Briefings in Bioinformatics Advance Access published August 20, 2013

doi:10.1093/bib/bbt051

A comparative study of statistical methods used to identify dependencies between gene expression signals

Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, Asuka Nakata and André Fujita

Submitted: 18th April 2013; Received (in revised form): 10th June 2013

Abstract

One major task in molecular biology is to understand the dependency among genes to model gene regulatory networks. Pearson's correlation is the most common method used to measure dependence between gene expression signals, but it works well only when data are linearly associated. For other types of association, such as non-linear or non-functional relationships, methods based on the concepts of rank correlation and information theory-based measures are more adequate than the Pearson's correlation, but are less used in applications, most probably because of a lack of clear guidelines for their use. This work seeks to summarize the main methods (Pearson's, Spearman's and Kendall's correlations; distance correlation; Hoeffding's ***D*** measure; Heller–Heller–Gorfine measure; mutual information and maximal information coefficient) used to identify dependency between random variables, especially gene expression data, and also to evaluate the strengths and limitations of each method. Systematic Monte Carlo simulation analyses ranging from sample size, local dependence and linear/non-linear and also non-functional relationships are shown. Moreover, comparisons in actual gene expression data are carried out. Finally, we provide a suggestive list of methods that can be used for each type of data set.

Keywords: *correlation; dependence; network; data analysis*

INTRODUCTION

In bioinformatics, the notion of dependence is central to model gene regulatory networks. The functional relationships and interactions among genes and their products are usually inferred by using statistical methods that identify dependence among signals [1–4].

One well-known method to identify dependence between gene expression signals is the Pearson's correlation. But Pearson's correlation, although it is one of the most ubiquitous concepts in modern molecular biology, is also one of the most misunderstood concepts. Some of the confusion may arise from the literary use of the word to cover any notion of dependence. To a statistician, correlation is only one particular measure of stochastic dependence among many. It is the canonical measure in the world of multivariate normal distributions, and more generally for spherical and elliptical distributions. However, empirical research in molecular biology shows that the distributions of gene expression signals may not belong in this class [4]. To identify associations not limited to linear associations, but dependent in a broad sense, several methods have been developed, most of them based on ranks or information theory.

The main aim of this article is to clarify the essential mathematical ideas behind several methods [Pearson's correlation coefficient [5,6], Spearman's correlation coefficient [7], Kendall's correlation

Corresponding author. André Fujita, Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, Rua do Matão, 1010–Cidade Universitária–SP, 05508-090, Brazil. Tel.: + 55 11 3091 5177; Fax.: + 55 11 3091 6134; E-mail: fujita@ime.usp.br

Suzana de Siqueira Santos obtained her BS in Computer Science in 2012 at the University of São Paulo. She is currently a master course student in Computer Science at the same university.

Daniel Yasumasa Takahashi received his Ph.D. in Bioinformatics in 2009 at the University of São Paulo. He is currently a post-doc at Princeton University.

Asuka Nakata received her Ph.D. in Biosciences in 2009. She is currently a post-doc at the University of Tokyo.

André Fujita received his Ph.D. in Bioinformatics in 2007. He is currently Assistant Professor at the University of São Paulo.

coefficient [8], distance correlation [9], Hoeffding's D measure [10], Heller–Heller–Gorfine (HHG) measure [11], mutual information (MI) [12] and maximal information coefficient (MIC) [13]] used to identify dependence between random variables that anyone wishing to model dependent phenomena should know. Furthermore, we illustrate by Monte Carlo simulations where we varied sample size, local dependence and linear/non-linear and also non-functional relationships, the strengths and limitations concerning the different measures used to identify dependent signals. Finally, we illustrate the application of the methods in actual gene expression data with known dependence structure between the genes. Thus we hope to provide the necessary elements for a better comprehension of the methods and also the choice of a suitable dependence test method, based on practical constraints and goals.

STATISTICAL INDEPENDENCE BETWEEN TWO RANDOM VARIABLES

Statistical independence indicates that there is no relation between two random variables. If the variables are statistically independent, then the distribution of one of them is the same no matter at which fixed levels the other variable is considered, and observations for such variables will lead correspondingly to nearly equal frequency distributions. On the other hand, if there is dependence, then the levels of one of the variables vary with changing levels of the other. In other words, under independence, knowledge about one feature remains unaffected by information provided about the other, whereas under dependence, it follows which level of one variable occurs as soon as the level of the other variable is known.

Formally, two random variables X and Y with cumulative distribution functions $F_X(x)$ and $F_Y(y)$, and probability densities $f_X(x)$ and $f_Y(y)$, are independent if and only if the combined random variable (X, Y) has a joint cumulative distribution function $F_{X,Y}(x, y) = F_X(x)F_Y(y)$, or, equivalently, a joint density $f_{X,Y}(x, y) = f_X(x)f_Y(y)$. We say that two random variables X and Y are dependent if they are not independent. The problem then is how to measure and detect dependence from the observation of the two random variables.

MEASURES OF DEPENDENCE BETWEEN RANDOM VARIABLES

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of joint n observations from two univariate random variables X and Y.

The test of dependence between X and Y is described as a hypothesis test as follows:

H_0: X and Y are not dependent (null hypothesis).
H_1: X and Y are dependent (alternative hypothesis).

In the next section, we will describe frequently used methods to identify dependent data and also show by simulations that some methods such as Pearson's, Spearman's and Kendall's correlations can only detect linear or non-linear monotonic (strictly increasing or strictly decreasing function, i.e. a function that preserves the given order) relationships, whereas others such as distance correlation, HHG measure, Hoeffding's D measure and MI are able to identify non-monotonic and non-functional relationships also. Furthermore, we will see that although MIC is not mathematically proven to be consistent against all general alternatives, it can detect some non-monotonic relationships.

Pearson's product-moment correlation coefficient

The Pearson's product-moment correlation or simply Pearson's correlation [5,6] is a measure of linear dependence, as the slope obtained by the linear regression of Y by X is Pearson's correlation multiplied by that ratio of standard deviations.

Let $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ be the means of X and Y, respectively. Then, the Pearson's correlation coefficient ρ_{Pearson} is defined as follows:

$$\rho_{\text{Pearson}}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

For joint normal distributions, Pearson's correlation coefficient under H_0 follows a Student's t-distribution with $n - 2$ degrees of freedom. The t statistic is as follows:

$$t = \frac{\rho_{\text{Pearson}}(X, Y)\sqrt{n - 2}}{\sqrt{1 - \rho^2_{\text{Pearson}}(X, Y)}}$$

When the random variables are not jointly normally distributed, the Fisher's transformation is used to get an asymptotic normal distribution.

