ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

BioSystems xxx (2003) xxx–xxx

BioSystems

www.elsevier.com/locate/biosystems

Genetic network inference: the effects of preprocessing

Angelica Lindlöf*, Björn Olsson

Department of Computer Science, University of Skövde, 541 28 Skovde, Sweden

Received 21 January 2003; received in revised form 3 July 2003; accepted 3 July 2003

Abstract

Clustering of gene expression data and gene network inference from such data has been a major research topic in recent years. In clustering, pairwise measurements are performed when calculating the distance matrix upon which the clustering is based. Pairwise measurements can also be used for gene network inference, by deriving potential interactions above a certain correlation or distance threshold. Our experiments show how interaction networks derived by this simple approach exhibit low—but significant—sensitivity and specificity. We also explore the effects that normalization and prefiltering have on the results of methods for identifying interactions from expression data. Before derivation of interactions or clustering, preprocessing is often performed by applying normalization to rescale the expression profiles and prefiltering where genes that do not appear to contribute to regulation are removed. In this paper, different ways of normalizing in combination with different distance measurements are tested on both unfiltered and prefiltered data, different prefiltering criteria are considered.

Keywords: Gene expression data; Normalization; Prefiltering; Genetic networks

1. Introduction

In recent years, gene expression profiling has been a major research area and methods have been developed to analyze such data. The data from microarray experiments are often abundant and therefore require efficient computational tools for the analysis. Among the methods proposed for this purpose are different clustering techniques and gene network inference methods (D'haeseleer et al., 2000).

Clustering techniques are used for grouping genes that are co-expressed over some specific time period or experiment, and therefore have similar expression profiles. In addition to being co-expressed, genes in the same cluster are also assumed to be regulated by the same gene(s) (D'haeseleer et al., 2000; Heyer et al., 1999). Furthermore, clustering has also been proposed to reveal features shared by genes in the same group, such as similar function or regulatory elements (Zhu and Zhang, 2000).

Methods for gene network inference are used to derive information about the underlying network from the expression data, and thereby identify potential interactions between genes (D'haeseleer et al., 2000). One of the first methods introduced was based on the Boolean network approach (Somogyi et al., 1997), and was elaborated on by several other researchers (Liang et al., 1998; Akutsu et al., 1999; D'haeseleer et al., 2000). Butte and Kohane (2000) proposed a method for gene network inference by calculating pairwise mutual information (MI) on the expression data. Gene pairs having expression profiles with MI above a certain threshold were considered as potentially associ-

* Corresponding author.
E-mail address: anglelica@ida.his.se (A. Lindlöf).

0303-2647/$ – see front matter © 2003 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/S0303-2647(03)00164-3

ated. For an overview of the developments in methods of gene network inference from expression data, see van Someren et al. (2002).

In clustering, pairwise measurements are also performed when calculating the distance matrix, upon which the clustering is based. Therefore, these measurements could also be used for gene network inference, by calculating pairwise correlation between profiles and deriving potential interactions by using a correlation or distance threshold (Lindlöf and Olsson, 2002).

There are different measurements that can be used for calculating the distance, e.g. Euclidean distance (D'haeseleer et al., 2000). Before using distance measurements, normalization is often performed on the data. The purpose is to rescale the expression profiles, which brings the transformed profiles closer to each other. This does not mean that the original pairwise expression profiles are close, but their normalized versions are probabilistically close (Chen et al., 1999). Gene pairs with similar profiles that are originally far apart would receive a high distance value and thereby not be included in the derived network, whereas their transformed profiles would receive a low distance value, and hence they would be included in the derived network. There are different techniques for normalization, and the question is which technique to use when considering expression data and to what extent normalization affects network inference.

Another preprocessing step is the use of prefiltering, where genes that do not appear to contribute to regulation are removed (Chen et al., 1999). Genes removed are those which have very low expression levels or show very little variation over time, since these genes are probably not involved in regulation. These genes could give rise to false positive interactions, since they could be misinterpreted due to noise or lack of variation and thereby falsely inferred. The remaining genes are assumed to reveal true positive interactions, which would result in the inference of a reliable gene network.

This paper presents testing of different normalization techniques, in combination with different distance measurements for gene network inference from both unfiltered and prefiltered data. The experiments do not aim at testing whether or not distance measurements are appropriate for gene network inference, but explore the effects that different normalization techniques have on the identification of interactions from expression data, and whether or not prefiltering of data makes a contribution. The prefiltering should actually be made before any normalization is carried out, since expression data are known to contain noise and the normalization may enhance the level of noise. Removing genes that do not seem to contain any actual information reduces the risk of the results being affected by noise. Therefore, a combination of normalization and prefiltering would be optimal and produce a more reliable gene network than using any of the techniques separately.

Expression data from Cho et al. (1998) were used as test data, containing expression values during the mitotic cell cycle from the organism *Saccharomyce cerevisiae* (Baker's yeast). The choice of data was made mainly because this is one of the most characterized organisms for which public data are available and with considerable literature on gene and protein interactions. The derived networks were evaluated against interactions verified by literature, as reported in the KEGG database release 23 (Ogata et al., 1999) and the YPD (Costanzo et al., 2000).

The paper is organized as follows. First, the distance measurements and preprocessing techniques used are presented and defined. Thereafter, the method for inferring gene interactions from expression data and the evaluation of inferred interactions are presented. Then, the results are analyzed, and finally a discussion is offered.

