0878**
MKDIP-R	0.1564	0.1247	0.1118	0.1031	0.0930

(b) $c = 0.5$, estimated precision factor

Method/ n	15	30	45	60	75
Jeffreys'	0.2285	0.1716	0.1429	0.1242	0.1114
RMEP	0.2218	0.1578	**0.1280**	0.1095	**0.0981**
RMDIP	0.2217	0.1575	0.1281	**0.1094**	**0.0981**
REMLP	0.1845	0.1505	0.1366	0.1235	0.1133
MKDIP-E	0.2149	0.1565	0.1282	0.1096	**0.0981**
MKDIP-D	0.2149	0.1564	0.1281	0.1096	**0.0981**
MKDIP-R	**0.1733**	**0.1410**	0.1281	0.1171	0.1082

(c) $c = 0.6$, optimal precision factor

Method/ n	15	30	45	60	75
Jeffreys'	0.2319	0.1723	0.1438	0.1262	0.1137
RMEP	0.1476	0.1222	0.1090	0.0987	0.0923
RMDIP	0.1474	0.1220	0.1087	0.0985	0.0921
REMLP	0.1751	0.1332	0.1192	0.1077	0.0980
MKDIP-E	0.1457	0.1215	0.1086	0.0985	0.0922
MKDIP-D	**0.1452**	**0.1211**	**0.1084**	**0.0983**	**0.0920**
MKDIP-R	0.1574	0.1217	0.1093	0.1010	0.0926

(d) $c = 0.6$, estimated precision factor

Method/ n	15	30	45	60	75
Jeffreys'	0.2319	0.1723	0.1438	0.1262	0.1137
RMEP	0.2182	0.1599	0.1304	**0.1144**	0.1032
RMDIP	0.2179	0.1597	0.1303	**0.1144**	**0.1031**
REMLP	0.1937	0.1522	0.1363	0.1235	0.1144
MKDIP-E	0.2165	0.1586	0.1304	0.1147	0.1036
MKDIP-D	0.2164	0.1585	0.1303	0.1147	0.1035
MKDIP-R	**0.1758**	**0.1418**	**0.1274**	0.1158	0.1086

The lowest distance for each sample size is written in bold

Table 8 Expected errors of different Bayesian classification rules in the mixture model for the mammalian cell-cycle network. Expected true error (left) and expected error on unlabeled training data (right), with $c_0 = 0.6$

Method/ n	30	60	90	120	150	Method/ n	30	60	90	120	150
PDCOTP	**0.3216**	**0.3246**	**0.3280**	0.3309	0.3334	PDCOTP	**0.3236**	**0.3270**	**0.3314**	0.3355	0.3339
Jeffreys'	0.4709	0.4743	0.4704	0.4675	0.4654	Jeffreys'	0.4751	0.4621	0.4681	0.4700	0.4645
RMEP	0.3417	0.3340	0.3307	0.3300	0.3299	RMEP	0.3447	0.3409	0.3366	0.3323	0.3316
RMDIP	0.3408	0.3336	0.3300	0.3305	0.3301	RMDIP	**0.3442**	0.3404	0.3342	0.3344	0.3343
REMLP	0.3754	0.3835	0.3882	0.3857	0.3844	REMLP	0.3748	0.3821	0.3908	0.3826	0.3812
MKDIP-E	0.3411	0.3341	**0.3297**	0.3297	0.3306	MKDIP-E	0.3457	0.3386	0.3351	0.3312	0.3320
MKDIP-D	**0.3407**	**0.3330**	0.3306	0.3304	0.3303	MKDIP-D	0.3482	0.3387	0.3381	0.3342	0.3334
MKDIP-R	0.3457	0.3342	0.3299	**0.3286**	**0.3289**	MKDIP-R	0.3449	**0.3343**	**0.3330**	**0.3306**	**0.3275**

The lowest error for each sample size and the lowest error among practical methods is written in bold

and PI3K-AKT signaling pathways, and also from [63], as shown in Fig. 4. The corresponding regulating functions are shown in Table 10.

