Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 1 2012 *Article* 2

Transcriptional Network Inference from Functional Similarity and Expression Data: A Global Supervised Approach

Jérôme Ambroise, *Université Catholique de Louvain*
Annie Robert, *Université Catholique de Louvain*
Benoit Macq, *Université Catholique de Louvain*
Jean-Luc Gala, *Université Catholique de Louvain*

Recommended Citation:
Ambroise, Jérôme; Robert, Annie; Macq, Benoit; and Gala, Jean-Luc (2012) "Transcriptional Network Inference from Functional Similarity and Expression Data: A Global Supervised Approach," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 1, Article 2.
DOI: 10.2202/1544-6115.1695
Available at: http://www.bepress.com/sagmb/vol11/iss1/art2

©2012 De Gruyter. All rights reserved.

Transcriptional Network Inference from Functional Similarity and Expression Data: A Global Supervised Approach

Jérôme Ambroise, Annie Robert, Benoit Macq, and Jean-Luc Gala

Abstract

An important challenge in system biology is the inference of biological networks from postgenomic data. Among these biological networks, a gene transcriptional regulatory network focuses on interactions existing between transcription factors (TFs) and and their corresponding target genes. A large number of reverse engineering algorithms were proposed to infer such networks from gene expression profiles, but most current methods have relatively low predictive performances. In this paper, we introduce the novel TNIFSED method (Transcriptional Network Inference from Functional Similarity and Expression Data), that infers a transcriptional network from the integration of correlations and partial correlations of gene expression profiles and gene functional similarities through a supervised classifier. In the current work, TNIFSED was applied to predict the transcriptional network in *Escherichia coli* and in *Saccharomyces cerevisiae*, using datasets of 445 and 170 affymetrix arrays, respectively. Using the area under the curve of the receiver operating characteristics and the F-measure as indicators, we showed the predictive performance of TNIFSED to be better than unsupervised state-of-the-art methods. TNIFSED performed slightly worse than the supervised SIRENE algorithm for the target genes identification of the TF having a wide range of yet identified target genes but better for TF having only few identified target genes. Our results indicate that TNIFSED is complementary to the SIRENE algorithm, and particularly suitable to discover target genes of "orphan" TFs.

KEYWORDS: transcriptional network, supervised microarray

Author Notes: The authors gratefully acknowledge Philippe Baret for his advice concerning gene identifiers. Funding: Jérôme Ambroise is funded by Nanotic/Tsarine, a project of the Region Wallonne of Belgium (convention number: 516250). Jérôme Ambroise and Benoit Macq are members of the ICTEAM Institute; Annie Robert and Jean-Luc Gala are members of the IREC Institute.

1 Introduction

The success of genome sequencing projects in a range of organisms has contributed to a progressive identification of every single expressed gene. The next important task is to assign a function to each of these genes. In this context, an important issue remains the inference of biochemical pathways and regulatory networks from postgenomic data. Genes and gene products are indeed tightly connected in structured biological networks. In transcriptional regulatory networks, trans-acting transcription factors (TFs) bind to cis-regulatory sequence elements of their respective target genes, hence modulating mRNA transcription. In order to compensate for the lack of comprehensive catalog of these transcriptional regulatory elements including cis-elements and transcription factors, in silico predictive methods have been proposed. They aim to reconstruct such regulatory networks from available genomic and post-genomic data.

A large number of algorithms have been proposed to infer a transcriptional regulation network from gene expression profiles (Markowetz and Spang, 2007, Bansal, Belcastro, Ambesi-Impiombato, and Di Bernardo, 2007). Some of these algorithms are specific to time-series data or data obtained from experimental interventions and perturbations. These approaches include Boolean network, ordinary and stochastic partial differential equation (Akutsu, Miyano, and Kuhara, 2000, Bickel, 2005, Perrin, Ralaivola, Mazurie, Bottani, Mallet, and d'Alche Buc, 2003, Martin, Zhang, Martino, and Faulon, 2007, Hickman and Hodgman, 2009, Gardner, di Bernardo, Lorenz, and Collins, 2003, di Bernardo, Thompson, Gardner, Chobot, Eastwood, Wojtovich, Elliott, Schaus, and Collins, 2005). In the current work, we rather focus on transcriptional network construction from steady-state expression data.

Among the algorithms that aim to reconstruct transcriptional network from steady-state data, the Relevance network method, proposed by Butte and Kohane (2000), is based on pairwise association scores. Scores like the Pearson correlation or the mutual information are computed for all gene pairs using expression profiles obtained from microarrays experiments. The main disadvantage of RN is its relative inability to distinguish between direct and indirect interactions. To circumvent this problem, Basso, Margolin, Stolovitzky, Klein, Dalla-Favera, and Califano (2005) developed the ARACNE algorithm which discards any potentially indirect statistical dependency (Margolin, Wang, Lim, Kustagi, Nemenman, and Califano, 2006, Hartemink, 2005). Another popular solution in the literature implies to compute the partial correlation coefficients. These coefficients measure the correlation between two genes conditional to either one, several or all other genes in the network (Kramer, Schafer, and Boulesteix, 2009). In this context, De La Fuente, Bing, Hoeschele, and Mendes (2004) proposed a method that organizes genes in undi-

rected dependency graphs based on partial coefficients up to order 2. If a multivariate Gaussian distribution is assumed, the estimation of full conditional partial correlation matrix, i.e the Graphical Gaussian Model (GGM), involves either the inversion of the sample covariance matrix, or the estimation of p least squares regression problems. If the number p of variables (genes) is much larger than the number n of experiments (microarrays), such approaches are inappropriate. Suitable alternatives are based either on regularized estimation of the inverse covariance matrix, or on regularized high dimensional regression (Schafer and Strimmer, 2005a, Li and Gui, 2006, Schafer and Strimmer, 2005b, Peng, Wang, Zhou, and Zhu, 2009, Friedman, Hastie, and Tibshirani, 2008, Zhou, Van De Geer, and Buhlmann, 2009, Castelo and Roverato, 2009). As Relevance Network and GGM assume both that experiments are statistically independent from each other, each approach was designed to integrate steady-state gene expression profiles. However, they can also be applied to time-series expression data if the sampling time is long enough to assume that each point is independent from the previous one. Statistical methods such as Bayesian networks were also applied to reconstruct transcriptional network. A Bayesian network is a graphical model establishing the probabilistic relationship among a set of random variables. One of the first publications promoting this class of algorithm was dedicated to the inference of a gene regulatory network in *Saccharomyces cerevisae* from gene expression profiles using Bayesian networks (Friedman, Linial, Nachman, and Pe'er, 2000). Several authors have then elaborated methods to integrate microarray gene expression data and others biological significances in Bayesian networks (Imoto, Higuchi, Goto, Tashiro, Kuhara, and Miyano, 2003, Werhli and Husmeier, 2007).

Algorithms have largely been compared in the literature (Soranzo, Bianconi, and Altafini, 2007, Werhli, Grzegorczyk, and Husmeier, 2006, Hickman and Hodgman, 2009, Cho, Choo, Jung, Kim, Choi, and Kim, 2007, Luo, Hankenson, and Woolf, 2008, Meyer, Kontos, Lafitte, and Bontempi, 2007, Faith, Hayete, Thaden, Mogno, Wierzbowski, Cottarel, Kasif, Collins, and Gardner, 2007). Faith et al. (2007) compared several approaches, including Bayesian networks (Friedman et al., 2000), ARACNE (Basso et al., 2005) and the context likelihood of relatedness (CLR) algorithm, a method that extends the relevance network algorithm (Butte and Kohane, 2000). Each algorithm was applied on a dataset of 445 *Escherichia coli* Affymetrix arrays and performances were assessed by comparing the algorithm predictions with a list of 3216 known *E. coli* regulatory interactions listed in RegulonDB (Gama-Castro, Jimenez-Jacinto, Peralta-Gil, Santos-Zavaleta, Penaloza-Spinola, Contreras-Moreira, Segura-Salazar, Muniz-Rascado, Martinez-Flores, Salgado et al., 2008). Accordingly, it was observed that CLR outperformed all other methods in terms of prediction accuracy. Furthermore, some new predictions resulting from the CLR model were experimentally validated. The fraction of

DOI: 10.2202/1544-6115.1695

known regulatory interactions that are correctly identified with CLR is nevertheless low (338/3216) at high precision (60%) (Faith et al., 2007).

