Including transcription factor information in the superparamagnetic clustering of microarray data

M. P. Monsiváis-Alonso[a], J. C. Navarro-Muñoz[a], L. Riego-Ruiz[b], R. López-Sandoval[a], H.C. Rosu[a]

[a] *Division of Advanced Materials, IPICyT, Instituto Potosino de Investigacion Cientifica y Tecnologica, San Luis Potosí, S.L.P., Mexico*
[b] *Division of Molecular Biology, IPICyT, Instituto Potosino de Investigacion Cientifica y Tecnologica, San Luis Potosí, S.L.P., Mexico*

Abstract

In this work, we modify the superparamagnetic clustering algorithm (SPC) by adding an extra weight to the interaction formula that considers which genes are regulated by the same transcription factor. With this modified algorithm that we call SPCTF, we analyze Spellman *et al.* microarray data for cell cycle genes in yeast, and find clusters with a higher number of elements compared with those obtained with the SPC algorithm. Some of the incorporated genes by using SPCFT were not detected at first by Spellman *et al.* but were later identified by other studies, whereas several genes still remain unclassified. The clusters composed by unidentified genes were analyzed with MUSA, the motif finding using an unsupervised approach algorithm, and this allow us to select the clusters whose elements contain cell cycle transcription factor binding sites as clusters worth of further experimental studies because they would probably lead to new cell cycle genes. Finally, our idea of introducing available information about transcription factors to optimize the gene classification could be implemented for other distance-based clustering algorithms.

Keywords: Superparamagnetic clustering, similarity measure, microarrays, cell cycle genes, transcription factors.

paper-physa-3 20100901.tex

Physica A 389(24), 5689-5697 (2010)
doi: 10.1016/j.physa.2010.09.006

Email addresses: monsivais@ipicyt.edu.mx (M. P. Monsiváis-Alonso), jcarlos@ipicyt.edu.mx (J. C. Navarro-Muñoz), lina@ipicyt.edu.mx (L. Riego-Ruiz), sandov@ipicyt.edu.mx (R. López-Sandoval), hcr@ipicyt.edu.mx (H.C. Rosu)

1. Introduction

DNA microarrays allow the comparison of the expression levels of all genes in an organism in a single experiment, which often involve different conditions (*i.e.* health-illness, normal-stress), or different discrete time points (*i.e.* cell cycle) [1, 2]. Among other applications, they provide clues about how genes interact with each other, which genes are part of the same metabolic pathway or which could be the possible role for those genes without a previously assigned function. DNA microarrays also have been used to obtain accurate disease classifications at the molecular level [3, 4, 5]. However, transforming the huge amount of data produced by microarrays into useful knowledge has proven to be a difficult key step [6].

On the other hand, clustering techniques have several applications, ranging from bioinformatics to economy [7, 8, 9]. Particularly, data clustering is probably the most popular unsupervised technique for analyzing microarray data sets as a first approach. Many algorithms have been proposed, hierarchical clustering, k-means and self-organizing maps being the most known [10, 11]. Clustering consists of grouping items together based on a similarity measure in such a way that elements in a group must be more similar between them than between elements belonging to different groups. The similarity measure definition, which quantifies the affinity between pairs of elements, introduces *a priori* information that determines the clustering solution. Therefore, this similarity measure could be optimized taking into account additional data acquired, for example, from real experiments. Some works with *a priori* inclusion of bioinformation in clustering models can be found in [12, 13].

In the case of gene expression clustering, the behavior of the genes reported by microarray experiments is represented as N points in a D-dimensional space, being N the total number of genes, and D the number of conditions. Each gene behavior (or point) is then described by its coordinates (its expression value for each condition). Genes whose expression pattern is similar will appear closer in the D-space, a characteristic that is used to classify data in groups. In our case, we have used the Superparamagnetic Clustering Algorithm (SPC) [14, 15, 16, 17], which was proposed in 1996 by Domany and collaborators as a new approach for grouping data sets. However, this methodology has difficulties dealing with different density clusters, and in order to ameliorate this, we report here some modifications of the original algorithm that improve cluster detection. Our main contribution consists on increasing the similarity measure between genes by taking advantage of transcription factors, special proteins involved in the regulation of gene expression.

The present paper is organized as follows: in Section 2, the SPC algorithm is introduced, as well as our proposal to include further biological information and our considerations for the selection of the most natural clusters. Results for a real data set, as well as performance comparisons, are presented in Section 3. Finally, Section 4 is dedicated to a summary of our results and conclusions.

2. Method

2.1. Superparamagnetic Clustering Algorithm (SPC)

A Potts model can be used to simulate the collective behavior of a set of interacting sites using a statistical mechanics formalism. In the more general inhomogeneous Potts model, the sites are placed on an irregular lattice. Next, in the SPC idea of Domany *et al.* [14], each gene's expression pattern is represented as a site in an inhomogeneus Potts model, whose coordinates are given by the microarray expression values. In this way, a particular lattice arrangement is spanned for the entire data set being analyzed.

