Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 79879, 9 pages
doi:10.1155/2007/79879

Research Article

Information-Theoretic Inference of Large Transcriptional Regulatory Networks

Patrick E. Meyer, Kevin Kontos, Frederic Lafitte, and Gianluca Bontempi

ULB Machine Learning Group, Computer Science Department, Université Libre de Bruxelles, 1050 Brussels, Belgium

Received 26 January 2007; Accepted 12 May 2007

Recommended by Juho Rousu

The paper presents MRNET, an original method for inferring genetic networks from microarray data. The method is based on maximum relevance/minimum redundancy (MRMR), an effective information-theoretic technique for feature selection in supervised learning. The MRMR principle consists in selecting among the least redundant variables the ones that have the highest mutual information with the target. MRNET extends this feature selection principle to networks in order to infer gene-dependence relationships from microarray data. The paper assesses MRNET by benchmarking it against RELNET, CLR, and ARACNE, three state-of-the-art information-theoretic methods for large (up to several thousands of genes) network inference. Experimental results on thirty synthetically generated microarray datasets show that MRNET is competitive with these methods.

1. INTRODUCTION

Two important issues in computational biology are the extent to which it is possible to model transcriptional interactions by large networks of interacting elements and how these interactions can be effectively learned from measured expression data [1]. The reverse engineering of transcriptional regulatory networks (TRNs) from expression data alone is far from trivial because of the combinatorial nature of the problem and the poor information content of the data [1]. An additional problem is that by focusing only on transcript data, the inferred network should not be considered as a biochemical regulatory network but as a gene-to-gene network, where many physical connections between macromolecules might be hidden by shortcuts.

In spite of these evident limitations, the bioinformatics community made important advances in this domain over the last few years. Examples are methods like Boolean networks, Bayesian networks, and Association networks [2].

This paper will focus on information-theoretic approaches [3–6] which typically rely on the estimation of mutual information from expression data in order to measure the statistical dependence between variables (the terms "variable" and "feature" are used interchangeably in this paper). Such methods have recently held the attention of the bioinformatics community for the inference of very large networks [4–6].

The adoption of mutual information in probabilistic model design can be traced back to Chow-Liu tree algorithm [3] and its extensions proposed by [7, 8]. Later [9, 10] suggested to improve network inference by using another information-theoretic quantity, namely multi-information.

This paper introduces an original information-theoretic method, called MRNET, inspired by a recently proposed feature selection technique, the maximum relevance/minimum redundancy (MRMR) algorithm [11, 12]. This algorithm has been used with success in supervised classification problems to select a set of nonredundant genes which are explicative of the targeted phenotype [12, 13]. The MRMR selection strategy consists in selecting a set of variables that has a high mutual information with the target variable (maximum relevance) and at the same time are mutually maximally independent (minimum redundancy between relevant variables). The advantage of this approach is that redundancy among selected variables is avoided and that the trade-off between relevance and redundancy is properly taken into account.

Our proposed MRNET strategy, preliminarily sketched in [14], consists of (i) formulating the network inference problem as a series of input/output supervised gene selection procedures, where one gene at the time plays the role of the target output, and (ii) adopting the MRMR principle to perform the gene selection for each supervised gene selection procedure.

The paper benchmarks MRNET against three state-of-the-art information-theoretic network inference methods, namely relevance networks (RELNET), CLR, and ARACNE. The comparison relies on thirty artificial microarray datasets synthesized by two public-domain generators. The extensive simulation setting allows us to study the effect of the number of samples, the number of genes, and the noise intensity on the inferred network accuracy. Also, the sensitivity of the performance to two alternative entropy estimators is assessed.

The outline of the paper is as follows. Section 2 reviews the state-of-the-art network inference techniques based on information theory. Section 3 introduces our original approach based on MRMR. The experimental framework and the results obtained on artificially generated datasets are presented in Sections 4 and 5, respectively. Section 6 concludes the paper.

2. INFORMATION-THEORETIC NETWORK INFERENCE: STATE OF THE ART

This section reviews some state-of-the-art methods for network inference which are based on information-theoretic notions.

These methods require at first the computation of the mutual information matrix (MIM), a square matrix whose i, j element

$$\text{MIM}_{ij} = I(X_i;X_j) = \sum_{x_i\in\mathcal{X}}\sum_{x_j\in\mathcal{X}} p(x_i,x_j)\log\left(\frac{p(x_i,x_j)}{p(x_i)p(x_j)}\right) \tag{1}$$

is the mutual information between X_i and X_j, where $X_i \in \mathcal{X}$, $i = 1,\dots,n$, is a discrete random variable denoting the expression level of the ith gene.

2.1. Chow-Liu tree

The Chow and Liu approach consists in finding the maximum spanning tree (MST) of a complete graph, where the weights of the edges are the mutual information quantities between the connected nodes [3]. The construction of the MST with Kruskal's algorithm has an $O(n^2 \log n)$ cost. The main drawbacks of this method are: (i) the minimum spanning tree has typically a low number of edges also for non sparse target networks and (ii) no parameter is provided to calibrate the size of the inferred network.

