Journal of Neurochemistry, 2006, **97** (Suppl. 1), 87–91 doi:10.1111/j.1471-4159.2005.03537.x

SPECIAL ISSUE

Identification of conserved modes of expression profiles during hippocampal development and neuronal differentiation *in vitro*

Michal Dabrowski,* Alicja Adach,* Stein Aerts,† Yves Moreau‡ and Bozena Kaminska*

**Laboratory of Transcription Regulation, Department of Cell Biology, The Nencki Institute of Experimental Biology, Warsaw, Poland*

†*Laboratory of Neurogenetics, Department of Human Genetics, VIB and Katholieke Universiteit Leuven, Belgium*

‡*Department of Electrical Engineering, Katholieke Universiteit Leuven, Heverlee (Leuven), Belgium*

Abstract

Gene expression profiles can be regarded as sums of simpler modes, analogous to the modes of a vibrating violin string. Decomposition of temporal gene expression profiles into modes by singular value decomposition (SVD) was reported before, but the question as to what degree the SVD modes can be interpreted in terms of biology remains open. We report and compare the results of SVD of published datasets from hippocampal development, neuronal differentiation *in vitro*, and a control time-series hippocampal dataset. We demonstrate that the first SVD mode reflects the magnitude of expression, interpretable on the Affymetrix platform. In the datasets from gene profiling of hippocampal development and neuronal differentiation, the second mode reflects a monotonous change in expression, either up- or down-regulation, in the time course of experiment. We demonstrate that the top two SVD modes are conserved between datasets and therefore, likely reflect properties of the underlying system (gene expression in hippocampus) rather than of a particular experiment or dataset. Our results also indicate that the magnitude of expression, and the direction of change in expression during hippocampal development, are uncorrelated, suggesting that they are regulated by largely independent mechanisms.

Keywords: development, hippocampus, magnitude, mode, neuronal differentiation, singular value decomposition.

J. Neurochem. (2006) **97** (Suppl. 1), 87–91.

Gene profiling experiments have documented changes in gene expression in several regions or cell populations of the developing brain (Mody *et al*. 2001; Diaz *et al*. 2002; Dabrowski *et al*. 2003; Gurok *et al*. 2004). Analysis of those data holds promise of identifying genes with common regulation. Gene expression profiles can conveniently be regarded as sums of modes, analogous to the modes of a vibrating violin string. Decomposition of temporal gene expression profiles into modes by singular value decomposition (SVD) was previously reported in yeast, fibroblasts and the immune response (Alter *et al*. 2000, 2001; Holter *et al*. 2000, 2001; Raychaudhuri *et al*. 2000; Dewey and Galas 2001; Wall *et al*. 2001; Horn and Axel 2003; Guthke *et al*. 2005). However, the question as to what degree the resulting SVD modes can be interpreted in terms of biological processes underlying gene expression remains unanswered. Alter *et al*. (2000) hypothesized that ''at least some SVD modes represent independent regulatory processes contributing to the overall gene expression''. On the other hand, Kuruvilla *et al*. (2002) stated about the SVD modes:

Received June 21, 2005; revised manuscript received September 19, 2005; accepted September 20, 2005.

Address correspondence and reprint requests to Michal Dabrowski, Laboratory of Transcription Regulation, Nencki Institute, ul. Pasteura 3, 02–093 Warsaw, Poland. E-mail: m.dabrowski@nencki.gov.pl

Abbreviations used: GO, Gene Ontology; SVD, singular value decomposition.

© 2005 International Society for Neurochemistry,

"While very efficient basis vectors, the vectors themselves are completely artificial and do not correspond to actual profiles".

SVD is a linear transformation (Strang 1993) of expression vectors from the original basis of time-points to a new orthogonal basis of modes. The use of an orthogonal basis usually simplifies description of data, which often facilitates discovery of patterns. The mathematical properties of modes, allowing them to reflect naturally independent and additive effects on gene expression, are particularly interesting in the context of the combinatorial nature of gene regulation (Ihmels *et al*. 2004) and recent experimental evidence suggesting modularity of metazoan *cis*-regulatory regions (Davidson 2001).

With only a limited number of measurements performed, a single SVD mode will likely reflect multiple processes with similar effects on expression. However, it does not preclude the possibility that only a subset of all regulatory processes will be reflected by a particular SVD mode. This is an interesting prospect because then, regulation of such a mode would be simpler than regulation of the whole expression profile (a sum of modes), and a complex problem of analysis of gene regulation could be split into several simpler ones.

