Biological Network Inference Using Redundancy Analysis

Patrick E. Meyer, Kevin Kontos, and Gianluca Bontempi

{pmeyer,kkontos,gbonte}@ulb.ac.be
ULB Machine Learning Group
Computer Science Department
Université Libre de Bruxelles
1050 Brussels – Belgium
http://www.ulb.ac.be/di/mlg/

Abstract. The paper presents MRNet, an original method for inferring genetic networks from microarray data. This method is based on maximum relevance/minimum redundancy (MRMR), an effective information-theoretic technique for feature selection.
MRNet is compared experimentally to relevance networks (RelNet) and ARACNE, two state-of-the-art information-theoretic network inference methods, on several artificial microarray datasets. The results show that MRNet is competitive with the reference information-theoretic methods on all datasets. In particular, when the assessment criterion attributes a higher weight to precision than to recall, MRNet outperforms the state-of-the-art methods.

1 Introduction

The data flood phenomenon that biology is experiencing has propelled scientists toward the view that biological systems are fundamentally composed of two types of information: genes, encoding the molecular machines that execute the functions of life, and transcriptional regulatory networks (TRNs), specifying how and when genes are expressed [1].

Two of the most important challenges in computational biology are the extent to which it is possible to model these transcriptional interactions by large networks of interacting elements and the way that these interactions can be effectively learnt from measured expression data [2]. The reverse engineering of 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 [2]. However, progress has been made over the last few years and effective methods have been developed. Well-known state-of-the-art methods used for TRN inference are *Boolean Network Models* [3–5], *Bayesian network models* [6–8] and *Association Network Models* [9–12].

This paper will focus on information theoretic approaches which typically rely on the estimation of mutual information [9, 12] from expression data in order to measure the statistical dependence between genes.

Information-theoretic network inference methods have recently held the attention of the bioinformatics community also for very large networks [9, 12]. This paper introduces an original information-theoretic method, MRNet, inspired by a recently proposed feature selection algorithm, the maximum relevance/minimum redundancy (MRMR) algorithm [13, 14]. This algorithm has been used with success in supervised classification problems to select a set of non redundant genes which have explanatory power for the targeted phenotype [14, 15]. The MRMR selection strategy consists in selecting a set of variables that both have high mutual information with the target variable (maximum relevance) and are mutually maximally dissimilar (minimum redundancy). 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.

The proposed MRNet strategy consists in (i) formulating the network inference problem as a series of input/output supervised gene selection procedures where each gene in turn 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 rationale is that the MRMR selection relies on a square matrix of bivariate mutual information that can be computed once for all, making the network inference computationally affordable for large numbers of genes.

The paper benchmarks MRNet against two state-of-the-art information-theoretic network inference methods, namely relevance networks and ARACNE. The comparison relies on six different artificial microarray datasets obtained with two different generators.

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

2 Information-Theoretic Network Inference

Let $X_i \in \mathcal{X}$, $i = 1, \ldots, n$, be a discrete random variable denoting the expression level of a gene, where $\mathcal{X}$ is the set of expression levels of all genes and n is the number of genes. Network inference methods based on information theory [9, 12] rely on the availability of the mutual information matrix (MIM) whose generic element

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

is the mutual information between the genes X_i and X_j. This measure has the important advantage of making no assumptions about the form of dependence (e.g. linear) between variables.

2.1 Relevance Network (RelNet) Models

The relevance network [9] approach consists in inferring a genetic network where each pair of genes $\{X_i, X_j\}$ is linked by an edge whose weight is the mutual information $I(X_i; X_j)$. A threshold value I_0 is used to infer the network by eliminating all edges whose weights are beyond that threshold. As a result, the complexity of this method is $O(n^2)$ since all pairwise interactions are considered. This method was successfully applied in [16] to infer relationships between RNA expression and chemotherapeutic susceptibility.

