Published in final edited form as:

Med Phys Mex Symp Med Phys. 2006 September 8; 854(1): 25–30. doi:10.1063/1.2356392.

Using Mutual Information to Discover Temporal Patterns in Gene Expression Data

Sergei Chumakov¹, Efren Ballesteros¹, Jorge E. Rodriguez Sanchez¹, Arturo Chavez¹, Meizhuo Zhang², B. Montgomery Pettit³, and Yuriy Fofanov^{2,4}

¹ Department of Physics, University of Guadalajara, Guadalajara, Jalisco 44420, Mexico

² Department of Computer Science, University of Houston, Houston, TX 77204-3058, USA

³ Department of Chemistry, University of Houston, Houston, TX 77204-5001, USA

⁴ Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5001, USA

Abstract

Finding relations among gene expressions involves the definition of the similarity between experimental data. A simplest similarity measure is the Correlation Coefficient. It is able to identify linear dependences only; moreover, is sensitive to experimental errors. An alternative measure, the Shannon Mutual Information (MI), is free from the above mentioned weaknesses. However, the calculation of MI for continuous variables from the finite number of experimental points, N , involves an ambiguity arising when one divides the range of values of the continuous variable into boxes. Then the distribution of experimental points among the boxes (and, therefore, MI) depends on the box size. An algorithm for the calculation of MI for continuous variables is proposed. We find the optimum box sizes for a given N from the condition of minimum entropy variation with respect to the change of the box sizes. We have applied this technique to the gene expression dataset from Stanford, containing microarray data at 18 time points from yeast *Saccharomyces cerevisiae* cultures (Spellman et al., [3]). We calculated MI for all of the pairs of time points. The MI analysis allowed us to identify time patterns related to different biological processes in the cell.

Keywords

Gene expression; Mutual information

INTRODUCTION

An important problem in bioinformatics is that of finding relations among experimental data. A typical task is to find pairs of variables in an experimental dataset, which depend on each other, usually in the presence of a strong noise. A common way to characterize relations among experimental data is the correlation coefficient (CC). Unfortunately, it is designed to reveal linear dependence and fails to identify dependences which are significantly non-linear. Also, it is sensitive to outliers, *i.e.* points located strongly apart of the majority of the points. These exceptional points are often experimental errors and they can dramatically affect CC values. A well-known alternative to the CC, the Shannon mutual information (MI) [1], is free from these shortcomings. MI is widely used to describe dependences among discrete variables (see, *e.g.* [2]). However, the application of MI to continuous variables is less straightforward, due to some ambiguities which arise here. The problem is that the number of experimental points is always finite, $N < \infty$. Hence, one is forced to divide the range of values of the continuous variable into bins or boxes and to introduce the auxiliary discrete variable N_k equal to the

number of points in every box. The resulting distribution depends on additional parameters: number of experimental points N and the box sizes. This dependence strongly affects MI.

Here, we seek a hopefully unique solution to the problem, of how, for a given value of N , to find the best values of the box sizes. Our motivation is to apply the potentially powerful tool of MI to the analysis of gene expression data. We confine our discussion by using the publicly available RNA expression dataset from Stanford, containing microarray data at 18 time points from yeast *Saccharomyces cerevisiae* cultures synchronized by α factor arrest (Spellman et al., [3]). This dataset has been thoroughly investigated and clusters of genes that represent similar behavior were found, using CC to measure similarity in gene expression patterns (Eisen et al., [4]). More general dependencies including time shifts and causal relationships (Local Invariants) for this dataset have been found (Fofanov et al., [5]). Clustering using the mutual information instead of the correlation coefficient to measure similarity among genes behavior has been also performed (Butte and Kohane, [6]). In section 2, we propose a method of defining the optimal box size for calculation of MI. In section 3, we apply this technique to the yeast dataset. We calculate the mutual information for expression values for all pairs of time points and discuss the time evolution of gene expression. The MI analysis allows us to identify three different time patterns: periodic behavior due to the cell division cycle, descending behavior due to the response to the initial synchronization (e.g. α factor arrest) and stationary behavior due to the group of genes not involved directly into the cell division cycle or response to initial synchronization.

