LU TP 00-18
May 18, 2000
Revised March 2001

Shuffling Yeast Gene Expression Data

Sven Bilke[1]

Complex Systems Division, Department of Theoretical Physics
University of Lund, Sölvegatan 14A, S-22362 Lund, Sweden
http://www.thep.lu.se/complex/

Abstract

A new method to sort gene expression patterns into functional groups is presented. The method is based on a sorting algorithm using a non-local similarity score, which takes all other patterns in the data-set into account. Therefore the procedure is robust with respect to noise. Using the expression data for yeast, we extract information about functional groups. Without prior knowledge of parameters the cell cycle regulated genes in yeast can be identified. Furthermore a second, independent cell clock is identified. The capability of the algorithm to extract information about signal flow in the regulatory network underlying the expression patterns is demonstrated.

[1]eMail: sven@thep.lu.se

1 Introduction

DNA microarray technology [1] has greatly facilitated the study of gene expressions. With a single microarray, the expression of thousands of genes can be measured simultaneously. Based on the central dogma it is reasonable to understand these expression vectors as a description of the functional state of the cell. The dynamics of the state-trajectory observed in expression time series reveals much information about the regulatory network underlying gene expression. A detailed knowledge of this network would allow for the analysis of possible states and trajectories, including, say, transitions from disease to a healthy state.

Cluster algorithms have been used successfully in the analysis of expression data. Using for example hierarchical clustering it has been demonstrated [2] that many genes, which on biological grounds are known to be related, are located nearby in the similarity tree. It is however difficult to identify genes which belong to a larger functional context, like for example cell cycle regulated genes. If two of the corresponding patterns are expressed with a phase difference close to $\pi/4$, they are uncorrelated and therefore placed on remote sites in the similarity tree. The prior knowledge of the cell cycle frequency ν_{cc} was used to identify the cell cycle regulated genes by inspection [3]. In [4] a spectral filter was used for this purpose. Expression patterns, for which the spectral energy at frequency $\nu \approx \nu_{cc}$ is larger than some threshold, were selected as cell cycle regulated.

In this work a new method, re-shuffling, is used to give a global viewpoint on the expression data. It allows to identify sets of genes which belong into a larger functional context, even if expression patterns are mutually uncorrelated. The algorithm sorts the data based on a global similarity score, which makes it robust with respect to noise. The method does not aim to distribute the data into clusters, because in many cases cluster boundaries are artificial. Instead the structure found in the data is reflected in a re-ordered sequence of expression patterns.

Using this algorithm we are able to identify the cell cycle regulated genes in the budding yeast *S. cerevisae* without referring to prior knowledge. Furthermore we find a second, independent clock in the cell. The reordered sequence of expression patterns reflects the propagation of a signal in the data, *i.e.* Patterns, responding to the same, however, deformed signal are grouped together.

2 Algorithm

Re-shuffling is in spirit similar to the frequently used self-organizing maps, in that it provides a (one dimensional) map reflecting properties of the data analyzed. In contrast to SOM's, which map P different patterns to localized activity patterns of N nodes, the re-shuffling algorithm maps P patterns to P different positions in a one dimensional sequence. In the output, the distance in position of two patterns on this list reflects the similarity of patterns. In other words, in this list similar patterns are closer than less similar patterns.

To describe the algorithm lets start with a matrix C_{ij} encoding the similarity between expression patterns i and j. This similarity can, for example, be the mutual information or correlation for the two patterns. The purpose of the algorithm is to find a relabeling $i \to \sigma(i)$ such that similar patterns i, j, i.e. $|C_{ij}| \approx 1$, get similar labels: the distance $|\sigma(i) - \sigma(j)|$ is close to one. This is, however, not achieved by performing local, mutual, comparisons, but rather by letting expression patterns move freely in a "force-field" generated by all other patterns. The field is is described by the cost function

$$S_\sigma(\alpha, \gamma, \lambda) = -\sum_{i,j} \mathrm{sgn}(C_{ij})^\gamma |C_{ij}|^\alpha \exp -\frac{d(\sigma(i), \sigma(j))^2}{N\lambda}. \qquad (1)$$

The optimal sorting σ in the sense described above is the one minimizing (1). The parameter α controls the importance of the similarity in the sorting procedure. It turns out that this parameter can be changed in a wide range without a strong effect on the results, we therefore use $\alpha = 2$. The variable λ is a localization parameter. For small λ, mainly similarities close in index space contribute to the energy. This leads to a local optimization. For large λ a more global optimization is achieved. The parameter γ is used to switch between maintaining ($\gamma = 1$) or ignoring ($\gamma = 0$) the sign of C_{ij}.