The informative prior construction methods utilize the knowledge in the pathways in Fig. 4, and the MKDIP methods also use the regulating relationships in Table 10 in order to construct prior distributions. The incidence rate of the two subtypes, LUA and LUS, varies based on demographic factors. Here, we approximate the class probability $c = P(Y = \text{LUA})$ as $c \approx 0.57$, based on the latest statistics of the American Cancer Society for NSCLC, and also based on a weighted average of the rates for 11 countries given in [66]. In each Monte Carlo repetition, n sample points by stratified sampling, i.e. $n_0 = \lceil cn \rceil$ and $n_1 = n - n_0$ sample points, are randomly taken from preprocessed LUA (class 0) and LUS (class 1) datasets, respectively, for prior construction (if used by the method) and classifier training, and the rest of the sample points are held out for error estimation. For each n, 400 Monte Carlo repetitions are done to calculate the expected classification error. In the prior construction methods, first the optimization is solved for both classes with the precision factors $\alpha_0^y = 200, y \in \{0, 1\}$, and then their optimal values are estimated using the training points. For REMLP and MKDIP-R, which use a fraction of the training data in their prior construction procedure, $\min\big(20, \max(6, \lfloor 0.25 n_y \rfloor)\big)$ sample points from the training data of each class ($y \in \{0, 1\}$) are used for prior construction, and all the training data are used for inference. The regularization parameters λ_1 and λ_2 are set to 0.5 and 0.25, respectively. The results are shown in Table 11. In the table, the best performance among Hist, CART, RF and SVM is shown as Best Non Bayesian method. Best RM represents the best performance among RMEP, RMDIP, and REMLP. Best MKDIP denotes the best performance among the MKDIP methods.

The best performing rule for each sample size is written in bold. As can be seen from the table, OBC with MKDIP prior construction methods has the best performance among the classification rules. It is also clear that the classification performance can be significantly improved when pathway prior knowledge is integrated for constructing prior probabilities, especially when the sample size is small.

Implementation remarks

The results presented in this paper are based on Monte Carlo simulations, where thousands of optimization

Table 9 Expected errors of different Bayesian classification rules in the mixture model for the TP53 network. Expected true error (left) and expected error on unlabeled training data (right), with $c_0 = 0.6$

Method/ n	15	30	45	60	75	Method/ n	15	30	45	60	75
PDCOTP	**0.2746**	**0.2824**	**0.2829**	**0.2996**	**0.2960**	PDCOTP	**0.2762**	**0.2818**	**0.2900**	**0.3027**	**0.2900**
Jeffreys'	0.4204	0.4324	0.4335	0.4432	0.4361	Jeffreys'	0.4220	0.4314	0.4381	0.4419	0.4348
RMEP	**0.3274**	**0.3204**	0.3327	**0.3402**	0.3422	RMEP	**0.3471**	0.3350	0.3487	0.3543	0.3529
RMDIP	0.3297	0.3260	0.3327	0.3406	0.3432	RMDIP	0.3504	0.3423	0.3496	0.3551	0.3545
REMLP	0.3637	0.3687	0.3706	0.3658	0.3653	REMLP	0.3489	0.3579	0.3709	0.3593	0.3556
MKDIP-E	0.3312	0.3246	0.3322	0.3428	0.3386	MKDIP-E	0.3502	0.3378	0.3486	0.3585	0.3492
MKDIP-D	0.3321	**0.3204**	**0.3306**	0.3436	**0.3366**	MKDIP-D	0.3551	**0.3329**	**0.3473**	0.3570	0.3475
MKDIP-R	0.3872	0.3749	0.3667	0.3607	0.3586	MKDIP-R	0.3613	0.3583	0.3589	**0.3539**	**0.3462**

The lowest error for each sample size and the lowest error among practical methods is written in bold

Table 10 Regulating functions corresponding to the signaling pathways in Fig. 4. In the Boolean functions {AND, OR, NOT} = $\{\wedge, \vee, -\}$

Gene	Node name	Boolean regulating function
EGFR	v_1	-
PIK3CA	v_2	$v_1 \vee v_4$
AKT	v_3	v_2
KRAS	v_4	-
RAF1	v_5	$v_4 \wedge \overline{v_3}$
BAD	v_6	$\overline{v_3}$
P53	v_7	-
BCL2	v_8	$\overline{v_6} \vee \overline{v_7}$