2.2 ARACNE

The Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) method [12] is based on the Data Processing Inequality [17]. This inequality states that, if gene X_1 interacts with gene X_3 through gene X_2, then

$$I(X_1; X_3) \leq \min\left[I(X_1; X_2), I(X_2; X_3)\right] \quad . \tag{2}$$

Like relevance networks, the ARACNE inference procedure begins by assigning to each pair of nodes a weight equal to the mutual information. It then scans all triplets of nodes and removes the weakest edge of each triplet, which is interpreted as a redundant indirect relationship. Eventually, a threshold value is used to eliminate the weakest edges. Note that the method is proved to recover the underlying network (from the exact MIM) provided that the network is a tree and has only pairwise interactions. ARACNE's complexity is $O(n^3)$ since the algorithm considers all triplets of genes. Note that in [12], the method has been able to recover components of the TRN in mammalian cells. Furthermore, ARACNE has shown higher experimental performances than Bayesian networks and relevance networks [12].

3 Our Proposal: Minimum Redundancy Network (MRNet) Models

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 [13, 14] together with a best-first search strategy for performing filter selection in supervised learning problems. Consider an input/output learning task where the output is denoted by Y and V is the set of input variables. The method aims at selecting a set $S \subset V$ of inputs that have high mutual information with the target Y (maximum relevance) and low mutual information between them (minimum redundancy). The method is initialised by setting $S = X_0$ where $X_0 = \arg\max_{X_i \in V} I(X_i; Y)$. Then for a given set S of selected variables, the criterion updates S by choosing the variable

$X_j \in V \setminus S$ that maximises the score $s_j = u_j - z_j$, where u_j is a relevance term and z_j is a redundancy term. More precisely,

$$u_j = I(X_j; Y) \tag{3}$$

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

$$z_j = \frac{1}{|S|} \sum_{X_k \in S} I(X_j; X_k) \tag{4}$$

expresses the average redundancy of X_j to the already selected variables $X_k \in S$. At each step of the algorithm, the variable

$$X_j^{\text{MRMR}} = \arg \max_{X_j \in V \setminus S} (u_j - z_j) \tag{5}$$

is selected. This variable is expected to allow an efficient trade-off between relevance and redundancy. It has been shown in [18] that the MRMR criterion is an optimal "pairwise" approximation of $I(X_i; X_j|S)$.

The MRNet approach consists in repeating the selection procedure for each target gene by putting $Y = X_i$ and $V = \mathcal{X} \setminus \{X_i\}$, $i = 1, \ldots, n$, where $\mathcal{X}$ is the set of expression levels of all genes. For each pair $\{X_i, X_j\}$, the MRMR method returns a score according to (5) where $Y = X_i$. A specific network can then be inferred by deleting all edges whose score lies below a given threshold I_0.

Note that this approach infers directed networks. For comparison purposes, however, the direction of the edges is ignored.

An effective implementation of MRMR can be performed by using the approach proposed in [19]. This implementation demands a $O(f \times n)$ complexity for selecting f features using a best first search strategy. It follows that MRNet requires a $O(f \times n^2)$ complexity since the feature selection step is repeated for each of the n genes. This means that the method can infer edges in a network with a complexity ranging between $O(n^2)$ and $O(n^3)$ according to the value of f. Note that the lower this value, the lower the number of incoming edges per node to infer and consequently the lower the resulting complexity.

4 Experiments

The experimental framework consists of four steps (see Fig. 1): the artificial network and data generation, the computation of the mutual information matrix, the inference of the network and finally 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 benchmarked it on artificially generated microarray datasets. This

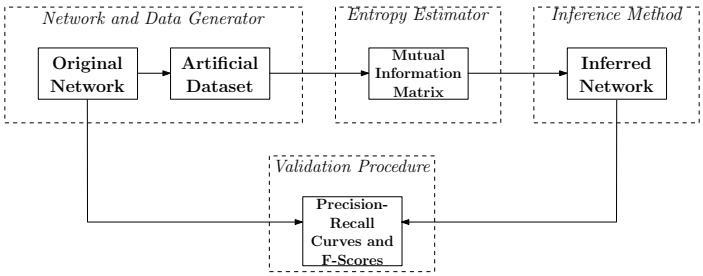


Fig. 1. An artificial microarray dataset is generated from an original network. The inferred network can then be compared to this *true* network.

approach allows us to compare an inferred network to the *true* underlying network, i.e. the one used to (artificially) generate the microarray dataset (see Fig. 1).

