BIOINFORMATICS

Vol. 18 Suppl. 2 2002
Pages S231–S240

The mutual information: Detecting and evaluating dependencies between variables

R. Steuer[1], J. Kurths[1], C. O. Daub[2], J. Weise[2] and J. Selbig[2]

[1]University Potsdam, Nonlinear Dynamics Group, Am Neuen Palais 10, 14469 Potsdam, Germany and [2]Max-Planck-Institute for Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany

Received on April 8, 2002; accepted on June 15, 2002

ABSTRACT

Motivation: Clustering co-expressed genes usually requires the definition of 'distance' or 'similarity' between measured datasets, the most common choices being Pearson correlation or Euclidean distance. With the size of available datasets steadily increasing, it has become feasible to consider other, more general, definitions as well. One alternative, based on information theory, is the mutual information, providing a general measure of dependencies between variables. While the use of mutual information in cluster analysis and visualization of large-scale gene expression data has been suggested previously, the earlier studies did not focus on comparing different algorithms to estimate the mutual information from finite data.

Results: Here we describe and review several approaches to estimate the mutual information from finite datasets. Our findings show that the algorithms used so far may be quite substantially improved upon. In particular when dealing with small datasets, finite sample effects and other sources of potentially misleading results have to be taken into account.

Contact: steuer@agnld.uni-potsdam.de

INTRODUCTION

The ability to monitor whole-genome gene expression in a parallel and quantitative way represents one of the latest breakthroughs in experimental molecular biology (Brazma and Vilo, 2000; Nature, 1999; Schena *et al.*, 1995). Simultaneously with experimental progress, increasing methodologies are available for conceptualizing and unraveling the functional relationships implicit in these datasets, see D'haeseleer *et al.* (2000) and references therein for a short review. One of the most widely used concepts is, to group together genes with similar patterns of expression (Eisen *et al.*, 1998). This clustering of co-expression is thought to allow the inference of shared regulatory inputs and functional pathways (D'haeseleer *et al.*, 2000). Even a more modest description would still assume that it is fairly safe to hypothesize that co-expressed genes may have something in common in their regulatory mechanism. Almost all clustering algorithms rely on a definition of pair-wise distances between measured expression profiles and it is widely recognized that the choice of the distance may be as crucial as the choice of the clustering algorithm itself (D'haeseleer *et al.*, 2000). However, as pointed out by Brazma and Vilo (2000), the appropriateness of similarity measures has not been systematically explored and these measures are used on an ad-hoc basis. In this work, we investigate the use of mutual information as a measure of distance between variables. This approach is not entirely new: An early attempt to use information-theoretic concepts in this context was given in (Michaels *et al.*, 1998), a more recent analysis was applied by Butte and Kohane (2000). We supplement this earlier work by focusing on different algorithms to estimate the mutual information for small datasets. As will be shown below, the simple algorithms used so far may be quite substantially improved upon. In particular, we will point out potential pitfalls in the analysis and discuss strategies to overcome them. The results will be exemplified using a publicly available dataset corresponding to up to 300 diverse mutations and chemical treatments in *S. cerevisae*, see (Hughes *et al.*, 2000) for further details.

We shall point out that the detection of relationships between two or more variables is not restricted to the analysis of gene expression, but is of great interest in many areas of science. Variables which are not statistically independent suggest the existence of some functional relation between them. While there are several approaches to quantify the *linear* dependence between variables, the framework of information theory (Shannon, 1948) provides a *general* measure of dependencies between variables. In particular, a vanishing Pearson correlation does not imply that two variables are independent. The mutual information therefore provides a better and more general criterion to investigate relationships between variables.

THE MUTUAL INFORMATION

We begin with a brief review of information theory and Shannon entropy. Following Shannon (1948), all definitions will be given in terms of discrete systems. As we shall later see, for the application on gene-expression data, this implies that we have to partition the continuous values into discrete bins.

The Shannon entropy

Consider a system A with M_A possible states. That is, a measurement performed on A will yield one of the possible values $a_1, \ldots, a_{M_a}$, each with its corresponding probability $p(a_i)$. The average amount of information gained from a measurement that specifies one particular value a_i is given by the *entropy* $H(A)$ of the system (Shannon, 1948; Cover and Thomas, 1991).

$$H(A) = -\sum_{i=1}^{M_A} p(a_i) \log p(a_i) \tag{1}$$

As stated by Faser and Swinney (1986), the entropy $H(A)$ could be described as the 'quantity of surprise you should feel upon reading the result of a measurement'. We summarize some properties of $H(A)$:

- Assume the outcome of the measurement is completely determined to be a_l, that is, the probability $p(a_l)$ is one and all other probabilities $p(a_i)$ with $i \neq l$ are zero. In this case we get $H(A) = 0$.
- For equiprobable events the entropy $H(A)$ is maximal.

$$p(a_i) = \frac{1}{M_A} \; \forall i \implies H(A) = \log M_A \tag{2}$$

- The Shannon entropy remains unchanged, when adding impossible events.
- The logarithm in Equation (1) always refers to the natural logarithm except otherwise noted. However, this is a matter of definition only. If the logarithm to base M_A is used, the entropy is normalized.

$$0 \leq H(A) \leq 1 \tag{3}$$

The joint entropy $H(A, B)$ of two discrete systems A and B is defined analogously

$$H(A, B) := -\sum_{i=1}^{M_A} \sum_{j=1}^{M_B} p(a_i, b_j) \log p(a_i, b_j) \tag{4}$$

