FEBS Letters 451 (1999) 142–146 FEBS 21985

Analysis of gene expression data using self-organizing maps

Petri Törönen[a], Mikko Kolehmainen[b], Garry Wong[a], Eero Castrén[a,c,*]

[a] *A.I. Virtanen Institute, University of Kuopio, Box 1627, 70211 Kuopio, Finland*
[b] *Department of Environmental Sciences, University of Kuopio, Box 1627, 70211 Kuopio, Finland*
[c] *Department of Psychiatry, University of Kuopio, Box 1627, 70211 Kuopio, Finland*

Received 3 April 1999

Abstract DNA microarray technologies together with rapidly increasing genomic sequence information is leading to an explosion in available gene expression data. Currently there is a great need for efficient methods to analyze and visualize these massive data sets. A self-organizing map (SOM) is an unsupervised neural network learning algorithm which has been successfully used for the analysis and organization of large data files. We have here applied the SOM algorithm to analyze published data of yeast gene expression and show that SOM is an excellent tool for the analysis and visualization of gene expression profiles.

Key words: Self-organizing map; Sammon's mapping; Gene expression data; Yeast; Cluster analysis

1. Introduction

The rapid development of DNA microarray technology has led to an explosion of gene expression data and much of this information is available in public databases [1]. The yeast genome has been sequenced completely and a draft version of the human genome is predicted to be completed by the spring of 2000. DNA chips containing nearly the entire repertoire of yeast genes have been used to investigate changes in gene expression during a diauxic shift [2], sporulation [3] and the cell cycle [4]. Similar DNA chips with thousands of mammalian genes have been produced and are being used to characterize changes in gene expression patterns in a variety of cell lines and experimental conditions [1,5].

Currently, the major obstacle is the lack of efficient methods to catalogue these data into useful and functionally meaningful groups. Initially, a mainly visual analysis was used to find genes with similar expression patterns. Although visual search has proven successful in grouping genes into functionally relevant classes [2,3], this method is labor intensive, prone to errors and is not well suited for the analysis of very large data sets. Cluster analysis methods have been used to more systematically group related gene expression patterns together [6]. Although these methods can effectively cluster together genes with closely related expression profiles, there is no direct relationship between different branches.

A self-organizing map (SOM, Kohonen's map) is an unsupervised neural network algorithm [7] which has successfully been used to analyze very large data files in various fields, such as process monitoring and visualization [8], exploratory data analysis [9] and simulation of brain-like feature maps [10]. SOM allows easy visualization of complex data and is robust to minor experimental variation.

We have here investigated whether SOM and Sammon's [11] mapping algorithms can be applied to the analysis of gene expression data. We have analyzed a publicly available database on the changes in gene expression during a diauxic shift (shift from anaerobic to aerobic metabolism [2]) in yeast using prototype software (GenePoint)[1] running in a Windows environment. Our results show that SOM rapidly and reliably clusters this gene expression data set into groups that not only show similar gene expression profiles but also contain functionally related genes. Moreover, the algorithm places genes with similar, but not identical profiles in neighboring groups creating a smooth transition of related profiles over the whole matrix.

2. Materials and methods

The SOM is one of the best known unsupervised neural learning algorithms [7]. The goal of the SOM algorithm is to find prototype vectors that represent the input data set and at the same time realize a continuous mapping from input space to a lattice. This lattice consists of a defined number of 'neurons' and may, for example, be a two-dimensional map (Fig. 1a) that is easily visualized. The basic principle behind the SOM algorithm is that the weight vectors of neurons which are first initialized randomly (Fig. 1b), come to represent a number of original measurement vectors during an iterative data input process, as illustrated in Fig. 1c and d.

A variation of the SOM, called tree-structured SOM, was used in this work [12]. The software implementation consists of several SOMs that are organized hierarchically in a pyramid-like fashion in several layers. The number of neurons at a higher level is four times the number of the previous level. However, the visual inspection of the measurement data is directed to one level at the time and the results are comparable to those achieved by 'standard' SOM. We have here used a prototype version of GenePoint software, which is based on tree-structured SOM and runs in a Windows 95/NT/98 environment.

