INSTITUTO POTOSINO DE INVESTIGACIÓN CIENTÍFICA Y TECNOLÓGICA, A.C.

POSGRADO EN CIENCIAS APLICADAS

Introducing Biological Information in the Superparamagnetic Clustering Algorithm of Gene Expression Data

Tesis que presenta

M. C. María del Pilar Monsiváis-Alonso

Para obtener el grado de

Doctora en Ciencias Aplicadas

En la opción de

Nanociencias y Nanotecnología

Codirectores de la Tesis:

Dra. Lina Raquel Riego Ruiz

Dr. Haret-Codratian Rosu Barbus

Dr. Román López-Sandoval

San Luis Potosí, S.L.P., Enero de 2012

Constancia de Aprobación de la Tesis

La tesis **"Introducing Biological Information in the Superparamagnetic Clustering Algorithm of Gene Expression Data"** presentada para obtener el Grado de Doctorado en Ciencias Aplicadas en la opción de Nanociencias y Nanotecnología fue elaborada por **María del Pilar Monsiváis Alonso** y aprobada el **9 de enero del 2012** por los suscritos, designados por el Colegio de Profesores de la División de Materiales Avanzados del Instituto Potosino de Investigación Científica y Tecnológica, A.C.

Dra. Lina Raquel Riego Ruiz
Codirectora de la tesis

Dr. Haret C. Rosu Barbus
Codirector de la tesis

Dr. Román López Sandoval
Codirector de la tesis

Dr. Braulio Gutiérrez Medina
Asesor

Dr. Raúl Garibay Alonso
Asesor Externo

Créditos Institucionales

Esta tesis fue elaborada bajo la codirección de los doctores Lina Raquel Riego Ruiz (División de Biología Molecular), Haret C. Rosu Barbus y Román López Sandoval (División de Materiales Avanzados) del Instituto Potosino de Investigación Científica y Tecnológica, A.C.

Durante la realización del trabajo el autor recibió una beca académica del Consejo Nacional de Ciencia y Tecnología (No. de registro 182493) y apoyo del Comité de Becas del Instituto Potosino de Investigación Científica y Tecnológica, A. C.

Certificado de Grado

IPICYT

Instituto Potosino de Investigación Científica y Tecnológica, A.C.

Acta de Examen de Grado

El Secretario Académico del Instituto Potosino de Investigación Científica y Tecnológica, A.C., certifica que en el Acta 037 del Libro Primero de Actas de Exámenes de Grado del Programa de Doctorado en Ciencias Aplicadas en la opción de Nanociencias y Nanotecnología está asentado lo siguiente:

En la ciudad de San Luis Potosí a los 9 días del mes de enero del año 2012, se reunió a las 17:00 horas en las instalaciones del Instituto Potosino de Investigación Científica y Tecnológica, A.C., el Jurado integrado por:

Dr. Braulio Gutiérrez Medina	**Presidente**	**IPICYT**
Dr. Raúl Garibay Alonso	**Secretario**	**UAC**
Dr. Haret-Codratian Rosu Barbus	**Sinodal**	**IPICYT**
Dra. Lina Raquel Riego Ruiz	**Sinodal**	**IPICYT**
Dr. Román López Sandoval	**Sinodal**	**IPICYT**

a fin de efectuar el examen, que para obtener el Grado de:

DOCTORA EN CIENCIAS APLICADAS
EN LA OPCIÓN DE NANOCIENCIAS Y NANOTECNOLOGÍA

sustentó la C.

María del Pilar Monsivais Alonso

sobre la Tesis intitulada:

Introducing Biological Information in the Superparamagnetic Clustering Algorithm of Gene Expression Data

que se desarrolló bajo la dirección de

Dra. Lina Raquel Riego Ruiz
Dr. Haret-Codratian Rosu Barbus
Dr. Román López Sandoval

El Jurado, después de deliberar, determinó

APROBARLA

Dándose por terminado el acto a las 19:00 horas, procediendo a la firma del Acta los integrantes del Jurado. Dando fe el Secretario Académico del Instituto.

A petición de la interesada y para los fines que a la misma convengan, se extiende el presente documento en la ciudad de San Luis Potosí, S.L.P., México, a los 9 días del mes de enero de 2012.

Dr. Marcial Bonilla Marín
Secretario Académico

INSTITUTO POTOSINO DE INVESTIGACIÓN CIENTÍFICA Y TECNOLÓGICA, A.C.
IPICYT
SECRETARIA ACADEMICA

Mtra. Ivonne Lizette Cuevas Vélez
Jefa del Departamento del Posgrado

Acknowledgments

First of all, I would like to thank my advisors, specially PhD Lina Raquel Riego Ruiz for her dedication, guidance, support and friendship during the development of this thesis. In the same spirit, I would like to thank Dr. Haret Codratian Rosu Barbus and Román López Sandoval for their continous encouragement and commitment.

I also want to acknowledge the PhD student Jorge Carlos Navarro Muñoz for his important collaboration in the development of this thesis.

I would like to thank in a special way to my parents, who always have been a support for me in everything, as well as the IPICyT staff and my friends, in particular Pedro Palomares and Jaime Pérez.

My final thanks go to CONACyT for the fellowship (no. 182493) during the years 2006-2008.

Contents

List of Figures

List of Tables

Resumen

Introducción de Información Biológica en el Algoritmo de Clustering Superparamagnético para Datos de Expresión Génica.

Los microarreglos proporcionan información de la actividad a nivel transcripcional de los genes de un organismo, bajo distintas circunstancias. Esto puede llevar al descubrimiento de genes clave en procesos celulares, clasificación molecular de enfermedades o identificar funciones para los genes, entre otras cosas. En el proceso de obtención de esta información, los algoritmos de clustering son una pieza importante al ayudar en la clasificación de los datos provenientes de microarreglos.

En este trabajo modificamos el algoritmo de Clustering Superparamagnético añadiendo un peso extra en la fórmula de interacción que aprovecha la información que se tiene sobre los genes regulados por un mismo factor de transcripción. Con este algoritmo modificado, que nombramos SPCTF, analizamos los datos de microarreglos de Spellman et al. para ciclo celular en levadura (Saccharomyces cerevisiae) y encontramos clusters con un número mayor de integrantes, comparando con el algoritmo original SPC. Algunos de los genes que pudimos incorporar no fueron detectados por Spellman et al. en un principio, pero fueron identificados por otros estudios posteriormente. Otros de los genes que fueron incorporados aún no han sido clasificados, por lo que analizamos los clusters compuestos en su mayoría por estos genes sin identificar con el algoritmo MUSA y esto nos permitió seleccionar aquellos cuyos genes contienen sitios de unión a factores de transcripcion correspondientes a ciclo celular. Estos clusters pueden ser estudiados ahora de manera experimental para descubrir nuevos genes involucrados en el ciclo celular. La idea de introducir la información biológica ya disponible para optimizar la clasificación de genes puede ser implementada para otros algoritmos de clustering.

Palabras Clave: Agrupamiento, Microarreglos, Factores de Transcripción, Ciclo Celular, Levadura

Abstract

Introducing Biological Information in the Superparamagnetic Clustering Algorithm of Gene Expression Data.

Microarray technology allow researchers to examine the transcriptional activity of thousands of genes under different conditions. Microarrays have been used, for example, to discover key genes involved in cellular processes, disease classification, drug development and gene function annotation. Clustering algorithms have become an important step in the microarray data analysis in order to discover biologically relevant information.

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. This combined similarity measure for two genes relies on two types of information: their expression profiles generated by a microarray, and the number of shared transcription factors that have been proved (experimentally) to bind to their promoters. With this modified algorithm which we call SPCTF, we analyze the Spellman et al. microarray data for cell cycle genes in yeast (Saccharomyces cerevisiae), 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 worthy of further experimental studies because they would probably lead to new cell cycle genes. Our idea of introducing the available information about transcription factors to optimize the gene classification could be implemented for other distance-based clustering algorithms.

KEY WORDS: Clustering, Microarrays, Transcription Factors, Cell Cycle, Yeast

Introduction

"To study men, we must look close by;
to study man, we must learn to look afar;
if we are to discover essential characteristics,
we must first observe differences."
Jean-Jacques Rousseau,
Essai sur l'origine des langues, 1781

Machine learning and data mining have naturally appeared in the past years to overcome the necessity of extracting valuable information from large data bases. Actually, with the aid of computers, different kind of data is being acquired and stored at tremendous speed, but unfortunately the employed techniques to analyze those data do not accomplish the same rate of velocity. Consequently, it has become particularly important to develop better and more reliable algorithms to substract information, and also to understand different techniques, its advantages and disadvantages.

Pattern recognition, the act of taking raw data and assign them appropriate labels using algorithms, is often divided in supervised and unsupervised data learning according to the procedure followed. In supervised learning some of the input values are already correctly labeled, and then a classifier is built in such a way that it performs well on the training data and that it could be also generalize as well as possible to new data. In unsupervised learning, no a priori information about the data is known, and a structure or pattern must be found that will help to label the data. In chapter 1, we will briefly describe the main problem in unsupervised data learning, the clustering. In chapter 2, a description of the SPC algorithm is provided, to finally introduce our proposed SPCTF algorithm in chapter 3, where results of both algorithms are compared when applying them to a classic yeast data set.

Chapter 1

Clustering

Contents

1.1 The Clustering Problem

Cluster analysis or clustering refers to the process in which a set of objects is divided into subsets, often called groups or clusters, with similar characteristics. The term was first used by Tryon in 1939 [1], and nowadays encompasses thousands of algorithms and methods proposed for grouping objects. This problem is encountered across all areas of study when the need of organizing observed data into a meaningful structure arises, and thus cluster analysis is used to discover natural categories that can guide the way to possible interpretations or at least perform a classification of those data, useful for future observations.

The capacity of clustering incoming information is essential. For example, in 1676, Antonie van Leeuwenhoek designed a single lens microscope and observed organisms never seen before: bacteria. He and the microbiologists that followed him had to catalog the new living forms, and as they observed spherical, rod and spiral shaped bacteria forming together chains or bunches, they first distinguished them by their shape and named them as cocci, bacilli etc [2]. Later, physiology and biochemistry features were included in addition to morphology, and more recently, features at the genetic level are being used to improve bacteria classification [3]. Other examples involving clustering are the correct diagnosis of diseases according to clustered symptoms [4], classification of antique pieces in archeology [5], sound classification in linguistics [6], face recognition [7] and classification of soil-type in ecology. Romesburg (1984) provides many examples of cluster analysis in diverse areas [9].

The basic steps intervening on cluster analysis are the correct collection and standardization of data, the definition of a similarity or distance measure, the execution of the clustering algorithm, and finally, validation of results (see Figure 1.1). Each of these steps can be done by different techniques and have large influence in the obtained results [10].

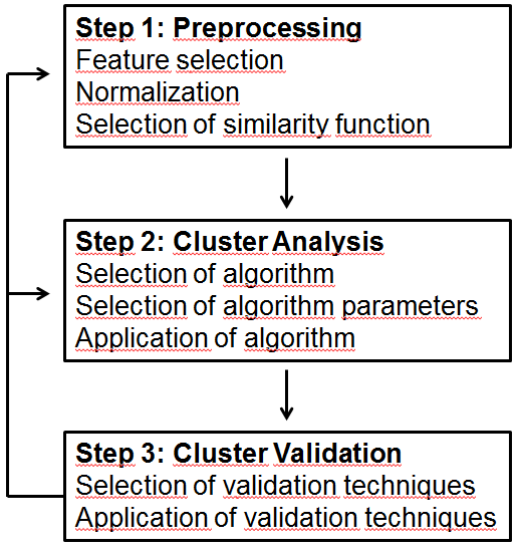


Figure 1.1: Overview of the basic steps in cluster analysis, taken from [11]

The initial data set, once standardized, is always arranged in a matrix, where usually rows stand for the objects to group, and columns are the selected attributes or features that characterize each one of the objects. It is important to mention that features can be either quantitative or qualitative. Weight and height are quantitative features, while gender and skin color, are qualitative. According to Gowda and Diday [12], they can be further subdivided into:

- Quantitative features: *e.g.*, (a) continuous values (*e.g.*, weight), (b) discrete values (*e.g.*, the number of computers), (c) interval values (*e.g.*, the duration of an event).
- Qualitative features: (a) nominal or unordered (*e.g.*, color), (b) ordinal (*e.g.*, military rank or qualitative evaluations of temperature ("cool" or "hot") or sound intensity ("quiet" or "loud")).

Data sets could be entirely quantitative, entirely qualitative or mixed, and consequently, there are several strategies to follow in each case. In this thesis, only quantitative continuous data would be treated.

As an example, we could be interested in grouping people with features as weight and height. Cluster analysis has to find out which persons are similar and in consequence which ones are dissimilar. Once data is collected, it is organized in a matrix as such one presented in Table 1.1:

	Weight	Height
P1	45	155
P2	90	180
P3	55	165
P4	60	160
P5	80	185
P6	85	175

Table 1.1: Example of a data matrix

For such a small feature matrix like this, one can easily find the solution by visually inspecting data, but for a large matrix with hundreds of objects and hundreds of attributes the process must be automatized. Commonly, the similarity is calculated by a mathematical formula that computes a distance for each pair of objects, as objects are represented by vectors in a high-dimensional space spanned by their n attributes. The Euclidean distance coefficient is the most used and measures the absolute distance between two items [13]. It is calculated as follows:

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}. \tag{1.1}$$

The closest vectors (with small coefficient values) are the most similar and must be grouped together. Objects in our example matrix are now represented in a two-dimensional

space, as shown in Figure 1.2, and the corresponding matrix with the Euclidean similarity measure between each pair of objects is also given in table 1.2. (As it is a pair-wise relation, only the inferior diagonal is needed)

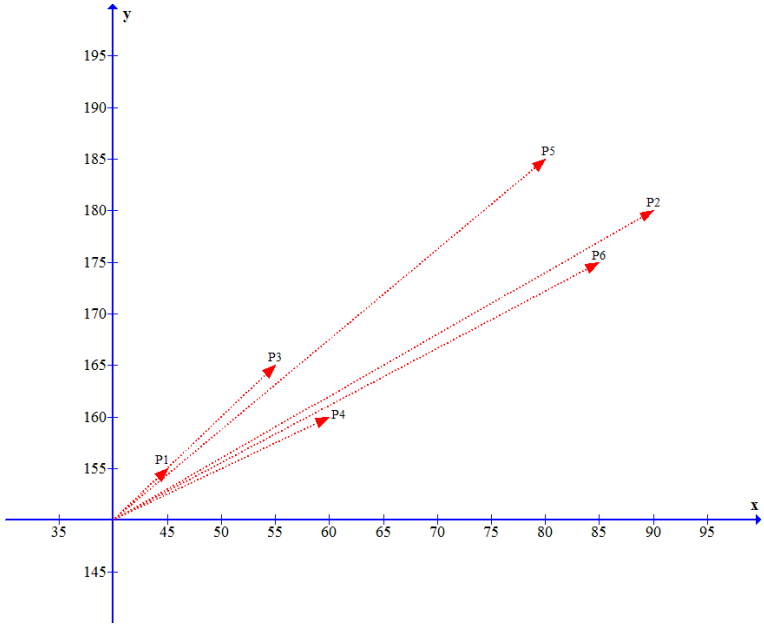


Figure 1.2: Objects from the example matrix represented as vectors in a 2 dimensional space

	P1	P2	P3	P4	P5	P6
P1	0	51.48	14.14	15.81	46.1	44.72
P2	0	0	38.08	36.06	11.18	7.07
P3	0	0	0	7.07	32.02	31.62
P4	0	0	0	0	32.02	29.15
P5	0	0	0	0	0	11.18
P6	0	0	0	0	0	0

Table 1.2: Similarity matrix calculated with Euclidean distance

Besides Euclidean distance, there are other similarity measures based on calculating "distance" between vectors, and these typically fall into two general classes: metric and semi-metric. To be classified as metric, a distance measure d_{ij} between two vectors, x and y, must obey the following rules: the distance must be a positive quantity $d_{xy} \geq 0$, it must be symmetric, $d_{xy} = d_{yx}$ (it is the same distance from x to y as from y to x), the distance between an object and itself is zero $d_{xx} = 0$, and has to follow the triangle inequality: when considering three objects, x, y and z, the distance from x to z is at most as large as the sum of the distance from x to y, and the distance from y to z, $d_{xz} \leq d_{xy} + d_{yz}$. Semi-metric distance measures accomplish the first three consistency rules but they fail to maintain the triangle inequality [14]. There are a large number of metric and semi-metric distance coefficients (see [14], [15] for a review) and next some of the most used (not necessary the best ones) distance coefficients are described.

1. The Euclidean squared distance metric is similar to Euclidean distance metric, the only difference being that it does not take the square root and in consequence, it is faster to compute but tends to give more weights to the outliers [11]. The output of some clustering algorithms is not affected if the Euclidean distance is replaced with Euclidean squared, as in the case of K-means. However, the output of hierarchical clustering is likely to change.

2. The Pearson distance is defined as $d_P = 1 - r$, where r is the Pearson correlation coefficient (also known as the centered Pearson correlation coefficient). For two vectors, x and y, representing items described by n attributes, this coefficient is calculated from the sample values and their standard deviations and it is given by:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n})}\sqrt{(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n})}}, \tag{1.2}$$

   This coefficient measures the strength of the linear relationship between two variables, ranging between -1 and $+1$. When the two variables have a perfect correlation, *i.e.*, if the variable x increases, then also y increases, r will take the value $+1$. Instead, when it is -1, the variables are perfectly opposite or anticorrelated, *i.e.* when x increases, then y decreases. A zero value means that there is absolutely no correlation between them. If x and y are conceptualized as vectors in a N-dimensional space, Pearson coefficient tells us how large is the angle between them, and if x and y are plotted as independent curves, r measures how similar the shapes of the two curves are. This is because the correlation coefficient is invariant under scalar transformation, and thus it will assign a correlation of $+1$ for curves different in magnitude but with identical shapes [13].

   The uncentered Pearson correlation coefficient is calculated basically with the same equation, but taking sample means as 0, even when they are not. In contrast, it does not assign a correlation of $+1$ for two objects with the same shape curve if they are shifted. It may be appropriate if there is a zero reference state.

   If the above mentioned coefficients are used, anticorrelated objects will be clustered

apart from correlated ones. Two other similarity metrics which involve absolute values of the previous correlation functions, the squared or absolute Pearson correlation coefficients (centered and uncentered), take values in the range 0 and 1 and assign the value $+1$ both to correlated and anticorrelated objects [16, 17].

3. City block or Manhattan distance (also known as boxcar distance or absolute value distance). It calculates the distance between two points if a grid-like path is followed, in which only movements along axes and right angles are allowed. The name comes from the distance one would travel in crossing a large city, such as Manhattan, in which the streets are laid out in a regular, rectangular grid. The Manhattan distance is calculated as the sum of the absolute distances between the components of each vector:

$$\sum_{i=1}^{n} |x_i - y_i|. \tag{1.3}$$

   In most cases, this metric distance measure yields results similar to the simple Euclidean distance [18].

4. Jackknife correlation. It was proposed by Heyer *et al.* [19] to cluster gene expression data, improving Pearson correlation, which tends to assign a high value to two unrelated objects if they display a very big or short value in the same component. Jackknife correlation computes several correlation values between two objects, but for each value it leaves out one of the total n dimensions. The final measure for the correlation corresponds to the minimum value found. The Jacknife distance for the pair i,j is defined as:

$$d_{ij} = 1 - min(\rho_{ij}^{(1)}, ..., \rho_{ij}^{(n)}) \tag{1.4}$$

   where $\rho_{ij}^{(l)}$ denotes the Pearson correlation of the pair i,j computed with the l-th component deleted. It could be used for other distance measures, not only with Pearson correlation, thus one could encounter Jacknife Euclidean distance or Jacknife Manhattan distance.

5. The Spearman rank correlation between two items is calculated using the Pearson correlation equation, but instead of using standard deviations, Spearman correlation uses difference in ranks. Thus, each data value is replaced by its rank once each component vector has been ordered (tied scores are given an average rank). As in the case of the Pearson correlation a distance measure can be defined corresponding to the Spearman rank correlation as: $d_S = 1 - r_S$, where r_S is the Spearman rank correlation [20].

6. Cosine angle metric measures the angular separation of two vectors in n dimension. Formally, it is defined as follows:

$$d_{ij} = 1 - \frac{\sum_{i=i}^{n} x_i y_i}{|x||y|} \tag{1.5}$$

One important property of vector cosine angle is that it gives a metric of similarity between two vectors unlike Manhattan distance and Euclidean distance, both of which give metrics of dissimilarities [14].

Other distance metrics have been recently proposed, such as Tukey's biweight [21, 22], however, a best distance measure does not exist as yet. Commonly, a distance measure which performs well for some kind of clustering problem does not perform well for others. The choice should depend on each particular set studied.

1.2 Clustering Techniques

1.2.1 Hierarchical Clustering Algorithm

Along with K-means algorithm, these methods were of the first to appear, and they are still in widespread usage. Typically, they are separated in two types: agglomerative (bottom-up) and divisive (top-down), being the most common agglomerative algorithms. In that approach, each object is taken as an individual cluster at the beginning and the distance between every pair is computed according to a chosen metric. Next, the two objects with the minimum distance between them are merged in one cluster, and the distances between this new cluster and the other objects that remain unchanged are recalculated. This process is iteratively done until all elements finished up into one single cluster. On the contrary, in a divisive approach all objects are placed together in one cluster at the beginning, and then they are iteratively separated into two clusters.

At the end of both approaches, the entire clustering process is graphically displayed as a tree-like structure called dendrogram (see Figure 1.3), in which each node represents two joined items and the length of the corresponding branches is proportional to objects similarity, *i.e.*, the shortest branches are assigned for the most similar clusters and branch length increases as similarity decreases [23].

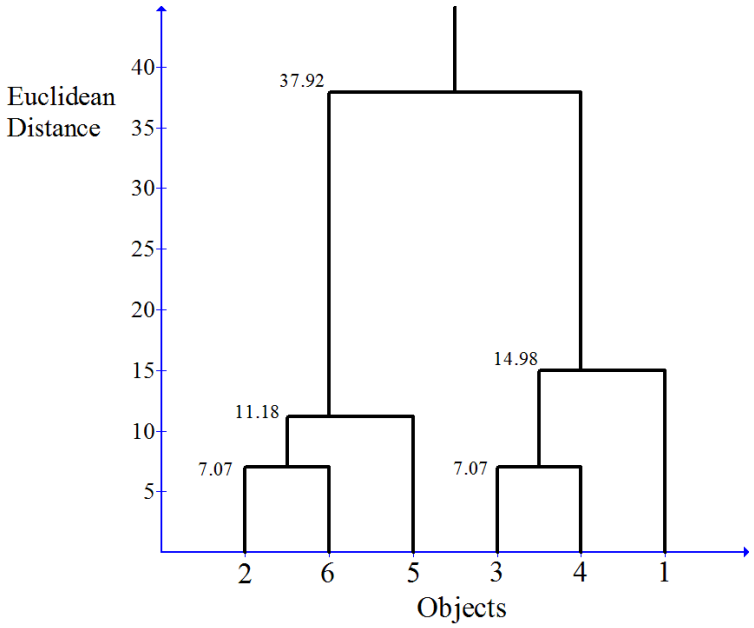


Figure 1.3: Dendogram for our weight-height example matrix, calculated by an agglomerative hierarchical clustering.

The determination of the final cluster division is set arbitrarily by cutting the tree at a certain level or height, which is equivalent to putting a threshold on the pairwise distance between clusters. The memory complexity of hierarchical clustering is quadratic in the number of objects n to be grouped, $O(n^2)$, which limits the size of data sets that can be analyzed [24].

The estimation of proximity between two clusters is the key element for this algorithm, as in every step the two closest clusters have to be merged. These are four of various possible methods:

1. Single linkage or MIN [25]: The distance between two clusters is taken as the distance between the two closest data points, provided that each point belongs to a different cluster.

2. Complete linkage or MAX [26]: The distance between clusters is the same as that between the two furthest points, each point belonging to a different group. As in single linkage, once the distance matrix is known, no more new distances need to be calculated.

3. Average linkage [27]: Inter-cluster distance is defined as the average of the distances between all pairs of points between the two clusters .

4. Centroid linkage [27]: A cluster centroid is defined as the point whose coordinates are the average values of all items forming the cluster. Thus, in this case, the distance between clusters is that calculated between their centroids. Figure 1.3 is a representation of the three above mentioned methods.

5. Ward's method [28]: at each step of the analysis, the union of every possible cluster pair is considered but the ones that are finally merged are those ones that minimize the variance associated with the merging. This variance is called "information loss" and is defined by Ward as the change in the sum-of-squared error, ESS, before and after the union. In this way, Ward's method assesses the quality of the merged clusters at each step of the agglomerative procedure.

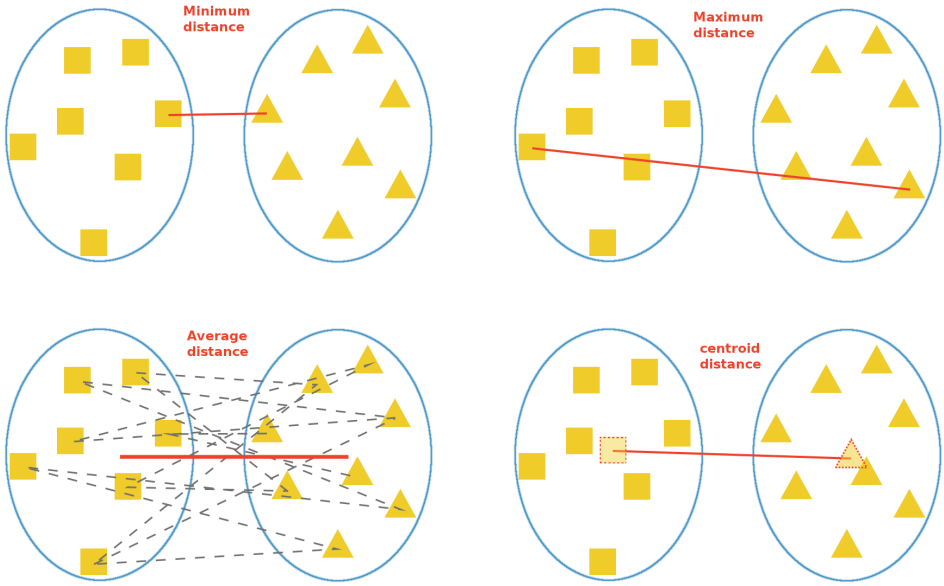


Table 1.3: Graphical Representation of single, complete, average and centroid linkage

The selection of a distance measure between clusters will affect the overall clustering result. For example, single linkage tends to produce long chains of points, while complete linkage tends to produce compact clusters, and both single and complete linkage are sensitive to outliers. Centroid linkage it is not appropriated when the distance metric is based on the Pearson correlation because no normalization is included when the cluster centroid is calculated, whereas the vectors are implicitly normalized when the Pearson correlation-based distance is calculated [29]. Wards method, although less well known, often produces the most satisfactory results [30].

It is worth mentioning that agglomerative clustering is good at identifying small clusters, but can provide sub-optimal performance for identifying a few large clusters. Conversely, divisive methods are good at identifying a few large clusters, but weaker at many small clusters. Chipman *et al.* [31] have proposed a hybrid hierarchical clustering method that combines the strengths of both approaches, modifying top-down procedures with information gained from a preliminary bottom-up clustering.

1.2.2 K-means

K-means clustering idea first appeared in 1957 [32] and it is still one of the most popular partitioning method due to its simplicity, although the number of clusters K in the data is needed as an input for the algorithm, information that is not commonly encountered for real data sets.

It proceeds by proposing K initial vectors at random as ”cluster centers”. Next, each vector from data is assigned to the closest of these initially proposed ”seeds”, which are then updated to be the mean of their constituent vectors. This process is iteratively done, refining more and more the cluster centers. It converges when there is no further change in assignment of vectors to clusters (*i.e.*, the cluster center remains stationary), or when the given number of iterations is exceeded [33].

As the number of clusters K has to be proposed arbitrarily by the user, some algorithms to locate the best value of K has been proposed. Also, this algorithm suffers from the problem that it can yield very different results depending on the initial vectors . It typically converges to one of the many local optima, rather than the global optimum [34]. Hartigan and Wong [35] give a more complicated algorithm which is more likely to provide a good local optimum. Whatever algorithm is used, it is advisable to repeatedly start the algorithm with different initial values, increasing the chance that a good local optimum is found [36].

1.2.3 SOM Clustering

A Self-Organizing Map, SOM, ([37]) is a type of artificial neural network that is trained to produce a lower dimensional representation (one-dimensional or two-dimensional) for the input space of data vectors. This makes SOM also useful for visualizing high-dimensional data with low-dimensional views. It consists of components called nodes or neurons, which are the intersections of a two-dimensional grid (usually of hexagonal or rectangular geometry). The dimension of the grid (*e.g.*, lattice of 6x5 nodes) needs to be specified a priori and each node represents a reference or weight vector (similar to the mean vectors in the K-means algo-

rithm) of the same dimension as the input data vectors and also a position in the map space. The nodes compete for representation of the samples changing themselves by learning to become more like samples. It is this selection and learning process that makes the weights organize themselves into a map representing similarities [38].

The algorithm begins by selecting a vector from original data and its Euclidean distance to all weight vectors is computed. The node with weight vector most similar to the input vector is called the best matching unit (BMU), and its weight and also the weight of its neighboring neurons in the SOM lattice are adjusted in order to become more similar to the selected sample vector as a reward. This process is iteratively repeated, but the number of neighbors and how much each weight can learn decreases over time. A number of different methods can be used to determine which weights are considered as neighbors: concentric squares, hexagons, Gaussian function, and the learning function or how much each node can become more like the sample vector, is based on their distance from the BMU and time.

The intuition for this learning process is that the reference vectors that are close enough to vector data will be pulled towards it, and the stiffness of the grid structure will propagate some of impact to neighbouring nodes. As a result, a reference vector is pulled more towards input vectors that are closer to the reference vector itself and is less influenced by the input vectors located further away. In the meantime, this adaptation procedure of reference vectors is reflected on the output nodes (nodes associated with similar reference vectors are pulled closer together on the output grid). The algorithm terminates when convergence of the reference vectors is achieved or after completing a pre-defined number of training iterations. The more neighbors you use the better similarity map you will get, but the number of distances the algorithm needs to compute increases exponentially [39]. Figure 1.4 is a visual representation of SOM.

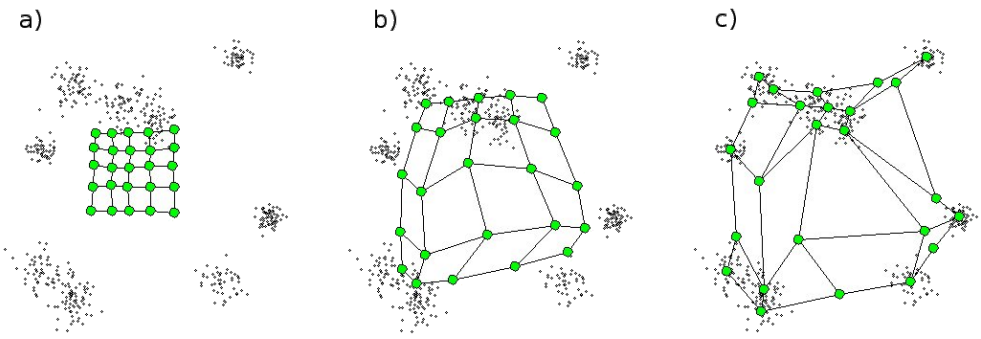


Figure 1.4: Representation of a SOM lattice adjustment to clusters data. Modified from [40]

It has been shown that while self-organizing maps with a small number of nodes behave in a way that is similar to K-means, larger self-organizing maps rearrange data in a way that is fundamentally topological in character. Choosing the geometry of the output grid is not as crucial as the choice of the number of clusters for K-means method, but the initial choice of nodes also influences the final clustering result. A good way to seed the reference vectors is to use the result from a principal component analysis (PCA) [30]. Also, a lot of maps can be

constructed in order to get the best one.

1.2.4 SOTA Clustering

The Self-Organising Tree Algorithm (SOTA) [41] combines the advantages of hierarchical and SOM clustering techniques, being a divisive (top-down) clustering method and also a neural network that grows adopting the topology of a binary tree.

SOTA starts the classification with a root node with two leaves (cells), whose reference vectors are updated through some neighbourhood weighting parameters, every time an input data vector is assigned to the most similar leaf. When all data vectors are associated with the leaf whose reference vector is located closest, the algorithm proceeds by creating two new leaves, expanding the tree from the node with the most heterogeneous population (which do not necessarily correspond to the more populated node). The heterogeneity is measured using a dispersion value, defined as the mean value of the distances between the node's own vector and the data vectors associated with the node. The values of the two new leaves are identical to the node that generate them and the whole procedure starts again [41].

If the growth of the network is not restricted, the outcome would be a binary tree which contains only one profile in each leaf, but generally, the growth ends up when the maximum dispersion value among all the terminal nodes reaches a certain threshold set by the user. A criterion to stop the growing of the tree based on the approximate distribution of probability obtained by randomization of the original data set, can be used [42]. Figure 1.5 is a schematic representation of the SOTA algorithm.

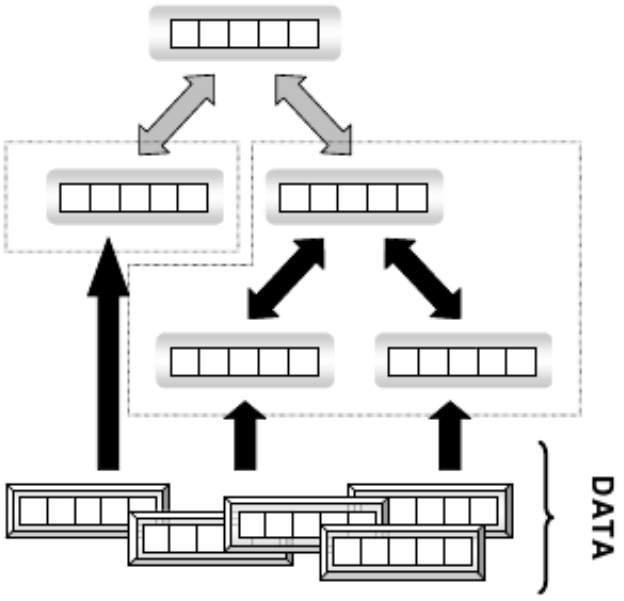


Figure 1.5: In SOTA, the cluster with most variability is divided in two new leaves. Taken from [43]

The algorithm generates a hierarchical cluster structure at the desired level of resolution, allowing a visual representation of the clusters. The number of nodes in SOTA is not fixed from the beginning (in contrast to SOM), and the obtained clustering is proportional to the heterogeneity of the data, instead of the number of items (SOTA is distribution preserving while SOM is topology preserving). Besides, the final structure can be asymmetrical, including branches with different numbers of nodes, which represent the averages of the patterns contained in the clusters [43].

Since SOTA runtimes are approximately linear with the number of items to be classified, it is especially suitable for dealing with huge amounts of data and it is one of the fastest available algorithms for performing hierarchical clustering. Moreover, since the comparison operations are performed between the data and the average profiles in the nodes, the absence of some points (missing values) in a vector corresponding to a particular data will have a negligible effect on the whole process of the network training. This renders unnecessary the use of methods for estimating missing values, required if average linkage or similar methods are used [44].

1.2.5 Fuzzy Clustering

Above mentioned algorithms assign each data object only to one cluster, thus they are unable to identify data similar to multiple distinct groups. In contrast, fuzzy clustering allows each data to belong to two or more clusters, by assigning them a degree of belonging to clusters, as in fuzzy logic. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster.

According to Ali *et al.*, [45], fuzzy clustering algorithms are broadly classified into two groups: i) Classical and ii) Shape-based. There exist many classical fuzzy clustering algorithms in the literature, among the most popular and widely used being: i) Fuzzy C-means (FCM), ii) Suppressed fuzzy C-means (SFCM), iii) Possibilistic C-means (PCM), and (iv) Gustafson-Kessel (GK), while from a shape-based fuzzy clustering viewpoint, well established and popular algorithms include: i) Circular shape-based, ii) Elliptical shape-based, and (iii) Generic shape-based techniques.

Let X_i be object i of the input data set, where $i = 1, ..., n$, such that n is the total number of input vectors. Any object X_i has a set of coefficients u_{ij}, $j = 1, ..., k$, where k is the total number of possible groups, that represent the degree of membership between data X_i and cluster j. These coefficients are numbers between 0 and 1 and are written in a matrix named U, with columns accounting for the clusters and rows for the objects. This degree of belonging, u_{ij}, is related inversely to the distance from X_i to the center of each cluster and it also depends on a parameter m, any real number greater than 1, that controls how much weight is given to the closest center [46].

The algorithm is composed of the following steps:

1. Choose a number of clusters and randomly initialize matrix U

2. For each step, calculate the center vectors C_j. With fuzzy C-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$C_j = \frac{\sum_{i=1}^{n} u_{ij} x_i}{\sum_{i}^{n} x_i} \tag{1.6}$$

3. Update U

$$u_{ij} = \frac{1}{\sum_{j=1}^{k} (\frac{D_{ij}}{D_{kj}})^{\frac{2}{m-1}}} \tag{1.7}$$

   where D_{ij} is the distance between data X_i and cluster centre vector C_j

4. Repeat from step 2 until the algorithm converges, that is when the values of the coefficients do not present modifications more than a given sensitivity threshold ε between two iterations.

Fuzzy C-means minimizes intra-cluster variance as well, but has the same problems as K-means: the minimum is a local minimum, and the results depend on the initial choice of weights. Other recent proposed algorithms with better results can be reviewed in [47].

1.2.6 Biclustering

Clustering can be applied to either the rows or the columns of the data matrix, separately. Biclustering, first mentioned by Cheng and Church [48] in gene expression data analysis, refers to a distinct class of clustering algorithms that perform simultaneous row and column clustering. A bicluster is defined thus as a submatrix spanned by a set of rows and a set of columns (compare Figure 1.6) [49].

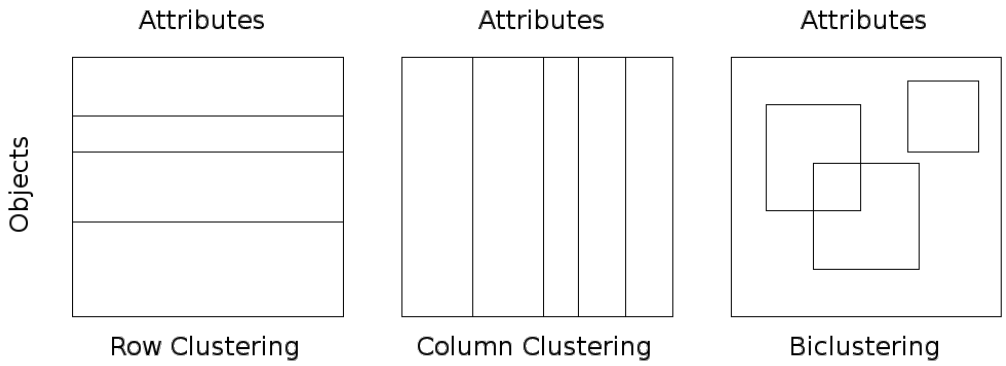


Figure 1.6: In contrast to row and column clustering, biclustering can group a subset of objects and a subset of attributes or features. Taken from [49]

Along with the Cheng-Church algorithm, another of the earliest biclustering formulations is that introduced by Hartigan [33], also known as block clustering. More recently, coupled two-way clustering (CTWC), [50] defines a generic scheme for transforming a one-dimensional

clustering algorithm into a biclustering algorithm. The algorithm relies on having a one-dimensional (standard) clustering algorithm that can discover significant or stable clusters. Given such an algorithm, the coupled two-way clustering procedure will recursively apply the one-dimensional algorithm to submatrices. For a more detailed review of biclustering algorithms see [49].

We can then conclude that, unlike clustering algorithms, biclustering algorithms identify groups of objects that show similarity under a specific subset of features [51].

1.2.7 Model Based Clustering

Model Based Clustering assumes that the data is generated by a finite mixture of underlying probability distributions such as multivariate normal distributions. In this case, each cluster k is represented by a multivariate Gaussian model f_k in d dimensions:

$$f_k(y_j|\mu_k, \sum_k) = \frac{e^{-1/2(y_j-\mu_k)^T \sum_k^{-1}(y_j-\mu_k)}}{\sqrt{(det(2\pi\sum_k)}}, \tag{1.8}$$

where y_j is an object and μ_k and $\sum_k$ are the mean and covariance matrix of the multivariate normal distribution, respectively [52].

The covariance matrix $\sum_k$ can be represented by its eigenvalue decomposition, which in general takes the following structure:

$$\sum_k = \lambda_k D_k A_k D_k^T, \tag{1.9}$$

where D_k is the orthogonal matrix of the eigenvectors of $\sum_k$, A_k is a diagonal matrix whose elements are proportional to the eigenvalues of $\sum_k$, and λ_k is the constant of proportionality. This decomposition implies a nice geometric interpretation of the clusters: D_k controls the orientation, A_k controls the shape, and λ_k controls the volume of the cluster. Note that simpler forms of the covariance structure can be used (e.g., by having some of the parameters take the same values across clusters), thereby decreasing the number of parameters that have to be estimated but also decreasing the model flexibility (capacity to model more complex data structures).

The mixture model p itself takes then the following form:

$$p(y_j) = \sum_{k=1}^{K} \pi_k p_k(y_j|\mu_k, \sum_k), \tag{1.10}$$

where K is the total number of clusters and π_k is the prior probability that an object belongs to cluster k so that:

$$\sum_{k=1}^{K} \pi_k = 1. \tag{1.11}$$

In practice we would like, given a collection of objects $y_j (j = 1, ..., n)$, to estimate all the parameters (π_k, μ_k, $\sum_k$ ($k = 1, ..., K$), and K itself) of this mixture model. In a first step they are estimated with an Expectation Maximization algorithm (EM algorithm), using a fixed

value for K and a fixed covariance structure [36]. In the EM algorithm, the Expectation steps and Maximization steps alternate. In the E step, the probability of each object belonging to each cluster is estimated conditionally on the current parameter estimates. In the M step the model parameters are estimated given the current group membership probabilities. When the EM algorithm converges, each object is assigned to the group with the maximum conditional probability [52]. This parameter estimation is then repeated for different values for K and different covariance structures. The result is thus a collection of different models fitted to the data and all having a specific value for K and a specific covariance structure. In a second step the best model in this group of models is selected (i.e., the most appropiate number of parameters and a covariance structure is chosen here). This model selection step involves the calculation of the Bayesian information criterion (BIC) for each model [53], which is not further discussed here.

A good implementation for model based clustering (called MCLUST) is available at www.stat.washington.edu/fraley/mclust. Yeung et al. reported good results using this software on several synthetic data sets and real expression data sets. McLachlan et al. [54] have also implemented model-based clustering in a Fortran program called EMMIX, which is also freely available from the web at http://www.maths.uq.edu.au/g̃jm/emmix/EMMIX.f.

1.2.8 Quality-Based Algorithms

Quality-based algorithms produce clusters with a quality guarantee that ensures that all members of a cluster are coexpressed (this property is called transitivity). This concept was introduced by Heyer, Kruglyak and Yooseph, ([19]) and their implementation is called QT_Clust. The quality of a cluster k is defined as a diameter (equal to $1 - min_i, j \in s_{ij}$, where s_{ij} is the jackknife correlation between expression profile i and j), but the method can be easily extended to other definitions.

The algorithm considers every object in the data set as a cluster seed (one could also call this a cluster center) and iteratively assigns the objects to these clusters that cause a minimal increase in diameter until the diameter threshold, i.e., quality guarantee, is reached. Note that at this stage every object is made available to every candidate cluster and that there are as many candidate clusters as there are objects. At this point, the candidate cluster that contains the most objects is selected as a valid cluster and removed from the data set whereafter the whole process starts again. The algorithm stops when the number of points in the largest remaining cluster falls below a threshold. Note that this stop criterion implies that the algorithm will terminate before all objects are assigned to a cluster.

This approach has some advantages, for example it is possible to find clusters containing highly correlated objects, and these clusters might therefore be good seeds for further analysis. Moreover, objects not really correlated with other members of the data set are not included in any of the clusters. Some disadvantages are that the quality guarantee of the clusters is a user-defined parameter hard to estimate, it is hard to use by biologists, needs extensive parameter fine-tuning, and produces clusters all having the same fixed diameter not optimally adapted to the local data structure [36].