In the case of perfect linear dependence, we have $\rho_{\text{Pearson}}(X, Y) = \pm 1$. The Pearson correlation is $+1$ in the case of a perfect positive (increasing) linear relationship and -1 in the case of a perfect negative (decreasing) linear relationship. In the case of linearly independent random variables, $\rho_{\text{Pearson}}(X, Y) = 0$, and in the case of imperfect linear dependence, $-1 < \rho_{\text{Pearson}}(X, Y) < 1$. These last two cases are the ones for which misinterpretations of correlation are possible because it is usually assumed that non-correlated X and Y means independent variables, whereas in fact, they may be associated in a non-linear fashion that Pearson's correlation coefficient is not able to identify.

The R function for Pearson's test is cor.test with parameter method='pearson' (package *stats*). The *stats* package can be downloaded from the R [14] Web page (http://www.r-project.org).

Spearman's rank correlation coefficient

Unlike the Pearson's correlation coefficient, Spearman's rank correlation or simply Spearman's correlation [7] does not require assumptions of linearity in the relationship between variables, nor should the variables be measured at interval scales, as it can be used for ordinal variables.

Let ρ_{Spearman} be simply the application of Pearson's correlation in the data converted to ranks before calculating the coefficient. Thus, Spearman's rank correlation can capture monotonic relationships, i.e. if values of Y tend to increase (or decrease) when values of X increase.

Another simpler procedure used to calculate ρ_{Spearman} is to convert the raw values of x_i and y_i to ranks, and calculate the differences d_i between the ranks of x_i and y_i and calculate the Spearman's rank correlation coefficient as:

$$\rho_{\text{Spearman}}(X, Y) = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

The Spearman's correlation coefficient under H_0 can be asymptotically approximated by a Student's t-distribution with $n - 2$ degrees of freedom:

$$t = \frac{\rho_{\text{Spearman}}\sqrt{n-2}}{\sqrt{1 - \rho_{\text{Spearman}}^2}}$$

Similarly to Pearson's correlation coefficient, ρ_{Spearman} assumes values between -1 and 1, where Spearman's correlation is $+1$ in the case of a perfect monotonically increasing relationship (for all x_1 and x_2 such that $x_1 < x_2$, we have $y_1 < y_2$), and -1 in the case of a perfect monotonically decreasing relationship (for all x_1 and x_2 such that $x_1 < x_2$, we have $y_1 > y_2$). In the case of monotonically independent random variables, $\rho_{\text{Spearman}}(X, Y) = 0$, and in the case of imperfect monotonically dependence, $-1 < \rho_{\text{Spearman}}(X, Y) < 1$. Again, similarly to Pearson's correlation coefficient, $\rho_{\text{Spearman}}(X, Y) = 0$ does not mean that random variables X and Y are independent, but only that they are monotonically independent.

The R function for Spearman's test is cor.test with parameter method='spearman' (package *stats*). The *stats* package can be downloaded from the R Web page (http://www.r-project.org).

Kendall τ rank correlation coefficient

The Kendall τ rank correlation coefficient or simply Kendall's correlation [8] is an alternative method to Spearman's correlations, i.e. it also identifies monotonic relationships.

Kendall's correlation is defined as [8]:

$$\tau(X, Y) = \frac{\left\{\begin{array}{l}(\text{number of concordant pairs}) \\ -(\text{number of discordant pairs})\end{array}\right\}}{0.5n(n-1)}$$

where *concordant* means if the ranks for both elements agree: that is, if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. They are said to be *discordant*, if $x_i > x_j$ and $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant. Specifically, for a pair of objects taken at random, τ can be interpreted as the difference between the probability for these objects to be in the same order and the probability of these objects being in a different order.

A null hypothesis test can be performed by transforming τ into a Z value as $Z_\tau = \frac{\tau}{\sigma_\tau}$, where $\sigma_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$ is the standard deviation of τ. This Z value is asymptotically normally distributed with a mean of 0 and a standard deviation of 1. The Kendall's rank correlation coefficient varies from -1 to 1, and the interpretations are the same for the Spearman's correlation coefficient.

The R function for Kendall's τ test is cor.test with parameter method='kendall' (package *stats*). The package *stats* can be downloaded from the R Web page (http://www.r-project.org).

Distance correlation

Distance correlation was introduced by Szekely *et al.* (2007) [9] to identify non-linear relationships between two random variables. The name distance correlation comes from the fact that it is based on the concept of energy distances (a statistical distance between probability distributions). The distance correlation is given by dividing the distance covariance between X and Y by the product of their distance standard deviations. In other words, let $\|\cdot\|$ be the Euclidian distance, $a_{k,l} = \|x_k - x_l\|$ and $b_{k,l} = \|y_k - y_l\|$ for $k, l = 1, 2, \ldots, n$. Define $\overline{a}_k$ as the kth row mean, $\overline{a}_{,l}$ as the lth column mean and $\overline{a}$ as the grand mean of the distance matrix of X; $\overline{b}_k$ as the kth row mean, $\overline{b}_{,l}$ as the lth column mean and $\overline{b}$ as the grand mean of the distance matrix of Y; $A_{k,l} = a_{k,l} - \overline{a}_k - \overline{a}_{,l} + \overline{a}$, and $B_{k,l} = b_{k,l} - \overline{b}_k - \overline{b}_{,l} + \overline{b}$. Then, the distance covariance between X and Y, the distance variance of X and the distance variance of Y are defined as $\mathrm{dCov}(X, Y) = \sqrt{\frac{1}{n^2}\sum_{k=1,\,l=1}^{n} A_{k,l}B_{k,l}}$, $\mathrm{dVar}(X) = \mathrm{dCov}(X, Y)$ and $\mathrm{dVar}(Y) = \mathrm{dCov}(Y, Y)$, respectively.

Using aforementioned definitions, the distance correlation is defined as follows:

$$\mathrm{dCor}(X, Y) = \frac{\mathrm{dCov}(Y, Y)}{\sqrt{\mathrm{dVar}(X)\mathrm{dVar}(Y)}}$$

The value of dCor ranges between 0 and 1, where distance correlation is 1 in the case of a perfect linear dependence and 0 in case the random variables are not dependent. In the case of imperfect linear dependence, $0 < \mathrm{dCor}(X, Y) < 1$.

To estimate the P-value under H_0, a permutation test [9,11] can be used to test if $\mathrm{dCor} = 0$ (which occurs if and only if $\mathrm{dCov} = 0$).

The permutation procedure to test independence between two random variables X and Y is as follows:

(i) Construct a permutated data set under the null hypothesis $(x_1, y_1^*), (x_2, y_2^*), \ldots, (x_n, y_n^*)$ by fixing x_i and permuting y_i;
(ii) Calculate the test statistic dCov on this permutated data set (x, y^*);
(iii) Repeat steps (i) and (ii) until the desired number of permutated replications is obtained.

The P-value from the permutation test is the fraction of replicates of dCor on the permutated data set (x, y^*) that are at least as large as the observed statistic on the original data set (x, y).