problems are solved for each sample size for each problem. Thus, the regularization parameters and the number of sample points used in prior construction are preselected for each problem. One can use cross validation to set these parameters in a specific application. It has been shown in [36] that by assuming precision factors greater than 1 ($\alpha_0^y > 1, y \in \{0, 1\}$), all three objective functions used are convex for the class of Dirichlet prior probabilities for multinomial likelihood functions. But unfortunately, we cannot guarantee the convexity of the feasible space due to the convolved constraints. Therefore, we have employed algorithms for nonconvex optimization problems and there is no guarantee of convergence to the global optimum. The method used for solving the optimization framework of the prior construction is based on the interior-point algorithm for nonlinear constrained optimization [67, 68] implemented in the `fmincon` function in MATLAB. In this paper, since the interest is in classification problems with small training sample sizes (which is often the case in bioinformatics) and also due to Monte Carlo simulations, we have only shown performance results on small networks with only a few genes. In practice, there would be no problem using the proposed method for larger networks, since there would then be a single one-time analysis. One should also note that with small sample sizes, one needs feature selection to keep the number of features small. In the experiments in this paper, feature selection is automatically done by focusing on the most relevant network by biological prior knowledge.

Conclusion

Bayesian methods have shown promising performance in classification problems in the presence of uncertainty and small sample sizes, which often occur in translational genomics problems. The impediment in using these methods is prior construction to integrate existing prior biological knowledge. In this paper we have proposed a knowledge-driven prior construction method with a

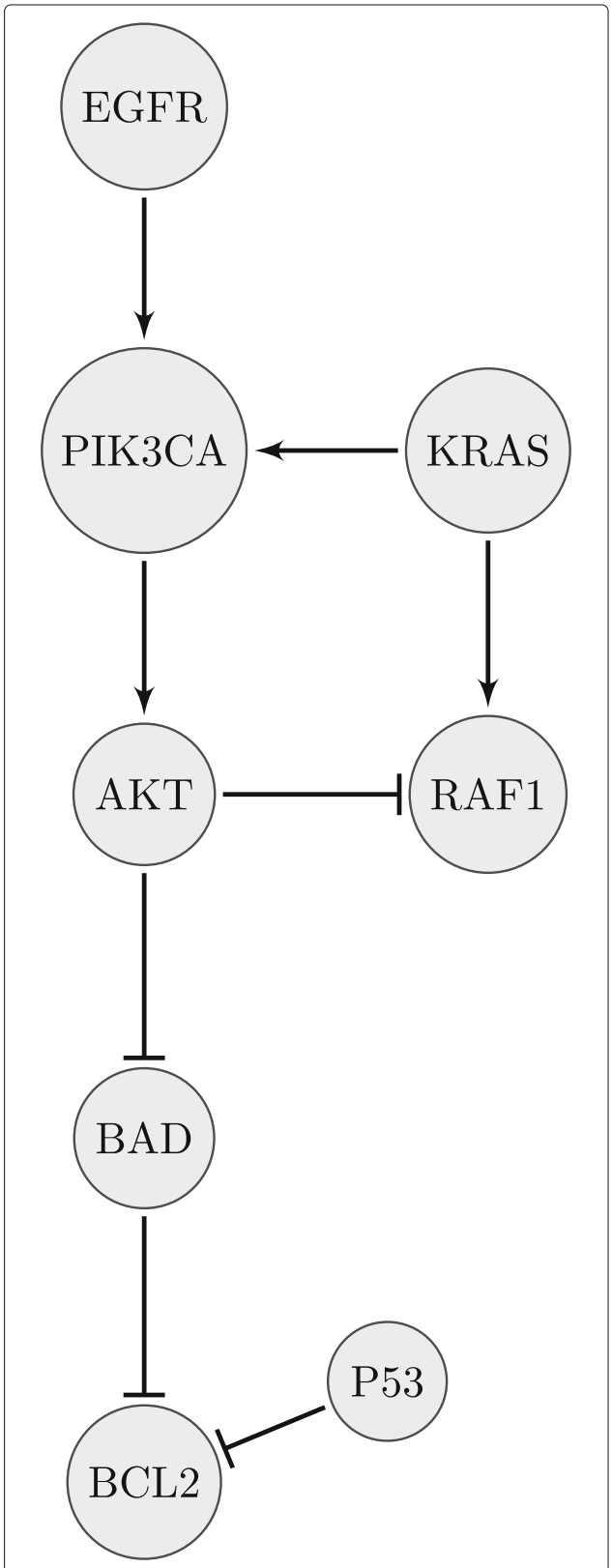

Fig. 4 Signaling pathways corresponding to NSCLC classification. The pathways are collected from KEGG Pathways for NSCLC and PI3K-AKT pathways, and from [63]