The data that SOM was applied to consist of gene expression profiles of 6400 yeast genes during a diauxic shift (http://cmgm.stanford.edu/pbrown/explore/diauxsearch.html). Ratios between the expression level at the starting point and at seven time points at 2-h intervals were calculated and log-base2 of these seven values were used as the training pattern. At the beginning, each neuron of the SOM was randomly assigned a weight vector with seven variables corresponding to an expression profile of random values within a specified range. During the learning phase, the expression profile of each yeast gene is repeatedly compared with the random profiles in neurons. The weight vectors of the best matching neuron and its four neighbors are moved towards the values of the input vectors such that neurons come to represent a group of similar expression profiles (Fig. 1). While the learning proceeds, the adjustment of weight vectors is diminished. Finally, each gene is placed into a neuron which best describes its expression profile and the average expression profile is displayed on each neuron as a bar graph.

Sammon's mapping is an iterative method based on a gradient

*Corresponding author. Fax: (358) (17) 163 030.
E-mail: eero.castren@uku.fi

[1] For more information about the GenePoint software, see http://www.visipoint.fi

PII: S0014-5793(99)00524-4

search [11]. The aim of the algorithm is to represent points in a n-dimensional space, usually in two dimensions. The algorithm finds the locations in the target space so that as much as possible of the original structure of the measurement vectors in the n-dimensional space is conserved. The numerical calculation is more time consuming than the SOM algorithm, however, which can be a problem with a massive data set. On the other hand, it is able to represent the relative distances of vectors in a measurement space and is thus useful in determining the shape of clusters and the relative distances between them. It is therefore of benefit to combine Sammon with SOM algorithm. Sammon's mapping can be applied to the stage where the SOM algorithm has already achieved a substantial data reduction by replacing the original data vectors with fewer representative prototype vectors.

3. Results

We have analyzed the published expression data of 6400 yeast genes at seven different time points during a diauxic shift using a tree-structured SOM algorithm [12], a modification of the classical Kohonen SOM [7]. The algorithm organizes the data into a two-dimensional matrix by an iterative process based on the relative similarity of the expression pattern of genes. Initial test learning trials using various two-dimensional matrices showed that a 16×16 matrix consisting of 256 neurons produced good separation of different patterns (Fig. 2a). Using this level of separation, the number of genes in individual neurons varied between 10 and 49. The bar graph displayed on each neuron represents the average gene expression profile of genes grouped in this neuron (Fig. 2b). In Fig. 2c, SOM is further modified using Sammon's mapping algorithm [11], where spatial distance correlates with the difference in average expression profile and the circle size with the number of genes within the neuron. The principal direction of change within the data set is represented as an axis in the Sammon map. In the particular example shown in Fig. 2c, the main axis reflects the general increase or decrease in the expression pattern. Neurons with increasing expression profiles are separated on the left side and genes which are suppressed during diauxic shift are located on the right side. Genes not regulated by diauxic shift tend to cluster in the center of the field.

We selected four neurons with different expression patterns for the analysis of their gene content (Table 1), two of which show an increasing pattern (C1 and C2), one a decreasing pattern (C3) and one no change (C4). The first three resemble in their expression profile the groups of genes that were visually selected in the original analysis of the same data [2], which allows comparison between these two categorizing

