Modelling Gene Expression Time-Series with Radial Basis Function Neural Networks

Carla S. Möller-Levet  
Dept of Electrical Engineering and Electronics,  
University of Manchester  
Institute of Science and Technology,  
Manchester M60 1QD, U.K.  
E-mail: c.moller-levet@postgrad.umist.ac.uk

Kwang-Hyun Cho  
School of Electrical Engineering, University of Ulsan,  
Ulsan 680-749, South Korea  
E-mail: ckh@mail.ulsan.ac.kr

Hujun Yin  
Dept of Electrical Engineering and Electronics,  
University of Manchester  
Institute of Science and Technology,  
Manchester M60 1QD, U.K.  
E-mail: h.yin@umist.ac.uk

Olaf Wolkenhauer  
Dept of Computer Science, University of Rostock,  
18051 Rostock, Germany  
E-mail: wolkenhauer@informatik.uni-rostock.de

Abstract—Gene expression time-series are discrete, noisy, short and usually unevenly sampled. Most existing methods used to compare expression profiles operate directly on the time points. While modelling the profiles can lead to more generalised, smooth characterisation of gene expressions. In this paper a Radial Basis Function neural network is employed to model gene expression time-series. The Orthogonal Least Square method, used for selection of centres, is further combined with a width optimisation scheme. The experiments on a number of expression datasets have shown the advantages of the approach in terms of generalisation and approximation. The results on known datasets have indeed coincided with biological interpretations.

I. INTRODUCTION

Microarrays revolutionise the traditional way of one gene per experiment in molecular biology [1], [2]. With microarray experiments it is possible to measure simultaneously and over time the activity levels of thousands of genes. Gene expression time-series are generally noisy, short and usually unevenly sampled. To eliminate these undesirable characteristics, the modelling of gene expression profiles with the Radial Basis Function (RBF) neural networks is proposed.

Other modelling techniques have been proposed recently, e.g., in [3], [4], gene expression time-series are modelled using mixed-effects models within a mixture model based clustering. Equally spaced cubic spline bases 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 necessarily clustering related.

The RBF is an important and popular model of feedforward neural networks [5], for which fast, linear learning algorithms exist. A neural network is an attempt to simulate biological nervous systems, it is formed by multiple layers of simple processing elements called neurons [6], [7]. Each neuron is linked to its neighbors with varying degrees of connectivity, stored as weight-values. The progressive adaptation of the connection weights of a neural network causes the network to learn the solution to a problem.

Feedforward neural networks are a class of universal approximators [8]. A trained multilayer feedforward network with at least one hidden layer and sufficient number of hidden neurons is capable of approximating any measurable function within an arbitrary accuracy. 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 limited by model assumptions and linearity, as well as the noise, irregular sampling and shortness of the time-series. Once the profiles are modelled, noise reduction can be achieved by smoothing techniques applied to the functions.

II. OPTIMISATION OF RADIAL BASIS FUNCTIONS

Radial basis networks have a single hidden layer where the nodes are Gaussian kernels, and an output linear layer, as illustrated in Figure 1.

The radial basis function can then be generalised to produce a continuous relationship between input and output:

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

where $x$ is the input vector, $\phi(\cdot)$ is a Gaussian function, $\| \cdot \|$ 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 number of centres. The Gaussian kernels $\phi(\cdot)$ are defined as:

$$\phi(\|c_i - x\|) = e^{-\frac{\|c_i - x\|^2}{\sigma^2}}$$  \hspace{1cm} (2)$$
where $\sigma$ is the width factor.

The problem of RBF approximation is to find appropriate centres, 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. The weights can be determined using the linear least squares method. There are several algorithms available in the literature for calculating the centres [9], [10] and widths [11] of the basis functions. The centres are often chosen to be a subset of the data by clustering techniques or random selections. The width is usually set equally for all nodes to the average distance between centres. The expectation maximisation (EM) algorithm could also be used for the estimation of centres and widths, particularly in probability estimation problems [12]. The weights of the linear layer are usually calculated using singular value decomposition (SVD) [13]. Other methods to optimise the parameters of the RBF include the gradient descent algorithm [14] and evolutionary strategies [15].

We adopt the orthogonal least squares (OLS) learning algorithm [16] for training the RBF 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. It is not necessary to use the same number of centres as number of time points. However, this method considers radial basis with equal widths. Fixing the width is not evident, the widths determine the degree of overlapping between the Gaussian kernels which compromise the locality and smoothness of the function. The optimal width value depends on the function to approximate, the dimension of the input set, as well as on the data distribution [11]. Moreover, fixing a single width for all the kernels is inadequate when the data is not evenly distributed (i.e., unevenly sampled). In order to improve the approximation, we complemented the OLS learning algorithm with a 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.