The R function for distance correlation test is dcov.test (package *energy*). The *energy* package can be downloaded from the R Web page (http://www.r-project.org).

Hoeffding's *D* measure

The intuitive idea of Hoeffding's D measure [10] is to test the independence of the data sets by calculating the distance between the product of the marginal distributions under the null hypothesis and the empirical bivariate distribution.

Let R_i and S_i be the rank of x_i and y_i, respectively, and Q_i be the number of points with both x and y values less than the ith point, i.e. $Q_i = \sum_{j=1}^{n} \phi(x_j, x_i)\phi(y_j, y_i)$, where $\phi(a, b) = 1$ if $a < b$ and $\phi(a, b) = 0$ otherwise. In other words, the quantity Q_i is the number of bivariate observations (x_j, y_j) for which $x_j < x_i$ and $y_j < y_i$. Set $D_1 = \sum_{i=1}^{n} Q_i(Q_i - 1)$, $D_2 = \sum_{i=1}^{n} (R_i - 1)(R_i - 2)(S_j - 1)(S_j - 2)$ and $D_3 = \sum_{i=1}^{n} (R_i - 2)(S_i - 2)Q_i$

Then, the formula for Hoeffding's $\boldsymbol{D}$ measure is given by:

$$D(X, Y) = \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}.$$

Asymptotically, the test for independence can be carried out as follows: let α be the desired level of significance, P be the probability distribution under H_0 and ρ_n be the smallest number satisfying the inequality $P\{D(X, Y) > \rho_n\} \leq \alpha$. Reject the hypothesis H_0 of independence if and only if $P\{D(X, Y) > \rho_n$, where $\rho_n = \frac{1}{30}\sqrt{\frac{2(n^2+5n-32)}{9n(n-1)(n-3)(n-4)\alpha}}$. Hoeffding's D measure ρ_n varies from $-\frac{1}{60}$ to $\frac{1}{30}$.

Differently of Pearson's, Spearman's and Kendall's correlation measures, the positive and negative signs of ρ_n do not have any interpretations, because Hoeffding's $\boldsymbol{D}$ measure identifies non-monotonic relationships also.

The R function for Hoeffding's D test is hoeffd (package *Hmisc*). The package *Hmisc* can be downloaded from the R Web page (http://www.r-project.org).

Heller, Heller and Gorfine measure

Heller, Heller and Gorfine [11] propose a test of independence based on the distances among values of X and Y, i.e. $d(x_i, x_j)$ and $d(y_i, y_j)$ for $i, j \in \{1, \ldots n\}$, respectively. Intuitively, note that if

X and Y are dependent and have a continuous joint density, then there exists a point (x_i, y_i) in the sample space of (X, Y), and radii around x_i and y_i, respectively, such that the joint distribution of X and Y differs from the product of the marginal distributions in the Cartesian product of balls around (x_i, y_i) [11].

Heller *et al.* perform the test in the following manner. For each observation i and each $j \neq i$, $i \leq n, j \leq n$, define:

$$\begin{aligned}
A_{11}(i,j) &= \sum_{k=1,k\neq i,k\neq j}^{n} I\{d(x_i,x_k) \leq d(x_i,x_j)\} \\
&\quad I\{d(y_i,y_k) \leq d(y_i,y_j)\}, \\
A_{12} &= \sum_{k=1,k\neq i,k\neq j}^{n} I\{d(x_i,x_k) \leq d(x_i,x_j)\} \\
&\quad I\{d(y_i,y_k) > d(y_i,y_j)\}, \\
A_{21} &= \sum_{k=1,k\neq i,k\neq j}^{n} I\{d(x_i,x_k) > d(x_i,x_j)\} \\
&\quad I\{d(y_i,y_k) \leq d(y_i,y_j)\}, \\
A_{22} &= \sum_{k=1,k\neq i,k\neq j}^{n} I\{d(x_i,x_k) > d(x_i,x_j)\} \\
&\quad I\{d(y_i,y_k) > d(y_i,y_j)\},
\end{aligned}$$

where $I\{.\}$ is the indicator function.

Let $S(i,j) = \dfrac{(n-2)\left\{A_{12}(i,j)A_{21}(i,j) - A_{11}(i,j)A_{22}(i,j)\right\}^2}{A_{1.}(i,j)A_{2.}(i,j)A_{.1}(i,j)A_{.2}(i,j)}$

where $A_{m.} = A_{m1} + A_{m2}$ and $A_{.m} = A_{1m} + A_{2m}$ for $m = 1, 2$.

To test the independence between two random variables X and Y, Heller *et al.* suggested the following statistic:

$$T = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq 1}}^{n} S(i,j)$$

where the expectation of T is $n(n-1)$ under the null hypothesis (for independent data).

The permutation test for HHG measure is the same performed for the distance correlation in the section Distance correlation.

The R function for HHG test is pvHHG (package *HHG2x2*). The package *HHG2x2* can be downloaded from Ruth Heller's Web page (http://www.math.tau.ac.il/~ruheller/Software.html).

Mutual information

MI is one of many quantities that measure how much one random variable tells us about another. The MI of two continuous random variables X and Y can be defined as:

$$MI(X,Y) = \iint f_{X,Y}(x,y) \log_2\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right) dxdy$$

MI can assume only positive values. High MI indicates a large reduction in uncertainty, low MI indicates a small reduction and zero MI between two random variables means the variables are independent [12]. Notice that if X and Y are independent, by the definition of dependent data given in the section STATISTICAL INDEPENDENCE BETWEEN TWO RANDOM VARIABLES ($f_{X,Y}(x,y) = f_X(x)f_Y(y)$), we have that $\log_2\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right) = 0$, and consequently, $MI(X,Y) = 0$.

As an analytical statistical test for MI is in general not available [15], the significance of the MI should be tested by assuming strong constraints on the data or by using a permutation procedure. The permutation test used for MI is the same as described in the section Distance correlation.

There are several algorithms to estimate MI [15–17]. For discrete data, density functions $f_X(x)$, $f_Y(y)$ and $f_{X,Y}(x,y)$ can be estimated by simply counting the events. For continuous data, one well-known method is based on estimating the density function by the Gaussian kernel regression using the Nadaraya–Watson estimator [18], and normalizing the integral under the curve to 1. The choice of a different estimator or set of parameters for MI may vary the estimations considerably. Here, we chose the kernel density estimator because it was already described to be better than the standard histogram-based estimator [19] and also provided good results in other comparative studies [20]. We do not discuss further details about the methods to estimate MI because it is not the scope of this work. For a good review, see [21].

The R function for MI is mi.empirical (package *entropy*). The package *entropy* can be downloaded from the R Web page (http://www.r-project.org).

