Analysing periodic phenomena by circular PCA

Matthias Scholz

Competence Centre for Functional Genomics (CC-FG),
Institute for Microbiology, Ernst-Moritz-Arndt-University Greifswald, Germany

matthias.scholz@functional-genomics.de
WWW home page: http://www.functional-genomics.de

S. Hochreiter and R. Wagner (Eds.): Proceedings of the Conference on Bioinformatics Research and Development BIRD'07, LNBI Vol. 4414, pages 38–47, 2007.
©Springer-Verlag Berlin Heidelberg 2007 (author-created version)

Abstract. Experimental time courses often reveal a nonlinear behaviour. Analysing these nonlinearities is even more challenging when the observed phenomenon is cyclic or oscillatory. This means, in general, that the data describe a circular trajectory which is caused by periodic gene regulation.
Nonlinear PCA (NLPCA) is used to approximate this trajectory by a curve referred to as nonlinear component. Which, in order to analyse cyclic phenomena, must be a closed curve hence a circular component. Here, a neural network with circular units is used to generate circular components.
This circular PCA is applied to gene expression data of a time course of the intraerythrocytic developmental cycle (IDC) of the malaria parasite *Plasmodium falciparum.* As a result, circular PCA provides a model which describes continuously the transcriptional variation throughout the IDC. Such a computational model can then be used to comprehensively analyse the molecular behaviour over time including the identification of relevant genes at any chosen time point.

Key words: gene expression, nonlinear PCA, neural networks, nonlinear dimensionality reduction, *Plasmodium falciparum*

1 Introduction

Many phenomena in biology proceed in a cycle. These include circadian rhythms, the cell cycle, and other regulatory or developmental processes such as the cycle of repetitive infection and persistence of malaria parasites in red blood cells which is considered here. Due to an individual behaviour of molecules over time, the resulting data structure becomes nonlinear as shown, for example, in Scholz et al. (2005) for a cold stress adaptation of the model plant *Arabidopsis thaliana.* In this context, nonlinearity means that the trajectory of the data describes a curve over time. For periodic processes, this curve is closed and hence cannot be well described by a standard (open) nonlinear component.

Therefore, the objective is to visualise and analyse the potential circular structure of molecular data by a nonlinear principal component analysis which is able to generate circular components.

Nonlinear principal component analysis (NLPCA) is generally seen as a nonlinear generalisation of standard linear *principal component analysis* (PCA) (Jolliffe, 1986; Diamantaras and Kung, 1996). The principal components are generalised from straight lines to curves. Here, we focus on a neural network based nonlinear PCA, the *auto-associative neural network* (Kramer, 1991; DeMers and Cottrell, 1993; Hecht-Nielsen, 1995; Scholz and Vigário, 2002).

To generate circular components, Kirby and Miranda (1996) constrained network units to work in a circular manner. In the fields of atmospheric and oceanic sciences, this circular PCA is applied to oscillatory geophysical phenomena (Hsieh, 2004). Other applications are in the field of robotics to analyse and control periodic movements (MacDorman et al., 2004). Here, we demonstrate the potential of circular PCA to biomedical applications. The biological process, analysed here, is the intraerythrocytic developmental cycle (IDC) of *Plasmodium falciparum.*

P. falciparum is the most pathogenic species of the *Plasmodium* parasite, which causes malaria. The three major stages of *Plasmodium* development take place in the mosquito and upon infection of humans in liver and red blood cells. The infection of red blood cells (erythrocytes) recurs with periodicity of around 48 hours. This intraerythrocytic developmental cycle (IDC) of *P. falciparum* is responsible for the clinical symptoms of the malaria disease. A better understanding of the IDC may provide opportunities to identify potential molecular targets for anti-malarial drug and vaccine development.

2 NLPCA - nonlinear PCA

The nonlinear PCA (NLPCA), proposed by Kramer (1991), is based on a multi-layer perceptron (MLP) with an auto-associative topology, also known as an autoencoder, replicator network, bottleneck or sandglass type network. Comprehensive introductions to multi-layer perceptrons can be found in Bishop (1995) and Haykin (1998).