2. Distance measurements and preprocessing techniques

2.1. Distance measurements

Euclidean distance, squared Euclidean distance, and Hamming distance were used in our experiments. All three distance measurements identify positive correlations and thereby only similar or identical expression profiles, i.e. antagonistic profiles are not considered. The distance measurements are all similar to each other, with just some slight differences.

Euclidean distance ED is the geometric distance in the multidimensional space and is computed as

$$\mathrm{ED}(x, y) = \sqrt{\sum (x_i - y_i)^2} \quad (1)$$

where x_i and y_i are the expression levels for gene x and y at time point i.

Squared Euclidean distance SED is used when progressively greater weights on objects that are further apart is wanted, and is defined as

$$\mathrm{SED}(x, y) = \sum_i (x_i - y_i)^2 \tag{2}$$

which is simply the square of the standard Euclidean distance.

Hamming distance HD is the average difference across dimensions and almost always yields similar results as the more simple Euclidean distance. It is defined as

$$\mathrm{HD}(x, y) = \sum_i |x_i - y_i| \tag{3}$$

Hamming distance differs from ED and SED in that it does not increase the influence of outliers.

The measurements show strong resemblance to each other and almost always give the same results. However, their specific characters could result in some variation in the derived networks and they could hopefully complement each other in different situations.

2.2. Normalization techniques

There are several different normalization techniques for time series analysis that are suitable for gene expression data. Some of them are presented here and also used in the experiments. For example, one could consider the maximum and minimum expression level for each profile by using

$$N_{i,j}^{\min,\max} = \frac{E_{i,j} - \min(E_i)}{\max(E_i)} \tag{4}$$

where $E_{i,j}$ is the expression level for gene i at time point j, $\max(E_i)$ and $\min(E_i)$ are the maximum and minimum expression levels in that profile, and $N_{i,j}^{\min,\max}$ is the normalized expression of $E_{i,j}$. This normalization technique will hereafter be denoted $N^{\min,\max}$.

There are normalization techniques that consider the mean and standard deviation in a profile, which is the classical normalization and one of the techniques used most commonly. Normalization considering mean and standard deviation, here denoted N^{mean}, is defined as

$$N_{i,j}^{\mathrm{mean}} = \frac{E_{i,j} - \bar{m}_i}{\sigma(E_i)} \tag{5}$$

where $\bar{m}_i$ is the mean and $\sigma(E_i)$ the standard deviation of the expression values in profile i.

Another category of normalization measurements is the logarithmic function, where one must define the base of the logarithm. In the case of expression data, base 2 is commonly used, but other bases could be used as well. The definition of the logarithmic function $N^{\log}$ is

$$N_{i,j}^{\log} = \log(b, E_{i,j}) = {}^{b}\log E_{i,j} \tag{6}$$

where b is the logarithmic base and $E_{i,j}$ is the expression value of gene i at time point j. Note that expression levels must be greater than 0 when using the logarithmic function. If there are negative values in the time series, it can be ensured that all values are above 0 by using

$$N_{i,j}^{\log} = \log(b, E_{i,j} + \min(|E_i|) + 0.01) \tag{7}$$

where E_i is the expression profile for gene i. In our experiments, both $b = 2$ and $b = 10$ were tested, denoted by $N^{\log 2}$ and $N^{\log 10}$, and Eq. (7) was used since the data contained negative values.

2.3. Prefiltering

The prefiltering step was based on the absolute and relative expression criteria defined by Chen et al. (1999), i.e. the genes included in the analysis were those having:

- expression level >200 in at least one of the data points; or
- $(\max(E_i) - \mathrm{avg}(E_i))/\mathrm{avg}(E_i) > 0.1$.

These criteria were also tested separately to explore if either contributes more than the other and therefore is more important. These criteria will here be termed *absolute expression criterion* and *relative average expression criterion*, respectively. In addition, a relative expression criterion indicating a twofold increase in expression levels was tested, according to

- $\max(E_i)/\min(E_i) > 2$

which will here be termed *relative twofold expression criterion*.

3. Method

3.1. Experiments

Interactions were derived from the set of expression data made available by Cho et al. (1998), using the different distance methods and normalization techniques on both unfiltered and prefiltered data. The data contains expression values from 17 time points during the mitotic cell cycle in the organism *Saccharomyces cerevisiae* (Baker's yeast).

To evaluate the results, a network was created from interactions verified by literature. This network was taken as the correct answer, which the derived interactions were evaluated against. The verified network was constructed for 86 genes known to be involved in the cell cycle regulation of *S. cerevisiae*. Interactions reported in the KEGG database release 23 (Ogata et al., 1999) and the YPD (Costanzo et al., 2000) were used to construct the network. Interactions included were of a number of types (illustrated in Fig. 1). Two proteins, *A* and *B*, were considered to have an interaction if:

- *A* regulates *B* (e.g. CDC4 and SIC1 in Fig. 1).
- *A* and *B* participate in the same protein complex (e.g. CLB5 and CDC28 in Fig. 1).
- *A* participates in a protein complex that regulates *B* (e.g. CLB5 and CDC6 in Fig. 1).
- *A* regulates a protein complex in that *B* participates.
- *A* participates in a protein complex which regulates a complex in which *B* participates.