We used two different generators of artificial gene expression data, specifically the data generator described in [20], hereafter referred to as the sRogers generator, and the SynTReN generator [21]. 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 normalised 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 normalised and scaled microarray measurements is returned.

Generation. We generated six datasets, varying the number of genes and samples, using both generators (see Table 1 for details).

4.2 Mutual Information Matrix and Inference Methods

Each mutual information measure is computed using the Miller-Madow entropy estimator [22]. In order to use that estimator, the data are first quantised with

Table 1. The six artificial datasets generated, where n is the number of genes and N is the number of samples.

Dataset	Generator	Topology	n	N
dR1	sRogers	power-law tail	2000	1000
dR2	sRogers	power-law tail	1000	750
dR3	sRogers	power-law tail	600	600
dS1	SynTReN	*E. coli*	500	500
dS2	SynTRen	*E. coli*	300	300
dS3	SynTReN	*E. coli*	50	500

a well known unsupervised method: the equal frequency intervals [23]. The complexity of a mutual information estimation is $O(N)$, where N is the number of samples. As a result, the MIM computation is of $O(N \times n^2)$ complexity, where n is the number of genes. In order to benchmark MRNet versus RelNet and ARACNE, the same Miller-Madow MIM is used for the three inference approaches. Note however that other entropy estimators exist [22]. For example [12] adopted Gaussian Kernel Estimators [24].

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 not. Each pair of nodes is thus assigned a positive label (an edge) or a negative label (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 summarised by a confusion matrix (see Table 2).

Table 2. Confusion matrix.

edge	actual positive	actual negative
inferred positive	TP	FP
inferred negative	FN	TN

Provost et al. [25] recommends the use of 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 an algorithm's performance if there is a large skew in the class distribution, as typically encountered in network inference.

To tackle this problem, precision-recall (PR) curves have been cited as an alternative to ROC curves [26]. Let the precision quantity

$$p = \frac{TP}{TP + FP} \ , \tag{6}$$

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

$$r = \frac{TP}{TP + FN} \ , \tag{7}$$

also known 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 (specifically, we used the interpolated precision as defined in [27]) versus recall for different values of the threshold on a two-dimensional coordinate system.

The F_1-measure [28] is defined as the harmonic mean of precision and recall. It combines these quantities with an equal weight and is computed as

$$F_1(p, r) = \frac{2pr}{p + r} \ . \tag{8}$$

The weighted version of the F_1-measure is obtained by computing the weighted harmonic mean of precision and recall, the F_α-measure:

$$F_\alpha(p, r) = \frac{(1 + \alpha)pr}{\alpha p + r} \ , \tag{9}$$

where $\alpha \in (0, +\infty)$ is the weight of recall in the weighted harmonic mean; the weight of precision being 1. Consequently, by taking values of α smaller than 1, one attributes relatively more importance to precision than to recall.

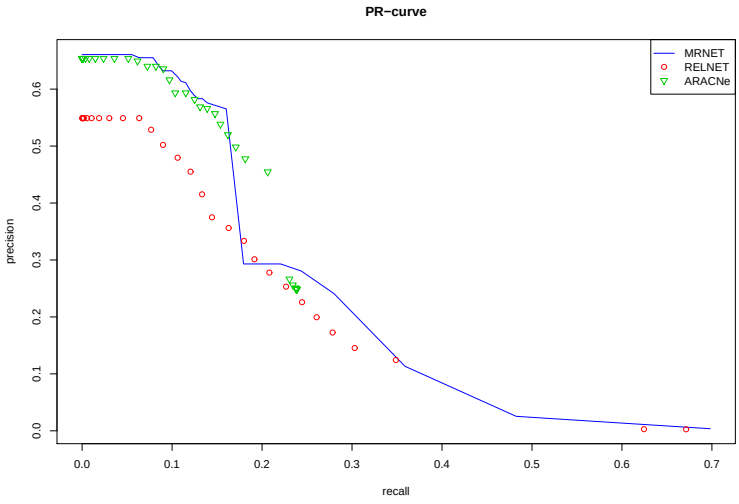
As precision and recall depend on the classification threshold, the F-measures vary with this threshold.