As the algorithms based on expression data predict only a limited number of regulations at high precision level, other approaches integrating expression data and complementary molecular information have been proposed. Hecker, Lambeck, Toepfer, van Someren, and Guthke (2009) reviewed approaches that enable the modeling of dynamic gene regulatory systems by using time-course gene expression data and various types of molecular biological information such as genome sequences and protein-DNA interaction data. Sun, Tuncay, Haidar, Ensman, Stanley, Trelinski, and Ortoleva (2007) developed a methodology, named TRND (Transcriptional regulatory network discovery) that integrates a preliminary transcriptional regulatory network to microarray data, gene ontology and phylogenic similarity in order to identify additional gene targets. Allocco, Kohane, and Butte (2004) have indeed shown that genes with similar functional annotations are also more likely to be bound by a common TF.

Another approach to improve the predictive performance was the Supervised Inference of Regulatory Networks (SIRENE) method developed by Mordelet and Vert (2008). SIRENE differs fundamentally from other approaches in the sense that it is a supervised method which requires the prior knowledge of a set of well defined regulations in order to train a Support Vector machine (SVM). For each TF, a local model is trained in order to discriminate the genes regulated by this TF from other genes, by considering their expression profiles. An important limitation of SIRENE is that the predictive performance tends to decrease proportionally to the number of known targets (Mordelet and Vert, 2008). If no target gene has yet been identified for a particular TF, this algorithm is unable to predict its potential target genes. SIRENE can therefore not be used to detect new TF nor to identify potential target genes of 'orphans' TFs.

In this paper, we introduce a new method named TNIFSED (Transcriptional Network Inference from Functional Similarity and Expression Data). TNIFSED integrates correlation and partial correlation of gene expression profiles with gene functional similarities trough a supervised classifier in order to identify target genes of known TFs. Unlike the SIRENE method, TNIFSED uses a single global model allowing identification of target genes related to 'orphan' TFs.

2 Methods

2.1 Data

2.1.1 *Escherichia coli*

TNIFSED was applied to infer the transcriptional Regulation network of *E. coli* by using the microarray dataset of 445 *E. coli* Affymetrix arrays presented by Faith et al. (2007). These expression data that have been collected under distinct experimental conditions such as growth phase, antibiotic exposure, heat shock, pH, and genetic alteration, were downloaded from the website (*http://m3d.bu.edu/*). The probe names of the expression dataset were converted in gene names and average profiles were computed for genes associated with several probes (Additional file 1). Among the 4,275 measured genes, we identified 177 TFs using the list of interacting gene pairs (Additional file 2) from the RegulonDB 7.2 (Gama-Castro et al., 2008) released in May 2011. As our method is applied on steady-state expression data, self-regulation of the TFs are not considered in this work and each TF has consequently 4,274 potential target genes. Among the 756,498 (177 * 4,274) gene pairs, 3,703 pairs were identified in the RegulonDB 7.2 and were considered interacting, hence belonging to positive training examples. The remaining 752,795 gene pairs were considered non-interacting.

2.1.2 *Saccharomyces cerevisiae*

TNIFSED was also applied to infer the transcriptional Regulation network of *S. cerevisiae* by using the microarray compendium dataset of 170 Affymetrix arrays presented by Knijnenburg, Daran, Van Den Broek, Daran-Lapujade, De Winde, Pronk, Reinders, and Wessels (2009). Raw expression data were downloaded from the GEO (Edgar, Domrachev, and Lash, 2002) website (Series GSE11452) and the *gcrma* Bioconductor package was used to obtain the expression matrix (Additional file 3). Using the list of known regulation presented by Balaji, Babu, Iyer, Luscombe, and Aravind (2006) as reference network (Additional file 4), we identified 155 TFs among the 6,058 measured genes. As each TF has 6,057 potential target genes, there is a total of 938,835 (155 * 6,057) gene pairs. Among those, 11,588 pairs were identified in the list of known regulations and were therefore considered interacting, hence to be part of the positive training examples. The remaining 927,247 gene pairs were considered non-interacting.

DOI: 10.2202/1544-6115.1695

2.2 TNIFSED

TNIFSED is a supervised inference algorithm integrating gene expression data and functional similarities to infer new regulation relationships between known or potential TFs and their target genes. Like with any other supervised classification algorithm, a training data set is needed, that is made of a set of positive and negative training examples. In TNIFSED algorithm, the list of examples includes all gene pairs involving at least one TF. In this list, positive training examples correspond to gene pairs that involve a TF and one of its target genes. The list of such interacting gene pairs can be downloaded from database such as RegulonDB for the *E. coli* organism. If a gene pair is not present in the database, this pair is considered to be non-interacting, i.e. a negative example. However, the absence of a gene pair in a database does not guarantee the lack of interaction between both genes. Accordingly, the list of negative examples contains true-negative as well as false-negative examples. To ensure optimal predictive performance of TNIFSED, the training dataset should include TFs for which a high proportion of target genes are already identified. A functional similarity is computed for each gene pair in the training dataset from each of the following ontology category: the cellular component (CC), the biological process (BP) and the molecular function (MF). Moreover, the absolute values of correlation and partial correlation coefficients of gene expression profiles are computed for every gene pairs. The three functional similarities and both correlation coefficients set up the continuous predictor variables of a logistic regression model (Figure 1). This logistic model is trained to discriminate interacting from non-interacting gene pairs based on the five predictor variables.

2.2.1 Gene functional similarity

The functional similarities of all gene pairs was computed for both datasets using a method developed by Wang, Du, Payattakool, Yu, and Chen (2007) which takes into account gene annotation information encoded in the GO-terms (Gene Ontology) semantic (Sun et al., 2007). In a first step, the algorithm of Wang *et al.* measures the semantic similarity of GO terms by taking into account the whole structure of the Directed Acyclic Graph encoding each ontology. In a second step, the similarities between GO terms are used in the algorithm designed to measure the functional similarities of a gene pair of interest. In this work, the method of Wang *et al.*, implemented in the 'GOSemSim' package (Yu, Li, Qin, Bo, Wu, and Wang, 2010) of the Bioconductor 2.8 project (published in April 2011) (Gentleman, Carey, Bates, Bolstad, Dettling, Dudoit, Ellis, Gautier, Ge, Gentry et al., 2004), was used to obtain one functional similarity for each gene pair and from each ontology (BP, CC and MF).