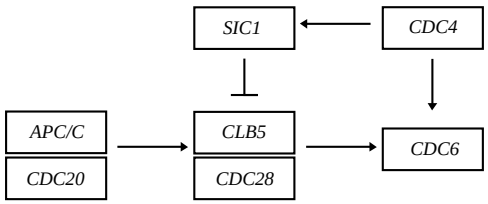

Fig. 1. Illustration of the verified interactions used in the evaluation of the experiments. APC/C is the Anaphase Promoting Complex/Cyclosome complex, interacting with gene CDC20 in a complex which up-regulates the complex of CLB5 and CDC28. The complex of CLB5 and CDC28 up-regulates CDC6. SIC1 down-regulates the complex of CLB5 and CDC28, and CDC4 up-regulates both CDC6 and SIC1.

The expression data were reduced to only contain the genes present in the verified network (see Appendix). Hence, the unfiltered data included 86 genes and for those genes 398 interactions were identified. When using prefiltering, additional genes and interactions were excluded: using both absolute expression and relative average expression criteria resulted in 67 genes and 281 interactions, using only the absolute expression criterion resulted in 67 genes and 277 interactions, using only the relative average expression criterion resulted in 85 genes and 281 interactions, and using relative twofold expression criterion resulted in 63 genes and 285 interactions.

To distinguish potential interactions a cutoff *C* was used in previous experiments, where pairwise profiles with a distance below *C* were considered to have an interaction and hence included in the derived network (Lindlöf and Olsson, 2002; Butte and Kohane, 2000). Setting the correct cutoff is of great importance as it affects the number of derived interactions. The trade-off here is between the size of the derived network and the number of true interactions that are derived. A more stringent cutoff result in fewer derived interactions than a looser cutoff, but the former should also result in a larger proportion of true positives than the latter. Since no general standard has been outlined regarding the cutoff, another approach was taken here: all network sizes from 2 to 600 interactions were evaluated.

3.2. Evaluation

Sensitivity SN and specificity SP were calculated for each derived network. These measurements are defined as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

where TP (true positives) is the number of interactions present in both the verified network and in the derived network, FN (false negatives) is the number of interactions present in the verified network but not in the derived network, and FP (False Positives) is the number of interactions present in the derived network but not in the verified network.

4. Results

The results from unfiltered non-normalized data show that all three distance measurements have problems in deriving the verified interactions, since the SN and SP for all are fairly low for all network sizes (Fig. 2). The plots show an average SP of approximately 11% for Euclidean distance and squared Euclidean distance, and approximately 10% for Hamming distance. Both SN and SP are lower for very small networks. SN progressively increases but SP varies around 10–12% after about 120 derived interactions. This is somewhat discouraging, but even if the measurements do not perform well at this stage, it is still interesting to explore if the normalizations and prefiltering have any effect. Also, even if SN and SP are rather low, they may at least be higher than expected by chance, which in that case means that the methods do contribute some knowledge, however limited. To explore this issue we selected one of the better performing method combinations, i.e. squared Euclidean distance combined with prefiltering using the relative average expression and absolute expression criteria and using data normalized by $N^{\log 10}$. The results were compared to those achieved by an algorithm that simply adds interactions to the network in random order. Ten runs were perfomed with the random algorithm, average SN and SP recorded, and the 95% confidence interval calculated. The results (shown in Fig. 3) demonstrate that the SN and SP achieved are at least higher than those achieved by the random algorithm for network