Maximal information coefficient

Intuitively, MIC [13] is based on the idea that if a relationship exists between two random variables, then a grid can be drawn on the scatterplot of the two variables that partitions the data to encapsulate that relationship. Thus, to calculate the MIC of a set of two-variable data, all grids up to a maximal grid resolution are explored, dependent on the sample

size, computing for every pair of integers, (a, b) the largest possible MI achievable by any a-by-b grid applied to the data. Then, these MI values are normalized to ensure a fair comparison between grids of different dimensions and to obtain modified values between 0 and 1. The characteristic matrix M is defined as $M = (m_{a,b})$, where $(m_{a,b})$ is the highest normalized MI achieved by an a-by-b grid, and the statistic MIC to be the maximum value in M.

Formally, for a grid G, let MI_G denote the MI of the probability distribution induced on the boxes of G, where the probability of a box is proportional to the number of data points falling inside the box. The (a, b)-th entry $m_{a,b}$ of the characteristic matrix equals $M_{a,b} = \frac{(MI_G)}{\log \min\{a, b\}}$, where the maximum is taken over all a-by-b grids G. MIC is the maximum of $m_{a,b}$ over ordered pairs (a, b) such that $ab < B$, where B is a function of sample size. Usually, set $B = n^{0.6}$.

The permutation test for MIC is the same as that performed for distance correlation in the section Distance correlation.

The Java code to compute MIC with an R wrapper can be downloaded from http://www.exploredata.net/Downloads/MINE-Application.

COMPARATIVE STUDIES

To compare the performance of the eight methods, both simulations and applications to actual biological data sets were carried out.

Simulations

To illustrate the strengths and limitations of each method, we performed a systematic simulation study that analyzes the effects of the number of observations and type of dependence (linear, non-linear monotonic/non-monotonic and non-functional). Figure 1 is an example of the types of relationships studied here. The construction of the scenarios is described in the supplementary material. Figure 1A is the case that two random variables are independent, i.e. under the null hypothesis. Linear association (alternative hypothesis) is represented by a line (Figure 1B), non-linear monotonic association (alternative hypothesis) is represented by an exponential curve (Figure 1C), non-linear non-monotonic associations (alternative hypotheses) are represented by quadratic (Figure 1D) and sine (Figure 1E) functions, and non-linear non-monotonic associations (alternative hypotheses) are represented by the circumference (Figure 1F), cross (Figure 1G) and square (Figure 1H) shape relationships. Moreover, we also illustrate the case of local correlation (alternative hypothesis), i.e. when part of the data (20% of data points) is linearly correlated (represented by crosses at Figure 1I) and the rest is independent.

For each scenario described in Figure 1, 1000 repetitions were carried out for different numbers of observations. The number of observations analyzed in this study was $n = 10, 30, 50$ for independent, linear, exponential, quadratic, sine, circumference and cross associations. For square association, the numbers of observations was $n = 40, 140$. For local correlation, $n = 100$. For further details regarding the simulations, refer to supplementary material.

To evaluate the performance of the methods, a receiver operating characteristic (ROC) curve was constructed for each scenario and each number of observations, and the area under this curve was calculated. The ROC curve is useful in evaluating the power of the test. It consists in a bi-dimensional plot of one minus the specificity in the x-axis versus sensitivity in the y-axis, where specificity = number of true-negatives/(number of true-negatives + number of false-positives) and sensitivity = number of true-positives/(number of true-positives + number of false-negatives). In our case, the P-value (nominal level) is on the x-axis and the proportion of rejected null hypothesis, i.e. the proportion of associations identified between two random variables, on the y-axis. The area under the ROC curve is a quantitative summary of the power of the employed test and it varies from 0 to 1. In other words, an area close to 1 denotes high power, whereas an area below 0.50 means that the method is not able to identify dependence. An area close to 0.50 is equivalent to random decisions. To calculate the area under the ROC curve, we computed the Riemman sum with intervals of 0.001. Table 1 describes the areas under ROC curves.

By analyzing the number of falsely identified dependencies between independent random variables, notice that all the eight methods present areas under ROC curves close to 0.50. In other words, it means that all the eight methods indeed control the rate of false-positives under the null hypothesis (i.e. the frequency of falsely rejected null hypothesis is proportional to the P-value threshold), as expected. One may notice that Hoeffding's D measure presents an area under the ROC curve slightly greater than 0.50. It can be explained by

