Modelling and Clustering of Gene Expressions Using RBFs and a Shape Similarity Metric

Carla S. Möller-Levet and Hujun Yin

Electrical Engineering and Electronics, University of Manchester
Institute of Science and Technology, Manchester M60 IQD, UK
c.moller-levet@postgrad.umist.ac.uk
h.yin@umist.ac.uk

Abstract. This paper introduces a novel approach for gene expression time-series modelling and clustering using neural networks and a shape similarity metric. The modelling of gene expressions by the Radial Basis Function (RBF) neural networks is proposed to produce a more general and smooth characterisation of the series. Furthermore, we identified that the use of the correlation coefficient of the derivative of the modelled profiles allows the comparison of profiles based on their shapes and the distributions of time points. The series are grouped into similarly shaped profiles using a correlation based fuzzy clustering algorithm. A well known dataset is used to demonstrate the proposed approach and a set of known genes are used as a benchmark to evaluate its performance. The results show the biological relevance and indicate that the proposed method is a useful technique for gene expression time-series analysis.

1 Introduction

With microarray experiments it is possible to measure simultaneously the activity levels of thousands of genes [1]. An appropriate clustering of gene expression data can lead to meaningful classification of diseases, identification of functionally related genes, logical descriptions of gene regulation, etc. Various statistical and clustering methods have been applied successfully to microarray data [2–5].

Gene expression time-series are generally noisy, short and usually unevenly sampled. To overcome these undesirable characteristics, we propose to model gene expression profiles, using the RBF neural networks [6]. Most existing methods used to compare expression profiles directly operate on the time points. While modelling the profiles can lead to more generalised, smooth characterisation of gene expressions. Standard time-series models are limited by the underlying assumptions of the model, such as stationarity or length of the time-series. In contrast, the use of artificial neural networks for modelling time-series is not restricted by model assumptions and linearity, as well as the noise, irregular sampling and shortness of the time-series. Other modelling techniques have been proposed recently, e.g., in [7, 8], gene expression time-series are modelled using mixed-effect models within a mixture model based clustering. Equally spaced

Z.R. Yang et al. (Eds.): IDEAL 2004, LNCS 3177, pp. 1–10, 2004.

cubic spline[1] are used for both mean and random effects, that is, sum of the smooth population mean spline function dependent on time and gene cluster, and a spline function with random coefficients for individual gene effects and Gaussian measurement noise. One of the advantages of the proposed modelling over the mixed-effects modelling is that each gene can be modelled independently, which make the models useful for different types of analysis, not just for clustering.

In microarray experiments, the absolute intensity of gene expression is not relevant, instead, the relative change of intensity characterised by the shape of the expression profile is regarded as characteristic and informative. In addition, biological processes are often sampled at short or long intervals of time when intense or moderate biological activity is taking place, leading to unevenly distributed sampling points. We identified that the use of the correlation coefficient of the derivative of the modelled profiles allows the comparison of profiles based on their shape and the distribution of their time points. This measure is further used in a correlation based fuzzy clustering algorithm.

2 Modelling with Radial Basis Function Neural Networks

Radial basis networks have a single hidden layer, where the nodes are Gaussian kernels, and a linear output layer. The radial basis function has the form:

$$f(x) = \sum_{i=1}^{n_r} w_i \phi(\|c_i - x\|) + b \tag{1}$$

where x is the input vector, $\phi(\cdot)$ is a Gaussian function kernel, $\|\cdot\|$ denotes the Euclidean norm, w_i are the weights of the second layer, c_i is the vector of the centre of the i-th kernel, and n_r is the total number of kernels.

The problem of RBF approximation is to find appropriate centres and widths of the hidden nodes, and weights of the linear layer. The network is linear in the parameters when all RBF centres and widths of the hidden layer are fixed. Then, the output layer linearly combines the output of the hidden layer and the only adjustable parameters are the weights. We chose the Orthogonal Least Squares (OLS) learning algorithm proposed in [9] for training the RBF neural networks. The algorithm allows the selection of the centres one by one in a rational procedure. Each selected centre maximises the increment to the explained variance of the desired output and it is not necessary to use all the time points as the centres. However, this method considers all kernels with an equal width, which is inadequate when the sampling points are not evenly distributed. In order to improve the approximation, we complemented the OLS learning algorithm with a heuristic search for the optimal width for each of the candidate centres, which implies the recalculation of the regression matrix before a new centre is selected. The optimal width minimises the mean square error of a piecewise linear fit of a segment of the series and the RBF model of the segment [10].

[1] Spline functions are piecewise polynomials of degree n that are connected together (at point called knots) so as to have $n-1$ continuous derivations.