Table 11 Expected error of different classification rules calculated on a real dataset. The classification is between LUA (class 0) and LUS (class 1), with $c = 0.57$

Method/ n	34	74	114	134	174
Best Non Bayesian	0.1764	0.1574	0.1473	0.1426	0.1371
Jeffreys'	0.1766	0.1574	0.1476	0.1425	0.1371
Best RM	0.1426	0.1289	0.1164	0.1083	0.1000
Best MKDIP	**0.1401**	**0.1273**	**0.1162**	**0.1075**	**0.0998**

general framework of mapping prior biological knowledge into a set of constraints. Knowledge can come from biological signaling pathways and other population studies, and be translated into constraints over conditional probabilities. This general scheme includes the previous approaches of using biological prior knowledge in prior construction. Here, the superior performance of this general scheme is shown on two important pathway families, the mammalian cell-cycle pathway and the pathway centering around TP53. In addition, prior construction and the OBC are extended to a mixture model, where data sets are with missing labels. Moreover, comparisons on a publicly available gene expression dataset show that classification performance can be significantly improved for small sample sizes when corresponding pathway prior knowledge is integrated for constructing prior probabilities.

Acknowledgements
Not applicable.

Funding
This work was funded in part by Award CCF-1553281 from the National Science Foundation, and a DMREF grant from the National Science Foundation, award number 1534534. The publication cost of this article was funded by Award CCF-1553281 from the National Science Foundation.

Availability of data and materials
The publicly available real datasets analyzed during the current study have been generated by the TCGA Research Network https://cancergenome.nih.gov/, and have been procured from http://www.cbioportal.org/.

About this supplement
This article has been published as part of BMC Bioinformatics Volume 18 Supplement 14, 2017: Proceedings of the 14th Annual MCBIOS conference. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-18-supplement-14.

Authors' contributions
SB developed mixture-model modeling and extracting knowledge from pathways and regulating functions, performed the experiments, and wrote the first draft. MSE structured the prior knowledge by integrating his previous prior methods into this new framework. XQ in conjunction with ERD proposed the new general prior structure and proofread and edited the manuscript. ERD oversaw the project, in conjunction with XQ proposed the new general prior structure, wrote the OBC section, and proofread and edited the manuscript. All authors have read and approved final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Electrical and Computer Engineering, Texas A&M University, MS3128 TAMU, 77843 College Station, TX, USA. [2]Division of Oncology and Center for Cancer Systems Biology, Stanford School of Medicine, 291 Campus Drive, 94305 Stanford, CA, USA.

