FEBS Letters 533 (2003) 110–114 FEBS 26857

Gene expression monitoring accurately predicts medulloblastoma positive and negative clinical outcomes

Michael J. Korenberg*

Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada K7L 3N6

Received 25 October 2002; accepted 8 November 2002

First published online 6 December 2002

Edited by Julio Celis

Abstract Prediction of medulloblastoma clinical outcome is crucial to personalizing treatment, both to identify high-risk patients for aggressive or alternative therapy and to spare those at low risk from excessive treatment. The best predictors [Pomeroy et al. (2002) Nature 415, 436–442], based on gene expression monitoring at diagnosis, have shown much less accuracy in recognizing patients with eventual failed outcomes – $<50\%$ for the predictor making fewest total errors – than those who would survive, while a single gene predictor exhibited reverse asymmetry. Such inaccuracy in recognizing one of the outcomes is a problem for clinical use. We hypothesized that a non-linear model could be built to significantly improve prediction of medulloblastoma outcome, thereby promoting use of gene-expression-based predictors in a clinical setting. In fact, this approach resulted in fewer errors and much less asymmetry in prediction, and bidirectional accuracy of about 80% could be obtained via its combination with other methods. Indeed, three combinations of methods were identified that yielded significantly better predictions of clinical outcome than previously attained, making feasible predictors of medulloblastoma treatment response with greatly improved bidirectional accuracy essential for clinical use.

Key words: Clinical outcome; Treatment response; Gene expression; Microarray; Medulloblastoma; Embryonal tumor

1. Introduction

Prediction of treatment response of medulloblastoma patients is crucial to personalizing therapy and improving clinical outcome [1]. Identifying patients likely to have poor response allows early institution of more aggressive or alternative therapy; recognizing other patients likely to have favorable outcome avoids over-treatment. A classic paper [2] introduced use of gene expression monitoring, and weighted voting (WV), to distinguish accurately between various acute leukemia classes, and motivated much further work with microarrays. Indeed, prediction of clinical outcome based on gene expression was recently achieved [1] for a group of 60 children with medulloblastoma, using several different classification algorithms including *k*-nearest neighbors (*k*-NN), WV, support vector machines (SVM), and IBM SPLASH. The *k*-NN made fewest total errors (13), but was much more accurate in recognizing eventual survivors (37 of 39 correct) than failed outcomes (10 of 21 correct). In fact, all these methods yielded strongly asymmetric predictors biased towards the survivor group. The single gene TRKC predictor showed reversed asymmetry (accuracy: 81% failed group, 59% survivor group). Two majority-vote combinations of predictors each reduced the number of errors to 12, but were also strongly asymmetric (accuracy: 61.9% failed group, 89.7% survivor group).

Here it is shown that a new approach, involving parallel cascade identification (PCI) [3], outperforms all individual methods considered in the original study [1], both in decreasing total errors and in diminishing the asymmetry in accuracy of recognizing the two outcomes. Most importantly, PCI combines well with metastatic staging, *k*-NN, TRKC, and SVM methods: errors are reduced by one-quarter compared to the best achievable, by the combination of predictors, in the original study [1], and asymmetry in performance is considerably reduced. In addition, PCI combined with staging and TRKC predictors to essentially eliminate asymmetry, achieving about 80% correct in predicting either failure or survivor outcomes. These results are important because, for the first time, they make feasible combination predictors of medulloblastoma outcome with good bidirectional accuracy, which is critical for clinical use.

Two points should be noted about the dataset in the original study [1]. First, because survivor outcomes considerably outnumber failures in the set, simply predicting every outcome to be survivor would make 65% of the predictions correct. Hence it is worthwhile to correct for the different sizes of the outcome groups by considering the breakdown of prediction errors for each outcome rather than simply the total number of errors. The *k*-NN correctly predicted 47.6% of failure and 94.9% of survivor outcomes, averaging 71%, the same overall accuracy as WV (52.3% failed group, 89.7% survivor group) and SVM (57.1% failed group, 84.6% survivor group). IBM SPLASH averaged 72% accuracy (61.9% failed group, 82.1% survivor group) though it made two more errors than *k*-NN.