For an open, non-cyclic list, the average distance $\overline{d}$ of indices from all other indices is not evenly distributed. It reaches its maximum on the border $i = 1, N$. For our purpose this non-flat distribution is not desired, therefore we use a cyclic distance measure. With

$$d(i,j) = \begin{cases} |i-j| & \text{if } |i-j| < N/2 \\ ||i-j| - N| & \text{if } |i-j| \geq N/2 \end{cases}, \qquad (2)$$

the first and the last pattern in the list are direct neighbors, the system has no boundary and therefore the $\overline{d}$ distribution is flat.

Unless C_{ij} has a very simple form, it is a non-trivial task to find the optimal sorting σ_0, for which equation (1) is minimal. We use simulated annealing [5] for this purpose. In this method, borrowed from statistical mechanics, a fictitious temperature parameter T is lowered. At the beginning, at high temperature, the global aspects of the structure contained in the data should be built into the order of expression patterns, while towards the end of the annealing procedure, at low temperature, the more local optimization takes place. Therefore the localization parameter λ is lowered together with the temperature from typically $\lambda = 1$ to $\lambda = 0.05$ in this procedure.

The time needed to perform the annealing procedure grows approximately quadratically with the number N of patterns: the time needed to calculate the change of the cost function in a single attempted exchange operation grows linearly with the volume. At the same time the number of attempts used in the simulation also grows linear with N. The time needed to perform reshuffling used to prepare figure 1 was 3 minutes on a 600 MHz Athlon processor.

3 Results

For this work we used the expression data for *S. cerevisae*, which is available on the internet [2] and described in [4]. Here we want to demonstrate the feasibility of our algorithm on a subset of this data set.

The original data consists of 82 experiments, which were done at different time points and/or boundary conditions. Each experiment provides measurements of 6177 expression ratios. The subset used here was extracted in the following way:

1. Experiments with more than 400 missing expression ratios were removed.

2. Expression patterns (in gene direction) with more than 8 missing ratios were removed.

3. From the remaining genes measured in $\tau = 69$ experiments we kept those $N = 803$ patterns with variance $\sigma > 0.5$.

[2] http://cellcycle-www.standford.edu

In the following the Pearson correlation

$$C_{ij} = \frac{\sum_t^\tau (D_{i,t} - \overline{D_i})(D_{j,t} - \overline{D_j})}{\tau \sqrt{\mathrm{Var}(D_i)\mathrm{Var}(D_j)}} \tag{3}$$

is used as the similarity score in equation (1). First the absolute value $|C_{ij}|$ is used ($\gamma = 0$), because we are interested in analyzing functional groups of genes, which show up by (anti)-correlated expression patterns. The result of the sorting procedure is visualized in figure 1, a graphical representation of the correlation matrix. In this diagram the intensity of the pixel at coordinate (i, j) is proportional to the absolute value $|C_{ij}|$. Red color represents correlation, green color anti-correlation.

At coordinates adjacent to the main diagonal of the matrix one clearly observes a grouping of gene expression patterns. With the help of the annotations for the genes involved one can verify that the grouping reflects a classification with respect to gene function. To this end, we plot in the lower part the integrated distribution of gene associations provided by the Gene Ontology Consortium [8]). More precisely,

$$S(x) = \frac{\sum_{i=0}^{x} f(i)}{\sum_{i=0}^{N} f(i)} \quad \text{with } f(i) = \begin{cases} 1 & \text{if gene } i \text{ is in the class} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

is shown where N is the number of genes used and x parametrizes the position on the horizontal axis. For a random pattern one expects a diagonal line from the lower left to the upper right corner A significantly different slope indicates a non-random distribution. In the extreme cases, a horizontal line indicates that *no* genes in this part of the sequence are part of the class, while a (almost) vertical line indicates that *all* genes belong to the class. More generally, a slope significantly larger than the one for random sequences indicates a region in which genes assigned to the chosen category are clustered. For the data presented here, one can not expect to observe the extreme case of a vertical line, because approximately only every third ORF used in this experiment has an annotation in the gene ontology. Nevertheless the curves show essentially flat and steep regions indicating a high degree of order in the resorted list.