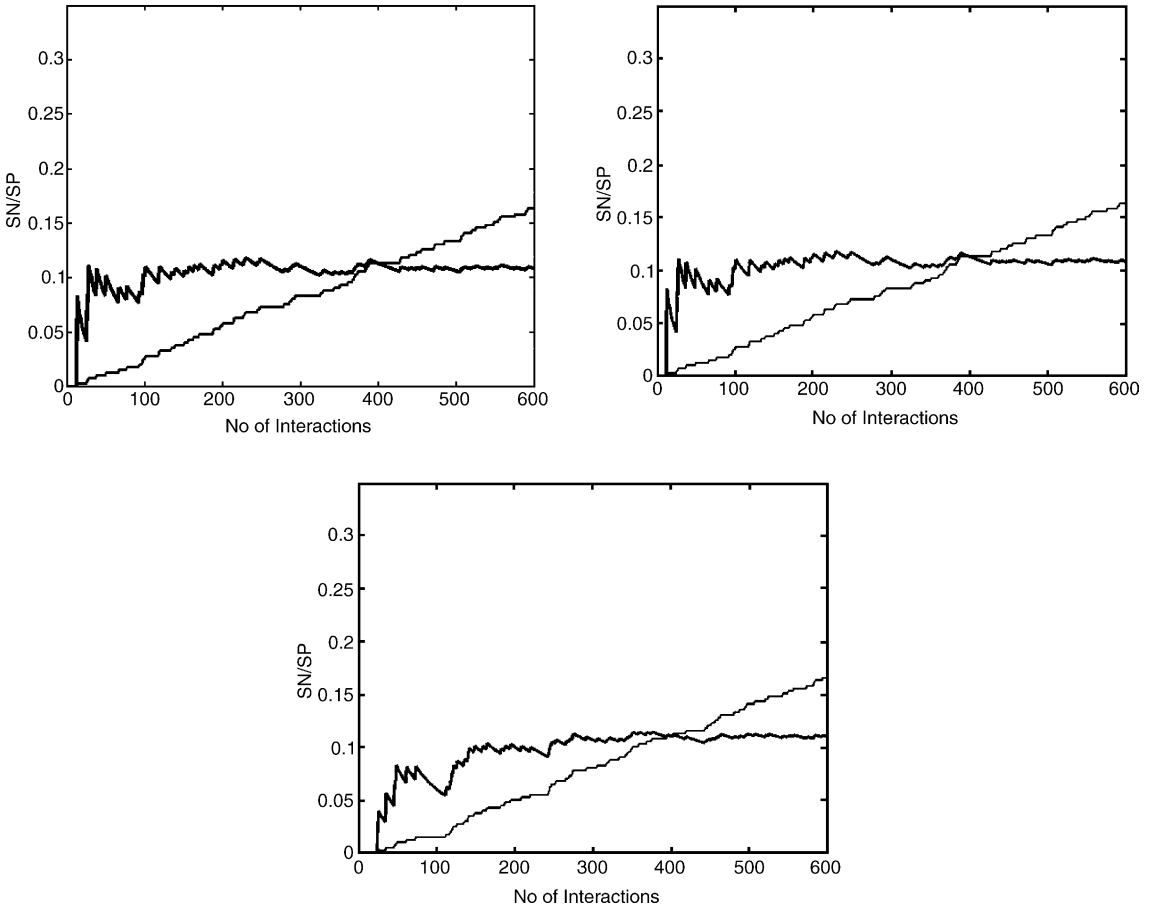

Fig. 2. SN and SP for (top left) Euclidean distance (ED), (top right) squared Euclidean distance (SED) and (bottom center) Hamming distance (HD) with non-normalized and unfiltered data. The thin line is SN and the thicker line is SP.

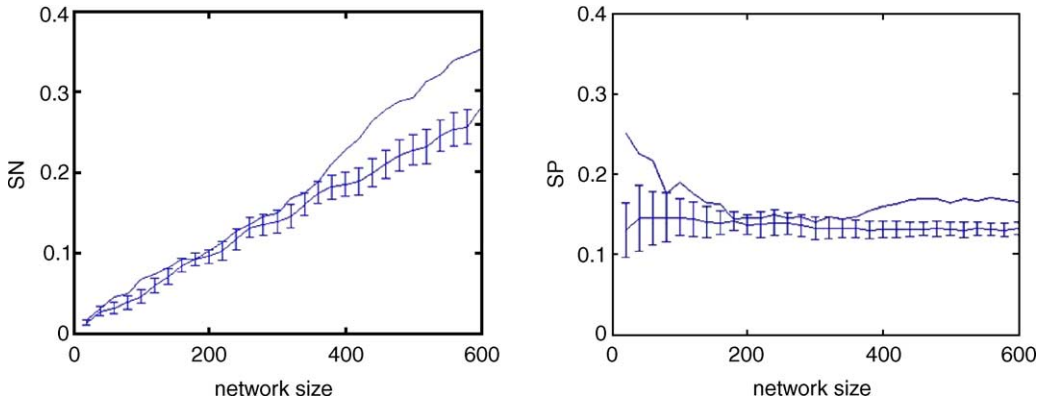

Fig. 3. SN and SP for the network derived by squared Euclidean distance compared to average SN and SP for random networks. Curves with error bars show average results for 10 networks containing randomly chosen interactions, where the bars represent a 95% confidence interval. The data were prefiltered using the relative average expression criterion in combination with the absolute expression criterion and $N^{\log 10}$ normalized.