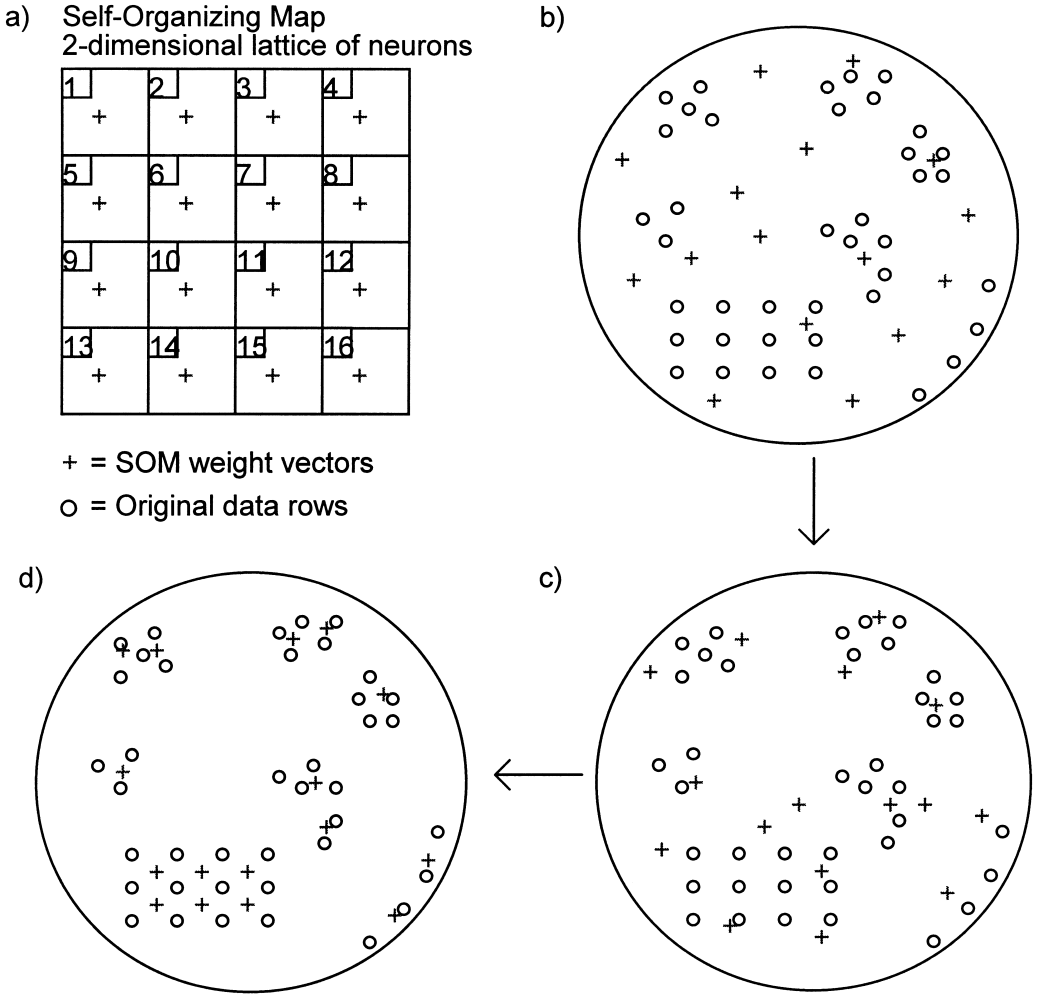

Fig. 1. Principle of Self-Organizing Map. a) Example of a SOM consisting of a 4×4 matrix of neurons. b) The weight vectors (+) of neurons are first initialized with random profiles. c) Intermediate configuration during the learning process where weight vectors are moving towards the data profiles (○). d) Finally, weight vectors come to represent groups of data profiles.