1.2.9 Adaptive Quality-Based Clustering

Adaptive quality-based clustering ([55]) consists of a two-step approach. In the first step, a quality-based approach is performed to locate a cluster center in an area where the density of objects is locally maximal using a preliminary estimate of the radius (*i.e.*, the quality) of the cluster. In the second step, called adaptive step, the algorithm re-estimates the radius of the cluster so that the objects belonging to it are, in a statistical sense, significantly correlated. To this end, a bimodal and one-dimensional probability distribution (the distribution consisting of two terms: one for the cluster and the other for the rest of the data) describing the Euclidean distance between the data points and the cluster center is fitted to the data using an expectation-maximization (EM) algorithm. Finally, step one and two are repeated, using the re-estimation of the quality as the initial estimate needed in the first step, until the relative difference between the initial and re-estimated quality is sufficiently small. The cluster is subsequently removed from the data and the whole procedure is restarted. Note that only clusters whose size exceeds a predefined number are presented to the user.

In adaptive quality-based clustering, users have to specify a threshold for quality control. This parameter has a strict statistical meaning and is therefore much less arbitrary (in contrast to the case in QT_Clust). It can be chosen independently of a specific data set or cluster and it allows for a meaningful default value (95%) that in general gives good results. This makes the approach user friendly without the need for extensive parameter fine-tuning. Furthermore, with the ability to allow the clusters to have different radius, adaptive quality-based clustering produces clusters adapted to the local data structure[36]. However, the method has some limitations like it does not converge in every situation. An application of Adaptive Quality-Based Clustering to nervous system is found in [56] and a server running the program is available at http://homes.esat.kuleuven.be/̃thijs/Work/Clustering.html

1.2.10 Some Physics Related Algorithms

There are also several physics related clustering algorithms, e.g. Deterministic Annealing [57] and Coupled Mass [58]. Deterministic Annealing uses the same cost function as K-means, but rather than minimizing it for a fixed value of clusters K, it performs a statistical mechanics type analysis, using a maximum entropy principle as its starting point. The resulting free energy is a complex function of the number of centroids and their locations, which are calculated by a minimization process. This minimization is done by lowering the temperature variable slowly and following minima that move and every now and then split (corresponding to a second order phase transition). Since it has been proven [59] that in the generic case the free energy function exhibits first order phase transitions, the deterministic annealing procedure is likely to follow one of its local minima [60].

Finally, it is important to stress that clustering methods have been used in a large variety of scientific disciplines and applications that include pattern recognition [61], learning theory [62], astrophysics [63], medical images and data processing [64], machine translation of text [65], satellite data analysis [66], as well as speech recognition [67].

1.3 Clustering of Gene Expression Data

With recent technology advances, biology researchers continually generate a growing amount of data at many levels, from genome sequences to protein structures and protein-protein interactions. In consequence, clustering algorithms have had an important impact as they have become an indispensable exploratory technique for many studies. Cluster analysis has been used to discover association between a gene behaviour and an outcome, annotation of functional genomics, discovering of protein families based on sequence similarity, physical location of transcription factor binding sites in a genome, discovering off drug targets, relating gene expression to chromosome location, phylogenetics and phylogenomics.

Gene expression studies are a particularly active area for clustering. Using microarray chips, it is possible to monitore the amount of mRNA produced by each gene in a cell, for several conditions or under many time points. Microarray data thus can provide a global view of the whole activity of all genes in a genome [1]. The outcome of this type of study is an mxn expression matrix with the m rows corresponding to genes and the n columns corresponding to different time points or different conditions. The expression matrix represents intensities of hybridization signals as provided by a DNA array. In reality, expression matrices usually represent transformed and normalized intensities rather than the raw intensities obtained as a result of a DNA array experiment. The element I_{ij} of the expression matrix represents the expression level of gene i in the experiment j; the entire ith row of the expression matrix is called the expression pattern of gene i. One can look for pairs of genes in an expression matrix with similar expression patterns, which be manifested as two similar rows. Therefore, if the expression patterns of two genes are similar, there is a good chance that these genes are somehow related, that is, they either perform similar functions or are involved in the same biological process. Accordingly, if the expression pattern of a newly sequenced gene is similar to the expression pattern of a gene with known function, a biologist may have reason to suspect that these genes perform similar or related functions [68]. Another important application of expression analysis is in the deciphering of regulatory pathways; similar expression patterns usually imply coregulation, hence grouping genes expression profiles provides a means for understanding gene function, gene regulation, and cellular processes [69].

For example, we can study a dataset matrix with 4000 genes (rows) and 100 lymphoma cancer profiles (columns), each matrix element representing the mRNA amount or expression level of a specific gene on certain profile. Clustering columns allows us to uncover various cancer subtypes based upon similarities between cancer specimens. Understanding the differences between cancer subtypes on a genetic level is crucial to understanding which types of treatments are most likely to be effective [70]. Alternatively, we can cluster rows or gene profiles. Similar row patterns would reveal information about genes whose products function together in pathways performing complex functions in the organism. The study of these pathways and their relationships to one another can then be used to build a complete model of the cell and its functions, bridging the gap between genetic maps and living organisms [71]. For more information on biology concepts and microarray technology see appendixes A and

[1] Expression analysis studies implicitly assume that the amount of mRNA (as measured by a DNA array) is correlated with the amount of its protein produced by the cell. We emphasized that number of processes affect the production of proteins in the cell (transcription, splicing, translation, post-transaltional modifications, protein degradation, etc.) and therefore this correlation may not be straight forward, but it is still significant.

B.

Many algorithms have been used for gene expression data analysis [69]. For example, Eisen *et al.* applied a variant of the hierarchical average-link clustering algorithm to identify groups of co-expression yeast genes [72]. Tamayo *et al.* used SOM to identify clusters in the yeast cell cycle and human hematopoietic differentiation data sets[73]. Tavazoie *et al.* used k-means method to analyze microarray data generated from studies of the yeast cell cycle [74]. In the next sections, we will focus on a particular clustering algorithm and a proposed improvement for it, and we will compare their perfomance when analyzing cell cycle yeast expression profiles.

Bibliography Chapter 1

[1] R. C. Tryon, *Cluster Analysis*, Oxford, England, Edwards Brothers, 1939.

[2] J. R. Porter, *Antony van Leeuwenhoek: tercentenary of his discovery of bacteria*, Oxford, England, Edwards Brothers 1939 Bacteriol Rev., Vol. 40(2), 1976, p. 260.

[3] K.H. Schleifer, *Classification of Bacteria and Archaea: past, present and future. Systematic and applied microbiology*, Elsevier, Vol. 32(8), 2009, p. 533.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M.A. Caligiuri, C. D. Bloomfield and E. S. Lander, *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*, Science, Vol. 286(5439), 1999, p. 531.

[5] F. X. Ricaut and M. Waelkens *Cranial discrete traits in a Byzantine population and eastern Mediterranean population movements*, Human biology an international record of research, Vol. 80(5), 2008, p. 535.

[6] M. Wieling and J. Nerbonne, *Hierarchical Spectral Partitioning of Bipartite Graphs to Cluster Dialects and Identify Distinguishing Features*, Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, p. 33.

[7] R. Chellappa, P. Sinha, and P. J. Phillips, *Face Recognition by Computers and Humans*, IEEE Computer, Vol. 43(2), 2010, p. 46.

[8] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys and E. Lambin *Digital Change Detection Methods in Ecosystem Monitoring: a review*, Int. J. Remote Sensing, Vol. 25(9), 2004, p. 1565.

[9] H. C. Romesburg, *Cluster Analysis for Researchers*, 1984.

[10] A. K. Jain, M. N. Murty and P.J. Flynn, *Data Clustering: A Review* ACM Computing Surveys, Vol. 31(3), 1999, p.264

[11] J. Handl, J. Knowles and D.B. Kell, *Cluster Validation in Post-genomic Data Analysis* Bioinformatics, Vol. 21(15), 2005, p. 3201.

[12] K.C. Gowda and E. Diday, *Symbolic Clustering Using a New Dissimilarity Measure*, Pattern Recognition, Vol. 24(6), 1991, p. 567.

[13] J. Quackenbush *Computational Analysis of Microarray Data*, Nat Rev Genet., Vol. 2(6), 2001, p.418.

[14] S. Draghici *Data Analysis Tools for DNA Microarrays*, Chapman and Hall/CRC Mathematical Biology and Medicine Series, Boca Ratón, USA, 2003.

[15] Everitt B. S., Landau S., Leese M. and Stahl D., *Cluster Analysis* Wiley

[16] `rana.lbl.gov/manuals/ClusterTreeView.pdf`

[17] `http://www.camo.com/resources/clustering.html`

[18] `cptweb.cpt.wayne.edu/miR-AT/miRAT-Help.pdf`

[19] L.J. Heyer, S. Kruglyak and S. Yooseph, *Exploring Expression Data: Identification and Analysis of Coexpressed Genes*, Genome Res., Vol. 9, 1999, p. 1106.

[20] M. H. Fulekar, *Bioinformatics: applications in life and environmental sciences*, Springer, 2009

[21] J. Hardin, A. Mitani, L. Hicks and B. VanKoten, *A Robust Measure of Correlation Between Two Genes on a Microarray* BMC Bioinformatics, Vol. 8, 2007, p. 220.

[22] K. Kim, S. Zhang, K. Jiang, L. Cai, I.B. Lee, L.J Feldman and H. Huang, *Measuring Similarities between Gene Expression Profiles Through New Data Transformation*, BMC Bioinformatics Vol. 8, 2007, p. 29.

[23] P. N. Tan, M. Steinbach and V. Kumar *Introduction to Data Mining* Addison-Wesley, 2005

[24] Y. Loewenstein, E. Portugaly, M. Fromer and M. Linial, *Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space* Bioinformatics, Vol. 24(13), 2008, p. i41.

[25] K. Florek, J. Lukaszewicz, J. Perkal and S. Zubrzycki, *Sur la Liaison et la Division des Points d'un Ensemble Fini*, Colloquium Mathematicae, Vol. 2, 1951, p. 282.

[26] T. Sorensen, *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danis commons*, Biol Skrifter, Vol.5, 1948, p. 1.

[27] R.R. Sokal and C.D. Michener, *A Statistical Method for Evaluating Systematic Relationships*, University of Kansas Science Bulletin, Vol. 38, 1958, p. 1409.

[28] J. H. Ward, *Hierarchical Grouping to Optimize an Objective Function*, Journal of the American Statistical Association, Vol. 58, 1963, p. 236.

[29] J. Hwan Do and D.K. Choi, *Clustering Approaches to Identifying Gene Expression Patterns from DNA Microarray Data*, Mol.Cells, Vol. 25(2), 2008, p.279.

[30] Q. Sheng, Y. Moreau, F. De Smet, K. Marchal, B. De Moor, *Advances in Cluster Analysis of Microarray data*, in Chapter 10 of Data analysis and Visualization in Genomics and Proteomics, (Azuaje F., and Dopazo J., eds.), Jonh Wiley and Sons Ltd. (Chichester, UK), 2005, p. 153.

[31] H. Chipman and R. Tibshirani, *Hybrid Hierarchical Clustering with Applications to Microarray Data*, Biostatistics, Vol. 7(2), 2006, p. 286.

[32] H. Steinhaus, *Sur la division des corps matériels en parties*, Bull. Acad. Polon. Sci., Vol. 4(12), 1957, p. 801.

[33] J. A. Hartigan, *Clustering Algorithms*, John Wiley and Sons, New York, 1975, p. 351.

[34] G.W. Milligan, *An Examination of the Effect of Six Types of Error Perturbation on fifteen Clustering Algorithms*, Psychometrika, Vol. 45(3), 1980, p. 325.

[35] J. A. Hartigan and M. A. Wong, *Algorithm AS 136: A k-means clustering algorithm*, Applied Statistics Vol. 28(1), 1979, p. 100.

[36] Y. Moreau, F. De Smet, G. Thijs, K. Marchal, B. De Moor, *Functional Bioinformatics of Microarray Data: From Expression to Regulation*, Proceedings of the IEEE, Vol. 90(11), 2002, p. 1722.

[37] T. Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, Vol. 30, 1995; Third, extended edition, New York, 2001.

[38] `http://davis.wpi.edu/~matt/courses/soms/#Introduction`

[39] R. Sharan, R. Elkon, R. Shamir, *Cluster Analysis and its Applications to Gene Expression Data*, Ernst Schering Research Foundation Workshop, Vol. 38: Bioinformatics and Genome Analysis, Editors: H.-W. Mewes, H. Seidel, B. Weiss, Springer-Verlag, Berlin Heidelberg, 2002, p. 83.

[40] `blog.peltarion.com/2007/06/13/the-self-organized-gene-part-2/`

[41] J. Dopazo and J. M. Carazo, *Phylogenetic Reconstruction Using an Unsupervised Growing Neural Network That Adopts the Topology of a Phylogenetic Tree*, J. Mol. Evol., Vol. 44(2), 1997, p. 226.

[42] `http://bioinfo.cipf.es/babelomicstutorial/clustering`

[43] J. Herrero, A. Valencia and J. Dopazo, *A Hierarchical Unsupervised Growing Neural Network for Clustering Gene Expression Patterns*, Bioinformatics, Vol. 17(2), 2001, p. 126.

[44] J. Tamames, D. Clark, J. Herrero, J. Dopazo, C. Blaschke, J. M. Fernandez, J. C. Oliveros and A. Valencia, *Bioinformatics Methods for the Analysis of Expression Arrays: data clustering and information extraction*, Journal of Biotechnology Vol. 98, 2002, p. 269.

[45] Ali, Ameer M.; Karmakar, Gour C. and Dooley, Laurence S, *Review on Fuzzy Clustering Algorithms*, Journal of Advanced Computations, Vol. 2(3), 2008, p. 169.

[46] `http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html`

[47] A. Baraldi and P. Blonda, *A Survey of Fuzzy Clustering Algorithms for Pattern Recognition*, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 29, NO. 6, DECEMBER 1999

[48] Y. Cheng and G.M. Church, *Biclustering of Expression Data*, Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology (ISMB '00), 2000, p. 93.

[49] A. Tanay, R. Sharan and R. Shamir, *Biclustering Algorithms: A Survey* Handbook of Computational Molecular Biology, edited by: Aluru S., Chapman & Hall/CRC, 2005.

[50] G. Getz, E. Levine and E. Domany, *Coupled Two-way Clustering Analysis of Gene Microarray Data*, PNAS, Vol. 97(22), 2000, p. 12079.

[51] S. C. Madeira and A. L. Oliveira, *Biclustering Algorithms for Biological Data Analysis: A Survey*, IEEE Transactions on Computational Biology and Bioinformatics, Vol. 1(1), 2004, p. 24.

[52] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery and W. L. Ruzzo, *Model Based Clustering and Data Transformations for Gene Expression Data*, Bioinformatics, Vol. 17, No. 10, 2001, p. 977.

[53] G. Schwarz, *Estimating the Dimension of a Model*, The Annals of Statistics, Vol. 6, No. 2, 1978, p. 461.

[54] G. J. McLachlan, R. W. Bean and D. Peel, *A Mixture Model-Based Approach to the Clustering of Microarray Expression Data*, Bioinformatics, Vol. 18, No. 3, 2002, p. 413.

[55] F. De Smet, J. Mathys, K. Marchal, G. Thijs, B. De Moor and Y. Moreau, *Adaptive Quality-Based Clustering of Gene Expression Profiles*, Bioinformatics, Vol. 18, No.5, 2002, p. 735.

[56] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith J. L. Barker and R. Somogyi, *Large-Scale Temporal Gene Expression Mapping of Central Nervous System Development*, Proc. Natl. Acad. Sci. USA, Vol. 95, Neurobiology, 1998, p. 334.

[57] K. Rose, E. Gurewitz and G. C. Fox, *Statistical Mechanics amd Phase Transitions in Clustering*, Phys. Rev. Lett., Vol. 65, No. 8, 1990, p. 945.

[58] L. Angelini, F. De Carlo, C. Marangi, M. Pellicoro, and S. Stramaglia, *Clustering Data by Inhomogeneous Chaotic Map Lattices*, Phys. Rev. Lett., Vol. 85, No. 3, 2000, p. 555.

[59] J. Schneider, *First Order Phase Transitions in Clustering*, Phys. Rev. E, Vol. 57, No. 2, 1998, p. 2449.

[60] E. Domany, *Cluster Analysis of Gene Expression Data*, Journal of Statistical Physics, Vol. 110, 2003, p. 1117.

[61] R. O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, New York, NY., Wiley and Sons., 1973.

[62] J. Moody and C.J. Darken, *Fast Learning in Networks of Locally-Tuned Processing Units*, Neural Computation, Vol. 1, No. 2, 1989, p. 281.

[63] A. Dekel and M. West, *On Percolation as a Cosmological Test*, Astrophys. J., Vol. 288, 1985, p. 411.

[64] W. E. Phillips, R. P. Velthuizen, S. Phuphanich, L.O. Hall, L.P. Clarke and M.L. Silbiger, *Application of fuzzy c–means segmentation technique for tissue differentiation in MR images of a hemorrhagic glioblastoma multiforme*, Magnetic Resonance Imaging, Vol. 13, 1995, p. 277.

[65] L. Cranias, H. Papageorgiou and S. Piperidis, *Clustering: A Technique for Search Space Reduction in Example–Based Machine Translation*, proceedings of the 1994 IEEE International Conference on Systems, Man, and Cybernetics. Humans, Information and Technology, 1, 1-6. IEEE, New York, 1994.

[66] A. Baraldi and F. Parmiggiani, *A Neural Network for Unsupervised Categorization of Multivalued Input Patterns: An Application to Satellite Image Clustering*, IEEE Transactions on Geoscience and Remote Sensing, Vol. 33, No. 2, 1995, p. 305.

[67] T. Kosaka and S. Sagayama, *Tree-Structured Speaker Clustering for Fast Speaker Adaptation*, proceedings of the 1994 IEEE International Conference on Acoustics, Speech and Signal Processing 1, 1994, IEEE, New York, p. 245.

[68] N.C. Jones and P.A. Pevzner, *An Introduction to Bioinformatics Algorithms*, MIT Press, 2004

[69] X. Ning and S. Zhang, *A Robust Clustering Technique for Grouping Biological Data: an Illustrative Study in Gene Expression Data*, The Third International Symposium on Optimization and Systems Biology (OSB'09), 2009, China, p. 267.

[70] L. Parsons, E Haque and H. Liu, *Subspace Clustering for High Dimensional Data: A Review*, ACM SIGKDD Explorations Newsletter, Vol. 6(1), 2004, p. 90.

[71] I. S. Kohane, A. Kho and A. J. Butte, *Microarrays for an Integrative Genomics* MIT Press, 2002.

[72] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, *Cluster analysis and display of genome-wide expression patterns*, Proc. Natl. Acad. Sci., Vol. 95(25), 1998, p. 14863.

[73] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander and T.R. Golub, *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation*, Proc. Natl Acad. Sci., Vol. 96, 1999, p. 2907.

[74] S. Tavazoie, J.D. Hughes, M. J. Campbell, R. J. Cho and G. M. Church, *Systematic determination of genetic network architecture* Nat Genet, Vol.22(3), 1999, p. 281.

Chapter 2

Superparamagnetic Gene Clustering with Transcription Factors

Contents

2.1 Description of the Superparamagnetic Clustering Algorithm

This method takes the data points generated by gene expression profiles as sites of an inhomogeneous Potts ferromagnet, and was first proposed by Eytan Domany *et al.* [1]. The presence of clusters in the data gives rise to magnetic grains, and working in the superparamagnetic phase, the SPC algorithm decides if a data point belong to the same grain using the pair correlation function of the Potts spins. Additionally, temperature controls the level of resolution obtained.

A Potts system is said to be homogeneous when its spins are on a lattice and all nearest neighbour couplings are equal, $J_{ij} = J$. This system exhibits two phases, at high temperatures is paramagnetic or disordered, and at low temperatures is ordered. In the disordered phase the correlation function G_{ij} decays to $1/q$ when the distance between points v_i and v_j is large (q is the number of possible states in the Potts model). This is the probability to find two completely independent Potts spins in the same state. At very high temperatures even neighbouring sites have $G_{ij} \approx 1/q$. As the temperature is lowered, the system undergoes a sharp transition to an ordered, ferromagnetic phase, meaning that one Potts state dominates the system. At very low temperatures $G_{ij} \approx 1$ for all pairs v_i, v_j, i.e. all spins have the same q [2].

In strongly inhomogeneous Potts models, spins form magnetic grains with very strong couplings between neighbours that belong to the same grain, and very weak interactions between all other pairs. At low temperatures such a system is also ferromagnetic, but as the temperature is raised the system may exhibit an intermediate, super-paramagnetic phase. In this phase strongly coupled grains are aligned (i.e. are in their respective ferromagnetic phases), while there is no relative ordering of different grains. This is illustrated in Fig. 2.1.

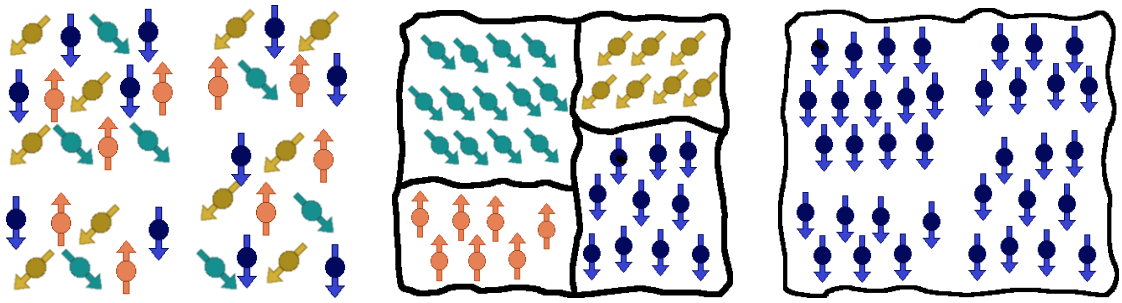


Figure 2.1: At high T all sites have different spin values, but as T is lowered, regions of aligned spins appears (superparamagnetic phase). At low T, the system is completely ordered.

At the transition temperature from the ferromagnetic to super-paramagnetic phase a pronounced peak of χ is observed [1]. As the temperature is further raised, the super-paramagnetic to paramagnetic transition is reached; each grain disorders and χ abruptly diminishes by a factor that is roughly the size of the largest cluster. Thus the temperatures where a peak of the susceptibility occurs and the temperatures at which χ decreases abruptly indicate the range of temperatures in which the system is in its super-paramagnetic phase. In principle, one can have a sequence of several transitions in the super-paramagnetic phase: as the temperature

is raised the system may break first into two clusters, each of them in turn breaks into more (macroscopic) sub-clusters and so on. Such a hierarchical structure of the magnetic clusters reflects a hierarchical organization of the data into categories and sub-categories [3].

In concreteness, SPC method consists on three stages. First, to specify the Hamiltonian which governs the system. Second, find the temperature range where the superparamagnetic phase take place, taking into acount the susceptibility behaviour. Finally, the correlation of neighbouring pairs of spins, G_{ij} is measured and, taking into account these values, the clusters are formed.

Each expression profile is represented as a point in a D dimensional space, and a random spin value σ_i, $i = 1,2,...,q$ is assigned to it. A small value q hinders the identification of the SPM clusters since different clusters are then forced to point into the same Potts direction. Too large q makes the calculations more cumbersome. However, the results depend only weakly on the value of q. In the next step, the neighbours of each spin v_i are calculated using the K mutual neighbour criterion. This criterion initially calculates the K nearest points of each site. If v_i has v_j among its K nearest points, and v_j, in turn, has v_i as one of its K nearest points, then v_i and v_j are considered as neighbours.

The average number of neighbours $\hat{K}$ and the average of all distances a between neighbouring pairs v_i and v_j are then computed, and finally the interaction couplings which will appear in the Hamiltonian will be calculated as follows:

$$J_{ij} = \begin{cases} \frac{1}{\hat{K}} e^{-\frac{d_{ij}^2}{2a^2}} & \text{if } v_i \text{ and } v_j \text{ are neighbours} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

Choosing J_ij in this way creates strong interactions between spins associated with the data from high density regions, and weak interactions between neighbours that are in low density regions [4].

Any different assignment of spins to data points S has an energy cost given by:

$$H(S) = \sum_{i,j} J_{ij} \delta_{\sigma_i,\sigma_j}, \tag{2.2}$$

where the sum is over neighbouring sites. The function $\delta_{\sigma_i,\sigma_j}$ is the Kroenecker symbol taking the value 1 when $\sigma_i = \sigma_j$ and 0 otherwise. The lowest possible energy cost, $H(S) = 0$ is attained when we assign the same spin to all points, which corresponds to all data points being assigned to the same cluster. Moreover, as one chooses interactions that are a decreasing function of the distance d_{ij}, then the closer two points are to each other, the more likely is for them to be in the same state. In summary, this Hamiltonian procedure penalizes placing spins at points i, j in different clusters, and this penalty decreases with the distance between the points [3].

The next step is the calculation of magnetization, susceptibility and correlation function for pairs of neighbours G_{ij} over a range of temperatures using Monte Carlo technique. The original creators of SPC used the Swendsen Wang algorithm.

As the temperature increases, M varies from 1 to 0 via sharp phase transitions. At low

temperatures the system is fully magnetized and the fluctuations in m are negligible. As T increases to the point where the single cluster breaks into subclusters (or become completely disordered), fluctuations become very large. Hence, one expect to identify the transitions at which clusters break up by the sharp peaks of the susceptibility [5].

The strategy is to vary T and measure $\chi(T)$. Transitions show up as peaks of χ. At temperatures between transitions, we expect to observe relatively stable phases that correspond to some clusters being ordered internally and uncorrelated with other clusters. Within each such phase, G_{ij} is measured. The value of G_{ij} is the probability to find the two Potts spins σ_i and σ_j in the same state, i.e. the probability to find them in the same cluster. By the relation to granular ferromagnets we expect that the distribution of G_{ij} is bimodal; if both spins belong to the same ordered grain (cluster), their correlation is close to 1; if they belong to two clusters that are not relatively ordered, the correlation is close to 0. Rather than thresholding the distances between pairs of points to decide their assignment to clusters, we use the pair correlations, which reflect a collective aspect of the data's distribution [3].

Clusters are identified in three steps:

1. Build the cores of the clusters using a thresholding procedure. If $G_ij > 0.5$, a link is set between the neighbour data points v_i and v_j. The resulting connected graph depends weakly in the value used in this thresholding, as long as it is bigger than $1/q$ and less than $1 - 2/q$ [3]. The reason is that the distribution of the correlations between two neighbouring spins peaks strongly at these two values and is very small between them.

2. Capture points lying in the periphery by linking each point to its neighbour of maximal correlation. Of course, some points were already linked in step one.

3. Data clusters are identified as the linked components of the graph obtained in the previous steps.

The temperature controls the resolution at which the data are clustered.

It is intuitively clear that if a set of data points form a dense cloud, isolated from the rest of the data, the corresponding spins will form a ferromagnetic domain at some low temperature, which will become paramagnetic and lose its correlations only at a high temperature. Hence the size of the temperature interval dT in which such a ferromagnetic domain exists can be used as a measure of the stability and significance of the corresponding data cluster.

Some of the demonstrated useful properties of SPC are the following: (a) the number of clusters is determined by the algorithm itself and not externally prescribed (as is done by SOM and K-means); (b) presents stability against noise; (c) generates a hierarchy (dendrogram) and provides a mechanism to identify in it robust, stable clusters (by the value of dT); (d) ability to identify a dense set of points forming a cloud of an irregular (non-spherical shape) as a cluster [26].

The SPC method has been used in various contexts, like computer vision [6], speech recognition [3] and identification of clusters of companies in stock indices [7]. Its first direct application to gene expression data has been for analysis of the temporal dependence of the expression levels in a synchronized yeast culture [8], identifying gene clusters whose variation reflects the cell cycle. Subsequently, the SPC was used to identify primary targets of p53 [9], the most important tumour suppressor that acts as a transcription factor of central importance

in human cancer. SPC has been used also to cluster protein sequences [10], and to classify or identify new genes associated with colon and skin cancer [11].

However, it has been reported that the main drawback of the SPC algorithm consists of dealing with data showing regions of different density [12, 13]. 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 [12] and a modularity approach [13].

2.2 Introduction of Transcription Factor Information in SPC: SPCTF

2.2.1 Cluster Stability Parameter

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 [23, 24, 25, 26], 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. The calculations performed for each temperature step yield a different cluster arrangement. In the original SPC algorithm, the final configuration was selected as the temperature located at halfway between the ferromagnetic to superparamagnetic transition and the superparamagnetic to paramagnetic transition. For yeast data, we found that no clear superparamagnetic to paramagnetic phase transition appeared. Additionally, 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. It is in this context that, in order to select the more representative clusters through all temperature steps, we assigned 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:

$$\mathcal{S}_v = \frac{col_v row_v}{|\, col_v - row_v| + \varepsilon}, \tag{2.3}$$

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. 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 $\mathcal{S}_v = \frac{n^2}{\varepsilon}$ instead of the infinity.

The advantage of using this stability parameter $\mathcal{S}_v$ is that it gives preference to clusters that survive several temperatures, but also have an acceptable number of elements. On the other

hand, it discards clusters that are too big but fragment rapidly, as well as clusters that maintain themselves for many temperature steps but consist of too few elements. Those situations are represented in our stability measure formula when $row_v >> col_v$ or $col_v >> row_v$. In those cases, the lesser value would be close to zero, and the cluster would be assigned a small stability index.

Summarizing, we were looking for an equilibrium between stable clusters running over several temperature steps and the clusters that achieve to gather several elements. The main goal is to have clusters with the most elements, which also maintained themselves for the most temperature steps. Cluster size is important because when we analyze clusters for biological relevance we are interested in groups of several elements with a similar function.

2.2.2 Improved Interaction

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 [28]. 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. (2.1) of the original algorithm, and has the following form:

$$J_{ij} = \begin{cases} \frac{F}{\tilde{K}} e^{-\frac{d_{ij}^2}{2(Fa)^2}} & \text{if } i \text{ and } j \text{ are neighbors,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

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. We actually tested some other values for the factor(f) that multiplies the common number of transcription factors (n) shared by two sites ($F = fn$). We present here the obtained susceptibility curves. As can be noticed, when we start increasing this factor, the curves shift towards the right but still hold their shape, however, with higher values the shape of the peaks start vanishing. We decided then to choose a value lesser than 2.5 to retain a well defined transition peak.

Additionally, we previously made several analysis using the data set we wanted to study and different factor values f. The table 2.3 compares cluster results that includes hits with Spellman *et al.* cell cycle reported genes on that study. The factor $f = 2.0$ gave us a first big cluster with somewhat lesser elements than in the case of SPC and other values. We wanted

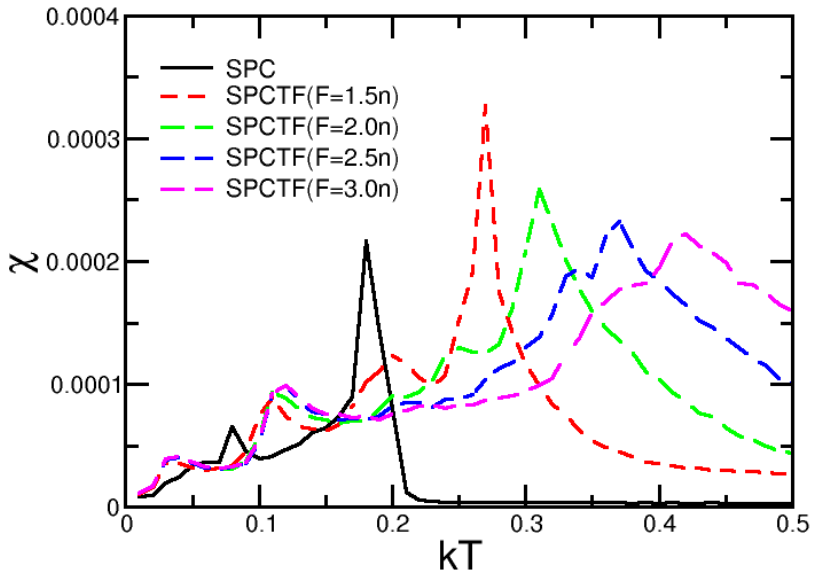


Figure 2.2: Susceptibility peaks for varios TF factors

to assure this fact as this cluster is discarded because it contains a majority of genes with no significant change in their expression. Combining the information of the susceptibility curves and the comparison table, we selected the factor $f = 2.0$ for the subsequent analysis.

Comparison for different TF factors (hits with Spellman *et al.*)

	First Cluster	hits	Cluster size ≥6	hits	Cluster size=5	hits	Cluster size=4	hits	Cluster size=3	hits	Cluster size=2	hits	Cluster size=1	hits	Total Genes	Total hits	Total clusters
SPC F=1	(2019 genes)	65	17(183)	85	7(35)	12	25(100)	11	50(150)	32	122(244)	47	1758	361	(4489)	613	1980
SPCTF F=1.5n	1(1667 genes)	56	24(306)	150	10(50)	19	31(124)	25	55(165)	33	180(360)	68	1817	262	(4489)	613	2118
SPCTF F=2.0n	1(1657 genes)	64	25(334)	168	11(55)	18	33(132)	22	63(189)	30	188(376)	65	1746	246	(4489)	613	2067
SPCTF F=2.5n	1(1726 genes)	72	24(325)	168	11(55)	21	32(128)	21	62(186)	30	184(368)	64	1701	237	(4489)	613	2015
SPCTF F=3.0n	1(1755 genes)	75	24(324)	170	11(55)	18	31(124)	19	61(183)	30	185(370)	64	1678	237	(4489)	613	1991

Figure 2.3: Comparison for different TF factors (hits with Spellman et al.)

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, the interaction between them becomes stronger, and this favors that the clusters including these sites remain stable for longer temperature ranges, with the corresponding increase of their stability values. In the next chapter we will described the results obtained with our proposed algorithm.

Bibliography Chapter 2

[1] M. Blatt, S. Wiseman and E. Domany, *Super-Paramagnetic Clustering of Data*, Phys. Rev. Lett., Vol. 76, 1996, pp. 3250-3255.

[2] S. Wiseman, M. Blatt and E. Domany, *Super-Paramagnetic Clustering of Data*, Phys. Rev. E, Vol. 57, 1998, pp. 3767-3787.

[3] M. Blatt, S. Wiseman and E. Domany, *Data Clustering Using a Model Granular Magnet*, Neural Computation, Vol. 9,1997, pp. 1805-1842. arxiv:cond-mat/9702072

[4] O. Barad, *Advanced Clustering Algorithm for Gene Expression Analysis using Statistical Physics Methods*, M.Sc Thesis conducted under the supervision of Prof. Eytan Domany Weizmann Institute of Science, December 2003 Chapter 4 Superparamagnetic Clustering-SPC.

[5] E. Domany, *Super-paramagnetic Clustering of Data- The Definitive Solution of an Ill-Posed Problem*, Physica A, Vol. 263, 1999, pp. 158-169.

[6] E. Domany, M. Blatt, Y. Gdalyahu and D. Weinshall, *Super Paramagnetic Clustering of Data: Aplication to Computer Vision*, Conference on Computational Physics, Granada, 1998; Comp. Phys. Comm., Vol. 121-122, 1999, p. 5.

[7] L. Kullmann, J. Kertész, R. N. Mantegna, *Identification of Clusters of Companies in Stock Indices Via Potts Super-Paramagnetic Transitions*, Physica A, Vol. 287, 2000, pp. 412-419.

[8] G. Getz, E. Levine, E. Domany and M.Q. Zhang, *Super-Paramagnetic Clustering of Yeast Gene Expression Profiles*, Physica A, Vol. 279, 2000, pp. 457-464.

[9] K. Kannan, N. Amariglio, G. Rechavi, J. Jakobo-Hirsch, I. Kela, N. Kaminski, G. Getz, E. Domany and D. Givol, *DNA Microarrays Identification of Primary and Secondary Target Genes Regulated by p53*, Oncogene, Vol. 20, 2001, pp. 2225-2234.

[10] I. Tetko, A. Facius, A. Ruepp and H-W Mewes, *Super Paramagnetic Clustering of Protein Sequences*, BMC Bioinformatics, Vol. 6, No. 1, 2005, p. 82.

[11] H. Gal, *Genome-Wide Expression Analysis using Novel Clustering Methods; Implications for Colon and Skin Cancer*, M.Sc Thesis conducted under the supervision of Prof. Eytan Domany and Prof. David Givol Weizmann Institute of Science January 2003 Chapter 3: Clustering Methods.

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

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

[14] Images of Microarray
http://www.imtek.de/anwendungen/content/workinggroups/topspotmikroarrayer/topspot_tech1.php

[15] Amol Prakash and Martin Tompa, *Discovery of Regulatory Elements in Vertebrates through Comparative Genomics*, Nature Biotechnology, Vol. 23, 2005, pp. 1249 - 1256.

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

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 3

Results and Conclusions

Contents

3.1 Comparison between SPC and SPCTF

We analyzed Spellman *et al.* [1] 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.* [1] 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.* [1]. 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.* [2] 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 [3].

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, or centroid, is obtained by averaging each coordinate between all its elements). Although different measures could have been used, this one performed adequately.

A possible concern for this approach would be when two clusters with very different shapes possess a very similar centroid. However this scenario is unlikely as more dimensions are added to the geometric problem and actually we noticed that the case mentioned never happens. We plotted the distance between the 28 most stable clusters from the SPCTF run against the distance with each of the four closest centroids from the SPC clusters in Fig. 3.1. It can be seen that most clusters from the SPCTF run had a very distinguishable equivalent in the SPC run (in the figure, that would be when the slope between the first and second closest SPC clusters is steep). This also can be seen by noticing that the closest clusters between the two runs generally do share most of their genes (please see the Supplementary Information Appendix).

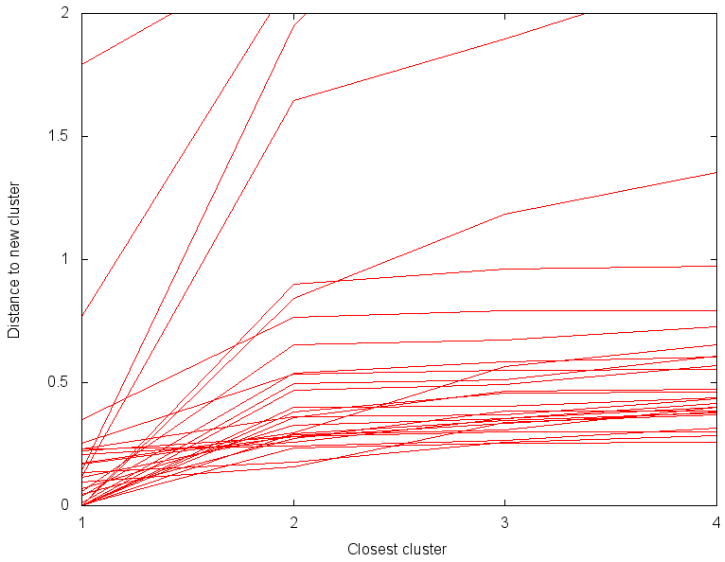


Figure 3.1: Here, each line depicts one of the 28 SPCTF clusters of size 6 and larger (the first one being the massive cluster discarded in the analysis), and we plot the distance of the closest four clusters from the SPC case.

Apart from the case discussed above (a steep slope between the first and second closest clusters) another case is worth mentioning: the slope between closest clusters in SPC is almost zero. This means that a new cluster has been formed by joining several SPC clusters, so its centroid is roughly at the same distance from the original, smaller clusters. In this case we also chose the closest one as the SPCTF equivalent. An example of this behavior can be analyzed in detail in the Supplementary Information Appendix section III, clusters 7 and 11, the former being of greater interest because it is formed by four smaller SPC clusters. All of them have Spellman-identified genes exclusively.

In Table 3.1, we present the differences in cluster size as well as the hits, the number of genes reported by Spellman *et al.* [1], 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.* [1] 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.

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). The 5 or less elements clusters are still well grouped but were not analyzed basically because they have a very reduced number of elements and do not offer relevant biological information as the majority of them do not have many hits with cell cycle genes (we have 2016 clusters of which 1723 are one element clusters). In the Supplementary Information Appendix, the elements and expression profiles of the first 27 clusters are showed, along with the information for clusters of 5 and 4 elements that have some hits with Spellman *et al.*.

Results of the comparison for the first 27 most stable clusters, discarding the first one, are

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 3.1: 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.* [1] appear in bold type. Hits with the 613 cell cycle genes reported by Spellman *et al.* [1] increase for clusters of size 6 and bigger, while decreasing in the first cluster and outliers.

shown in Fig. 3.2. Generally, these clusters incorporate more elements with SPCTF, including more cell cycle genes as those reported by Spellman *et al.* [1], thus improving the matching.

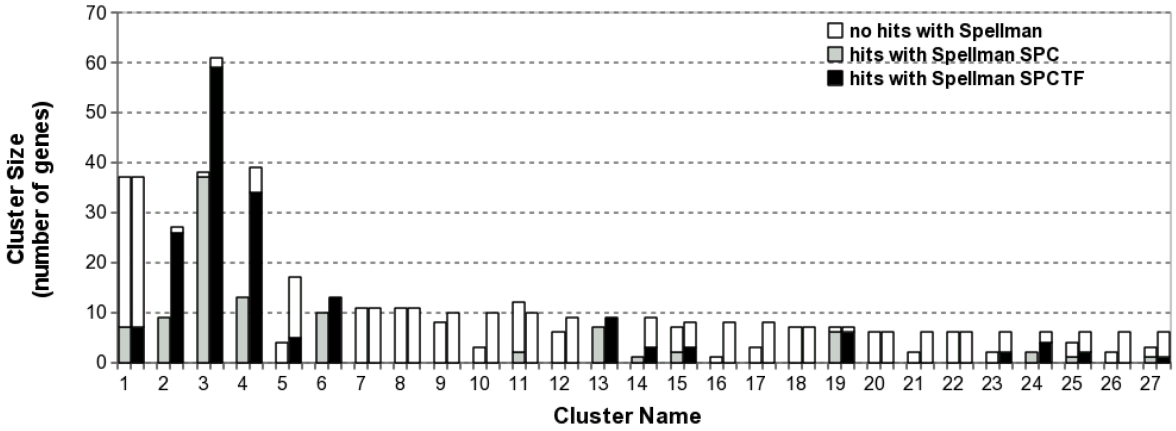


Figure 3.2: General comparison of the first 27 clusters, discarding the first one. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF. Groups tend to increase in size and also in hits with cell cycle genes reported by Spellman *et al.* [1], with the exception of cluster 11.

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. 3.3). The second type, mixed genes, M, contains clusters with non-reported genes as well as already known cell cycle genes (Fig. 3.4), 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. 3.4).

It is worth mentioning that more cell cycle experiments have been done since Spellman *et al.* [1] 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.* [4], Pramila *et al.* [5], Rowicka *et al.* [6] and Lichtenberg *et al.* [7]. We analize our 27 clusters taking now as hits, genes reported either by Spellman *et al.* [1] 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. 3.5–3.7 [8].

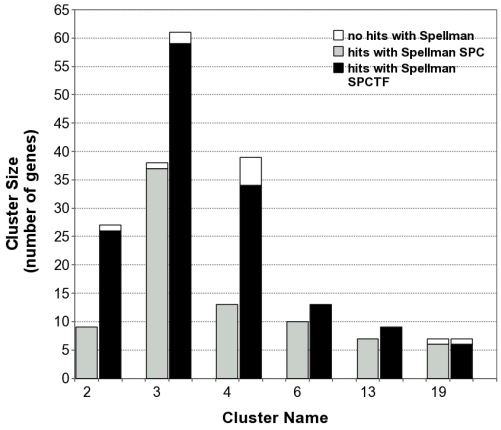


Figure 3.3: Comparison between the SPC and SPCTF results, showing the CC clusters. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF.

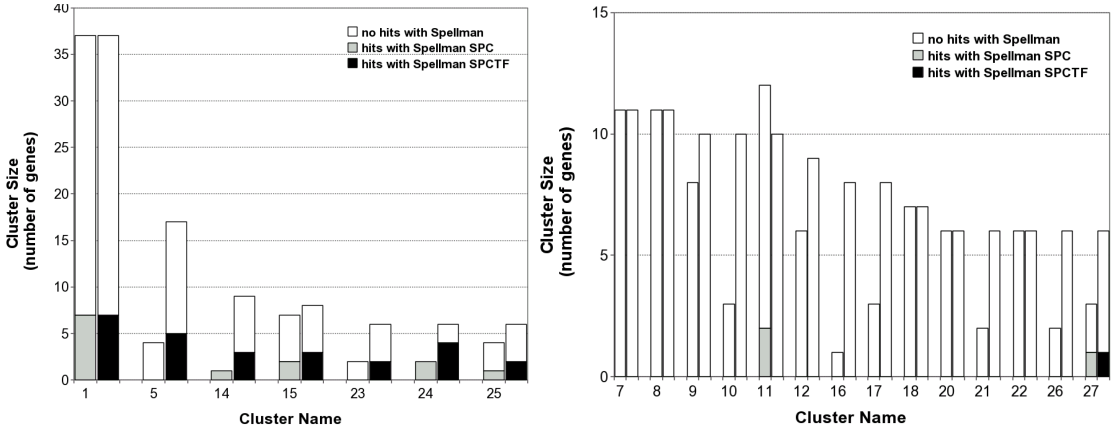


Figure 3.4: M and N clusters, left and right respectively. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF.