In previous work (Dabrowski *et al*. 2003), we demonstrated an overall high similarity of temporal gene expression profiles in neuronal differentiation *in vitro* and hippocampal development *in vivo* (Mody *et al*. 2001). In this work, we report and compare the results of SVD of two previously analysed datasets, and a control hippocampal dataset (Wilson *et al*. 2005), to identify biologically interpretable SVD modes.

Methods

Sources, format and annotation of expression data

The published dataset of 1926 intensity profiles of five time-points from expression profiling of hippocampal development (Mody *et al*. 2001) with the Affymetrix Mu11K chip was downloaded from http://braingenomics.princeton.edu/. The published dataset of 8799 intensity profiles (Wilson *et al*. 2005) from expression profiling of rat hippocampus with the Affymetrix RG-U34A chip, including a control time-series of five time-points following peripheral injection of saline (this time-series will be referred to here as 'the control dataset'), can be downloaded from http://pepr.cnmcresearch.org/. Our published dataset of 3216 ratio profiles of six time-points from expression profiling of the mouse primary hippocampal neuronal culture (Dabrowski *et al*. 2003) with cDNA microarrays can be downloaded from http://www.esat.kuleuven.ac.be/neurdiff/. The profiles obtained on the Affymetrix platform were identified by probe set identifiers and the profiles obtained on cDNA microarrays by GenBank/EMBL/DDBJ accession numbers. These primary identifiers were mapped to the Ensembl 27–3 gene_stable_ids, and to the NCBI Homologene homology_id. This resulted in mapping of 1926 hippocampal profiles to 1885 distinct gene_stable_ids, 3216 neuronal profiles to 1824 distinct gene_stable_ids and 8799 control profiles to 3123 gene_stable_ids. In total, 453 genes were common between the hippocampal and the neuronal dataset, and 306 genes (orthologs) were common between the mouse hippocampal dataset and the rat control dataset. We used only the profiles with no missing values from each dataset. Separately for either dataset, we computed a single average expression profile for each gene_stable_id, resulting in expression matrices: hippocampal D_H (1855×5), neuronal D_N (1824×6) and control D_C (2248×5).

SVD and comparison of loadings between two datasets

Before SVD, the matrices D_H, D_N and D_C were column-normalized (i.e. each column was divided by its vector norm) and then log-transformed, resulting in matrices A_H, A_N and A_C, respectively. SVD was performed separately on matrices A_H, A_N and A_C, resulting in matrices u_H, m_H, v_H; u_N, m_N, v_N; and u_C, m_C, v_C, respectively. For the comparison between the hippocampal and neuronal dataset, from the matrices $u_{H \text{ and }} u_N$ we selected the gene loading vectors for the genes common between these two datasets. This resulted in matrices u_{HN} and u_{NH}. Column k of matrix u_{HN} gives loadings of the k-th hippocampal mode $v_H(k)$ for all the genes common between the neuronal and hippocampal dataset. Column l of matrix u_{NH} gives loadings of the l-th neuronal mode $v_N(l)$ for all the genes common between the neuronal and hippocampal dataset. For comparison between the hippocampal and control dataset, the loading matrices u_{HC} and u_{CH} for the genes common between these two datasets were constructed as for the hippocampal–neuronal comparison. We calculated the Pearson correlation coefficient r between each pair of columns of u_{HN} and u_{NH}, and, separately, between each pair of columns of u_{HC} and u_{CH}. The two-sided p-values corresponding to these correlations were obtained from the Student t distribution, with the t statistics calculated with the formula $t = r[d/(1\text{-}r^2)]^{1/2}$, where d is the number of the degrees of freedom (Motulsky 1995).

Gene Ontology annotation

Separately for the neuronal and for the hippocampal dataset, Gene Ontology (GO) terms significantly over- or under-represented in the group of genes with positive loadings of the second mode as compared with the group of genes with negative loadings of this mode, were identified with program GOstat (Beissbarth and Speed 2004) available at http://gostat.wehi.edu.au/cgi-bin/goStat2.pl. GO terms with p-values < 0.01 after correction for the multiplicity of testing by the Benjamini (FDR) method are reported.