Here $p(a_i, b_j)$ denotes the joint probability that A is in state a_i and B is in state b_j. The number of possible states M_A and M_B may be different. If the systems A and B are statistically *independent* the joint probabilities factorize and the joint entropy $H(A, B)$ becomes

$$H(A, B) = H(A) + H(B) \tag{5}$$

In general, however, the joint entropy may be expressed in terms of the conditional entropy $H(A|B)$

$$H(A, B) = H(A|B) + H(B) \tag{6}$$

with $H(A|B)$ being defined as

$$H(A|B) := -\sum_{i=1}^{M_A} \sum_{j=1}^{M_B} p(a_i, b_j) \log p(a_i|b_j) \tag{7}$$

Since for arbitrary systems A and B

$$H(A|B) \leq H(A) \tag{8}$$

we get the relation

$$H(A, B) \leq H(A) + H(B) \tag{9}$$

instead of Equation (5). The *mutual information* $I(A, B)$ between the systems A and B is then defined as (Shannon, 1948; Kolmogorov, 1968)

$$I(A, B) := H(A) + H(B) - H(A, B) \; \geq 0 \tag{10}$$

The Kullback entropy

A different approach to the mutual information was given by Kullback (1959). The *Kullback* entropy $K(p|p^0)$ between two probability distributions $\{p\}$ and $\{p^0\}$ is

$$K(p|p^0) := \sum p_i \log \frac{p_i}{p_i^0} \tag{11}$$

The Kullback entropy can be interpreted as the *information gain* when replacing an initial probability distribution $\{p^0\}$ by a final distribution $\{p\}$. Therefore $K(p|p^0)$ establishes a measure of the distance between the distributions $\{p^0\}$ and $\{p\}$. However, the Kullback entropy is not symmetric and thus not a distance in the mathematical sense.

$$K(p|p^0) \neq K(p^0|p) \tag{12}$$

The Kullback entropy $K(p|p^0)$ is always greater than or equal to zero and vanishes if and only if the distributions $\{p\}$ and $\{p^0\}$ are identical (Cover and Thomas, 1991). In our case the *a priori* probability distribution $\{p^0\}$ is given by the hypothesis of statistical independence between the two systems A and B. Thus $p^0(a_i, b_j)$ is the product of the marginal distributions.

$$p^0(a_i, b_j) = p(a_i) \; p(b_j) \tag{13}$$

The elements of the final distribution $\{p\}$ are given by the actual joint probability distribution $p(x_i, y_j)$. Given this choice, the Kullback entropy $K(p|p^0)$ becomes[†]

$$K(p|p^0) = \sum_{ij} p(a_i, b_j) \log \frac{p(a_i, b_j)}{p(a_i)p(b_j)} \qquad (14)$$

Here $K(p|p^0)$ is a measure of 'distance' between our hypothesis that the systems are statistically independent and the actual joint probability distribution. It can be easily verified, that for our particular choice of $\{p\}$ and $\{p^0\}$, the Kullback entropy $K(p|p^0)$ corresponds to the mutual information $I(A, B)$, as defined in Equation (10). Of interest for us is, that the mutual information is zero *if and only if* the measurements on the systems A and B are statistically independent. This puts $I(A, B)$ in contrast to the more commonly used measures, such as Pearson correlation or Euclidean distance, which quantify linear dependencies only. A vanishing mutual information does imply that two variables are independent, while for the Pearson correlation this does *not* hold. Thus, the mutual information can be interpreted as a generalized measure of correlation, analogous to Pearson correlation, but sensitive to any functional relationship, not just linear dependencies. Before we continue with the numerical estimation of the above-defined quantities from finite dataset, we must remark, that the mutual information itself is not a distance in the mathematical sense. However, the definition

$$D(A, B) := H(A, B) - I(A, B) \qquad (15)$$

satisfies the necessary axioms. In this work, we will focus on the mutual information as a measure of similarity and will not explicitly use Equation (15).

NUMERICAL ESTIMATION

Up to now all definitions involved the explicit knowledge of the respective probability distributions. In general, these probabilities are not known, but have to be estimated from measurements. Therefore it remains crucial to investigate possible strategies to give reliable estimates of the mutual information for finite datasets.

The Naive Algorithm

Consider a collection of N simultaneous measurements of two continuous variables x and y (to be later identified with the expression of two genes X and Y under various conditions).

$$\text{measured:} \quad (x_i, y_i) \qquad i = 1, \ldots, N$$

[†] In the following $\sum_{ij}$ is used to denote $\sum_{i=1}^{M_A} \sum_{j=1}^{M_B}$.

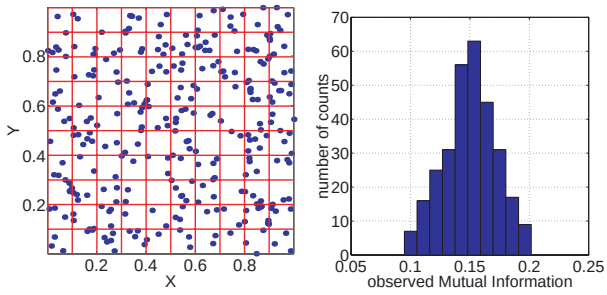

Fig. 1. Naive estimation of the mutual information for finite data. Left: The dataset consists of $N = 300$ artificially generated independent and equidistributed random numbers. The probabilities are estimated using a histogram which divides each axis into $M_x = M_y = 10$ bins. Right: The histogram of the estimated mutual information $I(X, Y)$ obtained from 300 independent realizations.