2.2. Relevance network (RELNET)

The relevance network approach [4] has been introduced in gene clustering problems and successfully applied to infer relationships between RNA expression and chemotherapeutic susceptibility [15]. The approach consists in inferring a genetic network, where a pair of genes $\{X_i, X_j\}$ is linked by an edge if the mutual information $I(X_i;X_j)$ is larger than a given threshold I_0. The complexity of the method is $O(n^2)$ since all pairwise interactions are considered.

Note that this method is prone to infer false positives in the case of indirect interactions between genes. For example, if gene X_1 regulates both gene X_2 and gene X_3, a high mutual information between the pairs $\{X_1, X_2\}$, $\{X_1, X_3\}$, and $\{X_2, X_3\}$ would be present. As a consequence, the algorithm would infer an edge between X_2 and X_3 although these two genes interact only through gene X_1.

2.3. CLR algorithm

The CLR algorithm [6] is an extension of RELNET. This algorithm computes the mutual information (MI) for each pair of genes and derives a score related to the empirical distribution of these MI values. In particular, instead of considering the information $I(X_i;X_j)$ between genes X_i and X_j, it takes into account the score $z_{ij} = \sqrt{z_i^2 + z_j^2}$, where

$$z_i = \max\left(0, \frac{I(X_i;X_j) - \mu_i}{\sigma_i}\right) \tag{2}$$

and μ_i and σ_i are, respectively, the mean and the standard deviation of the empirical distribution of the mutual information values $I(X_i, X_k)$, $k = 1,\dots,n$. The CLR algorithm was successfully applied to decipher the *E. coli* TRN [6]. Note that, like RELNET, CLR demands an $O(n^2)$ cost to infer the network from a given MIM.

2.4. ARACNE

The algorithm for the reconstruction of accurate cellular networks (ARACNE) [5] is based on the data processing inequality [16]. This inequality states that if gene X_1 interacts with gene X_3 through gene X_2, then

$$I(X_1;X_3) \leq \min\left(I(X_1;X_2), I(X_2;X_3)\right). \tag{3}$$

The ARACNE procedure starts by assigning to each pair of nodes a weight equal to their mutual information. Then, as in RELNET, all edges for which $I(X_i;X_j) < I_0$ are removed, where I_0 is a given threshold. Eventually, the weakest edge of each triplet is interpreted as an indirect interaction and is removed if the difference between the two lowest weights is above a threshold W_0. Note that by increasing I_0, we decrease the number of inferred edges while we obtain the opposite effect by increasing W_0.

If the network is a tree and only pairwise interactions are present, the method guarantees the reconstruction of the original network, once it is provided with the exact MIM. The ARACNE's complexity for inferring the network is $O(n^3)$ since the algorithm considers all triplets of genes. In [5], the method has been able to recover components of the TRN in mammalian cells and appeared to outperform Bayesian networks and relevance networks on several inference tasks [5].

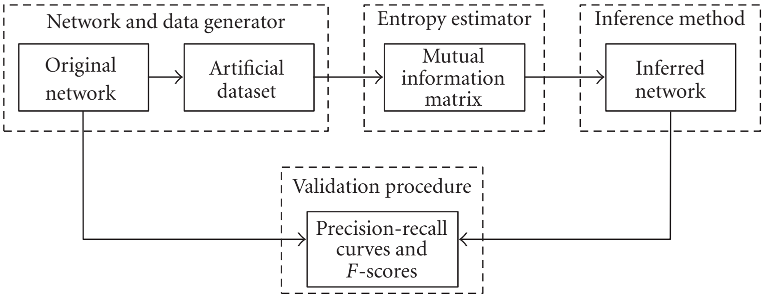

Figure 1: An artificial microarray dataset is generated from an original network. The inferred network can then be compared to this *true* network.

3. OUR PROPOSAL: MINIMUM REDUNDANCY NETWORKS (MRNET)

We propose to infer a network using the maximum relevance/minimum redundancy (MRMR) feature selection method. The idea consists in performing a series of supervised MRMR gene selection procedures, where each gene in turn plays the role of the target output.

The MRMR method has been introduced in [11, 12] together with a best-first search strategy for performing filter selection in supervised learning problems. Consider a supervised learning task, where the output is denoted by Y and V is the set of input variables. The method ranks the set V of inputs according to a score that is the difference between the mutual information with the output variable Y (maximum relevance) and the average mutual information with the previously ranked variables (minimum redundancy). The rationale is that direct interactions (i.e., the most informative variables to the target Y) should be well ranked, whereas indirect interactions (i.e., the ones with redundant information with the direct ones) should be badly ranked by the method. The greedy search starts by selecting the variable X_i having the highest mutual information to the target Y. The second selected variable X_j will be the one with a high information $I(X_j;Y)$ to the target and at the same time a low information $I(X_j;X_i)$ to the previously selected variable. In the following steps, given a set S of selected variables, the criterion updates S by choosing the variable