The correlation plot figure (1) does not only contain information about the dominant correlation, which clusters the genes into groups. Sub-dominant co-regulations can be seen in the more off-diagonal parts of the matrix. As

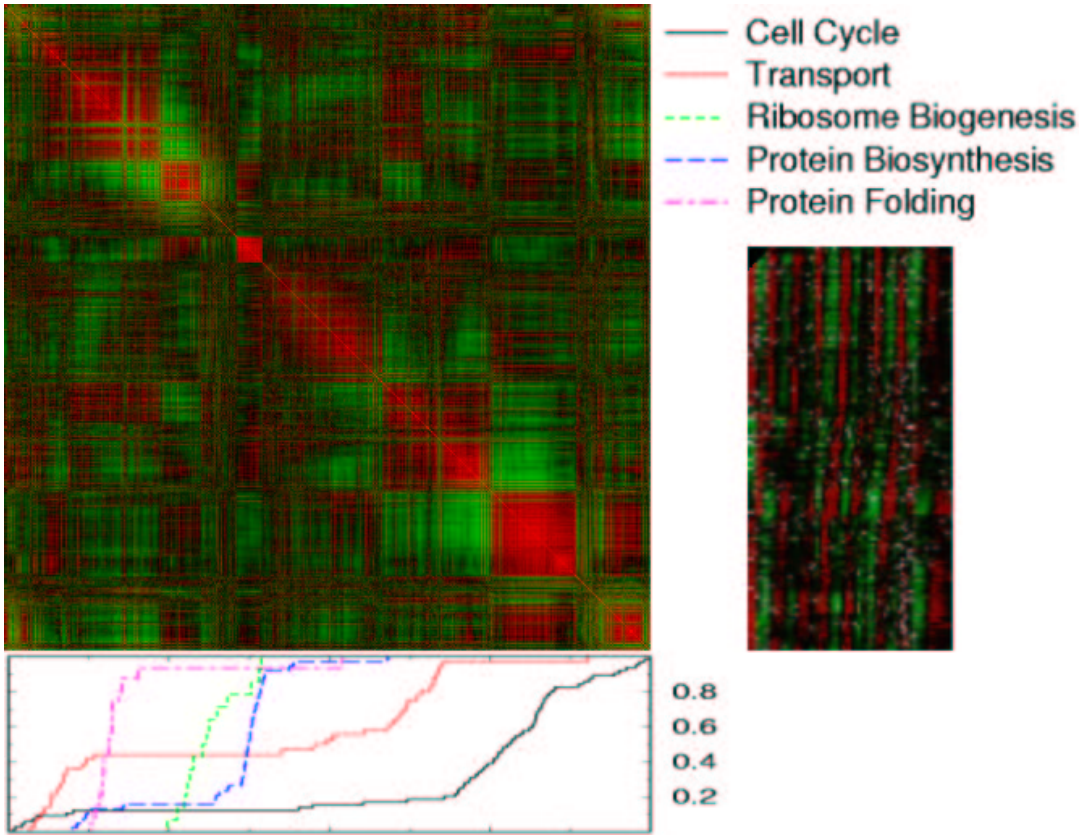

Figure 1: Correlation Matrix for the 800 most variant gene expression patterns (upper left). In the lower plot the integrated occurrence of genes assigned to a few representative gene-ontologies is plotted. The activity pattern for the Cell-Cycle regulated genes after reshuffling with $\gamma = 1$ is shown in the right diagram.

expected, the non-local interaction (eq. 1) sorts the groups on the main diagonal such that clusters with sub-dominant co-regulation group together. Therefore the distance of groups on the main diagonal reflects the relative strength of co-regulation. In the off-diagonal correlation coefficients an interesting fine structure can be observed: for example in the region marked with an arrow one sees that two groups, which are internally correlated, can be correlated and anti-correlated at the same time. From this observation one can infer a fine-structure into the groups on the diagonal.

The checker board patterns observed in the upper left and lower right in figure 1 are very interesting. They are generated by oscillatory processes: adjacent red and green blocks indicate co-regulated, but mutually exclusive expressed genes. These genes are active in different parts of a cycle. By inspection of gene annotations, the lower right functional group is identified as cell-cycle regulated genes. In [4] these genes were identified using a different method. Expression patterns, for which the Fourier component for frequencies close to the expected cell-cycle frequency were larger than some threshold were identified as cell-cycle regulated. For our method, the introduction of a threshold and knowledge of the oscillator frequency is not necessary.