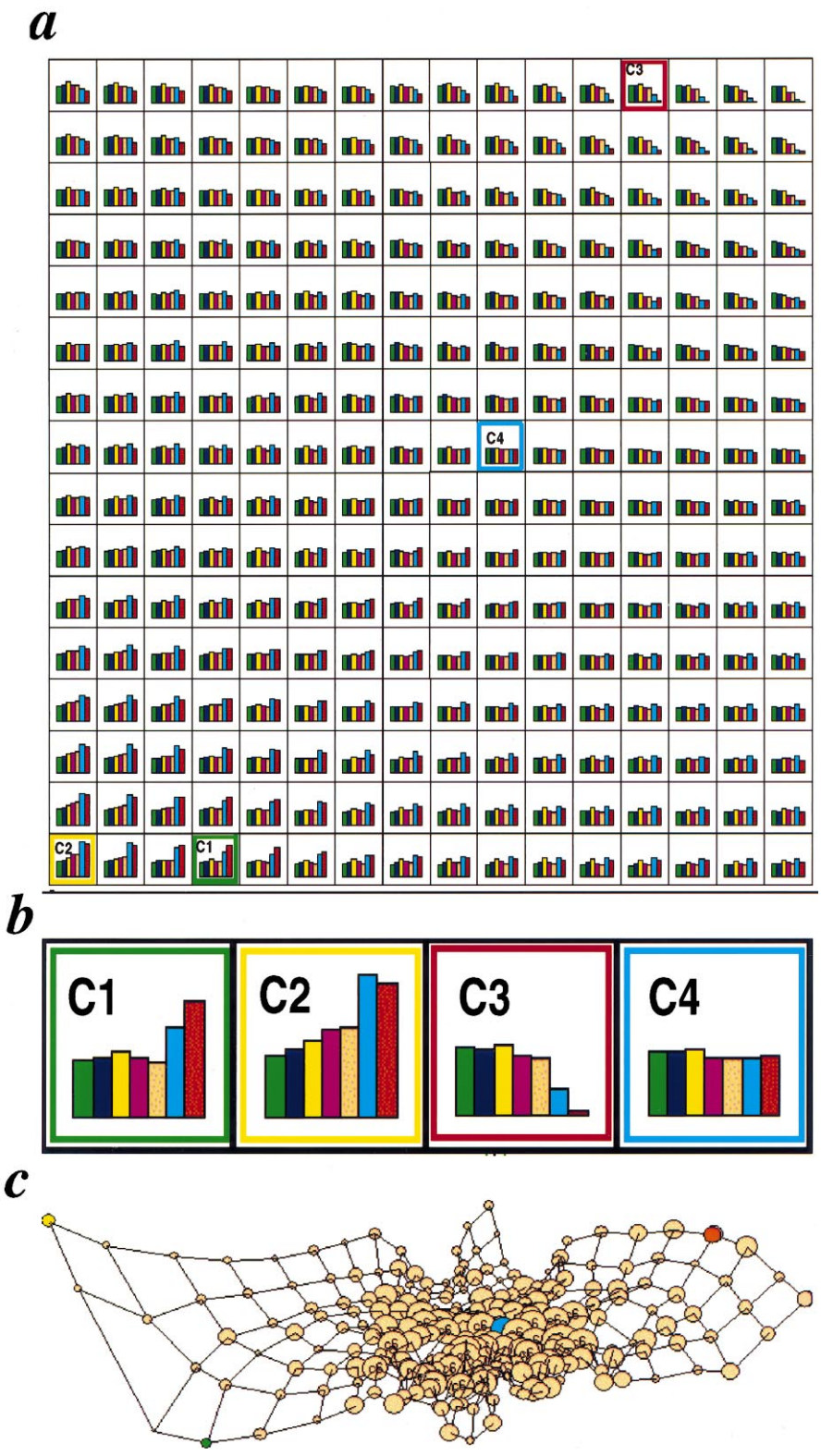

Fig. 2. SOM analysis of data of yeast gene expression during diauxic shift [2]. Data were analyzed by a prototype of GenePoint software. a: Genes with a similar expression profile are clustered in the same neuron of a 16×16 matrix SOM and genes with closely related profiles are in neighboring neurons. Neurons contain between 10 and 49 genes. b: Magnification of four neurons similarly colored in a. The bar graph in each neuron displays the average expression of genes within the neuron at 2-h intervals during the diauxic shift. c: SOM modified with Sammon's mapping algorithm. The distance between two neurons corresponds to the difference in gene expression pattern between two neurons and the circle size to the number of genes included in the neuron. Neurons marked in green, yellow (upper left corner), red and blue are similarly colored in a and b and had their gene content further analyzed in Table 1.