A spin value σ_i, arbitrarily chosen from q possibilities, is assigned to each site, where i corresponds to the site of the lattice $i = 1, 2, ..., N$. The main idea is to characterize the resulting spin configuration by the ferromagnetic Hamiltonian:

$$H = -\sum_{i,j} J_{ij}\delta_{\sigma_i,\sigma_j}, \qquad \sigma_i = 1, ..., q, \tag{1}$$

where the sum goes over all neighboring pairs, σ_i and σ_j are spin values of site i and site j respectively, and J_{ij} is their ferromagnetic interaction strength.

Each site interacts only with its neighbors, however since the lattice is irregular, it is necessary to assign the set of nearest-neighbors of each site using the so-called k-mutual-nearest-neighbor criterion [18]. The original interaction strength is as follows:

$$J_{ij} = \begin{cases} \frac{1}{\hat{K}} e^{-\frac{d_{ij}^2}{2a^2}} & \text{if } i \text{ and } j \text{ are neighbors} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

with $\hat{K}$ the average number of neighbors per site and a the average distance between neighbors. The interaction strength between two neighboring sites decreases in a Gaussian way with distance d_{ij} and therefore, sites that are separated by a small distance have more probability of sharing the same spin value during the simulation than the distant sites. On the other hand, said probability, $P_{ij} = (1 - e^{(-J_{ij}/T)})$, also depends on the temperature T, which acts as a control parameter. At low temperatures, the sites tend to have the same spin values, forming a ferromagnetic system. This configuration is preferred over others because it minimizes the total energy. However, the probability of encountering aligned spins diminishes as temperature increases, and the system could experience either a single transition to a totally disordered state (paramagnetic phase), or pass through an intermediate phase in which the system is partially ordered, which is known as the superparamagnetic phase. In the latter case, varios regions of sites sharing the same spin value emerge. Sites within these regions interact among them with a stronger force, exhibiting at the same time weak interactions with sites outside the region. These regions could fragment into smaller grains, leading to a chain of transitions within the superparamagnetic phase until the temperature is so high that the system enters the paramagnetic phase, where each spin behaves independently. This

hierarchical subdivision in magnetic grains reflects the organization of data into categories and subcategories. Regions of aligned spins emerging during simulation correspond to groups of points with similar coordinates, *i.e.*, similar gene expression patterns [14, 15, 16]. This subdivision can be simulated, for example, by using the Monte Carlo approach, by which one can compute and follow the evolution of system properties such as energy, magnetization and susceptibility, while the temperature is modified. In addition, the temperature ranges in which each phase transition takes place can be localized.

Rather than thresholding the distances between pairs of sites to decide their assignment to clusters, the pair correlation G_{ij}, indicating a collective aspect of the data distribution, is preferred . It can be calculated as follows [15]

$$G_{ij} = \frac{(q-1)(\langle \delta_{\sigma_i,\sigma_j} \rangle) + 1}{q-1} \ . \tag{3}$$

In this way, G_{ij} is the normalized probability for finding two Potts spins σ_i and σ_j sharing the same value for a given temperature step. If both spins belong to the same ordered region, their correlation value would be close to one, otherwise their correlation would be close to zero [17]. Thus, for each temperature step, two sites are assigned to the same cluster if their correlation exceeds a threshold value of $G_{ij} > 0.5$. If a site does not have a single correlation value greater than 0.5, it is joined with its neighbor showing the highest value.

2.2. Transcription Factors in SPC (SPCTF)

For our SPCTF algorithm, we also accept sites whose G_{ij} are larger than 0.5 in order to build a cluster. However, differently from the traditional SPC algorithm [14, 15, 16, 17], if two sites do not reach the G_{ij} value greater than 0.5 they are not connected. This is because with our data we have found that the original condition led to unnatural growth of some clusters when the temperature is increased.

As already mentioned, the data are fragmented in various clusters for each temperature value, and for higher temperatures, the number of clusters increases due to finer and finer segmentation. In order to select the more representative clusters through all temperature steps, we assign a stability value to each obtained cluster, based on its evolution. We define T_t as the number of temperature steps until the system reaches the paramagnetic phase and T_v as the number of temperature steps a cluster v survives, while I_t and I_v are defined as the total number of sites and the number of elements in a given cluster, respectively. We assign a stability parameter $\mathcal{S}_v$ to each cluster, as follows:

$$\mathcal{S}_v \ = \ \frac{col_v row_v}{\mid col_v - row_v \mid + \epsilon}, \tag{4}$$

where $col_v = \frac{T_v}{T_t}$ is the fraction of temperature steps a cluster v survives, while $row_v = \frac{I_v}{I_t}$ is the fraction of total elements belonging to v. The advantage

of using the stability parameter $\mathcal{S}_v$ is that it gives preference to clusters that survive several temperatures, but also have an acceptable number of elements. We added a small positive real number ϵ to the denominator in the expression of $\mathcal{S}_v$ for the special case when $col_v = row_v = n$, where n belongs to the range $(0, 1]$, leading to $\mathcal{S}_v = \frac{n^2}{\epsilon}$ instead of the infinity.