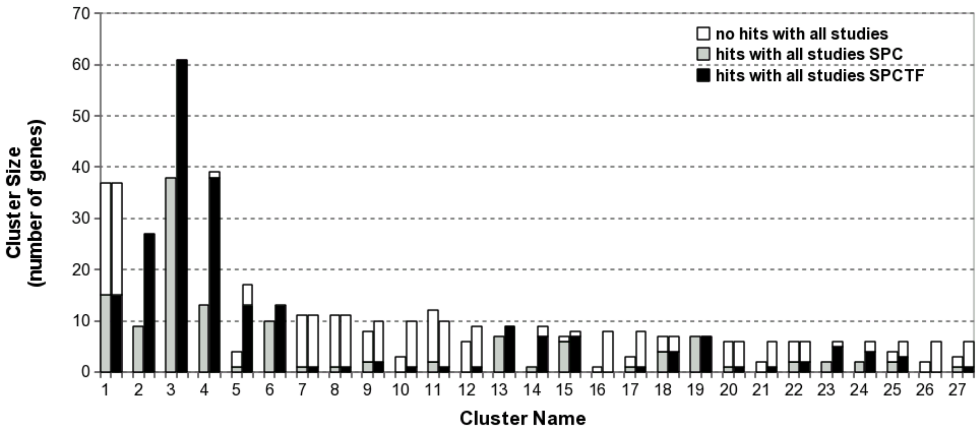


Figure 3.5: General comparison of the first 27 most stable clusters. Hits are now taken as cell cycle genes reported by all studies. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF.

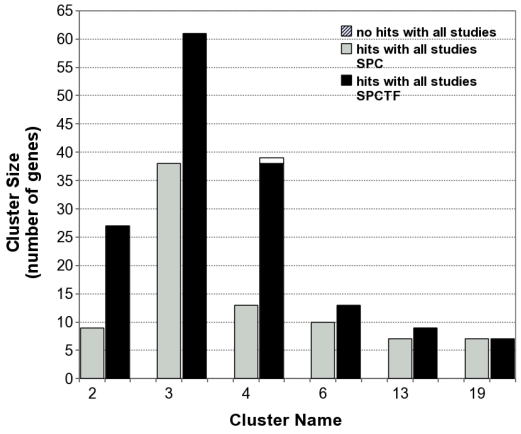


Figure 3.6: Comparison between SPC and SPCTF results, showing CC clusters. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF.

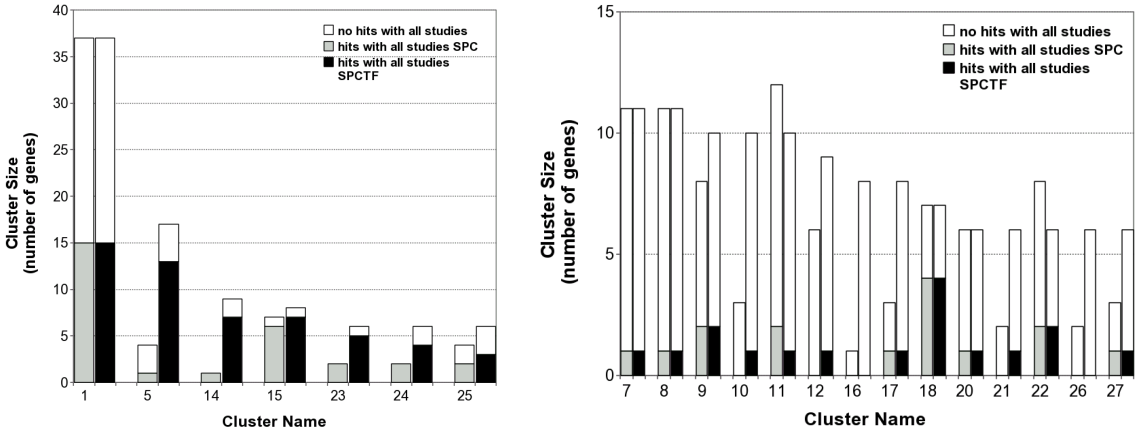


Figure 3.7: M and N clusters, left and right respectively. Gray bars correspond to the clusters obtained with the SPC algorithm and black bars to the equivalent clusters in SPCTF.

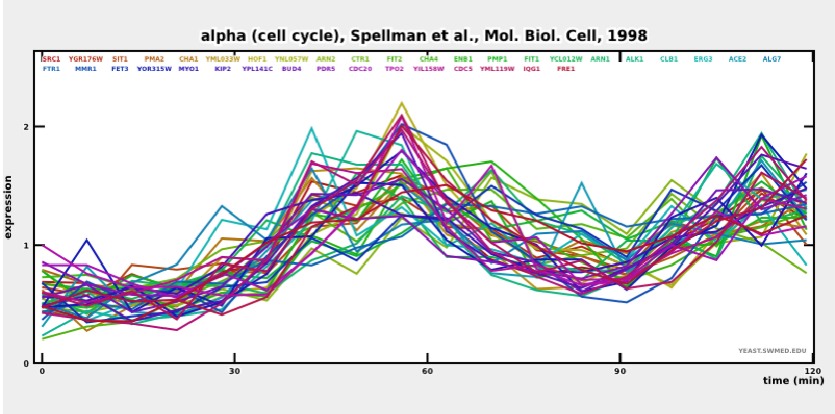
In addition, we analyze the expression profiles of the genes conforming each cluster using the SCEPTRANS tool [9], and we notice that all the genes grouped in the same cluster had the same expression pattern. This gives us further confidence that our algorithm is grouping data correctly. The expression profiles for a representative member of each cluster type are shown in Fig. 3.8.

We also find two clusters (21 and 27) that present an oscillating behaviour (see its expression profiles in the supplementary Information Appendix) that is due to an artifact in the manner the microarray experiment was performed, see [10, 11]. We contacted one of the authors of the former paper, Ph. D. Ahnert, who kindly gave us a list of 832 oscillating genes. We found that cluster 16 is not formed by any of these genes, however, cluster 21 and cluster 27 are; most of their elements being oscillating genes found by Ahnert *et al.* In table 3.9, we show the information of the number of genes in each cluster that are from Ahnert et al.'s list. This table can also be found in the Supplementary Information Appendix, along with the complete list of genes with oscillating behavior.

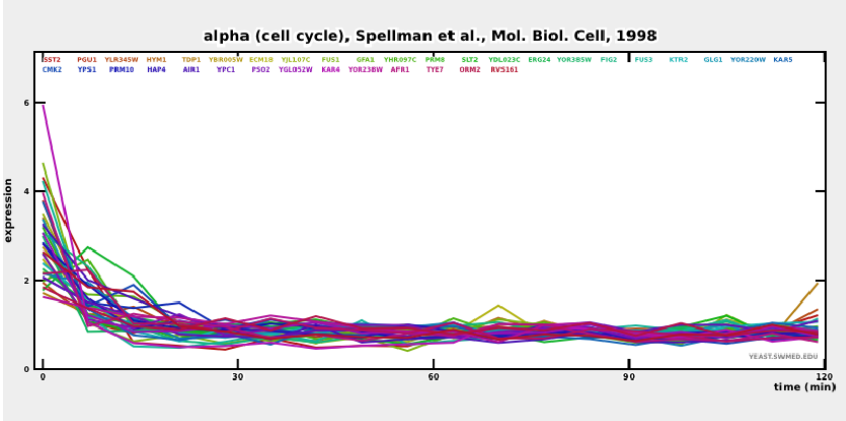
We also contacted one of the authors of the Spellman *et al.* article, Ph.D. Gavin Sherlock, and he mentioned that some genes that are alpha factor induced show strong expression at the beginning, but then fall off in expression. We actually found a cluster whose elements clearly show that behaviour, cluster 1.

The CC clusters are almost entirely composed of cell cycle regulated genes reported either by Spellman *et al.* [1] 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` and which is briefly explained in the next section.

CC: cluster 4



M: cluster 1



N: cluster 16

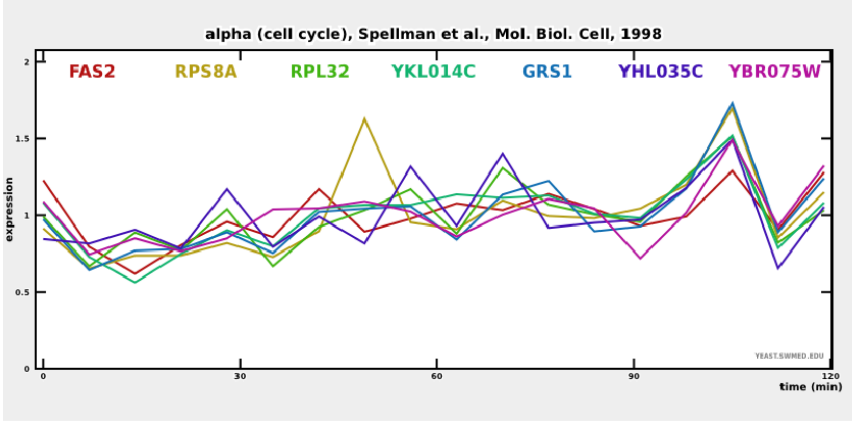


Figure 3.8: Expression profiles for a representative member of each cluster type using the SCEPTRANS tool. Expression profiles for all clusters are available in the supplementary information.

Oscillating Genes

Cluster Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Cluster Size	37	27	61	39	17	13	11	11	10	10	10	9	9	9	8	8	8	7	7	6	6	6	6	6	6	6	6
Number of Oscillating genes	0	0	0	1	2	0	1	1	0	1	2	1	0	1	0	0	0	1	1	0	4	0	0	0	1	2	5

Figure 3.9: Oscillating genes

3.2 MUSA

As already mentioned, many aspects of gene expression regulation involve transcription factors, proteins which bind to DNA in order to modulate the transcription rate of genes. Transcription factors search for specific positions in the upstream regulatory region (promoter) of a gene, thus, motif finding is the problem of discovering those promoter sequences and binding sites, usually referred to as consensus sequences or motifs, without any prior knowledge of their characteristics. Motif finding remains a difficult problem, one reason being that a single transcription factor might bind to regions which vary greatly in their sequence. In other words, although the binding sites for a particular transcription factor share some common pattern, the pattern is not specific, and thus finding it is a difficult task [12]. Another key feature of modern motif finders is the ability to extract complex motifs, i.e. motifs with gaps or spacers, otherwise known as structured motifs or multiads (dyads, triads, etc) [13].

After all the genes from an organism are clustered based on their expression patterns, an important next step is to examine the upstream region of genes in the same expression pattern group and look for sequence motifs. These motifs might be the regulatory signal (most likely a transcriptional regulatory site) that causes these genes to respond similarly to developmental or environmental changes. Information about expressions, regulatory motifs and functions provides substantial insight to the understanding of gene networks[15], [14].

In this spirit, the motivation is to analyze our obtained N type clusters with a motif finding algorithm, selecting those formed by genes with common regulatory sequence motifs.

Many algorithms have been proposed for the problem of finding biologically significant motifs in promoter regions. They can be classified into two large families: combinatorial methods and probabilistic methods. Probabilistic methods have been used more extensively, since they require less input from the user, and their output is easier to interpret. Combinatorial methods have the potential to identify hard to detect motifs, but their output is much harder to interpret, since it may consist of hundreds or thousands of motifs [13].

We selected MUSA algorithm because its performance is independent of the composite structure of the motifs being sought, making few assumptions about their characteristics. Additionally, this program also compares the found motifs on a set of genes to the transcription factor binding sites already described in yeastract database. MUSA propose a method that processes the output of combinatorial motif finders in order to find groups of motifs that represent variations of the same motif, thus reducing the output to a manageable size. This processing is done by building a graph that represents the co-occurrences of motifs, and finding clusters or communities in this graph. Structured motifs, i.e., motifs with gaps or spacers, are also processed. This innovative approach leads to a method that is as easy to use as a probabilistic motif finder, and as sensitive to low quorum motifs as a combinatorial motif

finder. The method was integrated with two combinatorial motif finders, and made available on the Web [13, 16].

MUSA searches in all the selected genes for overrepresented sequences in the promoter, that can be separated by different length gaps. It also constructs families of these sequences, including those which are very similar, except for some bases. Each extracted motif is associated to a p-value, which indicates its statistical significance. The given p-value corresponds to the lowest p-value attributed to the motifs in the family. The quorum is the number of genes that present a motif related to the total number of analyzed genes, and the alignment score quantifies how similar the encountered motif is with respect to an already reported transcription factor.

A comparison between a motif family and the transcription factor binding sites deposited in YEASTRACT can also be obtained using Position Weight Matrix descriptions. The MUSA algorithm proves to be very accurate providing predicted transcription factor binding sites, without any need for the user to manipulate search parameters [17]. A detailed explanation of the algorithm can be found at [13], [18].

3.3 MUSA Results

Results of this analysis are shown in Table 3.2, 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.

3.4 Conclusions

Using *a priori* information will undoubtedly affect the outcome of our procedure. However, it is desirable to use useful and available biological information and indeed in several papers, formulas in their algorithms are modified to make room for the bioinformation available. See, e.g., Ref [21, 22], and the many references cited in the first page of the latter paper.

Nevertheless, it could be inferred that post-processing the SPC results with the transcription factors information would yield the same results of the SPCTF approach, questioning thusly its validity. To look further into this matter, we also tried to see what comes out from post-processing the SPC results, joining clusters directly with only one criterion: whether they shared genes with common transcription factors. Basically, all the genes collapse in one massive cluster. The following figure illustrates this:

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 3.2: 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.

Basically if we add the transcription factor information to the SPC original clusters, all the genes collapse in one massive cluster. The following figure depicts a graphical sketch of this:

This can already be seen if we take the transcription factors data and have it run through the Hoshen-Kopelman algorithm: at one point or another, every gene can be traced to any other gene. On the other hand, by incorporating the TF information in the natural evolution of the SPC algorithm, only certain clusters will be enhanced. 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.

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 interaction between genes which share a transcription factor involved

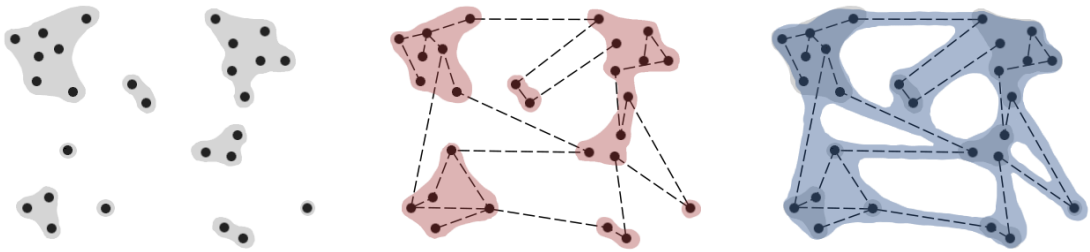


Figure 3.10: Left: original clusters (grey) found by SPC algorithm. Middle: new clusters (red) found by SPCTF algorithm. Some clusters have joined. Right: original SPC clusters joined using transcription factors a posteriori. Dashed lines represent transcription factors shared by two genes

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. 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.* [1] 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.

Bibliography Chapter 3

[1] P. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, *Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast S. cerevisiae by Microarray Hybridization*, Mol. Biol. Cell, Vol. 9, 1998, p. 3273.

[2] G. Getz, E. Levine, E. Domany, and M. Q. Zhang, *Super-paramagnetic Clustering of Yeast Gene Expression Profiles*, Physica A, Vol. 279, 2000, p. 457.

[3] 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, *A multi-step Approach to Time Series Analysis and Gene Expression Clustering*, Bioinformatics, Vol. 22, 2006, p. 589.

[4] 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, *A Genome-wide Transcriptional Analysis of the Mitotic Cell Cycle*, Mol. Cell, Vol 2, 1998, p. 65.

[5] T. Pramila, W. Wu, S. Miles, W. S. Noble, and L. L. Breeden *The Forkhead Transcription Factor Hcm1 Regulates Chromosome Segregation Genes and Fills the Sphase gap in the Transcriptional Circuitry of the Cell Cycle*, Genes Dev. Vol. 20, 2006, p. 2266.

[6] M. Rowicka, A. Kudlicki, B.P. Tu, and Z. Otwinowski, *High-resolution Timing of the Cell-cycle Regulated Gene Expression*, PNAS, Vol. 104(43), 2007, p. 16892.

[7] U. de Lichtenberg, R. Wernersson, T. S. Jensen, H. B. Nielsen, A. Fausboll, P. Schmidt, F. B. Hansen, S. Knudsen, and S. Brunak, *New Weakly Expressed Cell Cycle-regulated Genes in Yeast*, Yeast, Vol. 22(15), 2005, p. 1191.

[8] The list of the twenty-seven clusters with the names of all their genes and identified hits is available on Supplementary Information Appendix

[9] A. Kudlicki, M. Rowicka, and Z. Otwinowski, *SCEPTRANS: An Online Tool for Analyzing Periodic Transcription in Yeast*, Bioinformatics, Vol. 23(12), 2007, p. 1559.

[10] S. E. Ahnert, K. Willbrand, F. C. S. Brown, and T. M. A. Fink, *Unbiased Pattern Detection in Microarray Data Series* Bioinformatics, Vol. 22, 2006, p. 1471.

[11] E. M. Conlon, X. S. Liu, J. D. Lieb, and J. S. Liu, *Integrating Regulatory Motif Discover and Genome Wide Expression Analysis*, PNAS, Vol. 100, 2002, p. 3339.

[12] L. Hertzberg, O. Zuk, G. Getz and E. Domany, *Finding Motifs in Promoter Regions*, Journal of Computational Biology, Vol. 12(3), 2005, p. 314.

[13] N. D. Mendes, A. C. Casimiro, P. M. Santos, I. Sa-Correia, A. L. Oliveira, and A. T. Freitas, *MUSA: A Parameter Free Algorithm for the Identication of Biologically Signicant Motifs*, Bioinformatics, Vol. 22(24), 2006, p. 2996.

[14] X. Liu, D.L. Brutlag, and J.S. Liu, *Bioprospector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-Expressed Genes*, Pacific Symposium on Biocomputing Vol. 6, 2001, p. 127.

[15] J. Zhu and M.Q. Zhang, *Cluster, Function and Promoter: Analysis of Yeast Expression Array*, Pac Symp Biocomput., Vol. 5, 2000, p. 479.

[16] 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, *YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae*, Nucleic Acids Res., Vol 36, 2008, p. D132.

[17] `http://www.yeastract.com/discoverer/tutorial.php`

[18] A. M. Carvalho, A. T. Freitas, A. L. Oliveira, and M-F. Sagot, *An Efficient Algorithm for the Identification of Structured Motifs in DNA Promoter Sequences*, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, Vol. 3(2), 2006, p. 126.

Appendix A

Basic Concepts of Molecular Biology

A.1 Introduction: The Macromolecular Mechanisms of Life

Life on Earth, ranging from single-celled to complex organisms, seems extremely intrincated and diverse, therefore, is surprising to find that all biological systems share the same type of macromolecules and carry out processes in a similar way. The complexity arises in how these basic elements interact to attain the required high level of organization that makes life possible.

The cell, building block common for all organisms and elemental unit of life, is classified into two main types, eukaryotic and prokaryotic cells. An eukaryotic cell has its organelles and nucleus separated from the rest of the cell by membranes, while a prokaryotic cell, lacking this membrane-limited nucleus, is structurally simpler. Despite these differences, both kind of cells conserve some elemental processes and are made of the same building blocks. Moreover, creatures like humans are composed of collections of eukaryotic cells interacting in a coordinate manner to achieve an efficient performance of the organism, and, though each cell group is specialized in a particular task, leading to skin cells, neurons, blood cells or muscle cells, all of them work using common mechanisms [1]. How these processes are regulated and how is the interaction between genes are major focus of research today, and a good comprehension of them would lead to great advances in disease understanding and treatment. This appendix intend to give the basic notions of the mechanisms and macromolecules needed for cell functioning.

A.2 Proteins

Proteins are involved almost in every cell activity, hence, is not coincidence that they are the most abundant macromolecules in a cell. Many of them are enzymes in charge of catalyzing an incredible range of chemical reactions with extraordinary speed and specificity. Other roles of proteins are the transport of molecules across membranes, the transmission of information in cell-cell interactions and the detection and fighting of diseases (antibodies). They also provide structural rigidity, allow cells to move, and, as if that was not enough, proteins

carry out their own synthesis and that of other macromolecules controlling gene function [2]. The word protein is derived from the Greek proteios, whose meaning, “of the first rank”, reflects the importance of proteins to life.

The great number of tasks in which proteins participate is still more amazing if we consider that they are formed from only 20 different monomers, the amino acids. All amino acids have a general structure consisting of a central carbon atom (α carbon) bounded to four different chemical groups: an amino group (NH_2), a carboxyl group (COOH), an hydrogen atom (H), and one variable group, called side chain or R group (see Figure A.1). Side chains differ in size, shape, and chemical properties, thus confering to each amino acid its individual properties [1].

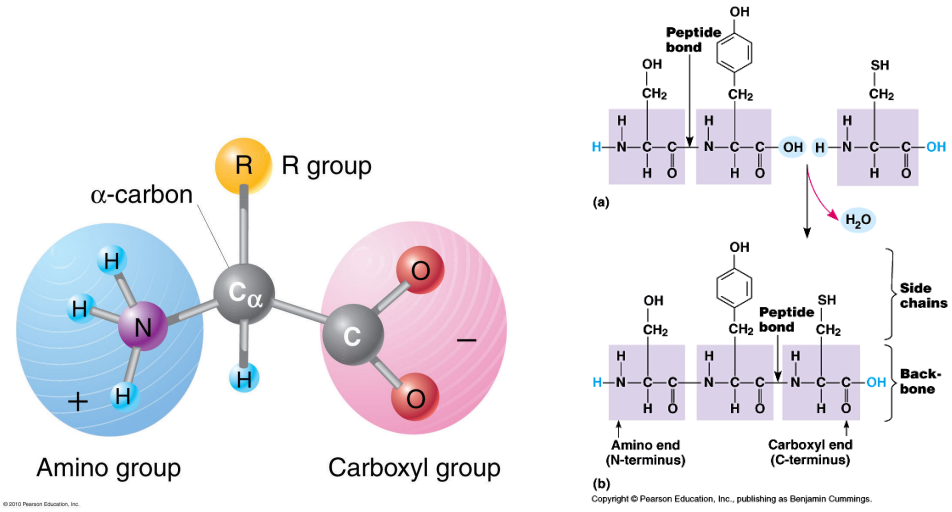


Figure A.1: Left: General structure of an aminoacid. Right: Formation of a peptide bond between two aminoacids. Images from [3], [4]

A protein molecule is formed by a linear sucesion of amino acids, linked through covalent peptide bonds (protein molecules are also known as polypeptides for this). Thus, amino acids can be combined in various sequences to form a huge variety of distinct proteins. The peptide bond occurs between the amino group of one amino acid and the carboxyl group of a second amino acid by a condensation reaction, leaving each protein molecule with two distinct “free” ends, the amino, or N terminus, and the carboxy or C terminus (see Figure A.1). As a polypeptide is synthesized from the amino to the carboxy terminus, the sequence of amino acids specifying a protein is written in the same order by convention [5]. The hormone insuline was the first protein whose amino acid sequence was determined in 1953 by Frederick Sanger [6].

The amino acid sequence is only the first element of the structure of a protein. For a correct performance, the three dimensional shape of the protein has to arise. Christian Anfinsen demonstrated that the conformation of a protein is the result of weak non covalent interactions between regions in the linear sequence, so the final shapes of proteins are also determined by their constituent amino acids [7]. The distinct chemical characteristics of the 20 amino acids play a critical role in the three dimensional conformation. For example, amino acids

with polar side groups tend to be on the surface of proteins, making them soluble in aqueous solutions, while amino acids with non polar side groups aggregate in a water insoluble core.

Protein structure is commonly described by four levels of organization [5]. The first level, the primary structure, is the sequence of amino acids. The secondary structure is the local arrangement of segments of the polypeptide chain. Without any stabilizing interactions, a polypeptide assumes a random coil structure, however, when hydrogen bonds form between the N-H and C=0 groups in the backbone, the chain can display, among others, two main types of secondary structure: α helix and β sheet. Both are particularly common, and were first proposed in 1951 by Linus Pauling and Robert Corey [8]. An α helix is formed when a single polypeptide chain twists around on itself and the C=O group of one peptide bond forms a hydrogen bond with the N-H group of an amino acid located four residues downstream, thus forming a spiral, rodlike structure. A β sheet, on the other hand, is a planar structure formed by several β strands, almost fully extended segments of polypeptide. This two structures can be seen in Figure A.2. As a single β strand is not very stable, it tends to interact with other strands via hydrogen bonds. These β sheets can arise either from strands running in the same orientation (parallel chain) or from strands running in opposite directions (antiparallel chains). These two structures generally appears in the hydrophobic cores of proteins, while loop regions connecting secondary structures are found on the surface of folded proteins.

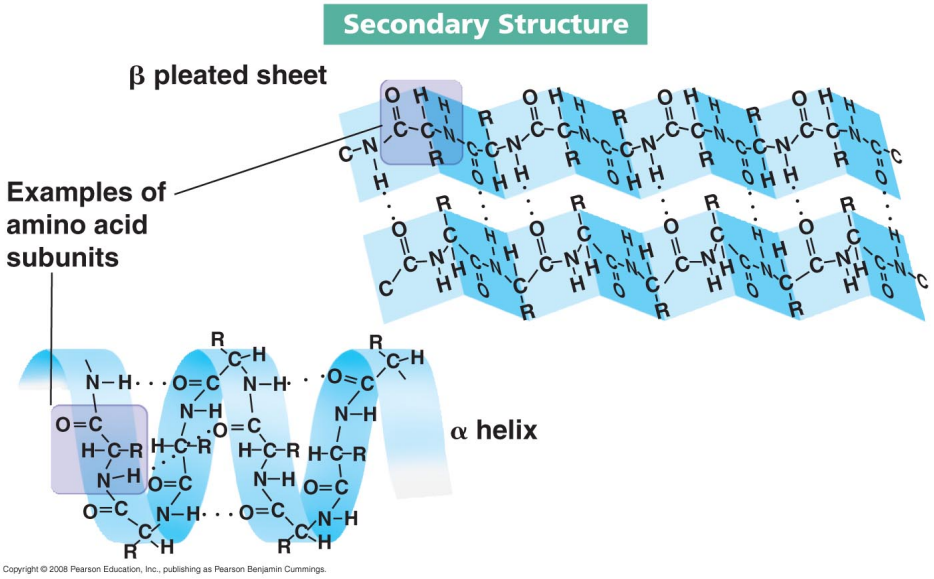


Figure A.2: The two principal secondary structures in proteins: α helix and β sheets. [9]

Tertiary structure refers to the folding of the polypeptide chain as a result of interactions between the side chains of amino acids, leading to a full three dimensional organization. Tertiary structure is stabilized by hydrophobic interactions between the non polar side chains and, in some proteins like cell-surface or secreted proteins, by covalent disulfide bonds between cysteine residues. Combinations of α helices, β sheets, and random coils fold independently into compact, stable globular structures called domains, which are the basic units of tertiary structure, and have probed to be of central importance in the correct functioning of proteins. Small proteins contain only a single domain; larger proteins may contain a number of different

domains, which are frequently associated with distinct functions. Thus, a protein conformation is dependent on the number, size, and arrangement of its secondary structures, determined in turn by its amino acid sequence.

Quaternary structure appears only in multimeric proteins composed by two or more polypeptide chains (subunits). This last level of organization describes the number and relative positions of the polypeptides in a multimeric protein (see Figure A.3). Hemoglobin, for example, is composed of four subunits held together by noncovalent bonds.

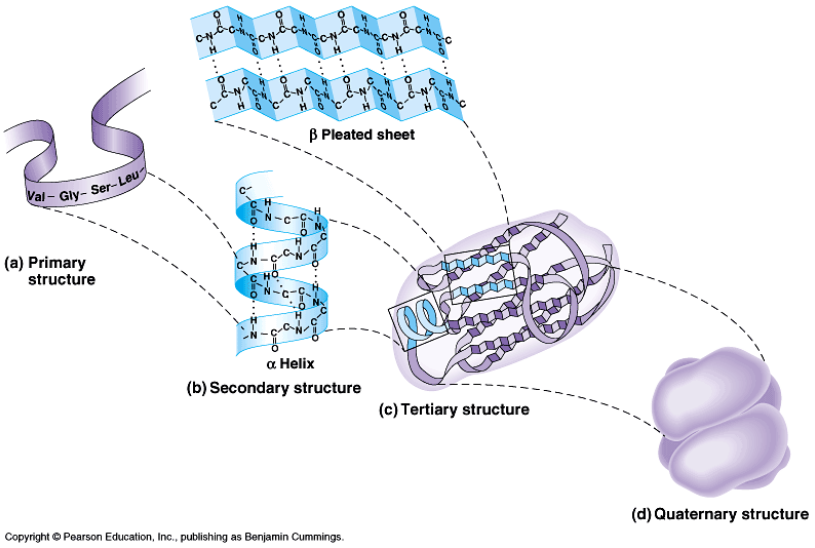


Figure A.3: The four basic levels of protein conformation [9]

Furthermore, proteins can associate to form larger structures termed macromolecular assemblies. Some examples include the protein coat of a virus, a bundle of actin filaments, the nuclear pore complex, and other large submicroscopic objects. Macromolecules in turn combine with other cell biopolymers like lipids, carbohydrates and nucleic acids to form complex cell organelles.

A.3 Nucleic Acids

Nucleic acid polymers exist as one of two similar chemical forms: ribonucleic acid (RNA) or deoxyribonucleic acid (DNA) [11]. The major responsability of nucleic acids is related with the storage and utilization of instructions to obtain proteins, basic molecules to build the cells and tissues of an organism. Deoxyribonucleic acid plays the role of the cellular library or “hard disk” of the cell, as it contains all these instructions. DNA is also the genetic material of the cell since duplication and passing of this molecule from generation to generation assures the genetic continuity of the species. On the other hand, ribonucleic acids are involved in various cell activities. Messenger RNA(mRNA), ribosomal RNA(rRNA) and transfer RNA(tRNA) are key elements in the process of translation, while other ribonucleic acids are capable of

catalyzing a number of chemical reactions intervening in both RNA processing and protein synthesis. Another example are small interfering RNAs (siRNAs), which hinder the expression of specific genes, and are involved in defense pathways against foreign nucleic acids.

DNA and RNA are polymers composed of monomers called nucleotides[1]. Nucleotides consist of two parts: a nitrogen-containing base linked to a five-carbon sugar (nucleoside), and one or more phosphate groups attached to the sugar, as shown in Figure A.4. There are two type of bases in nucleic acids, purine bases, which are formed by a pair of fused rings of nitrogen and carbon atoms, and pyrimidine bases, composed of a single ring. DNA contains the purines adenine(A) and guanine(G), and the pyrimidines cytosine(C) and thymine(T), leading to 4 different nucleotides. RNA contains A, G, and C too, but T is replaced by uracil(U), which lacks a methyl(CH_3) group. The other difference between the two polymers is that the sugar found in RNA is D-ribose, while the sugar found in DNA is 2'-deoxy-D-ribose. In this respect, RNA and DNA differ by the presence or abscence, respectively, of a hydroxyl(OH) group in one carbon of the sugar (see A.4). In few words, a single nucleic acid strand is a polymer (polyester) whose backbone is a sugar-phosphate sucession, with the bases coming out as side chains.

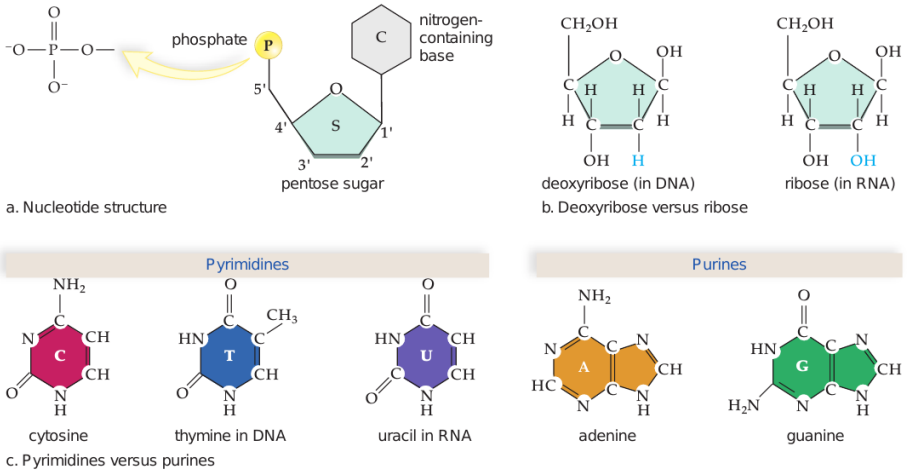


Figure A.4: Nucleotide structure, deoxyribose and ribose, and the different bases on nucleotides. [12].

Cellular RNA molecules could have many thousands of nucleotides, while DNA molecules can be as long as several hundred million nucleotides. The acidic character of nucleotides is due to the presence of phosphates, which dissociates at the pH found inside the cells, freeing hydrogen ions and leaving the phosphate negatively charged. Because these charges attract proteins, most nucleic acids in cells are associated with proteins.

When nucleotides polymerize to form nucleic acids, the hydroxyl group attached to the 3' carbon of a sugar of one nucleotide forms an ester bond with the 5' phosphate of another nucleotide (phosphodiester bond), eliminating a molecule of water. It is important to note that a

[1]Nucleotides also play other critical roles, for example, adenosine 5'-triphosphate(ATP) is the principal form of energy within cells; other nucleotides are carriers of reactive chemical groups or are important signaling molecules

polynucleotide chain has a direction: one end of the chain terminates with a free 5' phosphate group, while the other remains with a free hydroxyl group on the 3' carbon of the sugar (see Figure A.5). Furthermore, the synthesis proceeds always from 5' to 3' end, with nucleotides being added to the 3'OH group of a growing chain, which gives rise to the convention that nucleotide sequences are written and read in the 5' to 3' direction.

The linear sequence of nucleotides linked by phosphodiester bonds constitutes the primary structure of nucleic acids. Although the primary structures of DNA and RNA are very similar, their conformations are quite different. In contrast with RNA, which commonly exists as a single chain, DNA is formed by two inter-twined nucleotide strands running in opposite directions, and is chemically more stable. These differences are critical to the functions of the two types of nucleic acids.

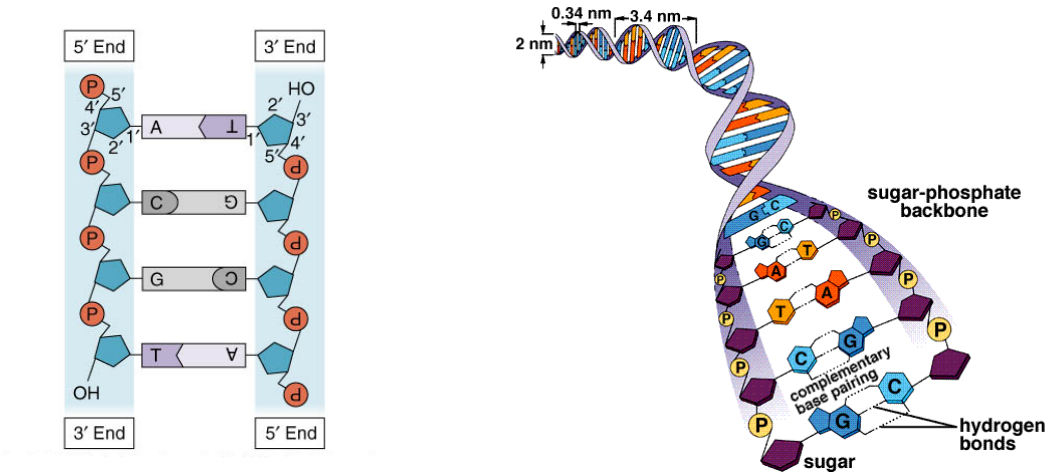


Figure A.5: General structure of DNA. Images from [12], [13]

The three dimensional structure of DNA was deduced in 1953 by James Watson and Francis Crick [14]. The central role of DNA as the genetic material had been demonstrated by that time, and the elucidation of its three dimensional structure was the next logical step to understand its function. Watson and Crick built their model taking into account Linus Pauling's description of hydrogen bonding, the alpha helix secondary structure, and experimental data from X-ray crystallography studies by Maurice Wilkins and Rosalind Franklin [15].

The DNA molecule structure turned out to be a double helix that rotate every 3.4 nm, with ten bases per turn of the helix (the distance between adjacent bases is 0.34 nm) and a diameter of approximately 2 nm. The sugar-phosphate backbones are on the outside of the molecule and the bases are on the inside, so that hydrogen bonds are formed between purines and pyrimidines on opposite chains (see Figure A.5). The covalently bonded hydrogens of amino(NH_2) and imino(NH) groups on the bases serve as donors, while the non bonded electron pairs of oxygen and ring nitrogen serve as acceptors. The bases are highly complementary, A always links with T and G with C. This explained earlier results of Erwin Chargaff, who had found, for various DNAs, that the amount of adenine was always equal to that of thymine and that the same happened for the amount of guanine and cytosine [17]. Base pairing through hydrogen bond formation is commonly referred to as secondary struc-

ture of nucleic acids.

The DNA structure above described is called B-form DNA, and is the most commonly found DNA structure. However, other kind of conformations for DNA can appear under special environmental conditions, such as pH, salinity or helical stress. Two such different helical conformations are A- and Z-form. Furthermore, a DNA molecule, although having a B-form in general, can acquire one of this different conformations in some regions, depending on its base sequence or the presence of proteins [18].

Unlike DNA, a RNA molecule generally interacts with itself rather than a second polynucleotide forming hairpins, and when RNA duplexes are formed, they acquire an A-form helix.

As a consequence of the specific base pairing, each strand on DNA contains all the information required to known the sequence of bases on the other strand. This property confers to DNA and RNA the capacity of direct their own self replication, as one strand of nucleic acid can direct the synthesis of a complementary strand. Furthermore, the redundancy helps to ensure that essential information is not lost, because if one copy is damaged, the other can serve as template for repairing.

In eukaryotes, DNA is divided in structures named chromosomes, which are single, long linear DNA molecules fold and packed with the aid of specialized proteins named histones. When the cell is ready to divide (mitosis), the chromosomes are even more compressed, and thus, easier to visualize. Many species carry more than one copy of their genetic material within each of their somatic cells (not germline cell). For example, all human cells contain 23 pairs of chromosomes, 46 in total, with the exception of the egg and sperm cells which have only one set. One chromosome pair is inherited from the mother, and the other one comes from the father. Organisms with only one copy of its genetic material are called haploid, those with two copies like human beings are called diploid, and those with more than two copies are called polyploid [1].

A.4 From DNA to Protein

The details of how a cell can produce proteins using its genetic material, were revealed in the years preceding the discovery of DNA structure. The first evidence relating a change on a DNA segment (a gene) with an alteration in a protein was made in 1957, when Dr. Vernon Ingram determined that there was only one different amino acid in the hemoglobin molecule of sickle-cell anemia patients, compared with normal hemoglobin molecules. Sickle-cell anemia is an inherited disease, thus, if the action of one gene variant was clearly to specify a particular sequence alteration in a protein, the immediate conclusion was that the probable function of all genes were to define particular amino acid sequences corresponding to distinct proteins [19]. For this to be true, DNA must be capable of specify in some way the 20 amino acids conforming proteins.

A DNA strand can be viewed as a linear code based on four letters, A,T,G and C, corresponding to the nucleotides. Four nucleotides taken individually could encode only four amino acids, and combining them in pairs they could encode only sixteen(4^2) amino acids, which still would be not enough. However, used as triplets, the four letters can display 64 (4^3) different combinations, more than enough to account for the 20 amino acids. Therefore, each triplet of

nucleotides, or codon, can codify for one amino acid, and a sequence of codons must define the sequence of amino acids corresponding to a protein. Indeed, this was experimentally confirmed and by middle sixties, in a truly stunning effort, scientists had deciphered the exact correspondence between codons and amino acids. In Figure A.6 the different existing codons are shown.

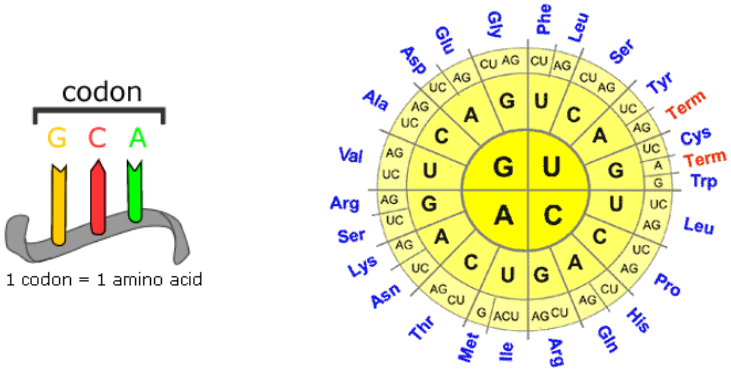


Figure A.6: Left: Three RNA bases are termed a codon. Right: Codons, read from the inside outward, are translated as amino acids. For example, the triplet CAC encodes the amino acid histidine (His) [20].

Another question that puzzled scientist was the fact that although DNA is located in the nucleus of eukaryotic cells, protein synthesis takes place outside it. Therefore, DNA was not capable of direct protein synthesis itself, some other molecule must transport the genetic information from DNA to the ribosomes. RNA appeared a logical candidate for such an intermediate because neither the change in sugar nor the substitution of U for T alters base pairing. This suggested that RNA could be synthesized directed by a DNA template, hypothesis effectively demonstrated in 1961 by Brenner et al. [21]. Later, a flow of genetic information was proposed by Crick, which states that the information is initially transfered from DNA to RNA molecules (transcription process), and that these RNA copies of segments of DNA are thus used to direct the fabrication of proteins (translation process). It also predicts that a protein cannot be used to specify or alter a particular nucleotide sequence [22]. This flow of genetic information is used nowadays, and receive the name of central dogma of biology [23].

A.4.1 Transcription

Although the concept of gene was conceived since the pioneer work of Mendel, it has been modified and adapted according to each new discovery. A gene is now defined as a sequence of nucleotides containing the information to synthesize a polypeptide or a functional RNA molecule. The whole set of genes that an organism could need during his entire life is stored in the DNA of each of its cells, and the word genome refers to the total nucleotide sequence of an organism, including both genes and non-coding regions. Various genomes have been completed until now, including the human genome of approximately 3 billion nucleotides. However, the number of genes in the human genome is constantly updated, being

current estimates around $25,000 - 30,000$.

The human average gene size is about $27,000$ nucleotides, but only about $1,300$ nucleotides are required to encode a protein of average size (about 430 amino acids in humans) [1]. The additional non-protein coding DNA on a gene is accommodated between the segments actually specifying for a protein, therefore the coding segments are called exons while non-coding sequences are named introns. Thus, a common human gene consists of a long string of alternating exons and introns, generally being introns much larger than exons. Additionally, a gene also contains a regulatory region controlling how often the product is synthesized. Some regulatory sequences includes promoters, enhancers and silencers. Some proteins, named transcription factors, recognize this sequences, bind to them and in this way promote or hamper the activity of the molecular machinery encharged of the transcription of the gene. The promoter region signals the start of transcription, being the TATA box the most common promoter sequence in eukaryotes. Enhancer sequences increase the velocity of the beginning of transcription, and silencers have the opposite effect. Most regulatory regions, but not all, are "upstream", that is, toward the 5' end of the transcription initiation site (see Figure A.7).

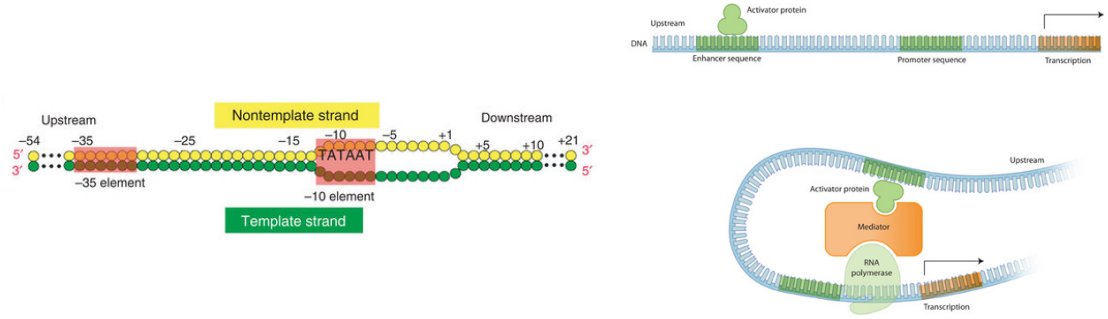


Figure A.7: Notable gene regions for transcription process. The transcription start site is denoted by +1. Positions upstream thus are negative numbers counting back from -1. Images from [24], [25]

A gene is active or is being expressed when its coding sequence is copied into messenger RNA (mRNA). The process is called transcription because the information to synthesize proteins specified on mRNA is essentially the same as in DNA.