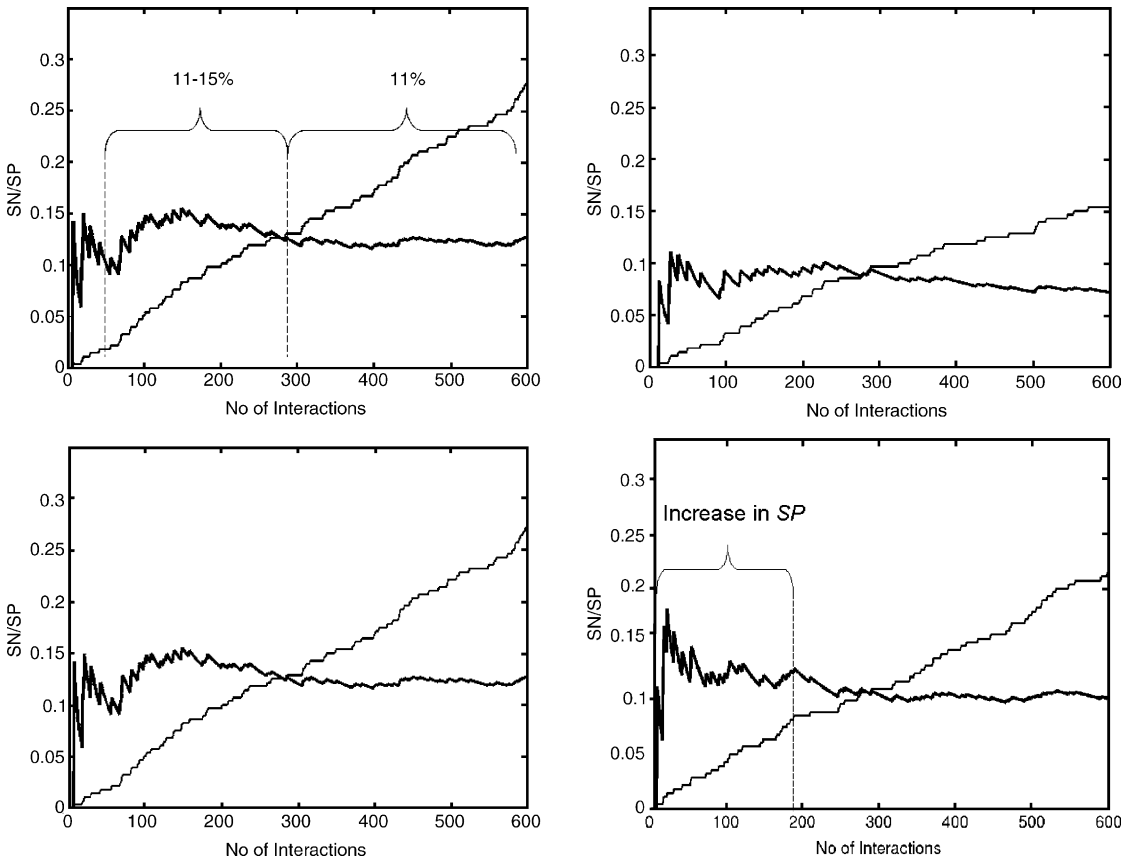

Fig. 4. SN and SP for Euclidean distance and non-normalized prefiltered data using the absolute expression criterion (top left), the relative average expression criterion (top right), the relative average expression and absolute expression criteria (lower left), and the relative twofold expression criterion (lower right).

sizes ranging from approximately 350 to 600 interactions. This can also be argued to be the relevant target range, given that the trusted network contains 398 interactions, of which 281 remained after filtering.

When comparing the results from unfiltered non-normalized data to prefiltered non-normalized data the following observations were made:

- For Euclidean and squared Euclidean distance, using only the absolute expression criterion gives similar results to using it in combination with the relative expression criterion (Fig. 4, only showing results for Euclidean distance). For network sizes of 70–300 interactions SP ranges from 11 to 15% and after about 300 interactions it is more or less stable around 11%.
- For Euclidean distance and squared Euclidean distance the prefiltering using relative average expression as criterion results in lower SP than using only absolute expression or in combination with the absolute expression criterion. This indicates that genes that show little variation also have low expression levels (Fig. 4).
- For Euclidean distance and squared Euclidean distance prefiltering with relative twofold expression as criterion gives an increase in SP for networks smaller than 200 interactions. For larger networks, no improvement was found when comparing to unfiltered data (Fig. 4).