It has been reported that the main drawback of the SPC algorithm consists of dealing with data showing regions of different density [19, 20]. In this case, either depending on temperature or the number of neighbors selected, some clusters will easily get prominent whereas the detection of others will be hindered. To overcome this problem, at least two techniques have been proposed *e.g.*, sequential superparamagnetic clustering [19] and a modularity approach [20]. Our idea is to take advantage of already available biological information to improve lattice connectivity in such a way that biologically significant clusters have more probability of being detected by the algorithm.

Indeed, at the transcriptional level, the expression of a gene could be promoted/suppressed by the binding of the proteins named transcription factors to specific sequences on the gene promoter region. Then, if a group of genes shows the same expression behavior in a microarray experiment, it is quite possible that they are being regulated by a specific transcription factor, forming a group of coregulated genes [21]. Thus, available information about which genes are targeted by the same transcription factors may be useful in the detection of groups of genes with similar expression profiles.

To make effective this idea, we downloaded from *www.yeastract.com* a list of yeast transcription factors that are well documented, and whenever two neighboring genes are controlled by the same transcription factor, we increased their interaction strength. It is important to note that the list provided by *www.yeastract.com* includes transcription factors associated with several processes and are not only cell cycle related. The formula that takes this into account replaces Eq. (2) of the original algorithm, and has the following form:

$$J_{ij} = \begin{cases} \frac{F}{K} e^{-\frac{d_{ij}^2}{2(Fa)^2}} & \text{if } i \text{ and } j \text{ are neighbors,} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Here, $F = fn$ is the number of common transcription factors shared by i and j (n, which varies for each pair of neighboring genes), multiplied by a factor f which was chosen to be 2.0 after comparing the results obtained with several other values. The selected value has the characteristic of preserving well-defined susceptibility peaks as well as obtaining larger clusters. The objective is to strengthen some connections without preventing the natural fragmentation of clusters caused by the temperature parameter. If two elements do not share a transcription factor, then $F = 1$, recovering the original SPC formula. Therefore, the modified interaction strength between each site and its neighbors is governed by two aspects: the distance between them, which comes from gene expression values generated through microarray experiments, and the number

of transcription factors regulating both genes, obtained from documented biological data. Any time two genes share a transcription factor, their interaction strength becomes larger, and this favors that the clusters including these sites remain stable for longer temperature ranges, with the corresponding increase of their stability values.

3. Results and Discussion

We analyzed Spellman *et al.* [22] microarray data in which gene expression values from synchronized yeast cultures were obtained at various time moments, aiming to identify cell cycle genes. Yeast cultures were synchronized by three methods: adding alpha pheromone, which arrests cells in the G1 phase; using centrifugal elutration for separating small G1 cells; and using a mutation that arrests cells late in mitosis at a given temperature. Combining the three experiments and using Fourier and correlation algorithms, Spellman *et al.* [22] reported 800 cell cycle regulated genes.

The goal was to compare the performance of SPC and SPC with transcription factors (SPCTF), which are algorithms that do not make assumptions about periodicity. Nonetheless, the overall analysis is time consuming and we only selected the data set treated with the alpha pheromone, available at *http://cellcycle-www.stanford.edu*. Genes with missing values were discarded, leaving an input matrix of 4489 genes and 18 time courses that included only 613 of the genes reported by Spellman *et al.* [22]. Furthermore, as we do not include the other two synchronization experiments, we expect to loose some of their cell cycle genes.

It is worth mentioning that Getz *et al.* [23] also analyzed the Spellman alpha synchronized set with the SPC algorithm. They took 2467 genes which have characterized functions and introduced a Fourier transform to take into account the oscillatory nature of the cell cycle. In our case, however, we decided not to introduce any considerations about the periodicity of the data, mainly because the time series cover only two cell cycle periods [24].

We obtain compact gene clusters implementing SPC original algorithm and SPCTF, both with parameter values $k = 8$ and $q = 20$. The cluster with the highest stability value contains an extremely large number of elements without a clear biological linkage between them. It is mainly composed of genes whose expression do not change significantly over time, thus it is possible that they are included here for this very reason. We discard this cluster from our analysis, although it could always be taken apart and analyzed again with SPCTF by choosing the appropiate number of neighbors to obtain more information.

To compare in more detail both approaches, it is necessary to correlate each cluster in the SPC method with its equivalent in SPCTF. In order to do this, we calculate the euclidian distance between the mean position vector of every cluster in each approach, and choose the pairs with the shortest distance between them. (We recall that the mean position vector of a cluster is obtained by averaging each coordinate between all its elements). Although different measures could have been used, this one performed adequately, as can be seen in the

supplementary information file, where we provide a more detailed comparison between SPCTF and SPC clusters. In Table 1, we present the differences in cluster size as well as the hits, the number of genes reported by Spellman *et al.* [22], which have been included in the clusters. When going through the SPCTF approach, one can see that the first largest cluster looses some genes, while the number of the rest of the clusters augments. Besides, hits or coincidences with Spellman *et al.* [22] cell cycle genes in clusters of six or more elements increase by 61%, from 108 to 174. Therefore, we were able to incorporate several genes to these clusters, mainly from outliers.

