BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 18 2012, pages 2349–2356
doi:10.1093/bioinformatics/bts361

Data and text mining

Advance Access publication June 27, 2012

Targeted retrieval of gene expression measurements using regulatory models

Elisabeth Georgii[1,*], Jarkko Salojärvi[2,3,*], Mikael Brosché[2], Jaakko Kangasjärvi[2] and Samuel Kaski[1,4,*]

[1]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, 00076 Aalto, Espoo, Finland, [2]Division of Plant Biology, Department of Biosciences, University of Helsinki, 00014 University of Helsinki, Helsinki, Finland, [3]Veterinary Microbiology and Epidemiology, Department of Veterinary Biosciences, University of Helsinki, 00014 University of Helsinki, Helsinki, Finland and [4]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, 00014 University of Helsinki, Helsinki, Finland

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: Large public repositories of gene expression measurements offer the opportunity to position a new experiment into the context of earlier studies. While previous methods rely on experimental annotation or global similarity of expression profiles across genes or gene sets, we compare experiments by measuring similarity based on an unsupervised, data-driven regulatory model around pre-specified genes of interest. Our experiment retrieval approach is novel in two conceptual respects: (i) targetable focus and interpretability: the analysis is targeted at regulatory relationships of genes that are relevant to the analyst or come from prior knowledge; (ii) regulatory model-based similarity measure: related experiments are retrieved based on the strength of inferred regulatory links between genes.

Results: We learn a model for the regulation of specific genes from a data repository and exploit it to construct a similarity metric for an information retrieval task. We use the Fisher kernel, a rigorous similarity measure that typically has been applied to use generative models in discriminative classifiers. Results on human and plant microarray collections indicate that our method is able to substantially improve the retrieval of related experiments against standard methods. Furthermore, it allows the user to interpret biological conditions in terms of changes in link activity patterns. Our study of the osmotic stress network for *Arabidopsis thaliana* shows that the method successfully identifies relevant relationships around given key genes.

Availability: The code (R) is available at http://research.ics.tkk.fi/mi/software.shtml.

Contact: elisabeth.georgii@aalto.fi; jarkko.salojarvi@helsinki.fi; samuel.kaski@hiit.fi

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on June 4, 2012; revised on June 4, 2012; accepted on June 18, 2012

1 INTRODUCTION

Considerable effort has been spent on collecting gene expression measurements into huge public repositories. This has opened up the door to large-scale comparisons and meta-analysis of data from different experiments. A crucial question is how to find datasets that are biologically relevant for a certain phenomenon under study. The common way to obtain relevant data from repositories is keyword search (for instance, Zhu *et al.*, 2008); however, the reliability and usefulness of this approach fully depend on the availability and quality of data annotation, and on the existence of good standardized keywords. Therefore, content-based retrieval methods have received attention recently. In addition to detecting other relevant datasets, such methods can assist in automatic prediction of annotations and in discovering previously unknown relationships between different biological conditions.

Previous data-driven approaches for relating gene expression profiles can be grouped by two criteria: the representation of gene expression profiles and the similarity measure for comparing profiles. Regarding the representation, there exist three common alternatives: (i) activity profiles over genes, often differential rather than absolute values if control measurements are available, or simply rankings of genes (Engreitz *et al.*, 2010; Fujibuchi *et al.*, 2007; Le *et al.*, 2010); (ii) activity profiles over known gene sets (also called pathways or signatures; Caldas *et al.*, 2009, 2012; Huttenhower and Troyanskaya, 2008; Segal *et al.*, 2004) and (iii) activity profiles over gene modules predicted from other data sources (e.g. protein–protein interactions; Suthram *et al.*, 2010). The most widely used similarity measures are Pearson and Spearman correlation (Engreitz *et al.*, 2010; Fujibuchi *et al.*, 2007); other search tools take signatures of up- and down-regulated genes as a query and compute rank-based scores against the measurements in a database (Feng *et al.*, 2009); neither approach is fitted in any way to the database.

In contrast, several approaches derive features from the database or learn a model of it; the results are then used when computing the similarity of the query to database entries. In the simplest case, variances of gene or gene set activity values are determined from the data compendium, and then exploited for calculating a weighted correlation coefficient (Engreitz *et al.*, 2010; Le *et al.*, 2010); beyond that, one can apply dimensionality reduction techniques to learn low-dimensional feature spaces that represent the data appropriately (Engreitz *et al.*, 2010). Caldas *et al.* (2009, 2012) proposed probabilistic generative

*To whom correspondence should be addressed.

latent variable models that inherently yield probabilistic similarity measures between gene expression profiles. In addition to these purely unsupervised methods, also supervised approaches have been proposed to relate gene expression experiments with each other (Huang *et al.*, 2010; Huttenhower and Troyanskaya, 2008).

The novelty of our approach lies in the fact that the model learned from the data compendium focuses on inferring the local regulatory network neighborhood of specific user-defined genes of interest. As a well-defined and intuitive model-based retrieval criterion, we use the Fisher kernel (FK) (Jaakkola and Haussler, 1999), which derives a similarity measure between data points from a generative model. Consequently, the comparison of biological samples is based on the differential activity of particular cellular processes only. The alternative of using global transcriptomic models and hence global similarities between data would easily obscure the interesting local processes shared between measurements. The alternative of using known ontologies, on the other hand, would miss unknown regulatory patterns. In practice, the user can focus the modeling by providing known key genes involved in the biological process under study or genes whose connection to known pathways and experimental conditions is yet unclear.