3 Correlation-Based Similarity Metric and Fuzzy Clustering

The Pearson correlation coefficient is a popular similarity measure to compare time-series. Yet, the correlation is not necessarily coherent with the shape and it does not consider the order of the time points and uneven sampling intervals. If the correlation, however, is performed on the derivatives of the modelled profiles[2], these drawbacks can be solved. The differentiation refers to the rates of change or shape of the profile.

The objective function that measures the desirability of partitions in fuzzy c-means clustering (FCM) [11] is described by,

$$J(x, v, u) = \sum_{i=1}^{n_c} \sum_{j=1}^{n_g} u_{ij}^m \, d^2(x_j, v_i) \tag{2}$$

where n_c is the number of clusters, n_g is the number of vectors to cluster, u_{ij} is the value of the membership degree of the vector x_j to the cluster i, $d^2(x_j, v_i)$ is the squared distance between vector x_j and prototype v_i, and m is the parameter that determines the degree of overlap of fuzzy clusters. The FCM optimisation process is operated on the inner product induced norms, the use of a different metric will require the recalculation of v and u to minimise the objective function. In [12], a correlation-based fuzzy clustering is presented by defining d as $d_2^2(x_j, v_i) = f(\rho(x_j, v_i))$, where f is a continuously decreasing function and ρ is the Pearson's correlation coefficient between x_j and v_i. If the time-series are normalised with mean values set to cero and norm set to one, every time point is localised at the perimeter of an hypersphere of radius one and the value of $\rho(x_j, v_i)$ is the cosine of the angle θ, between x_j and v_i. The chord d_2, formed by x_j and v_i, is $d_2^2(x_j, v_i) = 2(1 - \rho(x_j, v_i))$, which is the Euclidean distance in this particular space.

There are two user-defined parameters in the FCM algorithm, i.e., the number of clusters n_c and the fuzziness parameter m. In this paper, we have chosen the PBM-index, proposed in [13], to validate the number of clusters. In addition, to avoid the random initialisation of the partition matrix, the mountain clustering initialisation was implemented [14]. The selection of m for an optimal performance of fuzzy clustering for microarray data is addressed in [3], and can be aided by validity measures and membership plots (plot of membership degrees into grayscales in a hierarchical fashion).

4 Yeast Cell Cycle Dataset

In Spellman *et al* [2], cell-cycle-regulated genes of the Yeast *Saccharomayces cerevisiae* were identified by microarray hybridisation. Yeast cultures were synchronised by three independent methods: α factor arrest, elutriation and arrest

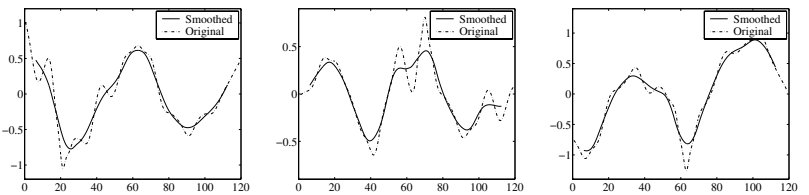
[2] The modelled profiles are further smoothed to reduce noise impact of the derivatives.

of a cdc15 temperature sensitive mutant. We utilised the temporal expression of the yeast culture synchronised by α factor arrest to illustrate the proposed method.

In the analysis of Spellman *et al* [2], 800 cell-cycle-regulated genes were identified using Fourier analysis of combined data of all three experiments. Among the 800 genes, 511 with no missing values for the α experiment are available from their web site[3]. Recently, Luan and Li [15] re-analysed the data using a shape-invariant model together with a false discovery rate procedure to identify periodically expressed genes. They identified 297 cell-cycle-regulated genes in the α experiment. Out of these 297 genes, 208 have no missing values. In addition, there are 104 genes determined to be cell-cycle-regulated by traditional methods. Out of these 104 genes, 71 with no missing values are available from the Spellman *et al* [2] dataset. In [2], 95 of the 104 genes were identified as cell-cycle-regulated, while in [15] 47 were identified. We utilised the 511, 208 and 71 genes corresponding to the cell-cycle-regulated genes identified by Spellman *et al* [2], Luan and Li [15], and traditional methods, respectively, to form three test datasets.

4.1 Modelling

The first step for the proposed approach is the modelling of the series with the RBF neural networks. The moving average smoothing technique was used to further smooth the modelled series. Figure 1 presents three example model profiles from the Spellman *et al* dataset and their corresponding smoothed expressions.



(a) YAL040C (b) YDR077W (c) YDL055C

Fig. 1. Example genes and their corresponding smoothed expression. In every figure the horizontal axis denotes time [mins] and the vertical axis denotes the expression level (log_2(ratio)).