Comparison between SPC and SPCTF

Method	First Cluster	Cluster size ≥ 6	Cluster size $= 5$	Cluster size $= 4$	Cluster size $= 3$	Cluster size $= 2$	Cluster size $= 1$	Total Clusters	Total Genes	**Total Hits**
SPC	1(2078) **68**	19(220) **108**	5(25) **2**	23(92) **11**	57(171) **39**	144(288) **49**	1615 **336**	1864	(4489)	**613**
SPCTF	1(1657) **64**	27(359) **174**	13(65) **23**	32(128) **22**	61(183) **30**	187(374) **60**	1723 **240**	2044	(4489)	**613**

Table 1: Number of clusters for different cluster size. The total number of genes for each cluster size appears in parentheses and their hits with Spellman *et al.* [22] appear in bold type. Hits with the 613 cell cycle genes reported by Spellman *et al.* [22] increase for clusters of size 6 and bigger, while decreasing in the first cluster and outliers.

In the following analysis, we focus on clusters of six or more elements, because we are interested in finding groups of several genes sharing the same expression pattern (coregulated genes). Results of the comparison for the first 27 most stable clusters, discarding the first one, are shown in Fig. 1. Generally, these clusters incorporate more elements with SPCTF, including more cell cycle genes as those reported by Spellman *et al.* [22] and thus improving the matching.

Depending on the available information about the genes, we classify the clusters in three groups. The first cluster type, cell cycle genes, CC, corresponds to groups formed in their majority ($\geq 85\%$) by already reported cell cycle genes (Fig. 2). The second type, mixed genes, M, contains clusters with non-reported genes as well as already known cell cycle genes (Fig. 3), and in the third type, no hits, N, we include the clusters that contain only one hit or are entirely composed of non-previously identified cell cycle genes (Fig. 3).

It is worth mentioning that more cell cycle experiments have been done since Spellman *et al.* [22] and new genes have been classified meanwhile as cell cycle regulated. Some of these newly reported cell cycle genes were obtained by Cho *et al.* [25], Pramila *et al.* [26], Rowicka *et al.* [27] and Lichtenberg *et al.* [28]. We analize our 27 clusters taking now as hits, genes reported either by Spellman *et al.* [22] or by one of the above mentioned studies. In this way, we gained thirty additional hits in the SPC clusters, while in SPCTF clusters we have fifty-two extra genes. The results including all the aforementioned cell cycle studies are presented in Figs. 4–6 [29].

In addition, we analyze the expression profiles of the genes conforming each cluster using the SCEPTRANS tool [30], and we notice that all the genes

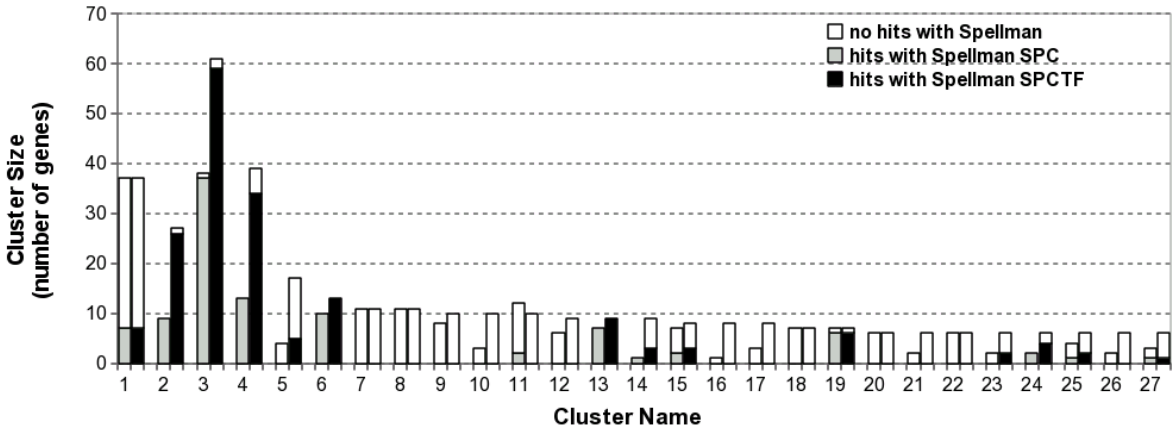


Figure 1: General comparison of the first 27 clusters, discarding the first one. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF. Groups tend to increase in size and also in hits with cell cycle genes reported by Spellman *et al.* [22], with the exception of cluster 11.

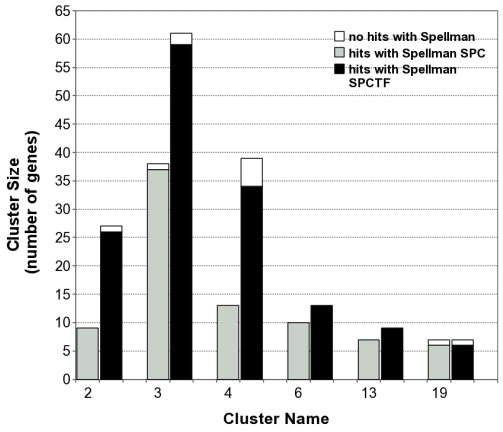


Figure 2: Comparison between the SPC and SPCTF results, showing the CC clusters. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF.