Results and discussion

Characterization of top two SVD modes

The datasets from gene profiling of neuronal differentiation *in vitro*, hippocampal development, and the control dataset from hippocampus of young rats following peripheral injection of saline, will be referred to as neuronal, hippocampal and control, respectively. The neuronal dataset was obtained on a cDNA microarray platform and the last two, on the Affymetrix platform. In this work, we concentrate on two highest-ranking SVD modes and on the hippocampal development, with the two other datasets used for comparisons

that enabled interpretation of the top two hippocampal SVD modes.

The distribution of the singular values for each dataset is dominated by the respective first modes (Modes 1) (Figs 1a, e and i). When the first singular value is omitted from the plot, it can be seen that the importance of the remaining modes is very similar between the two developmental datasets (Figs 1a and e zoom). Inspection of the profiles of Modes 1 (Figs 1c, g and k) reveals that they are constant in time and thus, their loadings capture differences in the magnitude of the corresponding linear expression vectors. Addition of a vector $\bar{c}$ with constant components c to an expression vector $\bar{x}$ in the $\log_b$ scale shifts the corresponding expression profile in the $\log_b$ scale by the constant c along the y ($\log_b$ expression) axis. A shift of an expression profile by c in the $\log_b$ scale corresponds to multiplication of the same expression profile in the linear scale by a factor $m = b^c$, i.e. to increasing the magnitude (vector norm) of the linear expression vector by the factor m, without changing the vector direction. If two genes i and j differ in their loadings of the first mode by $\Delta_{j,i}u_1$, then, for identical loadings of the remaining modes, the ratio of the magnitudes of their linear expression vectors is given by a factor $r_{j,i} = e^{\sigma_1 \cdot v_1 \cdot \Delta_{j,i} u_1}$, where σ_1 is the first singular value, v_1 is the loading of the of the first mode constant over the time points. The ratio of magnitudes was proposed to be a natural second metric for comparing two expression vectors (Kuruvilla *et al*. 2002), in addition to the usual comparison of vector directions (shapes of expression profiles). The magnitudes of linear expression vectors and hence, loadings of Modes 1 are only meaningful on the Affymetrix platform, not on the cDNA microarray platform. On the Affymetrix platform, the magnitude is proportional to the average absolute level of expression over all time-points. On the cDNA platform, the magnitude of linear expression vectors does not carry useful information (Moreau *et al*. 2003). Because in the hippocampal and control datasets the contribution of Modes 1 is so much larger than that of the remaining modes, as indicated by their singular values, the ratio of the magnitudes of two linear expression vectors of genes i and j differs approximately by the factor $r_{j,i}$ even without the condition of identical loadings of all the remaining modes. Therefore, in these two datasets, the linear magnitude of expression is captured, to a good approximation, by the respective first modes, with little input to the magnitude from the remaining modes.

In two developmental datasets (hippocampal and neuronal), the second modes (Modes 2) reflect components of monotonous change in expression, increase or decrease, for the positive or the negative sign of loadings, respectively (Figs 1d and h). Given that the second singular values are largest among the non-constant modes (Figs 1a and e zoom), the sign of the loading of the second mode captures the difference between an overall up- or down-regulation in the time-course of either experiment. In both developmental datasets, the expression vectors form two large clusters,