$$X_j^{\mathrm{MRMR}} = \arg\max_{X_j \in V \setminus S} (u_j - r_j) \tag{4}$$

that maximizes the score

$$s_j = u_j - r_j, \tag{5}$$

where u_j is a relevance term and r_j is a redundancy term. More precisely,

$$u_j = I(X_j;Y) \tag{6}$$

is the mutual information of X_j with the target variable Y, and

$$r_j = \frac{1}{|S|} \sum_{X_k \in S} I(X_j;X_k) \tag{7}$$

measures the average redundancy of X_j to each already selected variable $X_k \in S$. At each step of the algorithm, the selected variable is expected to allow an efficient trade-off between relevance and redundancy. It has been shown in [12] that the MRMR criterion is an optimal "pairwise" approximation of the conditional mutual information between any two genes X_j and Y given the set S of selected variables $I(X_j;Y \mid S)$.

The MRNET approach consists in repeating this selection procedure for each target gene by putting $Y = X_i$ and $V = X \setminus \{X_i\}$, $i = 1, \dots, n$, where X is the set of the expression levels of all genes. For each pair $\{X_i, X_j\}$, MRMR returns two (not necessarily equal) scores s_i and s_j according to (5). The score of the pair $\{X_i, X_j\}$ is then computed by taking the maximum of s_i and s_j. A specific network can then be inferred by deleting all the edges whose score lies below a given threshold I_0 (as in RELNET, CLR, and ARACNE). Thus, the algorithm infers an edge between X_i and X_j either when X_i is a well-ranked predictor of X_j ($s_i > I_0$) or when X_j is a well-ranked predictor of X_i ($s_j > I_0$).

An effective implementation of the *MRMR* best-first search is available in [17]. This implementation demands an $O(f \times n)$ complexity for selecting f features using a best-first search strategy. It follows that MRNET has an $O(f \times n^2)$ complexity since the feature selection step is repeated for each of the n genes. In other terms, the complexity ranges between $O(n^2)$ and $O(n^3)$ according to the value of f. Note that the lower the f value, the lower the number of incoming edges per node to infer and consequently the lower the resulting complexity.

Note that since mutual information is a symmetric measure, it is not possible to derive the direction of the edge from its weight. This limitation is common to all the methods presented so far. However, this information could be provided by edge orientation algorithms (e.g., IC) commonly used in Bayesian networks [7].

4. EXPERIMENTS

The experimental framework consists of four steps (see Figure 1): the artificial network and data generation, the computation of the mutual information matrix, the inference of the network, and the validation of the results. This section details each step of the approach.

4.1. Network and data generation

In order to assess the results returned by our algorithm and compare it to other methods, we created a set of benchmarks on the basis of artificially generated microarray datasets. In spite of the evident limitations of using synthetic data, this makes possible a quantitative assessment of the accuracy, thanks to the availability of the *true* network underlying the microarray dataset (see Figure 1).

We used two different generators of artificial gene expression data: the data generator described in [18] (hereafter referred to as the *sRogers* generator) and the *SynTReN* generator [19]. The two generators, whose implementations are freely available on the World Wide Web, are sketched in the following paragraphs.

sRogers generator

The *sRogers* generator produces the topology of the genetic network according to an approximate power-law distribution on the number of regulatory connections out of each gene. The normal steady state of the system is evaluated by integrating a system of differential equations. The generator offers the possibility to obtain $2k$ different measures (k wild type and k knock out experiments). These measures can be replicated R times, yielding a total of $N = 2kR$ samples. After the optional addition of noise, a dataset containing normalized and scaled microarray measurements is returned.

SynTReN generator

The *SynTReN* generator generates a network topology by selecting subnetworks from *E. coli* and *S. cerevisiae* source networks. Then, transition functions and their parameters are assigned to the edges in the network. Eventually, mRNA expression levels for the genes in the network are obtained by simulating equations based on Michaelis-Menten and Hill kinetics under different conditions. As for the previous generator, after the optional addition of noise, a dataset containing normalized and scaled microarray measurements is returned.

Generation

The two generators were used to synthesize thirty datasets. Table 1 reports for each dataset the number n of genes, the number N of samples, and the Gaussian noise intensity (expressed as a percentage of the signal variance).

4.2. Mutual information matrix estimation

In order to benchmark MRNET versus RELNET, CLR, and ARACNE, the same MIM is used for the four inference approaches. Several estimators of mutual information have been proposed in literature [5, 6, 20, 21]. Here, we test the Miller-Madow entropy estimator [20] and a parametric Gaussian density estimator. Since the Miller-Madow method requires quantized values, we pretreated the data with the equal-sized intervals algorithm [22], where the size $l = \sqrt{N}$. The parametric Gaussian estimator is directly computed by $I(X_i, X_j) = (1/2)\log(\sigma_{ii}\sigma_{jj}/|C|)$, where $|C|$ is the determinant of the covariance matrix. Note that the complexity of both estimators is $O(N)$, where N is the number of samples. This means that since the whole MIM cost is $O(N \times n^2)$, the MIM computation could be the bottleneck of the whole network inference procedure for a large number of samples ($N \gg n$). We deem, however, that at the current state of the technology, this should not be considered as a major issue since the number of samples is typically much smaller than the number of measured features.