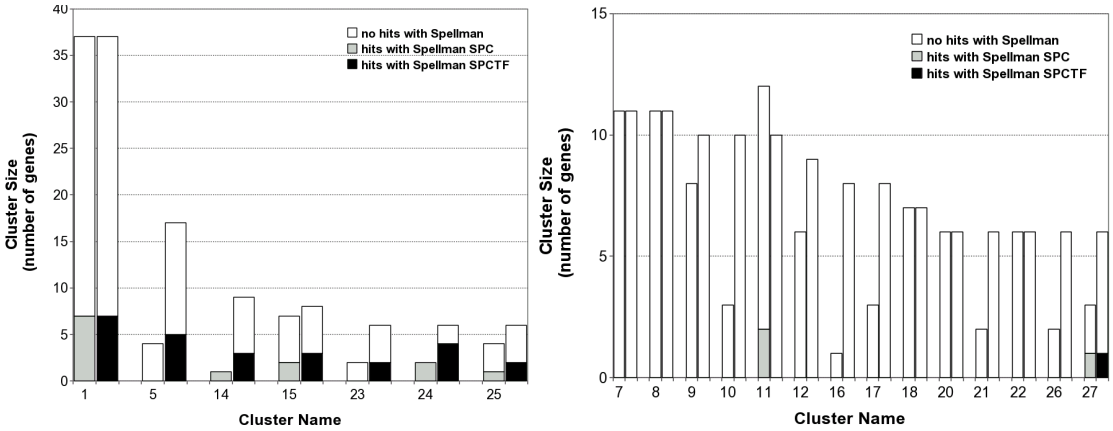


Figure 3: M and N clusters, left and right respectively. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF.

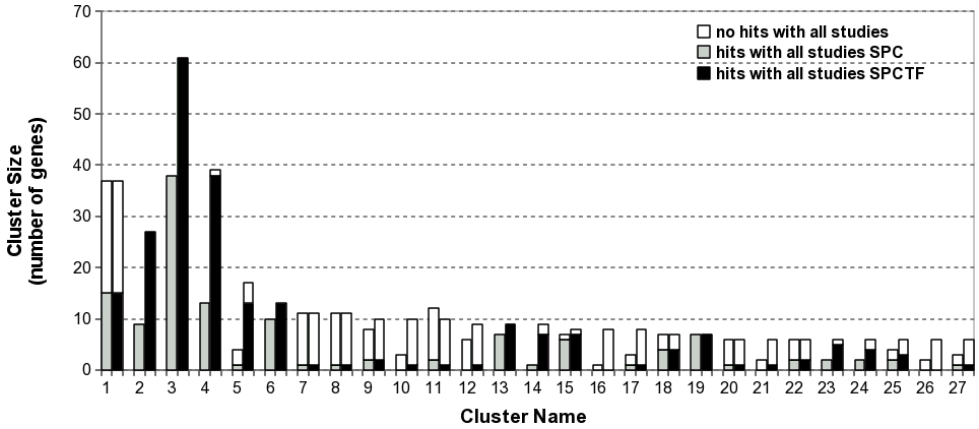


Figure 4: General comparison of the first 27 most stable clusters. Hits are now taken as cell cycle genes reported by all studies. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF.

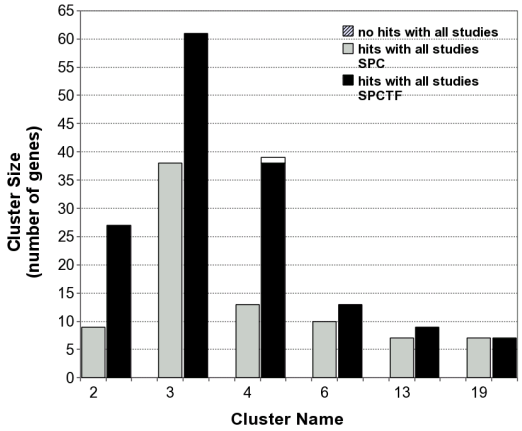


Figure 5: Comparison between SPC and SPCTF results, showing CC clusters. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF.

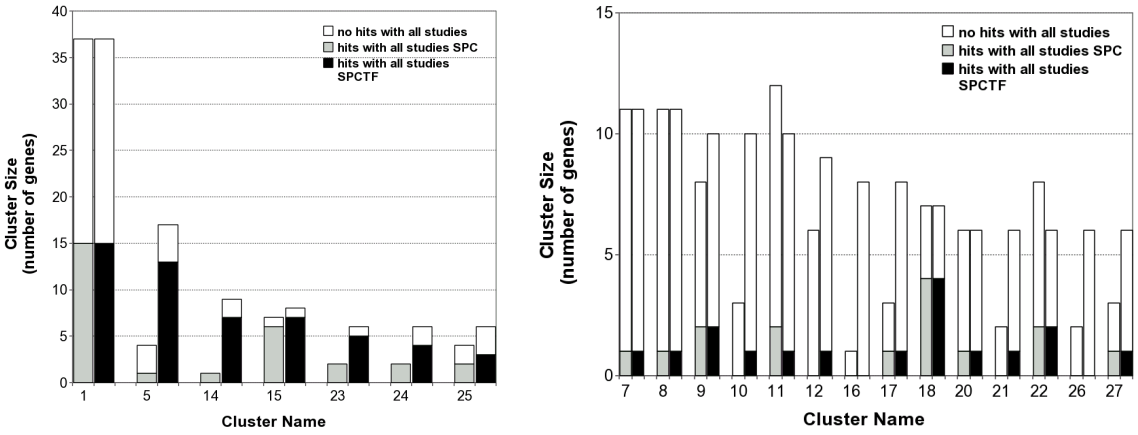


Figure 6: M and N clusters, left and right respectively. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF.