The most straightforward and widely used (Butte and Kohane, 2000; Michaels *et al.*, 1998) approach is, to use a histogram based technique. Given an origin o and a width h, the bins of the histogram for the variable x are defined through the intervals $[o + mh, o + (m + 1)h]$ with $m = 0, \ldots, M$. The data are thus partitioned into M discrete bins a_i and k_i denotes the number of measurements that lie within the bin a_i. The probabilities $p(a_i)$ are then approximated by the corresponding relative frequencies of occurrence

$$p(a_i) \;\rightarrow\; \frac{k_i}{N} \qquad (16)$$

and the mutual information $I(X, Y)$ between both datasets X and Y may be expressed as

$$I(X, Y) = \log N + \frac{1}{N} \sum_{ij} k_{ij} \log \frac{k_{ij}}{k_i k_j} \qquad (17)$$

Here k_{ij} denotes the number of measurements where x lies in a_i and y in b_j. To demonstrate the application of Equation (17) on experimental data we will now provide a simple numerical example. Our 'measurement' consists of N artificially generated independent and equidistributed random numbers x and y.

$$(x_i, y_i): \qquad x_i, y_i \in [0, 1] \quad \forall i = 1, \ldots, N$$

In this case the systems X and Y are independent and we know the true value of the mutual information $I(X, Y)$ to be zero.

Figure 1 (left) shows an example of $N = 300$ 'measurements' (x_i, y_i). The data was divided into $M_x = M_y = 10$ bins and the mutual information $I(X, Y)$ was calculated using Equation (17). By repeating this experiment with 300 independent realizations of X and Y we obtain a histogram of estimated values for $I(X, Y)$,

as shown in Figure 1 (right). What we observe is that the mutual information not only fluctuates around its true value, but gets *systematically* overestimated. In our case the estimated mutual information is

$$\langle I(X,Y)^{estimated} \rangle \approx 0.15 \pm 0.02$$

instead of the true value $I(X,Y)^{true} = 0$. In the next section we will discuss the reason for this observed systematic error.

Finite size effects

It is known that the estimation of entropies from finite samples may be affected by systematic errors (Grassberger, 1988). Herzel *et al.* (1994) showed that

$$\langle H^{observed} \rangle \approx H - \frac{M-1}{2N} \tag{18}$$

Here $H^{observed}$ denotes the estimated entropy using a finite sample of N datapoints to estimate the probabilities of M discrete states. It should be pointed out that in this approximation the systematic error is *independent* of the underlying probability distribution. Since the mutual information, as defined in Equation (10), is a sum of entropies, we may use this expression to estimate the systematic error of $I(X,Y)$ (Herzel and Grosse, 1995; Grosse, 1996; Roulston, 1999).

$$\langle I^{observed} \rangle \approx I(X,Y)^{\text{true}} + \Delta I(X,Y) \tag{19}$$

With

$$\Delta I(X,Y) = \frac{M_{xy} - M_x - M_y + 1}{2N} \tag{20}$$

Here M_x, M_y and M_{xy} denote the number of discrete states (histogram bins) with nonzero probability. In our previous example we used $M_x = M_y = 10$ and $M_{xy} = 100$. With the true value of $I(X,Y)$ being zero and $N = 300$ we get

$$\langle I^{observed} \rangle = \Delta I(X,Y) \approx 0.14 \tag{21}$$

which is in good agreement with the numerical result. Note that by plotting the observed mutual information as a function of $1/N$ Equation (20) corresponds to a straight line. With $M = M_x = M_y$ we get

$$\langle I^{observed} \rangle \approx I + \frac{(M-1)^2}{2}\frac{1}{N} \tag{22}$$

As seen in Figure 2 a linear extrapolation of the mutual information $I(X,Y)$ in the limit $1/N \to 0$ may improve the result considerably, compared to the uncorrected estimates using a straightforward application of Equation (17).

This is of particular importance for the application on gene-expression data. These datasets are often characterized by missing values, in pair-wise comparisons the mutual information might thus be estimated from datasets of different size. But as we learn from Equation (20), the finite-size corrections depend on the number of datapoints, leading to potentially spurious results. Finally, we must note that Equation (20) represents an *approximation*, which is believed to hold only as long as the number of datapoints is still considerably larger than the number of histogram bins. This must be verified prior to any application (Herzel *et al.*, 1994).

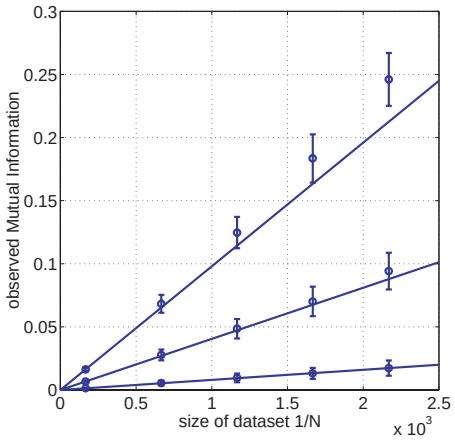

Fig. 2. The observed mutual information for finite data (artificially generated equidistributed random numbers, using M histogram bins on each axis). The straight lines denote the theoretical value for $M = 15, 10, 5$ (from top to bottom). The numerical values (circles) were averaged over an ensemble of 300 trials—the errorbars denote the standard deviation. The numerical simulations are in good agreement with the values predicted by Equation (20).

Adaptive partitioning

We shall now discuss one aspect of the mutual information, which has been neglected so far. As also accounted for by Michaels *et al.* (1998), the mutual information depends on the distribution of the individual datasets: $I(X,Y)$ is bounded by the individual entropies of X and Y.

$$I(X,Y) \leq \min\{H(X), H(Y)\}$$

Since our perspective is to do a comparison of the mutual information between datasets, we have to ascertain that the results are not blurred by the individual distributions of measured gene-expression values. The most straightforward strategy is to normalize all measured datasets to an identical reference distribution. Since correlations are preserved under such a transformation, this does not affect the validity of our analysis. Suppose the dataset X has an arbitrary distribution $p(x)$ which is generated by the rule $x = h(\xi)$ with ξ having the desired well defined simple probability distribution (e.g. uniform or Gaussian) and

h being a static monotonic function. The primary task is now to approximate the inverse function.

$$\xi = h^{-1}(x)$$

This is usually done by rank ordering the elements of $\{\xi\}$ to the same rank ordering as the elements of $\{x\}$ (Rapp *et al.*, 1994; Schreiber and Schmitz, 2000). Two sets have the same rank ordering if the jth element is the kth largest in both sets. In practical applications we rank-order the data to a uniform distribution with zero mean. Alternatively, one might use a equivalent method, which mimics the above described transformation. Instead of using fixed intervals to divide the axes into discrete bins, an adaptive partitioning method is applied. That is, each axis is partitioned into M discrete bins, each bin containing (approximately) $k = N/M$ datapoints Consequently, the width of each interval is determined by the local density of the measured dataset.