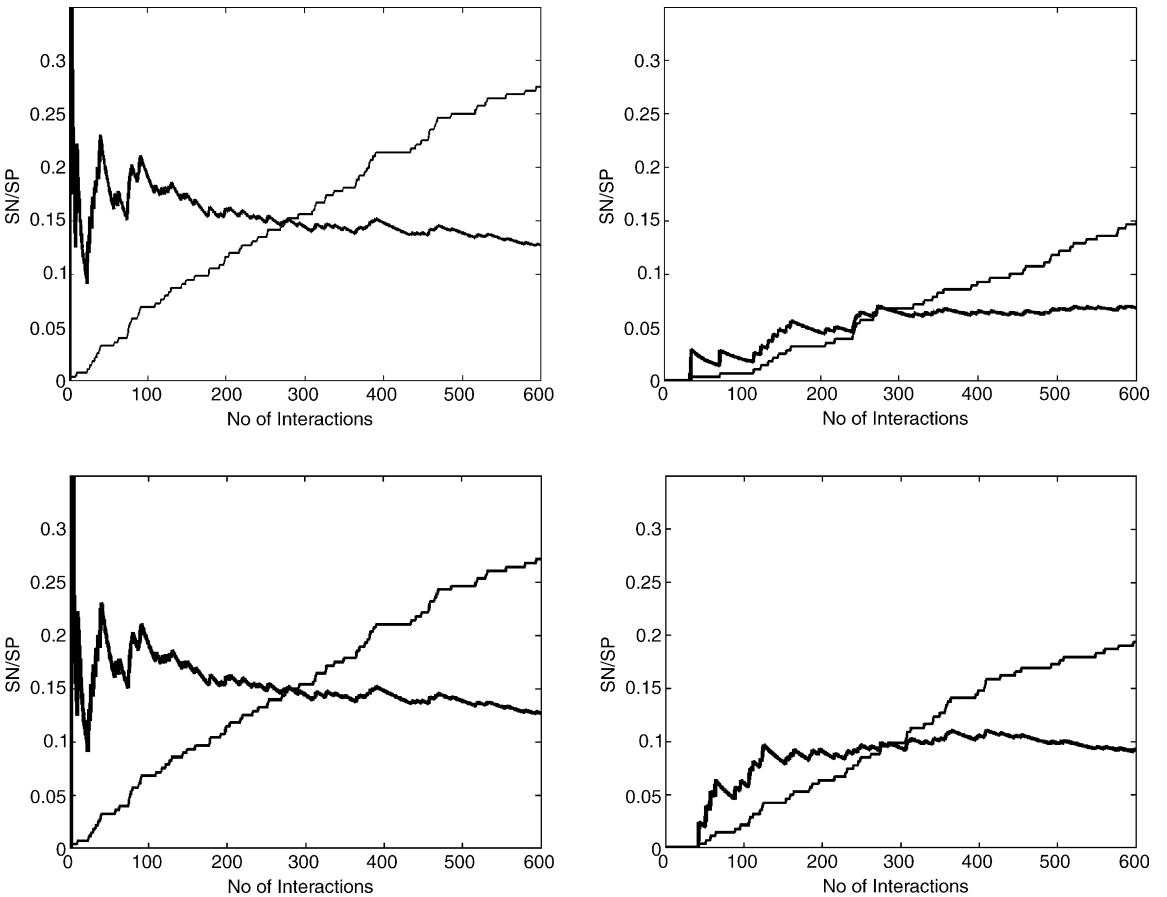

Fig. 5. SN and SP for Hamming distance and non-normalized prefiltered data using the absolute expression criterion (top left), the relative average expression criterion (top right), the relative average expression and absolute expression criteria (lower left), and the relative twofold expression criterion (lower right).

- SN increases more rapidly when filtering is used, and the largest effect is for filtering using either both absolute and relative average expression criteria, or only using the absolute expression criterion (compare Fig. 2 with Fig. 4).
- For Hamming distance the results are somewhat different. Both prefiltering using only absolute expression as criterion and in combination with the relative average expression criterion resulted in a higher average SP, but also with a highly fluctuating SP curve for networks smaller than 100 interactions. For larger networks, the curve decreases slowly to about 12% (Fig. 5). Both prefiltering using relative average expression and relative twofold expression as criterion gave worse results, especially when using the twofold expression criterion.

Thereafter, the results from unfiltered non-normalized data were compared to unfiltered normalized data and the following observations were made:

- Normalization of the data gave in general an increased SN and SP compared to non-normalized unfiltered data, with an increase in SP of about 2–3 percentage units. The exception was for Euclidean distance using mean and standard deviation as a normalization technique, which gave worse results (Fig. 6). The best result was obtained by Hamming distance using the mean and standard deviation as normalization (Fig. 6).

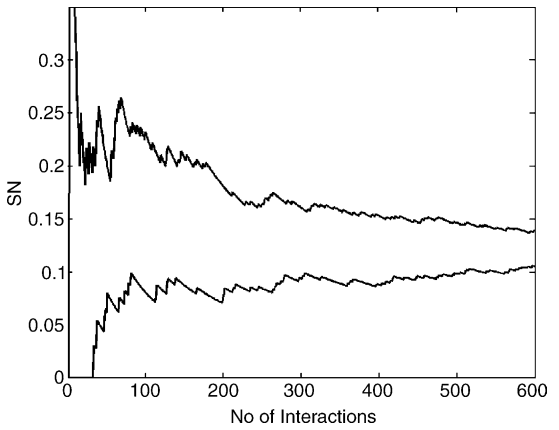

Fig. 6. SN for Euclidean distance and Hamming distance with N^{mean} normalized unfiltered data. The top curve shows Hamming distance and the bottom curve shows Euclidean distance.

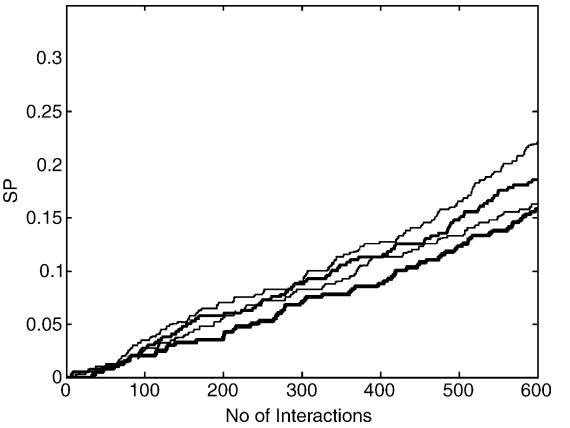

Fig. 7. SP for Euclidean distance with both non-normalized and normalized unfiltered data. The order of the curves is from top to bottom: $N^{\log 2}$, $N^{\log 10}$, $N^{\min,\max}$, and N^{mean} normalized values.

- In general, SN increases more rapidly when normalization is used, except for the case with Euclidean distance using mean and standard deviation as a normalization technique (Fig. 7).
- Normalization using logarithmic values with base 2 gave the same result as with base 10, except for squared Euclidean distance where base 10 showed a slightly better SP (no data shown).