Second, prediction of medulloblastoma outcome was difficult, and demanded a large training set for the tested methods [1] to be accurate. The above-reviewed results [1] were obtained by cross-validation, where a predictor was trained on 59 of the profiles and then tested on the remaining (held out) profile, with the procedure repeated until all profiles had been tested. To obtain the *k*-NN results, models with 1–200 genes were tested to find the optimal number of genes (8) for the

*Corresponding author. Fax: (1)-613-353 1729.
E-mail address: korenber@post.queensu.ca (M.J. Korenberg).

0014-5793/02/$22.00 © 2002 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(02)03777-8

model, and the value of k was similarly selected [1]. The large number of profiles required for training did not allow a subset to be set aside as an independent set. Hence the obtained accuracy still requires confirmation on independent sets as was pointed out [1]. Indeed, to date no published paper has tested a multi-gene-expression-based predictive model for medulloblastoma outcome on a dataset independent of that used to obtain the model.

Recently, it was shown that the treatment response of a group of acute myeloid leukemia (AML) patients could be predicted from their gene expression profiles using PCI [3]. See also review by Kirkpatrick [4]. In the present paper, the need for independent sets is recognized by adopting certain architectural parameter values for the parallel cascade model, the number of genes used, and the method of selecting the genes as previously published [3] for the AML study. While there is no reason that these choices from the latter study should be optimal for building the medulloblastoma outcome predictor, the aim was to approach as closely as possible a blind test. When the AML choices [3] are directly adopted, the resulting PCI model also obtains 71% accuracy in predicting medulloblastoma outcome, with the major difference that critical values were not chosen over the same set where the performance is measured. In addition, to put the comparison on the same footing as for the predictors in the earlier study [1], PCI results are also obtained when the architectural model parameters and the number of genes used are selected specifically for medulloblastoma outcome prediction. In this case, the performance is shown to surpass that of each individual method previously tested [1].

2. Method

The same profiles as in [1] (i.e. the raw values given after rescaling by Pomeroy et al.) were used here, and were taken at time of diagnosis. Each profile contained expression levels of 6817 human genes, but because of duplicates and additional probes in the Affymetrix microarray, in total 7129 gene expression levels were present in the profile. Profile nos. 1–21 were from medulloblastoma patients who ultimately had failed outcomes, while profiles nos. 22–60 were from patients who proved to be survivors.

Use of PCI for outcome prediction is described in [3] and now summarized. Briefly, given one or more expression profiles for both failed (F) and survivor (S) outcomes, begin by selecting genes that assist in distinguishing between the two outcomes. For AML response prediction [3], the genes selected were the 200 having greatest difference in raw expression levels between the first F and S profiles, and this was followed here. Accordingly, the first F profile (no. 1) and first S profile (no. 22) were compared to find the 200 genes with greatest difference in raw expression levels between the two profiles. The corresponding 200 raw values from profile no. 1 were appended, in the same order as they had in the profile, to form an F segment, and an S segment was similarly prepared from profile no. 22. The two segments were spliced together to form a 400-point training input, and a corresponding training output was defined as -1 over the F segment and 1 over the S segment of the input [3]. While only one F and one S profiles were employed here to select the genes to use, and to construct the training input, multiple exemplars can certainly be used for these purposes.

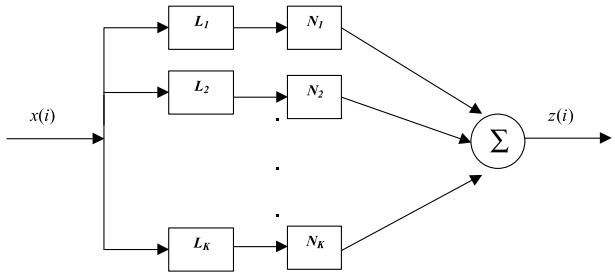

Fig. 1. Parallel cascade model used to predict medulloblastoma clinical outcome. Each L is a dynamic linear element and each N is a polynomial static non-linearity.