The following section presents the results by means of PR curves and F_α-measures for varying values of α.

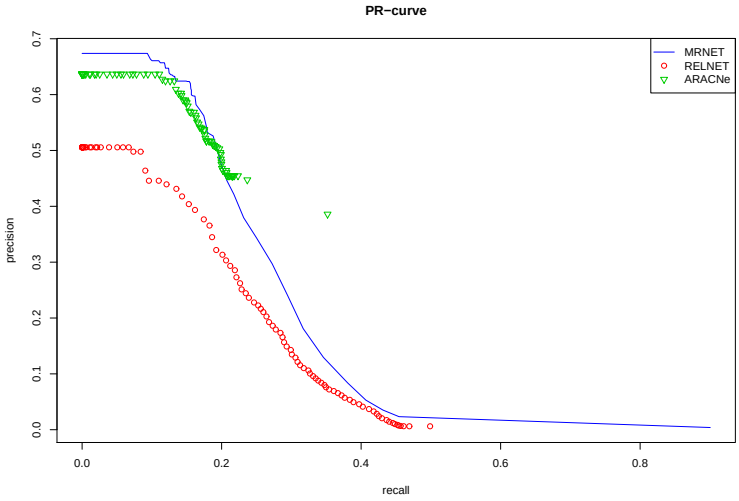
5 Results

PR-curves for the six artificial datasets generated (see Table 1) are shown in Figs. 2 and 3. For each algorithm, the best F-measure for each dataset is shown for varying values of α in Tables 3 and 4.

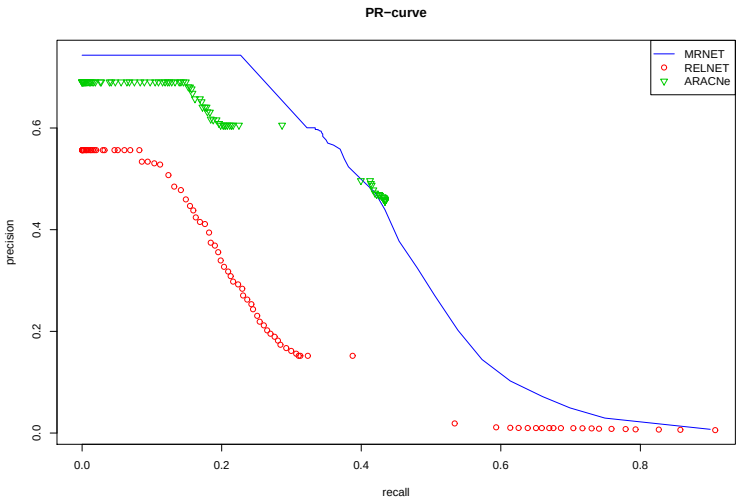
We observe that ARACNE often outperforms RelNet, in particular when precision has a higher weight than recall. MRNet appears competitive with ARACNE when precision is as important than recall (see Table 3). However, as precision is getting more important than recall, MRNet outperforms the two other methods (see Table 4). High precision is an interesting property in biological network inference since testing the existence of an affinity between a target gene and a regulator gene is expensive in time and materials.