Our model-driven targeted retrieval approach is intended to complement earlier tools for experiment retrieval targeted by query genes: Parkinson *et al.* (2009) retrieve experiments where the query genes are differentially expressed, Hibbs *et al.* (2007) retrieve datasets where the query genes are correlated, and Greene and Troyanskaya (2011) identify experimental conditions discriminating positive and negative query genes. As our approach models relationships of biological samples with respect to only a subset of genes, it shares some aspects with bicluster detection methods, which are very popular in gene expression analysis (Madeira and Oliveira, 2004); defining genes of interest and a query sample can be seen as providing a seed bicluster. However, in contrast to bicluster approaches, we bring the structure of gene networks into the similarity criterion. A related approach has been proposed by Lahti *et al.* (2010), but it assumes a given network whereas we infer the network around the target genes. Furthermore, none of the earlier biclustering-type approaches has been designed for retrieval.

In a nutshell, our approach consists of the following components, which will be described in detail in Section 2. First, a regulatory model of gene expression is learned from the database; to allow for more user interaction and better individual interpretability, the model focuses on relationships around user-defined genes of interest; here, we use a set of local dependency models learned by sparse linear regression as a simple regulatory model (Meinshausen and Bühlmann, 2006). Second, the model serves as a basis for comparing measurements: given a query measurement, the method ranks the measurements in the repository according to a model-based similarity criterion, here a variant of the FK (Jaakkola and Haussler, 1999; Shawe-Taylor and Cristianini, 2004). In Section 3, we show the effectiveness of our approach in a leukemia case study with a human microarray compendium and in an osmotic stress study with a collection of plant stress datasets. Section 4 discusses the work and describes future extensions.

2 METHODS FOR TARGETED COMPARISON OF GENE EXPRESSION MEASUREMENTS

Our approach for comparing gene expression measurements consists of two main steps: (i) learning a conditional expression model for a given target gene list, using a large data compendium and (ii) computing Fisher score similarities between measurements based on the model. In the following, we describe the steps in detail.

2.1 Targeted gene expression model

Our goal is to build a model for the expression of target genes, conditioned on the expression of other genes. Let us first introduce some notation. The $X = (X_1, \ldots, X_p)$ is a p-dimensional random variable that represents the gene expression profile, where p is the total number of genes. Let $\mathcal{T} \subset \{1, \ldots, p\}$ be the set of target gene indices and $-\mathcal{T}$ the set of all the remaining gene indices ($\{1, \ldots, p\} \setminus \mathcal{T}$). Then, $X_{\mathcal{T}}$ denotes the variable X limited to the subspace defined by $\mathcal{T}$. Furthermore, let $\mathbf{X}$ be the data matrix, containing in row i the i-th realization of X, which is denoted by $x^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)})$, $1 \le i \le n$. We are interested in modeling the expression values of the target genes given the expression values of the remaining genes, that is, modeling of the distribution $P(X_{\mathcal{T}} | X_{-\mathcal{T}})$. One prominent formalism for modeling conditional independence relationships between variables is Gaussian graphical models (GGMs), introduced by Dempster (1972); see Dobra *et al.* (2004) and Markowetz and Spang (2007) for the use of GGMs in gene network reconstruction. In that context, estimating the model structure is challenging due to the 'small n, large p' problem. A common approximation is to estimate the neighbors of each gene separately and combine these estimates into a so-called dependency network (Heckerman *et al.*, 2001).

We use a local dependency network composed of individual neighborhood estimates for each target gene $j \in \mathcal{T}$. Each neighborhood estimate is obtained by learning a Gaussian linear model (see, e.g. Alpaydin, 2010)

$$X_j = X_{-\{j\}}\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where β is a weight vector and σ^2 the variance of the Gaussian noise; the variables X_1,. . .,X_p are assumed to have been standardized to mean 0 and unit variance, and hence we need not consider an intercept term. In high-dimensional settings, it is desirable to obtain sparse estimates for β, i.e. solutions where most entries are 0. The variables corresponding to non-zero entries build the 'neighbor set' of the j-th variable, written as

$$N(j) = \{k \in \{1, \ldots, p\} \setminus \{j\} : \beta_{\mathrm{ind}(k)} \neq 0\}. \tag{2}$$

Here, ind(k) indicates the position of β that is mapped to the variable X_k. Sparse coefficient estimates are achieved by applying L_1-norm regularization or Lasso (Tibshirani, 1996), yielding the following solution:

$$\hat{\beta} = \arg\min_{\beta}\left(n^{-1}\|\mathbf{X}_j - \mathbf{X}_{-\{j\}}\beta\|_2^2 + \lambda\|\beta\|_1\right). \tag{3}$$

Meinshausen and Bühlmann (2006) provide an analytical choice for the penalty parameter λ that guarantees asymptotically consistent neighborhood estimates for sparse high-dimensional graphs. All results reported in this article use the analytical λ-value, with the constant α (bounding the probability of falsely joining two connectivity components) fixed to 0.05, as in the original article and in Schäfer and Strimmer (2005); it yielded sparser solutions than the λ-value chosen by cross-validation.