4.3. Validation

A network inference problem can be seen as a binary decision problem, where the inference algorithm plays the role of a classifier: for each pair of nodes, the algorithm either adds an edge or does not. Each pair of nodes is thus assigned a positive label (an edge) or a negative one (no edge).

A positive label (an edge) predicted by the algorithm is considered as a true positive (TP) or as a false positive (FP) depending on the presence or not of the corresponding edge in the underlying true network, respectively. Analogously, a negative label is considered as a true negative (TN) or a false negative (FN) depending on whether the corresponding edge is present or not in the underlying true network, respectively.

The decision made by the algorithm can be summarized by a confusion matrix (see Table 2).

It is generally recommended [23] to use receiver operator characteristic (ROC) curves when evaluating binary decision problems in order to avoid effects related to the chosen threshold. However, ROC curves can present an overly optimistic view of algorithm's performance if there is a large skew in the class distribution, as typically encountered in TRN inference because of sparseness.

To tackle this problem, precision-recall (PR) curves have been cited as an alternative to ROC curves [24]. Let the precision quantity

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{8}$$

measure the fraction of real edges among the ones classified as positive and the recall quantity

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{9}$$

also know as true positive rate, denote the fraction of real edges that are correctly inferred. These quantities depend on the threshold chosen to return a binary decision. The PR curve is a diagram which plots the precision (p) versus recall (r) for different values of the threshold on a two-dimensional coordinate system.

Table 1: Datasets with n the number of genes and N the number of samples.

Dataset	Generator	Topology	n	N	Noise
RN1	*sRogers*	Power-law tail	700	700	0%
RN2	*sRogers*	Power-law tail	700	700	5%
RN3	*sRogers*	Power-law tail	700	700	10%
RN4	*sRogers*	Power-law tail	700	700	20%
RN5	*sRogers*	Power-law tail	700	700	30%
RS1	*sRogers*	Power-law tail	700	100	0%
RS2	*sRogers*	Power-law tail	700	300	0%
RS3	*sRogers*	Power-law tail	700	500	0%
RS4	*sRogers*	Power-law tail	700	800	0%
RS5	*sRogers*	Power-law tail	700	1000	0%
RV1	*sRogers*	Power-law tail	100	700	0%
RV2	*sRogers*	Power-law tail	300	700	0%
RV3	*sRogers*	Power-law tail	500	700	0%
RV4	*sRogers*	Power-law tail	700	700	0%
RV5	*sRogers*	Power-law tail	1000	700	0%
SN1	*SynTReN*	*S. Cerevisae*	400	400	0%
SN2	*SynTReN*	*S. Cerevisae*	400	400	5%
SN3	*SynTReN*	*S. Cerevisae*	400	400	10%
SN4	*SynTReN*	*S. Cerevisae*	400	400	20%
SN5	*SynTReN*	*S. Cerevisae*	400	400	30%
SS1	*SynTReN*	*S. Cerevisae*	400	100	0%
SS2	*SynTReN*	*S. Cerevisae*	400	200	0%
SS3	*SynTReN*	*S. Cerevisae*	400	300	0%
SS4	*SynTReN*	*S. Cerevisae*	400	400	0%
SS5	*SynTReN*	*S. Cerevisae*	400	500	0%
SV1	*SynTReN*	*S. Cerevisae*	100	400	0%
SV2	*SynTReN*	*S. Cerevisae*	200	400	0%
SV3	*SynTReN*	*S. Cerevisae*	300	400	0%
SV4	*SynTReN*	*S. Cerevisae*	400	400	0%
SV5	*SynTReN*	*S. Cerevisae*	500	400	0%

Table 2: Confusion matrix.

Edge	Actual positive	Actual negative
Inferred positive	TP	FP
Inferred negative	FN	TN

Note that a compact representation of the PR diagram is returned by the maximum of the F-score quantity

$$F = \frac{2pr}{r+p}, \tag{10}$$

which is a weighted harmonic average of precision and recall. The following section will present the results by means of PR curves and F-scores.

Also in order to asses the significance of the results, a McNemar test can be performed. The McNemar test [25] states that if two algorithms A and B have the same error rate, then

$$P\left(\frac{(|N_{AB} - N_{BA}| - 1)^2}{N_{AB} + N_{BA}} > 3.841459\right) < 0.05, \tag{11}$$

where N_{AB} is the number of incorrect edges of the network inferred from algorithm A that are correct in the network inferred from algorithm B, and N_{BA} is the counterpart.

5. RESULTS AND DISCUSSION

A thorough comparison would require the display of the PR-curves (Figure 2) for each dataset. For reason of space, we decided to summarize the PR-curve information by the maximum F-score in Table 3. Note that for each dataset, the accuracy of the best methods (i.e., those whose score is not significantly lower than the highest one according to McNemar test) is typed in boldface.