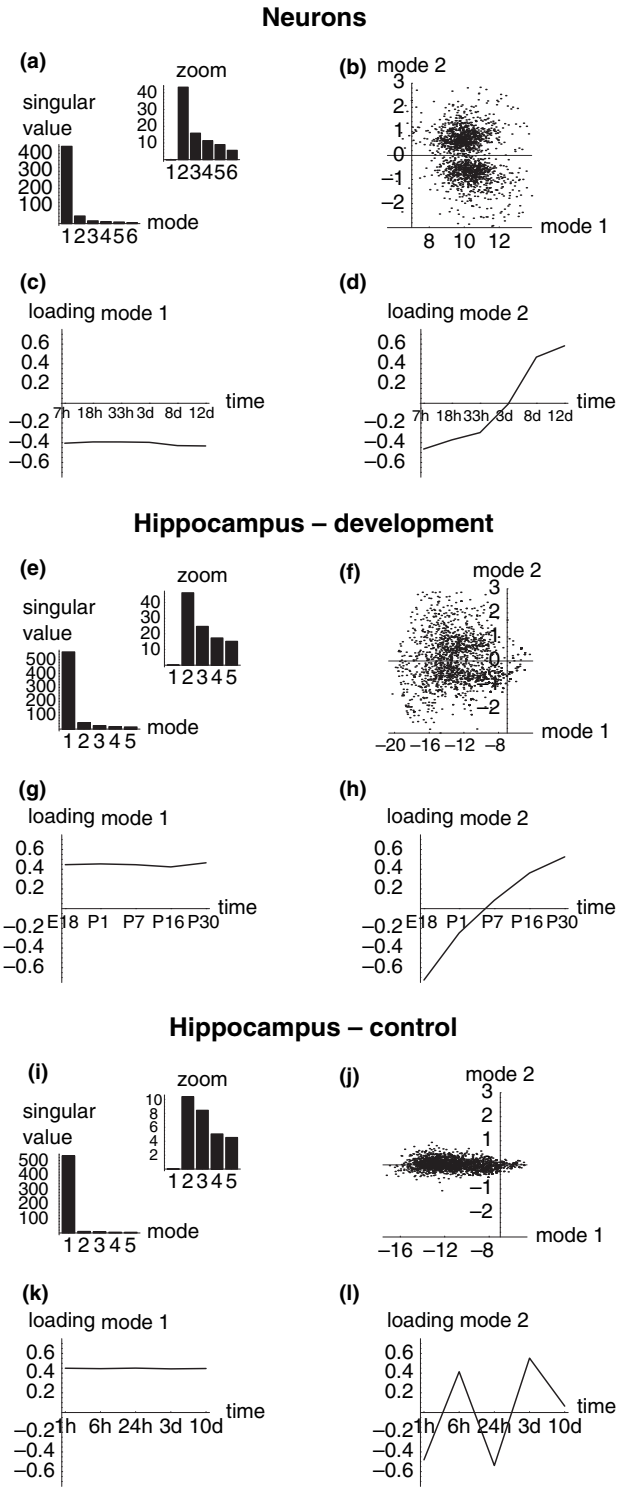


Fig. 1 Characterization of top two SVD modes. SVD was performed separately for the neuronal, hippocampal and control dataset. (a, e, i) Singular values. (c, d; g, h; k, l) Profiles of the top two modes. (b, f, j) Loadings of the top two modes to expression vectors of individual genes.

differing by the sign of loading of the respective Mode 2 (Figs 1b and f). Gene Ontology annotation of the genes represented in these clusters, provided as Supplementary Table S1, indicates that in both systems Modes 2 reflect changes of gene expression expected in the time-course of neuronal development, with up-regulation of expression of genes involved in synaptic vesicle transport, ion transport and energy metabolism, and down-regulation of genes involved in protein and nucleic acids biosynthesis, cell division and, interestingly, also regulation of transcription.

In the control dataset, the second mode has a biphasic temporal profile and its loadings do not show a bimodal distribution. This mode was included for comparison, to illustrate that the shape and the distributions of loadings, observed for the two other datasets, are data driven and likely reflect their developmental origin.

The dominance of the contribution of the first constant mode does not mean that expression of the majority of genes is constant in time. This is certainly not true for the two developmental datasets, where for any loadings of Modes 1 most genes simultaneously have non-zero loadings of Modes 2 (Figs 1b and f), so expression of the majority of genes changes over time, which was also shown previously by ANOVA (Mody *et al*. 2001; Dabrowski *et al*. 2003). On the other hand, expression of the majority of genes may well be constant in the control dataset, as indicated by clustering of their expression vectors around the zero loading of the second mode (Fig. 1j).

Conservation of modes between experiments

We assumed that some of the SVD modes represent independent regulatory mechanisms common to many genes, such as shared *cis*-regulatory features or a distance to an enhancer. If so, then the loadings of such modes across the genes should be maintained between different experiments in which these regulatory mechanisms are expected to operate.

We performed two independent comparisons: of the hippocampal dataset with the neuronal dataset; and of the hippocampal dataset with the control dataset. There were 453 common genes represented in both the hippocampal and neuronal dataset, and 306 common orthologs between the mouse hippocampal and rat control dataset. Separately for each comparison, we calculated Pearson correlation coefficients between columns (eigenarrays) of the loading matrices u from either of the compared datasets for the genes represented in both datasets (Fig. 2). In the comparison between the hippocampal and control datasets, the highest correlation ($r = 0.70$, d.f. $= 304$) was obtained between Modes 1 from either dataset (Fig. 2a, modes: h1, c1), representing the magnitude of expression.