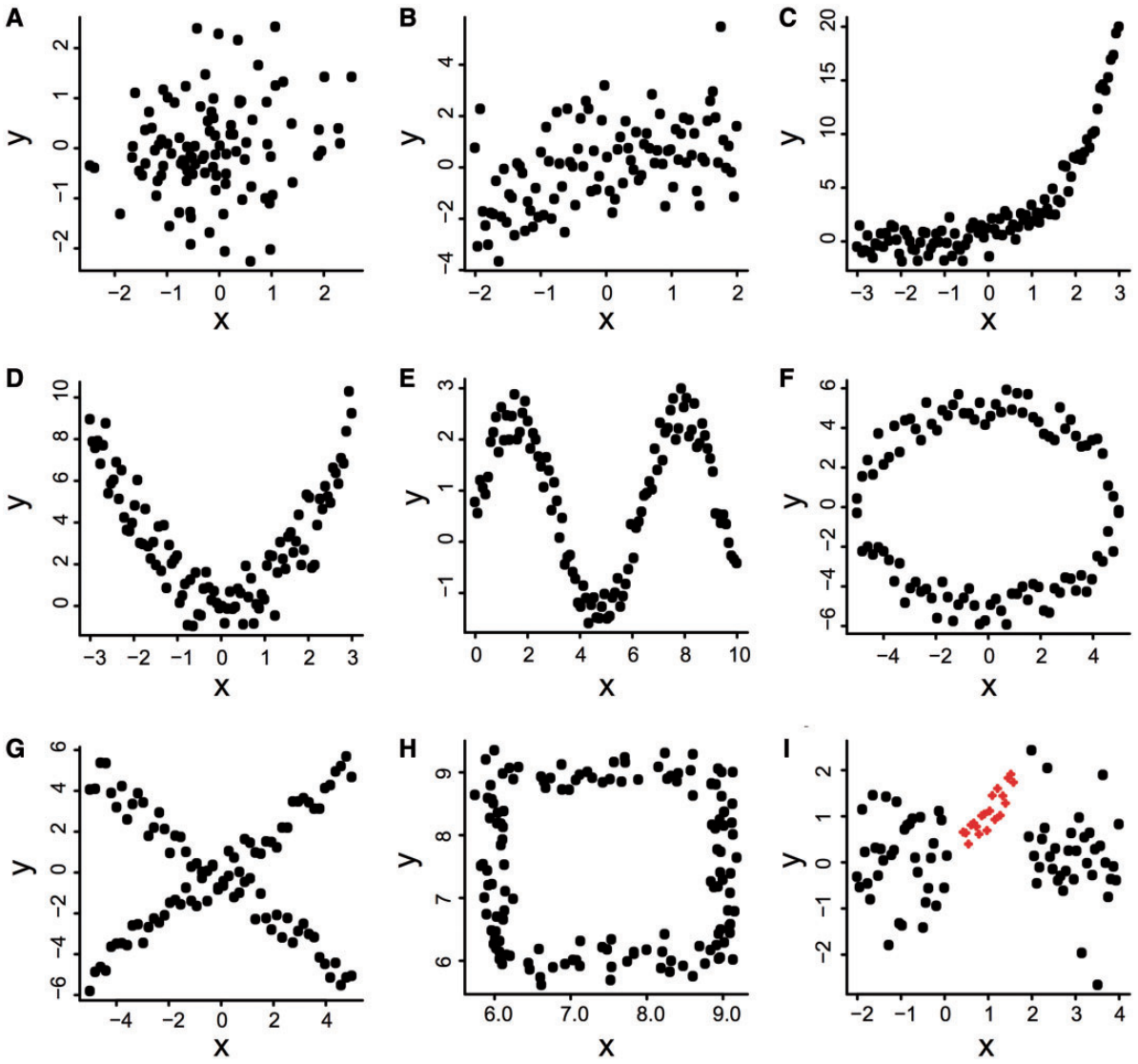

Figure 1: Simulations. (**A**) Independent data, (**B**) linear association, (**C**) exponential association - non-linear monotonic association, (**D**) quadratic association - non-linear non-monotonic, (**E**) sine association—non-linear non-monotonic, (**F**) circumference—non-functional association, (**G**) cross—non-functional association, (**H**) square—non-functional association and (**I**) local correlation—only part of the data is correlated, which is represented by crosses.

the fact that for *P*-values greater than 0.40, it overestimates the number of false-positives, as previously discussed [22]. However, as usually only *P*-values <0.05 are considered as statistically significant, it is not a cause of worry.

By analyzing the cases of relationships under the alternative hypothesis, the majority of the methods were shown to be consistent according to the number of observations. The greater the number of observations, the greater the areas under the ROC curves (the power of the test). Exceptions are the Pearson's, Spearman's and Kendall's correlations for non-linear non-monotonic (quadratic, sine) and also non-functional (circumference, cross and square) relationships. These results mean that independent of the number of observations, these methods are not able to identify these types of associations.

MI, MIC, HHG measure and Hoeffding's *D* measure are able to identify the majority of relationships studied here, including linear, non-linear monotonic/non-monotonic functions and also non-functional relationships (indicated by the areas under the ROC curves close to 1). Exception is the square association that was identified only by the HHG method. We note that distance correlation did not identify non-functional relationships in our

Table 1: Areas under ROC curves (2.5% quantile, mean and 97.5% quantile) obtained by applying each method (Pearson's correlation, Spearman's correlation, Kendall's correlation, distance correlation, HHG measure, Hoeffding's ***D*** measure, mutual information - MI and maximal information coefficient - MIC) on 12 different conditions and numbers of data points (***n***)

Type of association/ Method	*n*	Pearson			Dcor			Spearman			Kendall		
		2.50%	Mean	97.50%	2.50%	Mean	97.50%	2.50%	Mean	97.50%	2.50%	Mean	97.50%
Independent	10	0.49	0.5	0.52	0.49	0.51	0.53	0.48	0.5	0.51	0.44	0.46	0.48
	30	0.48	0.5	0.51	0.47	0.49	0.51	0.48	0.5	0.52	0.47	0.49	0.5
	50	0.49	0.51	0.53	0.49	0.5	0.52	0.49	0.51	0.53	0.5	0.51	0.53
Linear	10	0.81	0.82	0.84	0.79	0.8	0.82	0.78	0.8	0.82	0.76	0.77	0.79
	30	0.98	0.98	0.98	0.97	0.97	0.98	0.97	0.98	0.98	0.97	0.98	0.98
	50	1	1	1	0.99	0.99	1	0.99	1	1	0.99	1	1
Exponential	10	0.88	0.88	0.88	0.99	0.99	0.99	0.94	0.94	0.95	0.93	0.93	0.94
	30	0.96	0.96	0.96	1	1	1	1	1	1	1	1	1
	50	0.99	0.99	0.99	1	1	1	1	1	1	1	1	1
Quadratic	10	0.15	0.16	0.17	0.71	0.71	0.71	0.15	0.16	0.17	0.14	0.14	0.15
	30	0.19	0.2	0.21	0.99	0.99	0.99	0.16	0.17	0.18	0.2	0.21	0.22
	50	0.2	0.21	0.22	1	1	1	0.18	0.18	0.19	0.22	0.23	0.23
Sine	10	0.27	0.28	0.29	0.28	0.29	0.3	0.3	0.32	0.33	0.23	0.24	0.25
	30	0.34	0.35	0.37	0.88	0.88	0.89	0.37	0.38	0.39	0.34	0.35	0.37
	50	0.38	0.4	0.41	0.98	0.98	0.98	0.41	0.42	0.44	0.41	0.42	0.43
Circumference	10	0.09	0.09	0.1	0.1	0.1	0.11	0.23	0.24	0.24	0.19	0.2	0.2
	30	0.08	0.09	0.09	0.17	0.18	0.18	0.15	0.15	0.16	0.17	0.18	0.19
	50	0.09	0.09	0.09	0.38	0.38	0.39	0.15	0.15	0.16	0.17	0.18	0.19
Cross	10	0.08	0.09	0.09	0.02	0.03	0.03	0.13	0.14	0.15	0.1	0.11	0.11
	30	0.11	0.11	0.12	0.45	0.46	0.46	0.11	0.11	0.12	0.11	0.11	0.12
	50	0.11	0.12	0.12	0.77	0.77	0.78	0.11	0.11	0.11	0.11	0.11	0.12
Square	40	0.24	0.26	0.26	0.17	0.18	0.19	0.26	0.27	0.27	0.25	0.26	0.28
	140	0.24	0.25	0.26	0.44	0.45	0.46	0.25	0.26	0.27	0.24	0.25	0.26
Local correlation	100	0.28	0.29	0.3	1	1	1	0.41	0.43	0.44	0.39	0.41	0.42
Type of association/ Method	***n***	**Hoeffding**			**HHG**			**MI**			**MIC**		
		2.50%	**Mean**	**97.50%**	**2.50%**	**Mean**	**97.50%**	**2.50%**	**Mean**	**97.50%**	**2.50%**	**Mean**	**97.50%**
Independent	10	0.48	0.5	0.52	0.49	0.51	0.52	0.35	0.37	0.39	0.33	0.35	0.37
	30	0.53	0.54	0.56	0.47	0.49	0.51	0.47	0.48	0.5	0.48	0.5	0.52
	50	0.56	0.57	0.59	0.5	0.51	0.53	0.49	0.51	0.52	0.49	0.51	0.53
Linear	10	0.74	0.76	0.78	0.67	0.69	0.71	0.41	0.43	0.45	0.61	0.62	0.64
	30	0.96	0.97	0.97	0.89	0.9	0.91	0.67	0.68	0.7	0.86	0.87	0.88
	50	0.99	0.99	1	0.97	0.98	0.98	0.78	0.79	0.81	0.91	0.91	0.92
Exponential	10	0.97	0.97	0.98	0.99	0.99	0.99	0	0	0	0.71	0.72	0.74
	30	1	1	1	1	1	1	0.86	0.86	0.86	1	1	1
	50	1	1	1	1	1	1	0.9	0.9	0.9	1	1	1
Quadratic	10	0.89	0.9	0.91	0.96	0.97	0.97	0.53	0.54	0.56	0.51	0.52	0.52
	30	1	1	1	1	1	1	0.99	1	1	1	1	1
	50	1	1	1	1	1	1	1	1	1	1	1	1
Sine	10	0.49	0.51	0.52	0.32	0.34	0.35	0.31	0.33	0.34	0.17	0.19	0.2
	30	0.96	0.96	0.96	0.98	0.98	0.98	0.92	0.93	0.93	0.99	0.99	0.99
	50	0.99	0.99	1	1	1	1	1	1	1	1	1	1
Circumference	10	0.63	0.64	0.64	0.7	0.71	0.72	0.5	0.51	0.52	0.09	0.1	0.1
	30	0.87	0.88	0.88	0.96	0.96	0.96	0.72	0.74	0.76	0.69	0.71	0.72
	50	0.94	0.95	0.95	0.99	0.99	0.99	0.95	0.96	0.96	0.93	0.94	0.95
Cross	10	0	0	0	0.65	0.66	0.68	0.41	0.42	0.44	0.02	0.02	0.02
	30	0.42	0.42	0.43	1	1	1	0.96	0.96	0.97	0.57	0.57	0.58
	50	0.84	0.85	0.85	1	1	1	1	1	1	0.97	0.97	0.97
Square	40	0.26	0.27	0.28	0.9	0.9	0.91	0.31	0.33	0.34	0.36	0.38	0.39
	140	0.57	0.58	0.59	1	1	1	0.69	0.7	0.72	0.44	0.45	0.47
Local correlation	100	0.99	0.99	0.99	1	1	1	1	1	1	1	1	1