We may summarize the results as follows.

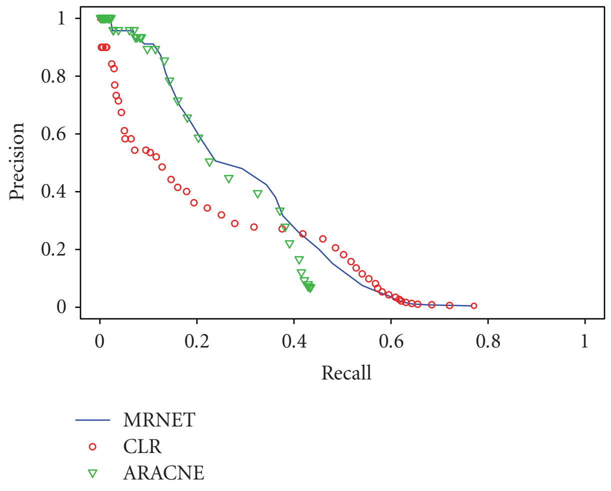

Figure 2: PR-curves for the RS3 dataset using Miller-Madow estimator. The curves are obtained by varying the rejection/acceptation threshold.

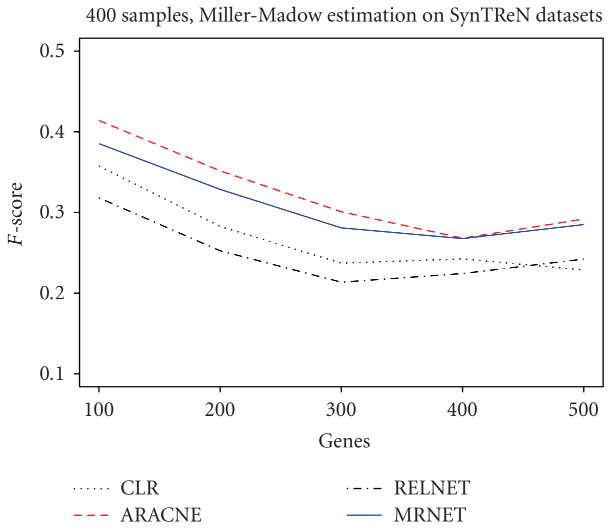

Figure 3: Influence of the number of variables on accuracy (*SynTReN* SV datasets, Miller-Madow estimator).

Accuracy sensitivity to the number of variables.

The number of variables ranges from 100 to 1000 for the datasets RV1, RV2, RV3, RV4, and RV5, and from 100 to 500 for the datasets SV1, SV2, SV3, SV4, and SV5. Figure 3 shows that the accuracy and the number of variables of the network are weakly negatively correlated. This appears to be true independently of the inference method and of the MI estimator.

Accuracy sensitivity to the number of samples.

The number of samples ranges from 100 to 1000 for the datasets RS1, RV2, RS3, RS4, and RS5, and from 100 to 500 for the datasets SS1, SS2, SS3, SS4, and SS5. Figure 4 shows how the accuracy is strongly and positively correlated to the number of samples.

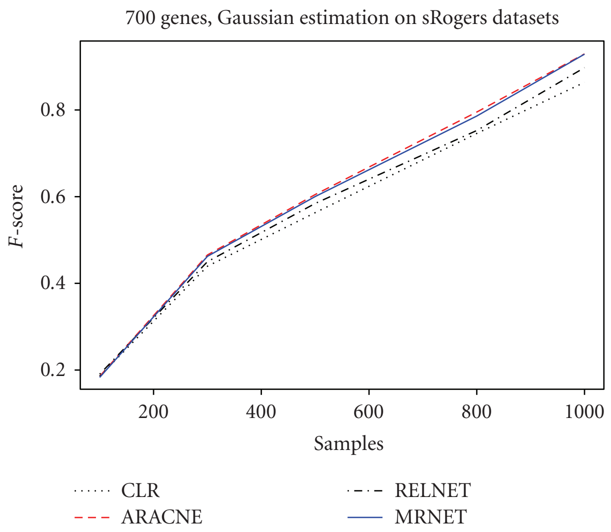

Figure 4: Influence of number of samples on accuracy (*sRogers* RS datasets, Gaussian estimator).

Accuracy sensitivity to the noise intensity.

The intensity of noise ranges from 0% to 30% for the datasets RN1, RN2, RN3, RN4, and RN5, and for the datasets SN1, SN2, SN3, SN4, and SN5. The performance of the methods using the Miller-Madow entropy estimator decreases significantly with the increasing noise, whereas the Gaussian estimator appears to be more robust (see Figure 5).

Accuracy sensitivity to the MI estimator.

We can observe in Figure 6 that the Gaussian parametric estimator gives better results than the Miller-Madow estimator. This is particularly evident with the *sRogers* datasets.

Accuracy sensitivity to the data generator.

The *SynTReN* generator produces datasets for which the inference task appears to be harder, as shown in Table 3.

Accuracy of the inference methods.