The auto-associative network performs the identity mapping. The output $\hat{x}$ is enforced to equal the input x with high accuracy. This is achieved by minimising the square error $\| x - \hat{x} \|^2$. This is no trivial task, as there is a 'bottleneck' in the middle, a layer of fewer nodes than at the input or output, where the data have to be projected or compressed into a lower dimensional space Z.

The network can be considered as two parts: the first part represents the extraction function $\Phi_{extr} : \mathcal{X} \rightarrow \mathcal{Z}$, whereas the second part represents the inverse function, the generation or reconstruction function $\Phi_{gen} : \mathcal{Z} \rightarrow \hat{\mathcal{X}}$. A hidden layer in each part enables the network to perform nonlinear mapping functions.

In the following we describe the applied network topology by the notation [l_1-l_2-...- l_S] where l_s is the number of units in layer s. For example, [3-4-1-4-3] specifies a network with three units in the input and output layer, four units in both hidden layers, and one unit in the component layer, as illustrated in Figure 1.

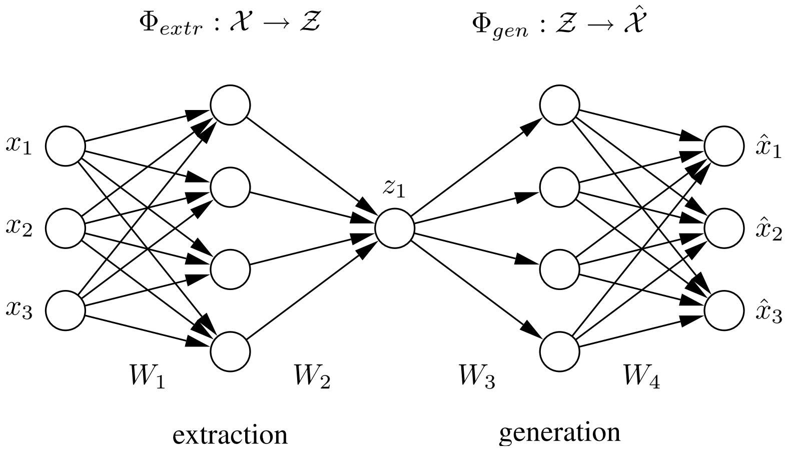

Figure 1: **Standard auto-associative neural network.** The network output $\hat{x}$ is required to be equal to the input x. Illustrated is a [3-4-1-4-3] network architecture. Biases have been omitted for clarity. Three-dimensional samples x are compressed (projected) to one component value z in the middle by the extraction part. The inverse generation part reconstructs $\hat{x}$ from z. The sample $\hat{x}$ is usually a noise-reduced representation of x. The second and fourth hidden layer, with four nonlinear units each, enable the network to perform nonlinear mappings. The network can be extended to extract more than one component by using additional units in the component layer in the middle.

2.1 Circular PCA

Kirby and Miranda (1996) introduced a circular unit at the component layer that describes a potential circular data structure by a closed curve. As illustrated in Figure 2, a circular unit is a pair of networks units p and q whose output values z_p and z_q are constrained to lie on a unit circle

$$z_p^2 + z_q^2 = 1 \tag{1}$$

Thus, the values of both units can be described by a single angular variable θ.

$$z_p = \cos(\theta) \qquad \text{and} \qquad z_q = \sin(\theta) \tag{2}$$

The *forward propagation* through the network is as follows: First, equivalent to standard units, both units are weighted sums of their inputs z_m given by the values of all units m in the previous layer.

$$a_p = \sum_m w_{pm} z_m \qquad \text{and} \qquad a_q = \sum_m w_{qm} z_m \tag{3}$$

The weights w_{pm} and w_{qm} are of matrix W_2. Biases are not explicitly considered, however, they can be included by introducing an extra input with activation set to one.