TfGene	RegGene	Pcorrelation	correlation	SimCC	SimBP	SimMF	Interaction
acrR	abgA	0.011	0.268	0.000	0.249	0	0
acrR	abgB	0.021	0.217	0.000	0.249	0	0
acrR	abgR	0.018	0.205	1.000	0.619	1	0
...	...	...	...	...	...	...	...
acrR	zur	0.005	0.284	1.000	1.000	1	0
ada	abgA	0.009	0.710	0.000	0.249	0	0
ada	abgB	0.002	0.696	0.000	0.249	0	0
ada	abgR	0.008	0.724	1.000	0.702	1	0
ada	abgT	0.003	0.743	0.487	0.000	0	0
...	...	...	...	...	...	...	...
ada	zur	0.001	0.560	1.000	1.000	1	0
adiY	abgA	0.015	0.422	0.000	0.249	0	0
adiY	abgB	0.003	0.415	0.000	0.249	0	0
adiY	abgR	0.002	0.456	0.000	1.000	1	0
...	...	...	...	...	...	...	...
adiY	zur	0.005	0.566	0.000	1.000	0.661	0
agaR	abgA	0.001	0.633	0.000	0.303	0	0
agaR	abgB	0.007	0.643	0.000	0.303	0	0
agaR	abgR	0.011	0.605	0.000	0.654	1	0
...	...	...	...	...	...	...	...
agaR	zur	0.023	0.749	0.000	1.000	1	0
allR	abgA	0.014	0.235	0.000	0.249	0	0
allR	abgB	0.000	0.238	0.000	0.249	0	0
allR	abgR	0.005	0.239	1.000	1.000	1	0
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

TRAINING OF THE LOGISTIC REGRESSION MODEL

$$\log\left[\frac{P(y=1|x_1,x_2,x_3,x_4,x_5)}{P(y=0|x_1,x_2,x_3,x_4,x_5)}\right] = \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5$$

TfGene	RegGene	Pcorrelation	correlation	SimCC	SimBP	SimMF	P(Interaction)
csiR	abgA	0.0114	0.3883	0.0000	0.2490	0.0000	0.011
csiR	abgB	0.0026	0.4329	0.0000	0.2490	0.0000	0.211
csiR	abgR	0.0018	0.4498	0.0000	1.0000	1.0000	0.114
...	...	...	...	...	...	...	...
csiR	zur	0.0015	0.4238	0.0000	1.0000	1.0000	0.002
cspA	abgA	0.0039	0.5364	0.0000	0.2490	0.0000	0.032
cspA	abgB	0.0137	0.5589	0.0000	0.2490	0.0000	0.811
cspA	abgR	0.0002	0.5340	1.0000	0.6190	1.0000	0.701

Figure 1: Pipeline of the TNIFSED algorithm. For each gene pair of the training dataset, functional similarities (i.e., cellular component, biological process and molecular function), as well as correlation and partial correlation of expression profiles are computed. The presence (1) or absence (0) of interaction is extracted from a database such as the RegulonDB for the *E. coli* organism. The training dataset is used to assess parameters of the logistic regression model with a maximum likelihood procedure.

DOI: 10.2202/1544-6115.1695

2.2.2 Correlation and partial correlation

For each gene pair, the absolute values of the correlation and partial correlation coefficients related to each expression profile were computed. Both correlation coefficients have their respective advantages. While correlation coefficients are weak criteria of dependence because they are unable to discriminate between direct and indirect interactions, partial correlation coefficients are a strong measure of dependence (Markowetz and Spang, 2007). However, unlike partial correlation networks (GGM), correlations networks can be accurately estimated even if the number of genes is much larger than the number of samples. In that case, the GGM computation requires a shrinkage approach, which potentially introduces a bias. In this work, the 'pcor.shrink' function of the *'corpcor'* package was used to obtain shrinkage estimates of partial correlations by using the method developed by Schafer and Strimmer (2005a).

2.2.3 Logistic regression

As previously explained, TNIFSED is a supervised approach that uses a logistic regression model in order to discriminate interacting from non-interacting gene pairs. The predictor variables are the three functional similarity scores and both correlation coefficients while the outcome is the presence (1) or absence (0) of an interacting gene pair. As a first step, the *logit* of the probability of interaction is modeled as a linear combination of the predictor variables. The five parameters of the model are assessed with a maximum likelihood procedure. The logistic model can then be used to predict the probability of interaction for gene pairs whose functional similarities and correlation coefficients were previously determined. In this work, logistic models were trained on both datasets using the *glm* R function. Because the number of negative training examples is much larger than the number of positive examples, weights were assigned to the negative training example in order to artificially reduce the total number of negative examples to the number of positive examples.

2.3 Comparison of predictive performances

TNIFSED was compared with unsupervised methods and with the SIRENE supervised method. For unsupervised methods, there is no parameter optimization using the training dataset. Regarding SIRENE, one local model is trained for each TF, so that n*p parameters are optimized using the training dataset, where n stands for the

number of conditions in the expression matrix and p for the number of TFs. Regarding the TNIFSED method, 5 similarity measures are computed using unsupervised approaches and then combined together by using 5 different weights that are optimized from the training dataset with a supervised approach. TNIFSED appears therefore to stand halfway between the fully data-driven SIRENE algorithm and the unsupervised methods. Accordingly, it makes sense to compare the TNIFSED predictive performance with both types of algorithms. In that respect, the continuous scores generated by all inference methods were compared for their power to discriminate presence and absence of interactions in the reference network, using the Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) and the maximum F-measure of the Precision Recall (PR) curve.

2.3.1 Predictions of unsupervised methods

Among the unsupervised algorithms, MRNET, CLR and ARACNE are implemented in the 'Minet' package developed by Meyer, Lafitte, and Bontempi (2008). The mutual information matrix associated to the expression profiles of the *E. coli* and of the *S. cerevisiae* organisms were computed with the 'mi.empirical' option and each algorithm was applied on the resulting matrix. These algorithms produced squared matrices of dimension n where n is the total number of genes in the expression matrices. From these matrices, we selected the 177 (155) columns corresponding to the TF in *E. coli* (*S. cerevisiae*). These columns represent indeed the scores of all gene pairs involving a least one TF. The same columns were also extracted from the correlation and the partial Correlation matrices and the absolute values were considered as the outputs of the Relevance network and the GGM. The partial correlation matrices were computed with the pcor.shrink function of the R *'corpcor'* package.

2.3.2 Predictions of TNIFSED

As TNIFSED is based on a global model, gene pair information of TFs can be shared both within a particular organism as well as between different organisms. Using the same data for training and for testing should nevertheless lead us to overestimate the predictive performance. In that respect, a 5-fold cross validation and an inter-organism generalization procedure were both applied for assessing the predictive performance of TNIFSED.

During the 5-fold cross-validation, the 177 TFs for *E. coli* and the 155 TFs for *S. cerevisiae* were divided in five parts. For both organisms, the logistic regression model was then trained on four parts and was used to predict the probability of

DOI: 10.2202/1544-6115.1695

interaction for the gene pairs involving TFs belonging to the remaining part. This procedure was repeated five times in order to predict interaction probabilities for each gene pair in the whole data sets. For the inter-organism generalization, the logistic model was trained on the data set from one organism and was used to predict the interaction probabilities of the gene pairs in the other organism.

As TNIFSED is based on a logistic regression model that includes only 5 variables which are plausible for classification from a biological perspective, no variable was preselected in the present analyses. A study of the predictive performances of embedded models was nevertheless performed in order to evaluate the relevance of each predictor variable.

2.3.3 Predictions of SIRENE

SIRENE is based on many supervised local models. For each TF, a local model is trained in order to discriminate the genes regulated by this TF from other genes, by considering their expression profiles. Using many independent local models prevents any information sharing between TFs of one or several organisms (Bleakley, Biau, and Vert, 2007). The SIRENE cross-validation must therefore be performed separately on each TF, as recommended by the authors (Mordelet and Vert, 2008). For each TF, a 3-fold cross-validation is performed. Genes are first separated in 3 subsets. For each iteration, a SVM is trained on two subsets and the SVM is then used to predict the target genes from the remaining subset.