Table 1
Genes grouped to clusters indicated in Fig. 2

C1 (green)		C2 (yellow)		C3 (red)		C4 (blue)	
ORF	Description	ORF	Description	ORF	Description	ORF	Description
YAL054C	**Acetyl-CoA synthetase**	**YBR072W**	**Heat shock protein 26**	**YBL027W**	**Ribosomal protein L19**	YBR153W	Riboflavin synthesis
YER065C	**Isocitrate lyase**	**YCR021C**	**Heat shock protein 30**	**YBR027W**	**Ribosomal protein S101**	YDR147W	pathway protein
YLR174W	**Isocitrate dehydrogenase**	**YDR171W**	**Heat shock protein 42**	**YER131W**	**Ribosomal protein 26B**	YDR507C	Ethanolamine Kinase
YLR377C	**Fructose 1,6-bisphosphatase**	**YFL014W**	**Heat shock protein 12**	**YGL103W**	**Ribosomal protein L28**	YER172c	Serine/threonine
YNL117W	**Malate synthase**	**YGR088W**	**Cytoplasmic catalase T**	**YGL123W**	**Ribosomal protein S2**	YFL009W	kinase
YIL125W	**α-Ketoglutarate dehydrogenase**	YFR015C	Glycogen synthase 1	**YHR203C**	**Ribosomal protein S7A**	YIL139C	ATP-dependent RNA
YKR097W	**P-enolpyruvate carboxyl kinase**	YLR258W	Glycogen synthase 2	**YIL018W**	**Ribosomal protein L5A**	YJL041W	helicase
YHR096C	**Hexose transporter**	YMR105C	Phosphoglucomutase	**YIL52C**	**Ribosomal protein L34B**	YKL209C	Part of a ubiquitin
YBR117C	**Transketolase**	YNR001C	Citrate synthase	**YIL133C**	**Ribosomal protein 22**	YLL048C	ligase complex
YEL012w	Ubiquitin-protein ligase	YKL085W	Malate dehydrogenase	**YJL136C**	**Ribosomal protein S25C**	YLR071c	Subunit of DNA
				YJL190C	**Ribosomal protein S24A**		polymerase-zeta
8 genes with unknown function		12 genes with unknown function		**YKR057W**	**Ribosomal protein S25A**		Nucleoskeletal protein
				YLR175W	**Ribosomal pseudouridine synthase**		ABC transporter
				YLR325C	**Ribosomal protein L38**		Yeast bile transporter
				YML063W	**Ribosomal protein 10B**		Component of RNA
				YMR121C	**Ribosomal protein L13B**		polymerase II
				YMR242C	**Ribosomal protein L18A**		
				YNL069C	**Ribosomal protein 23**	23 genes with unknown function	
				YOL121C	**Ribosomal protein S16B**		
				YKR059W	**Translation initiation factor 4A**		
				YAL038W	Pyruvate kinase		
				6 genes with unknown function			

ORF = open reading frames from *Saccharomyces* genome database. Genes in bold type share a functional pathway or regulatory element (see Section 3).

methods. The first cluster (C1, marked green in Fig. 2) contains genes that are strongly activated towards the end of the diauxic shift (note that the scale in bar graphs is logarithmic). This cluster contained 18 genes, 10 of these had a known function. Out of these 10 genes, nine encode enzymes in the glucose metabolism pathways. All the seven genes which were visually grouped into a late activated expression group by DeRisi et al. [2] are found in this green cluster.

C2 (yellow cluster) contains genes which are strongly activated during the diauxic shift but level off at the last time point. This cluster contains two kinds of genes. Five of the 10 genes with a known function are involved in glucose metabolism, as the majority of genes in C1 were. This highlights the ability of SOM to organize not only genes with related expression pattern, but also functionally related genes into neighboring neurons. The remaining five genes encode stress-activated proteins, four of which belong to the heat shock protein family. All these heat shock proteins as well as cytoplasmic catalase T and glycogen synthase 2 share the stress response element (STRE) in their promoter region [2], suggesting that this element predominantly regulates the expression of these genes during diauxic shift. All the five known genes visually placed to a same group by DeRisi et al. [2] are found in this cluster.