The Fraser–Swinney algorithm. Along similar lines, Faser and Swinney (1986) developed a sophisticated algorithm for the estimation of the mutual information in one of the 'classic' papers on this topic. Since this algorithm also involves the idea of adaptive partitioning and is sometimes used in bioinformatic analysis (Samoilov *et al.*, 2001), we shall shortly outline the essential ideas. Instead of choosing a fixed number of intervals, they construct a *hierarchy* of partitions P_m, which recursively divide the (x, y)-plane in smaller and smaller intervals (histogram bins). The crucial observation is, that regions in the (x, y)-plane, where the datapoints are equidistributed cease to contribute further to the estimated mutual information under refinement of the partition and there is no point in subdividing this region further. Thus, the partition in regions of the (x, y)-plane where the datapoints are rather dense becomes finer. Less occupied or empty regions are covered with larger boxes. However, as pointed out by Paluš (1993), this algorithm, though mathematically ingenious, does not lead to a substantial improvement in the estimation of mutual information compared to a simple adaptive partitioning approach.

Experimental data: examples

Before we continue with a more sophisticated approach to estimate the mutual information, we shall shortly outline the application of some of the above described concepts on experimental data. As an example we will use three-hundred full-genome expression profiles from *S. cerevisae* drawn from a publicly available dataset, see Hughes *et al.* (2000) for availability and experimental methods. The complete dataset contains up to $N = 300$ cDNA experiments, but due to missing values the number of simultaneously measured pairs might be slightly less. We start with visual inspection of the data: Figure 3 shows

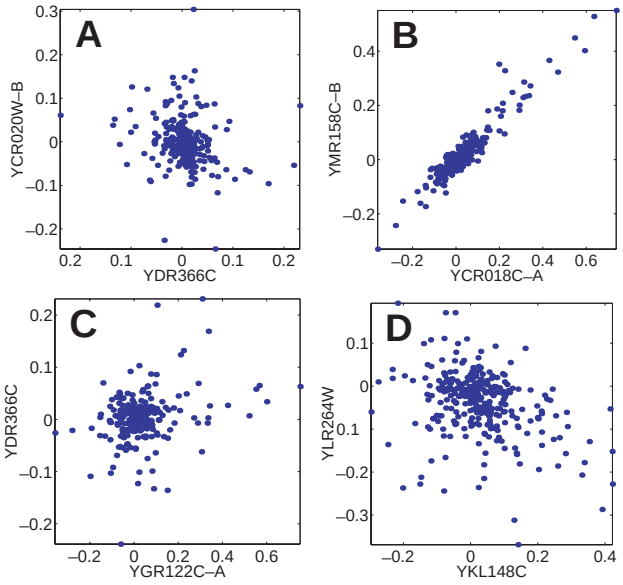

Fig. 3. Examples of simultaneous measurements of expression data. Each dot corresponds to the values (log-transformed) of two open reading frames (ORFs), given on the x- and y-axis, from one cDNA microarray experiment.

examples of simultaneous measurements of expression data.

Each dot corresponds to the log-transformed values of two open reading frames (ORFs), from one cDNA microarray experiment. As could be easily observed, the four examples show varying degrees of correlation. In particular, example B suggests a strong linear relationship, while the examples A, C, and D are not easily classified by eye. Not surprisingly, none of the datasets seems uniformly distributed. On the contrary, we observe large fluctuations and isolated datapoints, which make a computation of the mutual information using a fixed box size problematic. As already noted, it is preferable to have variables of known range and uniform density. Here we achieve this by replacing all values with their respective rank order.

$$(x_i, y_i) \longrightarrow (\mathrm{rk}(x_i), \mathrm{rk}(y_i)) \qquad i = 1, \ldots, N \leq 300$$

Figure 4 shows the rank-ordered version of our examples.

The data are now distributed homogeneously on the x- and y-axis, with the correlations within each dataset preserved.

Testing the significance. Prior to applying any algorithm to the datasets, we have to become clear about how the results of such an analysis should be interpreted. Generally speaking, we set up a *null hypothesis* and test, whether it is consistent with our data. Here, the most natural null hypothesis is to assume that the datasets X and Y are

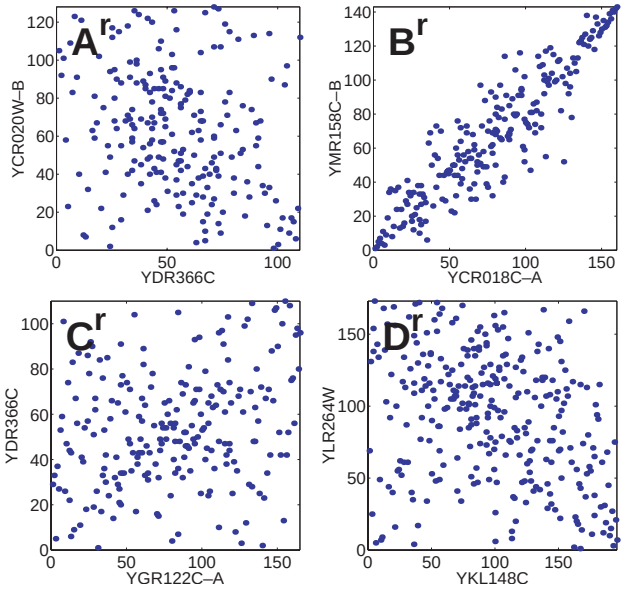

Fig. 4. The rank-ordered representation of the four datasets. Each value was replaced by its respective rank order. Note that the transformation preserved the correlations within each dataset. See text for details.