To obtain a model for the whole set of target genes, we simply concatenate the gene-specific models, treating them as independent. This is a widely used model approximation strategy, based on the so-called 'pseudo-likelihood' (Ambroise *et al.*, 2009; Besag, 1975; Schmidt *et al.*, 2008); the pseudo-likelihood regarding a specific data point $x^{(i)}$ is given by

$$\widetilde{\mathcal{P}}(x_{\mathcal{T}}^{(i)}|x_{-\mathcal{T}}^{(i)};\theta)=\prod_{j\in\mathcal{T}}P(x_j^{(i)}|x_{-\{j\}}^{(i)};\theta_j). \quad (4)$$

For compactness, we write in the following $\mathcal{L}(\theta_j; x^{(i)})$ instead of $P(x_j^{(i)}|x_{-\{j\}}^{(i)};\theta_j)$. Here, θ_j denotes the parameters of the j-th independent submodel, in our case consisting of the coefficient vector $\beta^{(j)}$ and the variance ${\sigma_j}^2$. The likelihood of a submodel with respect to the i-th data point is calculated as follows:

$$\mathcal{L}(\theta_j; x^{(i)})=\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{1}{2\sigma_j^2}\left(x_j^{(i)}-\sum_{k\in N(j)}\beta_{\mathrm{ind}(k)}^{(j)}x_k^{(i)}\right)^2\right).$$

2.2 Model-based similarity measure

Having learned a model for the expression of target genes, we exploit it for defining the similarity between gene expression profiles. For that purpose, we represent each expression profile (data point) by its Fisher score with respect to the model; the score is defined as the gradient of the log-likelihood for the profile – with respect to the parameters – at the learned values of the parameters (Shawe-Taylor and Cristianini, 2004). In our modeling setup, each submodel has its own parameters, and by concatenating the gradients of all submodels, we obtain a gradient of the global pseudo-log-likelihood. The parameters of biological interest in each submodel are the weight coefficients for the neighbors of the respective target gene; we ignore the derivative regarding the noise parameter σ_j, because the noise is assumed to be independent of the biological condition. The partial derivative of the log-likelihood with respect to ${\beta_{\mathrm{ind}(l)}}^{(j)}$ amounts to

$$\frac{\partial\log\mathcal{L}(\theta_j; x^{(i)})}{\partial\beta_{\mathrm{ind}(l)}^{(j)}}=\frac{1}{\sigma_j^2}\left(x_j^{(i)}-\sum_{k\in N(j)}\beta_{\mathrm{ind}(k)}^{(j)}x_k^{(i)}\right)x_l^{(i)}. \quad (5)$$

Concatenating these partial derivatives yields the Fisher score representation of the i-th data point, written as $s_{\hat{\theta}}(x^{(i)})$, where $\hat{\theta}$ denotes the estimated parameters of the overall model, which here are the L1-regularized maximum-likelihood estimates for β. The inner product in this new feature space is a model-dependent similarity measure K_θ, also called the simple Fisher kernel (FK) (Jaakkola and Haussler, 1999; Shawe-Taylor and Cristianini, 2004):

$$K_{\hat{\theta}}(x^{(i_1)}, x^{(i_2)})=s_{\hat{\theta}}(x^{(i_1)})^T s_{\hat{\theta}}(x^{(i_2)}). \quad (6)$$

To understand this measure, recall that the gradient of a function is sufficient for determining its first-order Taylor series approximation. In our case, the function is the (pseudo-)log-likelihood of the model. Its first-order Taylor series approximation evaluated at $\hat{\theta}$ is

$$\log\mathcal{L}(\theta; x^{(i)})\approx\log\mathcal{L}(\hat{\theta}; x^{(i)})+(\nabla\log\mathcal{L}(\theta; x^{(i)}))_{\hat{\theta}}(\theta-\hat{\theta}), \quad (7)$$

where $(\nabla\log\mathcal{L}(\theta; x^{(i)}))_{\hat{\theta}}$ denotes the gradient of the log-likelihood for $x^{(i)}$ at $\hat{\theta}$ and the $\hat{\theta}$ are the (maximum-likelihood) parameter estimates obtained on some training dataset D – not necessarily the ones maximizing the likelihood when a new data point $x^{(i)}$ is added to the dataset. Since the approximation is exact in the close neighborhood of $\hat{\theta}$, the gradient $(\nabla\log\mathcal{L}(\theta; x^{(i)}))_{\hat{\theta}}$, which is identical to the Fisher score $s_{\hat{\theta}}(x^{(i)})$, indicates the direction in which to update the parameters in order to maximize the log-likelihood for $x^{(i)}$, starting from the current parameters $\hat{\theta}$.

Hence, we get a parameter update for the extended dataset $D+x^{(i_1)}$ by gradient ascent

$$\hat{\theta}^{\mathrm{new}\,x^{(i_1)}}=\hat{\theta}+ds_{\hat{\theta}}(x^{(i_1)}), \quad (8)$$

where d is a suitably small step size. In the same way, we obtain an updated parameter estimate $\hat{\theta}^{\mathrm{new}\,x^{(i_2)}}$ for $D+x^{(i_2)}$. As the model parameters constitute the summary statistics of the dataset under the assumptions of the model, comparing $\hat{\theta}^{\mathrm{new}\,x^{(i_1)}}$ and $\hat{\theta}^{\mathrm{new}\,x^{(i_2)}}$ is sufficient for comparing the datasets $D+x^{(i_1)}$ and $D+x^{(i_2)}$. This is a well-defined way for doing a model-based comparison of two different data points $x^{(i_1)}$ and $x^{(i_2)}$ if D is the dataset from which the model has been learned. Furthermore, the informative part here is the comparison between $s_{\hat{\theta}}(x^{(i_1)})$ and $s_{\hat{\theta}}(x^{(i_2)})$, ignoring the step size (Shawe-Taylor and Cristianini, 2004).