The process is promoted by a multisubunit enzyme called RNA-polymerase, which catalyzes the formation of phosphodiester bonds to link nucleotides. The enzyme reads a DNA strand in the 3' to 5' direction and at the same time synthesizes a RNA strand from 5' to 3' sense. Besides, RNA-polymerase must recognize where to start and where to finish transcription. Bacteria contain a single type of RNA polymerase, while eukaryotic nuclei have three, called RNA-polymerase I, RNA-polymerase II and RNA-polymerase III. The three enzymes are structurally similar, sharing some common subunits and many structural features, but they transcribe different types of genes. While RNA-polymerase I and III transcribe the genes encoding transfer RNA, ribosomal RNA and various small RNAs, RNA-polymerase II transcribes the vast majority of genes, including all those that encode proteins, so we will focus on RNA-polymerase II.

The initial step of the transcription process in eukaryotes is the union of the so called general transcription factors to gene promoter sequences. TFIID (TF stands for transcription factor and II for RNA Polymerase II) is the first transcription factor to be recruited, since is composed by a TATA-box binding protein, capable of recognize the transcriptional start site, and some TBP-associated factors(TAF) that may help TBP in this task. RNA polymerase II and the remaining general transcription factors are consecutively incorporated to form a pre-initiation complex at the promoter. TFIIH is the last one to be attached, and is involved in the separation of DNA strands and the beggining of the elongation step. It is important to mention that other proteins are also involved in the overall process, including cromatin-remodeling complexes and histone acetyltransferases. Furthermore, enhancer sequences, which can be located several thousand bases upstream, downstream, or in the middle of the transcribed region, can also bind to proteins which stimulate transcription. These enhancer sequences are often tissue- and species-specific, explaining the regulation of genes in some tissues.

The carboxyl-terminal domain (CTD) of RNA-polymerase II is critical for elongation. In the initial phase, CTD is unphosphorylated, but once the pre-initiation complex has formed and RNA-polymerase II has incorporated a few nucleotides, TFIIH phosphorylate CTD causing the liberation of RNA-polymerase II (together with TFIIF). It starts to advance opening the following bases on DNA and adding new nucleotides to the 3' end of the mRNA chain, whose sequence is determined by the complementarity base-pairing between incoming nucleotides and the DNA strand used as template[2]. Once RNA-polymerase II go through, the mRNA chain is released from the DNA template, and the double helix re-arranges. In this phase, some transcription elongation factors interact with RNA-polymerase II to increase or reduce the rate of transcription elongation, or prevent the premature termination of the process.

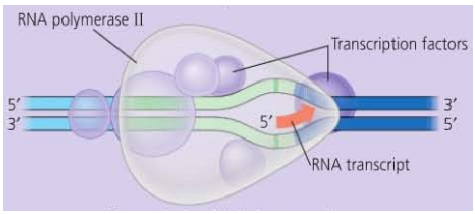


Figure A.8: In eukaryotic cells, proteins called transcription factors mediate the initiation of transcription by RNA polymerase II [9].

Termination mechanisms and its signalling sequences are less well characterized [26]. Transcription often terminates downstream of the poly(A) signal located downstream of the last exon and used to add a series of adenylate residues to mRNA. The CTD of RNA pol II recruites the enzymes CPSF (cleavage and polyadenylation specificity factor) and CstF (cleavage stimulation factor), which travel to mRNA once the poly(A) signal has been transcribed. CPSF, CstF, and other proteins cleave mRNA at the termination site, the sequence AAUAAA, generating the 3' end of the message to which now poly(A) polymerase adds approximately 250 adenine residues. Meanwhile, RNA-polymerase II continues with transcription until it stops and dissociates after several hundred nucleotides. The extra mRNA thus transcribed is

[2]The DNA strand whose sequence matches that of the mRNA is known as the coding strand, and the strand from which the mRNA was synthesized is the template strand.

degraded in the nucleus.

One of the models proposed to explain the dissociation of RNA-polymerase II postulates that cleavage of the mRNA chain at the poly(A) site leaves the following mRNA transcribed with an unprotected 5' end. Hence, this mRNA is degraded by exonucleases, destabilizating the template-transcript-Pol II complex, and leading to its final dissociation.

The chain of mRNA that results from direct transcription of the DNA encoding a gene is called the "primary transcript" or preliminary mRNA(pre-mRNA), and must be modified to be functional. Besides incorporation of the poly(A) tail, eukaryotic mRNA suffer other two changes during elongation phase: the incorporation of a protector nucleotide on the 5' terminal (cap), and the removal of introns [27].

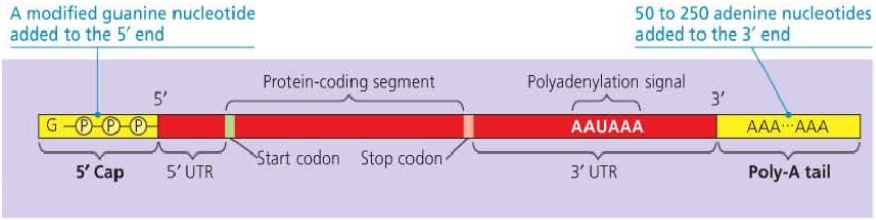


Figure A.9: Incorporation of the Poly(A) tail and 5' cap in pre-mRNA.[9]

The free 5' end is immediately modified at the beginning of transcription. The capping reaction begins with the escision, mediated by a phosphatase, of the 5'-end of the nascent transcript, converting the free triphosphate into a diphosphate. It follows the addition of a guanosine monophosphate by a 5'-5'-triphosphate linkage, type of bond occuring only in this process. The final reaction is methylation, yielding cap0 if only the guanosine is methylated, cap1 if also the second nucleotide is methylated, or cap2 if the guanosine, second and third nucleotide are methylated. The formation of this cap protects the 5' end from degradation due to nucleolytic and phosphorylytic activities (nucleases), while the poly(A) tail protects the 3' end, confering mRNA more stability and the possibility that the transcript could be translated several times, until degradation reaches the coding sequences [27].

In prokaryotic cells is common to find that adjacent genes are all transcribed in a single unit(operons), whose products are several proteins with related functions. By contrast, an eukaryotic primary transcript correspond only to one gene, but include intron sequences which will not be translated into protein. Reactions of splicing are in charge of remove introns and at the same time link the set of exons which will form the mature transcript. This is a major form of regulation in eukaryotic cells, because the primary transcript can be spliced in more than one way, through different arrangements of exons, to produce distinct mRNAs coding for a group of related proteins.

Intron elimination is determined by consensus sequences at the 5' and 3' splice sites, which define the boundaries of the intron, and an inner sequence, known as the branch site, about $20 - 50$ bases upstream from the 3' splice site. The transesterification reactions involved in the process are mediated by spliceosomes, RNA-protein complexes constituted by small nuclear ribonucleoprotein particles (snRNPs) and small cytoplasmic particles (scRNPs). Spliceosomes recognize and align the splice sites through base pairing, and prevent the

intermediates from leaving the complex until the reaction of cleavage and binding is complete. Most eukaryotes use this spliceosome-mediated splicing process, although some RNA molecules are able to do it alone (self-splicing RNA).

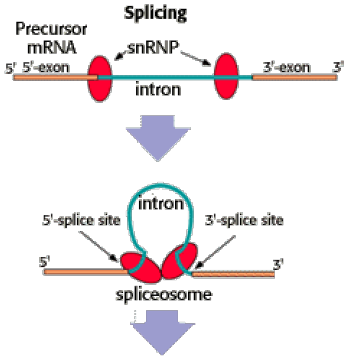


Figure A.10: Removing of introns by spliceosome. [28]

Once pre-mRNA is subject to 5'-end capping, splicing, 3'-end cleavage and polyadenilation, mature mRNA is ready to be transported to the ribosomes, where it will serve as the blueprint for protein synthesis. Associated with mRNA are always several proteins and small noncoding RNAs, all forming the messenger ribonucleoprotein particle (mRNP). Some of these associated factors are eliminated and others are incorporated during the entire mRNA lifetime, being this combination of factors who dictate almost everything that happens to mRNA, from its exit to the nucleus and transport to the ribosomes, until its total degradation (for a more complete review on this issue, consulte [29]).

A.4.2 Translation

Assembly of amino acids into proteins occurs by translation of the information stored on mRNA in the form of codons. However, each triplet does not bind directly to the amino acid that specifies, the process requires other type of RNA, called transfer RNA (tRNA), who plays the role of an adaptor molecule. Each tRNA consists about 80 nucleotides in length with a tertiary structure similar to a clover and two important regions: three unpaired nucleotides forming the anticodon, which would link to a complementary triplet in a mRNA molecule, and the acceptor site, where the corresponding amino acid is attached. The amino acid is bound by one of the 20 aminoacyl-tRNA synthetases, activating enzymes whose function is to recognize each tRNA and bind to it only the appropriate amino acid. Once a tRNA molecule is charged with an amino acid, receives the name of aminoacyl-tRNA. At least one kind of tRNA would be required for each of the 20 amino acids, but some of them can be attached to two or three different tRNAs. Taking the degeneracy of genetic code (there are 61 triplets to code only 20 amino acids), and the balance hypothesis, the theoretical minimum number of different tRNAs required is 31, plus a special initiator tRNA.

Protein synthesis takes place on ribosomes, complexes made of ribosomal proteins and rRNA. Besides, several nonribosomal molecules named protein factors are necessary during all process. For starting translation, a ribosome must be dissociated in a large and a

small subunit. The small subunit binds to the 5' cap of an mRNA molecule and proceeds downstream (5' to 3') until it finds the start signal, which in the majority of eukaryotic genes is the triplet ATG, embedded in a specific region, called Kozac sequence [3]. As ATG codes for methionine, all recently synthesized proteins start with this amino acid, which can be later eliminated in post-translational modifications. Once the small subunit encounters the start sequence, the large subunit and the special initiator aminoacyl-tRNA are incoporated. The complete ribosome is now functional, with three adjacent sites: the peptidyl binding site (P) in the middle, where the initiator aminoacyl-tRNA binds, the aminoacyl binding site(A), ready to accept a new aminoacyl-tRNA bearing the next amino acid, and an exit binding site (E), used later to release the uncharged tRNA molecules from the ribosome.(see Figure A.11)

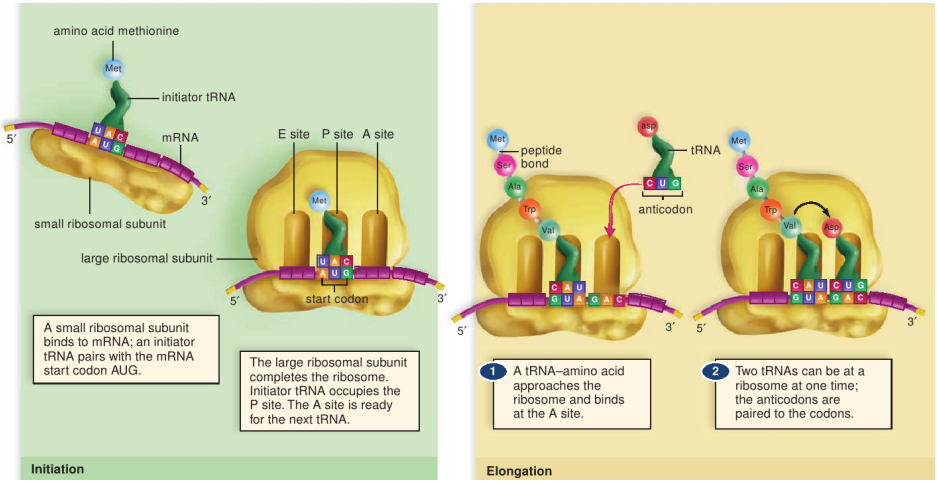


Figure A.11: Initiation and Elongation phase in the ribosome. [12]

The elongation phase starts with the incorporation of a second aminoacyl-tRNA molecule into the A site. If the pairing between anticodon and codon is correct, a peptide bond is formed between the carbonil group of the first amino acid and the amino group of the second one. The complete chain is now transferred from P to A site, leaving the uncharged initiator tRNA in the P site. Then, the ribosome advance one codon over mRNA in the 5' to 3' direction and as a consequence, the uncharged initiator tRNA is placed now over E site, ready to be released. The second tRNA carrying the complete chain is now on P site, while A site is ready to accept the third aminoacyl-tRNA. The entire process is repeated continually, and the protein grows from amino to carboxyl terminus. When one ribosome moves away from the initiation site, another ribosome can bind to the mRNA and begin translation again. Then, a single mRNA template can be translated simultaneously by several ribosomes.

When one of the three stop codons (UGA,UAA,UAG for nuclear eukaryotic mRNA) is reached, the ribosome stops because the corresponding amynoacil-tRNA does not exist. Instead, the eukaryotic release factor (eRF-1) recognize the stop codon, binds to A site and liberates the polypeptide chain from its tRNA at the P site with the aid of other release factors. Once the protein come out of the ribosome, the last tRNA is released and the ribosome dissociates in its two subunits, leaving free the mRNA template and finishing translation.

[3]The region between cap and AUG is known as the 5'-untranslated region [5'-UTR].

Translation is also regulated by the cell, through repressor proteins, localization of the mRNA chain or modulation of the activity of initiation factors among other mechanisms. During and after translation, the recently synthesized proteins must be modified and folded to acquire its three-dimensional structure and thus be functional. In many cases, various polypeptide chains must assemble into a single complex. Further modifications include cleavage and covalent attachment of carbohydrates and lipids, that are critical for the function and correct localization of proteins within the cell. Protein folding is often assisted by special proteins called molecular chaperones, which bind and stabilize partly folded polypeptide chains, preventing uncorrect folding or aggregation. Altough the final shape of the protein is completely specified by its amino acid sequence, chaperones facilitate the folding process making it more reliable.

A.5 Conclusion

Life is based on building blocks and mechanisms shared by all organisms. DNA is in charge of storing the information needed to synthesize the fundamental macromolecules required by cells. For this, the coding sequence in a gene is copied onto a RNA molecule. This molecule can be already functional, or be used to direct the synthesis of a protein following the rules of genetic code, in which sets of three nucleotides specify each protein aminoacid. In Figure A.12 a summary of the transcription and translation process is shown. The path between a gene and the emerging of a protein is not trivial, several processes occur with the intervention of many molecular complexes. Furthermore, regulatory mechanism are also involved at all steps making an intrincated network of interactions [1].

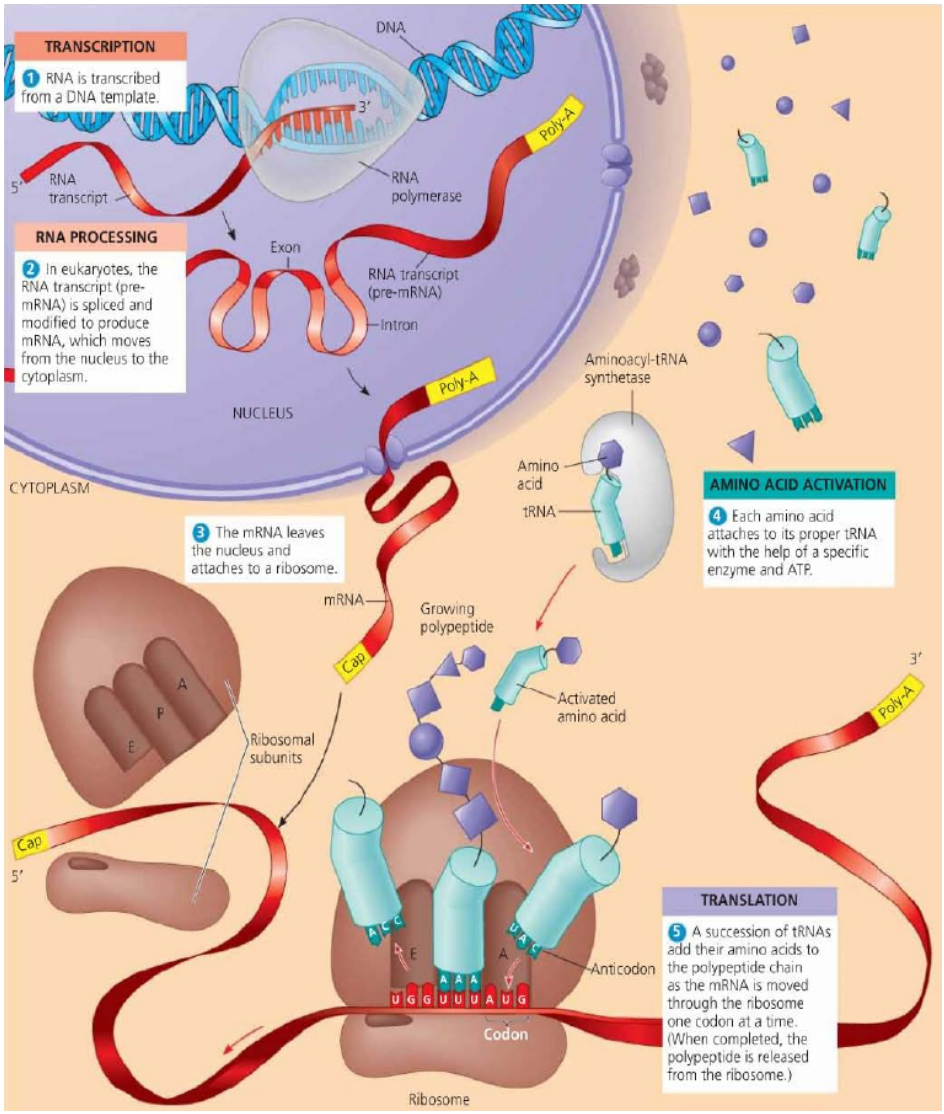


Figure A.12: This diagram shows the path from one gene to one protein [9].

Bibliography Appendix A

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Molecular Biology of the Cell*, Garland Science, 5th edition, New York USA, 2008.

[2] H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D Baltimore and J. Darnell, *Molecular Cell Biology*, W. H. Freeman, 4th edition, New York USA, 2000.

[3] Figure © 2010 PJ Russell, iGenetics 3rd ed.; all text material © 2010 by Steven M. Carr

[4] http://biochemistrycourse.blogspot.com/2011/05/peptide-bond.html

[5] A.J. Cozzone, *Proteins: Fundamental Chemical Properties*, Encyclopedia of Life Sciences, Macmillan Publishers Ltd, Nature Publishing Group, 2002.

[6] F. Sanger and E. O. P. Thompson, *The amino-acid sequence in the glycyl chain of insulin. 1. The investigation of lower peptides from partial hydrolysates*, Biochem. J., Vol 53, 1953, p. 353.

F. Sanger and E. O. P. Thompson, *The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates*, Biochem. J., Vol. 53, 1953, p. 366.

[7] C.B. Anfinsen, *The Molecular Basis of Evolution*; John Wiley & Sons, New York USA, 1959.

[8] L. Pauling and R.B. Corey, *Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets*, Proceedings of the National Academy of Sciences of the United States of America, Vol.37(11), 1951, p. 729.

[9] Campbell and Reese, (2008). Biology, 8th Edition. Benjamin Cummings

[10] http://academic.brooklyn.cuny.edu/biology/bio4fv/page/3d_prot.htm

[11] G.A. Soukup, *Nucleic Acids: General Properties*, Encyclopedia of Life Sciences, Macmillan Publishers Ltd, Nature Publishing Group, 2001.

[12] S.S. Mader, *Biology* McGraw-Hill, 6th Edition, 1998

[13] http://cae2k.com/bhagwan-photos-0/dna-nucleotides.html

[14] J.D. Watson and F.H.C. Crick, *A Structure for Deoxyribose Nucleic Acid* Nature, Vol. 171, 1953, p. 737.

[15] R. Franklin and R.G. Gosling, *Molecular Configuration in Sodium Thymonucleate*, Nature, Vol. 171, 1953, p. 740.

R. Franklin and R.G. Gosling, *Evidence for 2-Chain Helix in Crystalline Structure of Sodium Deoxyribonucleate*, Nature, Vol. 172, 1953, p. 156.

[16] M.H.F. Wilkins, A.R. Stokes and H.R. Wilson, *Molecular Structure of Deoxypentose Nucleic Acids*, Nature, Vol. 171, 1953, p. 738.

[17] E. Chargaff, S. Zamenhof and C. Green, (May 1950). *Composition of human desoxypentose nucleic acid*, Nature, Vol. 165(4202), 1950, p. 756.

[18] D.W. Ussery, *DNA Structure: A-,B- and Z-DNA Helix Families*, Encyclopedia of Life Sciences, Macmillan Publishers Ltd, Nature Publishing Group, 2002.

[19] V.M. Ingram, *Gene Mutations in Human Hemoglobin: The Chemical Difference between Normal and Sickle Hemoglobin*, Nature, Vol. 180(4581), 1957, p. 326.

[20] L. Merkel, N. Budisa, *Veränderung des Genetischen Codes*, BIOspektrum, Vol. 12, 2006, p. 41.

[21] S. Brenner, F. Jacob and M. Meselson, *An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein synthesis*, Nature, Vol 190(4776), 1961, p.576.

[22] F.H.C. Crick, *Central Dogma of Molecular Biology*, Nature, Vol. 227, 1970, p. 561.

[23] P.J. Pukkila, *Molecular Biology:The Central Dogma* Encyclopedia of Life Sciences, Macmillan Publishers Ltd, Nature Publishing Group, 2001.

[24] `http://www.nature.com/scitable/topicpage/gene-expression-14121669`

[25] C. Hyeon and D. Thirumalai, *Capturing the Essence of Folding and Functions of Biomolecules Using Coarse-grained Models*, Nature Communications, Vol. 2 (487), 2011, p. 1.

[26] A.R. Kornblihtt, *Shortcuts to the end*, Nature Structural & Molecular Biology Vol. 11, 2004, p. 1156

[27] A.J. Shatkin and J.L. Manley, *The Ends of the Affair: Capping and Polyadenylation*, Nat Struct Biol, Vol. 7, 2000, p. 838.

[28] `http://www.biology.arizona.edu/molecular_bio/problem_sets/mol_genetics_of_eukaryotes/06t.html`

[29] M.J. Moore, *From Birth to Death: The Complex Lives of Eukaryotic mRNAs*, Science, Vol. 309(5740), 2005, p. 1514.

Appendix B

DNA Microarrays

The first step to manufacture a needed protein in the cell is mRNA production, thus information about the transcript levels is needed as a first approach for understanding gene regulatory networks. DNA microarrays or DNA chips are one of the most important breakthroughs in experimental molecular biology precisely because they allow to monitor mRNA activity of thousands of genes at the same time.

Prior to this technology, researchers were limited to study a few genes per experiment and were able to assess interactions among genes under changing conditions on a much smaller scale. Microarray developed from *Southern blot*, a technique that was based on the principle that DNA and RNA strands could be labeled for detection and used to probe other nucleic acid molecules that were attached to a surface. This technique used at first a porous surface as support for DNA strands and radioactively labeled probes, but later, a glass surface and the application of fluorescente dyes instead of radioactivity for labelling greatly decreased the chemical reaction time and facilitate miniaturization as well. In the 1980s, numerous groups of researchers furthered the technology manufacturing arrays with high sensitivity, mechanizing and improving methods for their construction, leading to a decrease in production costs and industrializing the technology [1]. The first paper in which the term microarray was used in its current meaning were published in 1995 by the laboratory of Pat Brown at Stanford University, and they also were credited with engineering the first DNA microarray chip [2], while Stephen Fodor and colleagues at Affymetrix, Inc. patented the first DNA wafer chip in 1991 [3]. Since their conception, the use of microarrays has spread rapidly throughout the research community (see [4], [5] for a more detailed microarray development history).

A DNA microarray consists of a solid substrate similar to a microscope slide, which could be made of nylon, glass or plastic, hosting wells or spots arranged in a grid pattern. Typically, one DNA chip will provide hundreds or thousands of spots [1]. Each spot containes thousands of single stranded DNA or oligonucleotides, sometimes refered to as probes, that are specific for certain gene [1]. The technology works on the principle of base pairing and hybridization: a fluorescently labeled DNA or RNA solution coming from a sample (called target) is applied to the microarray and allowed to hybridize, then the slide is washed to remove nonspecific hybridization and it is read in a laser scanner collecting fluorescence intensities representing mRNA production from each gene. In other words, the light intensity emanating from each

[1] Microarrays containing thousands of spots are considered high-density microarrays

spot can be related to the amount of mRNA present in the tissue and in turn, with the amount of protein produced by the gene.

Many microarray systems have been developed by academic groups and commercial suppliers, however there are two mainly used systems according to the arrayed material, cDNA and oligonucleotide microarrays [6]. The term DNA microarrays usually refers to cDNA microarrays, whereas DNA chips or oligochips commonly refer to oligonucleotide arrays. CDNA microarrays spots contain sets of presynthesized complementary DNA sequences which are several hundred bases long (500-2000 bases) [7]. CDNAs for microarrays may include fully sequenced genes of known function or collections of partially sequenced cDNA derived from expressed sequence tags (ESTs) corresponding to the messenger RNAs of unknown genes. On the other hand, oligonucleotide microarrays are divided in two: short and long oligonucleotide chips. Each gene is represented on the spot by a set of approximately 20 different oligonucleotide probes designed to hybridize perfectly to some particular gene and these DNA sequences, termed perfect match sequences, are shorter than those in cDNA, approximately 25 bases for short oligochips and ranging from 50 until 70 bases for long oligochips [8]. Mismatch control oligonucleotides, identical to the perfect match sequences except for a single base-pair mismatch, are also included. These mismatch control oligonucleotides allow estimation of cross-hybridization, improving reproducibility and accuracy of RNA quantification, and reducing the rate of false-positives [10]. Oligonucleotides can be synthesized in situ with photolithography techniques, ink jet printing or electrochemical synthesis, or they can be also synthesized by standard methods and spotted latter onto the chip. Each technology has its advantages and limitations [11], [9]. (see Figure B.1)

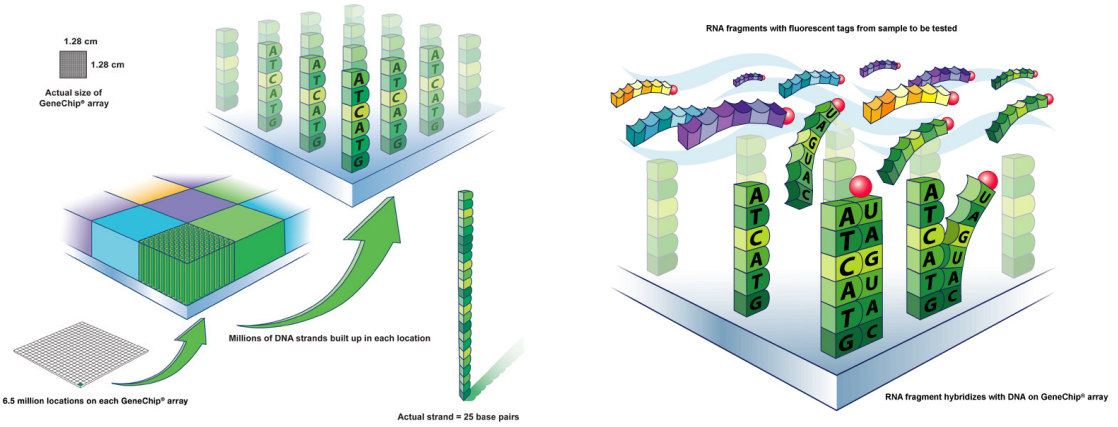


Figure B.1: Representation of a DNA chip and the hybridization of complementary DNA chains. Images courtesy of Affymetrix.[12]

One of the most popular experiments involving cDNA microarrays consists in compare mRNA abundance in two samples using one chip and two different fluorescent dyes (such as Cy3 and Cy5). The mRNA molecules extracted from two tissues of interest (e.g. tumour and normal tissue) are reverse transcribed from RNA to DNA and their concentration is enhanced. The resulting DNA is transcribed back into fluorescently marked single strand RNA: molecules coming from tumour tissue are labeled with a red dye while normal tissue molecules are labeled with a green one. The mixed solution of marked mRNA molecules (“targets”) is placed

on the chip and diffuses over the collection of single strand DNA probes at conditions promoting hybridization. When an mRNA encounters a part of the gene of which it is a perfect copy, it hybridizes to it with a high affinity (considerably higher than with a bit of DNA of which it is not a perfect copy) and when the mRNA solution is washed off, only those molecules that found their perfect match remain stuck to the chip. Subsequent illumination with appropriate laser wavelenghts will provide an image of the array in which the targets fluoresce. If RNA from tumour tissue is in abundance, the spot will emit red light, but if instead RNA from normal tissue is in abundance, it will appear green. If tumour and normal RNA bind equally, the spot will be yellow, while if neither binds, it will not fluoresce and appear black [13]. Fig B.2 is a simplified representation of these procedures. CDNA microarrays are a differential technique because only ratios between both fluorescence wavelengths give meaningful information and hence, only relative expressions levels are obtained.

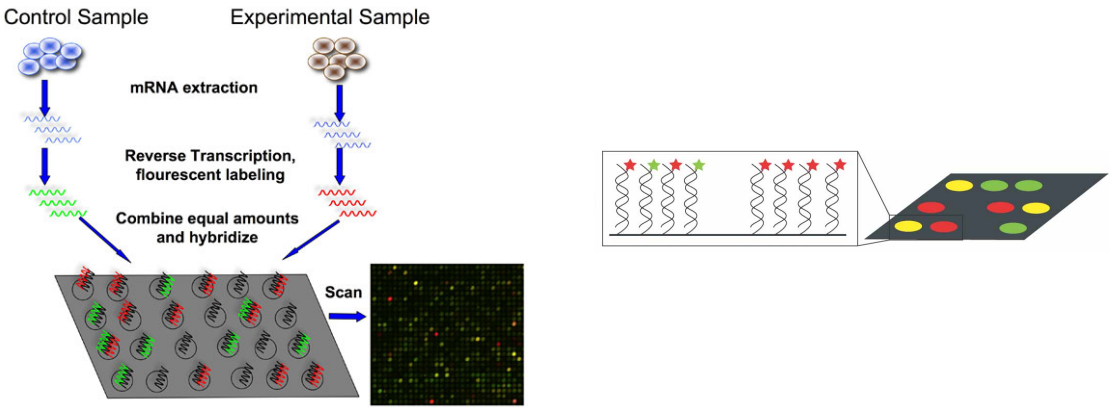


Figure B.2: Comparing normal and tumour gene expression levels with microarrays. Genes expressed only on tumour tissue appear red, while genes expressed only on normal tissue appear green. If the gen is expressed equally on both, the spot is yellow. Recovered from [21], [22]

The characterization of genes expressed differently in normal and their corresponding tumour cells has been a particularly important application of microarrays [14], but the potential of this technology is tremendous: monitoring gene expression levels in different developmental stages, tissue types, clinical conditions and different organisms can help understanding gene function and gene networks, assist in the diagnosis of disease conditions and reveal effects of medical treatments. For example, microaarrays have been used to discover transcribed regions in genomic DNA [15]; to detect polymorphism in copy number of regions of the genome[16], which may be a new and important class of mutation; and to analyze amplifications and deletions that are associated with oncogenic transformation and some inherited conditions [17], [18]. A few typical examples would include comparing diseased and drug-treated tissue, study cell phenomena over time as well as study the effect of various factors such as interferones, cytomegalovirus infection and oncogene transfection on the overall pattern of expression, but perhaps even more importantly than the success in any individual application, it has been shown that microarrays can be used to generate accurate, precise and reliable gene expression data [11]. In [7], [19] some other interesting microarray applica-

tions are reviewed and in [20] a very illustrative microarray virtual lab can be found.

Bibliography Appendix B

[1] G.J. McLachlan, K-A Do and C. Ambroise, *Analyzing Microarray Gene Expression Data*, John Wiley & Sons, New Jersey, USA, 2004.

[2] M. Schena, D. Shalon, R.W. Davis and P.O. Brown, *Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray*, Science, Vol. 270, 1995, p. 467.

[3] S.P.A. Fodor, J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu and D Solas, *Light-Directed, Spatially Addressable Parallel Chemical Synthesis*, Science, Vol 251, 1991, p. 767.

[4] T. Lenoir and E. Giannella, *The Emergence and Diffusion of DNA Microarray Technology*, Journal of Biomedical Discovery and Collaboration, 2006, Vol 1(10), p. 11.

[5] S.J. Wheelan, M.F. Martinez, J.D. Boeke, *The incredible shrinking world of DNA microarrays*, Mol Biosyst. Vol 4, 2008, p. 726.

[6] A. Schulze and J. Downward, *Navigating Gene Expression Using Microarrays — A Technology Review*, Nature Cell Biology, Vol. 3, 2001, p. E190.

[7] E.K. Lobenhofer, P. R. Bushel, C. A. Afshari and H.K. Hamadeh, *Progress in the Application of DNA Microarrays? - Reviews*, Environmental Health Perspectives, Vol. 109(9), 2001, p. 881.

[8] G. Gibson and S.V. Muse, *A Primer of Genome Science*, Sinauer Associates, Inc., Sunderland USA, 2004.

[9] H.C. Causton, J. Quackenbush and A. Brazma, *Microarray Gene Expression Data Analysis: a begginer's guide*, Wiley-Blackwell, Oxford UK, 2003.

[10] D. Gerhold, T. Rushmore, and C. T. Caskey, *DNA Chips: Promising Toys Have Become Powerful Tools*, TIBS, Vol. 24, 1999, p. 168.

[11] S. Draghici, *Data Analysis Tools for DNA Microarrays*, Chapman and Hall/CRC Mathematical Biology and Medicine Series, Boca Ratón, USA, 2003.

[12] `http://www.affymetrix.com`

[13] A. Brazma and J. Vilo, *Gene Expression Data Analysis*, FEBS letters, Vol. 480, 2000, p. 17.

[14] Golub,T.R. et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*, Science 286, 1999, p. 531.

[15] Hughes, T.R. et al., *Experimental Annotation of the Human Genome Using Microarray Technology*, Nature, Vol. 409, 2001, p. 922.

[16] Lucito, R. et al., *Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation*, Genome Res., Vol. 13, 2003, p. 2291.

[17] Anand, R. and Southern, E. M., *Pulsed Field Gel Electrophoresis*, in Gel Electrophoresis of Nucleic Acids, eds Rickwood, D. and Hames, B.D., IRL Press, Oxford, 1990, p.101.

[18] Edwin Southern, *Tools for Genomics*, Nature Medicine, Vol. 11(10), 2005, p. 1029.

[19] S. Khan, S. Chaturvedi, N. Goel, S. Bawa, and S. Drabu, *Review: DNA Microarray Technology and Drug Development*, Chron Young Sci, Vol. 1(1), 2010, p.1.

[20] `http://learn.genetics.utah.edu/content/labs/microarray/`

[21] Y. Grigoryev, *Introduction to DNA Microarrays*, Tech Tips, 2011.
`http://bitesizebio.com/articles/introduction-to-dna-microarrays/`

[22] K.B. Cederquist, S.L. Dean and C.D. Keating, *Encoded anisotropic particles for multiplexed bioanalysis* Nanomed Nanobiotechnol, Vol. 2(6), 2010, p. 578.
`http://wires.wiley.com/WileyCDA/WiresArticle/wisId-WNAN96.html`

Appendix C

Supplementary Information

Supplementary Information Appendix

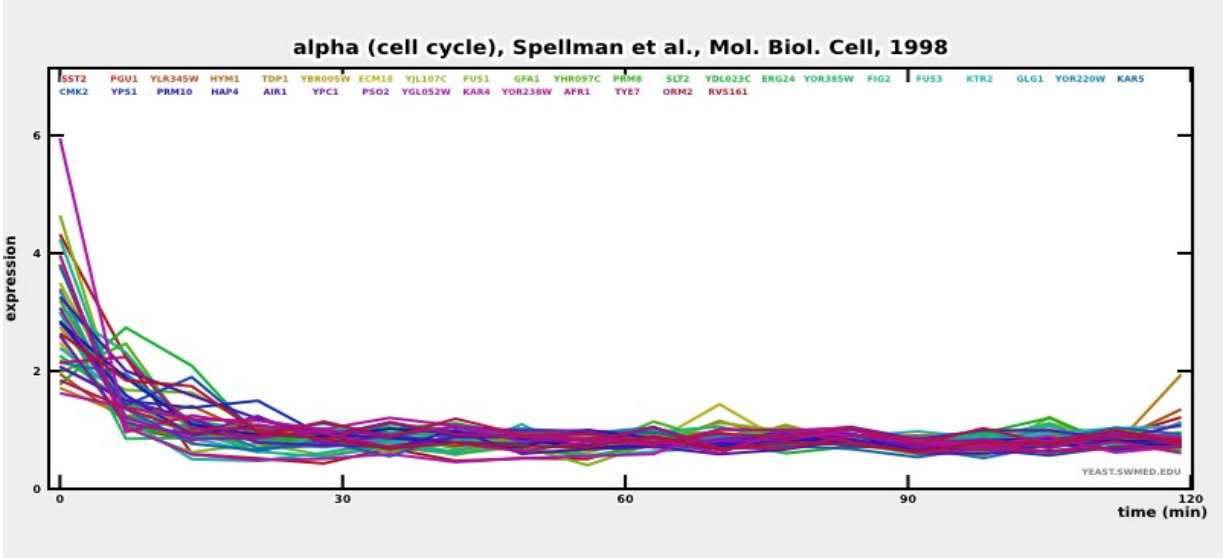
I.- Plots of temporal expression profiles for each cluster

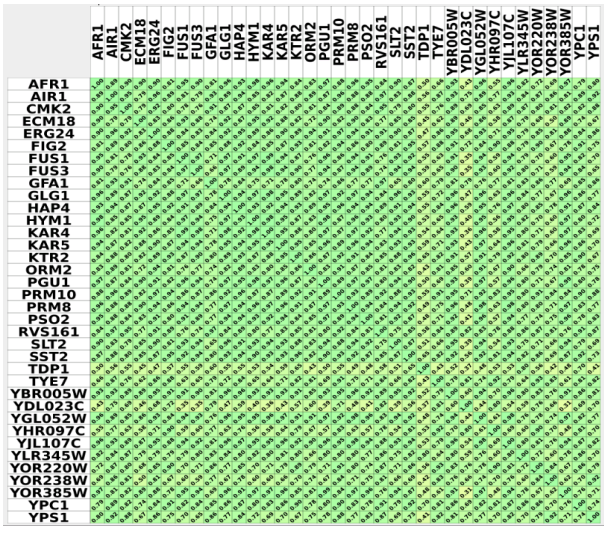
1.- Cluster 1037 M

YKL109W YHR030C YDR085C YLR120C YOL016C YDR125C YHR097C YCR089W YMR065W YKL104C YBR183W YDL023C YLR452C YCL027W YCL055W YCLX07W YBL016W YJR153W YCR009C YOR385W YIL079C YNL280C YKR058W YKL189W YKR061W YLR350W YJL108C YGL053W YGL052W YMR137C YBR223C YOR344C YBR005W YJL107C YLR345W YOR220W YOR238W

Cluster size: 37 genes
Hits with Spellman *et al.*: 7
Hits with all studies: 15
Unknown gene/ORF name: YCLX07W

PROPORTION: 15/37



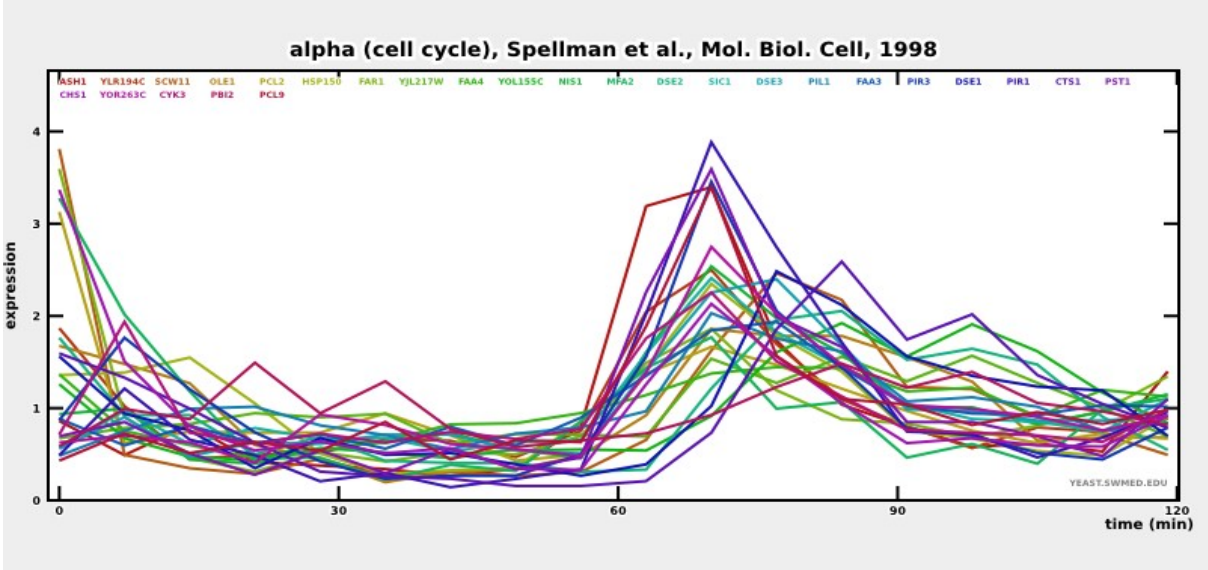


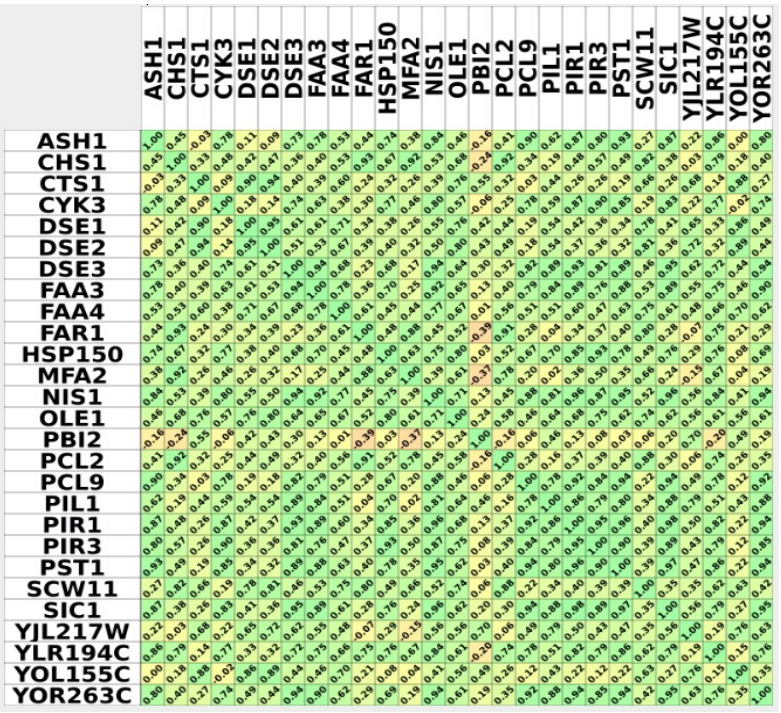
2.- Cluster 1098 CC

YDR055W YLR079W YGL055W YDL127W YKL185W YKL163W YNL192W YDL179W YKL164C YNL078W YJL159W YNL145W YOR264W YLR194C YIL009W YMR246W YJL157C YJL217W YGR086C YDL117W YOR263C YER124C YOL155C YGL028C YHR143W YLR286C YNL015W