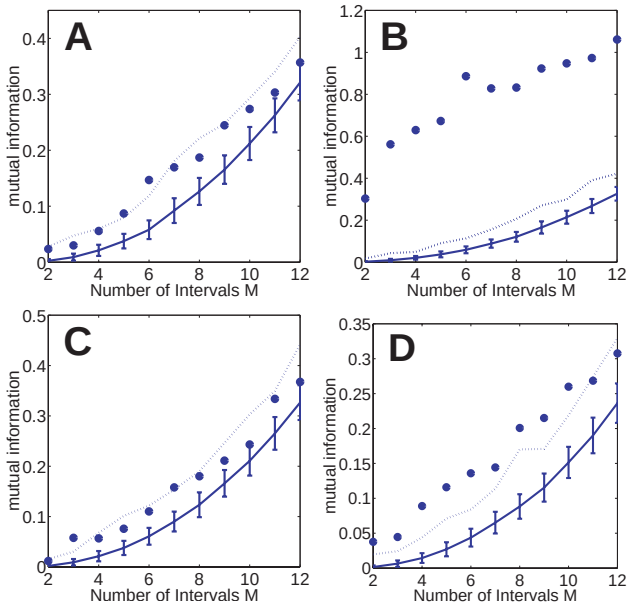

Fig. 5. The estimated mutual information $I(X,Y)$ using the naive algorithm on the rank ordered datasets as a function of the number of intervals $M = M_x = M_y$. The average mutual information obtained from an ensemble of 300 surrogates is drawn as a solid line, with errorbars denoting the standard deviation σ. The largest value found within the ensemble of surrogates is represented by the dotted line.

independent. Only if the observed mutual information is not consistent with this null hypothesis, we may claim that the null hypothesis is wrong, and the data thus contains significant correlation. However, the rejection of a given null hypothesis is never absolute, but always with respect to a certain significance level. Thus, in the interpretation of our test we have to follow Schreiber and Schmitz (2000): A rejection at a given significance level means that if the null hypothesis is true, there is still a certain small probability to see the structure we detected. Non-rejection means even less: Either the null hypothesis is true, or the statistics we use lacks the power to discriminate the data from it. To test our null hypothesis we construct an ensemble of *surrogate* datasets $\{X^s, Y^s\}$, consistent with the null hypothesis. There are two main approaches: *Typical* and *constraint* realization. Constraint realizations are obtained by creating random permutations of the original data X and Y. The values are constrained to take on the same values as the data, just in random order.

For a *typical* realization, we use the data to infer the marginal probability distributions and draw new datasets according to this distribution (Schreiber and Schmitz, 2000). Here, we will only use the first approach. The next step is then to estimate mean and standard deviation σ of the observed mutual information using the ensemble of surrogate data. The significance S is given by

$$S := \frac{I(X,Y)^{data} - \langle I(X,Y)^{surr} \rangle}{\sigma_{surr}} \tag{23}$$

Since this approach implicitly assumes knowledge about the distribution of the mutual information (e.g. $|S| \geq 2.6$ can be treated as significant at a level of 99% assuming a *Gaussian* distribution), it was suggested to use a more general reasoning (Schreiber and Schmitz, 2000; Theiler *et al.*, 1992): First a probability α of false rejection is selected. Then, for a one sided test, $M^s = \frac{1}{\alpha} - 1$ surrogates are generated. Including the original data, we now have an ensemble of $1/\alpha$ datasets. Thus, the probability that our measurement has the largest mutual information within the whole ensemble merely by coincidence is exactly α. The null hypothesis could therefore be rejected at a significance level $(1-\alpha)\cdot 100\%$.

We now give an example in term of the earlier described experimental datasets: Figure 5 shows the estimated mutual information $I(X,Y)$ using the naive algorithm on the rank ordered datasets as a function of the number of intervals $M = M_x = M_y$. The average mutual information obtained from an ensemble of 300 surrogates is drawn as a solid line, with errorbars denoting the standard deviation σ_{surr}. The largest value found within the

ensemble of surrogates is represented by the dotted line. As expected, dataset B shows strong deviations from its shuffled counterparts. Also for dataset D the hypothesis of statistical independence must be rejected, given our previously defined significance level.

However, we must comment on the interpretation of the chosen significance level: Our long-term goal is to identify correlated, and thus presumably co-regulated, genes within the complete genome. Consequently, we are not only concerned with individual pair-wise correlation measures, but with the application of such concepts on the complete dataset, usually involving up to $\sim 10^6$ pair-wise comparisons. Thus, even for a comparably large significance, the *absolute number* of non-correlated pairs which show a 'significant' mutual information (false positives) might still be large. When interpreting large amounts of data, we have to keep this in mind.