(a) dR1

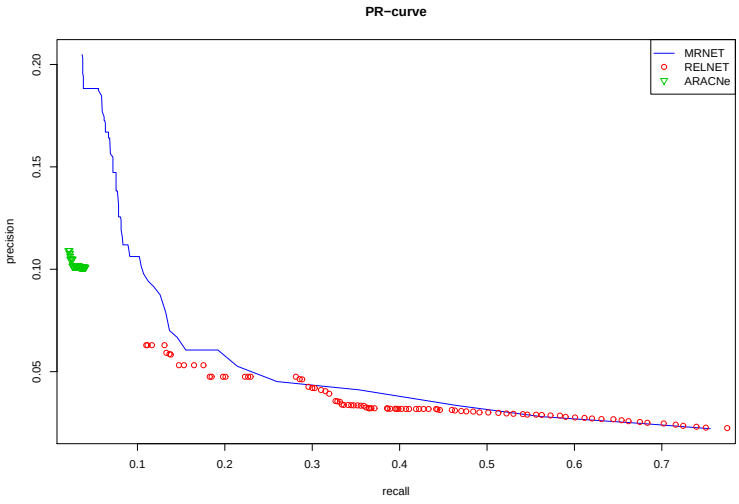


(b) dR2

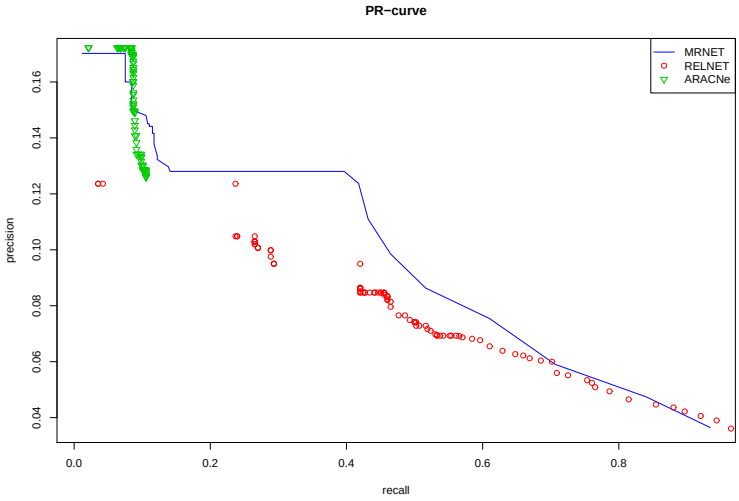


(c) dR3

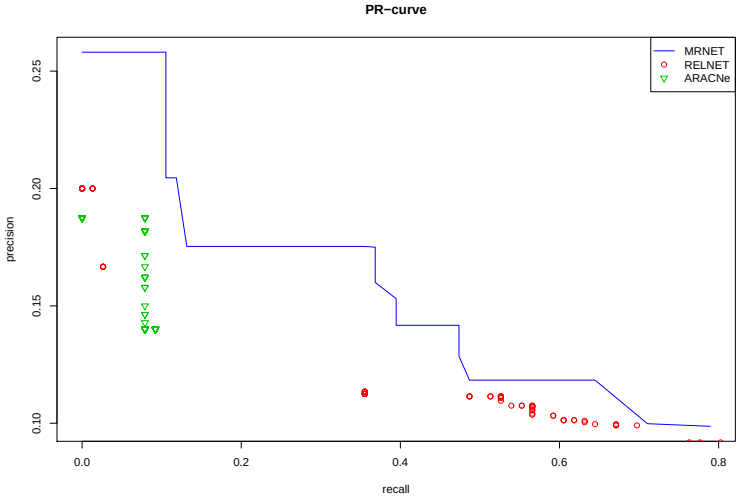
Fig. 2. PR-curves for the three artificial datasets generated with sRogers.



(a) dS1



(b) dS2



(c) dS3

Fig. 3. PR-curves for the three artificial datasets generated with SynTReN.

Table 3. Best F_α-measures with $\alpha = 1$ (precision as important as recall). For each dataset, the best measure is shown in boldface.

Dataset	RelNet	ARACNE	MRNet
dR1	0.24	**0.28**	0.26
dR2	0.25	**0.36**	0.29
dR3	0.25	**0.45**	**0.45**
dS1	0.09	0.06	**0.10**
dS2	0.16	0.12	**0.19**
dS3	0.18	0.11	**0.24**

Table 4. Best F_α-measures with $\alpha = \frac{1}{4}$ (precision more important than recall). For each dataset, the best measure is shown in boldface.

Dataset	RelNet	ARACNE	MRNet
dR1	0.29	0.37	**0.38**
dR2	0.31	0.38	**0.39**
dR3	0.32	0.49	**0.52**
dS1	0.07	0.08	**0.13**
dS2	0.13	0.14	**0.15**
dS3	0.13	0.15	**0.20**

6 Conclusion and Future Works

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.