### A. OLS learning algorithm

Considering the RBF network as a linear regression model:

$$d = P\theta + E$$

where $d = \{d(1)\cdots d(N)\}^T$ is the desired output, $P = [p_1 \cdots p_M]$, $p_i = [p_i(1) \cdots p_i(N)]^T$, are the regressors with $1 \leq i \leq M$, $\theta = [\theta_1 \cdots \theta_M]^T$ are the parameters and $E = [e(1) \cdots e(N)]^T$ is the error signal assumed to be uncorrelated with the regressors. The solution $\hat{\theta}$ satisfies the condition that $P\hat{\theta}$ be the projection of $d$ onto the space spanned by the basis vectors formed by the regressor vectors $p_i$. In order to obtain the individual contribution of each individual regressor to the total output, the set of $p_i$ is transformed into a set of orthogonal basis vectors. For this purpose, the regression matrix $P$ can be expressed as the product of a matrix $W$ with orthogonal columns $w_i$ and a matrix $A$ which is upper triangular, such that $P = WA$. The space spanned by the vectors $w_i$ is the same space spanned by the set of $p_i$, and (3) can be rewritten as:

$$d = Wg + E$$

The orthogonal least square solution $\hat{g}$ is:

$$\hat{g} = w_i^T d/(w_i^T w_i), \quad 1 \leq i \leq M. \tag{5}$$

The Gram-Schmidt is used to derive the triangular system $A\hat{\theta} = \hat{g}$ and solve for the LS estimate $\hat{\theta}$. The regressor selection is based in an error reduction ratio due to $w_i$ defined as:

$$[error]_i = g_i^2 w_i^T w_i/(d^T d), \quad 1 \leq i \leq M. \tag{6}$$

The selection of centres finishes when the error reaches a chosen tolerance, allowing the model to have fewer centres than time points.

### B. Search for an optimal width

In order to improve the approximation, especially for unevenly sampled time points, a search for the optimal width for each of the candidate centres is proposed. Extra time points are chosen halfway between the original points and added to the profiles in order to increase the linearity between time points. To aid the approximation of a sharp change, additional time points on both sides of the sharp turning point are used to form small angles, as illustrated in Figure 2.

In [17] the width factors $\sigma$ are computed by nearest neighbour heuristic, where the width of the radial basis $j$ centred at $c_j$ is set to the Euclidean distance between $c_j$ and its nearest neighbour $c_i$ multiplied by an overlap constant $r$, such that:

$$\sigma_j = r \min(\|c_j - c_i\|) \tag{7}$$

This method takes the variations of the distribution of the data into account. Using this approach, the search for the width is conducted within a fixed interval $[\sigma_{\text{min}}, \sigma_{\text{max}}]$, where:

$$\sigma_{\text{min}} = \min_j \min(\|c_j - c_i\|),$$

$$\sigma_{\text{max}} = \max_j \min(\|c_j - c_i\|). \tag{8}$$
Fig. 2. Extra time points (marked by squares) are added to the profiles in order to increase the linearity between time points, and to aid the approximation of sharp turning points, additional points (marked by asterisks) are used.

The optimal width minimises the mean square error (MSE) of a piecewise linear fit of a segment of the series and the RBF approximation. The segment is formed by the $h$ (typically $h = 4$) nearest centres. In general, performing the search using values around $r_{\text{min}} = 0.7$ and $r_{\text{max}} = 1.3$ lead to satisfactory results.

The search for the optimal width is done for each of the candidate centres before the selection of a further centre. Then, the regression matrix $P$ is recalculated with the optimal widths for the candidate centres and a new centre is selected.

III. CASE STUDIES: YEAST DATASETS

A. Evenly sampled data: yeast cell cycle dataset

In [18], 800 cell-cycle-regulated genes of the Yeast Saccharomyces cerevisiae were identified by microarray hybridisation\(^1\). Yeast cultures were synchronised by three independent methods: $\alpha$ factor arrest, elutriation and arrest of a cdc15 temperature sensitive mutant. We utilised the temporal expression of the yeast culture synchronised by $\alpha$ factor arrest to illustrate the method. The yeast cells were sampled every 7 minutes for 119 minutes, producing 18 evenly distributed sampling points. Among the 800 genes 511 have no missing values and are available from the web site. In [18], the genes are grouped based on their periodicity and classified according to different cell-cycle phases.