KERNEL DENSITY ESTIMATION

We will now resume our considerations towards an effective algorithm for estimating the mutual information between two variables. Until now we have focused on a histogram-based approach, dividing each axis into M discrete non-overlapping intervals. An alternative to this method, based on kernel density estimation (KDE), was suggested by Moon *et al.* (1995) and was found to be superior to the histograms in terms of (i) a better mean square error rate of convergence of the estimate to the underlying density, (ii) an insensitivity to the choice of origin, and (iii) the ability to specify more sophisticated window shapes than the rectangular window for frequency counting (Moon *et al.*, 1995; Silverman, 1986). Since this approach aims at improving the estimate of the probability density $p(x)$ in Equation (14) and is applicable in many other situations apart from the estimation mutual information, we will give a short overview of kernel density estimation in the following, for a comprehensive account we refer the reader to Silverman (1986). The first step is to free the histogram from a particular choice of origin and bin positions. This results in the *naive* estimator

$$\hat{f}(x) = \frac{1}{2Nh} \sum_{i=1}^{N} \Theta(h - |x - x_i|) \tag{24}$$

where $\Theta(x)$ denotes the Heaviside function

$$\Theta(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases} \tag{25}$$

A graphical interpretation of Equation (24) is that the estimator is obtained by additivly putting boxes of width $2h$ and height $(2Nh)^{-1}$ on each observation.

Figure 6 shows an example for the naive estimator using $N = 600$ numerically generated datapoints drawn from

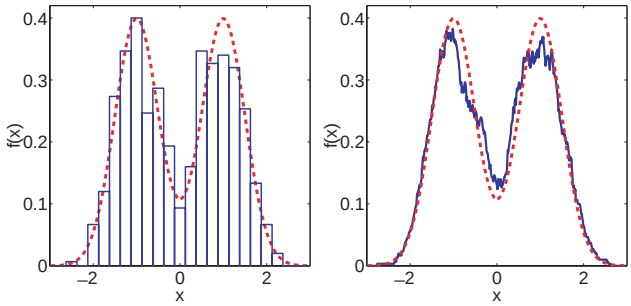

Fig. 6. Density estimation using two different approaches. The underlying bimodal probability distribution is shown as a dotted line (two Gaussian distributions at ± 1, both with standard deviation $\sigma = 0.5$). The probability density estimates $\hat{f}(x)$ were obtained from a artificially generated sample of $N = 600$ datapoints. Left: The histogram estimator using a bin width $h = 0.25$. Right: The naive estimator using $h = 0.2$.

a bimodal distribution (dotted line). There is no particular reason to stay with rectangular cubes as bins. Other shapes still lead to a valid estimate of the probability density. With a generalized weight or kernel function $K(x)$ the *kernel density estimator* $\hat{f}(x)$ is given by

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{26}$$

The parameter h is called *smoothing parameter* or *window width* and the kernel function $K(x)$ is required to be a (normalized) probability density. It may be easily verified that in this case $\hat{f}$ itself is a probability density. Further, $\hat{f}$ will inherit all the continuity and differentiability properties of the kernel K. For simplicity we focus on the Gaussian kernel. The density estimate then reads:

$$\hat{f}(x) = \frac{1}{N} \frac{1}{h\sqrt{2\pi}} \sum_{i=1}^{N} \exp\left(-\frac{(x - x_i)^2}{2h^2}\right) \tag{27}$$

Following the interpretation of the naive estimator, the Gaussian estimator may be explained as placing Gaussian 'bumps' at the position of each observation x_i. The Gaussian estimator is then given by the sum of 'bumps'. Crucial again is the choice of the bandwidth h: If h is chosen to small spurious fine structure becomes visible, while if h is too large all detail, spurious or otherwise, is obscured. While there are several methods for choosing an appropriate bandwidth available, most of them are associated with a considerable computational burden (Moon *et al.*, 1995). As a tradeoff between computational effort and performance one may choose the 'optimal' bandwidth as the one that minimizes the mean integrated square error, assuming the underlying

distribution is Gaussian. Following Silverman (1986), the optimal *Gaussian* bandwidth h_{opt} is given by

$$h_{opt} = \left(\frac{4}{3N}\right)^{\frac{1}{5}} \sigma \approx 1.06\sigma N^{-\frac{1}{5}} \tag{28}$$

where σ denotes the standard deviation of the data. For the estimation of the mutual information, we also need the two-dimensional probability density $p(x, y)$. With the abbreviation

$$d_i(x, y) = \sqrt{(x - x_i)^2 + (y - y_i)^2}$$

the two-dimensional Gaussian kernel estimate $\hat{f}_g(x, y)$ reads

$$\hat{f}_g(x, y) = \frac{1}{Nh^2}\frac{1}{2\pi}\sum_{i=1}^{N} \exp\left(-\frac{d_i(x, y)^2}{2h^2}\right) \tag{29}$$

As in the one-dimensional case, an appropriate value for h depends on the unknown density being estimated. Under the assumption that this density is Gaussian, an approximately optimal value is given by

$$h_{opt} \approx \sigma \left(\frac{4}{d+2}\right)^{1/(d+4)} N^{-1/(d+4)} \tag{30}$$

with $d = 2$ being the dimension of the dataset and σ the average marginal standard deviation (Silverman, 1986).

Estimating the mutual information

The mutual information $I(X, Y)$ is a functional of probability densities. Thus an obvious way to find an estimate for $I(X, Y)$ is to find estimates of the densities and then to substitute these into the required integral. However, this is not as trivial as it may seem. Up to now, we have always referred to the *probability* of discrete states, while by kernel estimation we obtain probability *densities* only. Remarkably enough, the discretization of the (x, y)-plane into infinitesimal bins of size $\Delta V = \Delta x\, \Delta y$ corresponds to the continuous form of the mutual information.

$$\hat{I}(X, Y) = \int_x \int_y \hat{f}(x, y) \log \frac{\hat{f}(x, y)}{\hat{f}(x)\,\hat{f}(y)} \mathrm{d}x\, \mathrm{d}y \tag{31}$$

Note that such a correspondence does *not* hold for the individual entropies used in Equation (10). Quite on the contrary: For a random variable with an arbitrary distribution the continuous expression for the entropy is finite, while the discrete diverges towards infinity as the bin size tends to zero[‡].