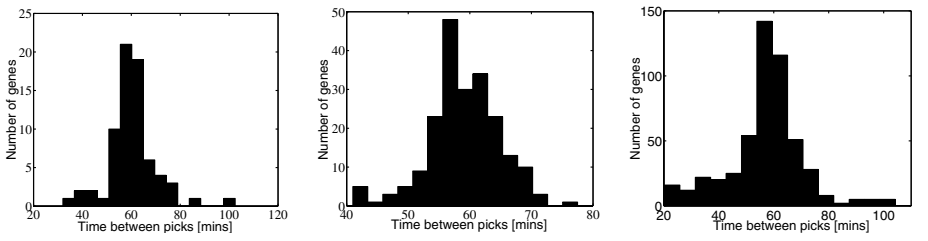
The cell cycle period for the α dataset has been identified with different methods, producing different results. In [16], a time-frequency analysis using wavelet transforms is used to identify the period. The authors conclude that the dominant period is not that of the cell cycle but, the higher frequency 30-40

[3] Dataset available from http://cellcycle-www.stanford.edu

min submultiples of the cycle period. In [4], the authors analyse the similarity of a time-series curve with itself and deduce that the period is 70 minutes. As in [16], the authors in [4] observe that there are very strong indications of underlying oscillatory phenomena with periods smaller that the observed cell cycles, around 20 and 40 minutes. Later, in [8], five clusters are identified and the times between peaks for four clusters are estimated to be 67, 63.4, 54.2, and 61.4 minutes. The modelling and smoothing of the profiles proposed in this paper allow the estimation of times between peaks. The estimated times are comparable to the previous period identification results, having higher occurrences at periods between 55 and 65 minutes. Table 1 presents the summery of the estimated times between two peaks for the three datasets, and Figure 2 presents the corresponding histograms.

Table 1. Summary of the estimated times (in minutes) between two peaks, for the 71, 208 and 511 genes datasets.

Dataset	Mean	Median	Std. Dev.
71 genes	60.2289	59.2500	10.1421
208 genes	59.1010	58.7500	5.7688
511 genes	57.0186	57.7500	13.6742



(a) (b) (c)

Fig. 2. Histograms of estimated times between peaks of genes of the three datasets used. (a) 71 genes, (b) 208 genes and (c) 511 genes.

4.2 Clustering

The PBM-index validates 4 clusters for all three datasets. Figure 3 plots the PBM-index as a function of the number of clusters for the three datasets. The genes were assigned to the clusters according to their highest membership degree. A hierarchical clustering of the partition matrix was performed to order the genes and obtain a complete visualisation of the results. In this way, the membership plot can be utilised to identify genes with similar distribution of membership degree across all the clusters. The clustering results are shown in Figure 4. We classified the clusters according to the different cell cycle phases used in [2]. The genes were classified based on the minimum error, comparing their peak times

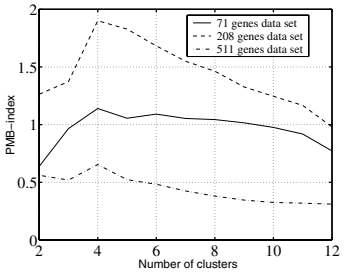


Fig. 3. The maximum value of the PBM-index indicates the formation of the smallest number of compact clusters with large separation.

to the peak times described in [2] for each phase. Tables 2 and 3 present the distributions of the genes in each cluster among the cell cycle phases. In the 71 and 208 genes datasets, the genes in each cluster belong mainly to one phase or the two adjacent phases. In the 511 genes dataset, the genes are more spread mainly due to the inclusion of genes which are not cell-cycle-regulated and have a low membership degree to all the clusters, for example genes in Figure 4(f), cluster II.

In the 71 genes identified by traditional methods, the phase of the genes with the highest membership degree to each cluster coincide with the phase represented by the cluster. The clustering results of these previously known genes, indeed show biological relevance, indicating that the proposed method is a useful technique for gene expression time-series analysis. In Figure 4 (a), a region of genes with an extremely high membership to cluster II can be observed. All those genes correspond to the Histons identified by traditional biological methods and their profiles are illustrated in Figure 5.

Table 2. Distribution of the 71 genes dataset, among the five different cell-cycle phases over the four clusters obtained using the proposed method. The column "Gene" corresponds to the gene with the highest membership to the cluster, and "Phase", corresponds to the phase of the gene identified by traditional methods.