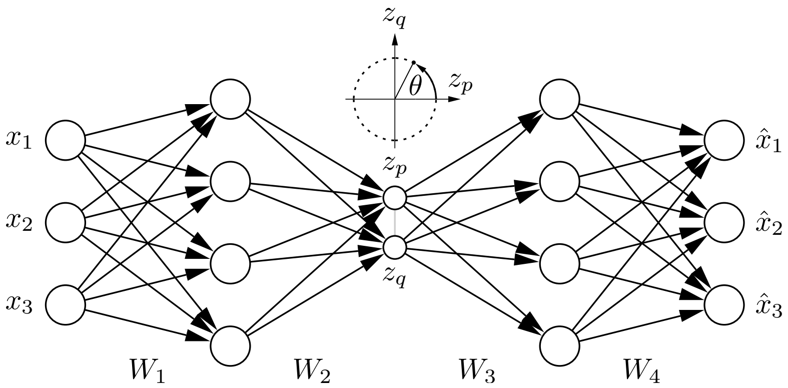

Figure 2: **Circular PCA network.** To obtain circular components, the auto-associative neural network contains a circular unit pair (p, q) in the component layer. The output values z_p and z_q are constrained to lie on a unit circle and hence can be represented by a single angular variable θ.

The sums a_p and a_q are then corrected by the radial value

$$r = \sqrt{a_p^2 + a_q^2} \tag{4}$$

to obtain circularly constraint unit outputs z_p and z_q

$$z_p = \frac{a_p}{r} \qquad \text{and} \qquad z_q = \frac{a_q}{r} \tag{5}$$

For *backward propagation*, we need the derivatives of the error function

$$E = \frac{1}{2} \sum_n^N \sum_i^d \left[x_i^n - \hat{x}_i^n\right]^2 \tag{6}$$

with respect to all network weights w. The dimensionality d of the data is given by the number of observed variables, N is the number of samples.
To simplify matters, we first consider the error e of a single sample x, $e = \frac{1}{2} \sum_i^d \left[x_i - \hat{x}_i\right]^2$ with $x = (x_1, \ldots, x_d)^T$. The resulting derivatives can then be extended with respect to the total error E given by the sum over all n samples, $E = \sum_n e^n$.
While the derivatives of weights of matrices W_1, W_3, and W_4 are obtained by standard back-propagation, the derivatives of the weights w_{pm} and w_{qm} of matrix W_2 which connect units m of the second layer with the units p and q are obtained as follows: We first need the partial derivatives of e with respect to z_p and z_q:

$$\tilde{\sigma}_p = \frac{\partial e}{\partial z_p} = \sum_j w_{jp} \sigma_j \qquad \text{and} \qquad \tilde{\sigma}_q = \frac{\partial e}{\partial z_q} = \sum_j w_{jq} \sigma_j \tag{7}$$

where σ_j are the partial derivatives $\frac{\partial e}{\partial a_j}$ of units j in the fourth layer.
The required partial derivatives of e in respect to a_p and a_q of the circular unit pair are

$$\sigma_p = \frac{\partial e}{\partial a_p} = (\tilde{\sigma}_p z_q - \tilde{\sigma}_q z_p)\frac{z_q}{r^3} \qquad \text{and} \qquad \sigma_q = \frac{\partial e}{\partial a_q} = (\tilde{\sigma}_q z_p - \tilde{\sigma}_p z_q)\frac{z_p}{r^3} \tag{8}$$

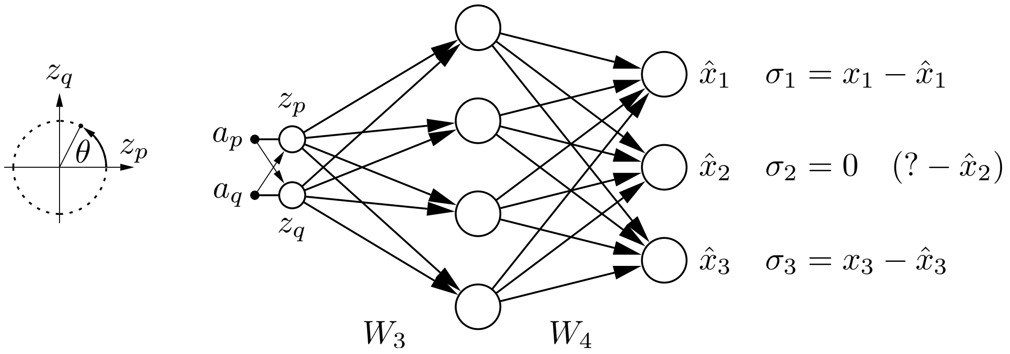
z_q
θ
z_p
a_p
a_q
z_p
z_q
W_3
W_4
$\hat{x}_1$ $\sigma_1 = x_1 - \hat{x}_1$
$\hat{x}_2$ $\sigma_2 = 0 \quad (? - \hat{x}_2)$
$\hat{x}_3$ $\sigma_3 = x_3 - \hat{x}_3$