Published: 28 December 2017

References

1. Dougherty ER, Zollanvari A, Braga-Neto UM. The illusion of distribution-free small-sample classification in genomics. Current Genomics. 2011;12(5):333.
2. Dougherty ER, Dalton LA. Scientific knowledge is possible with small-sample classification. EURASIP J Bioinforma Syst Biol. 2013;2013(1):1–12.
3. Jaynes ET. What is the question? In: Bernardo JM, deGroot MH, Lindly DV, Smith AFM, editors. Bayesian Stat. Valencia: Valencia Univ. Press. 1980. p. 618–629.
4. Jeffreys H. An invariant form for the prior probability in estimation problems. Proc Royal Soc London Ser A Math Phys Sci. 1946;186(1007):453–61.
5. Zellner A. Past and Recent Results on Maximal Data Information Priors. Working paper series in economics and econometrics. University of Chicago, Graduate School of Business, Department of Economics, Chicago. 1995.
6. Rissanen J. A universal prior for integers and estimation by minimum description length. Ann Stat. 1983;11(2):416–31.
7. Rodríguez CC. Entropic priors. Albany: Department of Mathematics and Statistics, State University of New York; 1991.
8. Berger JO, Bernardo JM. On the development of reference priors. Bayesian Stat. 1992;4(4):35–60.
9. Spall JC, Hill SD. Least-informative Bayesian prior distributions for finite samples based on information theory. Autom Control IEEE Trans. 1990;35(5):580–3.
10. Bernardo JM. Reference posterior distributions for Bayesian inference. J Royal Stat Soc Ser B Methodol. 1979;41(2):113–147.
11. Kass RE, Wasserman L. The selection of prior distributions by formal rules. J Am Stat Assoc. 1996;91(435):1343–1370.
12. Berger JO, Bernardo JM, Sun D. Objective priors for discrete parameter spaces. J Am Stat Assoc. 2012;107(498):636–48.
13. Jaynes ET. Information theory and statistical mechanics. Physical Rev. 1957;106(4):620.
14. Jaynes ET. Prior probabilities. Syst Sci Cybern IEEE Trans. 1968;4(3):227–41.
15. Zellner A. Models, prior information, and Bayesian analysis. J Econ. 1996;75(1):51–68.
16. Burg JP, Luenberger DG, Wenger DL. Estimation of structured covariance matrices. Proc IEEE. 1982;70(9):963–74.
17. Werner K, Jansson M, Stoica P. On estimation of covariance matrices with kronecker product structure. Signal Proc IEEE Trans. 2008;56(2):478–91.
18. Wiesel A, Hero AO. Distributed covariance estimation in Gaussian graphical models. Signal Proc IEEE Trans. 2011;60(1):211–220.
19. Wiesel A, Eldar YC, Hero AO. Covariance estimation in decomposable Gaussian graphical models. Signal Process IEEE Trans. 2010;58(3):1482–1492.
20. Breslin T, Krogh M, Peterson C, Troein C. Signal transduction pathway profiling of individual tumor samples. BMC Bioinforma. 2005;6(1):163.
21. Zhu Y, Shen X, Pan W. Network-based support vector machine for classification of microarray samples. BMC Bioinforma. 2009;10(1):21.
22. Svensson JP, Stalpers LJ, Esveldt–van Lange RE, Franken NA, Haveman J, Klein B, Turesson I, Vrieling H, Giphart-Gassler M. Analysis of gene expression using gene sets discriminates cancer patients with and without late radiation toxicity. PLoS Med. 2006;3(10):422.
23. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol. 2008;4(11):1000217.