MRNet was compared experimentally to two state-of-the-art information-theoretic network inference methods, namely relevance networks and ARACNE. Six different artificial microarray datasets were generated with two different generators. Our preliminary results appear promising. Indeed, MRNet is competitive with the other information-theoretic methods proposed in the literature on all datasets. In particular, when precision has a higher weight than recall, MRNet outperforms the other methods.

Generation of a large number of artificial datasets for the significance assessment of the results is being conducted. Experiences on real microarray datasets to infer the regulatory networks of *E. coli* and *S. cerevisiae* are also in progress. The sensitivity of the MRNet algorithm to the mutual information estimator and to the noise component will be the topic of future investigations.

Acknowledgements

This research was partially supported by the Communauté Française de Belgique through an *ARC* project (no. 04/09-307) entitled "Integrating experimental and

theoretical approaches to decipher the molecular networks of nitrogen utilisation in yeast".

References

1. Ideker, T., Galitski, T., Hood, L.: A new approach to decoding life: Systems biology. Annual Review of Genomics and Human Genetics **2** (2001) 343–372
2. van Someren, E.P., Wessels, L.F.A., Backer, E., Reinders, M.J.T.: Genetic network modeling. Pharmacogenomics **3** (2002) 507–525
3. Liang, S., Fuhrman, S., Somogyi, R.: REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E., eds.: Pacific Symposium on Biocomputing. Volume 3., Singapore, World Scientific Publishing (1998) 18–29
4. Akutsu, T., Miyano, S., Kuhara, S.: Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics **16** (2000) 727–734
5. Shmulevich, I., Dougherty, E., Kim, S., Zhang, W.: Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics **18** (2002) 261–274
6. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian Networks to Analyze Expression Data. Journal of Computational Biology **7** (2000) 601–620
7. Pe'er, D., Regev, A., Elidan, G., Friedman, N.: Inferring subnetworks from perturbed expression profiles. Bioinformatics **17** (2001) S215–S224
8. van Berlo, R.J.P., van Someren, E.P., Reinders, M.J.T.: Studying the conditions for learning dynamic bayesian networks to discover genetic regulatory networks. SIMULATION **79** (2003) 689–702
9. Butte, A.J., Kohane, I.S.: Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurments. Pacific Symposium on Biocomputing **5** (2000) 415–426
10. de la Fuente, A., Bing, N., Hoeschele, I., Mendes, P.: Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics **20** (2004) 3565–3574
11. Schäfer, J., Strimmer, K.: An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics **21** (2005) 754–764
12. Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A.: ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics **7** (2006) S7
13. Tourassi, G., Frederick, E., Markey, M., Floyd Jr, C.: Application of the mutual information criterion for feature selection in computer-aided diagnosis. Medical Physics **28** (2001) 2394
14. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology **3** (2005) 185–205
15. Meyer, P.E., Bontempi, G.: On the use of variable complementarity for feature selection in cancer classification. Lecture Notes in Computer Science **3907** (2006) 91–102
16. Butte, A., Tamayo, P., Slonim, D., Golub, T., Kohane, I.: Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proceedings of the National Academy of Sciences **97** (2000) 12182–12186

17. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley, New York (1990)
18. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1226–1238
19. Merz, P., Freisleben, B.: Greedy and local search heuristics for unconstrained binary quadratic programming. Journal of Heuristics **8** (2002) 1381–1231
20. Rogers, S., Girolami, M.: A Bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics **21** (2005) 3131–3137
21. Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., De Moor, B., Marchal, K.: SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics **7** (2006) 43
22. Paninski, L.: Estimation of entropy and mutual information. Neural Computation **15** (2003) 1191–1253
23. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. Proceedings of the Twelfth International Conference on Machine Learning **202** (1995) 194–202
24. Beirlant, J., Dudewica, E., Gyofi, L., van der Meulen, E.: Nonparametric entropy estimation: An overview. Journal of Statistics (97)
25. Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA (1998) 445–453
26. Bockhorst, J., Craven, M.: Markov networks for detecting overlapping elements in sequence data. In Saul, L.K., Weiss, Y., Bottou, L., eds.: Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, MA (2005) 193–200
27. Chakrabarti, S.: Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann (2003)
28. van Rijsbergen, C.J.: Information Retrieval. Buttersworth, London (1979)