Figure 3: **Inverse circular PCA network.** As inverse model, only the second part of the auto-associative neural network (Figure 2) is used. Now, the values a_p and a_q are unknown inputs and have to be estimated together with all weights w of matrices W_3 and W_4. This is done by propagating the partial errors σ_i back to the input (component) layer. Beside a higher efficiency, the main advantage is that the inverse model can be applied to incomplete data. If one value x_i of a sample vector x is missing, the corresponding partial error σ_i is set to zero, thereby ignoring the missing value but still back-propagating all others.

The final back-propagation formulas for all n samples are

$$\frac{\partial E}{\partial w_{pm}} = \sum_n \sigma_p^n z_m^n \qquad \text{and} \qquad \frac{\partial E}{\partial w_{qm}} = \sum_n \sigma_q^n z_m^n \tag{9}$$

2.2 Inverse NLPCA model

In this work, NLPCA is applied as an inverse model (Scholz et al., 2005). Only the second part, the generation or reconstruction part, of the auto-associative neural network is modelled, see Figure 3. The major advantage is that NLPCA can be applied to incomplete data. Another advantage is a higher efficiency since only half of the network weights have to be estimated.

Optimising the second part as inverse model means that the component layer becomes the input layer. Thus, in circular PCA as inverse model we have to find suitable values for all network weights as well as for a_p and a_q as input. Hence, the error function E depends on both the weights w and the component layer inputs a_p and a_q

$$E(w, a_p, a_q) = \frac{1}{2} \sum_n^N \sum_i^d \left[x_i^n - \hat{x}_i^n(w, a_p, a_q)\right]^2 \tag{10}$$

The required partial derivatives of E with respect to the weights w of matrix W_3 and W_4 can be obtained by standard back-propagation, see Scholz et al. (2005), while the derivatives with respect to a_p and a_q are given by equation (8).

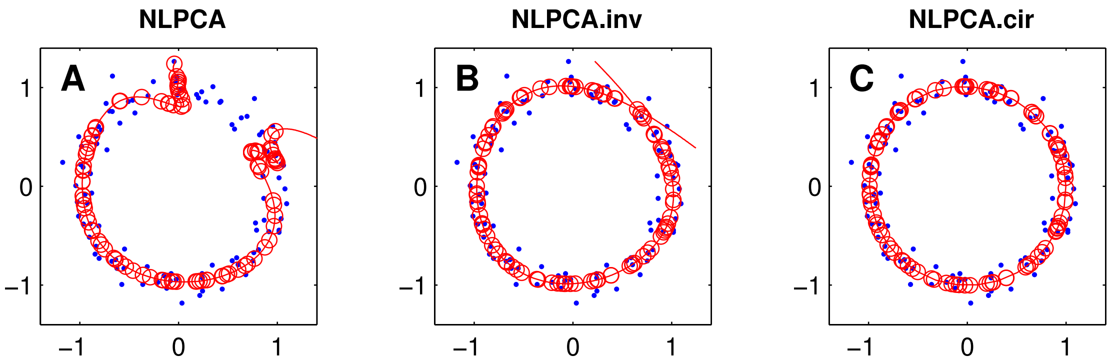

Figure 4: **Nonlinear PCA (NLPCA).** Shown are results of three variants of NLPCA applied to a two-dimensional artificial data set of a noisy circle. **(A)** The standard NLPCA cannot describe a circular structure completely. There is always a gap. **(B)** The inverse NLPCA can provide self-intersecting components and hence approximates the circular data structure already quite well but the circular PCA **(C)** is most suitable since it is able to approximates the data structure continuously by a closed curve.