Table 3 supports the following three considerations: (i) MRNET is competitive with the other approaches, (ii) ARACNE outperforms the other approaches when the Gaussian estimator is used, *and* (iii) MRNET and CLR are the two best techniques when the nonparametric Miller-Madow estimator is used.

5.1. Feature selection techniques in network inference

As shown experimentally in the previous section, MRNET is competitive with the state-of-the-art techniques. Furthermore, MRNET benefits from some additional properties

Table 3: Maximum F-scores for each inference method using two different mutual information estimators. The best methods (those having a score not significantly weaker than the best score, i.e., P-value < .05) are typed in boldface. Average performances on *SynTReN* and *sRogers* datasets are reported, respectively, in the S-AVG, R-AVG lines.

	Miller-Madow				*Gaussian*			
	RELNET	CLR	ARACNE	MRNET	RELNET	CLR	ARACNE	MRNET
SN1	0.22	0.24	**0.27**	**0.27**	**0.21**	0.24	**0.3**	0.26
SN2	**0.23**	0.26	**0.29**	**0.29**	0.21	0.25	**0.31**	0.25
SN3	0.23	**0.25**	0.24	**0.26**	0.21	0.25	**0.31**	0.26
SN4	0.22	0.24	**0.26**	**0.26**	0.21	0.25	**0.28**	0.26
SN5	**0.21**	0.23	**0.24**	**0.24**	**0.2**	0.25	**0.27**	0.24
SS1	0.21	0.22	0.22	**0.23**	0.19	**0.24**	**0.24**	0.23
SS2	**0.21**	0.24	0.28	**0.29**	**0.2**	0.24	**0.27**	0.25
SS3	0.21	0.24	0.27	**0.28**	**0.2**	0.24	**0.28**	0.25
SS4	**0.22**	0.24	**0.27**	**0.27**	**0.21**	0.24	**0.3**	0.26
SS5	**0.22**	**0.24**	0.28	**0.29**	**0.21**	0.24	**0.3**	0.26
SV1	**0.32**	0.36	**0.41**	**0.39**	0.3	0.4	**0.44**	0.38
SV2	0.25	0.28	**0.35**	**0.33**	0.25	0.35	**0.36**	0.32
SV3	0.21	0.24	**0.3**	0.28	0.21	0.28	**0.3**	0.27
SV4	0.22	0.24	**0.27**	**0.27**	**0.21**	0.24	**0.3**	0.26
SV5	**0.24**	0.23	**0.29**	**0.29**	0.22	0.24	**0.31**	0.26
S-AVG	0.23	0.25	0.28	0.28	0.21	0.26	0.30	0.27
RN1	0.59	**0.65**	**0.6**	0.61	0.89	0.87	0.92	**0.93**
RN2	**0.5**	**0.57**	**0.5**	**0.49**	0.89	0.87	**0.92**	**0.92**
RN3	0.5	**0.55**	0.5	0.52	0.89	0.87	**0.92**	**0.92**
RN4	**0.46**	**0.51**	**0.47**	**0.47**	0.89	0.87	**0.92**	**0.91**
RN5	**0.42**	**0.46**	0.41	0.4	0.88	0.86	**0.91**	**0.91**
RS1	0.1	**0.11**	0.09	0.1	**0.19**	**0.19**	**0.19**	0.18
RS2	**0.35**	0.32	0.31	0.31	0.45	0.44	**0.47**	**0.46**
RS3	**0.38**	0.32	0.36	**0.38**	0.58	0.56	**0.6**	**0.6**
RS4	0.47	**0.54**	0.47	0.5	0.75	0.75	**0.8**	0.79
RS5	0.58	**0.68**	0.6	0.64	0.9	0.86	**0.93**	**0.93**
RV1	**0.52**	**0.38**	0.46	**0.46**	**0.72**	**0.75**	**0.72**	**0.72**
RV2	0.49	**0.53**	0.49	**0.53**	**0.71**	**0.71**	**0.71**	**0.71**
RV3	0.45	**0.5**	**0.45**	**0.48**	0.69	0.69	**0.71**	**0.71**
RV4	0.47	**0.51**	0.48	0.48	0.69	0.7	**0.74**	0.72
RV5	0.47	**0.52**	**0.47**	0.48	0.7	0.68	**0.74**	0.73
R-AVG	0.45	0.48	0.44	0.46	0.72	0.71	0.74	0.74
Tot-AVG	0.34	0.36	0.36	0.37	0.47	0.49	0.52	0.51

which are common to all the feature selection strategies for network inference [26, 27], as follows.

(1) Feature selection algorithms can often deal with thousands of variables in a reasonable amount of time. This makes inference scalable to large networks.

(2) Feature selection algorithms may be easily made parallel, since each of the n selections tasks is independent.

(3) Feature selection algorithms may be made faster by a priori knowledge. For example, knowing the list of regulator genes of an organism improves the selection speed and the inference quality by limiting the search space of the feature selection step to this small list of genes. The knowledge of existing edges can also improve the inference. For example, in a sequential selection process, as in the forward selection used with MRMR, the next variable is selected given the already selected features. As a result, the performance of the selection can be strongly improved by conditioning on known relationships.