simulations using less than 100 and 140 data points (circumference and square scenarios, respectively). But increasing the number of data points to 1000 (results not shown), distance correlation was able to identify both dependences as predicted by the theory.

By analyzing linear and non-linear monotonic relationships (exponential), Pearson, Spearman, Kendall, Hoeffding and HHG presented similar performances, whereas methods based on information theory (MI and MIC) presented the lowest power. Although, in theory, Pearson's correlation identifies only linear relationships, its performance in identifying monotonic associations such as the exponential association is satisfactory. It occurs due to the fact that non-linear monotonic relationships can usually be adjusted well by linear functions.

The analyses of non-linear non-monotonic relationships (quadratic and sine) show that Hoeffding's *D* measure and HHG are the most powerful methods, followed by distance correlation, and then by MI and MIC. For non-functional relationships, HHG is the most powerful method, followed by Hoeffding's *D* measure, MI, MIC and distance correlation.

Illustrative biological example

To illustrate an application of the eight methods in retrieving relevant relationships among gene expression signals, we applied them to a data set composed of 168 DNA microarrays derived from stage I lung tumor samples. This data set [23,24] is freely available and can be downloaded from Gene Expression Omnibus - GEO (http://www.ncbi.nlm.nih.gov/geo/) with id GSE31210. We chose Wnt as our illustrative model gene because it is known to be highly associated with lung cancer and several pathways have already been described in the literature [25].

We selected 81 genes that are already known to belong to the Wnt pathway (alternative hypothesis - H_1) and 62 control probe sets of the microarray that should not have any association with Wnt (null hypothesis - H_0).

To study the performance of the methods in different sample sizes, different numbers of observations ($n = 12, 25, 50, 100, 168$) were considered to construct the ROC curves under both the null (H_0) and alternative hypotheses. In ROC curves constructed under H_1, the y-axis is the proportion of relationships identified between Wnt and the 81 genes already described in the literature as belonging to its pathway. Under H_0, the y-axis is the proportion of associations identified between Wnt and the 62 controls probe sets. For each varied number of observations n ($n = 12, 25, 50, 100, 168$), we sampled n microarrays and applied the statistical tests. This procedure was repeated 100 times to construct 100 ROC curves. The average areas under the 100 ROC curves (under both H_0 and H_1) are described in Table 2.

Notice that all the eight methods control the rate of false-positives under H_0 when the sample size is large (the areas under the ROC curves were ~0.50) (Table 2) in this biological example as well as in our simulations. This means that although some hypotheses of the statistical tests are eventually not valid in actual biological data (sometimes all the hypotheses required by the method cannot be checked), the tests are still controlling the type I error. We also observed that, for all the eight methods, the areas under the ROC curves under alternative hypothesis were >0.50 (Table 2), meaning that, in fact, it is possible to retrieve at least part of the regulatory network by using methods that identify dependence between random variables. Corroborating the results obtained by simulations, the powers of MI and MIC were lower than other methods and the decrease of the power of the methods was proportional to the decrease of the number of observations n.

To verify how much is the overlap of associations identified by the methods, the number of co-identified associations was counted. Table 3 shows the number of co-identified relationships between Wnt and 81 genes belonging to its pathway by different methods assuming different *P*-value thresholds.

Because biomedical researchers usually try to find linear and monotonic relationships and then more complex relationships, we also count how many findings each of the methods such as distance correlation, HHG, Hoeffding, MI and MIC were able to identify above the union of the findings of Pearson's, Spearman's and Kendall's correlations. These results are described in Table 4.

Notice that the quantity of overlaps is close to the total amount of significant dependence identified by each method. Moreover, the number of dependences identified only by methods that are able to identify more general relationships than monotonic associations is low (equal or less than four). These results suggest that for this data set and genes analyzed, the majority of relationships can be considered as linear.