2.3 Artificial data

The performance of NLPCA is illustrated in Figure 4 for the three described variants: the standard auto-associative network (NLPCA), the inverse model with standard units (NLPCA.inv) and with circular units (NLPCA.cir). NLPCA is applied to data lying on a unit circle and disturbed by Gaussian noise with standard deviation 0.1. The standard auto-associative network cannot describe a circular structure completely by a nonlinear component due to the problem to map at least one point on the circle onto two different component values. This problem does not occur in inverse NLPCA since it is only a mapping from component values to the data. However, the resulting component is an intersecting circular loop with open ends. Thus, a closed curve solution as provided by circular PCA would be more appropriate to describe the circular structure of the data.

2.4 Experimental data

Circular PCA is used to analyse the transcriptome of the intraerythrocytic developmental cycle (IDC) of the malaria parasite *Plasmodium falciparum* (Bozdech et al., 2003), available at `http://malaria.ucsf.edu/`. The 48-hour IDC is observed by a sampling time of one hour thereby providing a series of time points 1, 2, ..., 48. Since two time points, 23 and 29, are missing, the total number of expression profiles (samples) is 46. Each gene is represented by one or more oligonucleotides on the microarray. The samples of individual time points (Cy5) were hybridised against a reference pool (Cy3). The $\log_2$(Cy5/Cy3) ratio is used in our analysis. Due to the sometimes large number of missing data in the total set of 7,091 oligonucleotides, we removed all oligonucleotides

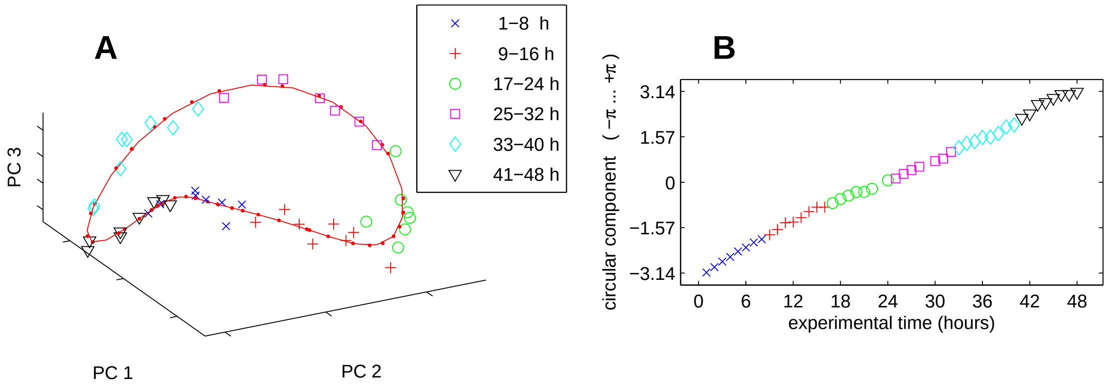

Figure 5: **Circular PCA. (A)** The data describe a circular structure which is approximated by a closed curve (the circular component). The circular component is one single curve in the 5,800 dimensional data space. Visualised is the reduced three dimensional subspace given by the first three components of standard (linear) PCA.
(B) The circular component (corrected by an angular shift) is plotted against the original experimental time. It shows that the main curvature, given by the circular component, explains the trajectory of the IDC over 48 hours.

of more than 1/3 missing time observations (more than 15 missing time points). The considered reduced data set contains the $\log_2$ ratios of hybridisations of 5,800 oligonucleotides at 46 time points.