In terms of differential geometry, the proper metric for comparing models on the Riemannian manifold of the model class is the Fisher information metric. The steepest ascent locally along the manifold is the natural gradient $J^{-1}s_{\hat{\theta}}(x^{(i_1)})$, where J is the Fisher information matrix. Replacing the gradient $s_{\hat{\theta}}(x^{(i_1)})$ in Equation (8) by the natural gradient and applying the metric, we obtain a kernel that is proportional to $s_{\hat{\theta}}(x^{(i_1)})^T J^{-1}s_{\hat{\theta}}(x^{(i_2)})$ (Jaakkola and Haussler, 1999). The simpler kernel $s_{\hat{\theta}}(x^{(i_1)})^{\mathcal{T}} s_{\hat{\theta}}(x^{(i_2)})$ is a suitable approximation (Jaakkola and Haussler, 1999) that makes the computation much more efficient.

In summary, the inner product between the Fisher scores of two different data points $x^{(i_1)}$ and $x^{(i_2)}$ is an approximate measure for the similarity of the resulting datasets, from the model's point of view, when adding either $x^{(i_1)}$ or $x^{(i_2)}$ to the data. In our case, we compare the Fisher scores with respect to the β coefficients, which indicate the strength and the sign of gene relationships (i.e. whether the co-regulation is in the same direction or in opposite directions). In retrieval, which is the primary application in this article, this similarity criterion is used to compare a new measurement profile against all previous ones, which results in a ranked list of the measurements available in the database. The underlying model is learned based on the whole database and does not require any experimental annotation. The retrieval is computationally inexpensive because the gene-specific submodels as well as the corresponding scores can be pre-computed offline on the data compendium. At query time, we only have to compute the scores of the query and then compute inner products between the query scores and previous measurement scores, where only scores of submodels for genes specified in the user-given target list are considered. These operations require $O(KM)$ time, where K is the number of non-zero coefficients in the model and M is the number of measurements in the database.

3 RESULTS

We applied our method in two case studies: human leukemia and plant stress. We show that targeted expression models are not only able to enhance the retrieval of related measurements but also allow for interpretation of underlying biological processes in terms of networks and network activation patterns. Prior knowledge in the form of key genes is taken into account, and relevant relationships (putative network neighbors) of the genes are inferred directly from the data.

3.1 Leukemia case study

For a proof-of-principle study of the new method's performance, we used a large compendium of human microarrays (Lukk *et al.*, 2010); it consists of 5372 measurements from different cell types tissue types and disease states, which have been collected from 206 different studies, all based on the same microarray platform (Affymetrix U133A). A subset of 567 samples from 15 studies are labeled as leukemia (according to the '15 metagroups' annotation). We tested the performance of the model-based method for the task of retrieval of relevant experiments (REx) by 5-fold cross-validation, where one-fifth of the leukemia samples were excluded from the model training and then taken as queries. The model was learned from the remaining leukemia samples and all non-leukemia samples, without using the label information. The ranked retrieval results were evaluated based on the given annotation (leukemia versus non-leukemia). The precision of the top-*k* results of a query is given by the fraction of leukemia samples among the first *k* samples.

As targets, we chose eight genes that are known to play a role in leukemia: BCR, ETV6, FLT3, HOXA9, MYST3, PRDM2, RUNX1 and TAL1. They were also reported as being differentially expressed in the 567 leukemia samples (among a list of 243 genes; Lukk *et al*., 2010). This preselection might give a positive bias to the precision values, which we controlled by testing whether this set of target genes would be a good biomarker for discriminating leukemia samples against non-leukemia samples. We observed that baseline approaches (Pearson correlation, Euclidean distance) restricted to this set of genes are clearly inferior to REx in the cross-validation performance (Table 1; rows 1, 7 and 8), and hence REx has a competitive advantage here.

We tested two variants of REx: the first one learned the expression model by choosing predictors from among all the other genes of the total set of 13 262 measured genes; the second took only the other target genes into account. Interestingly, the second variant is only slightly worse than the first at the high precision end (top1 and top5 results), and it is even better than the first when looking at more than five retrieval results (Table 1; rows 1 and 2). This indicates that already the relationships among the eight target genes (i.e. at most 56 links) are sufficient for almost perfect retrieval performance. Both variants of REx achieved an enormous dimensionality reduction compared with the baseline methods that use all genes (the first REx variant used on average 351.4 links, the second 44.4); in addition to increasing the biological interpretability, REx led to the best retrieval results (Table 1; rows 1, 2, 5 and 6).

Finally, we benchmarked the FK similarity used in REx against the simpler baseline similarity criteria (Pearson correlation, Euclidean distance) when controlling the dimensionality of data representations to be equal. For that purpose, we constructed hybrid approaches that take the predictor genes from the learned expression model and then compute Pearson correlation or Euclidean distance on their expression values. The similarity criterion of REx, which considers modeled relationships between predictor genes and target genes, achieved better results in the retrieval task (Table 1; rows 1, 3 and 4). Variances across the 567 queries were generally high, but with REx they were lower than with the other methods. For instance, at the top100 results, REx had slightly lower average precision, but also lower variance than the hybrid method with Pearson correlation, which had the best average precision at that level.

3.2 Plant stress case study

Besides being efficient and effective in finding relevant data, retrieval by targeted models can assist in analyzing molecular pathways related to specific biological processes. Here, we present a study on osmotic stress regulation in *Arabidopsis thaliana*.

3.2.1 Data Collection, preprocessing and annotation We downloaded 38 raw Affymetrix ATH1 datasets from NASCArrays, (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl) ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/), Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), and The Integrated Microarray Database System (http://ausubellab.mgh.harvard.edu/imds). All measurements were related to stress conditions in *A.thaliana*. By applying standard preprocessing methods, we obtained 141 differential expression profiles across 6658 differentially expressed genes (see Supplementary Material for detailed information about the data and preprocessing). Each profile was manually annotated with respect to 24 binary stress-related attributes.