As the SIRENE cross-validation must be performed on each TF separately, a particular attention must be paid to the existence of transcription units and operons (Mordelet and Vert, 2008). Considering that all genes within an operon have similar profiles and are regulated by the same TF, the SVM predictions are likely to be correct for any gene belonging to a particular operon if the SVM was trained on a dataset containing some genes of the operon. Performing classical cross-validation results therefore in a predictive performance evaluation which does not reflect the SIRENE performance to predict interactions involving new operons. In this context, contributors of SIRENE have developed a scheme that reliably estimates its predictive performance (Mordelet and Vert, 2008). The predictive performance obtained with classical cross-validation on *E. coli* was named SIRENE-BIAS while the predictive performance obtained with the contributors specific scheme was called SIRENE. This nomenclature is in line with the paper of the Authors (Mordelet and Vert, 2008). For *S. cerevisiae*, only the classical cross-validation, i.e. SIRENE-BIAS, was performed because complete information on operons is still lacking within the reference network (Balaji et al., 2006).

The SIRENE algorithm was downloaded from the author website and data were prepared as recommended (http://cbio.ensmp.fr/sirene/) for both organisms. The algorithm was slightly modified in order to perform cross-validation on all potential target genes (i.e. 4,274 for *E. coli* and 6,056 for *S. cerevisiae*).

2.3.4 Area under the curve of the Receiver Operating Characteristic

Continuous scores produced by unsupervised methods, by TNIFSED and by SIRENE were used to construct ROC curves. If the continuous prediction for a gene pair is higher than a determined threshold, this gene pair is considered as interacting. Inference methods generate therefore one network for each selected threshold. In order to compare the predictive performances of transcriptional network inference methods, a range of thresholds was applied to the scores produced with each method in order to generate ROC curves and to measure their AUC. The ROC curve plots the $sensitivity$ $(TP/(TP+FN))$ against $1-specificity$ $(FP/(FP+TN))$ for the consecutive thresholds (Steyerberg, Vickers, Cook, Gerds, Gonen, Obuchowski, Pencina, and Kattan, 2010). The AUC is equal to the probability that a predictive model assigns a higher score for a genuine interacting pair than for a non-interacting gene pair. Consequently, an AUC of 0.5 corresponds to a non informative model whereas an AUC of 1 indicates a perfect model. In this work, two average AUC's were computed. The micro-average AUC is based on a global ROC curve constructed from the predictions of gene pairs involving all TFs of the organism. On the other hand, the macro-average AUC is equal to the average of the AUC of individual ROC curves constructed for each TF separately. These two average AUC were computed with the MKmisc package (Kohl and Kohl, 2010) for both organisms and each inference method.

2.3.5 Precision-Recall and F-measure

Continuous scores produced by the methods were also used to plot Precision-Recall curves and to compute the maximum F-measure. The precision-recall curve plots the precision $(TP/(TP+FP))$ against the recall $(TP/(TP+FN)=sensitivity)$ for the consecutive thresholds. The precision-recall F-measure is a weighted harmonic mean of precision and recall $(F=(2\,Prec.Rec.)/(Prec.+Rec.))$. As one threshold corresponds to one F-measure, predictive performance of each inference method was measured as the maximum of the F-measures obtained on the whole threshold range. Two average versions of the F-measure were computed in this work. The micro-average F-measure is based on gene pair predictions of the entire and global set of TFs from one organism. Alternatively, the macro-average F-measure is equal

DOI: 10.2202/1544-6115.1695

to the average of the F-measures computed separately for each TF of the organism. These two average F-measures were computed with the ROCR package (Sing, Sander, Beerenwinkel, and Lengauer, 2005) for both organisms and each inference method.

2.3.6 Identification of false-negative interactions in a reference network

As SIRENE and TNFISED are two supervised methods, another performance comparison was performed on *E. coli*, in order to assess whether both methods were able to detect false-negative interactions in a reference network. TF-gene regulations reported in the RegulonDB can indeed safely be taken as true-positive interactions but TF-gene pairs which have not yet been reported in the database can indiscriminately be attributed either to true-negative or false-negative interactions (Mordelet and Vert, 2008).

In that respect, TNIFSED and SIRENE were both trained using the interactions related in RegulonDB 6.7 as reference network (released in March 2010, Additional file 5). For the TNIFSED method, functional similarities were computed using Bioconductor 2.6 (released in June 2010) and classical 5-fold cross validation was performed. For SIRENE, predictions were separately performed for each TF, using the following scheme as previously recommended (Mordelet and Vert, 2008). All genes known to be regulated by a TF formed a set of positive examples while the others were split in three subsets of approximately equal size. The SVM was then trained with all positive examples and the set of genes from 2 of the 3 subsets containing the negative examples. In the next step, The SVM was used to predict a score for the genes in the last subset. This procedure was repeated 3 times to obtain the SIRENE score for all genes that were apparently not TF-regulated.

Predictions obtained with TNIFSED, SIRENE and unsupervised methods were used to detect false-negative interactions in the reference network. The aim was to assess the power of each method for detecting gene pairs that had been wrongly assigned to the negative class in the reference network. These false-negative interactions were absent in RegulonDB 6.7 (reference network released in March 2010) but present in RegulonDB 7.2 (new network released in May 2011) so that we can assume that their experimental identification and confirmation took place sometime between March 2010 and May 2011. AUC and maximum F-measure were computed for each TF having at least one false-negative interaction in the reference network (RegulonDB 6.7).

3 Results and Discussion

3.1 Logistic Regression Model

The logistic regression model was first trained on the whole *E. coli* dataset, i.e., on the 756,498 gene pairs. Parameters estimations associated to both correlation coefficients and to the three functional similarities are presented in Table 1.

Table 1: Coefficients of the logistic model with *E. coli*

Parameter	Coefficient	Std. error	-log10 P-value
Intercept	-1.333	0.006	> 16
Correlation	1.204	0.013	> 16
Part. Cor.	75.638	0.591	> 16
Funct. Sim. BP	1.861	0.010	> 16
Funct. Sim. CC	0.700	0.007	> 16
Funct. Sim. MF	-1.505	0.015	> 16

Table 2: Coefficients of the logistic model with *S. cerevisiae*

Parameter	Coefficient	Std. error	-log10 P-value
Intercept	-0.373	0.006	> 16
Correlation	0.445	0.013	> 16
Part. Cor.	42.822	1.451	> 16
Funct. Sim. BP	0.543	0.013	> 16
Funct. Sim. CC	-0.051	0.008	> 16
Funct. Sim. MF	0.230	0.010	> 16

Variables with positive parameters are correlated to the presence of a positive training example, i.e. an interacting gene pair. The very high parameter noted with partial correlation is partially related to the small standard deviation of this variable (0.003 compared to 0.172 for Correlation). The parameter associated with the Molecular Function similarity is negative. Consecutively, any increase in the Molecular Function similarity results in a decrease of the interaction probability.

The logistic regression model was also trained on the whole *S. cerevisiae* dataset, i.e. on the 938,835 gene pairs. Parameters estimations of the logistic model are presented in Table 2. Parameters associated to both correlation coefficients and to the BP similarity were correlated with those observed in the model trained on the *E. coli* dataset. On the other hand, coefficients associated with the CC and MF similarities were not correlated with those observed for *E. coli*. The differences

DOI: 10.2202/1544-6115.1695

observed between parameters estimated on both organisms result from a genuine difference between both organisms, an insufficient knowledge of the genuine regulatory network and from imprecise estimations of the continuous predictor variables.

3.2 Predictive performances comparison on the *E. coli* organism

Micro- and macro-average F-measures and AUC were computed on the *E. coli* dataset with each of the following algorithms: TNIFSED, SIRENE and unsupervised methods. Regarding SIRENE, predictive performance obtained by the cross-validation scheme recommended by the authors (Mordelet and Vert, 2008) is given in Table 3, while predictive performance obtained using the classical cross-validation (named SIRENE-bias) is detailed later in the text.