Identifying the optimal curve (the circular component) in the very high-dimensional data space of 5,800 variables is difficult or even intractable with a number of 46 data points. Therefore, the 5,800 variables are linearly reduced to 12 principal components, each of which is a linear combination of all oligonucleotides. To handle missing data, a PCA algorithm, based on a linear neural network working in inverse mode (Scholz et al., 2005), is used. Alternatively, *probabilistic PCA* (PPCA) (`http://lear.inrialpes.fr/~verbeek/software`) by Verbeek et al. (2002), based on Roweis et al. (2002), can be used as PCA algorithm for missing data.

Circular PCA is applied to the reduced data set of 12 linear components. It describes a closed curve explaining the circular structure of the data, as shown in Figure 5 and 6. To achieve circular PCA, a network of a [2-5-12] architecture is used, where the two units in the first layer are the circularly constrained unit pair (p, q). Using the inverse 12 eigenvectors the curve can be mapped back into the 5,800-dimensional original data space. The circular component represents the 48 hour time course of the IDC observation, as shown in Figure 5B.

Thus, circular PCA provides a model of the IDC, which gives us to any chosen time point, including interpolated time points, the corresponding gene expression values. The neural network model is given by a function $\hat{x} = \Phi_{gen}(\theta)$ which maps any time point, represented by a angular value θ onto a 5,800-dimensional vector $\hat{x} = (\hat{x}_1, ..., \hat{x}_{5800})^T$ representing the response of the original variables. Thus, circular PCA provides ap-

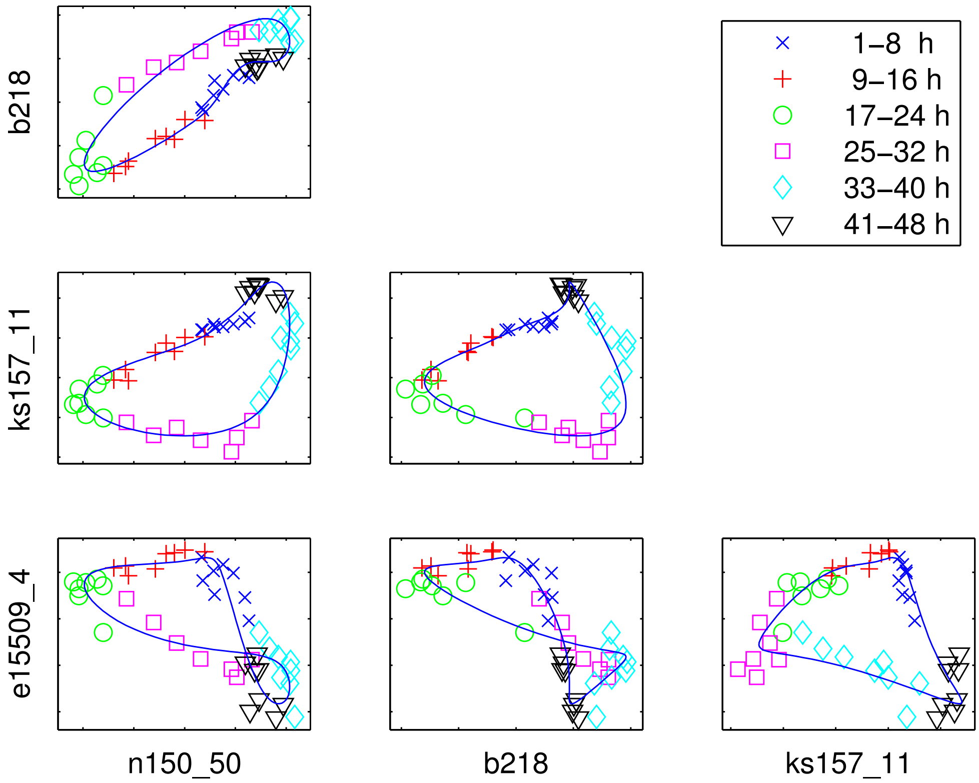

Figure 6: Pair-wise scatter plot of four selected oligonucleotides of importance at 12, 24, 36, and 48 hours respectively, see also Table 1. The curve represents the circular component which approximates the trajectory of the 48 hour IDC.

proximated response curves of all oligonucleotides, as shown in Figure 7 for the top 50 oligonucleotides of genes of highest response (highest relative change at their expression level).