MUTUAL INFORMATION AND BOX SIZES

We start with the one-dimensional case, recalling the definitions of Shannon entropy for the discrete and continuous random variable. Let K be a *discrete random variable* and k , its value. We have N experimental points and N_k of them correspond to the value k . The probability of k and entropy the probability distribution are defined as

$$P_k = \lim_{N \rightarrow \infty} \frac{N_k}{N}, \quad S_k = - \sum_k P_k \log P_k. \quad (1)$$

For the case of a *continuous random variable*, X , which takes the values x , we introduce an auxiliary discrete variable K - the number of a box Δ_k of size Δ to which a given value of X belongs: $K=k$, if $k\Delta \leq x < (k+1)\Delta$. Then, N_k is the number of experimental points in the box k . Probability distributions P_k and $P(x)$ are connected as follows,

$$P_k = \int_{k\Delta}^{(k+1)\Delta} dx P(x) = P(\bar{x})\Delta, \quad \bar{x} \in \Delta_k, \quad \sum_k P(\bar{x})\Delta = \int dx P(x) = 1. \quad (2)$$

$P(\bar{x})$ goes to $P(x)$ in the limit $\Delta \rightarrow 0$. (Note, that first the limit $N \rightarrow \infty$ and second, $\Delta \rightarrow 0$ have to be taken). We define the entropy of a continuous variable X as

$$S_x = - \int dx P(x) \log P(x) = \lim_{\Delta \rightarrow 0} \lim_{N \rightarrow \infty} (S_k + \log \Delta), \text{ i.e.,}$$

$$S_x = \lim_{\Delta \rightarrow 0} \lim_{N \rightarrow \infty} S_x(N, \Delta), \quad S_x(N, \Delta) = \log N + \log \Delta - \frac{1}{N} \sum_k N_k \log N_k. \quad (3)$$

In the case of two random variables x and y , with a joint distribution,

$$P(x, y) = \lim_{\Delta \rightarrow 0} \lim_{N \rightarrow \infty} \frac{N_{kj}}{N\Delta},$$

where N_{kj} is a number of experimental points in a box of size $\Delta = \Delta_x \times \Delta_y$,

$$(n\Delta_x \leq x < (n+1)\Delta_x) \times (j\Delta_y \leq y < (j+1)\Delta_y),$$

joint entropy may be defined as

$$S_{xy}(N, \Delta_x, \Delta_y) = \log N + \log(\Delta_x \Delta_y) - \frac{1}{N} \sum_{kj} N_{kj} \log N_{kj}. \quad (4)$$

Mutual information is defined as follows,

$$E(N, \Delta) = S_x + S_y - S_{xy},$$

where marginal entropies S_x and S_y have the form (3) (with $N_k = \sum_j N_{kj}$, $N_j = \sum_k N_{kj}$). Note, that we have from experiments samples of a random variable, and we need to work with finite values of N and Δ .

To choose the best box size for given N , we suggest that changing twice the sizes

$\Delta'_x = \Delta_x/2$, $\Delta'_y = \Delta_y/2$, of the rectangular box which contains the point (x, y) does not significantly affect the value of $P(x, y)$. We calculate the corresponding variation of the entropy, $\delta(N, \Delta) = S(N, \Delta') - S(N, \Delta)$. Let us denote $f(x) = x \log x$. It is clear that $f(x+y) > f(x) + f(y)$, and moreover,

$$f\left(\sum_{i=1}^4 x_i\right) - \sum_{i=1}^4 f(x_i) \approx c 4 \log 4 - \frac{1}{2c} \sum_{i=1}^4 \delta_i^2$$

where $c = \frac{1}{4} \sum_{i=1}^4 x_i$ and $x_i = c + \delta_i$. Using this equation to calculate the change of the entropy and reducing rectangular boxes from $\Delta = \Delta_x \Delta_y$ to $\Delta/4 = (\Delta_x/2)(\Delta_y/2)$ we have

$$S(N, \Delta) - S(N, \Delta') = \frac{2}{N} \sum_{kj} \frac{\delta_1^2 + \delta_2^2 + \delta_3^2 + \delta_4^2}{N_{kj}} = \delta(N, \Delta_x, \Delta_y), \quad (5)$$

where there are N_{kj} experimental points inside the box (kj) and they are distributed among sub-boxes as follows, $N_{i,kj} = N_{kj}/4 + \delta_i(k, j)$, $i = 1, 2, 3, 4$.

It is natural to choose the box sizes which minimize the entropy variation $\delta(N, \Delta_x, \Delta_y)$. This leads to a simple and fast algorithm, which is in a very good correspondence with exactly solvable examples, such as normal or exponential distributions.