C3 (red cluster) contains genes that are suppressed towards the end of the diauxic shift. Out of 21 genes in this cluster, all but one encode proteins involved in protein synthesis and 19 of these are ribosomal proteins. In addition, many protein synthesis related genes are found in neighboring neurons (data not shown) and all the genes grouped into a late suppressed group in the previous analysis of these data [2] are found in this or neighboring clusters. Only six out of a total of 27 genes had an unknown function, and it is to be expected that many of these are involved in protein synthesis.

Genes that are not regulated by diauxic shift are clustered in a more random manner and showed less functional similarity. Thus, genes in the blue cluster (C4) have little obvious functional similarity to each other. In addition, the functions of these genes are often less clearly delineated and the majority of the genes (23 out of a total of 33) in this cluster do not have a known function.

4. Discussion

Our results demonstrate that SOM is a fast and convenient method to organize and interpret gene expression data. Expression profiles of a group of genes are represented by a common weight vector and the data are easily visualized as a two-dimensional matrix. Application of Sammon's algorithm transforms the data into a format where the relationship between individual neurons is even more clearly visualized.

Examination of the gene contents of individual neurons demonstrates that in many cases, the clustering achieved by SOM reliably predicts functional similarity. This is particularly the case for those genes which are clearly regulated by the applied treatment. Similarity reflects participation in a common pathway or regulation by a common regulatory element in a promoter region, or both. In contrast, genes which were not regulated at the transcriptional level during diauxic shift are grouped together in more random basis and the clustering of non-regulated genes into a same neuron does not appear to predict functional similarity.

The yeast genome still contains thousands of genes with unknown function. Elucidation of the functional role of these genes is currently being attempted by examining the phenotype of yeast with a null mutation of these genes. Problems arise when null mutants do not show any clear phenotype. Our results suggest an alternative approach to find a functional role for an unknown gene. Gene expression profiles from yeast exposed to a variety of different environmental conditions are becoming publicly available. Applying SOM to these different data sets may reveal conditions under which this particular gene is regulated and because clustering of regulated genes appears to predict functional similarity, comparison with other genes in the same neurons gives valuable hints about the possible functional role of the unknown gene.

Many mammalian genes with an unknown function have a homologue in the yeast genome. Localization of this homologue in yeast SOMs may also suggest a possible functional role. Furthermore, the entire genomic sequences of *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens* are being finalized. As the genomes of these species are far more complex than that of yeast, the need for efficient methods to analyze and categorize gene expression data in these species is obvious. Since SOMs typically work better when more data are provided, this method may become a valuable tool in the organization and interpretation of mammalian gene expression data.

While this article was being finalized, another paper appeared which had used SOM to analyze yeast cell cycle data using software running under Unix and requiring a Web browser [13]. The conclusions of that paper generally agree with ours.

References

[1] Brown, P.O. and Botstein, D. (1999) Nature Genet. 21, 33–37.
[2] DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Science 278, 680–686.
[3] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. (1998) Science 282, 699–705.
[4] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Mol. Biol. Cell 9, 3273–9327.
[5] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson Jr., J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D. and Brown, P.O. (1999) Science 283, 83–87.
[6] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Proc. Natl. Acad. Sci. USA 95, 14863–14868.
[7] Kohonen, T. (1995) Self-Organizing Maps, Springer, Berlin.
[8] Simula, O. and Kangas, J. (1995) in: Neural Networks for Chemical Engineers (Bulsari, A.B., Ed.), pp. 371–384, Elsevier Science, Amsterdam.
[9] Ultsch, A. and Siemon, H.P. (1990) in: Proc. INNC'90, Int. Neural Network Conf., Dordrecht, Netherlands, pp. 305–308.
[10] Kohonen, T. and Hari, R. (1999) Trends Neurosci. 22, 135–139.
[11] Sammon Jr., J.W. (1969) IEEE Trans. Computers C-18, 401–409.
[12] Koikkalainen, P. (1994) in: Proceedings of the 11th European Conference on Artificial Intelligence (Cohn, A., Ed.), pp. 211–215, Wiley and Sons, New York.
[13] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Proc. Natl. Acad. Sci. USA 96, 2907–2912.