In standard PCA we can present the variables that are most important to a specific component by a rank order given by the absolute values of the corresponding eigenvector, sometimes termed loadings or weights. As the components are curves in nonlinear (circular) PCA, no global ranking is possible. The rank order is different for different positions on the curved component, meaning that the rank order depends on time. The rank order for each individual time point is given by the values of the tangent vector $v = \frac{d\hat{x}}{d\theta}$ on the curve at a specific time θ. To compare different times, we use l_2-normalised tangents $\hat{v}_i = v_i / \sqrt{\sum_i |v_i|^2}$ such that $\sum_i (\hat{v}_i)^2 = 1$. Large values $\hat{v}_i$ point to genes of high changes on their expression ratios and hence may have an importance at the considered time point. A list of 10 most important genes at 12 hours and 36 hours is exemplarily given in Table 1.

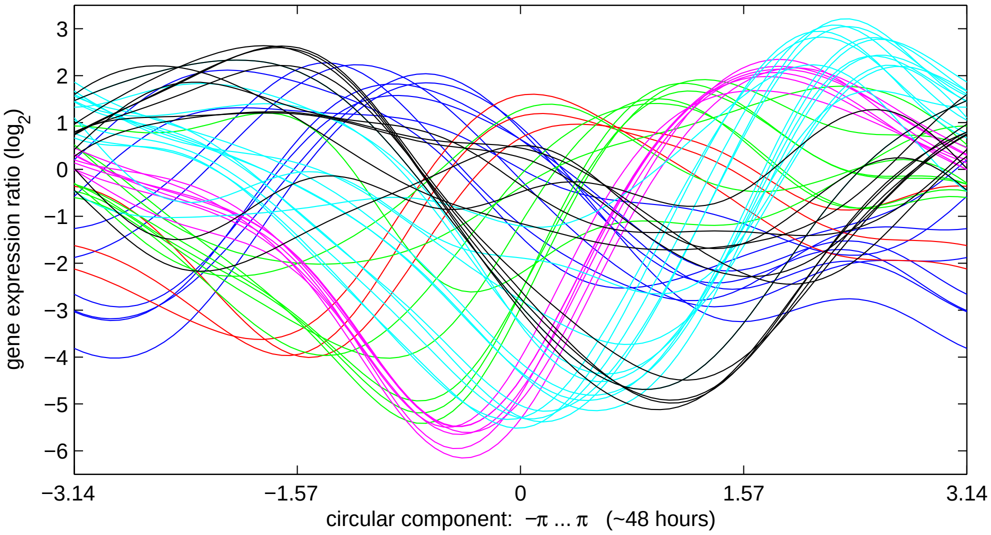

Figure 7: **Gene expression curves.** Plotted are shapes of oligonucleotide response curves over the 48-hour IDC time course. Shown are the top 50 oligonucleotides of genes of highest response at any time. Nearly all of them show a period of 48 hours: one up- and down-regulation within the 48-hour time course. However, the time of activation is differently for individual genes.

3 Conclusions

Circular PCA as special case of nonlinear PCA (NLPCA) was applied to gene expression data of the intraerythrocytic developmental cycle (IDC) of the malaria parasite *Plasmodium falciparum*. The data describe a circular structure which was found to be caused by the cyclic (nonlinear) behaviour of gene regulation.

The extracted circular component represents the trajectory of the IDC. Thus, circular PCA provides a noise reduced model of gene responses continuously over the full time course. This computational model can then be used for analysing the molecular behaviour over time in order to get a better understanding of the IDC.

With the increasing number of time experiments, nonlinear PCA may become more and more important in the field of molecular biology. This includes the analysis of both: non-periodic phenomena by standard NLPCA and periodic phenomena by the circular variant.

Acknowledgement

This research is funded by the German Federal Ministry of Education and Research (BMBF) within the programme of *Centres for Innovation Competence* of the BMBF initiative *Entrepreneurial Regions* (Project No. ZIK 011).