Table 3: Macro- and micro-average AUC and F-measures of the inference methods on *E. coli* datasets

Method	macro-aver. AUC (Std. Err.)	macro-aver. F-meas. (Std. Err.)	micro-aver. AUC (Std. Err.)	micro-aver. F-meas.
Relevance Network	0.669 (0.019)	0.168 (0.019)	0.579 (0.005)	0.071
GGM	0.705 (0.017)	0.207 (0.021)	0.581 (0.005)	0.066
MRNET	0.661 (0.017)	0.097 (0.014)	0.623 (0.005)	0.049
CLR	0.665 (0.017)	0.141 (0.018)	0.621 (0.005)	0.069
ARACNE	0.621 (0.008)	0.114 (0.015)	0.564 (0.005)	0.046
SIRENE	0.692 (0.014)	0.140 (0.017)	**0.787** (0.005)	**0.246**
TNIFSED 5-fold	**0.801** (0.015)	**0.253** (0.015)	0.697 (0.005)	0.099
TNIFSED inter-orga.	0.783 (0.016)	0.226 (0.022)	0.671 (0.005)	0.076

Macro-average AUC of the unsupervised methods ranged between 0.621 for ARACNE and 0.705 for the GGM. Compared with the GGM, the SIRENE macro-average AUC (0.692) was slightly lower. The best macro-average AUC was obtained with the TNIFSED five-fold cross validation (0.801). The TNIFSED inter-organism generalization AUC was slightly lower. Indeed, when a logistic model was trained on the *S. cerevisiae* dataset and applied to identify interacting gene pairs in *E. coli*, a macro-average AUC value of 0.783 was obtained, which is better than the value obtained with any other method, except the biased version of SIRENE (0.836). Regarding the macro-average F-measure and not considering the biased version of SIRENE (0.469), the two best inference methods were also TNIFSED and the GGM.

Most inference methods produced micro-average indexes that were lower than their macro-average counterparts. Unlike all other methods, SIRENE produced

micro-average AUC and F-measure that were higher than the corresponding macro-average indexes. SIRENE micro-average AUC and F-measure appeared therefore to be better that the results produced by other methods, especially when classical cross-validation (SIRENE-bias) was used (micro-average AUC and F-measure of 0.920 and 0.612, respectively). As explained in the next section, the SIRENE local models favor the micro-average indexes.

3.2.1 Comparison of TNIFSED and SIRENE

The high micro-average AUC and F-measure obtained with SIRENE can be explained by the following effect. As depicted in the right part of Figure 2, the average score given to gene pairs of a particular TF increases with the number of its identified target genes. Gene pairs of TFs having a large number of already known target genes produce therefore the highest scores. As such gene pairs include a high proportion of interacting pairs and because SIRENE performs particularly well on those gene pairs (see Figures 5 and 6), most top scored gene pairs are consequently expected to be true-positive interactions. The SIRENE local models favor therefore the micro-average AUC and F-measure because the average score produced by each local model tends to increase with the number of already identified target genes.

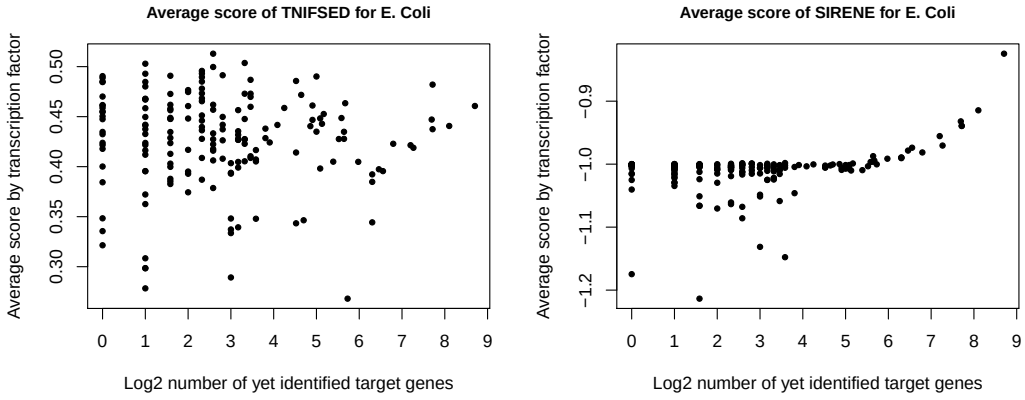

Figure 2: Average score of the gene pairs obtained with TNIFSED (left) and SIRENE (right) for each TF as a function of the number of already known target genes.

For the TNIFSED method, average scores are far less dependent on the number of target genes because a single global model is used (left part of Figure 2). This TNIFSED feature produced a global ROC curve and a global precision-recall

DOI: 10.2202/1544-6115.1695

curve which are lower than the SIRENE curves (Figure 3 and 4). This TNIFSED feature makes however the method particularly suitable for predicting new interactions with 'orphan' TFs. The impact of the number of yet identified target genes on predictive performance was indeed assessed for TNIFSED, SIRENE and Relevance Network and illustrated in Figure 5 and 6. From these figures, one can see that the predictive performance of Relevance Network for a TF tends to decrease with the number of its target genes.

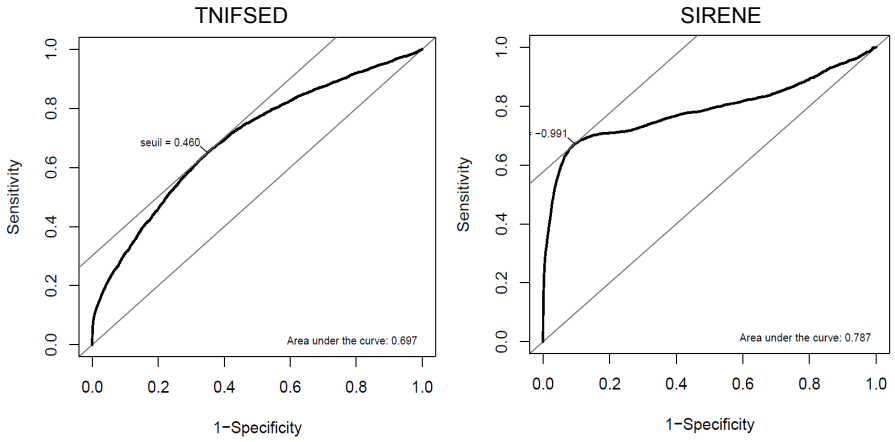

Figure 3: Receiver operating characteristic curves obtained with the results of the TNIFSED and the SIRENE algorithms. The micro-average AUC values with TNIFSED and SIRENE are 0.697 and 0.787, respectively.

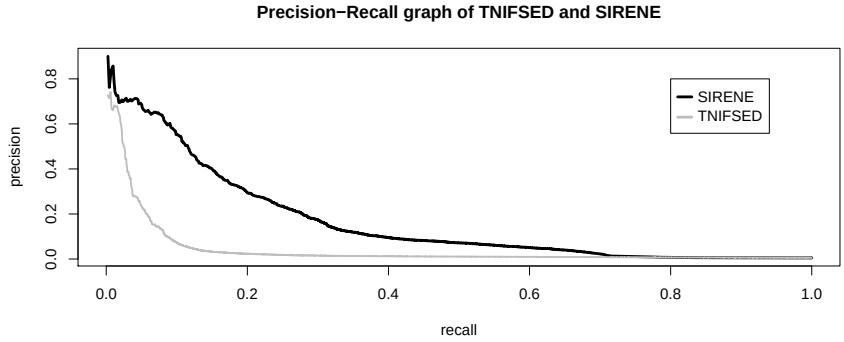

Figure 4: Precision-recall curves obtained with the results of the TNIFSED and the SIRENE algorithms. The micro-average F-measure with TNIFSED and SIRENE are 0.099 and 0.246, respectively.