Cluster size: 27 genes
Hits with Spellman *et al.*: 26
Hits with all studies: 27

PROPORTION: 27/27





3.- Cluster 2288 CC

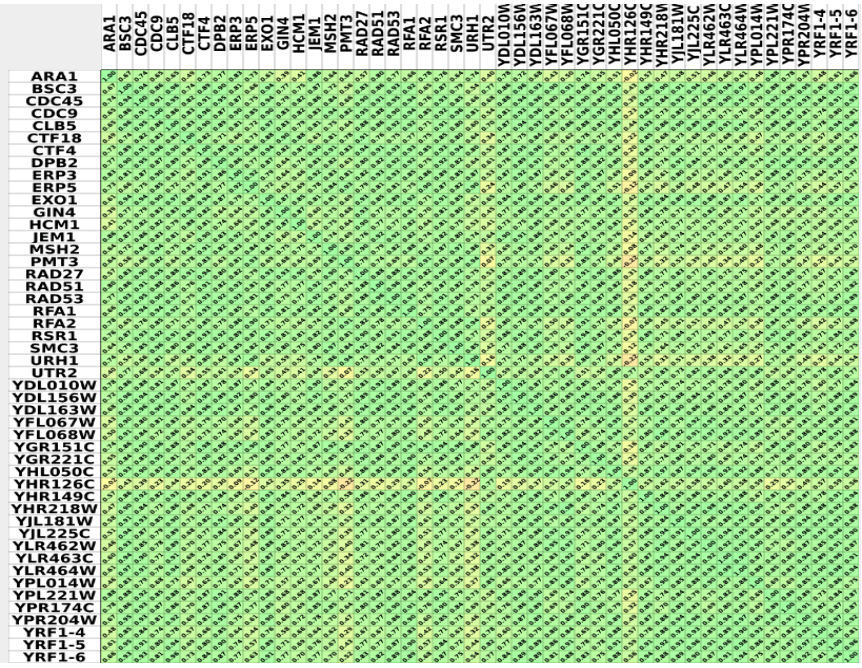
YLR462W YLR467W YLR466W YLR463C YER189W YLR464W YPR204W YPR203W YNL339C YEL077C YLR465C YPR120C YFL068W YFL064C YFL067W YPL283C YHR218W YHL049C YFL066C YEL076C YEL075C YJL225C YHL050C YGR296W YPR202W YHR149C YKL113C YDR507C YCR065W YMR078C YPL221W YGR152C YGR151C YPR174C YPR175W YLR103C YDL164C YPR135W YOL090W YDL163W YJL074C YGR221C YDL018C YJL073W YHR110W YDL156W YAR007C YDL010W YBL113C YBL111C YBR149W YPL014W YEL040W YNL312W YHR126C YER095W YPL153C YJL181W YOR321W YDR400W YOR033C

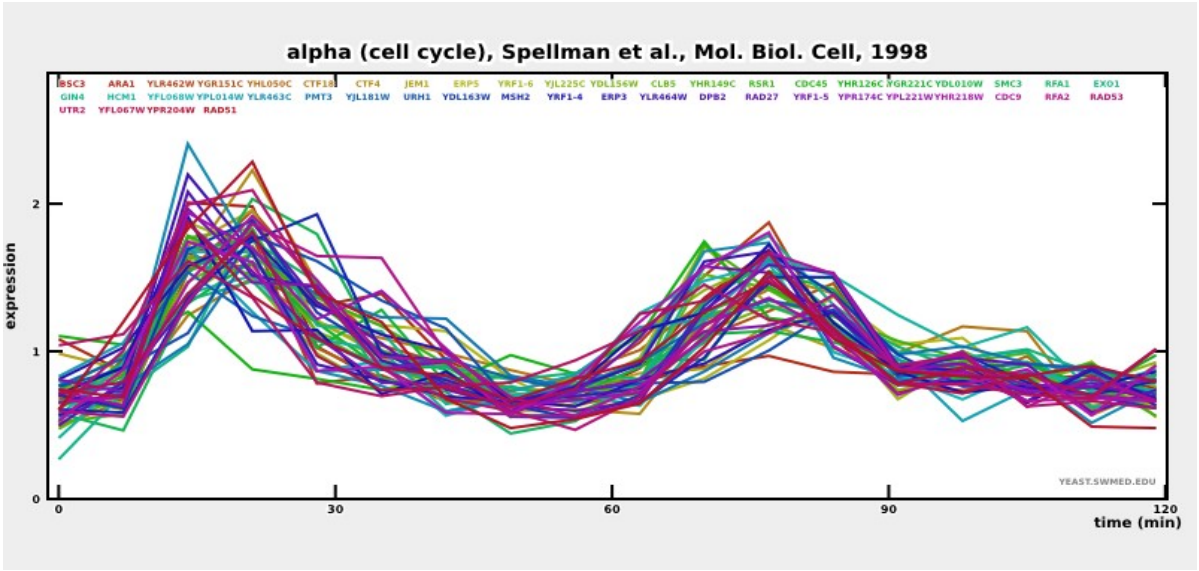
Cluster size: 61 genes
Hits with Spellman *et al.*: 59
Hits with all studies: 61
Genes included on this analysis: 48

PROPORTION: 61/61

List of not included genes, clasified by Spellman:

YBL111C	YBL111c
YBL113C	YBL113c
YEL075C	YEL075c
YEL076C	YEL076c
YEL077C	YEL077c
YER189W	YER189w
YFL064C	YFL064c
YFL066C	YFL066c
YGR296W	YRF1-3
YHL049C	YHL049c
YPL283C	YRF1-7
YPR202W	YPR202w
YPR203W	YPR203w



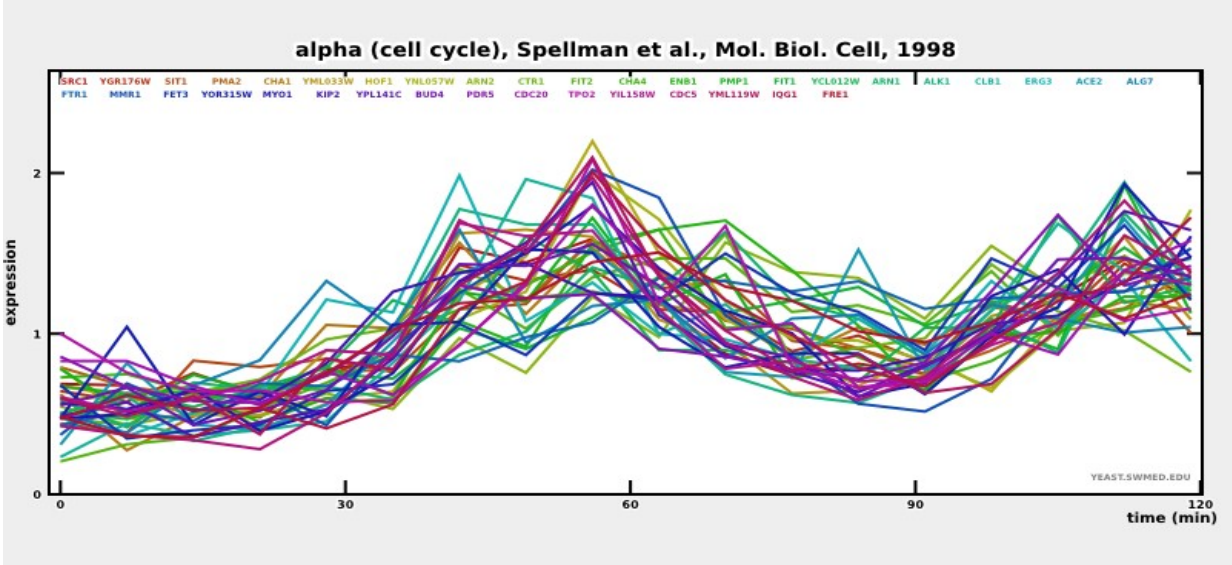


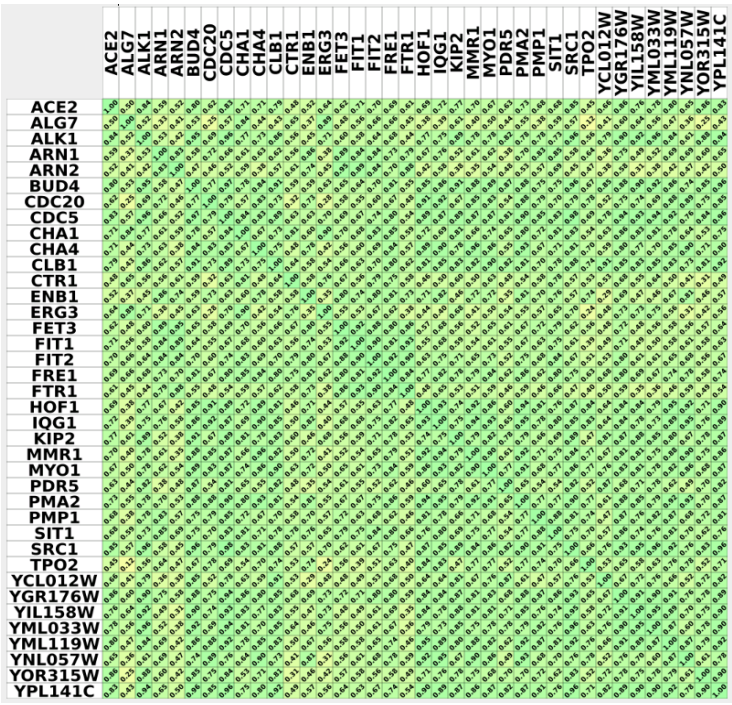
4.- Cluster 1225 CC

YJR092W YML119W YGR108W YML034W YMR032W YPL155C YLR131C YGL021W YPL141C YOR315W YPL242C YLR190W YHR023W YIL158W YMR001C YOR153W YNL057W YML033W YDR534C YEL065W YOR382W YHL040C YCL064C YMR058W YHL047C YLR214W YOL158C YER145C YLR056W YCR024C-A YBR243C YCL013W YGR176W YCL012W YPL036W YGL116W YGR138C YLR098C YPR124W

Cluster size: 39
Hits with Spellman *et al.*: 34
Hits with all studies: 38
Unknown gene/ORF name: YCL013W (classified by Spellman et al.)

PROPORTION: 38/39



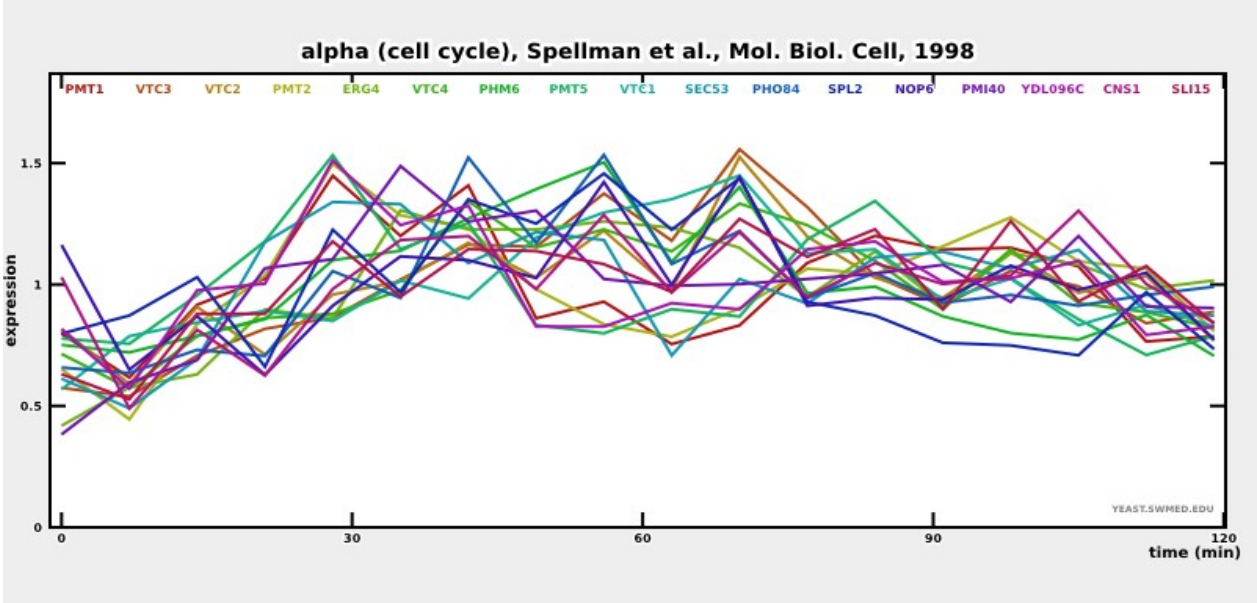


5.- Cluster 2183 M

YML123C YER072W YDR281C YHR136C YBR156C YDL213C YER003C YGL012W YDL095W YAL023C YFL045C YDL093W YDL096C YBR155W YFL004W YPL019C YJL012C

Cluster size: 17
Hits with Spellman *et al.*: 5
Hits with all studies: 13

PROPORTION: 13/17



	CNS1	ERG4	NOP6	PHM6	PHO84	PMI40	PMT1	PMT2	PMT5	SEC53	SLI15	SPL2	VTC1	VTC2	VTC3	VTC4	YDL096C
CNS1	1.00	0.60	0.82	0.54	0.66	0.47	0.40	0.43	0.07	0.51	0.57	0.41	0.38	0.74	0.63	0.63	0.39
ERG4	0.60	1.00	0.43	0.66	0.75	0.83	0.31	0.45	0.15	0.64	0.69	0.42	0.73	0.59	0.78	0.77	0.30
NOP6	0.82	0.43	1.00	0.68	0.64	0.16	0.05	0.10	-0.19	0.25	0.48	0.64	0.52	0.81	0.65	0.62	0.03
PHM6	0.54	0.66	0.68	1.00	0.83	0.54	0.20	0.20	0.12	0.50	0.59	0.89	0.75	0.70	0.77	0.75	0.19
PHO84	0.66	0.75	0.64	0.83	1.00	0.53	0.33	0.30	0.13	0.48	0.66	0.77	0.63	0.70	0.77	0.81	0.27
PMI40	0.47	0.83	0.16	0.54	0.53	1.00	0.57	0.69	0.43	0.84	0.56	0.25	0.44	0.38	0.55	0.55	0.57
PMT1	0.40	0.31	0.05	0.20	0.33	0.57	1.00	0.87	0.88	0.72	0.52	0.06	-0.04	0.26	0.19	0.34	0.94
PMT2	0.43	0.45	0.10	0.20	0.30	0.69	0.87	1.00	0.72	0.81	0.66	0.04	0.06	0.35	0.26	0.36	0.85
PMT5	0.07	0.15	-0.19	0.12	0.13	0.43	0.88	0.72	1.00	0.55	0.43	0.03	0.01	0.11	0.10	0.23	0.90
SEC53	0.51	0.64	0.25	0.50	0.48	0.84	0.72	0.81	0.55	1.00	0.63	0.18	0.28	0.38	0.44	0.44	0.64
SLI15	0.57	0.69	0.48	0.59	0.66	0.56	0.52	0.66	0.43	0.63	1.00	0.47	0.65	0.81	0.75	0.82	0.51
SPL2	0.41	0.42	0.64	0.89	0.77	0.25	0.06	0.04	0.03	0.18	0.47	1.00	0.63	0.64	0.62	0.61	0.11
VTC1	0.38	0.73	0.52	0.75	0.63	0.44	-0.04	0.06	0.01	0.28	0.65	0.63	1.00	0.73	0.88	0.80	0.00
VTC2	0.74	0.59	0.81	0.70	0.70	0.38	0.26	0.35	0.11	0.38	0.81	0.64	0.73	1.00	0.88	0.88	0.27
VTC3	0.63	0.78	0.65	0.77	0.77	0.55	0.19	0.26	0.10	0.44	0.75	0.62	0.88	0.88	1.00	0.94	0.21
VTC4	0.63	0.77	0.62	0.75	0.81	0.55	0.34	0.36	0.23	0.44	0.82	0.61	0.80	0.88	0.94	1.00	0.33
YDL096C	0.39	0.30	0.03	0.19	0.27	0.57	0.94	0.85	0.90	0.64	0.51	0.11	0.00	0.27	0.21	0.33	1.00

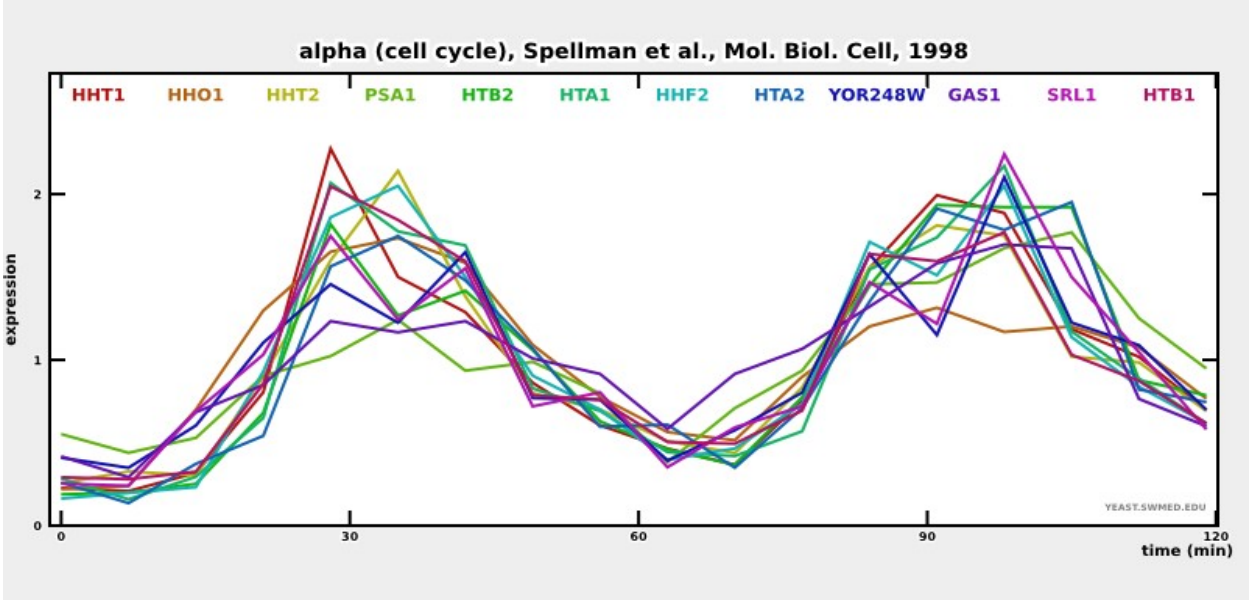
6.- Cluster 1137 CC

YPL127C YNL031C YOR248W YBR009C YDR225W YBR010W YDR224C YDL055C YNL030W YBL003C YOR247W YBL002W YMR307W

Cluster size: 13
Hits with Spellman *et al.*: 13
Hits with all studies: 13
Genes included in this analysis: 12

This one was left out:
YBR009C: HHF1: One of two identical histone H4 proteins (see also HHF2); core histone required for chromatin assembly and chromosome function; contributes to telomeric silencing; N-terminal domain involved in maintaining genomic integrity

PROPORTION: 13/13



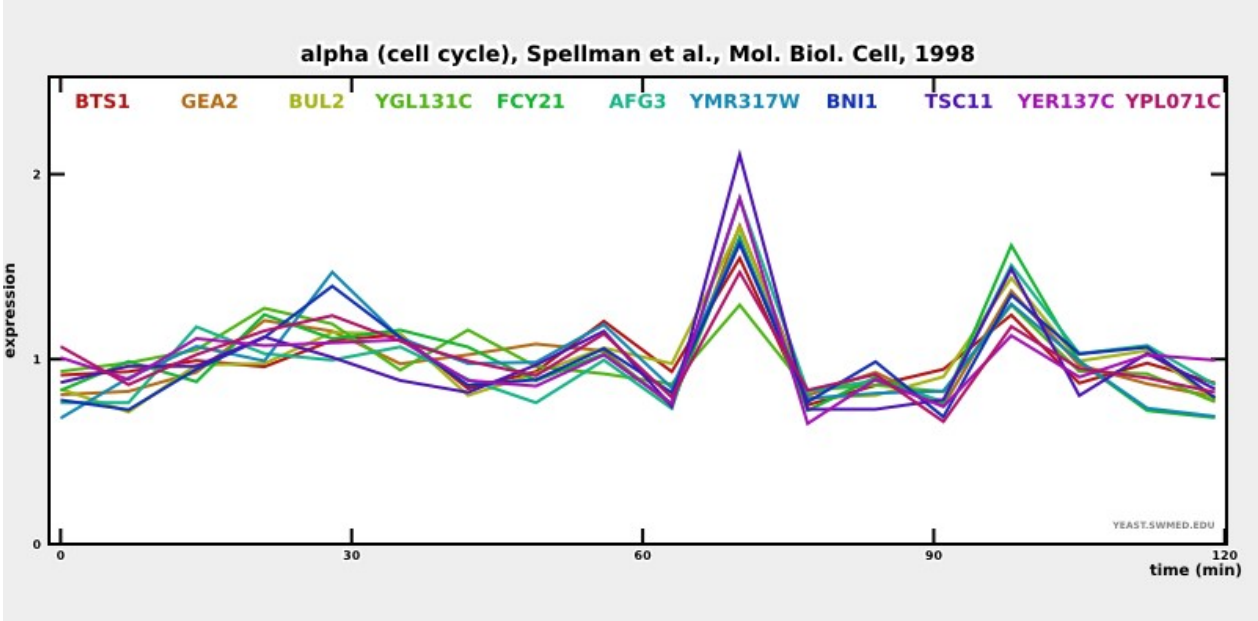
	GAS1	HHF2	HHO1	HHT1	HHT2	HTA1	HTA2	HTB1	HTB2	PSA1	SRL1	YOR248W
GAS1	1.00	0.82	0.72	0.81	0.78	0.83	0.90	0.78	0.92	0.87	0.87	0.82
HHF2	0.82	1.00	0.88	0.93	0.97	0.97	0.89	0.98	0.88	0.76	0.90	0.91
HHO1	0.72	0.88	1.00	0.84	0.87	0.84	0.82	0.88	0.79	0.65	0.80	0.79
HHT1	0.81	0.93	0.84	1.00	0.92	0.96	0.89	0.96	0.93	0.76	0.88	0.85
HHT2	0.78	0.97	0.87	0.92	1.00	0.94	0.89	0.96	0.85	0.75	0.82	0.83
HTA1	0.83	0.97	0.84	0.96	0.94	1.00	0.91	0.98	0.91	0.75	0.92	0.91
HTA2	0.90	0.89	0.82	0.89	0.89	0.91	1.00	0.87	0.97	0.86	0.84	0.80
HTB1	0.78	0.98	0.88	0.96	0.96	0.98	0.87	1.00	0.87	0.70	0.88	0.88
HTB2	0.92	0.88	0.79	0.93	0.85	0.91	0.97	0.87	1.00	0.88	0.89	0.84
PSA1	0.87	0.76	0.65	0.76	0.75	0.75	0.86	0.70	0.88	1.00	0.80	0.79
SRL1	0.87	0.90	0.80	0.88	0.82	0.92	0.84	0.88	0.89	0.80	1.00	0.97
YOR248W	0.82	0.91	0.79	0.85	0.83	0.91	0.80	0.88	0.84	0.79	0.97	1.00

7.- Cluster 1459 N

YER017C YER137C YML111W YNL271C YEL022W YPL071C YPL069C YGL131C YER060W YER093C YMR317W

Cluster size: 11
Hits with Spellman *et al.*: 0
Hits with all studies: 1

PROPORTION: 1/11



Correlations in alpha:

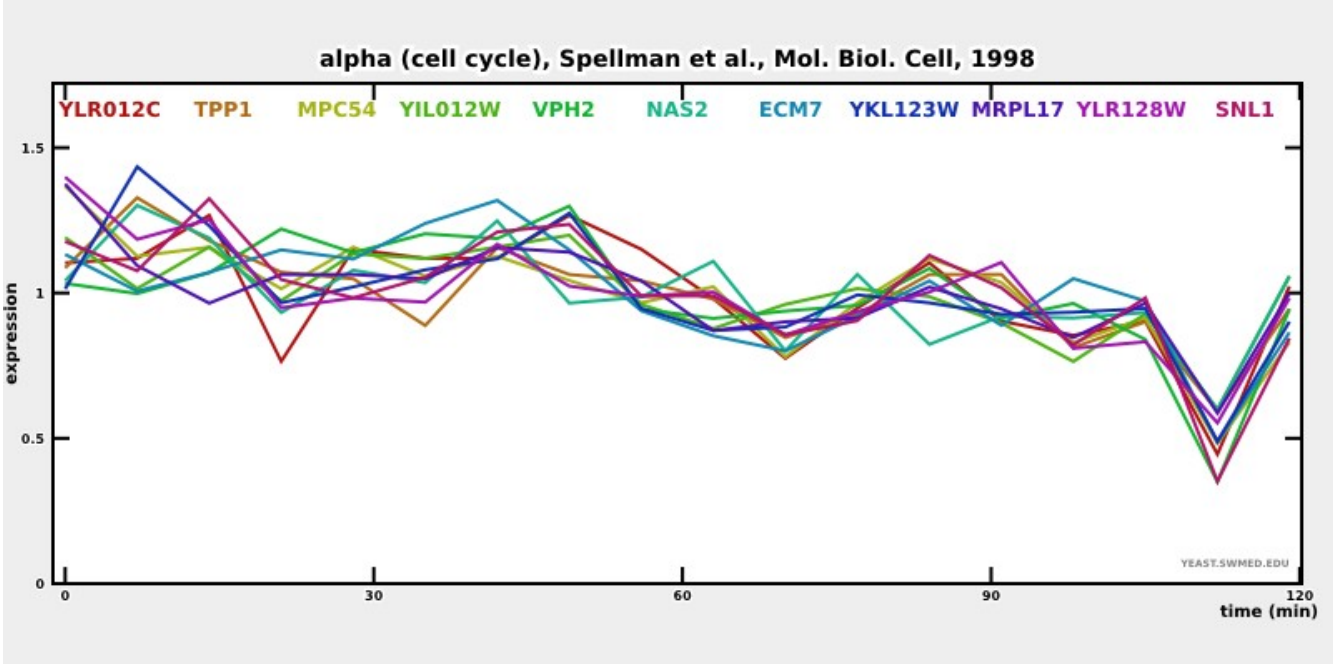
	AFG3	BNI1	BTS1	BUL2	FCY21	GEA2	TSC11	YER137C	YGL131C	YMR317W	YPL071C
AFG3	1.00	0.86	0.84	0.91	0.83	0.87	0.90	0.88	0.69	0.76	0.78
BNI1	0.86	1.00	0.83	0.90	0.83	0.89	0.80	0.82	0.74	0.87	0.88
BTS1	0.84	0.83	1.00	0.92	0.79	0.81	0.91	0.87	0.59	0.84	0.80
BUL2	0.91	0.90	0.92	1.00	0.82	0.86	0.88	0.82	0.65	0.83	0.77
FCY21	0.83	0.83	0.79	0.82	1.00	0.92	0.83	0.74	0.86	0.83	0.83
GEA2	0.87	0.89	0.81	0.86	0.92	1.00	0.91	0.83	0.84	0.86	0.87
TSC11	0.90	0.80	0.91	0.88	0.83	0.91	1.00	0.90	0.71	0.77	0.80
YER137C	0.88	0.82	0.87	0.82	0.74	0.83	0.90	1.00	0.64	0.73	0.85
YGL131C	0.69	0.74	0.59	0.65	0.86	0.84	0.71	0.64	1.00	0.75	0.80
YMR317W	0.76	0.87	0.84	0.83	0.83	0.86	0.77	0.73	0.75	1.00	0.83
YPL071C	0.78	0.88	0.80	0.77	0.83	0.87	0.80	0.85	0.80	0.83	1.00

8.- Cluster 1934 N

YIL016W YOR177C YLR012C YMR156C YLR128W YNL252C YKL123W YLR443W YIL007C YIL012W YKL119C

Cluster size: 11
Hits with Spellman *et al.*: 0
Hits with all studies: 1

PROPORTION: 1/11



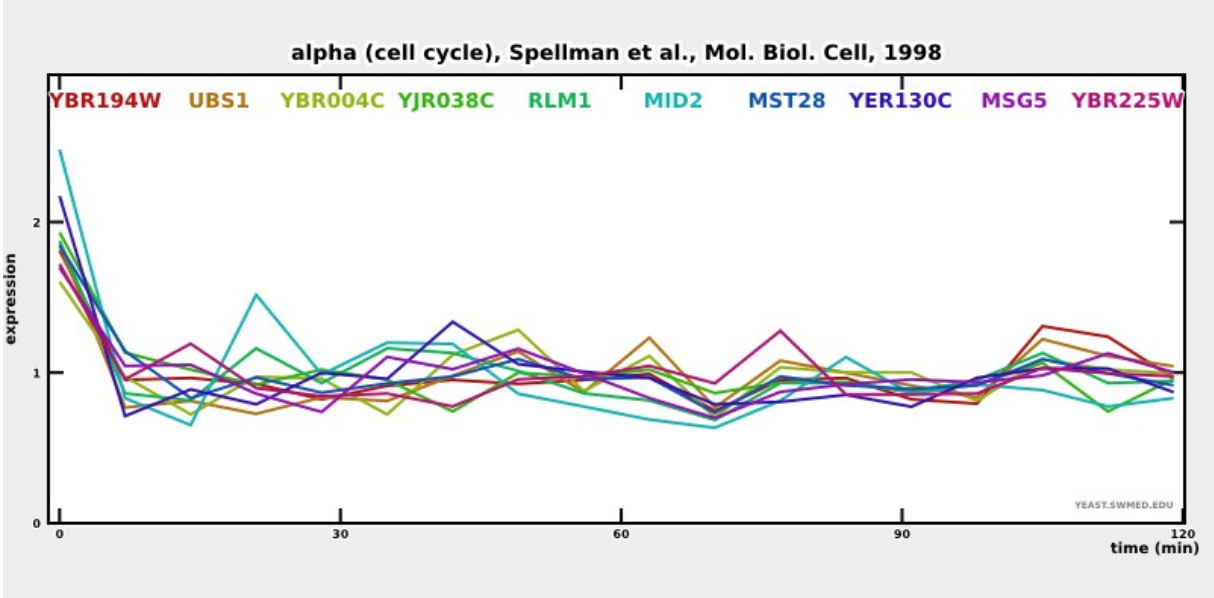
	ECM7	MPC54	MRPL17	NAS2	SNL1	TPP1	VPH2	YIL012W	YKL123W	YLR012C	YLR128W
ECM7	1.00	0.76	0.78	0.59	0.81	0.61	0.90	0.76	0.70	0.69	0.60
MPC54	0.76	1.00	0.86	0.67	0.88	0.82	0.72	0.79	0.70	0.78	0.90
MRPL17	0.78	0.86	1.00	0.56	0.78	0.71	0.74	0.84	0.65	0.71	0.81
NAS2	0.59	0.67	0.56	1.00	0.66	0.78	0.54	0.65	0.77	0.71	0.73
SNL1	0.81	0.88	0.78	0.66	1.00	0.83	0.83	0.83	0.81	0.85	0.83
TPP1	0.61	0.82	0.71	0.78	0.83	1.00	0.64	0.67	0.84	0.74	0.85
VPH2	0.90	0.72	0.74	0.54	0.83	0.64	1.00	0.81	0.73	0.73	0.58
YIL012W	0.76	0.79	0.84	0.65	0.83	0.67	0.81	1.00	0.73	0.84	0.76
YKL123W	0.70	0.70	0.65	0.77	0.81	0.84	0.73	0.73	1.00	0.80	0.69
YLR012C	0.69	0.78	0.71	0.71	0.85	0.74	0.73	0.84	0.80	1.00	0.74
YLR128W	0.60	0.90	0.81	0.73	0.83	0.85	0.58	0.76	0.69	0.74	1.00

9.- Cluster 1324 N

YPL089C YNL053W YAR033W YBR004C YBR194W YBR165W YJR038C YER130C YLR332W YBR225W

Cluster size: 10
Hits with Spellman *et al.*: 0
Hits with all studies: 2

PROPORTION: 2/10



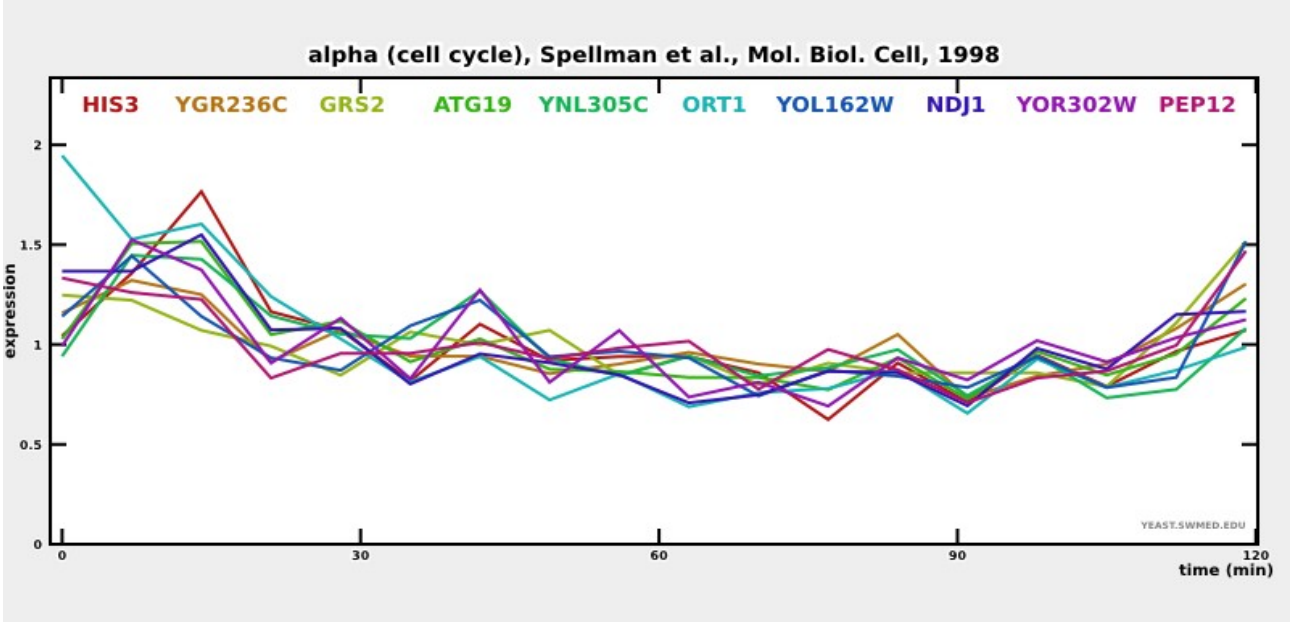
	MID2	MSG5	MST28	RLM1	UBS1	YBR004C	YBR194W	YBR225W	YER130C	YJR038C
MID2	1.00	0.71	0.81	0.95	0.60	0.64	0.69	0.55	0.80	0.75
MSG5	0.71	1.00	0.87	0.81	0.73	0.67	0.82	0.68	0.81	0.74
MST28	0.81	0.87	1.00	0.88	0.83	0.84	0.89	0.77	0.85	0.89
RLM1	0.95	0.81	0.88	1.00	0.72	0.72	0.81	0.62	0.87	0.77
UBS1	0.60	0.73	0.83	0.72	1.00	0.87	0.88	0.77	0.82	0.73
YBR004C	0.64	0.67	0.84	0.72	0.87	1.00	0.74	0.58	0.74	0.65
YBR194W	0.69	0.82	0.89	0.81	0.88	0.74	1.00	0.79	0.82	0.77
YBR225W	0.55	0.68	0.77	0.62	0.77	0.58	0.79	1.00	0.67	0.82
YER130C	0.80	0.81	0.85	0.87	0.82	0.74	0.82	0.67	1.00	0.76
YJR038C	0.75	0.74	0.89	0.77	0.73	0.65	0.77	0.82	0.76	1.00

10.- Cluster 2635 N

YOR130C YOL082W YNL305C YGR236C YOR036W YPR081C YOL162W YOR202W YOR302W YOL104C

Cluster size: 10
Hits with Spellman *et al.*: 0
Hits with all studies: 1

PROPORTION: 1/10



	ATG19	GRS2	HIS3	NDJ1	ORT1	PEP12	YGR236C	YNL305C	YOL162W	YOR302W
ATG19	1.00	0.57	0.91	0.87	0.72	0.69	0.86	0.88	0.73	0.88
GRS2	0.57	1.00	0.40	0.61	0.50	0.87	0.73	0.44	0.84	0.40
HIS3	0.91	0.40	1.00	0.83	0.71	0.54	0.71	0.84	0.55	0.80
NDJ1	0.87	0.61	0.83	1.00	0.88	0.73	0.81	0.66	0.59	0.75
ORT1	0.72	0.50	0.71	0.88	1.00	0.65	0.68	0.60	0.52	0.60
PEP12	0.69	0.87	0.54	0.73	0.65	1.00	0.87	0.52	0.85	0.56
YGR236C	0.86	0.73	0.71	0.81	0.68	0.87	1.00	0.64	0.73	0.71
YNL305C	0.88	0.44	0.84	0.66	0.60	0.52	0.64	1.00	0.72	0.77
YOL162W	0.73	0.84	0.55	0.59	0.52	0.85	0.73	0.72	1.00	0.67
YOR302W	0.88	0.40	0.80	0.75	0.60	0.56	0.71	0.77	0.67	1.00

11.- Cluster 2864 N

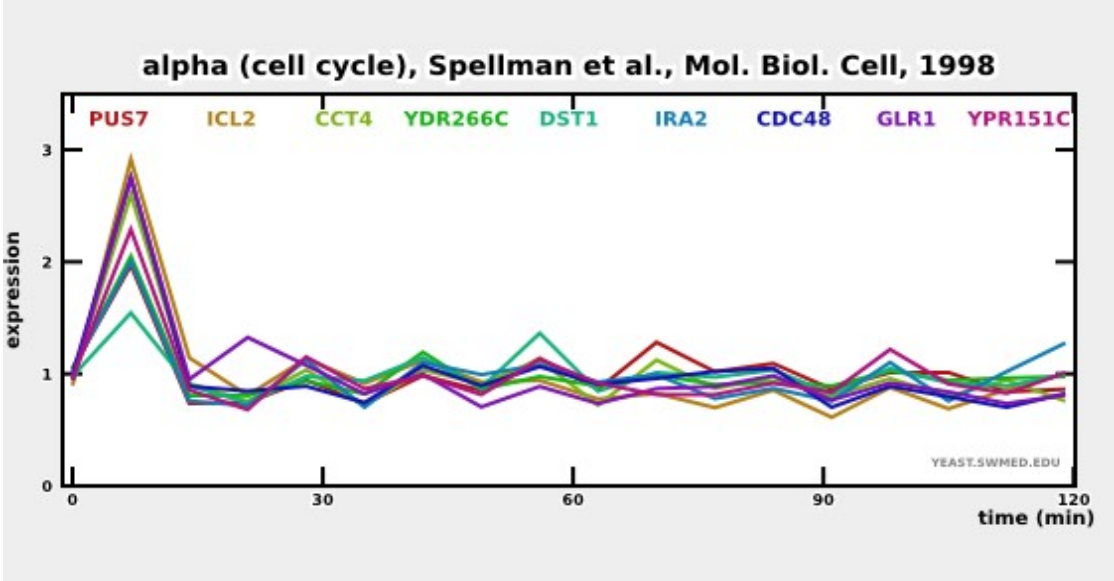
YDL143W YIL172C YDL126C YPR006C YPL091W YOR243C YDR266C YGL043W YPR151C YOL081W

Cluster size: 10
Hits with Spellman *et al.*: 0
Hits with all studies: 1

Genes included in this analysis: 9

This gene was not included:
YIL172C: Putative protein of unknown function with similarity to glucosidases; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm

PROPORTION: 1/10



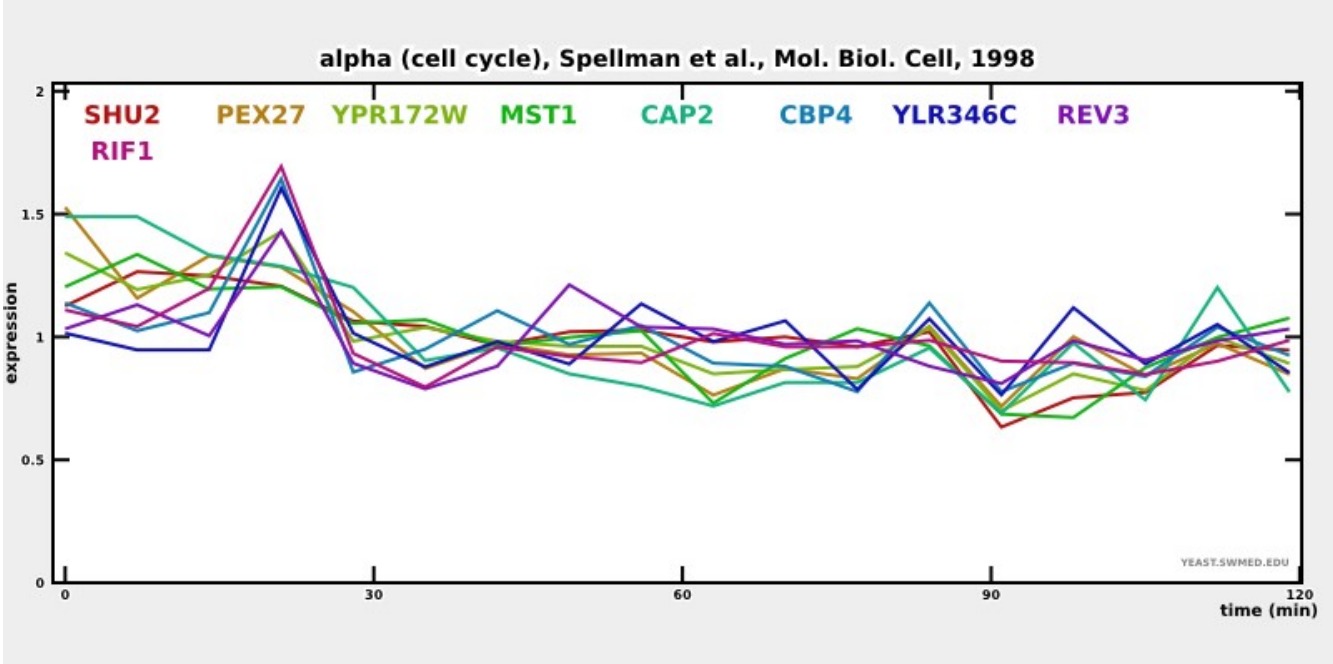
	CCT4	CDC48	DST1	GLR1	ICL2	IRA2	PUS7	YDR266C	YPR151C
CCT4	1.00	0.97	0.82	0.92	0.95	0.87	0.93	0.95	0.93
CDC48	0.97	1.00	0.80	0.94	0.94	0.84	0.91	0.95	0.92
DST1	0.82	0.80	1.00	0.67	0.75	0.80	0.84	0.81	0.83
GLR1	0.92	0.94	0.67	1.00	0.93	0.77	0.81	0.89	0.87
ICL2	0.95	0.94	0.75	0.93	1.00	0.88	0.79	0.92	0.93
IRA2	0.87	0.84	0.80	0.77	0.88	1.00	0.81	0.91	0.91
PUS7	0.93	0.91	0.84	0.81	0.79	0.81	1.00	0.91	0.86
YDR266C	0.95	0.95	0.81	0.89	0.92	0.91	0.91	1.00	0.93
YPR151C	0.93	0.92	0.83	0.87	0.93	0.91	0.86	0.93	1.00

12.- Cluster 1437 N

YBR275C YLR346C YGR174C YPL167C YIL034C YOR193W YPR172W YDR078C YKL194C\\

Cluster size: 9
Hits with Spellman *et al.*: 0
Hits with all studies: 1

PROPORTION: 1/9



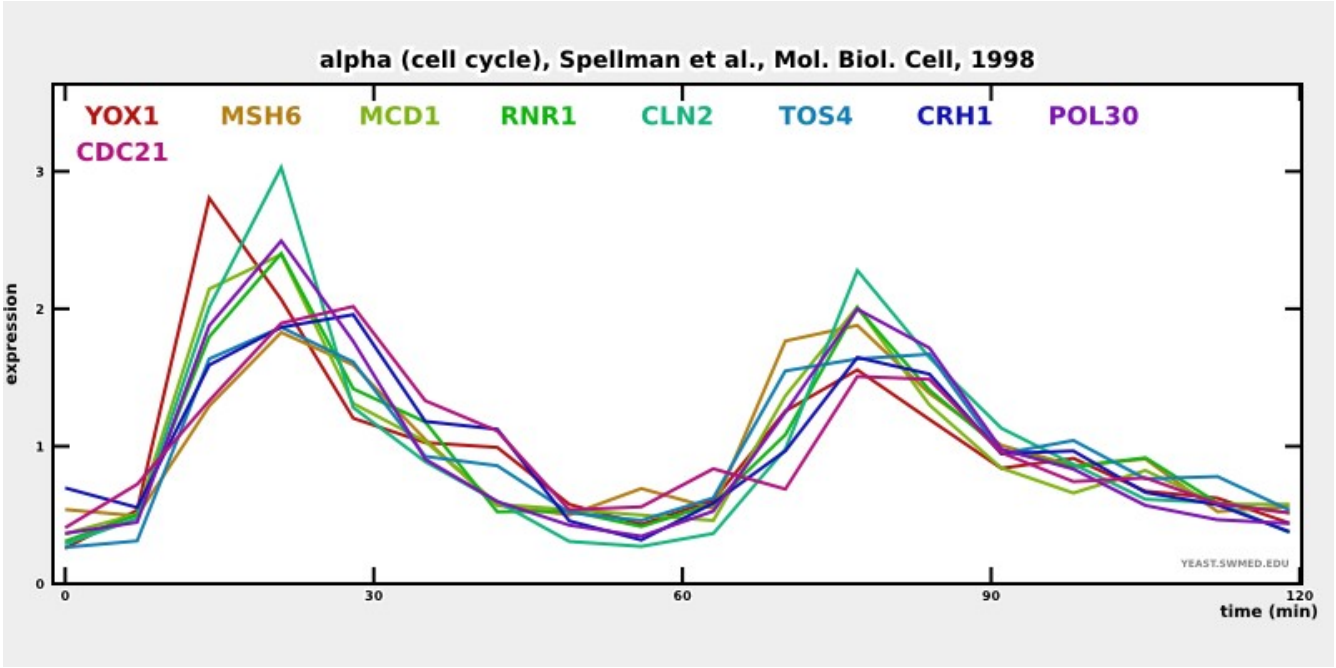
	CAP2	CBP4	MST1	PEX27	REV3	RIF1	SHU2	YLR346C	YPR172W
CAP2	1.00	0.55	0.73	0.91	0.37	0.51	0.72	0.38	0.85
CBP4	0.55	1.00	0.52	0.65	0.67	0.84	0.60	0.83	0.83
MST1	0.73	0.52	1.00	0.68	0.43	0.46	0.87	0.22	0.81
PEX27	0.91	0.65	0.68	1.00	0.41	0.60	0.70	0.47	0.90
REV3	0.37	0.67	0.43	0.41	1.00	0.76	0.51	0.64	0.57
RIF1	0.51	0.84	0.46	0.60	0.76	1.00	0.56	0.74	0.75
SHU2	0.72	0.60	0.87	0.70	0.51	0.56	1.00	0.38	0.85
YLR346C	0.38	0.83	0.22	0.47	0.64	0.74	0.38	1.00	0.58
YPR172W	0.85	0.83	0.81	0.90	0.57	0.75	0.85	0.58	1.00