Comparison between the hippocampal and neuronal dataset revealed an even higher correlation ($r = 0.76$, d.f. $= 451$)

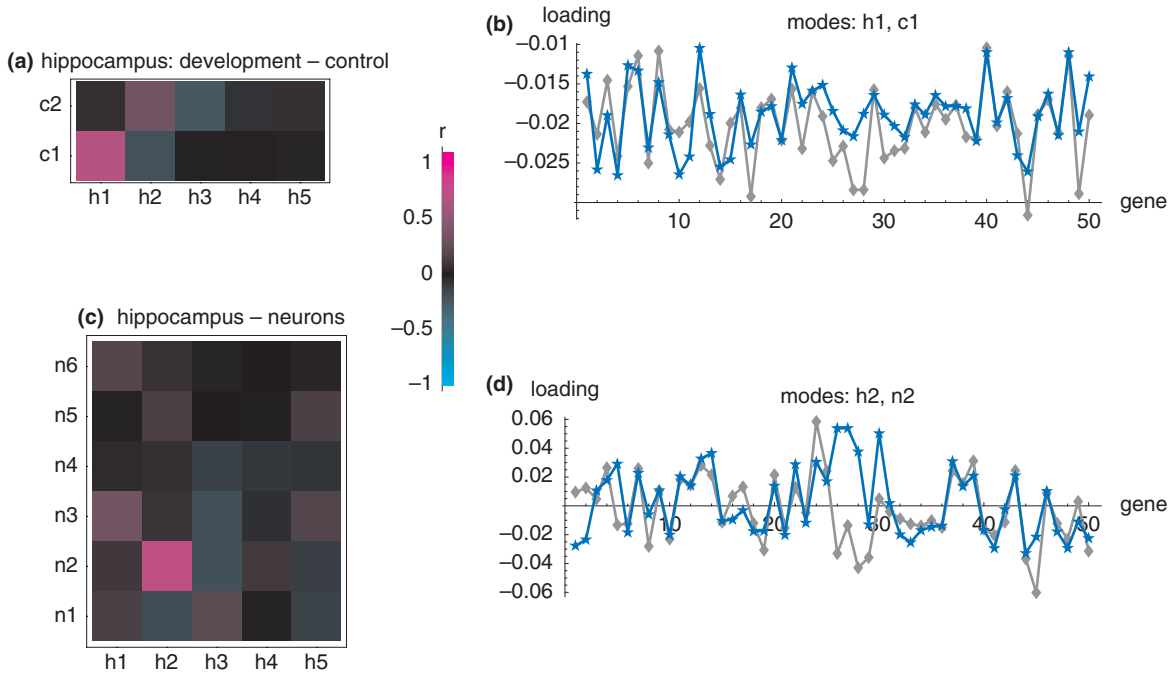


Fig. 2 Conservation of modes between experiments. (a, c) Pearson correlation coefficients r between pairs of columns of the loadings matrices u from the compared datasets, represented as correlation matrices in two-colour scale. The letters before the mode numbers indicate the dataset of origin: c = control, n = neuronal, h = hippocampal. (b) Loadings of the respective first modes, from the hippocampal and control dataset, for a representative 50 of 306 genes common between the two datasets. (d) Loadings of the respective second modes, from the hippocampal, and neuronal dataset, for a representative 50 of the 453 genes common between the two datasets.

between Modes 2 (Fig. 2c, modes: h2, n2), capturing the direction of change in expression in the time-course of development. These correlations are highly significant, with *t*-distribution *p*-values of 10^{-46} and 10^{-83}, for the conservation of Modes 1 and 2, respectively. Because the formula for the *t*-statistics does not take into account dependencies between genes, we checked the meaning of the obtained correlations by plotting the loadings of the corresponding modes from the compared systems; this confirmed a remarkably good agreement (Figs 2b and d). We demonstrated that the top two SVD modes are conserved between the datasets and therefore likely reflect properties (magnitude, direction of change in expression) of the underlying system (gene expression in hippocampus) rather than a particular experiment or dataset.

Uncoupling of magnitude and direction of change

Because, in the hippocampal dataset, the magnitude of expression and the direction of change in expression during development are captured by different SVD modes, they are uncorrelated. This suggests that during hippocampal development, the magnitude of gene expression and the direction of change in expression are regulated by largely independent mechanisms. Our finding has implications for studies of gene co-regulation, because it is in agreement with modular models of promoter organization (Yuh *et al*. 1998) and some of the models of enhancer action (Blackwood and Kadonaga 1998).