grouped in the same cluster had the same expression pattern. This gives us further confidence that our algorithm is grouping data correctly. The expression profiles for a representative member of each cluster type are shown in Fig. 7. We also find two clusters (21 and 27) that present an oscillating behaviour that is due to an artifact in the manner the microarray experiment was performed, see [31, 32]. In the supplementary information file, we include the list of oscillating genes identified in [31] and the number of these genes inside each of our first 27 clusters. We also include the expression profiles of these clusters as well as those of size 5 and 4 which contain hits with cell cycle genes identified by Spellman *et al.* [22]. These clusters have also similar expression profiles but were not further analyzed because of their low number of elements. In the case of gene annotation, it is important to have clusters of many elements to effectively assure that an unknown gene shares the biological function already assigned to the other genes in the same cluster.

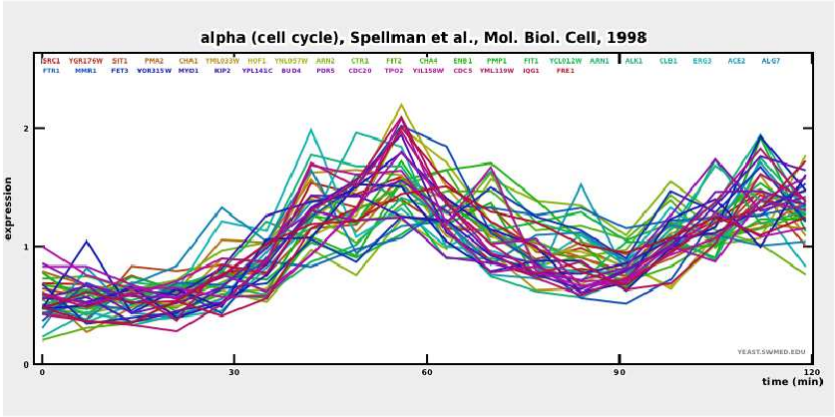
The CC clusters are almost entirely composed of cell cycle regulated genes reported either by Spellman *et al.* [22] or by other authors, besides, their expression patterns are similar, which leaves no doubt on their validity. For the M and N clusters, we know that they are well grouped because their elements share the same expression patterns, but in order to select those of worth for further analysis (for example in a laboratory experiment) we analyze them through MUSA, motif finding using an unsupervised approach algorithm, that can be found at *www.yeastract.com*. This program searches for the most common sequences (motifs) in the regulatory region of a set of genes, and compare them to the transcription factor binding sites already described in yeastract database [33, 34]. Results of this analysis are shown in Table 2, which includes the quorum or percentage of genes containing a motif in each cluster, and the alignment score, which quantifies the level of similarity between the encountered motif and the known transcription factor associated with it. The clusters that probably would give us the best results would be those associated with cell cycle transcription factors with high percentages and scores. We select in this way, the clusters 1, 5, 9, 12, 16 and 24 because they have percentages higher than 70% and scores higher than 80%.

In order to validate the MUSA analysis, we also constructed various clusters with sizes ranging from six to thirty-seven genes that were composed by genes selected at random from the original data. When analyzing these random clusters in the same way in MUSA, we obtain at most two cell cycle transcription factor coincidences.

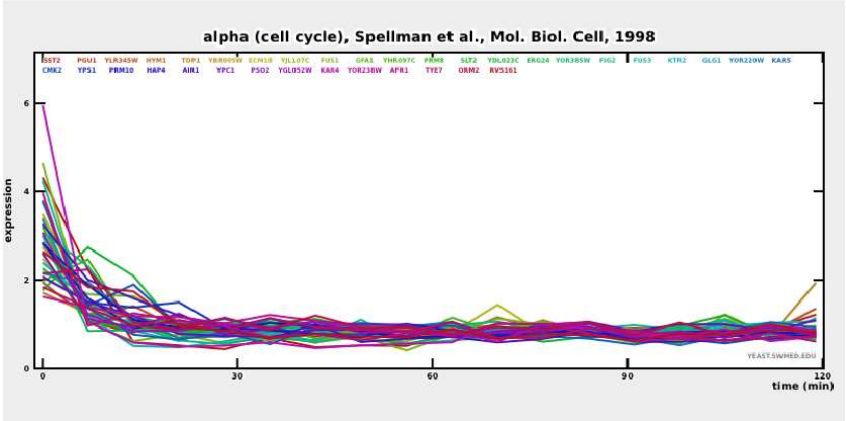
4. Summary and conclusions

Large amounts of biological information are constantly obtained by throughput techniques and clustering algorithms have taken an important place in the unraveling of this information. However, the clustering analyses offer a difficult challenge because any data set can be grouped in numerous ways, depending on the level of resolution asked for and the applied similarity measure. In this work, we propose the use of available biological information in order to strengthen the