13.- Cluster 4922 CC

YPL256C YER070W YDL003W YLR183C YML027W YBR088C YDR097C YOR074C YGR189C

Cluster size: 9
Hits with Spellman *et al.*: 9
Hits with all studies: 9

PROPORTION: 9/9



	CDC21	CLN2	CRH1	MCD1	MSH6	POL30	RNR1	TOS4	YOX1
CDC21	1.00	0.81	0.94	0.77	0.75	0.87	0.84	0.81	0.70
CLN2	0.81	1.00	0.86	0.96	0.85	0.97	0.98	0.88	0.86
CRH1	0.94	0.86	1.00	0.84	0.83	0.93	0.88	0.90	0.79
MCD1	0.77	0.96	0.84	1.00	0.89	0.96	0.97	0.90	0.92
MSH6	0.75	0.85	0.83	0.89	1.00	0.91	0.89	0.92	0.72
POL30	0.87	0.97	0.93	0.96	0.91	1.00	0.97	0.95	0.86
RNR1	0.84	0.98	0.88	0.97	0.89	0.97	1.00	0.91	0.86
TOS4	0.81	0.88	0.90	0.90	0.92	0.95	0.91	1.00	0.84
YOX1	0.70	0.86	0.79	0.92	0.72	0.86	0.86	0.84	1.00

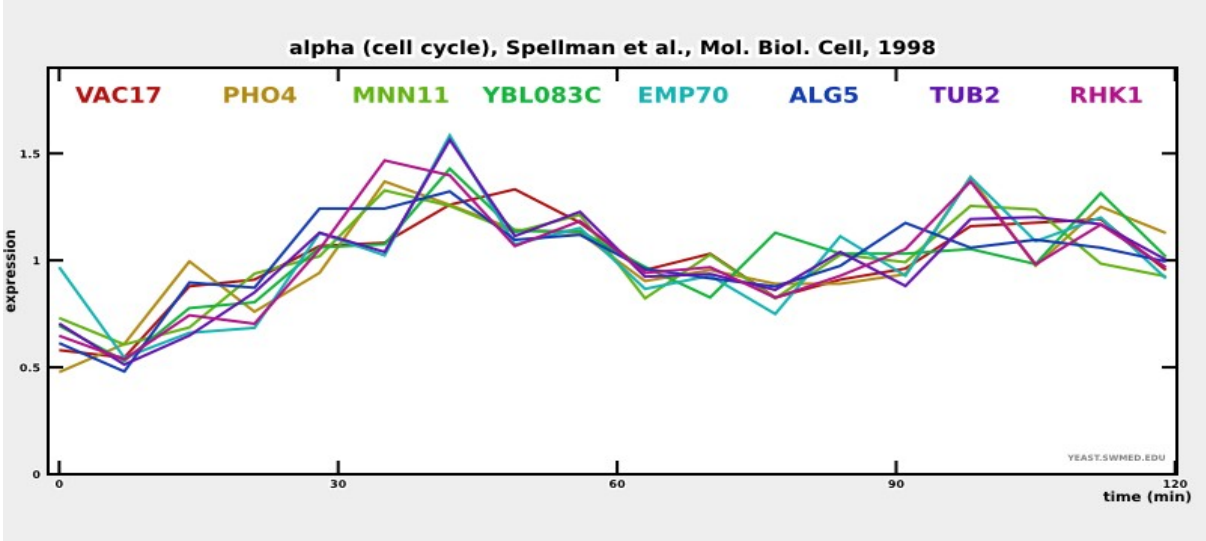
14.- Cluster 2719 M

YCL063W YPL227C YBL082C YJL183W YFL037W YCL062W YLR083C YBL083C YFR034C

Cluster size: 9
Hits with Spellman *et al.*: 3
Hits with all studies: 7
Genes included in this analysis: 8

Gene not included:
YCL062W: Protein involved in vacuole inheritance; acts as a vacuole-specific receptor for myosin Myo2p

PROPORTION: 7/9



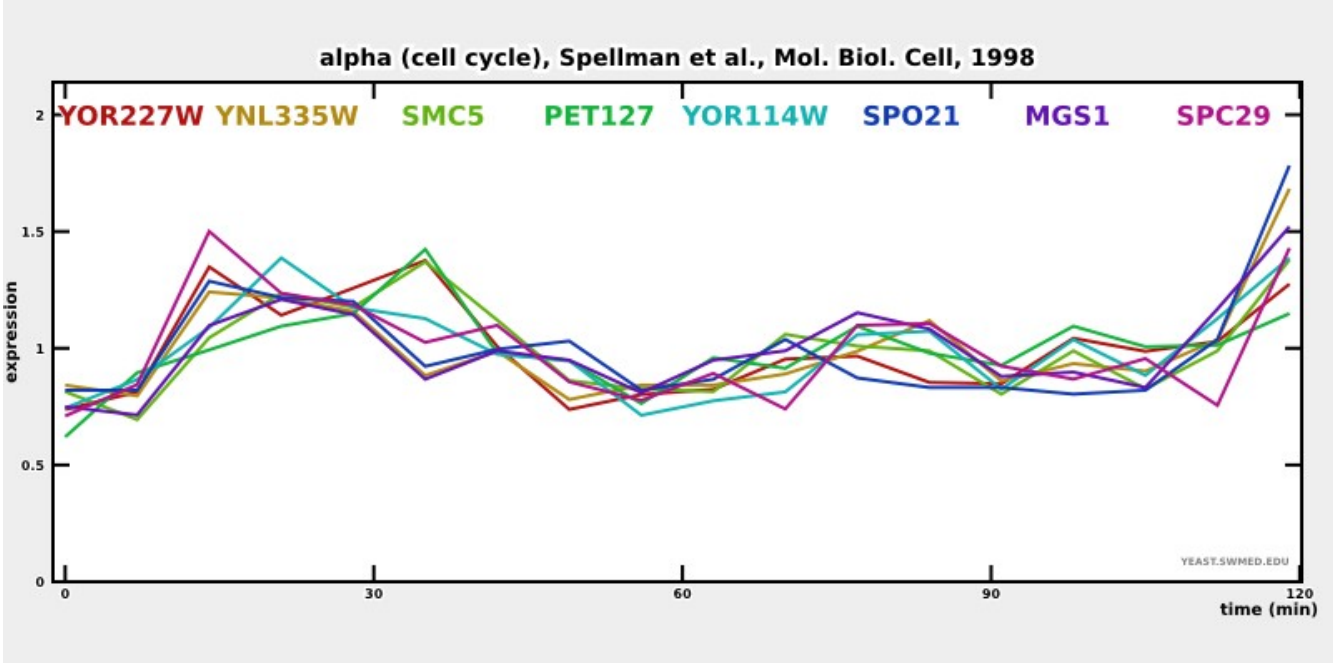
	ALG5	EMP70	MNN11	PHO4	RHK1	TUB2	VAC17	YBL083C
ALG5	1.00	0.70	0.81	0.77	0.86	0.82	0.85	0.82
EMP70	0.70	1.00	0.77	0.64	0.81	0.91	0.72	0.77
MNN11	0.81	0.77	1.00	0.75	0.88	0.85	0.83	0.67
PHO4	0.77	0.64	0.75	1.00	0.90	0.72	0.82	0.74
RHK1	0.86	0.81	0.88	0.90	1.00	0.82	0.80	0.79
TUB2	0.82	0.91	0.85	0.72	0.82	1.00	0.86	0.87
VAC17	0.85	0.72	0.83	0.82	0.80	0.86	1.00	0.79
YBL083C	0.82	0.77	0.67	0.74	0.79	0.87	0.79	1.00

15.- Cluster 1374 M

YNL335W YOL091W YPL124W YNL218W YOL034W YOR114W YOR017W YOR227W

Cluster size: 8
Hits with Spellman *et al.*: 3
Hits with all studies: 7

PROPORTION: 7/8



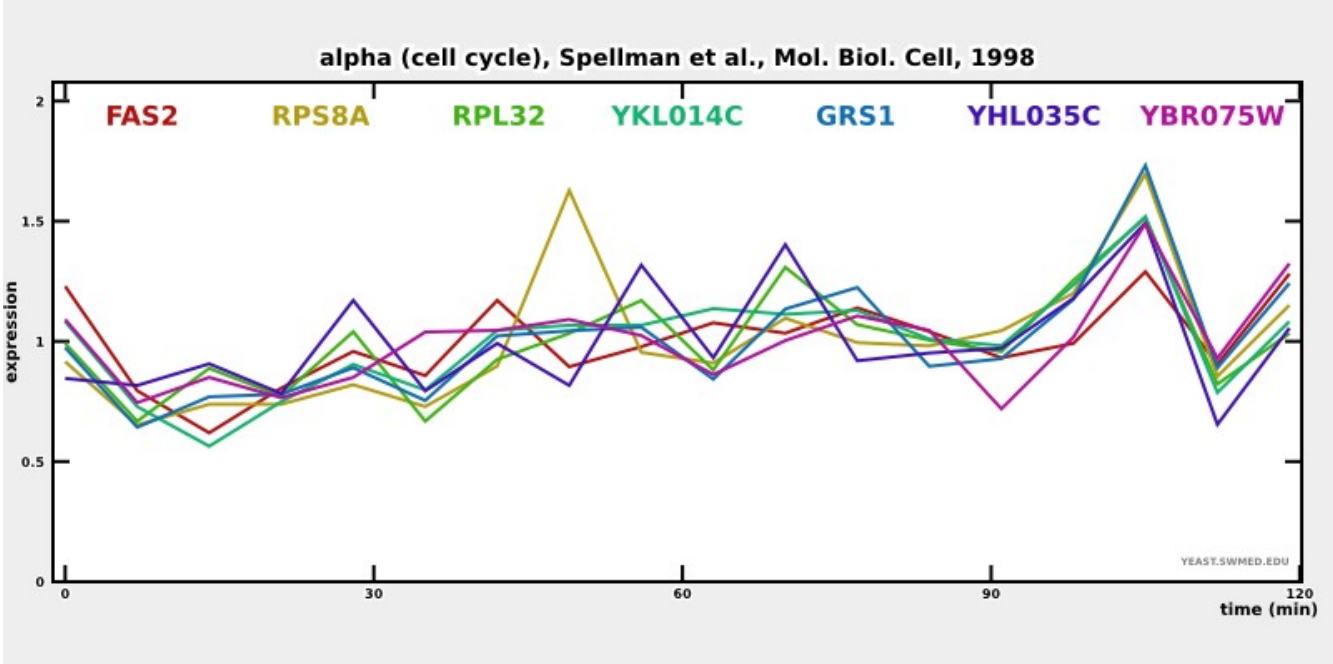
	MGS1	PET127	SMC5	SPC29	SPO21	YNL335W	YOR114W	YOR227W
MGS1	1.00	0.45	0.69	0.69	0.84	0.90	0.82	0.55
PET127	0.45	1.00	0.75	0.50	0.35	0.38	0.70	0.78
SMC5	0.69	0.75	1.00	0.63	0.69	0.69	0.81	0.83
SPC29	0.69	0.50	0.63	1.00	0.71	0.81	0.73	0.72
SPO21	0.84	0.35	0.69	0.71	1.00	0.89	0.72	0.63
YNL335W	0.90	0.38	0.69	0.81	0.89	1.00	0.81	0.65
YOR114W	0.82	0.70	0.81	0.73	0.72	0.81	1.00	0.74
YOR227W	0.55	0.78	0.83	0.72	0.63	0.65	0.74	1.00

16.- Cluster 2097 N

YBL072C YBR121C YBL092W YKL014C YPL231W YBR075W YHL035C YAR074C

Cluster size: 8
Hits with Spellman *et al.*: 0
Hits with all studies: 0
Genes included in this analysis: 7
Unknown gene/ORF name(s), 'YAR074C'

PROPORTION: 0/8

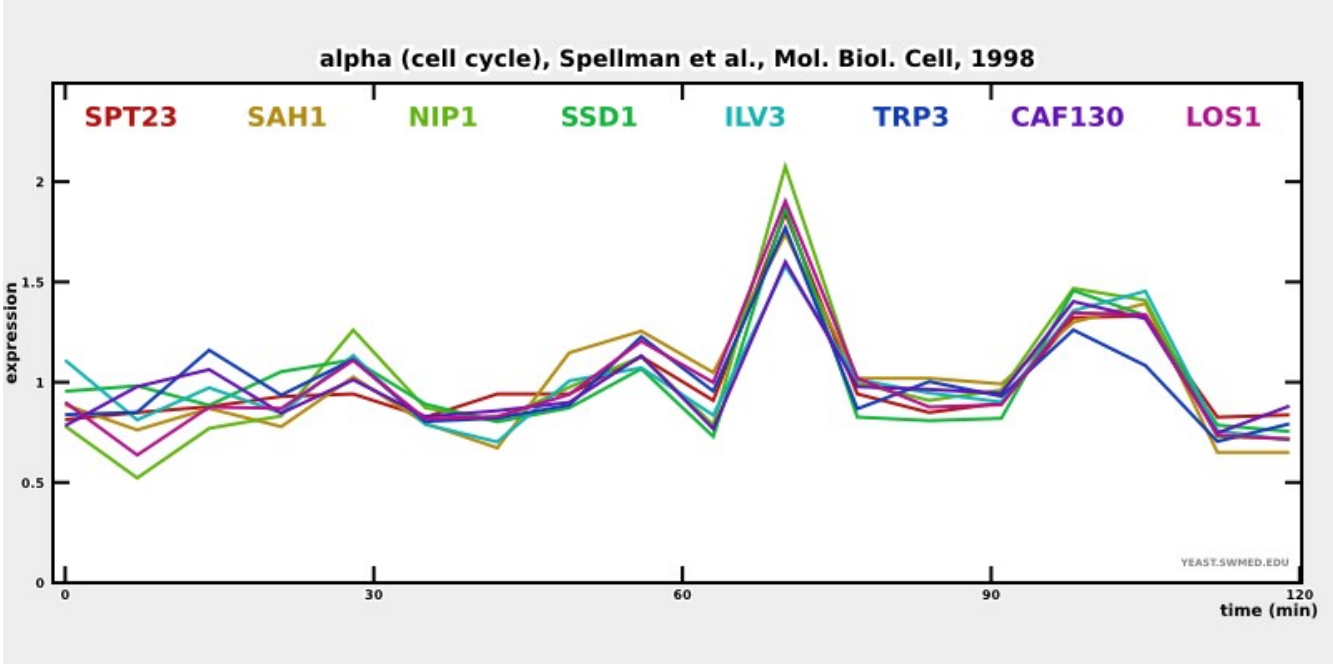


	FAS2	GRS1	RPL32	RPS8A	YBR075W	YHL035C	YKL014C
FAS2	1.00	0.72	0.55	0.48	0.74	0.43	0.82
GRS1	0.72	1.00	0.88	0.81	0.84	0.70	0.87
RPL32	0.55	0.88	1.00	0.75	0.63	0.87	0.82
RPS8A	0.48	0.81	0.75	1.00	0.69	0.49	0.78
YBR075W	0.74	0.84	0.63	0.69	1.00	0.46	0.71
YHL035C	0.43	0.70	0.87	0.49	0.46	1.00	0.66
YKL014C	0.82	0.87	0.82	0.78	0.71	0.66	1.00

17.- Cluster 2565 N

YKL020C YMR309C YKL205W YJR016C YKL211C YER043C YGR134W YDR293C
Cluster size: 8
Hits with Spellman *et al.*: 0
Hits with all studies: 1

PROPORTION: 1/8



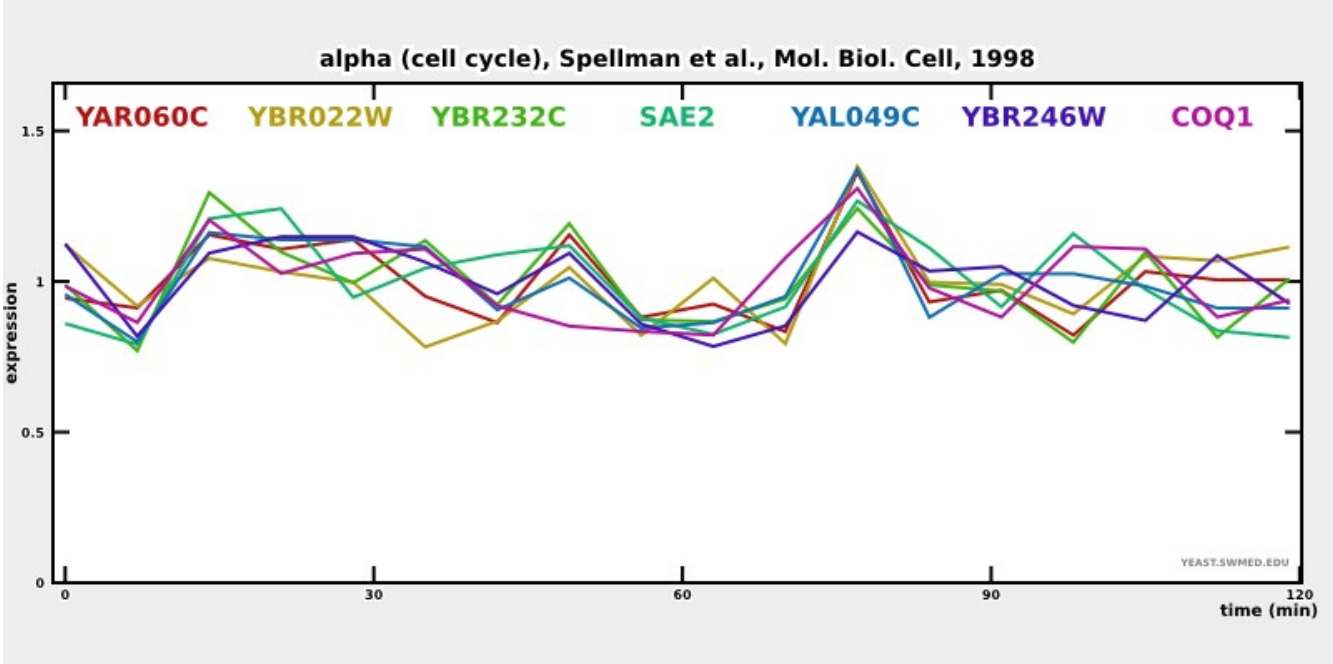
	CAF130	ILV3	LOS1	NIP1	SAH1	SPT23	SSD1	TRP3
CAF130	1.00	0.87	0.88	0.88	0.85	0.92	0.90	0.89
ILV3	0.87	1.00	0.91	0.89	0.91	0.85	0.88	0.81
LOS1	0.88	0.91	1.00	0.97	0.94	0.95	0.89	0.91
NIP1	0.88	0.89	0.97	1.00	0.90	0.94	0.90	0.87
SAH1	0.85	0.91	0.94	0.90	1.00	0.88	0.80	0.86
SPT23	0.92	0.85	0.95	0.94	0.88	1.00	0.93	0.88
SSD1	0.90	0.88	0.89	0.90	0.80	0.93	1.00	0.86
TRP3	0.89	0.81	0.91	0.87	0.86	0.88	0.86	1.00

18.- Cluster 863 N

YAL049C YAR060C YBR003W YGL175C YBR232C YBR022W YBR246W

Cluster size: 7
Hits with Spellman *et al.*: 0
Hits with all studies: 4

PROPORTION: 4/7



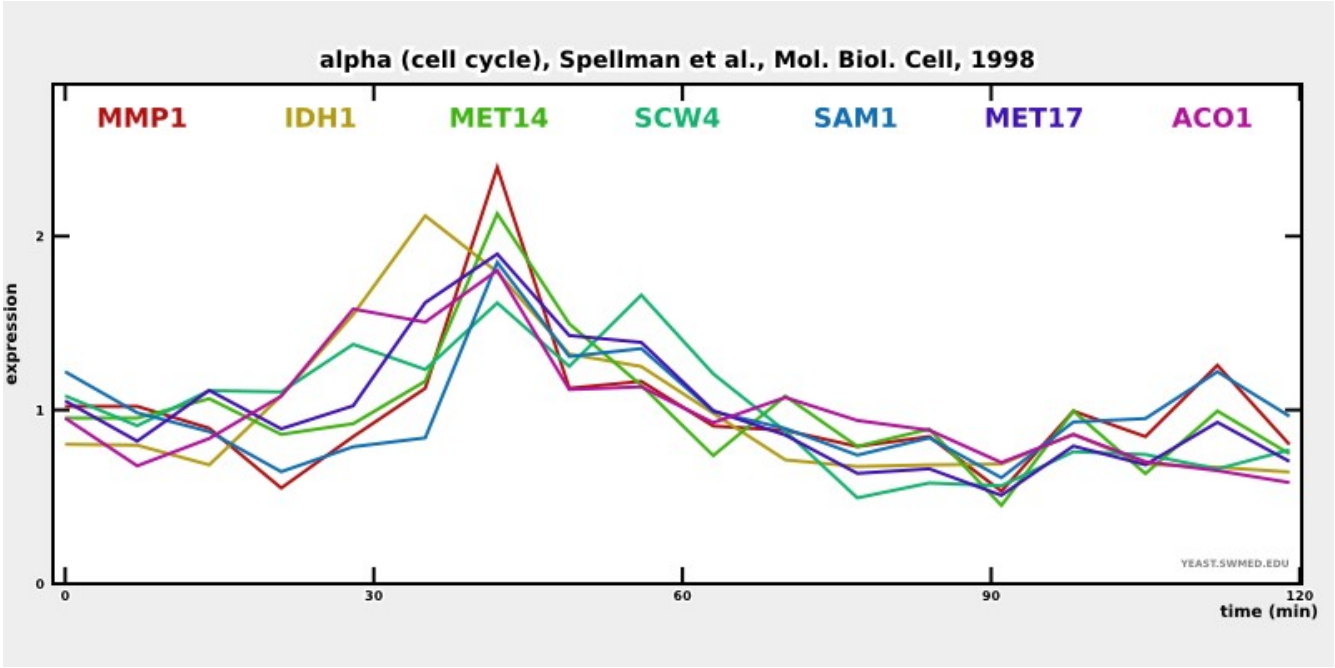
	COQ1	SAE2	YAL049C	YAR060C	YBR022W	YBR232C	YBR246W
COQ1	1.00	0.63	0.82	0.49	0.35	0.60	0.41
SAE2	0.63	1.00	0.71	0.49	0.25	0.64	0.53
YAL049C	0.82	0.71	1.00	0.78	0.50	0.74	0.71
YAR060C	0.49	0.49	0.78	1.00	0.79	0.76	0.67
YBR022W	0.35	0.25	0.50	0.79	1.00	0.48	0.48
YBR232C	0.60	0.64	0.74	0.76	0.48	1.00	0.59
YBR246W	0.41	0.53	0.71	0.67	0.48	0.59	1.00

19.- Cluster 1120 CC

YGR279C YLR303W YKL001C YLR304C YNL037C YLL061W YLR180W

Cluster size: 7
Hits with Spellman *et al.*: 6
Hits with all studies: 7

PROPORTION: 7/7



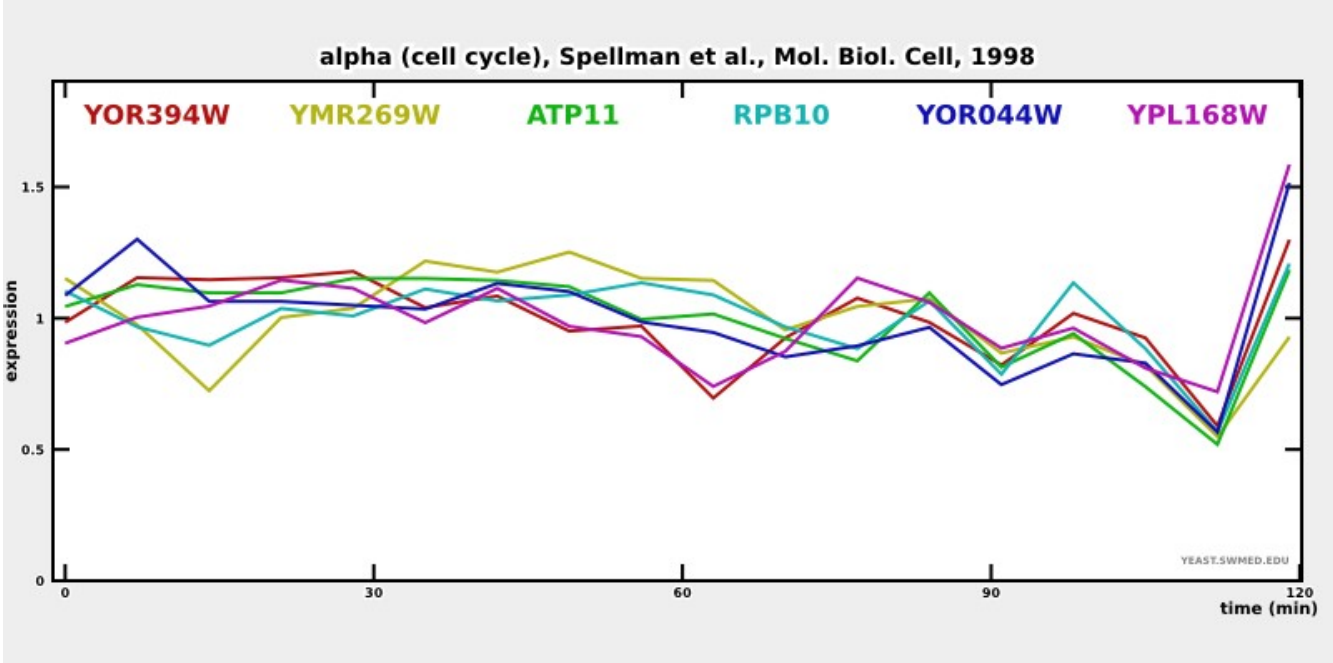
	ACO1	IDH1	MET14	MET17	MMP1	SAM1	SCW4
ACO1	1.00	0.89	0.70	0.78	0.58	0.38	0.74
IDH1	0.89	1.00	0.62	0.84	0.53	0.35	0.75
MET14	0.70	0.62	1.00	0.87	0.89	0.81	0.65
MET17	0.78	0.84	0.87	1.00	0.78	0.72	0.85
MMP1	0.58	0.53	0.89	0.78	1.00	0.90	0.54
SAM1	0.38	0.35	0.81	0.72	0.90	1.00	0.58
SCW4	0.74	0.75	0.65	0.85	0.54	0.58	1.00

20.- Cluster 1212 N

YPL168W YOR394W YOR044W YNL315C YMR269W YOR210W

Cluster size: 6
Hits with Spellman *et al.*: 0
Hits with all studies: 1

PROPORTION: 1/6



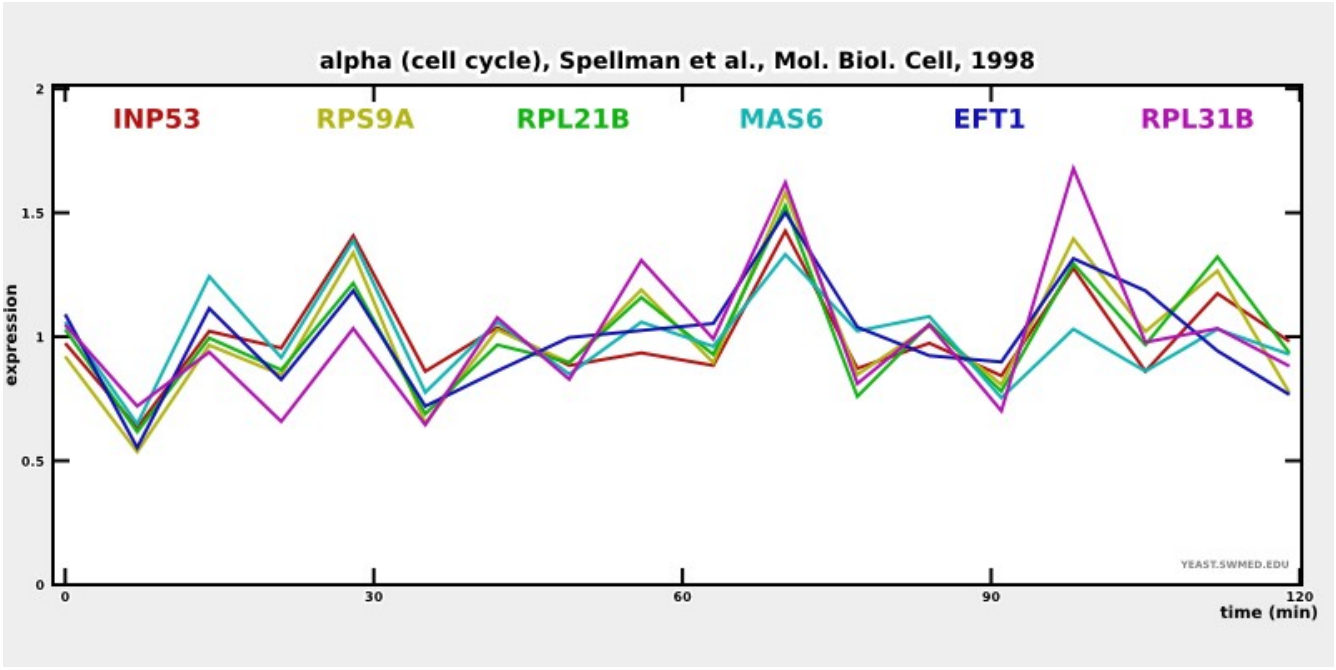
	ATP11	RPB10	YMR269W	YOR044W	YOR394W	YPL168W
ATP11	1.00	0.80	0.65	0.84	0.72	0.59
RPB10	0.80	1.00	0.76	0.69	0.53	0.47
YMR269W	0.65	0.76	1.00	0.42	0.24	0.15
YOR044W	0.84	0.69	0.42	1.00	0.79	0.76
YOR394W	0.72	0.53	0.24	0.79	1.00	0.85
YPL168W	0.59	0.47	0.15	0.76	0.85	1.00

21.- Cluster 1563 N

YLR406C YOR133W YPL081W YPL079W YNR017W YOR109W

Cluster size: 6
Hits with Spellman *et al.*: 0
Hits with all studies: 1

PROPORTION: 1/6



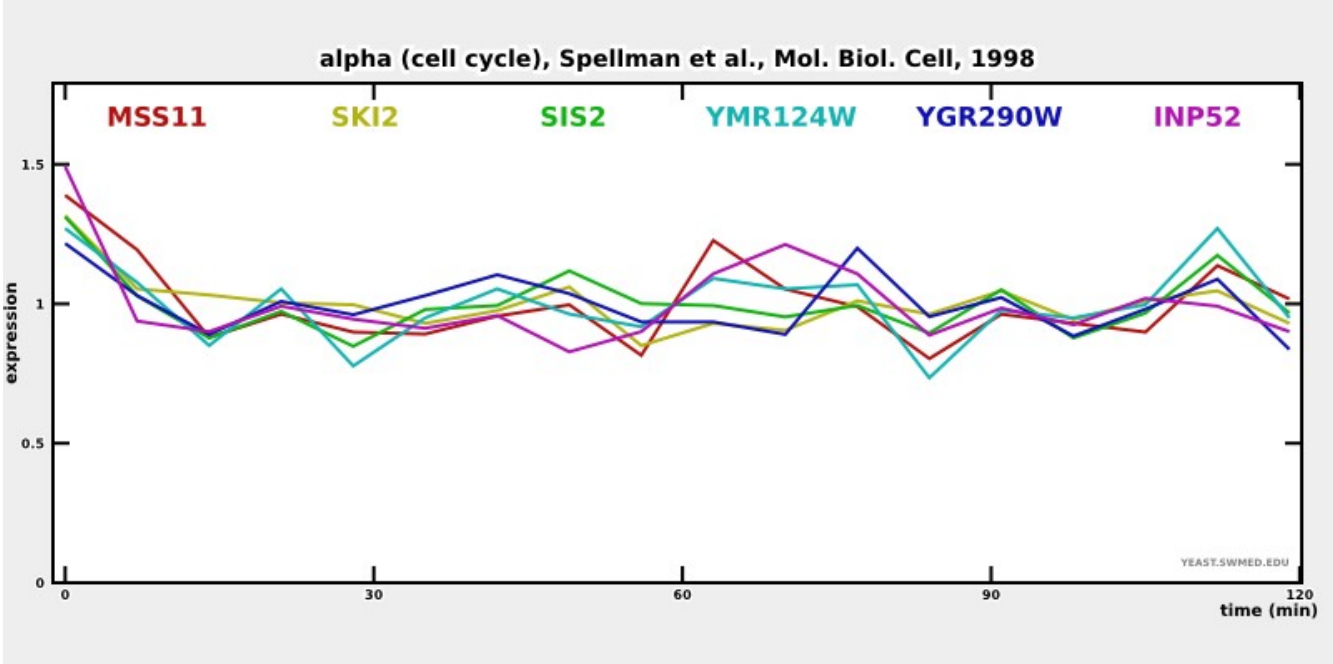
	EFT1	INP53	MAS6	RPL21B	RPL31B	RPS9A
EFT1	1.00	0.72	0.71	0.77	0.77	0.83
INP53	0.72	1.00	0.83	0.88	0.71	0.90
MAS6	0.71	0.83	1.00	0.76	0.60	0.78
RPL21B	0.77	0.88	0.76	1.00	0.86	0.96
RPL31B	0.77	0.71	0.60	0.86	1.00	0.86
RPS9A	0.83	0.90	0.78	0.96	0.86	1.00

22.- Cluster 810 N

YLR398C YMR164C YGR290W YKR072C YNL106C YMR124W

Cluster size: 6
Hits with Spellman *et al.*: 0
Hits with all studies: 2

PROPORTION: 2/6



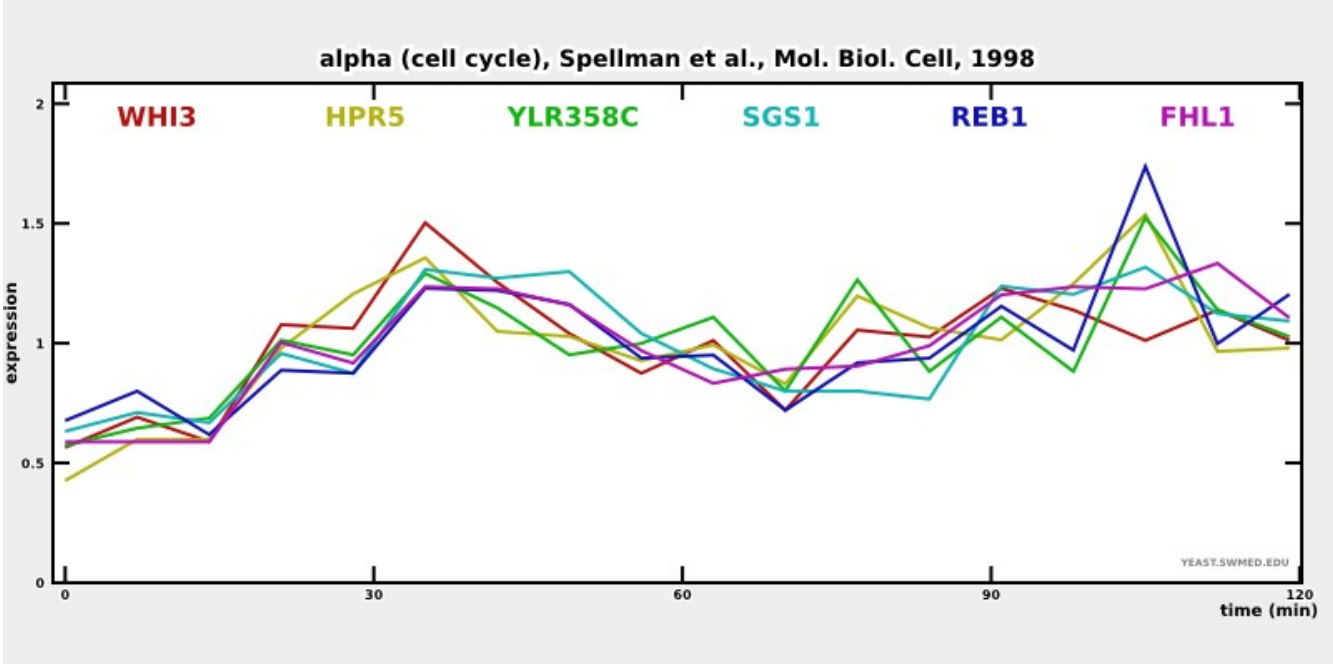
	INP52	MSS11	SIS2	SKI2	YGR290W	YMR124W
INP52	1.00	0.72	0.56	0.58	0.48	0.65
MSS11	0.72	1.00	0.72	0.63	0.43	0.81
SIS2	0.56	0.72	1.00	0.71	0.68	0.79
SKI2	0.58	0.63	0.71	1.00	0.67	0.49
YGR290W	0.48	0.43	0.68	0.67	1.00	0.59
YMR124W	0.65	0.81	0.79	0.49	0.59	1.00

23.- Cluster 1620 M

YJL092W YNL197C YBR049C YMR190C YPR104C YLR358C

Cluster size: 6
Hits with Spellman *et al.*: 2
Hits with all studies: 5

PROPORTION: 5/6



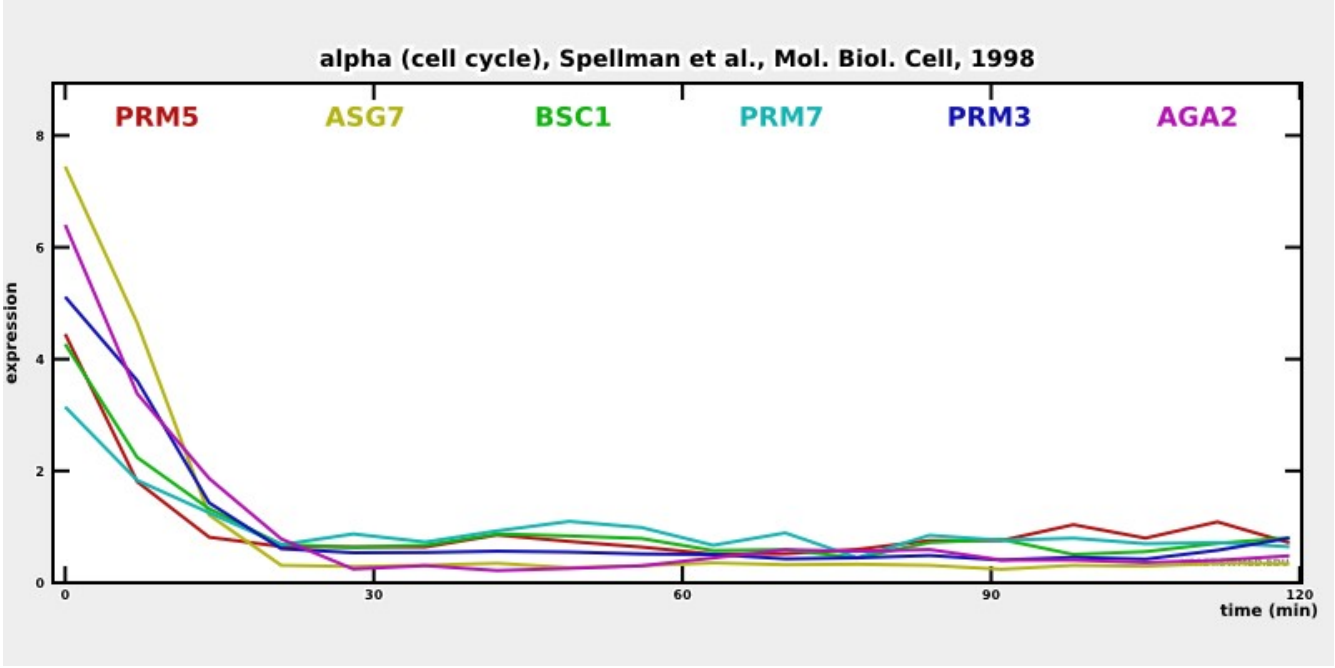
	FHL1	HPR5	REB1	SGS1	WHI3	YLR358C
FHL1	1.00	0.75	0.73	0.90	0.83	0.71
HPR5	0.75	1.00	0.76	0.69	0.77	0.85
REB1	0.73	0.76	1.00	0.83	0.61	0.82
SGS1	0.90	0.69	0.83	1.00	0.77	0.69
WHI3	0.83	0.77	0.61	0.77	1.00	0.73
YLR358C	0.71	0.85	0.82	0.69	0.73	1.00

24.- Cluster 1002 M

YGL032C YDL037C YIL117C YDL039C YJL170C YPL192C

Cluster size: 6
Hits with Spellman *et al.*: 4
Hits with all studies: 4

PROPORTION: 4/6



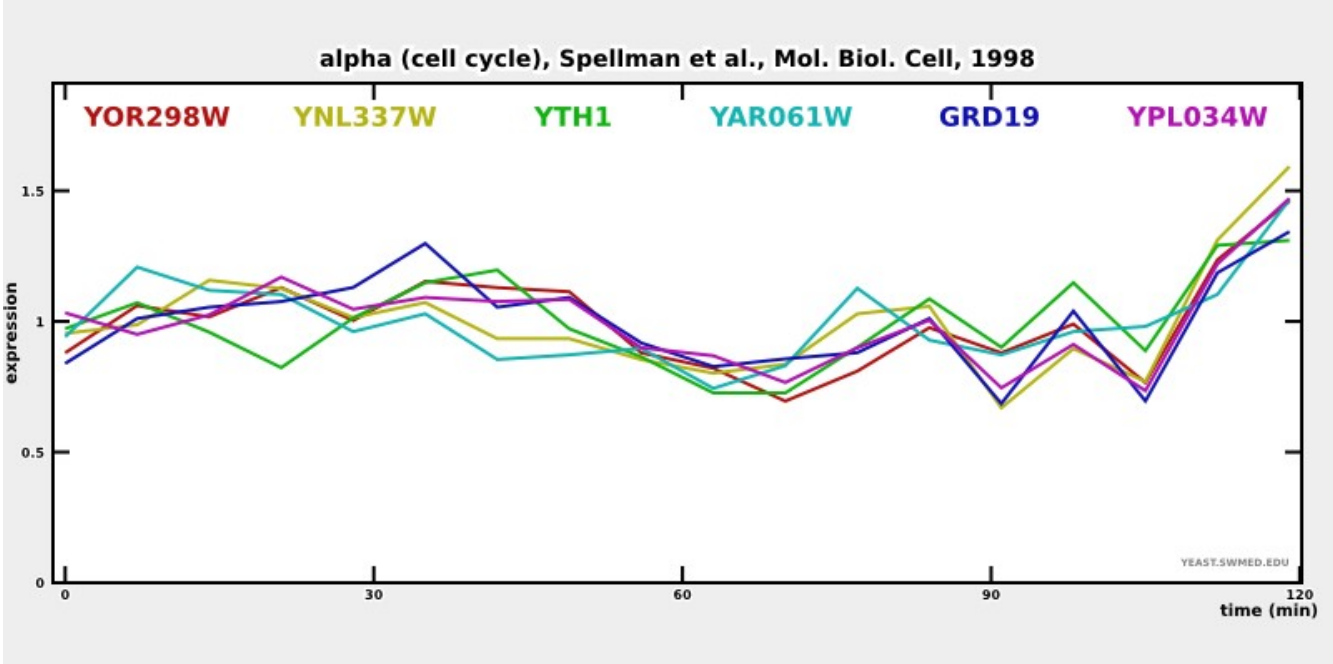
	AGA2	ASG7	BSC1	PRM3	PRM5	PRM7
AGA2	1.00	0.99	0.98	0.98	0.94	0.96
ASG7	0.99	1.00	0.98	1.00	0.95	0.96
BSC1	0.98	0.98	1.00	0.98	0.96	0.98
PRM3	0.98	1.00	0.98	1.00	0.92	0.95
PRM5	0.94	0.95	0.96	0.92	1.00	0.94
PRM7	0.96	0.96	0.98	0.95	0.94	1.00