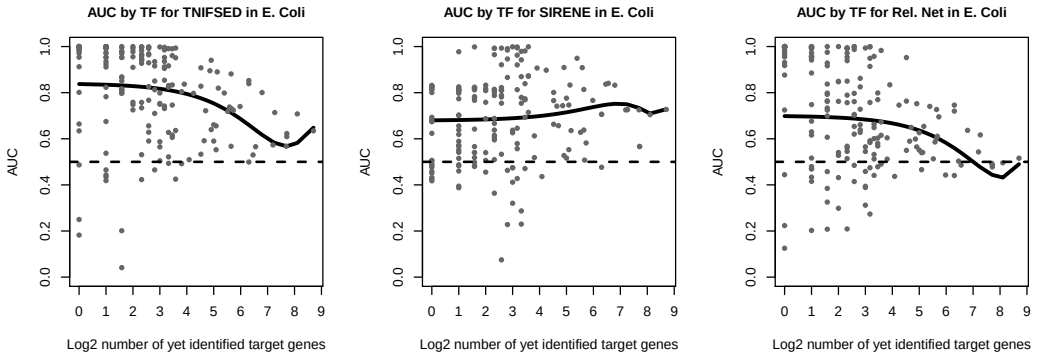

Figure 5: AUC obtained with TNIFSED (left), with SIRENE (center) and with Relevance Network (right) for each TF in *E. coli* as a function of the number of already known target genes.

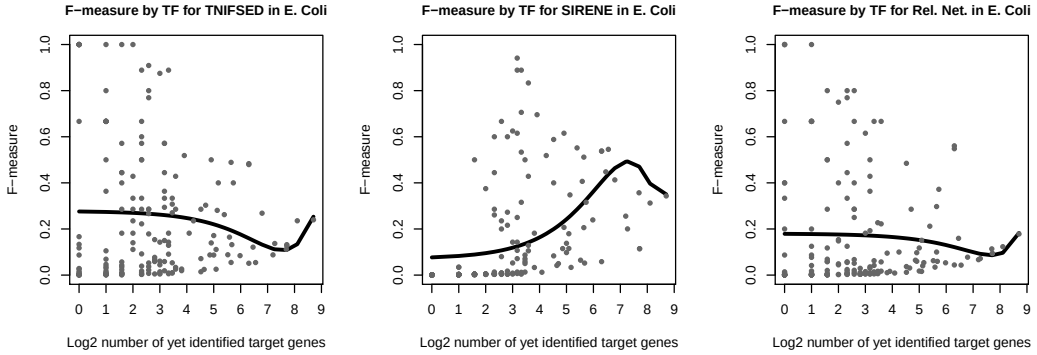

Figure 6: F-measure obtained with TNIFSED (left), with SIRENE (center) and with Relevance Network (right) for each TF in *E. coli* as a function of the number of already known target genes.

These results bring complementary information to those provided by Altay and Emmert-Streib (2010) who showed that regulations involving in-hub target genes (i.e., genes regulated by more than 3 TFs) are more difficult to infer than regulations involving leaf target genes (i.e., genes regulated by a single TF). For SIRENE, the predictive performance obtained for a particular TF tends to increase with the number of its target genes. Compared to SIRENE and Relevance Network, the TNIFSED average performance was better (Table 3). Furthermore, TNIFSED outperformed SIRENE and unsupervised algorithms when TFs have only few identified target genes.

DOI: 10.2202/1544-6115.1695

3.2.2 Predictive performances of embedded models

The relevance of the five TNIFSED predictor variables was studied by computing micro- and macro-average AUC and F-measures of embedded models (Table 4). Logistic models including only correlation coefficients or partial correlation coefficients produced macro-average indexes that were equal to those obtained from Relevance Network and from GGM, respectively. The ranking of the TF-gene scores within each TF were indeed not influenced by the cross-validation. On the other hand, the cross-validation modified the ranking of the TF-gene pairs across the different TFs and consequently the micro-average indexes.

Table 4: Micro- and macro-average AUC and F-measures obtained from the embedded models of TNIFSED on the *E. coli* dataset

Variables	macro-aver. AUC (Std. Err.)	macro-aver. F-meas. (Std. Err.)	micro-aver. AUC (Std. Err.)	micro-aver. F-meas.
Corr.	0.669 (0.019)	0.168 (0.019)	0.539 (0.005)	0.039
Part. Corr.	0.705 (0.017)	0.207 (0.021)	0.530 (0.005)	0.063
SimBP	0.705 (0.016)	0.039 (0.004)	0.638 (0.005)	0.027
SimCC	0.593 (0.010)	0.015 (0.002)	0.572 (0.005)	0.017
SimMF	0.512 (0.008)	0.012 (0.002)	0.380 (0.005)	0.010
Corr. + Part. Corr.	0.720 (0.018)	0.235 (0.023)	0.568 (0.005)	0.079
SimBP + SimMF + SimCC	0.722 (0.014)	0.062 (0.006)	0.667 (0.005)	0.038
Part. Corr. + SimBP	0.786 (0.015)	0.228 (0.021)	0.666 (0.005)	0.061
Complete Model	0.801 (0.015)	0.253 (0.022)	0.697 (0.005)	0.099

Considering the micro-average indexes, both correlation coefficients and their combination produced low AUC but high F-measures. Conversely, BP and CC functional similarities and their combination produced high AUC but low F-measures. High AUC and F-measures were obtained by combining correlation coefficients and functional similarities. The maximum micro- and macro-average AUC and F-measures were obtained when using the complete model. Consequently, a variable selection step was not implemented within TNIFSED.

3.2.3 Identification of false-negative interactions in a reference network

Identification of false-negative interactions in a reference network was performed with the scheme proposed in the method section. One AUC was obtained for each TF having at least one false-negative interaction in the reference network (Table 5). In average, TNIFSED performed better than all other methods when all TFs were taken into account to compute the mean AUC. As SIRENE can not predict the target genes of 'orphan' TFs, AUC were therefore set to 0.5 (equivalent to random prediction) for such TFs.

Table 5: AUC of the ROC obtained for the identification of false-negative interactions in RegulonDB 6.7. For each TF, N Targets 2010 corresponds to the number of related target genes in RegulonDB 6.7. N New Targets 2011 corresponds to the number of related target genes that are reported in RegulonDB 7.2 but were absent from the previous version (RegulonDB 6.7) and have consequently been discovered between March 2010 and May 2011