The 511 profiles have been modelled using the proposed RBF networks described in the previous section. Figure 3(a) presents the histogram of MSE for the RBF approximation, Figure 3(b), (c) and (d) show the profiles with the minimum, median and maximum MSE respectively.

Out of the 511 genes, 71 genes have been identified to be cell-cycle regulated by traditional biological methods. The RBF models of these gene profiles were sphere normalised and then clustered using fuzzy c-means. Table I presents the distribution of the genes in each cluster among the cell-cycle phases identified\(^2\). In the six clusters the genes belong

\(^1\)The dataset is available from http://cellcycle-www.stanford.edu
\(^2\)In Table I, G1a corresponds to G1 SCB regulated genes and G1b corresponds to G1 MCB regulated genes.
mainly to one phase and the two adjacent phases. Cluster I is essentially formed by M/G1 and some genes from phase G2/M. Cluster II is mainly formed by M/G1 genes, and has some genes belonging to the G1a and S/G2 phases. Cluster III is formed by G1b genes and has one gene in phases M/G1 and G1a. Cluster IV is mostly formed by G1b genes and has one gene in phases G1a and G2/M. Cluster V is formed by genes expressed in the S phase and has genes from the adjacent phases, G1a, G1b and S/G2. Cluster VI is formed by genes in the S/G2 and G2/M phases. The clustering results indeed show biological relevance, indicating that RBF network is a useful technique for gene expression time-series modelling.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>M/G1</th>
<th>G1a</th>
<th>G1b</th>
<th>S</th>
<th>S/G2</th>
<th>G2/M</th>
</tr>
</thead>
<tbody>
<tr>
<td>I(10)</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>II(10)</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>III(21)</td>
<td>1</td>
<td>1</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IV(10)</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>V(13)</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>VI(7)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

B. Unevenly sampled data: budding yeast dataset

In [19], DNA microarrays containing 97% of the known or predicted genes of *Saccharomyces cerevisiae* were used to explore the temporal program of gene expression during meiosis and spore formation. Changes in the concentrations of mRNA transcripts from each gene were measured at seven uneven time intervals. The authors [19] distinguished seven temporal patterns of induced transcription. They chose a set of representative genes from each of the seven expression patterns, and the average for each set was calculated to create seven model profiles. We used this set of chosen genes in this experiment. The profiles were modelled using the proposed RBFs. Figure 4(a) presents the histogram of MSE for the RBF approximation, Figure 4(b), (c) and (d) show the profiles with the minimum, median and maximum MSE respectively.

The resulting RBF models were differentiated (for shape characterisation), sphere normalised and clustered using fuzzy c-means. The clustering results are shown in Figure 5. Cluster IV is formed by genes corresponding to the Metabolic and Early profiles, which are characterised by a sudden initial induction. The profiles forming Cluster I have a strong induction from time point 2hr. to 7hr. The profiles forming Cluster II have a negative induction at time 0.5hr., followed by an induction at time 2hr. and at 7hr. Cluster III is characterised by genes having a continuous induction until time 7hr. Clusters I, II and II correspond to Early II, Early-Middle and Middle

---

*Fig. 4.* Summary of the RBF approximation of the dataset in [19]. In figures (b), (c) and (d) the horizontal axis represents time (hrs) and the vertical axis denotes the expression level.

---

5The dataset is available from http://cmgm.stanford.edu/pbrown/sporulation
Fig. 5. Clustering results. In every figure the horizontal axis denotes time [0, 0.5, 2, 5, 7, 9 and 11.5 hrs] and the vertical axis denotes the expression level.

IV. CONCLUSIONS

In this paper gene expression time-series are modelled by using Radial Basis Function neural networks. The Orthogonal Least Square method is used for the selection of centres. A search for the optimal width for each of the candidate centres is further incorporated. Two benchmark datasets were used to test the proposed approach. The biological relevance of the clustering results on modelled profiles had shown the advantages of the approach in terms of generalisation (eliminating time points) and approximation in both even and unevenly sampled gene expression time-series.

ACKNOWLEDGMENT

C. S. Moller-Levet was supported in part by grants from ABB Ltd. U.K., an Overseas Research Studentship (ORS) award and Consejo Nacional de Ciencia y Tecnologia (CONACYT); K.-H. Cho acknowledges the support received by a grant from the Korea Ministry of Science and Technology (Korean Systems Biology Research Grant, M10309000006-03B5000 -00211) and also by grant No. (R05-2004-000-10549-0) from the Korea Ministry of Science and Technology; and O.Wolkenhauer was supported by a grant from the UK Department for the Environment, Food and Rural Affairs (DEFRA).

REFERENCES