CC: cluster 4



M: cluster 1



N: cluster 16

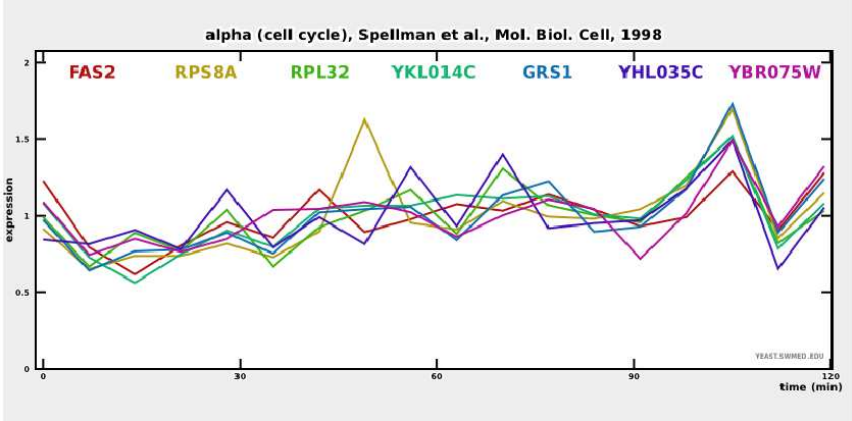


Figure 7: (Color online) Expression profiles for a representative member of each cluster type using the SCEPTRANS tool. Expression profiles for all clusters are available in the supplementary information.

MUSA analysis

Cluster Name	Cluster Type	Transcription Factor Association with Promoter (alignment score/maximum possible score)	Percentage of cluster genes sharing a motif, *i.e.* quorum
1	M	Cup2p, Mig3p, Mig2p, Mig1p, Arg80p (5/6)	91.67 %
		Swi4p (6/6), Azf1p, **Ime1p**, Dal82p, Dal81p (5/6)	88.89 %
		Ste12p (7/7), Rox1p (6/7)	83.33 %
5	M	Hac1p (6/6), **Rme1p**, Arg80p, Mot3p (5/6)	76.47 %
7	N	Azf1p, Zap1p (6/7)	81.82 %
8	N	Low scores	Low percentages
9	N	Mig3p, Mig1p, Crz1p, Mig2p (5/6)	80 %
		Rfx1p, Arg81p (6/7)	70 %
10	N	Low scores	
11	N	Azf1p (6/7)	
12	N	Azf1p (7/8)	88.89 %
		Rfx1p, Cup2p (5/6)	77.78 %
14	M	Azf1p (7/8)	75 %
15	M	Low scores	Low percentages
16	N	**Mcm1p** (5.25/6), Crz1p (5/6)	100 %
		Hap1p (5/6)	71.43 %
17	N	Arg81p, Upc2p, Sip4p, Rox1p, Crz1p, Zap1p (5/6)	100
		Pdr8p (5.33/6)	87.5 %
18	N	Azf1p, Zap1p (6/7)	100 %
20	N	Low scores	
21	N	Low scores	
22	N		Low percentages
23	M	Low scores	
24	M	Hap1p (6/6), Ecm22p, Upc2p (5/6)	100 %
		Rfx1p (6/7)	83.33 %
25	M	Hac1p (6/7)	83.33 %
26	N	Dal80p, Gat1p, Gln3p, Gzf3p (6/7)	83.33 %
27	N	Ino4p (6.5/7), Ino2p (6/7)	100 %

Table 2: Results for quorum higher than 70% and scores higher than 80%. Transcription factors associated to cell cycle are shown in bold. The most confident clusters are taken as those that included cell cycle transcription factor.

interaction between genes which share a transcription factor involved in any metabolic process, improving the similarity measure. This information is introduced in the natural evolution of the SPC algorithm, and in this way, we are able to enhance the creation and endurance of groups of possible coregulated genes. As the network spanned by the transcription factors information connects all genes, clustering directly *a posteriori* using only this information in the present case results into a single massive cluster (See section IV of the Supplementary Information). However, by having the distance play an important weight in the interaction formula, the far-located clusters will not join, despite sharing transcription factors between their genes.

With this in mind, we have modified the SPC algorithm, and applied both the original and modified SPCTF algorithm to one of the three Spellman *et al.* [22] data sets of the yeast cell cycle. The expression profiles of the genes in all resulting clusters show a similar behavior, but we obtain larger clusters with SPCTF. We classified them in three types, CC, M, and N, depending on the amount of cell cycle reported elements inside each cluster. With SPCTF, the CC type clusters increase in size including more cell cycle genes, and for the M and N type clusters, we also looked for common sequences in its regulatory regions and selected various groups worth of further research in order to report possible new

cell cycle genes. As expected, some of these clusters include already known cell cycle genes sharing a transcription factor, *but more importantly, at the predictive level, they promote the inclusion of new genes with similar expression patterns.* It is also important to note that the modified algorithm can be applied to any data set, and the followed methodology leads to the selection of the potential gene subsets feasible to be experimentally investigated. Our work can serve as an example of how the inclusion of available biological information, such as transcription factors, and bioinformatic tools, such as MUSA, can lead to better and more confident results, aiding in the analysis of data coming from microarray experiments.