3.2.2 Retrieval of osmotic stress measurements One widely studied type of plant stress is osmotic stress, which relates to dehydration of the plant caused by environmental conditions such as drought, high salt concentration or cold (Boudsocq and Laurière, 2005). The transcription factor DREB2A has been found to play a central role in osmotic stress response in *A.thaliana*; it interacts with a dehydration-responsive element (DRE). Sakuma *et al*. (2006) experimentally induced overexpression of a constitutively active form of DREB2A and published a list of 36 genes that responded with expression changes greater than eight times, relative to control plants. We refer to this list, together with DREB2A, as 'Sakuma-all' later on. Note that the underlying experiment is not included in the data compendium we use in our analysis. Nine out of these strongly overexpressed genes are annotated as water-stress-related and have a DRE element in the upstream region. We call this list, with the addition of DREB2A, 'Sakuma-water'. Moreover, we downloaded from STIFDB (http://caps.ncbs.res.in/stifdb/browse.html#se)

Table 1. Retrieval performance of targeted model-based retrieval of REx and baseline methods in a leukemia case study

	Mean precision (and sdev) across all queries (in %)	top1	top5	top10	top20	top50	top100
1	REx: predictors among all other genes	**97.53** (15.53)	**93.93** (15.68)	91.75 (17.86)	89.40 (18.24)	85.78 (18.69)	82.21 (21.01)
2	REx: predictors among other targets	95.59 (20.55)	93.40 (18.21)	**92.42** (17.34)	**91.02** (17.59)	**87.31** (20.36)	81.41 (23.34)
3	Hybrid: Pearson correlation on predictors	84.30 (36.41)	83.63 (31.41)	82.77 (31.21)	82.62 (30.12)	83.22 (27.09)	**83.38** (25.07)
4	Hybrid: Euclidean distance on predictors	80.07 (39.98)	77.04 (36.52)	75.31 (36.53)	72.74 (36.49)	68.12 (37.29)	63.28 (38.16)
5	Baseline: Pearson correlation on all genes	79.72 (40.25)	77.78 (36.78)	76.40 (36.60)	75.45 (36.16)	73.15 (34.39)	70.69 (33.65)
6	Baseline: Euclidean distance on all genes	78.13 (41.37)	74.67 (37.32)	72.59 (37.21)	70.00 (37.28)	66.00 (37.14)	62.20 (36.73)
7	Baseline: Pearson correlation on targets only	57.50 (49.48)	52.73 (40.19)	51.08 (38.21)	48.20 (35.97)	42.34 (31.77)	36.17 (26.43)
8	Baseline: Euclidean distance on targets only	73.90 (43.96)	71.15 (40.23)	70.49 (39.76)	68.84 (39.64)	65.64 (39.10)	61.88 (38.40)

Abbreviations for the methods used later on in the article: 1. REx, 2. REx (targets), 3. Corr. (predictors), 4. Eucl. (predictors), 5. Corr., 6. Eucl., 7. Corr. (targets) and 8. Eucl. (targets).
Mean average precision among the top-*k* results is given for several *k* (taken across all queries). The best value in each column is marked in bold.

(Shameer *et al.*, 2009) a list of 41 genes annotated as 'drought-salt-cold', which originally has been derived by integrating various microarray datasets. This list is named 'STIFDB' hereafter. 'Sakuma-all' and 'STIFDB' share only four genes, since DREB2A is not activated by cold stress.

Our data collection contained 31 osmotic stress samples from five different datasets (comprising at least six samples). We investigated the retrieval performance of REx by cross-validation in a leave-one-dataset-out manner, using the same six comparison methods as in the leukemia case above. Figure 1 shows the resulting precision–recall curves for the three different lists of input genes: Sakuma-water, Sakuma-all and STIFDB. In all cases, REx performs better than the baseline methods. The best retrieval results are obtained with Sakuma-water and REx, closely followed by the REx variant restricted to only operate on the target genes; the top false positives are shown in the Supplementary Material. For the other two gene lists, restricting the REx model only to predictors among the target genes was actually beneficial for the retrieval. One possible reason for this observation is that these lists are larger than Sakuma-water and thereby can already explain by themselves the expression of some central target genes sufficiently well, whereas less relevant external predictors can decrease the performance. Next, we take a closer look at the osmotic stress network learned around Sakuma-water targets.

3.2.3 Analysis of osmotic stress gene network The REx method extracted 29 additional genes with putative relationships to genes in Sakuma-water; they form three networks which we refer to as DREB2A, RD29A and AT2G46140 after the central target of each network (Fig. 2). Target genes are depicted as boxes. Arrows go from putative predictors to targets; they do not indicate regulatory relationships. Only one negative (i.e. suppressive) gene relationship was found (dashed edge). Black edges indicate relationships that are strengthened in osmotic stress relative to the base model (i.e. the corresponding coefficients move further away from zero in >75% of the osmotic stress samples). The remaining, gray edges mark relationships with larger heterogeneity.

Figure 3 provides an overall view of the expression changes across all datasets in terms of model-based Fisher scores with respect to the three networks. The RD29A network is the largest of the three networks; it is activated by drought and osmotic stresses, by cold stress (which has an osmotic stress component), and by the plant hormone abscisic acid, also a regulator of abiotic stress. The DREB2A network shows some osmotic stress regulation and regulation by UV-B radiation. Closer inspection of the genes in this network indicated that they could be part of the plant heat shock response. We validated this hypothesis against an independent dataset containing a time course heat shock experiment (Supplementary Fig. S2). Indeed, the genes showed rapid expression response to heat shock. The role of DREB2A as regulator in osmotic stress and heat shock response is well established (Yoshida *et al.*, 2008). The smallest network, centered around AT2G46140, is a novel finding suggesting that DREB2A and some of its targets also have a role in responses to pathogen infection and pathogen elicitors.