A parallel cascade model (Fig. 1) was then identified to approximate the input/output relation, using the method [5] previously applied to protein family prediction [6]. In Fig. 1, each L is a dynamic linear element, each N is a polynomial static non-linearity. This parallel LN model is related to a parallel LNL structure proposed earlier by Palm [7] where the static non-linearities were logarithmic and exponential functions rather than the polynomials used here. Fig. 2A shows the training input, and Fig. 2B (solid line) the corresponding training output. When these data were used to identify a parallel LN model (Fig. 1), the model mean square error (MSE) was 4.1%, expressed relative to variance of the training output. Fig. 2B (dashed line) shows the calculated output of the identified model, when evoked by the training input. Notice that the latter output is predominately negative over the F segment and positive over the S segment of the training input. Hence the identified model is able to distinguish between F and S profiles, at least with respect to the two training exemplars.

To identify the PCI model, certain parameter values relating chiefly to its architecture had to be pre-specified [5]:

1. Memory length of dynamic linear element L that began each cascade. The memory has length R+1 if the element's output depends on the present value of its input and on the R previous input values.
2. Degree of polynomial static non-linearity N that followed the linear element.
3. Maximum number of cascades allowed in the model.
4. A threshold concerning minimum reduction in MSE required before allowing a candidate cascade into the model.

As noted above, the present choices for these parameters were taken directly from a different cancer prediction study [3]. There it was found that a PCI model could be identified to predict treatment response of an AML patient group if the memory length (R+1) of each linear element was 12, and the degree of each polynomial static non-linearity was 7, with seven cascades in the model. These values and the same threshold of 11 were used here. While it seemed unlikely that these parameters would also be optimal for medulloblastoma outcome prediction, it allowed the remaining 58 profiles that had not been used for training to form an independent set for testing the PCI model. The latter set comprised profile nos. 2–21 and 23–60.

To classify a test profile, the raw expression values from the previously selected genes were appended, in the same order as used above, to form a 200-point input signal, which was then fed to the identified parallel cascade model to obtain a corresponding output signal. Since the model had a memory length