To evaluate Equation (31), we have to integrate over a smooth function. The choice of the integration steps Δx and Δy could thus be based entirely on standard procedures for numerical integration. Note that the discretization introduced by the numerical integration does not correspond to our earlier attempts to partition the data. Figure 7 shows the estimated mutual information for the rank ordered datasets A to D as a function of the smoothing parameter h. The results were compared to uncorrelated datasets, using an ensemble of 100 randomly shuffled realizations. As could be observed, the standard deviation obtained from the ensemble of surrogates is much smaller compared to the histogram-based algorithm. The discriminability between the correlated and uncorrelated datasets is thus enhanced. Further, the estimated mutual information is much less sensitive to the choice of the smoothing parameter h, than the probability density itself. For all datasets the 'optimal' smoothing parameter h_{opt} (denoted with an arrow) seems to be an appropriate choice.

A simpler algorithm. Depending on the application, the numerical integration of Equation (31) might put high demands on computational power. So we may ask for strategies to simplify the algorithm. As already noted in the introduction, entropy measures represent an *average* over a probability distribution. According to Equation (14), the estimated mutual information may thus be written as

$$\hat{I}(X, Y) = \left\langle \log \frac{\hat{f}(x, y)}{\hat{f}(x)\,\hat{f}(y)} \right\rangle \tag{32}$$

Under the assumption that our dataset is a faithful sample of the underlying probability distribution, we get

$$\hat{I}(X, Y) = \frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{\hat{f}(x_i, y_i)}{\hat{f}(x_i)\,\hat{f}(y_i)}\right] \tag{33}$$

However, we must note that Equation (33) should be applied with caution. For example, the individual datapoints have to be independent realizations of the underlying distribution—a requirement not always fulfilled by experimental data.

NETWORK ANALYSIS AND RESULTS

In the last section, we discuss the potential results obtained from a cluster analysis based on mutual information. First of all, clustering based on Pearson correlation or Euclidean distance is probably the most widely used method for analyzing and visualizing expression data. It has already been shown extensively, that the results obtained from such an analysis leads to biologically relevant insights (Eisen *et al.*, 1998). Also, direct applications of mutual information as a measure of distance showed that it groups together genes of known similar function (Butte and Kohane, 2000; Daub *et al.*, 2002). In the present work,

[‡] However, in Equation (10) these terms cancel out.

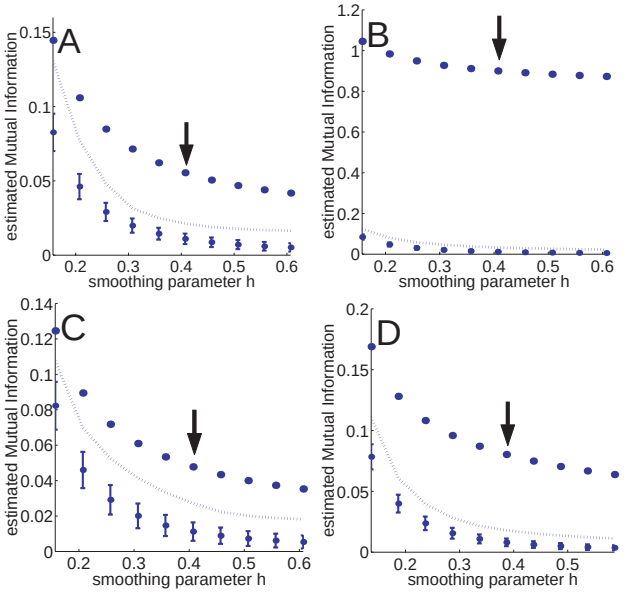

Fig. 7. The estimated mutual information for the (rank-ordered) datasets A to D as a function of the smoothing parameter h, using a Gaussian kernel density estimator. Each result was compared to an ensemble of 100 shuffled surrogates (lower dots) with errorbars denoting the standard deviation. The dotted line gives the maximal mutual information found within the ensemble of surrogates. In each plot the optimal smoothing parameter h_{opt} is denoted with an arrow.

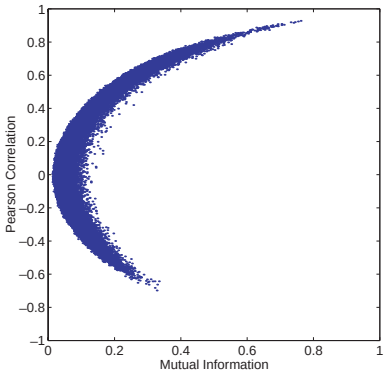

Fig. 8. A comparison between the mutual information $I(X,Y)$ and the Pearson correlation $\hat{C}_{xy}$. Each dot corresponds to the tuple $(\hat{I}(X,Y),\ \hat{C}_{xy})$ for a pair of genes X and Y. Plotted is only a representative fraction of the $L(L-1)/2$ possible pair-wise comparisons. See text for details.

we will therefore focus solely on a *comparison* of the mutual information to the Pearson correlation. Do we detect non-linear relationships in the data, which were previously missed by linear measures? To answer this question, we evaluate both, the pair-wise mutual information and the Pearson correlation (correlation coefficient), defined as

$$\hat{C}_{xy} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \langle x\rangle}{\sigma_x}\right)\left(\frac{y_i - \langle y\rangle}{\sigma_y}\right) \qquad (34)$$

where $\langle\cdot\rangle$ and σ denotes the mean and standard deviation respectively, for the previously described experimental dataset (Hughes *et al.*, 2000). The dataset contains up to $N = 300$ expression values for 6314 genes. Genes with less than 300 expression measurements were discarded from the analysis, resulting in $L = 5345$ fully defined rows. For each of the $L(L-1)/2$ pair-wise comparisons, we evaluated the mutual information according to Equation (33), and the correlation coefficient $\hat{C}_{xy}$. To compare both measures, we plot the tuple $(\hat{I}(X,Y),\ \hat{C}_{xy})$ for each pair of genes X and Y, see Figure (8).