We checked the significance of the inferred gene relationships in a bootstrapping experiment (see Supplementary Material). The most prevalent relationships were well supported by functional annotations. In addition, concordant expression between the transcription factor DREB2A and the predictors AT3G62260 and ZAT12 was validated in two independent datasets, giving rise to interesting biological hypotheses (see Supplementary Material). While further experimental studies are needed to untangle regulatory mechanisms of stress genes, our results show that the relationships discovered by REx are relevant in recognizing osmotic stress conditions.

3.2.4 Smallest discriminative set of target genes In Figure 1, we observed that the list of targets (Sakuma-water) is more discriminative on our data compendium than the larger lists Sakuma-all and STIFDB. Next, we were interested whether even smaller lists of target genes have discriminative power. For that purpose, we tested the retrieval performance with reduced versions of Sakuma-water. More specifically, for given sizes of the target list, we used cross-validation on the training set to choose the best subset of Sakuma-water satisfying the size constraint, where the quality criterion was average precision (across the whole recall range). This best subset was used to learn a model with all training data, which was then tested on left-out test data. Training and test data were defined by cross-validation in exactly the same way as for Figure 1. Supplementary Figure S4a shows the precision–recall curves obtained on the test data (averaged across all queries) for different sizes of the target list. Target lists of size 3 outperformed the original target list of size 10. The exact composition of the subset differed between the cross-validation folds, but only the following genes occurred (in decreasing order): RD29A, LEA7, COR15A, AT3G17520 and LSR3. Remarkably, a model with a single target gene performed very well, except in the top precision region. In all cases, RD29A was the selected gene (*r*esponsive to *d*ehydration). However, while selected subsets of targets can be very powerful, on average the retrieval performance decreases monotonically when reducing the size of the target list (Supplementary. Fig. S4b).

3.2.5 Robustness against nuisance target genes A further question is whether errors in the target list are harmful for the retrieval. To study this, we added randomly picked genes to Sakuma-water; they were chosen among the other 6648 genes (genes showing only minor differential expression in all datasets were removed in the preprocessing). We did 50 repeats on each number of additional genes. Supplementary Figure S4c shows the average retrieval performance. As expected, a larger number of added random genes led to a stronger decrease in precision. However, the change was not very dramatic, implying that a reasonable number of unrelated genes can be tolerated quite well as long as the target list contains also discriminative genes.

4 DISCUSSION

We introduced a novel approach for targeted model-based retrieval of gene expression measurements. The model we proposed is suitable for efficient retrieval; due to the decomposition into gene-specific submodels, learning can be done offline, prior to the queries. Although we used L1-regularized regression to learn

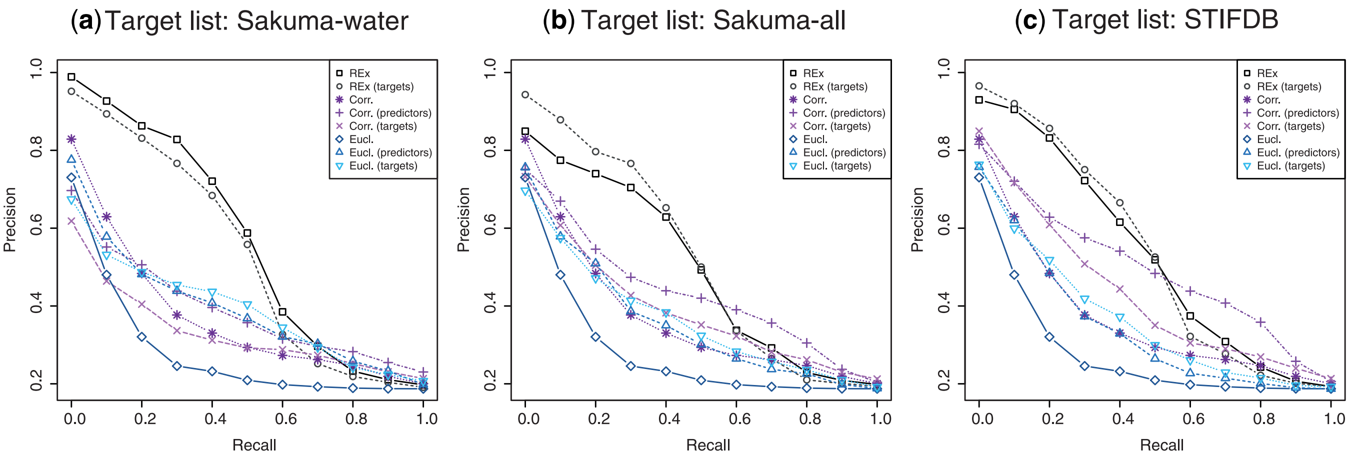

Fig. 1. Osmotic stress retrieval performance of several methods for three different gene lists of interest (see text for details). For the meaning of method abbreviations, see Table 1