24. Su J, Yoon BJ, Dougherty ER. Accurate and reliable cancer classification based on probabilistic inference of pathway activity. PLoS ONE. 2009;4(12):8161.
25. Eo HS, Heo JY, Choi Y, Hwang Y, Choi HS. A pathway-based classification of breast cancer integrating data on differentially expressed genes, copy number variations and microrna target genes. Mol Cells. 2012;34(4):393–8.
26. Wen Z, Liu ZP, Yan Y, Piao G, Liu Z, Wu J, Chen L. Identifying responsive modules by mathematical programming: An application to budding yeast cell cycle. PloS ONE. 2012;7(7):41854.
27. Kim S, Kon M, DeLisi C, et al. Pathway-based classification of cancer subtypes. Biology direct. 2012;7(1):1–22.
28. Khunlertgit N, Yoon BJ. Identification of robust pathway markers for cancer through rank-based pathway activity inference. Advances Bioinforma. 2013;Article ID 618461:8.
29. Wei P, Pan W. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. Bioinforma. 2007;24(3):404–11.
30. Wei P, Pan W. Network-based genomic discovery: application and comparison of Markov random-field models. J Royal Stat Soc Ser C Appl Stat. 2010;59(1):105–25.
31. Wei P, Pan W. Bayesian joint modeling of multiple gene networks and diverse genomic data to identify target genes of a transcription factor. Annals Appl Stat. 2012;6(1):334–55.
32. Gatza ML, Lucas JE, Barry WT, Kim JW, Wang Q, Crawford MD, Datto MB, Kelley M, Mathey-Prevot B, Potti A, et al. A pathway-based classification of human breast cancer. Proc Natl Acad Sci. 2010;107(15):6994–999.
33. Nevins JR. Pathway-based classification of lung cancer: a strategy to guide therapeutic selection. Proc Am Thoracic Soc. 2011;8(2):180.
34. Wen Z, Liu ZP, Liu Z, Zhang Y, Chen L. An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. J Am Med Inform Assoc. 2013;20(4):659–67.
35. Esfahani MS, Dougherty ER. Incorporation of biological pathway knowledge in the construction of priors for optimal Bayesian classification. IEEE/ACM Trans Comput Biol Bioinforma. 2014;11(1):202–18.
36. Esfahani MS, Dougherty ER. An optimization-based framework for the transformation of incomplete biological knowledge into a probabilistic structure and its application to the utilization of gene/protein signaling pathways in discrete phenotype classification. IEEE/ACM Trans Comput Biol Bioinforma. 2015;12(6):1304–1321.
37. Boluki S, Esfahani MS, Qian X, Dougherty ER. Constructing pathway-based priors within a Gaussian mixture model for Bayesian regression and classification. IEEE/ACM Trans Comput Biol Bioinforma. 2017. In press.
38. Guiasu S, Shenitzer A. The principle of maximum entropy. Math Intell. 1985;7(1):42–8.
39. Hua J, Sima C, Cypert M, Gooden GC, Shack S, Alla L, Smith EA, Trent JM, Dougherty ER, Bittner ML. Tracking transcriptional activities with high-content epifluorescent imaging. J Biomed Opt. 2012;17(4): 0460081–04600815.
40. Dalton LA, Dougherty ER. Optimal classifiers with minimum expected error within a Bayesian framework–part I: Discrete and Gaussian models. Pattern Recog. 2013;46(5):1301–1314.
41. Dalton LA, Dougherty ER. Optimal classifiers with minimum expected error within a Bayesian framework–part II: Properties and performance analysis. Pattern Recog. 2013;46(5):1288–1300.
42. Dalton LA, Dougherty ER. Bayesian minimum mean-square error estimation for classification error–part I: Definition and the bayesian MMSE error estimator for discrete classification. Signal Process IEEE Trans. 2011;59(1):115–29.
43. MacKay DJC. Introduction to Monte Carlo methods. In: Jordan MI, editor. Learning in Graphical Models. NATO Science Series. Dordrecht: Kluwer Academic Press. 1998. p. 175–204.
44. Casella G, George EI. Explaining the Gibbs sampler. Am Stat. 1992;46(3): 167–74.
45. Robert CP, Casella G. Monte Carlo Statistical Methods. New York: Springer; 2004.
46. Zellner A. Maximal Data Information Prior Distributions, Basic Issues in Econometrics. Chicago: The University of Chicago Press; 1984.
47. Ebrahimi N, Maasoumi E, Soofi ES. In: Slottje DJ, editor. Measuring Informativeness of Data by Entropy and Variance. Heidelberg: Physica-Verlag HD; 1999, pp. 61–77.
48. Dougherty ER, Brun M, Trent JM, Bittner ML. Conditioning-based modeling of contextual genomic regulation. Comput Biol Bioinforma IEEE/ACM Trans. 2009;6(2):310–20.
49. Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol. 1969;22(3):437–67.
50. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinforma. 2002;18(2):261.
51. Fauré A, Naldi A, Chaouiya C, Thieffry D. Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. Bioinformatics. 2006;22(14):124.
52. Weinberg R. The Biology of Cancer. New York: Garland science; 2013.
53. Esfahani MS, Yoon BJ, Dougherty ER. Probabilistic reconstruction of the tumor progression process in gene regulatory networks in the presence of uncertainty. BMC Bioinformatics. 2011;12(10):9.
54. Layek RK, Datta A, Dougherty ER. From biological pathways to regulatory networks. Mol BioSyst. 2011;7:843–51.
55. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. Boca Raton: Chapman & Hall/CRC; 1984.
56. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.
57. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273–97.
58. Kecman V. Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. Cambridge: MIT Press; 2001.
59. American Cancer Society. Cancer Facts and Figures 2017. Atlanta: American Cancer Society; 2017.
60. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling. 2013;6(269):1–1.
61. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.
62. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8.
63. West L, Vidwans SJ, Campbell NP, Shrager J, Simon GR, Bueno R, Dennis PA, Otterson GA, Salgia R. A novel classification of lung cancer into molecular subtypes. PLOS ONE. 2012;7(2):1–11.
64. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
65. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):457–62.
66. Lortet-Tieulent J, Soerjomataram I, Ferlay J, Rutherford M, Weiderpass E, Bray F. International trends in lung cancer incidence by histological subtype: Adenocarcinoma stabilizing in men but still increasing in women. Lung Cancer. 2014;84(1):13–22.
67. Waltz RA, Morales JL, Nocedal J, Orban D. An interior algorithm for nonlinear optimization that combines line search and trust region steps. Math Program. 2006;107(3):391–408.
68. Byrd RH, Hribar ME, Nocedal J. An interior point algorithm for large-scale nonlinear programming. SIAM J Optim. 1999;9(4):877–900.