However, there is a disadvantage in using a feature selection technique for network inference. The objective of feature selection is selecting, among a set of input variables, the ones that will lead to the best predictive model. It has been

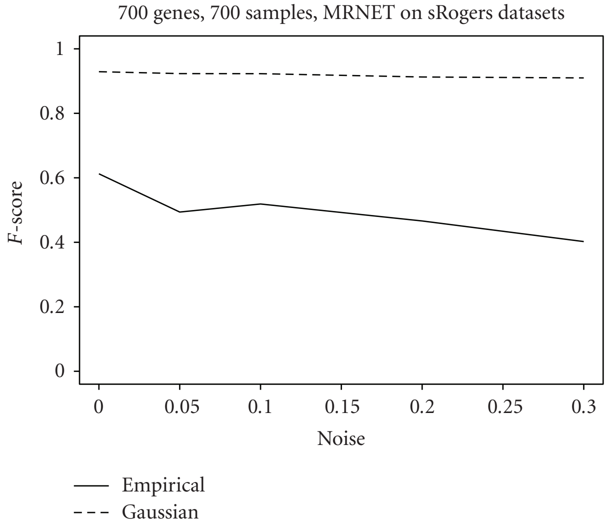

Figure 5: Influence of the noise on MRNET accuracy for the two MIM estimators (*sRogers* RN datasets).

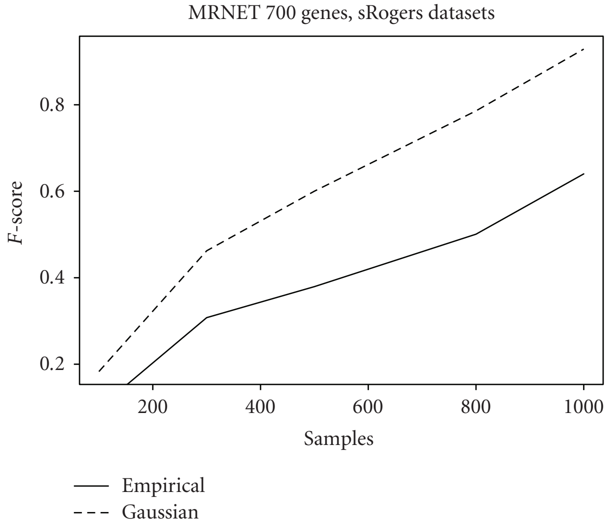

Figure 6: Influence of MI estimator on MRNET accuracy for the two MIM estimators (*sRogers* RS datasets).

proved in [28] that the minimum set that achieves optimal classification accuracy under certain general conditions is the Markov blanket of a target variable. The Markov blanket of a target variable is composed of the variable's parents, the variable's children, and the variable's children's parents [7]. The latter are indirect relationships. In other words, these variables have a conditional mutual information to the target variable Y higher than their mutual information. Let us consider the following example. Let Y and X_i be independent random variables, and $X_j = X_i + Y$ (see Figure 7). Since the variables are independent, $I(X_i; Y) = 0$, and the conditional mutual information is higher than the mutual information, that is, $I(X_i; Y \mid X_j) > 0$. It follows that X_i has some information to Y given X_j but no information to Y taken alone. This behavior is colloquially referred to as *explaining-away effect* in the Bayesian network literature [7]. Selecting variables, like X_i, that take part into indirect interactions reduce the accuracy of the network inference task. However, since MRMR relies only on pairwise interactions, it does not take into account the gain in information due to conditioning. In our example, the MRMR algorithm, after having selected X_j, computes the score $s_i = I(X_i; Y) - I(X_i; X_j)$, where $I(X_i; Y) = 0$ and $I(X_i; X_j) > 0$. This score is negative and is likely to be badly ranked. As a result, the MRMR feature selection criterion is less exposed to the inconvenient of most feature selection techniques while sharing their interesting properties. Further experiments will focus on this aspect.

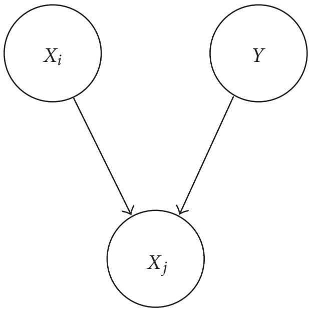

Figure 7: Example of indirect relationship between X_i and Y.

6. CONCLUSION AND FUTURE WORK

A new network inference method, MRNET, has been proposed. This method relies on an effective method of information-theoretic feature selection called MRMR. Similarly to other network inference methods, MRNET relies on pairwise interactions between genes, making possible the inference of large networks (up to several thousands of genes).

Another advantage of MRNET, which could be exploited in future work, is its ability to benefit explicitly from a priori knowledge.