Cluster	M/G1	G1	S	G2	M	Gene	Phase
I(14)	0	29	0	0	0	YKL113C	G1
II(7)	0	1	13	0	0	YNL030W	S
III(8)	7	6	0	0	3	YNL192W	M/G1
IV(12)	3	3	0	2	7	YMR001C	G2/M

In the case of the dataset formed by the 208 genes identified as cell-cycle-regulated by [15], the gene with the highest membership degree to the first cluster, YPR174C, has not been characterized. Table 4, presents the genes with high similarity to YPR174C according to their membership distribution. YPR019W, essential for initiation of DNA replication, presents the highest membership de-

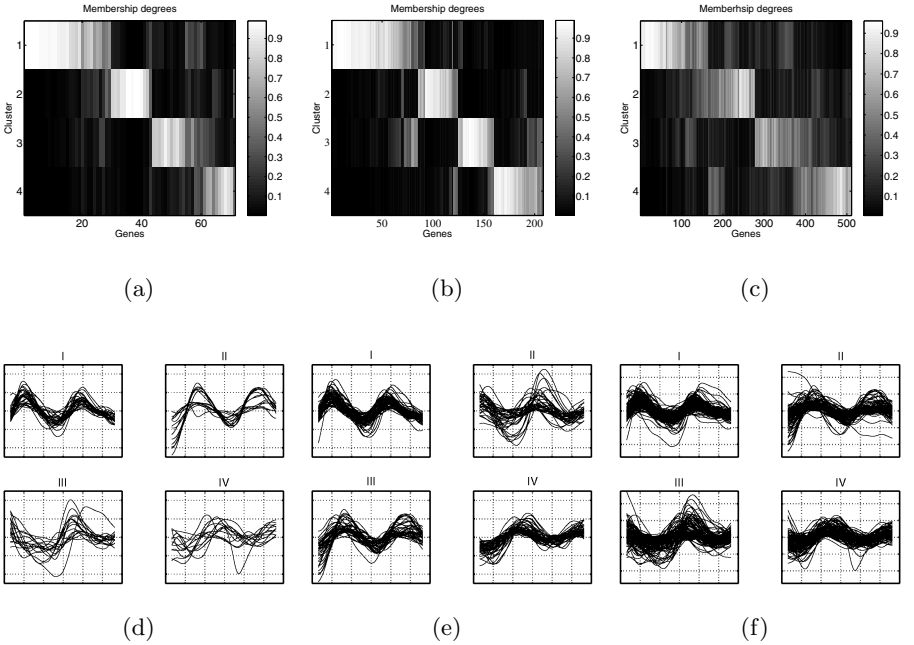


Fig. 4. Clustering results (membership plots and time-series plots) for the 71 genes dataset (figures (a) and (d)), 208 genes dataset (figures (b) and (e)) and 511 genes dataset (figures (c) and (f)). In figures d, e, and f, the horizontal axis denotes time [0, 20, 40, 60, 80, 100, and 120 mins] and the vertical axis denotes the expression level (log_2(ratio)).

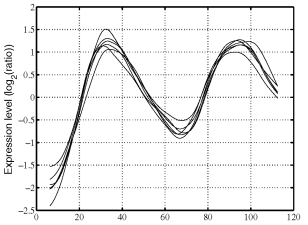


Fig. 5. Genes with high membership to cluster II, illustrated in Figure 4 (a).

gree to the second cluster. YNL030W, core histone required for chromatin assembly and chromosome function, presents the highest membership degree to the third cluster which mainly represents genes of the S phase. Figure 4(b) shows a sharp boundary between clusters II and III. The corresponding time-series plots, Figure 4(e), show that these clusters are shifted almost half cell cycle, thus, presenting opposite shapes. Finally, YGL021W, presents the highest membership degree to the fourth cluster.

In the dataset formed by the 511 genes identified as cell-cycle-regulated by [2], cluster I mainly contains genes belonging to phase G1, and the gene with the

Table 3. Distribution of the 208 and 511 genes datasets, among the five different cell cycle phases over the four clusters obtained using the proposed method.

208 genes dataset						511 genes dataset					
Cluster	M/G1	G1	S	G2	M	Cluster	M/G1	G1	S	G2	M
I(78)	0	78	0	0	0	I(158)	5	148	1	2	2
II(37)	28	2	0	0	7	II(107)	4	33	39	31	0
II(42)	0	10	24	8	0	II(113)	49	20	4	10	30
IV(51)	0	0	0	33	18	IV(133)	27	4	5	53	44

Table 4. Genes with high similarity to YPR174C according to their membership distribution. "Phase" corresponds to the phase identified by traditional methods, "U" stands for uncharacterised gene.