Acknowledgements

This work was supported by European Marie-Curie program grant MERG-CT-2004–510153, and Polish State Committee for Scientific Research grant SPUB-M. SA is funded by Fund for Scientific Research Flanders (FHO). YM is supported by grants BelSpo IUAP V-22, KUL-GOA Ambiorics, IDO IOTA & Genet. Net., FWO G.0115.01 & G.0388.03, FP5 CAGE.

Supplementary Material

The following supplementary material is available for this article online.

Table S1. GO annotation.

This material is available as part of the online article from http://www.blackwell-synergy.com

References

Alter O., Brown P. O. and Botstein D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA* **97**, 10 101–10 106.

Alter O., Brown P. O. and Botstein D. (2001) Processing and modeling genome-wide expression data using singular value decomposition. *Proc. SPIE* **4266**, 171–186.

Beissbarth T. and Speed T. P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**, 1464–1465.

Blackwood E. M. and Kadonaga J. T. (1998) Going the distance: a current view of enhancer action. *Science* **281**, 61–63.

Dabrowski M., Aerts S., Van Hummelen P., Craessaerts K., De Moor B., Annaert W., Moreau Y. and De Strooper B. (2003) Gene profiling of hippocampal neuronal culture. *J. Neurochem.* **85**, 1279–1288.

Davidson E. H. (2001) *Genomic Regulatory Systems*. Academic Press, San Diego, CA, USA.

Dewey T. G. and Galas D. J. (2001) Dynamic models of gene expression and classification. *Funct. Integr. Genomics* **1**, 269–278.

Diaz E., Ge Y., Yang Y. H., Loh K. C., Serafini T. A., Okazaki Y., Hayashizaki Y., Speed T. P., Ngai J. and Scheiffele P. (2002) Molecular analysis of gene expression in the developing pontocerebellar projection system. *Neuron* **36**, 417–434.

Gurok U., Steinhoff C., Lipkowitz B., Ropers H. H., Scharff C. and Nuber U. A. (2004) Gene expression changes in the course of neural progenitor cell differentiation. *J. Neurosci.* **24**, 5982–6002.

Guthke R., Moller U., Hoffmann M., Thies F. and Topfer S. (2005) Dynamic network reconstruction from gene expression data applied to immune response during bacterial infection. *Bioinformatics* **21**, 1626–1634.

Holter N. S., Mitra M., Maritan A., Cieplak M., Banavar J. R. and Fedoroff N. V. (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl Acad. Sci. USA* **97**, 8409–8414.

Holter N. S., Maritan A., Cieplak M., Fedoroff N. V. and Banavar J. R. (2001) Dynamic modeling of gene expression data. *Proc. Natl Acad. Sci. USA* **98**, 1693–1698.

Horn D. and Axel I. (2003) Novel clustering algorithm for microarray expression data in a truncated SVD space. *Bioinformatics* **19**, 1110–1115.

Ihmels J., Bergmann S. and Barkai N. (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20**, 1993–2003.

Kuruvilla F. G., Park P. J. and Schreiber S. L. (2002) Vector algebra in the analysis of genome-wide expression data. *Genome Biol.* **3**, RESEARCH0011.

Mody M., Cao Y., Cui Z., Tay K. Y., Shyong A., Shimizu E., Pham K., Schultz P., Welsh D. and Tsien J. Z. (2001) Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc. Natl Acad. Sci. USA* **98**, 8862–8867.

Moreau Y., Aerts S., De Moor B., De Strooper B. and Dabrowski M. (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.* **19**, 570–577.

Motulsky H. (1995) *Intuitive Biostatistics*. Oxford University Press, Oxford.

Raychaudhuri S., Stuart J. M. and Altman R. B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series, in Proceedings of the Pacific Symposium on Biocomputing, (Altman R. B., Lauderdale K., Dunker A. K., Hunter L. and Klein T. A. eds) pp. 455–466. World Scientific Publishing, Singapore.

Strang G. (1993) *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, MA.

Wall M. E., Dyck P. A. and Brettin T. S. (2001) SVDMAN – singular value decomposition analysis of microarray data. *Bioinformatics* **17**, 566–568.

Wilson D. N., Chung H., Elliott R. C., Bremer E., George D. and Koh S. (2005) Microarray analysis of postictal transcriptional regulation of neuropeptides. *J. Mol. Neurosci.* **25**, 285–298.

Yuh C. H., Bolouri H. and Davidson E. H. (1998) Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**, 1896–1902.