25.- Cluster 2139 M

YNL337W YOR298W YPR107C YOR357C YAR061W YPL034W

Cluster size: 6
Hits with Spellman *et al.*: 2
Hits with all studies: 3

PROPORTION: 3/6



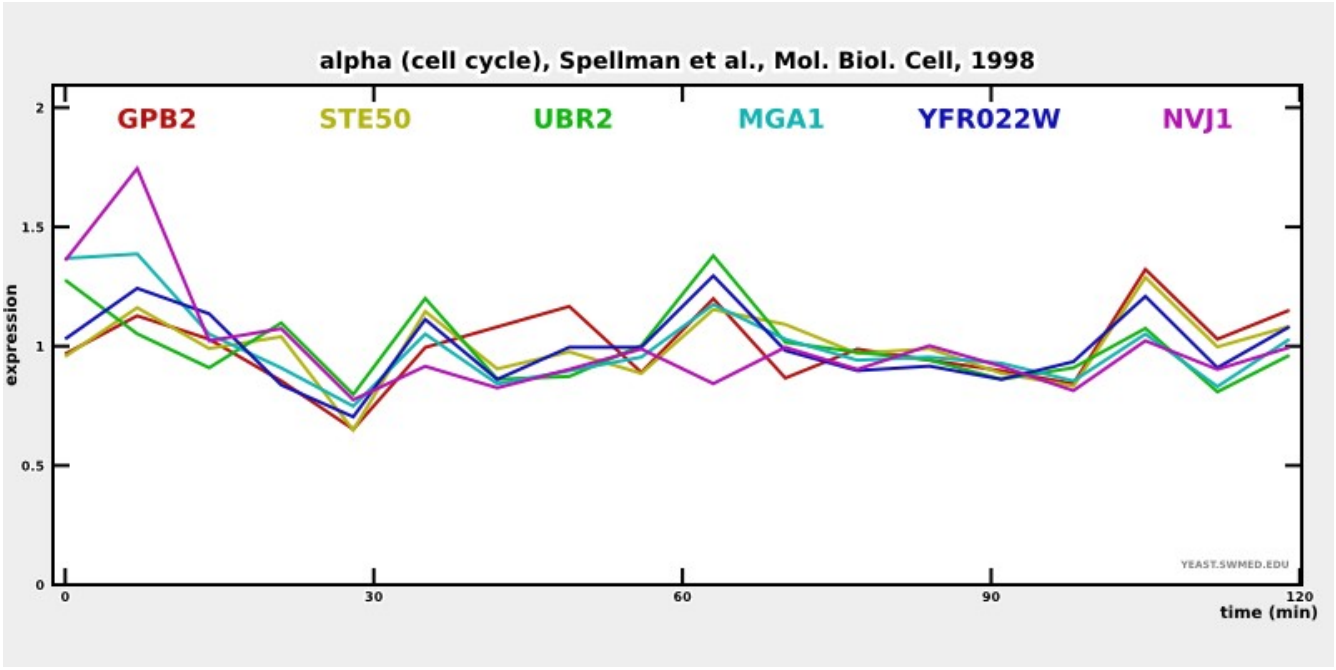
	GRD19	YAR061W	YNL337W	YOR298W	YPL034W	YTH1
GRD19	1.00	0.56	0.81	0.88	0.87	0.72
YAR061W	0.56	1.00	0.83	0.65	0.65	0.54
YNL337W	0.81	0.83	1.00	0.82	0.91	0.65
YOR298W	0.88	0.65	0.82	1.00	0.92	0.80
YPL034W	0.87	0.65	0.91	0.92	1.00	0.69
YTH1	0.72	0.54	0.65	0.80	0.69	1.00

26.- Cluster 2903 N

YFR022W YGR249W YHR195W YLR024C YCL032W YAL056W

Cluster size: 6
Hits with Spellman *et al.*: 0
Hits with all studies: 0

PROPORTION: 0/6



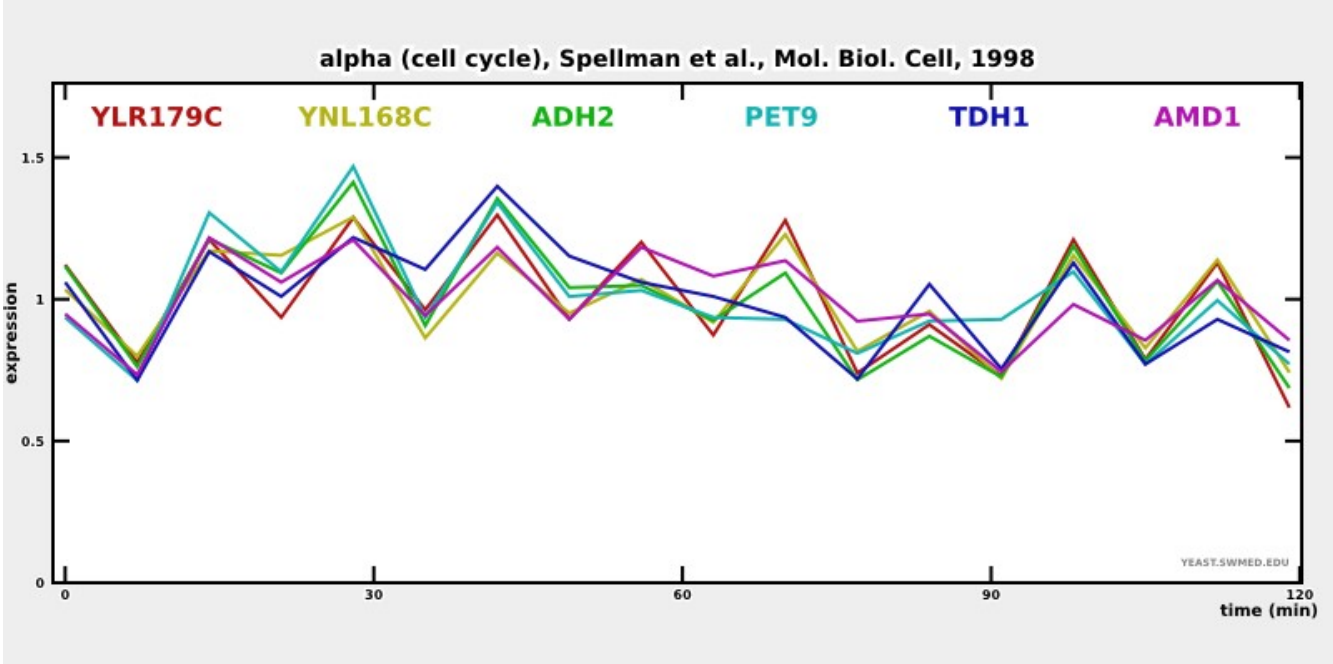
	GPB2	MGA1	NVJ1	STE50	UBR2	YFR022W
GPB2	1.00	0.42	0.22	0.76	0.31	0.75
MGA1	0.42	1.00	0.81	0.57	0.72	0.75
NVJ1	0.22	0.81	1.00	0.39	0.33	0.44
STE50	0.76	0.57	0.39	1.00	0.58	0.79
UBR2	0.31	0.72	0.33	0.58	1.00	0.64
YFR022W	0.75	0.75	0.44	0.79	0.64	1.00

27.- Cluster 3140 N

YBL030C YMR303C YLR179C YNL168C YJL052W YML035C

Cluster size: 6
Hits with Spellman *et al.* : 1
Hits with all studies: 1

PROPORTION: 1/6



	ADH2	AMD1	PET9	TDH1	YLR179C	YNL168C
ADH2	1.00	0.81	0.90	0.86	0.92	0.93
AMD1	0.81	1.00	0.78	0.72	0.82	0.86
PET9	0.90	0.78	1.00	0.83	0.76	0.79
TDH1	0.86	0.72	0.83	1.00	0.76	0.69
YLR179C	0.92	0.82	0.76	0.76	1.00	0.92
YNL168C	0.93	0.86	0.79	0.69	0.92	1.00

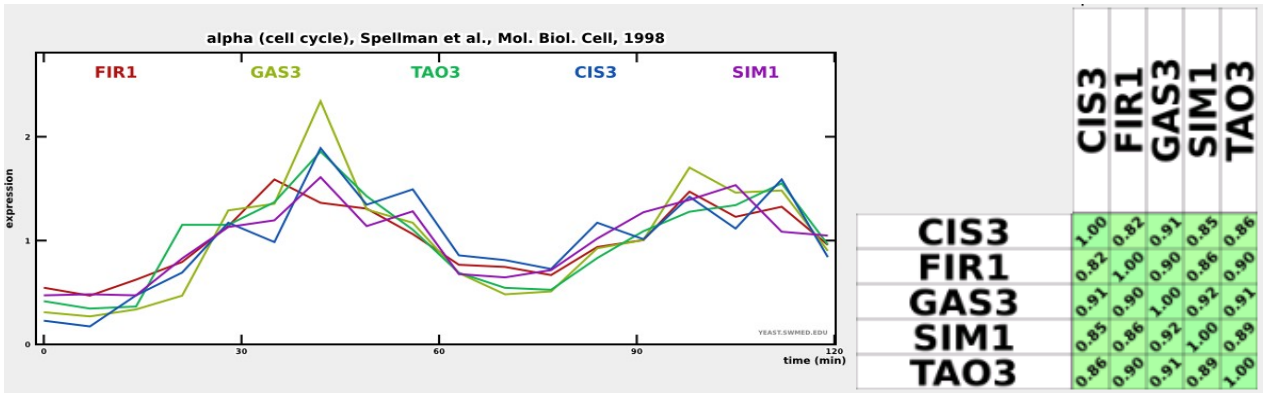
Clusters of size 5 and 4 which have hits with Spellman *et al.*:

28.- Cluster 477 CC

YJL158C YMR215W YIL129C YER032W YIL123W

Cluster size: 5
Hits with Spellman *et al.* : 5
Hits with all studies: 5

PROPORTION: 5/5



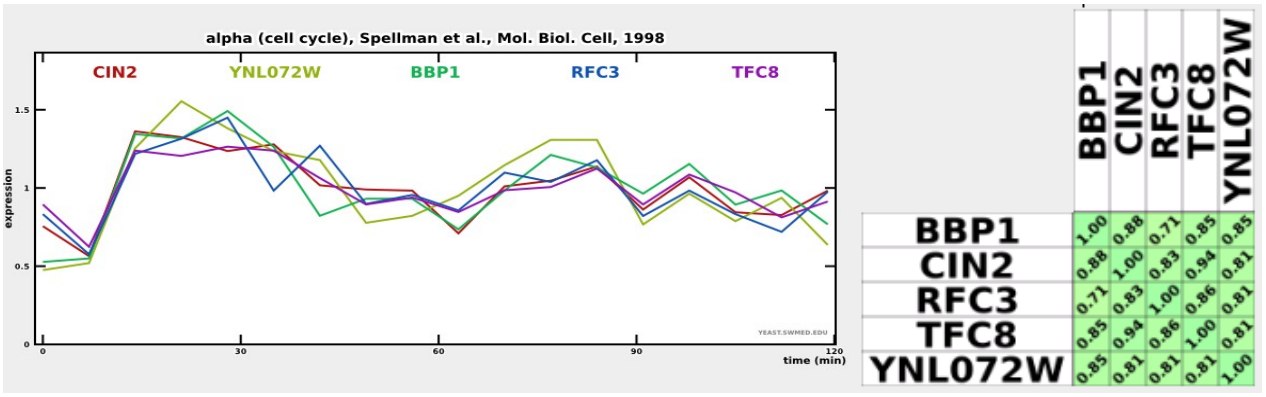
	CIS3	FIR1	GAS3	SIM1	TAO3
CIS3	1.00	0.82	0.91	0.85	0.86
FIR1	0.82	1.00	0.90	0.86	0.90
GAS3	0.91	0.90	1.00	0.92	0.91
SIM1	0.85	0.86	0.92	1.00	0.89
TAO3	0.86	0.90	0.91	0.89	1.00

30.- Cluster 1965 M

YPL007C YPL241C YPL255W YNL072W YNL290W

Cluster size: 5
Hits with Spellman *et al.* : 3
Hits with all studies: 5

PROPORTION: 5/5



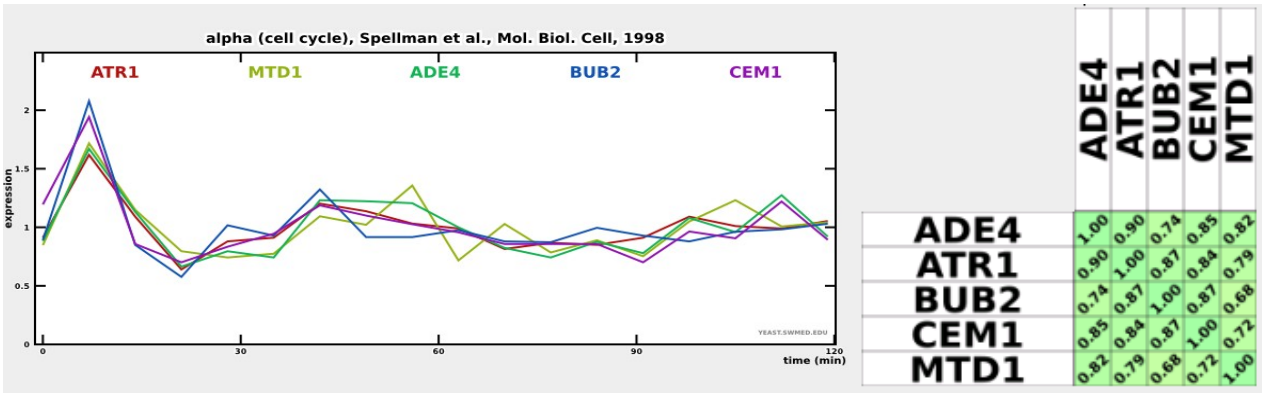
	BBP1	CIN2	RFC3	TFC8	YNL072W
BBP1	1.00	0.88	0.71	0.85	0.85
CIN2	0.88	1.00	0.83	0.94	0.81
RFC3	0.71	0.83	1.00	0.86	0.81
TFC8	0.85	0.94	0.86	1.00	0.81
YNL072W	0.85	0.81	0.81	0.81	1.00

34.- Cluster 2384 M

YML116W YMR055C YMR300C YKR080W YER061C

Cluster size: 5
Hits with Spellman *et al.* : 2
Hits with all studies: 2

PROPORTION: 2/5



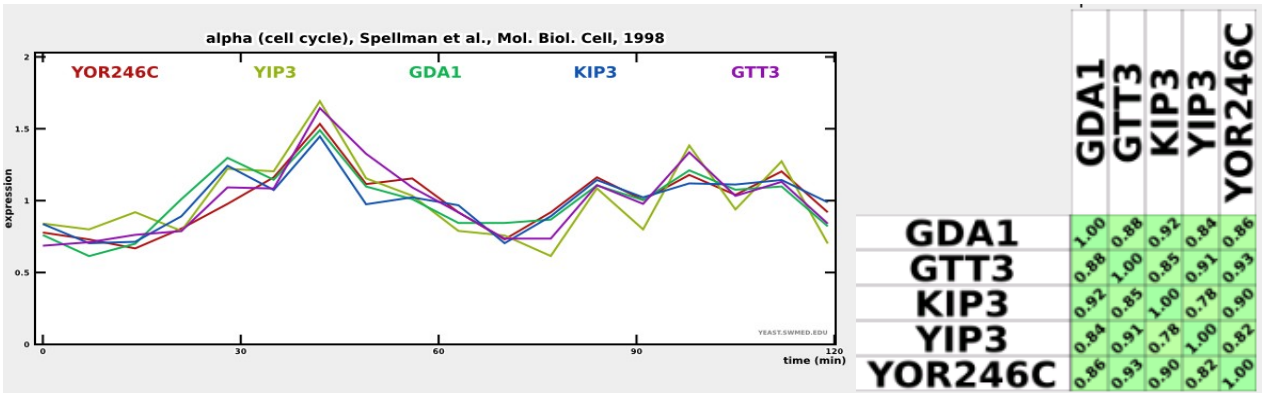
	ADE4	ATR1	BUB2	CEM1	MTD1
ADE4	1.00	0.90	0.74	0.85	0.82
ATR1	0.90	1.00	0.87	0.84	0.79
BUB2	0.74	0.87	1.00	0.87	0.68
CEM1	0.85	0.84	0.87	1.00	0.72
MTD1	0.82	0.79	0.68	0.72	1.00

35.- Cluster 2741 M

YEL042W YEL017W YOR246C YGL216W YNL044W

Cluster size: 5
Hits with Spellman *et al.* : 3
Hits with all studies: 4

PROPORTION: 4/5



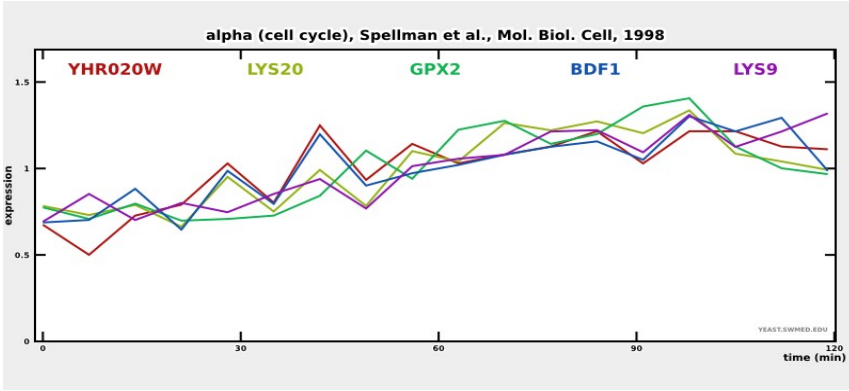
	GDA1	GTT3	KIP3	YIP3	YOR246C
GDA1	1.00	0.88	0.92	0.84	0.86
GTT3	0.88	1.00	0.85	0.91	0.93
KIP3	0.92	0.85	1.00	0.78	0.90
YIP3	0.84	0.91	0.78	1.00	0.82
YOR246C	0.86	0.93	0.90	0.82	1.00

36.- Cluster 1279 N

YNR050C YDL182W YBR244W YLR399C YHR020W

Cluster size: 5
Hits with Spellman *et al.* : 1
Hits with all studies: 1

PROPORTION: 1/5



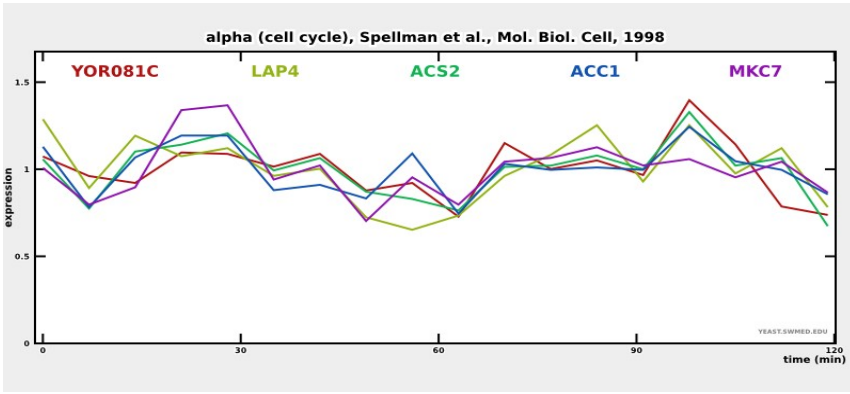
	BDF1	GPX2	LYS20	LYS9	YHR020W
BDF1	1.00	0.68	0.82	0.75	0.88
GPX2	0.68	1.00	0.84	0.71	0.62
LYS20	0.82	0.84	1.00	0.80	0.80
LYS9	0.75	0.71	0.80	1.00	0.72
YHR020W	0.88	0.62	0.80	0.72	1.00

37.- Cluster 1858 M

YNR016C YKL103C YDR144C YOR081C YLR153C

Cluster size: 5
Hits with Spellman *et al.* : 2
Hits with all studies: 4

PROPORTION: 4/5



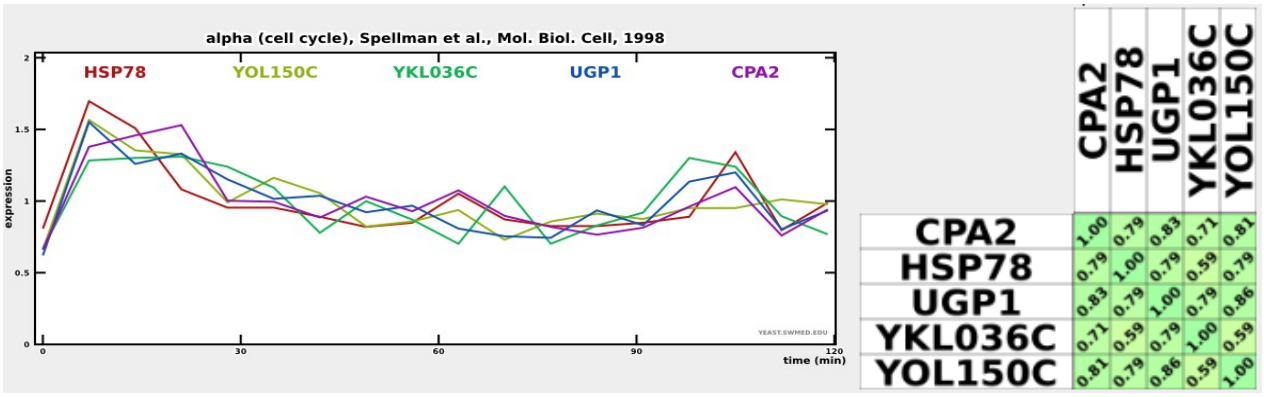
	ACC1	ACS2	LAP4	MKC7	YOR081C
ACC1	1.00	0.80	0.62	0.75	0.67
ACS2	0.80	1.00	0.82	0.70	0.76
LAP4	0.62	0.82	1.00	0.57	0.56
MKC7	0.75	0.70	0.57	1.00	0.50
YOR081C	0.67	0.76	0.56	0.50	1.00

39.- Cluster 924 M

YKL035W YOL150C YJR109C YDR258C YKL036C

Cluster size: 5
Hits with Spellman *et al.* : 2
Hits with all studies: 3

PROPORTION: 3/5



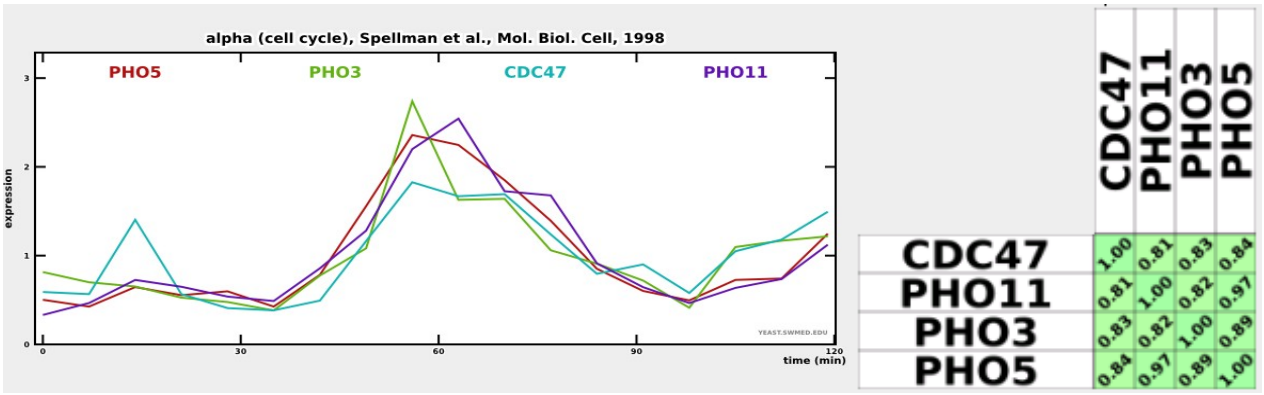
40.- Cluster 413 CC

YHR215W YBR093C YBR092C YBR202W YAR071W

Cluster size: 5
Hits with Spellman *et al.* : 5
Hits with all studies: 5

Gene not included in this analysis:
YHR215W: One of three repressible acid phosphatases, a glycoprotein that is transported to the cell surface by the secretory pathway; nearly identical to Pho11p; upregulated by phosphate starvation

PROPORTION: 5/5

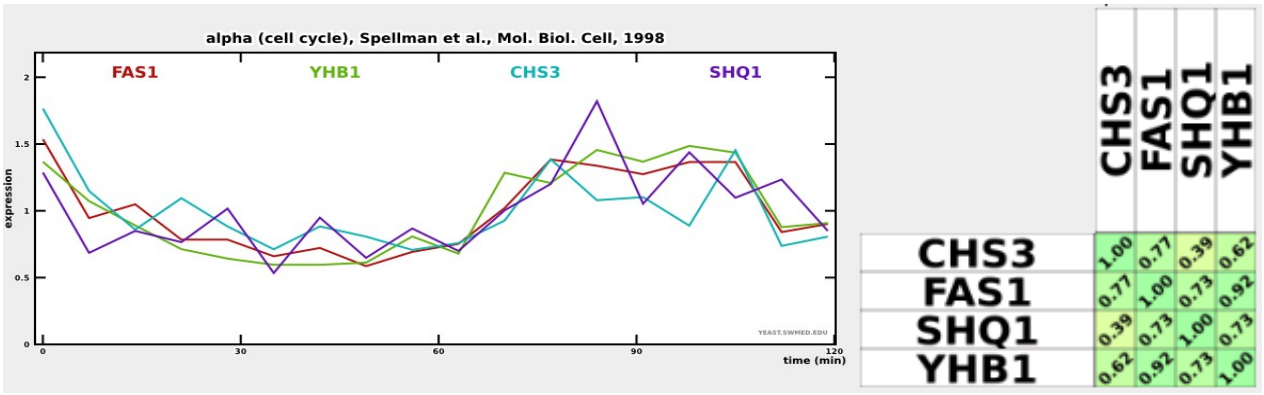


41.- Cluster 834 M

YGR234W YKL182W YIL104C YBR023C

Cluster size: 4
Hits with Spellman *et al.* : 3
Hits with all studies: 3

PROPORTION: 3/4



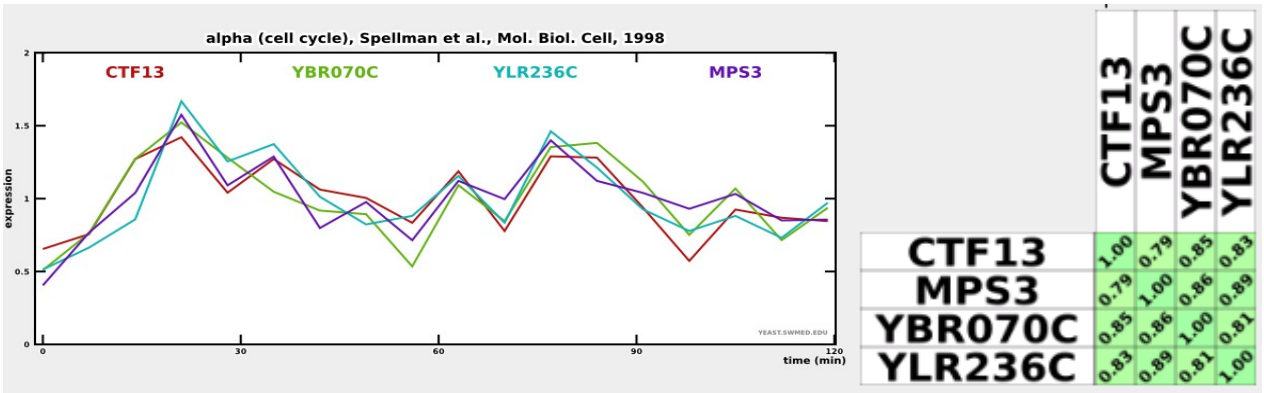
	CHS3	FAS1	SHQ1	YHB1
CHS3	1.00	0.77	0.39	0.62
FAS1	0.77	1.00	0.73	0.92
SHQ1	0.39	0.73	1.00	0.73
YHB1	0.62	0.92	0.73	1.00

42.- Cluster 778 M

YJL019W YBR070C YLR236C YMR094W

Cluster size: 4
Hits with Spellman *et al.* : 3
Hits with all studies: 4

PROPORTION: 4/4



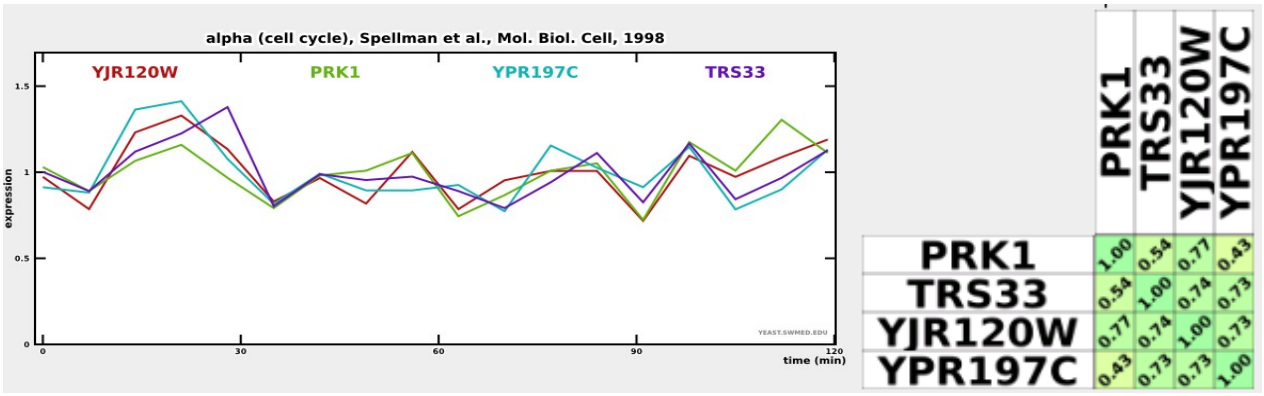
	CTF13	MPS3	YBR070C	YLR236C
CTF13	1.00	0.79	0.85	0.83
MPS3	0.79	1.00	0.86	0.89
YBR070C	0.85	0.86	1.00	0.81
YLR236C	0.83	0.89	0.81	1.00

45.- Cluster 190 N

YOR115C YIL095W YJR120W YPR197C

Cluster size: 4
Hits with Spellman *et al.* : 1
Hits with all studies: 1

PROPORTION: 1/4

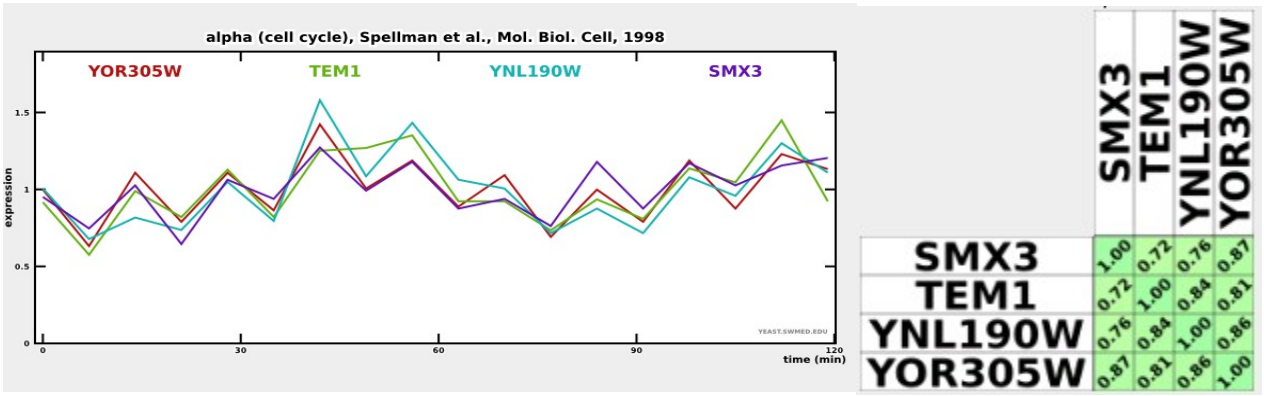


49.- Cluster 537 N

YML064C YNL190W YOR305W YPR182W

Cluster size: 4
Hits with Spellman *et al.* : 1
Hits with all studies: 1

PROPORTION: 1/4

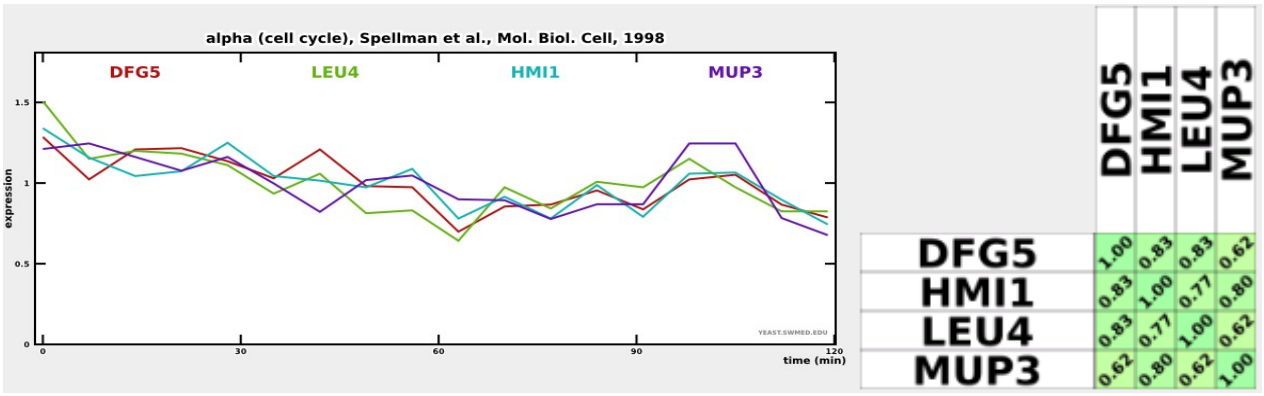


50.- Cluster 1030 N

YMR238W YNL104C YOL095C YHL036W

Cluster size: 4
Hits with Spellman *et al.* : 1
Hits with all studies: 1

PROPORTION: 1/4



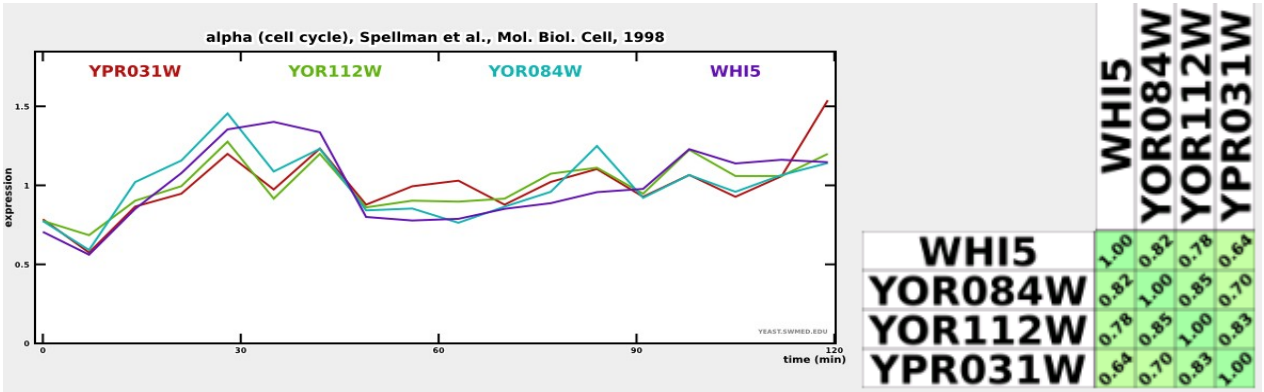
	DFG5	HMI1	LEU4	MUP3
DFG5	1.00	0.83	0.83	0.62
HMI1	0.83	1.00	0.77	0.80
LEU4	0.83	0.77	1.00	0.62
MUP3	0.62	0.80	0.62	1.00

53.- Cluster 1590 M

YOR084W YOR083W YOR112W YPR031W

Cluster size: 4
Hits with Spellman *et al.* : 2
Hits with all studies: 3

PROPORTION: 3/4



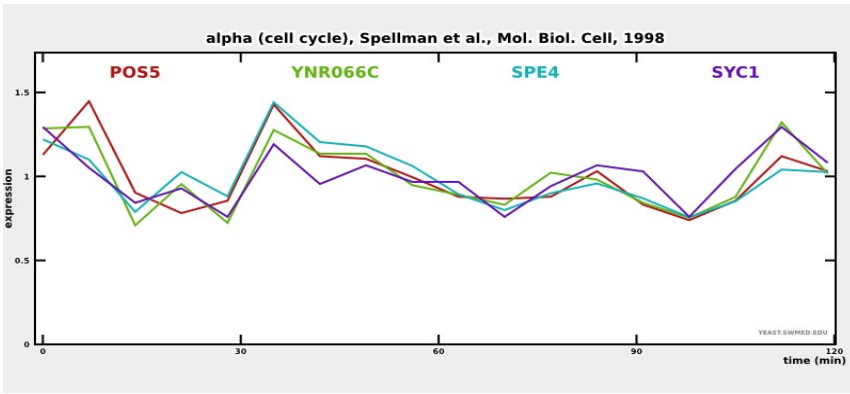
	WHI5	YOR084W	YOR112W	YPR031W
WHI5	1.00	0.82	0.78	0.64
YOR084W	0.82	1.00	0.85	0.70
YOR112W	0.78	0.85	1.00	0.83
YPR031W	0.64	0.70	0.83	1.00

55.- Cluster 1878 N

YNR066C YPL188W YOR179C YLR146C

Cluster size: 4
Hits with Spellman *et al.* : 1
Hits with all studies: 2

PROPORTION: 2/4



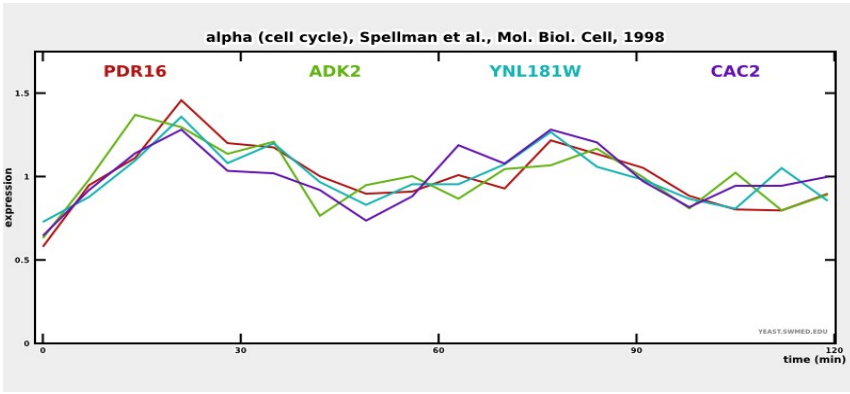
	POS5	SPE4	SYC1	YNR066C
POS5	1.00	0.81	0.63	0.84
SPE4	0.81	1.00	0.67	0.83
SYC1	0.63	0.67	1.00	0.83
YNR066C	0.84	0.83	0.83	1.00

57.- Cluster 2004 CC

YNL231C YML102W YER170W YNL181W

Cluster size: 4
Hits with Spellman *et al.* : 4
Hits with all studies: 4

PROPORTION: 4/4



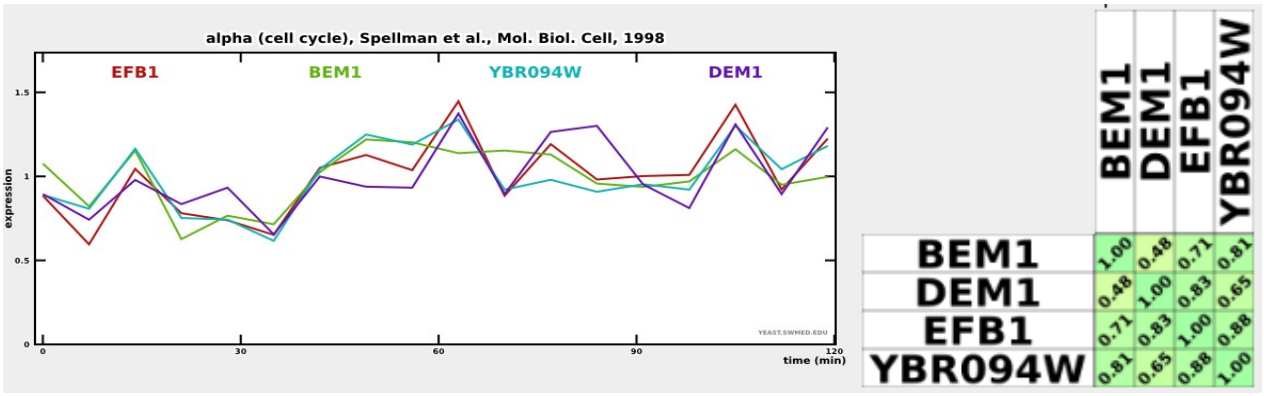
	ADK2	CAC2	PDR16	YNL181W
ADK2	1.00	0.67	0.78	0.70
CAC2	0.67	1.00	0.80	0.80
PDR16	0.78	0.80	1.00	0.87
YNL181W	0.70	0.80	0.87	1.00

61.- Cluster 1268 M

YBR200W YBR094W YAL003W YBR163W

Cluster size: 4
Hits with Spellman *et al.* : 2
Hits with all studies: 2

PROPORTION: 2/4

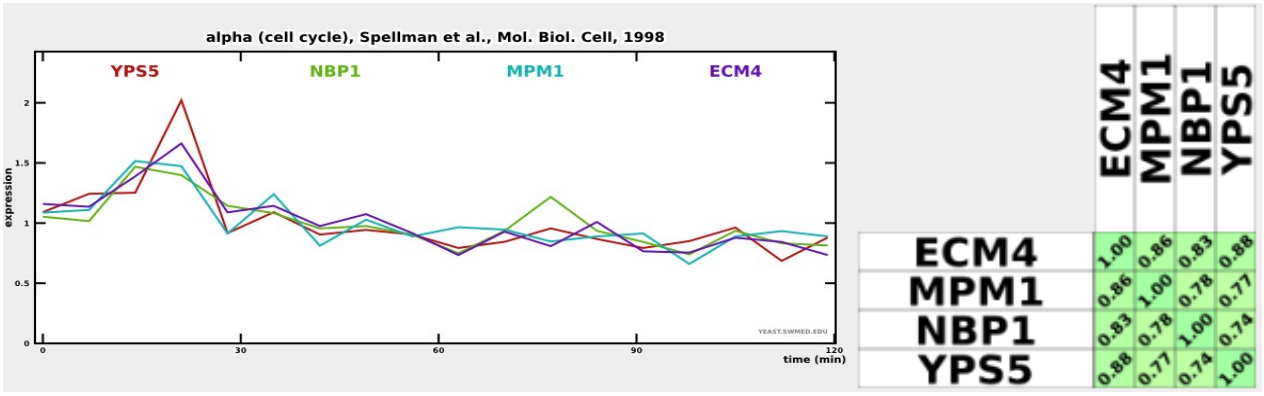


62.- Cluster 1185 N

YLR457C YGL259W YKR076W YJL066C

Cluster size: 4
Hits with Spellman *et al.* : 1
Hits with all studies: 2

PROPORTION: 2/4

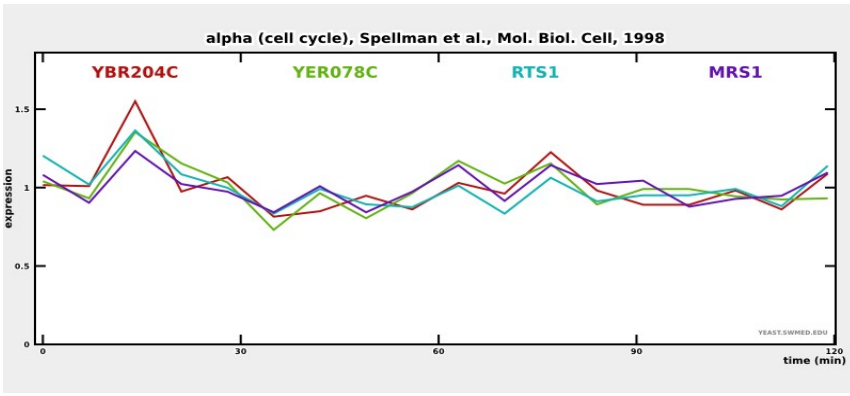


70.- Cluster 2486 N

YBR204C YIR021W YER078C YOR014W

Cluster size: 4
Hits with Spellman *et al.* : 1
Hits with all studies: 1

PROPORTION: 1/4



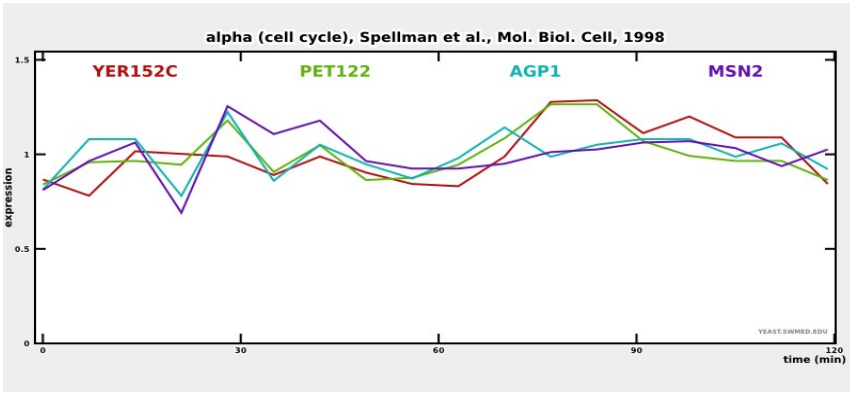
	MRS1	RTS1	YBR204C	YER078C
MRS1	1.00	0.77	0.73	0.80
RTS1	0.77	1.00	0.81	0.72
YBR204C	0.73	0.81	1.00	0.75
YER078C	0.80	0.72	0.75	1.00

71.- Cluster 2170 M

YER152C YER153C YCL025C YMR037C

Cluster size: 4
Hits with Spellman *et al.* : 2
Hits with all studies: 3

PROPORTION: 3/4



	AGP1	MSN2	PET122	YER152C
AGP1	1.00	0.66	0.59	0.33
MSN2	0.66	1.00	0.42	0.21
PET122	0.59	0.42	1.00	0.74
YER152C	0.33	0.21	0.74	1.00

II.- Genes with an oscillatory behaviour identified by Ahnert *et al*.