Finally, the results from combining normalization and prefiltering were evaluated. In this step only relative twofold expression criteria and the combination of absolute and average expression criteria were investigated, since the absolute and average expression criteria separately showed a similar or worse result than in combination. The analysis resulted in the following observations:

- In all cases the SP curve fluctuates highly for networks with fewer than 100 inferred interactions, while the curve thereafter in most cases stabilizes or in a few cases decreases (Fig. 8).
- An average SN and SP for unprocessed data versus preprocessed data calculated over the range of 100–600 derived interactions demonstrates that preprocessing improves the results (Table 1).
- Logarithmic normalization generates clearly better results compared to normalization using any of the other techniques, and prefiltering using the combination of the relative expression criterion and the

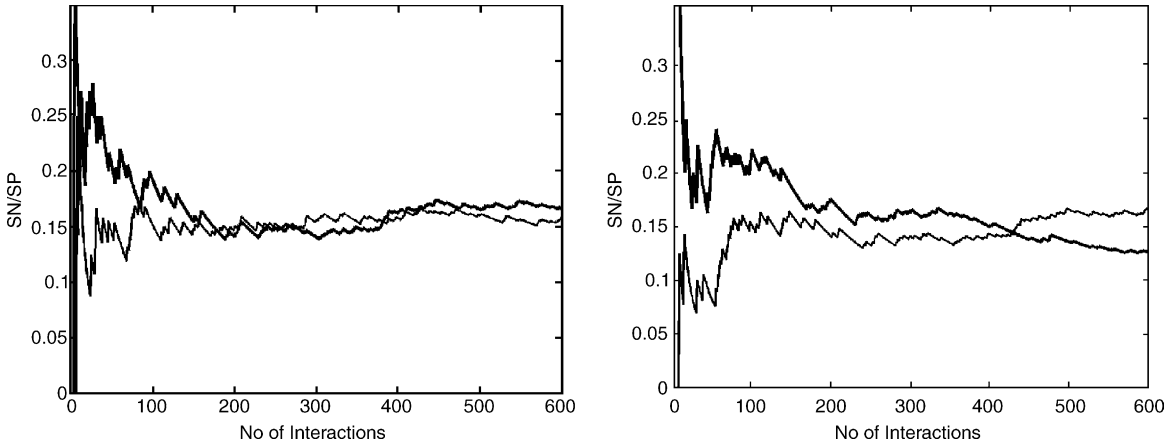

Fig. 8. SP curves showing a highly fluctuating result for networks containing fewer than 100 interactions, which thereafter either stabilize or decrease. The graph on the right shows Hamming distance with N^{mean} and the relative expression criterion in combination with the absolute expression criterion (bold line), and Euclidean distance with $N^{\log 2}$ and the twofold expression criterion (thin line). The graph on the left shows squared Euclidean distance with $N^{\log 10}$ and the twofold expression criterion (bold line) and Hamming distance with $N^{\log 10}$ and the relative expression criterion in combination with the absolute expression criterion (thin line).

Table 1
Average SN and SP for each distance measure and for unprocessed vs. preprocessed data

	Unpreprocessed		Preprocessed	
	SN	SP	SN	SP
ED	0.10	0.11	0.15	0.12
HD	0.09	0.10	0.16	0.16
SED	0.10	0.11	0.15	0.12

Averages are calculated over the range of 100–600 derived interactions for the different normalization techniques and prefiltering techniques.

Table 2
Average SN and SP for different normalization and prefiltering techniques

	SN	SP
Normalization technique		
$N^{\log 10}$	0.19	0.15
$N^{\log 2}$	0.19	0.15
N^{mean}	0.13	0.11
$N^{\text{min,max}}$	0.13	0.11
Prefiltering technique		
Relative + absolute	0.17	0.14
Twofold	0.15	0.12

Averages are calculated over the range of 100–600 derived interactions and over all distance measures.

absolute expression criterion generates higher SN and SP than the relative twofold expression criterion (Table 2). It also indicates that logarithmic normalization is more important than prefiltering since it yields higher average SN and SP than any of the prefiltering techniques.

5. Discussion

The results indicate that logarithmic normalization seems to generate the best result, irrespective of filtering technique. This is of course considering this dataset; further analyses with other datasets must be made before any general conclusion can be made. In addition, the consequence of using prefiltering is that a number of genes are lost (here, about one-third). In this case, these genes are known to be involved in cell cycle regulation, but prefiltering will exclude them from the derived network. The trade-off here is that filtering increases the sensitivity and specificity of the method by resulting in a larger proportion of correct interactions, but it also results in the loss of genes that are known to be part of the true network.

In conclusion, the results show that normalization should be carried out (which is not a surprise)

and, considering this dataset, preferably with logarithmic normalization as it generated the best results. The particular base that is used seems to be of less importance, since bases 2 and 10 almost always yielded the same result. Prefiltering should be done with caution, since the results indicate that it is possible to exclude genes important to the process under investigation. This has, however, to be further investigated with other prefiltering techniques, such as measurements for most differentially regulated genes or according to background noise (excluding genes with expression profiles similar to the background signal) (Lu et al., 2002; Draghici, 2002). In these experiments, the absolute and average expression criteria in combination show the best results.

The same normalization technique can give very different results depending on distance measurement used, e.g. comparing Euclidean distance with Hamming distance when the data have been normalized using the mean and standard deviation (Fig. 6). This shows that it is worthwhile to consider several distance techniques and compare the results.

Even if preprocessing in general gave better results, there were also high fluctuations in specificity, especially for networks with less than 100 interactions. This indicates that smaller networks have a larger proportion of false positive interactions.

All distance measurements had problems in deriving the verified interactions, as the SN and SP in general were fairly low, even after preprocessing had been made. Here, regulatory interactions and protein complexes are considered to be plausible for the methods to identify. There is some evidence that similar expression profiles reflect protein interactions (Ge et al., 2001; Mrowka et al., 2003). However, a more thorough study is needed to investigate what types of interactions that can really be derived using the presented approach. In addition, other data sources than YPD and KEGG should also be considered, such as MIPS (Mewes et al., 2002), as well as other types of interactions, such as genes expressed in the same phase in the cell cycle. Additional information in form of regulatory patterns that indicate associations between genes (Spellman et al., 1998; Chen et al., 2000; Kielbasa et al., 2001; van Helden et al., 2000; Tyson et al., 2002) should also be considered.