TF	N Targets 2010	N New Targets 2011	Rel. Net.	G.G.M.	MRNET	CLR	ARACNE	SIRENE	TNIFSED
dpiA	0	10	0.813	0.649	0.362	0.355	0.386	0.500	0.770
mcbR	0	3	0.982	0.929	0.878	0.904	0.627	0.500	0.865
mlrA	0	8	0.814	0.545	0.786	0.757	0.584	0.500	0.839
ttdR	0	3	0.982	0.762	0.661	0.800	0.475	0.500	0.831
yqhC	0	1	0.999	0.997	0.980	0.996	0.996	0.500	0.989
yqjI	0	1	0.768	0.996	0.251	0.258	0.403	0.500	0.998
adiY	1	7	0.562	0.503	0.795	0.791	0.694	0.459	0.594
kdgR	1	1	0.133	0.997	0.717	0.671	0.463	1.000	0.869
mqsA	1	1	0.764	0.496	0.583	0.611	0.395	0.993	0.570
mqsR	1	1	0.996	0.917	0.939	0.938	0.960	0.069	0.985
hdfR	2	3	0.291	0.574	0.703	0.705	0.604	0.225	0.591
sgrR	3	1	0.185	0.249	0.697	0.643	0.643	0.616	0.061
sdiA	4	1	0.176	0.996	0.911	0.853	0.450	0.314	0.960
fadR	10	1	0.041	0.882	0.145	0.146	0.442	0.996	0.894
leuO	11	6	0.232	0.309	0.529	0.416	0.612	0.287	0.105
paaX	11	1	0.924	0.285	0.556	0.629	0.448	0.996	0.985
rob	17	6	0.494	0.631	0.504	0.480	0.490	0.841	0.522
oxyR	18	7	0.489	0.639	0.404	0.400	0.437	0.589	0.574
pdhR	18	1	0.774	1.000	0.965	0.973	0.889	0.946	1.000
nagC	19	13	0.510	0.508	0.379	0.370	0.483	0.642	0.528
rcsA	20	10	0.416	0.518	0.535	0.537	0.478	0.509	0.636
rcsB	21	13	0.302	0.582	0.395	0.376	0.444	0.457	0.650
gadX	22	1	0.563	0.698	1.000	0.996	0.999	0.017	0.384
iscR	25	1	0.713	0.578	0.689	0.802	0.907	0.627	0.388
gadE	29	24	0.569	0.667	0.668	0.649	0.545	0.486	0.655
marA	29	6	0.739	0.742	0.662	0.660	0.715	0.870	0.676
soxS	31	2	0.619	0.730	0.813	0.754	0.682	0.924	0.662
phoP	34	16	0.552	0.569	0.582	0.563	0.547	0.461	0.594
fruR	39	24	0.477	0.431	0.653	0.659	0.663	0.648	0.715
fur	80	14	0.407	0.555	0.512	0.492	0.478	0.739	0.594
lrp	87	1	0.974	0.699	0.829	0.867	0.408	0.846	0.928
hns	127	20	0.350	0.612	0.457	0.443	0.495	0.738	0.592
arcA	153	1	0.275	0.870	0.843	0.867	0.871	0.740	0.659
fis	204	3	0.892	0.616	0.660	0.737	0.796	0.907	0.969
ihfA	205	5	0.891	0.440	0.835	0.750	0.548	0.201	0.557
ihfB	205	5	0.953	0.448	0.812	0.793	0.866	0.183	0.703
fnr	273	2	0.226	0.817	0.384	0.516	0.637	0.798	0.497
crp	417	2	0.718	0.351	0.512	0.524	0.647	0.910	0.595
mean			0.594	0.652	0.647	0.650	0.611	0.606	0.684

TNIFSED remains also the best method if 'orphan' TFs are discarded to compute the mean AUC. Finally, TNIFSED is the second method (after SIRENE) for the TFs that have more than 10 target genes in the reference network but the

DOI: 10.2202/1544-6115.1695

best method for the TFs that have less than 10 target genes. These results are in line with those obtained with both methods using the cross-validation on all interactions of RegulonDB7.2 (Figure 5 and 6).

3.3 Predictive performances comparison on the *S. cerevisiae* organism

With the *S. cerevisiae* organism, the only method that produced AUC values well above 0.5 was the biased version of the SIRENE algorithm (micro- and macro-average AUC of 0.722 and 0.616 ,respectively). The micro- and macro-average F-measures of the biased version of SIRENE were 0.132 and 0.133, respectively. Predictive performances of the methods are far below those observed for *E. coli*. These observations are in line with the results reported by Michoel, De Smet, Joshi, Van de Peer, and Marchal (2009) and by Zampieri, Soranzo, and Altafini (2008).

Table 6: Micro- and macro-average AUC and F-measures of the inference methods on the *S. cerevisiae* dataset

Method	macro-aver. AUC (Std. Err.)	macro-aver. F-meas. (Std. Err.)	micro-aver. AUC (Std. Err.)	micro-aver. F-meas.
Relevance Network	0.512 (0.009)	0.052 (0.005)	0.517 (0.004)	0.027
GGM	0.526 (0.009)	0.051 (0.005)	0.513 (0.004)	0.025
MRNET	0.534 (0.008)	0.042 (0.003)	0.513 (0.004)	0.027
CLR	0.534 (0.008)	0.047 (0.004)	0.509 (0.004)	0.026
ARACNE	0.506 (0.001)	0.043 (0.003)	0.496 (0.004)	0.026
SIRENE	/		/	
TNIFSED 5-fold	**0.542** (0.010)	0.058 (0.009)	**0.527** (0.004)	0.027
TNIFSED inter-orga.	0.541 (0.010)	0.059 (0.006)	0.522 (0.004)	0.026

4 Conclusion

In this paper, we presented TNIFSED, a new method for gene transcriptional network inference from functional similarity and gene expression data. The method is designed to predict the probability of interaction between a TF and a potential target gene. Parameter estimations of logistic regression model trained on both complete datasets (*E. coli* and *S. cerevisiae*) were first presented. The differences observed between parameters estimated on both organisms result from a combination of the three following effects. The first effect results from the genuine difference between

both organisms. This effect produces lower predictive performance in the inter-organism generalization than in the intra-organism cross-validation. The second effect is a lack of knowledge regarding the genuine regulatory network. Negative examples represent indeed the combination of true-negative and false-negative interactions. The weak predictive performance obtained with the cross-validation on the *S. cerevisiae* dataset is probably another consequence of this lack of knowledge. However, the number of false-negatives decreases progressively with each successive version of RegulonDB and other databases; this is expected to impact favorably on the inter-organism generalization and on cross-validation predictive performance in future. The third and last effect results from imprecise estimations of the continuous predictor variables. As Gene Ontology annotations are more and more detailed, the precision of functional similarities tends to increase with the successive versions of Bioconductor. In parallel, the number of samples in gene expression benchmarks will also increase in the future, which should improve the precision of both correlation coefficients.

Predictive performance of TNIFSED was then compared to those obtained with unsupervised methods and with the supervised SIRENE algorithm. On the *E. coli* organism, TNIFSED produced higher macro-average AUC and F-measure than any other method. This implies that TNIFSED is the best performing tool when the issue is to identify the most probable target genes for any well defined TF. The impact of the number of yet identified target genes on predictive performance was assessed for TNIFSED, SIRENE and Relevance Network. TNIFSED was the best performing method for TFs having only scarce identified target genes. Conversely, SIRENE showed higher predictive performance indexes for TF with numerous already known target genes. This effect was also observed when both methods were applied for detecting false-negative interactions in a reference network, so that TNIFSED and SIRENE appear really as two complementary useful tools.

A potential practical application of the TNIFSED method is the discovery of genes that are the target of new hypothetical TFs, as predicted by another computational method (Pérez-Rueda and Collado-Vides, 2000) listed in the regulonDB database website (http://regulondb.ccg.unam.mx). In this application, the TNISFED method can bring valuable indications when designing ChIP-chip or ChIP-Sequencing experiences that aim to confirm predicted interactions. Indeed, we showed in this study that TNIFSED was the best method to detect false-negative interactions involving TFs that have few identified target genes in a reference network. Another practical application could be the identification of interacting gene pairs from an organism that has so far been poorly investigated. In this study, we showed indeed that inter-organism performance of TNIFSED is close from intra-organism cross-validation performance. Parameters obtained for the *E. coli* organ-

DOI: 10.2202/1544-6115.1695

ism could therefore be used to combine predictor variables computed on a poorly investigated organism in order to detect its potential interacting gene pairs. For this organism, both correlation coefficients have to be computed from a benchmark of expression profiles. In addition, gene functional similarities have to be computed with the 'geneSim' function. In the current version of Bioconductor, 'geneSim' can be applied on 20 organisms but this number is rapidly growing in its successive versions.

References

Akutsu, T., S. Miyano, and S. Kuhara (2000): "Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function," *Journal of Computational Biology*, 7, 331–343.

Allocco, D., I. Kohane, and A. Butte (2004): "Quantifying the relationship between co-expression, co-regulation and gene function," *BMC bioinformatics*, 5, 18.

Altay, G. and F. Emmert-Streib (2010): "Revealing differences in gene network inference algorithms on the network level by ensemble methods," *Bioinformatics*, 26, 1738.