Table 2: Application of the eight methods (Pearson's correlation, Spearman's correlation, Kendall's correlation, distance correlation, HHG measure, Hoeffding's ***D*** measure, mutual information - MI and maximal information coefficient - MIC) in an expression data set composed of 168 stage I lung tumors microarrays

	n	Pearson			Dcor			Spearman			Kendall		
		2.5%	Mean	97.5%	2.5%	Mean	97.5%	2.5%	Mean	97.5%	2.5%	Mean	97.5%
Under H_0	12	0.62	0.67	0.72	0.52	0.57	0.64	0.59	0.65	0.70	0.55	0.64	0.71
	25	0.51	0.58	0.65	0.59	0.65	0.72	0.58	0.66	0.74	0.62	0.68	0.74
	50	0.28	0.34	0.40	0.22	0.28	0.39	0.38	0.43	0.48	0.40	0.45	0.51
	100	0.39	0.45	0.52	0.33	0.39	0.44	0.34	0.41	0.47	0.35	0.41	0.48
	168	0.27	0.34	0.40	0.34	0.40	0.45	0.26	0.33	0.39	0.28	0.34	0.42
Under H_1	12	0.47	0.52	0.59	0.41	0.50	0.57	0.49	0.54	0.61	0.43	0.50	0.56
	25	0.53	0.60	0.66	0.57	0.63	0.69	0.51	0.59	0.65	0.51	0.57	0.64
	50	0.58	0.64	0.69	0.64	0.68	0.74	0.59	0.64	0.71	0.58	0.64	0.72
	100	0.56	0.63	0.69	0.62	0.66	0.72	0.55	0.61	0.66	0.54	0.60	0.67
	168	0.59	0.66	0.73	0.62	0.68	0.75	0.60	0.67	0.74	0.59	0.66	0.72
	n	Hoeffding			HHG			MI			MIC		
		2.5%	Mean	97.5%	2.5%	Mean	97.5%	2.5%	Mean	97.5%	2.5%	Mean	97.5%
Under H_0	12	0.49	0.59	0.66	0.36	0.41	0.47	0.24	0.30	0.37	0.34	0.40	0.46
	25	0.59	0.66	0.71	0.45	0.52	0.58	0.48	0.55	0.62	0.53	0.60	0.66
	50	0.33	0.39	0.45	0.26	0.32	0.39	0.37	0.43	0.48	0.45	0.50	0.55
	100	0.42	0.47	0.51	0.32	0.38	0.44	0.39	0.46	0.53	0.32	0.39	0.45
	168	0.48	0.52	0.55	0.39	0.44	0.51	0.40	0.48	0.55	0.42	0.49	0.58
Under H_1	12	0.45	0.52	0.60	0.42	0.49	0.55	0.29	0.36	0.42	0.32	0.39	0.45
	25	0.57	0.62	0.68	0.50	0.56	0.63	0.53	0.58	0.63	0.53	0.59	0.65
	50	0.64	0.70	0.76	0.54	0.61	0.67	0.47	0.52	0.58	0.51	0.58	0.64
	100	0.62	0.67	0.72	0.55	0.61	0.67	0.49	0.56	0.62	0.52	0.57	0.63
	168	0.66	0.72	0.78	0.57	0.63	0.68	0.53	0.59	0.64	0.50	0.57	0.63

The values of the average, 2.5 and 97.5% quantiles under the ROC curves were calculated in 100 repetitions and in a varied number of observations ($n = 12, 25, 50, 100, 168$). Under H_0: associations obtained by applying the methods on expression data of control probe sets (that do not present any association with the Wnt gene); Under H_1: associations obtained by applying the methods on expression data of genes belonging to Wnt pathway (which are already known to be associated with Wnt).

FINAL REMARKS

The use of each method depends essentially on the type of data or relationship one wants to identify and the number of observations. A summary of the method to be used depending on the characteristics of the data set is illustrated in a decision tree in Figure 2. We considered $n \geq 30$ and $n < 30$ as large and small data sets, respectively. This threshold was chosen based on the simulations results that showed a high accuracy for all the methods when $n \geq 30$. It is necessary to point out that this threshold may vary depending on data variance. For large data sets ($n \geq 30$), methods such as distance correlation, Hoeffding's *D* measure, HHG, MI and MIC could be more interesting because they identify broad types of relationships. Methods to identify non-functional relationships (Hoeffding's *D* measure, HHG, MI and MIC) are interesting in a theoretical point of view; however, in the analysis of gene expression signal, non-functional relationships are difficult to interpret and usually are ignored. On the other hand, local correlations are interesting in a biological point of view (one gene may be associated with another only in a specific expression range) but are usually ignored too. For small data sets ($n < 30$), HHG is recommended if one is interested in identifying non-functional relationships.

For the identification of non-linear and non-monotonic associations, Hoeffding's *D* measure and HHG are recommended.

If hypothesis of linearity or monotonicity can be assumed, the application of Spearman's or Kendall's correlations may be more useful than Pearson's correlation because they identify both linear and non-linear monotonic relationships with high power (even when the relationship is linear, the power is similar to Pearson's correlation).

Table 3: Number of co-identified relationships between Wnt and 81 genes belonging to its pathway by different methods and assuming different *P*-value thresholds

	P-value threshold	Pearson	Dcor	Spearman	Kendall	Hoeffding	HHG	MI	MIC
Pearson	0.01	12	8	10	10	10	4	3	2
	0.05	19	18	16	16	14	8	4	6
	0.1	28	25	23	23	23	14	10	9
Dcor	0.01		11	11	11	10	5	4	1
	0.05		22	17	17	16	11	5	6
	0.1		32	25	25	28	17	10	12
Spearman	0.01			15	15	14	5	4	2
	0.05			24	24	21	10	7	6
	0.1			30	30	27	14	11	11
Kendall	0.01				15	14	5	4	2
	0.05				24	21	10	7	6
	0.1				30	27	14	11	11
Hoeffding	0.01					14	5	4	2
	0.05					22	11	5	6
	0.1					31	18	10	12
HHG	0.01						7	4	1
	0.05						15	5	3
	0.1						20	6	7
MI	0.01							4	1
	0.05							8	1
	0.1							15	5
MIC	0.01								3
	0.05								8
	0.1								14

In gray are highlighted the total number of relationships identified by the respective method.

Table 4: Number of findings each of the methods such as distance correlation, HHG, Hoeffding, MI and MIC was able to identify over and above the union of the findings of Pearson's, Spearman's and Kendall's correlations assuming different *P*-value thresholds

P-value threshold	Dcor	Hoeffding	HHG	MI	MIC
0.01	0	0	2	0	1
0.05	2	1	4	1	2
0.10	3	3	4	3	3

It is important to clarify that only distance correlation, Hoeffding's *D* measure, HHG and MI have mathematically proven consistency against all alternatives (theoretically, they asymptotically can detect all situations of deviation from independence). Distance correlation did not present enough power to identify non-functional associations in our simulations, but by increasing the data points to 1000, it correctly identified them (data not shown) as predicted by theory. Methods that are applicable in multivariate scenarios are distance correlation and HHG [9,11].

Another point to be discussed is the relationship between correlation and causality. It is important to mention that correlation does not imply causation. In other words, a correlation between two random variables does not necessarily imply that one causes the other. The classic example is when there are two variables A and B, and one more unobserved variable C that causes A and B. In this example, by applying an independence test, one may conclude that A and B are correlated, but in fact, there is no causal influence between them.