The checker board in the upper left corner represents a second cell "clock". From the intensity of the correlation of this group with the cell-cycle, we conclude that this second clock is not strongly coupled to the cell cycle. Therefore this functional group is an independent oscillator. To elaborate this further we have analyzed the frequency spectrum for the expression patterns of genes in this group and for the cell-cycle regulated genes. Both spectral power distributions show a clear maximum at a frequency ν, which differs significantly for the two groups. For the second clock we find $\nu = 4/7\nu_{\rm CC}$, where $\nu_{\rm CC}$ is the frequency which maximizes the cell cycle power spectrum. The method used in [4] could therefore not identify the genes in this group as regulated by the cell-cycle. Unfortunately, the maximal time-span, for which experiments were done, contains only one complete cycle of this clock. It is therefore not possible to decide, if this is a continuous oscillator or a one-shot clock, possibly the orchestrated shock-response of the respective genes on the environmental conditions, like for example the procedure to synchronize the cells with respect to the cell cycle. Observing the onthology distribution in this part of the gene-list, it is plausible, that this clock controls the transcription of genes and the synthesis of proteins.

Next we want to show how the non-local part of the interaction in (1) influences the ordering of gene expression patterns. For each site i in the list

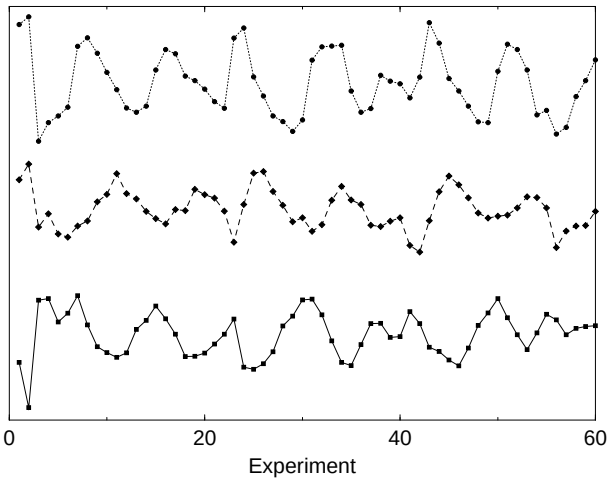

Figure 2: Expression patterns for a set of 260 genes Cell-Cycle regulated genes after reshuffling with $\gamma = 1$. The curves display the average expression for the genes at the reshuffled positions $10 - 40$ (solid), $100 - 130$ (dshed) and $190 - 220$ (dotted).

of expression patterns one can define an effective prototype expression P_i, which is induced by *all* expression patterns in the data set via the energy. This prototype is the pattern which minimizes the energy, the solution $D_{i,t}$ of

$$0 = \frac{\partial}{\partial D_{i,t}} S(\alpha, \beta), \qquad t = 1 \ldots \tau. \tag{5}$$

In general this optimal pattern will differ from site to site. Hence it can follow a signal which is deformed from a pattern A to a pattern B.

To demonstrate this property we choose the cell-cycle regulated genes, where the signal "activated" travels through the system. Differently from above, where the list of patterns was sorted with with respect to co-regulation, anti-correlated genes should not be grouped together in this case, because presumably they belong to opposite parts of the cell cycle. We therefore choose $\gamma = 1$, when relabeling the expression patterns. In figure 2 we show three representative examples for the average expression pattern observed at different parts of the reshuffled list of genes. The oscillatory signal reflects the activity of the respective genes in the different parts of the cell cycle. The three curves seem to follow the same signal, with some deformations and a phase shift. Nevertheless, in figure 1 these genes are sorted in one big group.

It is also interesting to observe, that the $\gamma = 1$ sorted data is essentially time ordered, as can be seen from the activity pattern shown in the same figure. Note that the algorithm does in no way explicitly refer to the time aspect in the data. In fact, the energy (1) is invariant under the exchange of experiments, different time-points. This observation confirms the capability of the algorithm to identify signal-pathway hidden in the data without prior knowledge.