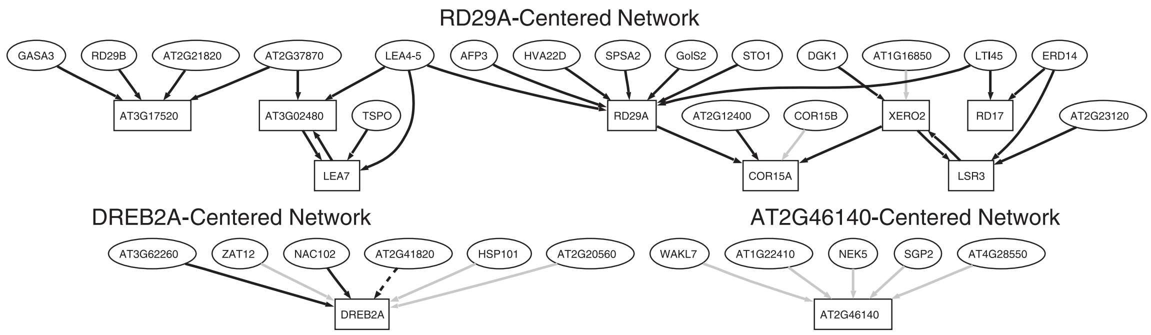

Fig. 2. Osmotic stress network learned around Sakuma-water targets (box-shaped). Arrows point from predictors to targets. The dashed edge indicates a negative relationship. Black edges are increased in weight for a majority of osmotic stress samples, compared with the background model. See text for details

a simple regulatory model in this work, the approach is not limited to a specific type of probabilistic models. For instance, it is straightforward to use the elastic net approach (Zou and Hastie, 2005), and it is also possible to exploit additional data (e.g. protein–protein interactions) or prior knowledge during learning. If the user-specified genes are modeled jointly (e.g. by common predictors), more computational effort is required at query time; alternatively, the model may be provided directly by the user. In contrast to condition-specific regulatory models (Shimamura *et al*., 2010; Zhang and Wang, 2010), our method does not rely on annotation of samples. The current modeling approach requires that data are comparable. To apply it in cross-platform analysis, previous methods to achieve comparability can be used (Stafford and Brun, 2007). An interesting direction for future work is retrieval across different platforms or species using an integrated model.

Besides allowing for unsupervised data-based retrieval of related measurements, the proposed method assists in investigating relationships among genes. As illustrated in the osmotic stress example, measurements with curated annotation can help to assess the quality of models and to detect condition-dependent activity changes. When applying targeted modeling, an important question is how to choose the target list for a biological process of interest. If it is too narrow, it might not have sufficient discriminative power; if it is too wide, the performance might suffer due to irrelevant genes. The retrieval framework allows the user to check different possibilities and to interactively improve the target list for her purposes. It also makes possible exploration of the types of retrieval results a particular target list yields, exploiting different kinds of annotations that are available for the database content. This can be useful if the user is interested in specific genes, but uncertain about the processes they are involved in. Finally, the targeted modeling and retrieval approach can be a useful analysis tool also for other omics-type high-throughput data.

ACKNOWLEDGEMENTS

We are grateful to A. Faisal for fruitful discussions, preparatory work and helpful comments. We thank L. Lahti, J. Rung, K. Overmyer, M. Wrzaczek and A. Brazma for very interesting discussions.

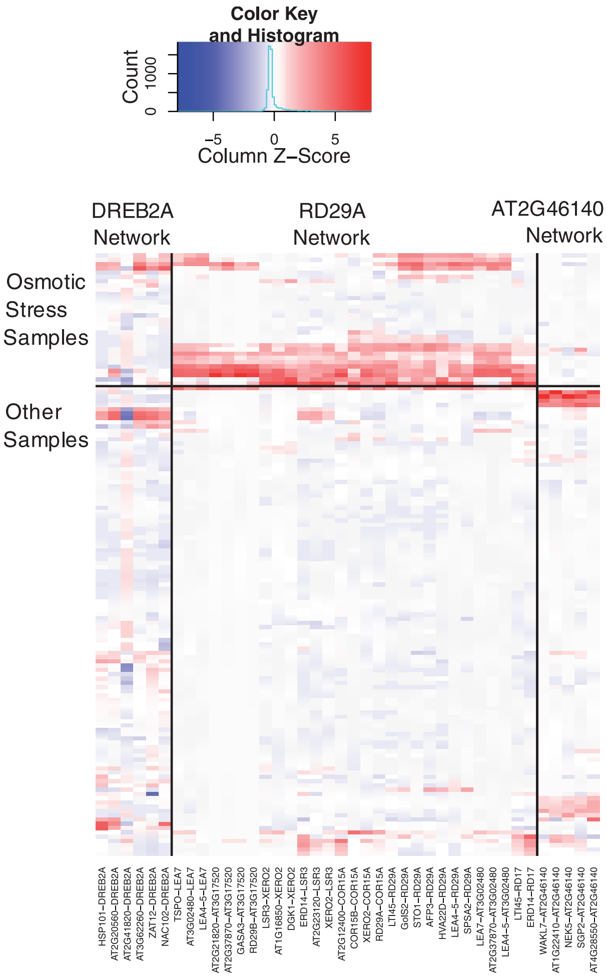

Fig. 3. Heatmap of Fisher scores of expression profiles (rows) with respect to the networks in Figure 2; columns represent network links. The horizontal line separates osmotic stress samples from the other samples, the vertical lines separate the networks. The RD29A network is highly activated in osmotic stress samples. The DREB2A network is activated in UV-B samples and some osmotic stress samples. The AT2G46140 network is induced by pathogen infection and treatment with pathogen elicitors. Annotation of each sample can be found in Supplementary Figure S1

Funding: The work was financially supported by Tekes–the Finnish Funding Agency for Technology and Innovation (40141/07 to E.G., J.S., J.K. and S.K.); the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170); the Academy of Finland Center of Excellence in Plant Signal Research 2006–2011; and the IST Programme of the European Community under the PASCAL2 Network of Excellence (IST-2007-216886). This publication only reflects the authors' views.