MRNET was compared experimentally to three state-of-the-art information-theoretic network inference methods, namely RELNET, CLR, and ARACNE, on thirty inference tasks. The microarray datasets were generated artificially with two different generators in order to effectively assess their inference power. Also, two different mutual information estimation methods were used. The experimental results showed that MRNET is competitive with the benchmarked information-theoretic methods.

Future work will focus on three main axes: (i) the assessment of additional mutual information estimators, (ii) the validation of the techniques on the basis of real microarray data, (iii) a theoretical analysis of which conditions should be met for MRNET to reconstruct the true network.

ACKNOWLEDGMENT

This work was partially supported by the Communauté Française de Belgique under ARC Grant no. 04/09-307.

REFERENCES

[1] E. P. van Someren, L. F. A. Wessels, E. Backer, and M. J. T. Reinders, "Genetic network modeling," *Pharmacogenomics*, vol. 3, no. 4, pp. 507–525, 2002.

[2] T. S. Gardner and J. J. Faith, "Reverse-engineering transcription control networks," *Physics of Life Reviews*, vol. 2, no. 1, pp. 65–88, 2005.

[3] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[4] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," *Pacific Symposium on Biocomputing*, pp. 418–429, 2000.

[5] A. A. Margolin, I. Nemenman, K. Basso, et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, supplement 1, p. S7, 2006.

[6] J. J. Faith, B. Hayete, J. T. Thaden, et al., "Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles," *PLoS Biology*, vol. 5, no. 1, p. e8, 2007.

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible*, Morgan Kaufmann, San Fransisco, Calif, USA, 1988.

[8] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu, "Learning Bayesian networks from data: an information-theory based approach," *Artificial Intelligence*, vol. 137, no. 1-2, pp. 43–90, 2002.

[9] E. Schneidman, S. Still, M. J. Berry II, and W. Bialek, "Network information and connected correlations," *Physical Review Letters*, vol. 91, no. 23, Article ID 238701, 4 pages, 2003.

[10] I. Nemenman, "Multivariate dependence, and genetic network inference," Tech. Rep. NSF-KITP-04-54, KITP, UCSB, Santa Barbara, Calif, USA, 2004.

[11] G. D. Tourassi, E. D. Frederick, M. K. Markey, and C. E. Floyd Jr., "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *Medical Physics*, vol. 28, no. 12, pp. 2394–2402, 2001.

[12] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.

[13] P. E. Meyer and G. Bontempi, "On the use of variable complementarity for feature selection in cancer classification," in *Applications of Evolutionary Computing: EvoWorkshops*, F. Rothlauf, J. Branke, S. Cagnoni, et al., Eds., vol. 3907 of *Lecture Notes in Computer Science*, pp. 91–102, Springer, Berlin, Germany, 2006.

[14] P. E. Meyer, K. Kontos, and G. Bontempi, "Biological network inference using redundancy analysis," in *Proceedings of the 1st International Conference on Bioinformatics Research and Development (BIRD '07)*, pp. 916–927, Berlin, Germany, March 2007.

[15] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane, "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 22, pp. 12182–12186, 2000.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 1990.

[17] P. Merz and B. Freisleben, "Greedy and local search heuristics for unconstrained binary quadratic programming," *Journal of Heuristics*, vol. 8, no. 2, pp. 197–213, 2002.

[18] S. Rogers and M. Girolami, "A Bayesian regression approach to the inference of regulatory networks from gene expression data," *Bioinformatics*, vol. 21, no. 14, pp. 3131–3137, 2005.

[19] T. van den Bulcke, K. van Leemput, B. Naudts, et al., "SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC Bioinformatics*, vol. 7, p. 43, 2006.

[20] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[21] J. Beirlant, E. J. Dudewica, L. Gyofi, and E. van der Meulen, "Nonparametric entropy estimation: an overview," *Journal of Statistics*, vol. 6, no. 1, pp. 17–39, 1997.

[22] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Proceedings of the 12th International Conference on Machine Learning (ML '95)*, pp. 194–202, Lake Tahoe, Calif, USA, July 1995.

[23] F. J. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proceedings of the 15th International Conference on Machine Learning (ICML '98)*, pp. 445–453, Morgan Kaufmann, Madison, Wis, USA, July 1998.

[24] J. Bockhorst and M. Craven, "Markov networks for detecting overlapping elements in sequence data," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds., pp. 193–200, MIT Press, Cambridge, Mass, USA, 2005.

[25] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[26] K. B. Hwang, J. W. Lee, S.-W. Chung, and B.-T. Zhang, "Construction of large-scale Bayesian networks by local to global search," in *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence (PRICAI '02)*, pp. 375–384, Tokyo, Japan, August 2002.

[27] I. Tsamardinos, C. Aliferis, and A. Statnikov, "Algorithms for large scale markov blanket discovery," in *Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference (FLAIRS '03)*, pp. 376–381, St. Augustine, Fla, USA, May 2003.

[28] I. Tsamardinos and C. Aliferis, "Towards principled feature selection: relevancy, filters and wrappers," in *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics (AI&Stats '03)*, Key West, Fla, USA, January 2003.