Appendix A

Genes known to be involved in cell cycle regulation and included in the experiments

APC2	CDC45	CRT1	MET30	RAD53
APC4	CDC46	DBF2	MIH1	RAD9
APC9	CDC53	DBF20	ORC1	SIC1
BUB1	CDC54	DBF4	MCM2	SSN6
BUB2	CDC6	DOC1	MCM3	PHO81
BUB3	CDC7	ESC5	MCM6	PHO85
CAK1	CLB4	ESP1	ORC2	RAD17
CDC14	CLB5	FAR1	ORC3	SWE1
CDC15	CLB6	FUS3	ORC4	SWI4
CDC16	CDH1	GRR1	ORC5	SWI5
CDC20	CKS1	HSL1	ORC6	SWI6
CDC26	CLB1	HSL7	PCL1	RAD24
CDC27	CLB2	LTE1	PCL2	TEM1
CDC28	CLB3	MAD1	PDS1	TUP1
CDC34	DDC1	MAD3	PHO2	UBA1
CDC47	CLN1	MBP1	PHO4	
CDC5	CLN2	MEC1	PHO5	
CDC4	CLN3	MEC3	PHO80	

References

Akutsu, T., Miyano, S., Kuhara, S., 1999. Identification of genetic networks from a small number of expression patterns under the Boolean network model. Pac. Symp. Biocomput. 4, 17–28.

Butte, A., Kohane, I., 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac. Symp. Biocomput. 5, 418–429.

Chen, Y., Bittner, M.L., Dougherty, E.R., 1999. Issues associated with microarray data analysis and integration. Nat. Genet. 22, 213–216.

Chen, K.C., Csikasz-Nagy, A., Gyorffy, B., Val, J., Novak, B., Tyson, J.J., 2000. Kinetic analysis of a molecular model of the budding yeast cell cycle. Mol. Biol. Cell 11, 369–391.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wolfsberg, T.G., Babrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W., 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 21, 65–73.

Costanzo, M.C., Hogan, J.D., Cusick, M.E., Davis, B.P., Fancher, A.M., Hodges, P.E., Kondu, P., Lengieza, C., Lew-Smith, J.E., Lingner, C., Roberg-Perez, K.J., Tillberg, M., Brooks, J.E., Garrels, J.I., 2000. The Yeast Proteome Database (YPD) and *Caenorhabditis elegans* Proteome Database (WormPD): comprehensive resources for the organization and comparison

of model organism protein information. Nucl. Acids Res. 28, 81–84.

D'haeseleer, P., Liang, S., Somogyi, R., 2000. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16, 707–726.

Draghici, S., 2002. Statistical intelligence: effective analysis of high-density microarray data. Drug Discov. Today 6/7 (11), S55–S63.

Ge, H., Liu, Z., Church, G.M., Vidal, M., 2001. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. Nat. Genet. 29, 482–486.

van Helden, J., André, B., Collado-Vides, J., 2000. A web site for the computational analysis of yeast regulatory sequences. Yeast 16, 177–187.

Heyer, L.J., Kruglyak, S., Yooseph, S., 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. 9, 1106–1115.

Kielbasa, S.M., Korbel, J.O., Beule, D., Schuchhardt, J., Herzel, H., 2001. Combining frequency and positional information to predict transcription binding sites. Bioinformatics 17, 1019–1026.

Liang, S., Furhman, S., Somogyi, R., 1998. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Pac. Symp. Biocomput. 3, 18–29.

Lindlöf, A., Olsson, B., 2002. Could correlation based methods be used for deriving genetic association networks. Inf. Sci. 146 (1–4), 103–113.

Lu, X.L., Olson, J.M., Zhao, L.P., 2002. A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. Hum. Mol. Genet. 11, 1977–1985.

Mewes, H.W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkoetter, M., Rudd, S., Weil, B., 2002. MIPS: a database for genomes and protein sequences. Nucl. Acids Res. 30, 31–41.

Mrowka, R., Liebermeister, W., Holste, D., 2003. Does mapping reveal correlation between gene expression and protein–protein interaction? Nat. Genet. 33, 15–16.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., 1999. KEGG: Kyoto encyclopedia of genes and genomes. Nucl. Acids Res. 27, 29–34.

van Someren, E.P., Wessels, L.F.A., Backer, E., Reinders, M.J.T., 2002. Genetic network modeling. Pharmacogenetics 3 (4), 507–525.

Somogyi, R., Fuhrman, S., Askenazi, M., Wuensche, A., 1997. The gene expression matrix: towards the extraction of genetic network architectures. Nonlinear Anal. Theor. Methods Appl. 3, 1815–1824.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell 9, 3273–3297.

Tyson, J.J., Csikasz-Nagy, A., Novak, B., 2002. The dynamics of cell cycle regulation. Bioessays 24, 1095–1109.

Zhu, J., Zhang, M.Q., 2000. Cluster, function and promoter: analysis of yeast expression array. Pac. Symp. Biocomput. 5, 476–487.