Conflict of Interest: none declared.

REFERENCES

Alpaydin,E. (2010) *Introduction to Machine Learning*. 2nd edn. MIT Press, Cambridge, MA.

Ambroise,C. *et al.* (2009) Inferring sparse Gaussian graphical models with latent structure. *Electron J. Stat.*, **3**, 205–238.

Besag,J. (1975) Statistical analysis of non-lattice data. *J. R. Stat. Soc. Ser. D Statist.*, **24**, 179–195.

Boudsocq,M. and Laurière,C. (2005) Osmotic signaling in plants. Multiple pathways mediated by emerging kinase families. *Plant Physiology*, **138**, 1185–1194.

Caldas,J. (2009) Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, **25**, i145–i153.

Caldas,J. *et al.* (2012) Data-driven information retrieval in heterogeneous collections of transcriptomics data links SIM2s to malignant pleural mesothelioma. *Bioinformatics*, **28**, 246–253.

Dempster,A.P. (1972) Covariance selection. *Biometrics*, **28**, 157–175.

Dobra,A. *et al.* (2004) Sparse graphical models for exploring gene expression data. *J. Multivariate Analy.*, **90** (1), 196–212.

Engreitz,J. *et al.* (2010) Content-based microarray search using differential expression profiles. *BMC Bioinformatics*, **11**, 603.

Feng,C. (2009) GEM-TREND: a web tool for gene expression data mining toward relevant network discovery. *BMC Genomics*, **10**, 411.

Fujibuchi,W. *et al.* (2007) CellMontage: similar expression profile search server. *Bioinformatics*, **23**, 3103–3104.

Greene,C.S. and Troyanskaya,O.G. (2011) Pilgrm: an interactive data-driven discovery platform for expert biologists. *Nucleic Acids Res.*, **39** (**Suppl. 2**), W368–W374.

Heckerman,D. *et al.* (2001) Dependency networks for inference, collaborative filtering, and data visualization. *J. Mach. Learn. Res.*, **1**, 49–75.

Hibbs,M.A. *et al.* (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.

Huang,H. (2010) Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc. Nat. Acad. Sci. USA*, **107**, 6823–6828.

Huttenhower,C. and Troyanskaya,O. (2008) Assessing the functional structure of genomic data. *Bioinformatics*, **24**, i330–i338.

Jaakkola,T.S. and Haussler,D. (1999) Exploiting generative models in discriminative classifiers. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II.* MIT Press, Cambridge, MA, USA, pp. 487–493.

Lahti,L. *et al.* (2010) Global modeling of transcriptional responses in interaction networks. *Bioinformatics*, **26**, 2713–2720.

Le,H.-S. *et al.* (2010) Cross-species queries of large gene expression databases. *Bioinformatics*, **26**, 2416–2423.

Lukk,M. *et al.* (2010) A global map of human gene expression. *Nat. Biotechnol.*, **28**, 322–324.

Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **1**, 24–45.

Markowetz,F. and Spang,R. (2007) Inferring cellular networks—a review. *BMC Bioinformatics*, **8** (**Suppl. 6**), S5.

Meinshausen,N. and Bühlmann,P. (2006) High dimensional graphs and variable selection with the Lasso. *Ann. Stat.*, **34**, 1436–1462.

Parkinson,H. *et al.* (2009) Arrayexpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37** (**Suppl. 1**), D868–D872.

Sakuma,Y. *et al.* (2006) Dual function of an Arabidopsis transcription factor DREB2A in water-stress-responsive and heat-stress-responsive gene expression. *Proc. Nat. Acad. Sci. USA*, **103**, 18822–18827.

Schäfer,J. and Strimmer,K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article 32.

Schmidt,M.W. *et al.* (2008) Structure learning in random fields for heart motion abnormality detection. *Computer Vision and Pattern Recognition (CVPR 2008)*, Anchorage, Alaska, USA.

Segal,E. *et al.* (2004) A module map showing conditional activity of expression modules in cancer. *Nat Genet*, **36**, 1090–1098.

Shameer,K. (2009) STIFDB—Arabidopsis stress responsive transcription factor database. *Int. J. Plant Genomics*, **2009**, 583429.

Shawe-Taylor,J. and Cristianini,N. (2004) *Kernel Methods for Pattern Analysis.*. Cambridge University Press, New York, NY, USA.

Shimamura,T. *et al.* (2010) Inferring dynamic gene networks under varying conditions for transcriptomic network comparison. *Bioinformatics*, **26**, 1064–1072.

Stafford,P. and Brun,M. (2007) Three methods for optimization of cross-laboratory and cross-platform microarray expression data. *Nucleic Acids Res.*, **35**, e72.

Suthram,S. *et al*. (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput Biol*., **6**, e1000662.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. Ser. B Methodol*., **58**, 267–288.

Yoshida,T. *et al*. (2008) Functional analysis of an Arabidopsis heat-shock transcription factor HsfA3 in the transcriptional cascade downstream of the DREB2A stress-regulatory system. *Biochem Biophys Res Commun*, **368**, 515–521.

Zhang,B. and Wang,Y. (2010) Learning structural changes of Gaussian graphical models in controlled experiments. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*. AUAI Press, Catalina Island, CA, USA, pp. 701–708.

Zhu,Y. *et al*. (2008) GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, **24**, 2798–2800.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. Ser. B Methodol*, **67**, 301–320.