Acknowledgments

The authors thank Drs. S. Ahnert and G. Sherlock for useful discussions and comments. We also thank CONACYT for providing support for two of the authors (M.P.M.A. and J.C.N.M.) and also the referees for helpful remarks and information. This work was partly supported through the project SEP-CONACYT-2005-49039.

References

[1] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, Science **270**, 467 (1995).

[2] A. Schulze and J. Downward, Nat. Cell Biol. **3**, E190 (2001).

[3] G. J. McLachlan, K-A. Do, and C. Ambroise, *Analyzing Microarray Gene Expression Data* (Wiley, Hoboken, New Jersey, 2004).

[4] V. Trevino, F. Falciani, and H. A. Barrera-Saldaña, Mol. Med. **13**, 527 (2007).

[5] S. Draghici, *Data Analysis Tools for DNA Microarrays* (Chapman and Hall/CRC Press, Boca Raton, Florida, 2003).

[6] A. Fadiel and F. Naftolin, Int. Arch. Biosci., 1111 (2003).

[7] J.A. Hartigan, *Clustering Algorithms* (Wiley, New York, 1975).

[8] H. C. Romesburg, *Cluster Analysis for Researchers* (Lulu Press, North Carolina, 2004).

[9] R. Xu and D. Wunsch II, *Clustering* (Wiley, Hoboken, New Jersey, 2009).

[10] R. Xu and D. Wunsch II, IEEE Trans. on Neural Networks **16**, 645 (2005).

[11] A. K. Jain, M. N. Murty, and P. J. Flynn, ACM Comput. Surv. **31**, 264 (1999).

[12] W. Pan, Bioinformatics **22**, 795 (2006).

[13] D. Huang and W. Pan, Bioinformatics **22**, 1259 (2006).

[14] M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996).

[15] M. Blatt, S. Wiseman, and E. Domany, Neural Comp. **9**, 1805 (1997).

[16] S. Wiseman, M. Blatt, and E. Domany, Phys. Rev. E **57**, 3767 (1998).

[17] E. Domany, Physica A **263**, 158 (1999).

[18] K. Chidananda Gowda and G. Krishna, Pattern Recognition **10**, 105 (1978).

[19] T. Ott, A. Kern, A. Schuffenhauer, M. Popov, P. Acklin, E. Jacoby, and R. Stoop, J. Chem. Inf. Comput. Sci. **44**, 1358 (2004).

[20] L. Angelini, D. Marinazzo, M. Pellicoro, and S. Stramaglia, J. Stat. Mech., L08001 (2007).

[21] H. Yu, N. Luscombe, J. Quian, and M. Gerstein, Trends. Genet. **19**, 422 (2003).

[22] P. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, Mol. Biol. Cell **9**, 3273 (1998).

[23] G. Getz, E. Levine, E. Domany, and M. Q. Zhang, Physica A **279**, 457 (2000).

[24] R. Amato, A. Ciaramella, N. Deniskina, C. Del Mondo, D. di Bernardo, C. Donalek, G. Longo, G. Mangano, G. Miele, G. Raiconi, A. Staiano and R. Tagliaferri, Bioinformatics **22**, 589 (2006).

[25] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, Mol. Cell **2**, 65 (1998).

[26] T. Pramila, W. Wu, S. Miles, W. S. Noble, and L. L. Breeden, Genes Dev. **20**, 2266 (2006).

[27] M. Rowicka, A. Kudlicki, B.P. Tu, and Z. Otwinowski, PNAS **104**, 16892 (2007).

[28] U. de Lichtenberg, R. Wernersson, T. S. Jensen, H. B. Nielsen, A. Fausboll, P. Schmidt, F. B. Hansen, S. Knudsen, and S. Brunak, Yeast **22**, 1191 (2005).

[29] The list of the twenty-seven clusters with the names of all their genes and identified hits is available on request from the authors.

[30] A. Kudlicki, M. Rowicka, and Z. Otwinowski, Bioinformatics **23**, 1559 (2007).

[31] S. E. Ahnert, K. Willbrand, F. C. S. Brown, and T. M. A. Fink, Bioinformatics, **22**, 1471 (2006).

[32] E. M. Conlon, X. S. Liu, J. D. Lieb, and J. S. Liu, PNAS **100**, 3339 (2002).

[33] N. D. Mendes, A. C. Casimiro, P. M. Santos, I. Sa-Correia, A. L. Oliveira, and A. T. Freitas, Bioinformatics **22**, 2996 (2006).

[34] P. T. Monteiro, N. D. Mendes, M. C. Teixeira, S. d'Orey, S. Tenreiro, N. P. Mira, H. Pais, A. P. Francisco, A. M. Carvalho, A. B. Lourenco, I. Sa-Correia, A. L. Oliveira, and A. T. Freitas, Nucleic Acids Res. **36**, D132 (2008).