Membership	Gene	Phase
0.9803	YAR007C	G1
0.9837	YLR103C	G1
0.9794	YDL156W	U
0.9882	YDL163W	U
0.9896	**YPR174C**	**U**
0.9773	YDL164C	G1
0.9812	YBR088C	G1
0.9739	YPL153C	G1
0.9739	YDL003W	G1

highest membership degree to cluster I, YLR103C, was identified to peak in the G1 phase by traditional methods. YOR248W, uncharacterised, has the highest membership degree to cluster II, which has genes associated to the G1, S and G2 phases. YHR005C, presents the highest membership degree to cluster III. Cluster IV is mainly formed by genes belonging to phases G2 and M. YGL021W, has the highest membership degree to this cluster. This gene, is also the highest member of cluster IV (formed by G2 and M phase genes) in the 208 genes dataset.

The membership and time-series plots indicate that the correlation based fuzzy clustering of the derivatives allow the grouping of profiles based on their shape. The distribution of the genes from the three datasets among the cell cycle phases indicates that the proposed method is able to extract meaningful groups with biological relevance.

5 Conclusions

In this paper, the modelling of gene expression profiles using the RBF neural networks is proposed, leading to a more generalised, smooth characterisation of gene expressions. An extended ORS method is proposed to optimise the network parameters. The models obtained are smoothed to reduce noise and differentiated to characterise the shapes of the expression profiles. Then, the use of the correlation coefficient of the derivatives of the modelled profiles as a similarity measure, allows the comparison of profiles based on their shapes and the

distributions of their time points. Finally, considering the advantages of fuzzy membership, a correlation based fuzzy clustering algorithm to group profiles according to the proposed similarity metric has been used. The well known dataset in [2] has been used to demonstrate the advantages of the proposed approach. The set of genes identified by traditional methods was used as a benchmark to evaluate the performance of the proposed approach with coherent biological meanings.

Acknowledgments

This research was supported by grants from ABB Ltd. U.K., an Overseas Research Studentship (ORS) award by Universities U.K. and Consejo Nacional de Ciencia y Tecnologia (CONACYT).

References

1. Brown, P., Botstein, D.: Exploring the new world of the genome with DNA microarrays. Nature Genetics supplement **21** (1999) 33–37
2. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of yeast *Saccharamyces cerevisiae* by microarray hybridization. Mol. Biol. Cell **9** (1998) 3273–3297
3. Dembélé, D., Kastner, P.: Fuzzy C-means method for clustering micoarray data. Bioinformatics **19** (2003) 973–980
4. Filkov, V., Skiena, S., Zhi, J.: Analysis techniques for microarray time-series data. Journal of Computational Biology **9** (2002) 317–330
5. Möller-Levet, C.S., Cho, K.H., Wolkenhauer, O.: Microarray data clustering based on temporal variation: FCV with TSD preclustering. Applied Bioinformatics **2** (2003) 35–45
6. Park, J., Sandberg, I.: Approximation and radial basis function networks. Neural Computing **5** (1993) 305–316
7. Bar-Joseph, Z., Gerber, G., Gifford, D.K., Jaakkola, T.S., Simon, I.: A new approach to analyzing gene expression time series data. In: Proceedings of RECOMB, Washington DC, USA (2002) 39–48
8. Luan, Y., Li, H.: Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics **19** (2003) 474–482
9. Chen, S., Cowan, C.F.N., Grant, P.M.: Orthogonal least squares learning algorithm for radial basis function networks. IEEE Transactions on Neural Networks **2** (1991) 302–309
10. Möller-Levet, C.S., Yin, H., Cho, K.H., Wolkenhauer, O.: Modelling gene expression time-series with radial basis function neural networks. In: Proceeding of the International Joint Conference on Neural Networks (IJCNN'2004). (to appear)
11. Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
12. Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., Boesiger, P.: A new correlation-based fuzzy logic clustering algorithm for fMRI. Magnetic Resounance Medicine **40** (1998) 249–260

13. Pakhira, M.K., Bandyopedhyay, S., Maulik, U.: Validity index for crisp and fuzzy clusters. Pattern recognition **37** (2003) 487–501
14. Yager, R.R., Filev, D.P.: Approximate Clustering Via the Mouintain Method. IEE Transactions on systems, man, and cybernetics **24** (1994) 1279–1284
15. Luan, H., Li, H.: Model-based methods for identifying periodically expressed genes based on time course microrray gene expression data. Bioinformatics **20** (2004) 332–339
16. Klevecz, R.R., Dowse, H.B.: Tuning in the transcriptome: basins of attraction in th yeast cell cycle. Cell Prolif **33** (2000) 209–218