12 hours			36 hours		
$\hat{v}_i$	**Oligo ID**	**PlasmoDB ID**	$\hat{v}_i$	**Oligo ID**	**PlasmoDB ID**
-0.06	i6851_1	—	0.06	ks157_11	PF11_0509
-0.06	n150_50	PF14_0102	0.06	a10325_32	PFA0110w
-0.05	e24991_1	PFE0080c	0.06	a10325_30j	—
-0.05	opfg0013	—	0.06	b70	PFB0120w
-0.05	c76	PFC0120w	0.06	a10325_30	PFA0110w
-0.05	n140_2	PF14_0495	0.06	i14975_1	PF07_0128
-0.05	opff72487	—	0.06	f739_1	PF07_0128
-0.05	kn5587_2	MAL7P1.119	0.06	opfl0045	PFL1945c
0.05	f24156_1	PFI1785w	0.05	a10325_29	PFA0110w
0.05	d44388_1	PF10_0009	0.05	ks75_18	PF11_0038
...	...	...	...	...	...

Table 1: **Candidate genes.** At specific times, exemplarily shown for 12 and 36 hours, the most important genes can be provided. Listed are the identified oligonucleotides and, if available, the corresponding PlasmoDB gene identifier of the *Plasmodium* genome resource *PlasmoDB.org* (Kissinger et al., 2002; Gardner et al., 2002). Note that a single gene may be represented by more than one oligonucleotide.

References

Bishop, C. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

Bozdech, Z., Llinas, M., Pulliam, B., Wong, E., Zhu, J., DeRisi, J. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum. PLoS Biology*, 1(1):E5, 2003.

DeMers, D., Cottrell, G.W. Nonlinear dimensionality reduction. In Hanson, D., Cowan, J., Giles, L., eds., *Advances in Neural Information Processing Systems 5*, pages 580–587, San Mateo, CA, 1993. Morgan Kaufmann.

Diamantaras, K., Kung, S. *Principal Component Neural Networks.* Wiley, New York, 1996.

Gardner, M., Hall, N., Fung, et al., E. Genome sequence of the human malaria parasite *plasmodium falciparum. Nature*, 419(6906):498–511, 2002.

Haykin, S. *Neural Networks - A Comprehensive Foundation.* Prentice Hall, 2nd edition, 1998.

Hecht-Nielsen, R. Replicator neural networks for universal optimal source coding. *Science*, 269:1860–1863, 1995.

Hsieh, W.W. Nonlinear multivariate and time series analysis by neural network methods. *Reviews of Geophysics*, 42(1):RG1003.1–RG1003.25, 2004.

Jolliffe, I.T. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

Kirby, M.J., Miranda, R. Circular nodes in neural networks. *Neural Computation*, 8(2): 390–402, 1996.

Kissinger, J., Brunk, B., Crabtree, J., Fraunholz, M., Gajria, et al., B. The plasmodium genome database. *Nature*, 419(6906):490–492, 2002.

Kramer, M.A. Nonlinear principal component analysis using auto-associative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

MacDorman, K., Chalodhorn, R., Asada, M. Periodic nonlinear principal component neural networks for humanoid motion segmentation, generalization, and generation. In *Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR), Cambridge, UK*, pages 537–540, 2004.

Roweis, S.T., Saul, L.K., Hinton, G.E. Global coordination of locally linear models. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds., *Advances in Neural Information Processing Systems 14*, pages 889–896, Cambridge, MA, 2002. MIT Press.

Scholz, M., Kaplan, F., Guy, C., Kopka, J., Selbig, J. Non-linear PCA: a missing data approach. *Bioinformatics*, 21(20):3887–3895, 2005.

Scholz, M., Vigário, R. Nonlinear PCA: a new hierarchical approach. In Verleysen, M., ed., *Proceedings ESANN*, pages 439–444, 2002.

Verbeek, J., Vlassis, N., Kröse, B. Procrustes analysis to coordinate mixtures of probabilistic principal component analyzers. Technical report, Computer Science Institute, University of Amsterdam, The Netherlands, 2002.