The following list of oscillatory genes was kindly provided by Ph. D. Ahnert (See Reference 31).

YBR016W	YPL145C	YPL211W	YCR043C	YGL085W	YER115C	YDR071C	YHL038C
YBR154C	YPL220W	YPL251W	YER075C	YGL129C	YGL058W	YEL009C	YHR132C
YBR196C	YPR198W	YPR028W	YGL137W	YML114C	YGR148C	YIL094C	YHR133C
YDL182W	YDL042C	YAL043C	YBL052C	YGL091C	YHR101C	YJL188C	YHR181W
YDL240W	YGL023C	YOL005C	YJL113W	YIL102C	YHR163W	YBL007C	YIL011W
YDR076W	YGL025C	YOL064C	YLR195C	YIL114C	YJL081C	YGL122C	YIL014W
YDR226W	YGR149W	YPR158W	YLR199C	YKL019W	YKL192C	YJR013W	YIL083C
YDR264C	YGR179C	YBR173C	YLR295C	YIL007C	YMR083W	YJR159W	YIL085C
YER099C	YJL197W	YER133W	YDR005C	YPR060C	YOR021C	YOR293W	YIL153W
YER131W	YMR061W	YFR055W	YDR120C	YAR061W	YOR063W	YOR328W	YJL013C
YGL173C	YPR056W	YGL001C	YDR308C	YBL008W	YOR209C	YJL096W	YJL032W
YGR094W	YBR289W	YNL135C	YGR001C	YBL010C	YOR251C	YJL202C	YJL055W
YHR054C	YCRX03C	YNL178W	YGR005C	YBR081C	YOR253W	YNL301C	YJL151C
YIL008W	YDL176W	YOL026C	YIL134W	YDR273W	YOR277C	YPL007C	YJL171C
YIR026C	YDL237W	YPL234C	YOL106W	YLL003W	YPL169C	YBL006C	YJR005W
YKL159C	YER068W	YBR039W	YGL218W	YLR001C	YPL225W	YBR063C	YJR068W
YLR060W	YER143W	YBR145W	YLR444C	YLR077W	YPR062W	YDR265W	YJR099W
YLR130C	YER148W	YGL249W	YNL124W	YLR191W	YPR082C	YGL113W	YJR101W
YLR427W	YFR038W	YLR245C	YDL191W	YPL115C	YPR086W	YAR002W	YJR115W
YML018C	YGL014W	YMR122C	YDR048C	YPR026W	YPR100W	YBL060W	YJR117W
YMR027W	YGL027C	YOL165C	YDR510W	YKR103W	YPR132W	YBR205W	YJR139C
YMR229C	YGL198W	YNL143C	YKL073W	YLL059C	YCR103C	YDL024C	YJR145C
YMR243C	YGR244C	YBR078W	YGR173W	YOR261C	YDL192W	YDL047W	YKL004W
YNL113W	YGR295C	YPR113W	YHL041W	YBL036C	YDR161W	YDR037W	YKL175W
YNL322C	YHL045W	YKR030W	YMR294W	YBR146W	YDR441C	YDR063W	YKL193C
YPL039W	YHR203C	YKL125W	YER117W	YDR177W	YGL220W	YDR150W	YKL199C
YDR465C	YIL104C	YKL156W	YIL017W	YOL075C	YIL137C	YDR227W	YKL219W
YER012W	YIL175W	YLR206W	YJR102C	YOR367W	YJL150W	YDR324C	YKR014C
YGL165C	YIL176C	YPR009W	YER179W	YDL172C	YLR421C	YER109C	YKR048C
YGR267C	YJR146W	YIL154C	YGL056C	YGL130W	YMR107W	YFR008W	YKR070W
YIL062C	YKL081W	YPL130W	YDL202W	YGL228W	YNR061C	YGR069W	YKR088C
YLR283W	YLR011W	YDR230W	YDR337W	YJR018W	YOL143C	YGR116W	YLL038C
YML040W	YLR037C	YPL224C	YHR195W	YJR024C	YOR355W	YJL046W	YLL054C
YMR034C	YLR357W	YDR417C	YPR168W	YJR128W	YPL013C	YJR047C	YLR059C
YOL121C	YNL185C	YLR367W	YHR127W	YKL064W	YDL045C	YKL024C	YLR087C
YOR193W	YNL268W	YBL106C	YIL127C	YPR118W	YDL086W	YKR043C	YLR089C
YPL037C	YOR271C	YMR266W	YJR020W	YDR323C	YDR231C	YLL040C	YLR179C
YER008C	YPL136W	YMR267W	YMR202W	YDR478W	YFL029C	YLR110C	YLR181C
YER176W	YPL187W	YNL130C	YNL038W	YGL026C	YGL176C	YLR309C	YLR229C
YFL033C	YPR144C	YOL066C	YOL040C	YNL338W	YGR091W	YLR461W	YLR277C
YGL079W	YCL049C	YOR302W	YOR167C	YAL015C	YGR139W	YMR109W	YLR305C
YGL221C	YCLX03C	YPR006C	YOR169C	YBR085W	YGR251W	YNL033W	YLR355C
YGR122W	YDL002C	YGL108C	YPR074C	YBR095C	YGR264C	YOL004W	YLR378C
YGR150C	YDL015C	YGL167C	YDL236W	YCR084C	YHL007C	YOR218C	YLR447C
YGR172C	YDL141W	YJL077C	YDR339C	YDL046W	YHR092C	YPL080C	YML028W
YGR266W	YDR474C	YML004C	YGR195W	YDL184C	YHR162W	YPR181C	YML035C
YHL046C	YER161C	YDL232W	YHR013C	YDL206W	YJR148W	YAL039C	YML092C
YHL047C	YFR051C	YDR151C	YJL215C	YDR118W	YKL224C	YAR009C	YML111W
YIL030C	YGL018C	YGL148W	YJR049C	YDR174W	YKL225W	YBL030C	YML115C
YJL008C	YGL104C	YJR142W	YLR369W	YDR486C	YLR136C	YBL056W	YMR064W

YJL010C	YGL110C	YNL017C	YOL029C	YGR178C	YLR351C	YBL102W	YMR091C
YJL042W	YGL152C	YCR107W	YBR032W	YGR191W	YML030W	YBR109C	YMR113W
YJL053W	YGR081C	YDL210W	YBR117C	YGR228W	YML123C	YBR111C	YMR134W
YKR086W	YGR255C	YKL105C	YDR403W	YHR015W	YMR010W	YBR131W	YMR146C
YNL148C	YHR004C	YKR094C	YDR446W	YJL025W	YNL052W	YBR159W	YMR172W
YAL045C	YIL133C	YLR155C	YFL001W	YJL219W	YNL259C	YBR181C	YMR220W
YBR082C	YJL072C	YLR393W	YGL064C	YJL221C	YNR015W	YBR285W	YMR241W
YDL005C	YKR097W	YMR170C	YGL143C	YKL001C	YOL041C	YCL039W	YMR297W
YDL208W	YLL013C	YNL020C	YGR141W	YKR066C	YOL063C	YCL041C	YMR303C
YDL212W	YLL047W	YBR113W	YGR194C	YLL036C	YOL129W	YCL046W	YMR319C
YDL241W	YLR113W	YBR197C	YHR150W	YLR203C	YOL133W	YCL048W	YMR321C
YDR059C	YLR116W	YCL006C	YIL032C	YLR375W	YOR011W	YCR004C	YMR323W
YDR276C	YLR456W	YDR067C	YLL057C	YLR448W	YOR184W	YCR033W	YNL048W
YDR280W	YNL074C	YER048C	YLR061W	YMR005W	YOR256C	YDL067C	YNL067W
YDR419W	YNL225C	YER086W	YML083C	YMR056C	YOR283W	YDR041W	YNL070W
YER032W	YOR022C	YHL017W	YBR206W	YNL042W	YOR304W	YDR105C	YNL071W
YER083C	YOR085W	YHR021C	YCR106W	YNL100W	YOR347C	YDR107C	YNL096C
YFL003C	YOR163W	YHR043C	YDR316W	YNL307C	YPL067C	YDR155C	YNL120C
YGL103W	YPL001W	YHR111W	YLL033W	YAR035W	YPL078C	YDR176W	YNL122C
YGL145W	YPL059W	YIL128W	YLR374C	YCLX06C	YPL082C	YDR209C	YNL168C
YGL197W	YPL199C	YKR038C	YOL052C	YCR002C	YPL107W	YDR233C	YNR017W
YGR121C	YAL042W	YKR085C	YPR128C	YCR105W	YPL125W	YDR321W	YNR021W
YHR063C	YBR172C	YML082W	YBR076W	YCRX01W	YPL147W	YDR397C	YNR041C
YIL035C	YDL121C	YNL162W	YBR133C	YDL235C	YPR012W	YDR471W	YOL139C
YIL064W	YDL137W	YBR033W	YBR193C	YDR152W	YPR029C	YDR487C	YOL161C
YIL157C	YDR081C	YCL032W	YCRX09C	YDR375C	YPR054W	YDR526C	YOR067C
YJL004C	YDR399W	YCL034W	YDR277C	YDR447C	YPR129W	YER132C	YOR108W
YJL206C	YDR463W	YDL008W	YDR394W	YDR542W	YPR182W	YER134C	YOR109W
YJL223C	YDR493W	YDR412W	YDR443C	YGL142C	YAL066W	YER159C	YOR157C
YJR001W	YEL063C	YBL070C	YGL088W	YGL214W	YBR274W	YER177W	YOR239W
YJR017C	YEL067C	YOR243C	YKL056C	YGR102C	YCL020W	YER178W	YOR254C
YKL002W	YGR200C	YPL042C	YLL056C	YGR117C	YDR119W	YGL077C	YOR260W
YKL097W-A	YGR214W	YDR153C	YLR285W	YGR167W	YDR200C	YGL132W	YOR305W
YKL129C	YHL015W	YBR013C	YLR391W	YHR207C	YDR418W	YGL147C	YOR323C
YKL152C	YHL037C	YDR476C	YOR276W	YIL084C	YFL010C	YGL189C	YOR331C
YKL173W	YHR065C	YGL095C	YBR019C	YJR003C	YGR284C	YGL191W	YPL028W
YKR065C	YHR089C	YLR055C	YBR137W	YJR141W	YHL004W	YGL199C	YPL079W
YLL006W	YHR183W	YLR071C	YDR021W	YKL025C	YHR098C	YGL215W	YPL081W
YLR039C	YIL087C	YMR306W	YJR107W	YKL158W	YIL121W	YGL219C	YPL129W
YLR043C	YIL088C	YOR123C	YKR104W	YLR160C	YJL014W	YGL261C	YPL131W
YLR062C	YJL179W	YBL066C	YML041C	YLR426W	YKL018W	YGR067C	YPL178W
YML089C	YJR070C	YBR216C	YPR102C	YML054C	YKL195W	YGR098C	YPL218W
YMR054W	YJR123W	YBR280C	YBR140C	YMR088C	YLR241W	YGR118W	YPL246C
YMR272C	YKL058W	YDL125C	YMR152W	YMR135C	YLR243W	YGR123C	YPL271W
YMR276W	YKL080W	YDR027C	YAR043C	YNL043C	YML031W	YGR124W	YPL273W
YMR298W	YKR092C	YDR315C	YBL005W-B	YNL103W	YMR176W	YGR128C	YPR016C
YNL001W	YLL004W	YDR354W	YFL050C	YNL171C	YNL236W	YGR192C	YPR033C
YNL069C	YLR065C	YDR499W	YGL159W	YPL213W	YOR009W	YGR216C	YPR036W
YNL170W	YLR083C	YER146W	YHR049W	YPL240C	YPR050C	YGR246C	YPR044C
YNL190W	YLR109W	YER162C	YJR090C	YPR087W	YAL044C	YGR270W	YPR114W
YNL336W	YLR441C	YIL022W	YPR189W	YPR109W	YAR060C	YGR285C	YPR127W
YNR043W	YNL002C	YDR356W	YDL085W	YBR011C	YJR150C	YGR292W	YPR150W
YOR020C	YNL114C	YGL081W	YDR171W	YDL213C	YKL114C	YGR294W	YPR154W
YOR159C	YPL105C	YHR199C	YFR029W	YDR301W	YOL055C	YHL009C	YPR176C

Number of oscillating genes inside the first 27 clusters:

Oscillating Genes

Cluster Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Cluster Size	37	27	61	39	17	13	11	11	10	10	10	9	9	9	8	8	8	7	7	6	6	6	6	6	6	6	6
Number of Oscillating genes	0	0	0	1	2	0	1	1	0	1	2	1	0	1	0	0	0	1	1	0	4	0	0	0	1	2	5

Clusters 21 and 27 are formed in its majority by genes with an oscillatory behavior representing a systematic bias in the preparation of data points. This can be also noticed in their corresponding temporal expression profiles.

III.- Similarity beween clusters from SPCTF to clusters from SPC

Comparison between SPCTF clusters of size 6 or more against the first 5 geometrically closest clusters from the SPC run. Bracketed names are genes identified by Spellman et al. Red names are genes found both in the SPCTF cluster as well as in the closest SPC cluster. Turquoise, green, gold and purple names are genes found both in the SPCTF cluster as well as in the second, third, fourth and fifth closest SPC cluster, respectively.

SPCTF Cluster **1.- 1037**
Size: **37 genes**
Genes: YBL016W YBR005W YBR183W YBR223C [YCL027W] [YCL055W] YCLX07W YCR009C YCR089W YDL023C [YDR085C] YDR125C YGL052W YGL053W [YHR030C] YHR097C YIL079C YJL107C YJL108C YJR153W [YKL104C] YKL109W YKL189W YKR058W YKR061W YLR120C YLR345W YLR350W [YLR452C] YMR065W YMR137C YNL280C [YOL016C] YOR220W YOR238W YOR344C YOR385W
~~~~~~
Centroid distance: **0.00000**
SPC Cluster: **1132**
Size: **37 gene(s)**
Genes: YBL016W YBR005W YBR183W YBR223C [YCL027W] [YCL055W] YCLX07W YCR009C YCR089W YDL023C [YDR085C] YDR125C YGL052W YGL053W [YHR030C] YHR097C YIL079C YJL107C YJL108C YJR153W [YKL104C] YKL109W YKL189W YKR058W YKR061W YLR120C YLR345W YLR350W [YLR452C] YMR065W YMR137C YNL280C [YOL016C] YOR220W YOR238W YOR344C YOR385W

Centroid distance: **0.84379**
SPC Cluster: **1545**
Size: **1 gene(s)**
Genes: [YDL048C]

Centroid distance: **1.18583**
SPC Cluster: **1837**
Size: **1 gene(s)**
Genes: YDL234C

Centroid distance: **1.35325**
~~~~~~

SPC Cluster: **1561**
Size: **1 gene(s)**
Genes: YHR011W

Centroid distance: **1.37545**
SPC Cluster: **1826**
Size: **1 gene(s)**
Genes: YER158C

SPCTF Cluster **2.- 1098**
Size: **27 genes**
Genes: [YDL117W] [YDL127W] [YDL179W] [YDR055W] [YER124C] [YGL028C] [YGL055W] [YGR086C] [YHR143W] [YIL009W] [YJL157C] [YJL159W] [YJL217W] [YKL163W] [YKL164C] [YKL185W] [YLR079W] [YLR194C] [YLR286C] [YMR246W] [YNL015W] [YNL078W] [YNL145W] [YNL192W] YOL155C [YOR263C] [YOR264W]
~~~~~~

Centroid distance: **1.79093**
SPC Cluster: **1060**
Size: **9 gene(s)**
Genes: [YDL179W] [YGR086C] [YIL009W] [YJL217W] [YLR079W] [YNL015W] [YNL078W] [YOR263C] [YOR264W]

Centroid distance: **2.28515**
SPC Cluster: **101**
Size: **1 gene(s)**
Genes: [YBR083W]

Centroid distance: **3.09176**
SPC Cluster: **146**
Size: **1 gene(s)**
Genes: [YMR246W]

Centroid distance: **3.37494**
SPC Cluster: **384**
Size: **1 gene(s)**
Genes: [YLR194C]

Centroid distance: **3.57494**
SPC Cluster: **884**
Size: **4 gene(s)**
Genes: YBR023C [YGR234W] [YIL104C] [YKL182W]

~~~~~~

SPCTF Cluster **3.- 2288**
Size: **61 genes**
Genes: [YAR007C] [YBL111C] [YBL113C] YBR149W [YCR065W] [YDL010W] [YDL018C] [YDL156W] [YDL163W] [YDL164C] [YDR400W] [YDR507C] [YEL040W] [YEL075C] [YEL076C] YEL077C [YER095W] [YER189W] [YFL064C] [YFL066C] [YFL067W] YFL068W [YGR151C] [YGR152C] [YGR221C] [YGR296W] [YHL049C] [YHL050C] [YHR110W] [YHR126C] [YHR149C] [YHR218W] [YJL073W] [YJL074C] [YJL181W] [YJL225C] [YKL113C] [YLR103C] [YLR462W] [YLR463C] [YLR464W] [YLR465C] [YLR466W] [YLR467W] [YMR078C] [YNL312W] [YNL339C] [YOL090W] [YOR033C] [YOR321W] [YPL014W] [YPL153C] [YPL221W] [YPL283C] [YPR120C] [YPR135W] [YPR174C] [YPR175W] [YPR202W] [YPR203W] [YPR204W]
~~~~~~
Centroid distance: **0.05714**
SPC Cluster: **2485**
Size: **38 gene(s)**
Genes: [YAR007C] [YDL156W] [YDL163W] [YDL164C] [YEL075C] [YEL076C] YEL077C [YER189W] [YFL064C] [YFL066C] [YFL067W] YFL068W [YGR151C] [YGR152C] [YGR296W] [YHL049C] [YHL050C] [YHR218W] [YJL074C] [YJL181W] [YJL225C] [YLR103C] [YLR462W] [YLR463C] [YLR464W] [YLR465C] [YLR467W] [YNL339C] [YOL090W] [YOR033C] [YPL283C] [YPR120C] [YPR135W] [YPR174C] [YPR175W] [YPR202W] [YPR203W] [YPR204W]

Centroid distance: **0.65543**
SPC Cluster: **2461**
Size: **1 gene(s)**
Genes: [YDL010W]

Centroid distance: **0.67327**
SPC Cluster: **1014**
Size: **1 gene(s)**
Genes: [YNL082W]

Centroid distance: **0.72679**
SPC Cluster: **2481**
Size: **1 gene(s)**
Genes: [YGR221C]

Centroid distance: **0.80292**
SPC Cluster: **119**
Size: **2 gene(s)**
Genes: [YNL233W] [YOL017W]

SPCTF Cluster **4.- 1225**
Size: **39 genes**
Genes: [YBR243C] [YCL012W] [YCL013W] YCL064C [YCR024C-A] YDR534C [YEL065W] [YER145C] [YGL021W] [YGL116W] [YGR108W] [YGR138C] [YGR176W] [YHL040C] YHL047C [YHR023W] [YIL158W] [YJR092W] [YLR056W] [YLR098C] [YLR131C] [YLR190W]
~~~~~~

[YLR214W] [YML033W] [YML034W] [YML119W] [YMR001C] [YMR032W] [YMR058W] [YNL057W] [YOL158C] [YOR153W] [YOR315W] YOR382W [YPL036W] [YPL141C] [YPL155C] [YPL242C] YPR124W
~~~~~~

Centroid distance: **0.35079**
SPC Cluster: **1367**
Size: **13 gene(s)**
Genes: [YCL012W] [YCL013W] [YGL021W] [YGR176W] [YIL158W] [YJR092W] [YLR098C] [YML033W] [YML034W] [YOR153W] [YPL036W] [YPL141C] [YPL155C]

Centroid distance: **0.76666**
SPC Cluster: **1239**
Size: **12 gene(s)**
Genes: [YBR243C] YCL064C [YCR024C-A] YDR534C [YEL065W] [YER145C] [YHL040C] YHL047C [YLR056W] [YLR214W] [YMR058W] YPR124W

Centroid distance: **0.79204**
SPC Cluster: **1449**
Size: **1 gene(s)**
Genes: [YDR150W]

Centroid distance: **0.79305**
SPC Cluster: **1227**
Size: **7 gene(s)**
Genes: [YHR023W] [YLR190W] [YML119W] [YMR032W] [YNL057W] [YOL158C] [YPL242C]

Centroid distance: **1.17446**
SPC Cluster: **1294**
Size: **1 gene(s)**
Genes: [YGL008C]

SPCTF Cluster **5.- 2183**
Size: **17 genes**
Genes: YAL023C YBR155W YBR156C [YDL093W] [YDL095W] [YDL096C] YDL213C YDR281C [YER003C] YER072W YFL004W YFL045C YGL012W YHR136C YJL012C [YML123C] YPL019C
~~~~~~

Centroid distance: **0.25519**
SPC Cluster: **1866**
Size: **4 gene(s)**
Genes: YBR156C YFL004W YJL012C YPL019C

Centroid distance: **0.53346**
SPC Cluster: **1907**
Size: **1 gene(s)**
Genes: YPR130C

Centroid distance: **0.54984**
SPC Cluster: **1487**
Size: **5 gene(s)**
Genes: YBR155W YDR299W YDR496C YOR116C YPL207W

Centroid distance: **0.55424**
SPC Cluster: **1961**
Size: **1 gene(s)**
Genes: YJL056C

Centroid distance: **0.56298**
SPC Cluster: **2697**
Size: **1 gene(s)**
Genes: YBR222C

SPCTF Cluster **6.- 1137**
Size: **13 genes**
Genes: [YBL002W] [YBL003C] [YBR009C] [YBR010W] [YDL055C] [YDR224C] [YDR225W] [YMR307W] [YNL030W] [YNL031C] [YOR247W] [YOR248W] [YPL127C]
~~~~~~

Centroid distance: **0.14234**
SPC Cluster: **1115**
Size: **10 gene(s)**
Genes: [YBL002W] [YBL003C] [YBR009C] [YBR010W] [YDR224C] [YDR225W] [YNL030W] [YNL031C] [YOR247W] [YOR248W]

Centroid distance: **1.94954**
SPC Cluster: **788**
Size: **1 gene(s)**
Genes: [YPL127C]

Centroid distance: **2.81350**
SPC Cluster: **1257**
Size: **1 gene(s)**
Genes: [YMR307W]

Centroid distance: **3.88026**
SPC Cluster: **359**
Size: **1 gene(s)**
Genes: [YDL055C]

Centroid distance: **3.93668**
SPC Cluster: **1260**
Size: **1 gene(s)**
Genes: [YNL126W]
~~~~~~

SPCTF Cluster **7.- 1459**
Size: **11 genes**
Genes: YEL022W YER017C YER060W YER093C YER137C YGL131C YML111W YMR317W YNL271C YPL069C YPL071C
~~~~~~
Centroid distance: **0.00000**
SPC Cluster: **1716**
Size: **11 gene(s)**
Genes: YEL022W YER017C YER060W YER093C YER137C YGL131C YML111W YMR317W YNL271C YPL069C YPL071C

Centroid distance: **0.47113**
SPC Cluster: **1928**
Size: **1 gene(s)**
Genes: YML103C

Centroid distance: **0.49263**
SPC Cluster: **2341**
Size: **1 gene(s)**
Genes: YLR341W

Centroid distance: **0.56805**
SPC Cluster: **2821**
Size: **1 gene(s)**
Genes: YDR319C

Centroid distance: **0.60510**
SPC Cluster: **1187**
Size: **1 gene(s)**
Genes: YGR097W

SPCTF Cluster **8.- 1934**
Size: **11 genes**
Genes: YIL007C YIL012W YIL016W YKL119C YKL123W YLR012C YLR128W YLR443W YMR156C YNL252C YOR177C
~~~~~~
Centroid distance: **0.00000**
SPC Cluster: **2440**
Size: **11 gene(s)**
Genes: YIL007C YIL012W YIL016W YKL119C YKL123W YLR012C YLR128W YLR443W YMR156C YNL252C YOR177C

Centroid distance: **0.35731**

SPC Cluster: **2426**
Size: **1 gene(s)**
Genes: YHR105W

Centroid distance: **0.46443**
SPC Cluster: **1407**
Size: **6 gene(s)**
Genes: YMR269W YNL315C YOR044W YOR210W YOR394W YPL168W

Centroid distance: **0.47389**
SPC Cluster: **2654**
Size: **1 gene(s)**
Genes: YNL136W

Centroid distance: **0.59443**
SPC Cluster: **1406**
Size: **1 gene(s)**
Genes: YGR080W

SPCTF Cluster **9.- 1324**
Size: **10 genes**
Genes: YAR033W YBR004C YBR165W YBR194W YBR225W YER130C YJR038C YLR332W YNL053W YPL089C
~~~~~~
Centroid distance: **0.03907**
SPC Cluster: **1466**
Size: **8 gene(s)**
Genes: YAR033W YBR004C YBR165W YBR194W YBR225W YJR038C YNL053W YPL089C

Centroid distance: **0.53801**
SPC Cluster: **1506**
Size: **2 gene(s)**
Genes: YGL051W YGL193C

Centroid distance: **0.58490**
SPC Cluster: **2690**
Size: **1 gene(s)**
Genes: YNL107W

Centroid distance: **0.60263**
SPC Cluster: **856**
Size: **6 gene(s)**
Genes: YGR290W YKR072C YLR398C YMR124W YMR164C YNL106C

Centroid distance: **0.63082**
SPC Cluster: **1486**
~~~~~~

Size: **1 gene(s)**
Genes: YER130C

SPCTF Cluster **10.- 2635**
Size: **10 genes**
Genes: YGR236C YNL305C YOL082W YOL104C YOL162W YOR036W YOR130C YOR202W YOR302W YPR081C
~~~~~~
Centroid distance: **0.20630**
SPC Cluster: **1909**
Size: **3 gene(s)**
Genes: YOL082W YOR202W YOR302W

Centroid distance: **0.25680**
SPC Cluster: **2674**
Size: **1 gene(s)**
Genes: YGR236C

Centroid distance: **0.35476**
SPC Cluster: **1765**
Size: **1 gene(s)**
Genes: YOL104C

Centroid distance: **0.41562**
SPC Cluster: **2755**
Size: **1 gene(s)**
Genes: YOR036W

Centroid distance: **0.45740**
SPC Cluster: **2191**
Size: **1 gene(s)**
Genes: YOL162W

SPCTF Cluster **11.- 2864**
Size: **10 genes**
Genes: YDL126C YDL143W YDR266C YGL043W YIL172C YOL081W YOR243C YPL091W YPR006C YPR151C
~~~~~~
Centroid distance: **0.17315**
SPC Cluster: **2029**
Size: **2 gene(s)**
Genes: YDL126C YDL143W

Centroid distance: **0.29730**
SPC Cluster: **2036**
Size: **12 gene(s)**
Genes: YBL006C YBR014C YDR266C YDR357C YGL043W YLR114C YLR195C [YML116W] [YMR055C] YNR057C YPL210C YPR151C

Centroid distance: **0.56457**
SPC Cluster: **1981**
Size: **1 gene(s)**
Genes: YOR243C

Centroid distance: **0.65199**
SPC Cluster: **1328**
Size: **1 gene(s)**
Genes: YDR523C

Centroid distance: **0.66983**
SPC Cluster: **1357**
Size: **1 gene(s)**
Genes: YOL081W

SPCTF Cluster **12.- 1437**
Size: **9 genes**
Genes: YBR275C YDR078C YGR174C YIL034C YKL194C YLR346C YOR193W YPL167C YPR172W
~~~~~~
Centroid distance: **0.06991**
SPC Cluster: **1689**
Size: **6 gene(s)**
Genes: YDR078C YGR174C YIL034C YKL194C YOR193W YPR172W

Centroid distance: **0.32602**
SPC Cluster: **1532**
Size: **2 gene(s)**
Genes: YBR275C YPL167C

Centroid distance: **0.35362**
SPC Cluster: **2245**
Size: **1 gene(s)**
Genes: YEL069C

Centroid distance: **0.40170**
SPC Cluster: **2344**
Size: **1 gene(s)**
Genes: YGR101W
~~~~~~

Centroid distance: **0.42400**
SPC Cluster: **2813**
Size: **1 gene(s)**
Genes: YDR137W

SPCTF Cluster **13.- 4922**
Size: **9 genes**
Genes: [YBR088C] [YDL003W] [YDR097C] [YER070W] [YGR189C] [YLR183C] [YML027W] [YOR074C] [YPL256C]
~~~~~~

Centroid distance: **0.11988**
SPC Cluster: **4970**
Size: **7 gene(s)**
Genes: [YBR088C] [YDL003W] [YDR097C] [YER070W] [YLR183C] [YML027W] [YPL256C]

Centroid distance: **1.64660**
SPC Cluster: **2354**
Size: **1 gene(s)**
Genes: [YKL113C]

Centroid distance: **1.89475**
SPC Cluster: **4981**
Size: **1 gene(s)**
Genes: [YGR189C]

Centroid distance: **2.17171**
SPC Cluster: **4982**
Size: **3 gene(s)**
Genes: [YER111C] [YKL045W] [YMR179W]

Centroid distance: **2.17658**
SPC Cluster: **1270**
Size: **2 gene(s)**
Genes: [YER095W] [YPL153C]

SPCTF Cluster **14.- 2719**
Size: **9 genes**
Genes: YBL082C YBL083C [YCL062W] [YCL063W] [YFL037W] YFR034C YJL183W YLR083C YPL227C
~~~~~~

Centroid distance: **0.23212**
SPC Cluster: **2640**

Size: **1 gene(s)**
Genes: [YCL063W]

Centroid distance: **0.24200**
SPC Cluster: **2702**
Size: **1 gene(s)**
Genes: YBL082C

Centroid distance: **0.26616**
SPC Cluster: **2699**
Size: **1 gene(s)**
Genes: YPL227C

Centroid distance: **0.31713**
SPC Cluster: **2680**
Size: **1 gene(s)**
Genes: [YCL062W]

Centroid distance: **0.59447**
SPC Cluster: **2553**
Size: **1 gene(s)**
Genes: YFR034C

SPCTF Cluster **15.- 1374**
Size: **8 genes**
Genes: YNL218W YNL335W [YOL034W] YOL091W YOR017W [YOR114W] YOR227W [YPL124W]
~~~~~~

Centroid distance: **0.01014**
SPC Cluster: **1591**
Size: **7 gene(s)**
Genes: YNL218W YNL335W [YOL034W] YOL091W YOR017W [YOR114W] YOR227W

Centroid distance: **0.49711**
SPC Cluster: **1355**
Size: **1 gene(s)**
Genes: [YPL124W]

Centroid distance: **0.51149**
SPC Cluster: **2068**
Size: **1 gene(s)**
Genes: YOR082C

Centroid distance: **0.60859**
SPC Cluster: **2060**
Size: **1 gene(s)**
~~~~~~

Genes: YPR015C

Centroid distance: **0.61969**
SPC Cluster: **1515**
Size: **1 gene(s)**
Genes: YOR284W

SPCTF Cluster **16.- 2097**
Size: **8 genes**
Genes: YAR074C YBL072C YBL092W YBR075W YBR121C YHL035C YKL014C YPL231W
~~~~~~

Centroid distance: **0.16815**
SPC Cluster: **2396**
Size: **1 gene(s)**
Genes: YBR121C

Centroid distance: **0.27988**
SPC Cluster: **2394**
Size: **1 gene(s)**
Genes: YAR074C

Centroid distance: **0.30055**
SPC Cluster: **2204**
Size: **1 gene(s)**
Genes: YKL014C

Centroid distance: **0.30168**
SPC Cluster: **2552**
Size: **1 gene(s)**
Genes: YBL026W

Centroid distance: **0.32553**
SPC Cluster: **1626**
Size: **1 gene(s)**
Genes: YBL092W

SPCTF Cluster **17.- 2565**
Size: **8 genes**
Genes: YDR293C YER043C YGR134W YJR016C YKL020C YKL205W YKL211C YMR309C
~~~~~~

Centroid distance: **0.13469**
SPC Cluster: **2737**
Size: **3 gene(s)**

Genes: YGR134W YKL020C YKL211C

Centroid distance: **0.17522**
SPC Cluster: **2589**
Size: **1 gene(s)**
Genes: YKL205W

Centroid distance: **0.25592**
SPC Cluster: **2593**
Size: **1 gene(s)**
Genes: YJR016C

Centroid distance: **0.28580**
SPC Cluster: **2734**
Size: **1 gene(s)**
Genes: YGL063W

Centroid distance: **0.34545**
SPC Cluster: **2741**
Size: **1 gene(s)**
Genes: YDR141C

SPCTF Cluster **18.- 863**
Size: **7 genes**
Genes: YAL049C YAR060C YBR003W YBR022W YBR232C YBR246W YGL175C
~~~~~~

Centroid distance: **0.00000**
SPC Cluster: **920**
Size: **7 gene(s)**
Genes: YAL049C YAR060C YBR003W YBR022W YBR232C YBR246W YGL175C

Centroid distance: **0.39962**
SPC Cluster: **256**
Size: **1 gene(s)**
Genes: YDR095C

Centroid distance: **0.40247**
SPC Cluster: **2558**
Size: **1 gene(s)**
Genes: YAL030W

Centroid distance: **0.43856**
SPC Cluster: **495**
Size: **1 gene(s)**
Genes: YLR396C
~~~~~~

Centroid distance: **0.44543**
SPC Cluster: **2433**
Size: **1 gene(s)**
Genes: [YLR233C]

SPCTF Cluster **19.- 1120**
Size: **7 genes**
Genes: [YGR279C] [YKL001C] [YLL061W] [YLR180W] [YLR303W] YLR304C [YNL037C]
~~~~~~
Centroid distance: **0.00000**
SPC Cluster: **1203**
Size: **7 gene(s)**
Genes: [YGR279C] [YKL001C] [YLL061W] [YLR180W] [YLR303W] YLR304C [YNL037C]

Centroid distance: **0.90013**
SPC Cluster: **1190**
Size: **3 gene(s)**
Genes: YDR384C [YJR148W] YOR135C

Centroid distance: **0.96211**
SPC Cluster: **1042**
Size: **1 gene(s)**
Genes: YDR059C

Centroid distance: **0.97290**
SPC Cluster: **913**
Size: **1 gene(s)**
Genes: YMR191W

Centroid distance: **0.99351**
SPC Cluster: **1838**
Size: **1 gene(s)**
Genes: [YLL062C]

SPCTF Cluster **20.- 1212**
Size: **6 genes**
Genes: YMR269W YNL315C YOR044W YOR210W YOR394W YPL168W
~~~~~~
Centroid distance: **0.00000**
SPC Cluster: **1407**
Size: **6 gene(s)**
Genes: YMR269W YNL315C YOR044W YOR210W YOR394W YPL168W

Centroid distance: **0.38049**

SPC Cluster: **2084**
Size: **1 gene(s)**
Genes: YPL002C

Centroid distance: **0.45938**
SPC Cluster: **2426**
Size: **1 gene(s)**
Genes: YHR105W

Centroid distance: **0.46175**
SPC Cluster: **1386**
Size: **1 gene(s)**
Genes: YNL127W

Centroid distance: **0.46443**
SPC Cluster: **2440**
Size: **11 gene(s)**
Genes: YIL007C YIL012W YIL016W YKL119C YKL123W YLR012C YLR128W YLR443W YMR156C YNL252C YOR177C

SPCTF Cluster **21.- 1563**
Size: **6 genes**
Genes: YLR406C YNR017W YOR109W YOR133W YPL079W YPL081W
~~~~~~

Centroid distance: **0.11617**
SPC Cluster: **2023**
Size: **2 gene(s)**
Genes: YPL079W YPL081W

Centroid distance: **0.28772**
SPC Cluster: **1936**
Size: **1 gene(s)**
Genes: YOR109W

Centroid distance: **0.30948**
SPC Cluster: **2375**
Size: **2 gene(s)**
Genes: YMR321C YNR021W

Centroid distance: **0.39926**
SPC Cluster: **1957**
Size: **1 gene(s)**
Genes: YNR017W

Centroid distance: **0.44130**
SPC Cluster: **1339**
Size: **1 gene(s)**
~~~~~~

Genes: YOR133W

SPCTF Cluster **22.- 810**
Size: **6 genes**
Genes: YGR290W YKR072C YLR398C YMR124W YMR164C YNL106C
~~~~~~

Centroid distance: **0.00000**
SPC Cluster: **856**
Size: **6 gene(s)**
Genes: YGR290W YKR072C YLR398C YMR124W YMR164C YNL106C

Centroid distance: **0.29243**
SPC Cluster: **2223**
Size: **1 gene(s)**
Genes: YGR023W

Centroid distance: **0.34637**
SPC Cluster: **2195**
Size: **1 gene(s)**
Genes: YLL018C

Centroid distance: **0.37770**
SPC Cluster: **673**
Size: **1 gene(s)**
Genes: YJR050W

Centroid distance: **0.44148**
SPC Cluster: **2824**
Size: **2078 gene(s)**
Genes: YAL014C YAL021C YAL036C YAL037W YAL042W YAL046C YAL063C YAR028W...

SPCTF Cluster **23.- 1620**
Size: **6 genes**
Genes: YBR049C [YJL092W] YLR358C YMR190C [YNL197C] YPR104C
~~~~~~

Centroid distance: **0.22920**
SPC Cluster: **1608**
Size: **2 gene(s)**
Genes: YMR190C YPR104C

Centroid distance: **0.36007**
SPC Cluster: **1629**
Size: **1 gene(s)**
Genes: [YNL197C]

Centroid distance: **0.36916**
SPC Cluster: **1524**
Size: **1 gene(s)**
Genes: YLR358C

Centroid distance: **0.43293**
SPC Cluster: **1517**
Size: **1 gene(s)**
Genes: YBR049C

Centroid distance: **0.57383**
SPC Cluster: **2326**
Size: **1 gene(s)**
Genes: [YGR113W]

SPCTF Cluster **24.- 1002**
Size: **6 genes**
Genes: [YDL037C] [YDL039C] [YGL032C] [YIL117C] YJL170C YPL192C
~~~~~~
Centroid distance: **0.76847**
SPC Cluster: **174**
Size: **1 gene(s)**
Genes: YPL192C

Centroid distance: **2.15339**
SPC Cluster: **1082**
Size: **2 gene(s)**
Genes: [YDL037C] [YDL039C]

Centroid distance: **3.23832**
SPC Cluster: **842**
Size: **1 gene(s)**
Genes: [YIL117C]

Centroid distance: **4.40884**
SPC Cluster: **1055**
Size: **1 gene(s)**
Genes: YJL170C

Centroid distance: **4.71721**
SPC Cluster: **1132**
Size: **37 gene(s)**
Genes: YBL016W YBR005W YBR183W YBR223C [YCL027W] [YCL055W] YCLX07W YCR009C YCR089W YDL023C [YDR085C] YDR125C YGL052W YGL053W [YHR030C] YHR097C YIL079C YJL107C YJL108C YJR153W [YKL104C] YKL109W YKL189W YKR058W YKR061W
~~~~~~

YLR120C YLR345W YLR350W [YLR452C] YMR065W YMR137C YNL280C [YOL016C] YOR220W YOR238W YOR344C YOR385W

SPCTF Cluster **25.- 2139**
Size: **6 genes**
Genes: YAR061W YNL337W [YOR298W] YOR357C YPL034W [YPR107C]
~~~~~~
Centroid distance: **0.04555**
SPC Cluster: **2667**
Size: **4 gene(s)**
Genes: YNL337W [YOR298W] YOR357C YPL034W

Centroid distance: **0.27199**
SPC Cluster: **2546**
Size: **1 gene(s)**
Genes: YPR077C

Centroid distance: **0.38555**
SPC Cluster: **1514**
Size: **1 gene(s)**
Genes: [YPR107C]

Centroid distance: **0.38615**
SPC Cluster: **1894**
Size: **1 gene(s)**
Genes: YOL108C

Centroid distance: **0.42936**
SPC Cluster: **1896**
Size: **1 gene(s)**
Genes: YPL072W

SPCTF Cluster **26.- 2903**
Size: **6 genes**
Genes: YAL056W YCL032W YFR022W YGR249W YHR195W YLR024C
~~~~~~
Centroid distance: **0.09453**
SPC Cluster: **2561**
Size: **2 gene(s)**
Genes: YGR249W YIR011C
Centroid distance: **0.15921**
SPC Cluster: **2624**
Size: **1 gene(s)**
Genes: YFR022W

Centroid distance: **0.34037**
SPC Cluster: **2118**
Size: **1 gene(s)**
Genes: YCL005W

Centroid distance: **0.36944**
SPC Cluster: **2567**
Size: **1 gene(s)**
Genes: YLR024C

Centroid distance: **0.43071**
SPC Cluster: **2073**
Size: **1 gene(s)**
Genes: YJL162C

SPCTF Cluster **27.- 3140**
Size: **6 genes**
Genes: [YBL030C] YJL052W YLR179C YML035C YMR303C YNL168C
~~~~~~

Centroid distance: **0.22395**
SPC Cluster: **2357**
Size: **3 gene(s)**
Genes: [YBL030C] YNL164C YPL240C

Centroid distance: **0.28114**
SPC Cluster: **2349**
Size: **1 gene(s)**
Genes: YJL052W

Centroid distance: **0.33627**
SPC Cluster: **1876**
Size: **1 gene(s)**
Genes: YMR134W

Centroid distance: **0.38248**
SPC Cluster: **2808**
Size: **1 gene(s)**
Genes: YPL059W

Centroid distance: **0.39218**
SPC Cluster: **952**
Size: **1 gene(s)**
Genes: YLR011W

~~~~~~

Appendix D

Published Paper

Physica A 389 (2010) 5689–5697

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

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

ARTICLE INFO

Article history:
Received 1 April 2010
Received in revised form 2 September 2010
Available online 18 September 2010

Keywords:
Superparamagnetic clustering
Similarity measure
Microarrays
Cell cycle genes
Transcription factors

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 which we call SPCTF, we analyze the 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 worthy of further experimental studies because they would probably lead to new cell cycle genes. Finally, our idea of introducing the available information about transcription factors to optimize the gene classification could be implemented for other distance-based clustering algorithms.

1. Introduction

DNA microarrays allow the comparison of the expression levels of all genes in an organism in a single experiment, which often involves 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 [3–5]. 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 [7–9]. 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 the additional data acquired,

* Corresponding author. Tel.: +52 4448342000; fax: +52 4448342010.
E-mail 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).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.09.006