4 Discussion

Re-shuffling is efficient in finding functional groups in the expression data. The philosophy of this algorithm is considerably different from cluster algorithm, which compare each pattern p locally with some prototypic pattern for each cluster and finally assign p to the cluster with the most similar prototype. In this way many of these algorithms do not use the information available about inter-cluster similarities. Re-shuffling is based on a global comparison with *all* patterns in the ensemble. In this way it makes use of the information contained in sub-leading similarities as well. We want to emphasize that this method is not restricted to the analysis of time expression data. It can also be used to detect patterns in static data to identify, for example, genes which are responsible for a certain phenotype or disease.

Self organizing maps use a non-local assignment of patterns to neurons. Therefore they can reflect inter-cluster similarities. They were used in [6, 7] to classify yeast gene expression patterns. With this method it is possible to identify the cell cycle oscillation as a dominant motif in the expression data [6]. However, the neurons most active for the corresponding patterns at different parts of the cell cycle were not grouped in an obvious way. It was not possible to identify these patterns as belonging to the same functional cycle.

The re-shuffling method is able to extract this information. Without any prior knowledge it can identify the cell cycle regulated genes. It is very interesting to observe that the algorithm extracts time information from the data without actually referring to the time aspect. This demonstrates that the pathways of signals can be extracted from the data using this method. This aspect can be very useful when analyzing functional groups which are not so well known as the cell cycle. A further example of the algorithms power is the identification of a second independent clock in yeast.

The cyclic distance measure (2) may seem inappropriate if the investigated data has no intrinsic cycle-structure. Consider for example a set of patterns, which is best sorted into a linear, *open* sequence. The cyclic distance measure forces this list into a contact which is inappropriate. This happens at the weakest point in the list, where the cost function penalty is smallest. It is easy to circumvent this problem by inserting a few uncorrelated dummy-patterns, which serve as a buffer between the free ends. A more severe constraint certainly is the one dimensional structure imposed by the method. It will therefore be interesting to generalize the method to a higher dimensional index space.

A general problem when analyzing expression data is noise. When measurements are easily available, the usual way to reduce noise is to increase the number M of measurements until the noise level $\sigma \propto 1/\sqrt{M}$ is small enough. Repeated measurements of the same system would also allow for a reliable estimation of the noise level. This knowledge is crucial when interpreting the results of an analysis. However, gene expression measurements are quite costly in time and are usually not repeated. Therefore the methods used to analyze the data have to be relatively insensitive to noise. The re-shuffling algorithm is very robust in this respect because the energy function (1) used in the sorting procedure averages over *all* patterns in the data set. Many cluster algorithms and self organizing maps only average over a subset of the data when extracting a prototype for a cluster (or a neighborhood of a neuron). They therefore tend to be more sensitive with respect to noise.

The visualization of the correlation matrix some insight into the connectivity of the underlying regulatory network. One may ask if it is possible to learn the full network from the data. One should be aware of the limitations in the data available so far. It is very possible that large parts of the network were inactive for the states observed. These "unexcited" parts of the network can not be deduced from the data. Furthermore important parts of the regulatory network, like e.g. inter- and intra-cell signalling, are observed only indirectly by back reaction on the gene expression pattern.

5 Acknowledgement

I want to thank Å. Borg, C. Peterson and M. Ringnér for fruitful discussions. This work was supported by the Swedish Foundation for Strategic Research.

References

[1] Schena, M., Shalon, M., Davis, R.W., Brown, P.O. "Quantitative Monitoring of Gene Expression Patterns with a Complimentary-DNA Microarray" (1995) *Science* **270**, 467-70

[2] Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. (1998), "Cluster Analysis of Genome Wide Expression Patterns", (1998) *Proc. Natl. Acad. Sci.* USA **95**, 14863-68

[3] Cho, R.J., Campbell, M.J., Winzeler, E.J., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W. "A genome-wide transcriptional analysis of the mitotic cell cycle", (1998) *Mol. Cell* **2**, 65-73

[4] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.N., Brown, P.O., Botstein, D., Futcher, B. "Identification of cell cycle regulated genes in yeast by DNA microarray hybridization" (1998) *Mol. Biol. Cell.* **9** , 3273-97

[5] Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., (1983) *Science* **220**, 671-80

[6] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Litareewan, S. Dmitrosky, E., Lander, E.S., Golub, T.R., "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation", (1999) *Proc. Natl. Acad. Sci.* USA **96**, 2907-12

[7] Törönen, P., Kolehmainen, M., Wong, G,. Castren, E. "Analysis of gene expression data using self-organizing maps", (1999) *FEBS Letters* **451**, 142-6

[8] http://www.geneontology.org