First, we confirm some well-known results: The Pearson correlation $\hat{C}_{xy}$ distinguishes between positive and negative correlations, while the mutual information does not. Positive correlations are much more frequent than negative ones. Further, the Pearson correlation $\hat{C}_{xy}$ is bound by the mutual information: Except for numerical or statistical errors, a situation with high Pearson correlation $|\hat{C}_{xy}|$ and low mutual information does not exist. However, more important for us is, that within the analyzed dataset there seems to be almost a one-to-one correspondence between mutual information and $|\hat{C}_{xy}|$ (apart from statistical fluctuations). As could be observed in Figure (8) we detect *no* genuinely non-linear correlation. This does have implications for further analysis. Most of all, it means that previous investigations using Pearson correlation as a measure of similarity for gene-expression measurements were justified and did not miss a significant fraction of possible correlations: The correlations between simulateously measured gene-expression values are—if any—essentially linear. Here we can only speculate about the reasons. To some extend this may reflect the inherent robustness in genetic networks: Many gene products are known to behave in a highly coordinated fashion. Further, the detection of truly non-linear relationships usually requires a large amount of *accurately* measured datapoints. It may be easily conceived, that measurement errors first affect the detectability of highly non-linear correlations, while linear relationships are still visible.

CONCLUSION

We presented several approaches to estimate the mutual information from finite data. Starting with a histogram-based method, we discussed the systematic errors due to the finite size of the dataset. As an alternative, a kernel-based approach was described and exemplified using artificially generated numbers, as well as publicly

available datasets of full-genome expression profiles. A comparison between the mutual information and the Pearson correlation for one particular full-genome dataset revealed that there are presently no genuinely non-linear correlations detectable in this dataset. Here we could only speculate about the reasons behind this finding, a more thorough discussion will be given elsewhere.

The authors would like to thank W.Ebeling (HU-Berlin), J.Kopka (MPIMP) and S.Kloska (Scienion AG, Berlin) for stimulating discussion.

REFERENCES

Brazma,A. and Vilo,J. (2000) Gene expression data analysis. *FEBS Lett.*, **480**, 17–24.

Butte,A. and Kohane,I.S. (2000) Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, **5**, 415–426.

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. Wiley, New York.

Daub,C.O., Weise,J., Steuer,R., Kopka,J. and Kloska,S. (2002) Using b-spline functions to introduce fuzzyness to mutual information based analysis of gene-expression data. (manuscript in preparation).

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

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Faser,A.M. and Swinney,H.L. (1986) Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, **33**, 2318–2321.

Grassberger,P. (1988) Finite sample corrections to entropy and dimension estimates. *Phys. Lett. A*, **128**, 369.

Grosse,I. (1996) Estimating entropies from finite samples. In Freund,JanA. (ed.), *Dynamik–Evolution–Strukturen*. Dr. Koster.

Herzel,H. and Grosse,I. (1995) Measuring correlations in symbols sequences. *Physica A*, **216**, 518–542.

Herzel,H., Schmitt,A.O. and Ebeling,W. (1994) Finite sample effects in sequence analysis. *Chaos, Solitons and Fractals*, **4**, 97–113.

Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D., Kidd,M.J., King,A.M., Meyer,M.R., Slade,D., Lum,P.Y., Stepaniants,S.B., Shoemaker,D.D., Gachotte,D., Chakraburtty,K., Simon,J., Bard,M. and Friend,S.H. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126. see http://www.rii.com/publications/cell_hughes.htm for additional information.

Kolmogorov,A.N. (1968) Logical basis for information theory and probability theory. *IEEE Trans. Information Theor.*, **14**, 662–664.

Kullback,S. (1959) *Information Theory and Statistics*. Wiley, New York.

Michaels,G., Carr,D., Askenazi,M., Fuhrman,S., Wen,X. and Somogyi,R. (1998) Cluster analysis and data visualization of large-scale gene expression data. *Pac. Symp. Biocomput.*, **3**, 42–53.

Moon,Y., Rajagopalan,B. and Lall,U. (1995) Estimation of mutual information using kernel density estimators. *Phys. Rev. E*, **52**, 2318–2321.

Nature (1999) The chipping forecast. *Nature Genet. (suppl.)*, **21**, 1–60.

Paluš,M. (1993) Identifying and quantifying chaos by using information-theoretic functionals. In Weigend,A.S. and Gershenfeld,N.A. (eds), *Time Series Prediction: Forecasting the Future and Understanding the past*, SFI Studies in the Sciences of Complexity, Proc. Vol. XV, Addison Wesley.

Rapp,P.E., Zimmerman,I.D., Vining,E.P., Cohen,N., Albano,A.M. and Jiménez-Montaño,M.A. (1994) The algorithmic complexity of neural spike trains increases during focal seizures. *J. Neurosci.*, **14**, 4731–4739.

Roulston,M.S. (1999) Estimating the errors on measured entropy and mutual information. *Physica D*, **125**, 285–294.

Samoilov,M., Arkin,A. and Ross,J. (2001) On the deduction of chemical reaction pathways from measurements of time series of concentrations. *Chaos*, **11**, 108–114.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schreiber,T. and Schmitz,A. (2000) Surrogate time series. *Physica D*, **142**, 346–382.

Shannon,C.E. (1948) A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423. *ibid*, 623–656

Silverman,B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Theiler,J., Eubank,S., Longtin,A., Galdrikian,B. and Farmer,J.D. (1992) Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, **58**, 77.