APPLICATION TO YEAST DATA

Now, we will apply this technique to the analysis of time dependent gene expression data from yeast *Saccharomyces cerevisiae* cultures. Yeast time-series microarray experiments were performed by Cho (Cho *et al.*, [7]) and Spellman (Spellman *et al.*, [3]) using several methods of synchronization: α factor arrest, elutriation, *cdc15* temperature sensitive mutant, and *cdc28*. There are many online sources from which these datasets can be downloaded. A common dataset contains 2467 genes. We briefly recall the experimental procedures from [3] for the α factor arrest. The α factor was added to asynchronous yeast culture for 120 min. The α factor was then removed, and the arrested cells were placed in fresh medium. Samples were taken every 7 min for 119 min. The synchronization experiments produced cell cycle synchrony through two cycles, as verified by bud count, FACS analysis and/or nuclear staining. Evidence for cell cycle count was not published for elutriation and α factor arrest, but has been performed and supports the cell cycle count (personal communication).

Here we restrict ourselves with the results for α factor arrest dataset. We calculated the mutual information for all of the pairs time points (*i.e.*, for all the pairs of microarray chips). Consider the left graph in Fig 1. Ideally, one would expect high dependence (correlation) between one time point and its immediate neighbor, when “ $\Delta t = 1$ ”. Dependence between a time point and another point two time-intervals down, $\Delta t = 2$, would expectedly be lower. Plotting MI versus Δt would produce a typical relaxation curve. However, since the α -experiments span approximately two cell cycles, and we expect the period of genes involved in cell cycle regulation to coincide with the length of the cell cycle, we see a second, but lower peak (hump) around $\Delta t = 9$. This time interval ($9 \times 7 \text{ min} = 63 \text{ min}$) indicates the period of one cell cycle, which is approximately half of the duration of the experiment (119 min). This reaffirms the authors’ claim that α factor experiment spanned two cell cycles. The correlation coefficient for the same data is shown on the right graph of Fig. 1.

Both, MI and CC graphs reveal the periodic behavior due to the cell cycle. It is interesting to note, that the correlation graphs of time point $t=1$ and time point $t=2$ drop dramatically versus Δt , and show virtually no correlation after $\Delta t=3$, whereas the rest of time points show a periodic trend. On the other hand, the MI graph for the point $t=1$ with the rest of time points, $t=2 \div 18$ does reveal the periodic behavior whereas the MI for the second time point $t=2$ with the rest of time points, $t=3 \div 18$ shows no recurrence at the cycle period.

The analysis of scatter plots of one time point against another shows that nonlinear dependences cannot play a significant role for this data. Thus, the difference between the MI and CC graphs is caused by outliers. Note, that there are two types of outliers here. Several genes (numbered as 1770, 183, 462, 1771, and 2305) reveal very high expression level (>5), many genes show the moderately high expression ($2 \div 4$), still the most of the genes are expressed at lower level, (<2). The exclusion of the five most highly expressed genes (Fig. 1) does not change qualitatively the figures. However, the numerous outliers from the second group destroy the recurrence of correlation between the first time point and the rest of the points at the cycle period time. This does not affect the MI, which is determined by the total number of genes expressed at a lower level.

Acknowledgments

The authors acknowledge the support of the Texas Learning and Computation Center. BMP acknowledges support from NIH. SCh is grateful to the University of Houston for hospitality.

References

1. Shannon CE. The Bell System Technical Journal 1948;27:379–423. 623–656. *ibid*.
2. Herzel H, Grosse I. Phys Rev E 1997;55:800–809.
3. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Molecular Biology of the Cell 1998;9:3273–3297. [PubMed: 9843569]
4. Eisen MB, Spellman PT, Brown PO, Botstein D. Proc Natl Acad Sci USA 1998;95(25):14863–14868. [PubMed: 9843981]
5. Fofanov YV, Montgomery Pettitt B. Journal of Biomedical Informatics 2002;35:343–351. (2003). [PubMed: 12968783]
6. Butte AJ, Kohane IS. Proc Pacific Symposium on Biocomputing 2000;5:415–426.
7. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. Molecular Cell 1998;2:65–73. [PubMed: 9702192]

FIGURE 1.

A pairwise Mutual Information (left) and Correlation Coefficient (right) as functions of the time shift, Δt , for the yeast dataset with α factor arrest. Five strongest outliers were removed. Every red point represents the pair of time points. The pink line marks the similarity in gene expression at the first time point with the rest of the points; the green line – the second time point with the rest of the points; the blue line – average over all the pairs of time points with a given value of Δt .