Although correlation does not imply causation, correlation can be used as a hint to identify causality between random variables. For example, suppose two time series A and B. The identification of a correlation between past values of A and future values of B may be an indication that A causes B (due to our intuitive concept that the cause never occurs after its effect). This type of correlation between lagged time series is known as Granger causality [26]. However, even Granger causality is not causality in a deep sense of the word because it is also based only on numeric predictions. So, how one can establish causality? This is a challenging problem,

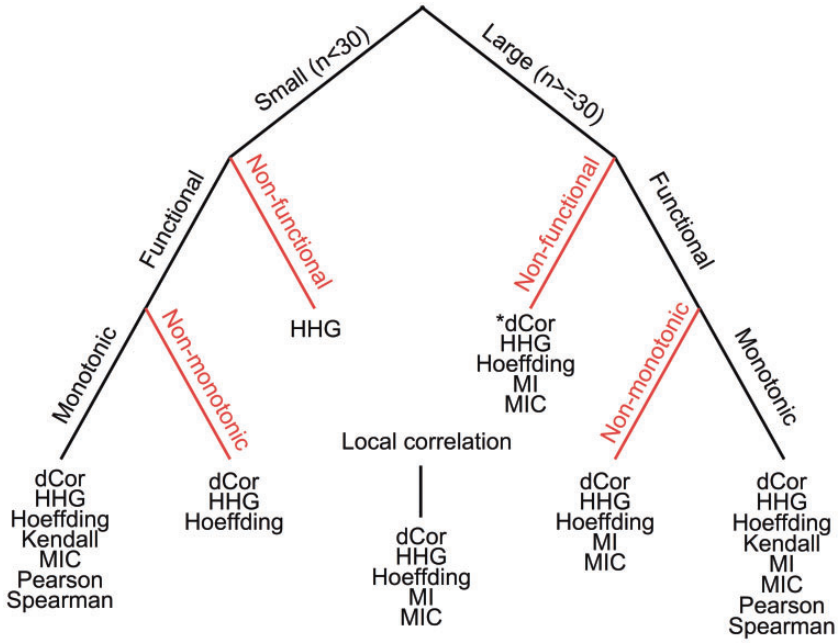

Figure 2: Decision tree. Summary of the relationships each method can identify depending on the number of observations and type of relationship. The order of the methods presented within each leaf is arbitrary. *Distance correlation detects some non-functional dependencies in only very large sample size.

especially in molecular biology. The most effective way to identify causality is through a well-controlled experiment. For example, two groups of cells whose are comparable in almost every way are submitted to different conditions. If the two groups of cells have statistically different outcomes, then the different condition may have caused the different outcome.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Uncorrelated random variables do not mean independency.
- Understanding the different measures to identify dependent genes is crucial to model gene regulatory networks.
- The choice of a suitable method to identify dependency between random variables depends on several constraints (sample size and type of dependency) and goals.

ACKNOWLEDGEMENTS

The authors thank the authors of the HHG measure, Yair Heller, Ruth Heller and Malka Gorfine, for kindly providing the R code.

FUNDING

A.F. was supported by FAPESP grants 11/07762-8, 11/50761-2, 13/03447-6, and CNPq306319/2010-1. S.S.S. was supported by FAPESP grant 12/25417-9. D.Y.T. was partially supported by Pew Latin America fellowship. A.N. was partially supported by JSPS KAKENHI 25830111.

References

1. Bhardwaj N, Lu H. Correlation between gene expression profiles and protein-protein interactions within and across genomes. *Bioinformatics* 2005;**21**:2730–8.
2. Bansal M, Belcastro V, Ambesi-Impiombato A, *et al*. How to infer gene networks from expression profiles. *Mol Syst Biol* 2007;**3**:78.
3. De la Fuente A, Bing N, Hoeschele I, *et al*. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 2004;**20**:3565–74.
4. Steuer R, Kurths J, Daub CO, *et al*. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* 2002;**18**:S231–40.
5. Galton F. Co-relations and their measurement, chiefly from anthropometric data. *Proc R Soc London* 1888;**45**:135–45.
6. Pearson K. Notes on the history of correlation. *Biometrika* 1920;**13**:25–45.
7. Spearman C. 'General intelligence', objectively determined and measured. *Am J Psychol* 1904;**15**:201–92.

8. Kendall M. A new measure of rank correlation. *Biometrika* 1938;**30**:81–9.
9. Szekely G, Rizzo M, Bakirov N. Measuring and testing independence by correlation of distances. *Ann Stat* 2007; **35**:2769–94.
10. Hoeffding W. A non-parametric test of independence. *Ann Math Stat* 1948;**19**:546–57.
11. Heller Y, Heller R, Gorfine M. A consistent multivariate test of association based on ranks of distances. *Biometrika* 2012. doi:10.1093/biomet/ass070 (Advance Access publication 4 December 2012).
12. Shannon CE, Weaver W. *The Mathematical Theory of communication*. Urbana, IL: University of Illinois Press, 1949.
13. Reshef DN, Reshef YA, Finucane HK, *et al*. Detecting novel associations in large data sets. *Science* 2011;**334**: 1518–24.
14. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.
15. Paninski L. Estimation of entropy and mutual information. *Neural Computation* 2003;**15**:1191–253.
16. Daub CO, Steuer R, Selbig J, *et al*. Estimating mutual information using B-spline functions—an improved similarity measure for analyzing gene expression data. *BMC Bioinformatics* 2004;**5**:118.
17. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. *Phys Rev E* 2004;**69**:066138.
18. Nadaraya EA. On estimating regression. *Theory Probab Appl* 1964;**10**:186–90.
19. Moon YI, Balaji R, Lall U. Estimation of mutual information using kernel density estimators. *Phys Rev E* 1995;**52**: 2318–21.
20. Khan S, Bandyopadhyay S, Ganguly AR, *et al*. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys Rev E* 2007;**76**:026209.
21. Cellucci CJ, Albano AM, Rapp PE. Statistical validation of mutual information calculations: comparison of alternative numerical algorithms. *Phys Rev E* 2005;**71**:066208.
22. Fujita A, Sato JR, Demasi MA, *et al*. Comparing Pearson, Spearman and Hoeffding's D measure for gene expression association analysis. *J Bioinform Comput Biol* 2009;**7**:663–84.
23. Okayama H, Kohno T, Ishii Y, *et al*. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res* 2012;**72**:100–11.
24. Yamauchi M, Yamaguchi R, Nakata A, *et al*. Epidermal growth factor receptor tyrosine kinase defines critical prognostic genes of stage I lung adenocarcinoma. *PLoS One* 2012;**7**:e43923.
25. Mazieres J, He B, You L, *et al*. Wnt signaling in lung cancer. *Cancer Lett* 2005;**222**:1–10.
26. Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 1969;**37**: 424–38.