Balaji, S., M. Babu, L. Iyer, N. Luscombe, and L. Aravind (2006): "Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast," *Journal of molecular biology*, 360, 213–227.

Bansal, M., V. Belcastro, A. Ambesi-Impiombato, and D. Di Bernardo (2007): "How to infer gene networks from expression profiles," *Molecular systems biology*, 3, 1.

Basso, K., A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano (2005): "Reverse engineering of regulatory networks in human b cells," *Nature genetics*, 37, 382–390.

Bickel, D. (2005): "Probabilities of spurious connections in gene networks: application to expression time series," *Bioinformatics*, 21, 1121.

Bleakley, K., G. Biau, and J. Vert (2007): "Supervised reconstruction of biological networks with local models," *Bioinformatics*, 23, i57.

Butte, A. and I. Kohane (2000): "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pac Symp Biocomput*, volume 5, Citeseer, volume 5, 418–429.

Castelo, R. and A. Roverato (2009): "Reverse engineering molecular regulatory networks from microarray data with qp-graphs," *Journal of Computational Biology*, 16, 213–227.

Cho, K., S. Choo, S. Jung, J. Kim, H. Choi, and J. Kim (2007): "Reverse engineering of gene regulatory networks," *Systems Biology, IET*, 1, 149–163.

De La Fuente, A., N. Bing, I. Hoeschele, and P. Mendes (2004): "Discovery of meaningful associations in genomic data using partial correlation coefficients," *Bioinformatics*, 20, 3565.

di Bernardo, D., M. Thompson, T. Gardner, S. Chobot, E. Eastwood, A. Wojtovich, S. Elliott, S. Schaus, and J. Collins (2005): "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks," *Nature Biotechnology*, 23, 377–383.

Edgar, R., M. Domrachev, and A. Lash (2002): "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic acids research*, 30, 207.

Faith, J., B. Hayete, J. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. Collins, and T. Gardner (2007): "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biol*, 5, e8.

Friedman, J., T. Hastie, and R. Tibshirani (2008): "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, 9, 432.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er (2000): "Using bayesian networks to analyze expression data," *Journal of computational biology*, 7, 601–620.

Gama-Castro, S., V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, et al. (2008): "RegulonDB(version 6. 0): gene regulation model of Escherichia coli K-12 beyond transcription, active(experimental) annotated promoters and Textpresso navigation," *Nucleic Acids Research*, 36, .

Gardner, T., D. di Bernardo, D. Lorenz, and J. Collins (2003): "Inferring genetic networks and identifying compound mode of action via expression profiling," *Science*, 301, 102.

Gentleman, R., V. Carey, D. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. (2004): "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, 5, R80.

Hartemink, A. (2005): "Reverse engineering gene regulatory networks," *Nature Biotechnology*, 23, 554–555.

Hecker, M., S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke (2009): "Gene regulatory network inference: Data integration in dynamic models : A review," *Biosystems*, 96, 86 – 103.

Hickman, G. and T. Hodgman (2009): "Inference of gene regulatory networks using booleannetwork inference methods," *Journal of bioinformatics and computational biology*, 7, 1013–1029.

DOI: 10.2202/1544-6115.1695

Imoto, S., T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano (2003): "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks," in *Proceedings of the 2003 IEEE Bioinformatics Conference, 2003. CSB 2003*, 104–113.

Knijnenburg, T., J. Daran, M. Van Den Broek, P. Daran-Lapujade, J. De Winde, J. Pronk, M. Reinders, and L. Wessels (2009): "Combinatorial effects of environmental parameters on transcriptional regulation in Saccharomyces cerevisiae: A quantitative analysis of a compendium of chemostat-based transcriptome data," *BMC genomics*, 10, 53.

Kohl, M. and M. Kohl (2010): "Package mkmisc," .

Kramer, N., J. Schafer, and A. Boulesteix (2009): "Regularized estimation of large-scale gene association networks using graphical Gaussian models," *BMC bioinformatics*, 10, 384.

Li, H. and J. Gui (2006): "Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks," *Biostatistics*, 7, 302.

Luo, W., K. Hankenson, and P. Woolf (2008): "Learning transcriptional regulatory networks from high throughput gene expression data using continuous three-way mutual information," *BMC bioinformatics*, 9, 467.

Margolin, A., K. Wang, W. Lim, M. Kustagi, I. Nemenman, and A. Califano (2006): "Reverse engineering cellular networks," *Nature Protocols*, 1, 662–671.

Markowetz, F. and R. Spang (2007): "Inferring cellular networks–a review," *BMC bioinformatics*, 8, S5.

Martin, S., Z. Zhang, A. Martino, and J. Faulon (2007): "Boolean dynamics of genetic regulatory networks inferred from microarray time series data," *Bioinformatics*, 23, 866.

Meyer, P., K. Kontos, F. Lafitte, and G. Bontempi (2007): "Information-theoretic inference of large transcriptional regulatory networks," *EURASIP Journal on Bioinformatics and Systems Biology*, 2007, 8–8.

Meyer, P., F. Lafitte, and G. Bontempi (2008): "minet: A r/bioconductor package for inferring large transcriptional networks using mutual information," *BMC bioinformatics*, 9, 461.

Michoel, T., R. De Smet, A. Joshi, Y. Van de Peer, and K. Marchal (2009): "Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks," *BMC Systems Biology*, 3, 49.

Mordelet, F. and J. Vert (2008): "SIRENE: supervised inference of regulatory networks," *Bioinformatics*, 24, i76.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009): "Partial correlation estimation by joint sparse regression models," *Journal of the American Statistical Association*, 104, 735–746.

Pérez-Rueda, E. and J. Collado-Vides (2000): "The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12," *Nucleic Acids Research*, 28, 1838.

Perrin, B., L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alche Buc (2003): "Gene networks inference using dynamic Bayesian networks," *Bioinformatics*, 19, 11138–1118.

Schafer, J. and K. Strimmer (2005a): "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical applications in genetics and molecular biology*, 4, 1175.

Schafer, J. and K. Strimmer (2005b): "An empirical Bayes approach to inferring large-scale gene association networks," *Bioinformatics*, 21, 754.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer (2005): "Rocr: visualizing classifier performance in r," *Bioinformatics*, 21, 3940.

Soranzo, N., G. Bianconi, and C. Altafini (2007): "Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data," *Bioinformatics*, 23, 1640.

Steyerberg, E., A. Vickers, N. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. Pencina, and M. Kattan (2010): "Assessing the performance of prediction models: a framework for traditional and novel measures," *Epidemiology*, 21, 128.

Sun, J., K. Tuncay, A. Haidar, L. Ensman, F. Stanley, M. Trelinski, and P. Ortoleva (2007): "Transcriptional regulatory network discovery via multiple method integration: application to e. coli K12," *Algorithms Mol. Biol*, 2, 2.

Wang, J., Z. Du, R. Payattakool, P. Yu, and C. Chen (2007): "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, 23, 1274.

Werhli, A., M. Grzegorczyk, and D. Husmeier (2006): "Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks," *Bioinformatics*, 22, 2523.

Werhli, A. and D. Husmeier (2007): "Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge," *Statistical applications in genetics and molecular biology*, 6, 15.

Yu, G., F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang (2010): "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, 26, 976.

Zampieri, M., N. Soranzo, and C. Altafini (2008): "Discerning static and causal interactions in genome-wide reverse engineering problems," *Bioinformatics*, 24, 1510.

Zhou, S., S. Van De Geer, and P. Buhlmann (2009): "Adaptive lasso for high dimensional regression and gaussian graphical modeling," *Arxiv preprint arXiv:0903.2515*.

DOI: 10.2202/1544-6115.1695