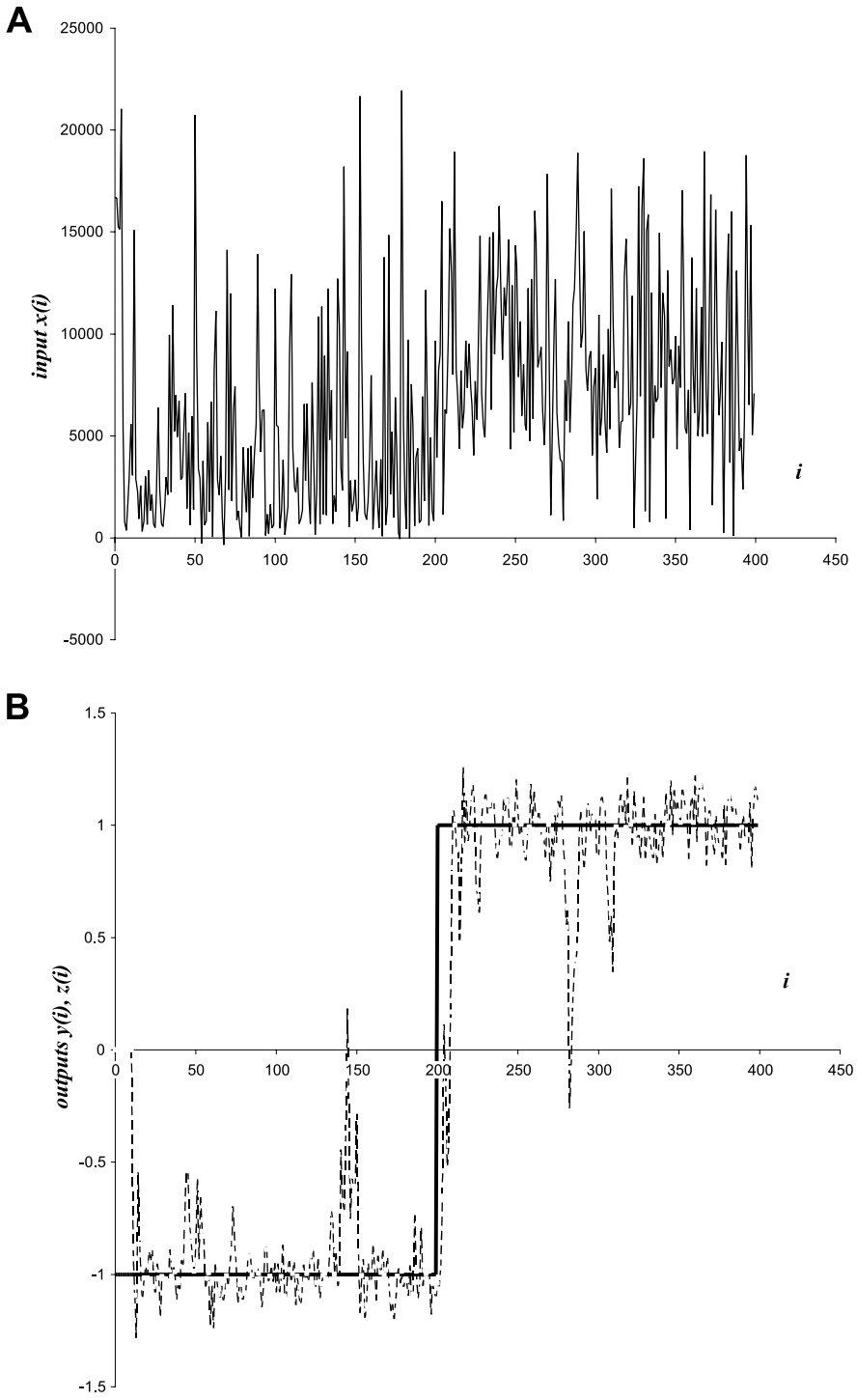

Fig. 2. A: Training input $x(i)$ formed by splicing together the raw expression levels of genes from the first 'failed outcome' profile no. 1 and first 'survivor outcome' profile no. 22. The genes used were the 200 having greatest difference in expression levels between the two profiles. B: Training output $y(i)$ (solid line) defined as -1 over the 'failed outcome' portion of the training input and 1 over the 'survivor outcome' portion. The training input and output were used to identify a parallel cascade model of the same form as in Fig. 1. The dashed line represents calculated output $z(i)$ when the identified model is stimulated by training input $x(i)$. Note that $z(i)$ is predominately negative (average value: -0.961) over the 'failed outcome' portion and positive (average value: 0.961) over the 'survivor outcome' portion, of the training input. This ability to separate failed and survivor outcome profiles is exploited by using the identified model to filter corresponding portions of novel profiles prior to their classification.

of 12, the first 11 points of the output signal were excluded to allow the model to 'settle', and only the last 189 points of each output signal were used to determine the class for the corresponding profile.

The class of each of the 58 test output signals was predicted as follows, using a leave-one-out protocol. Each time, the Euclidean distance was calculated of the query output signal from each of the remaining 57 output signals. That is, if $z^{(0)}(i)$

and $z^{(j)}(i)$ respectively represent the query output signal and one of the remaining 57 output signals, where $i=1,...,200$, and the first 11 output values are excluded as explained above, then one calculates distance:

$$D=\sqrt{\sum_{i=12}^{200}(z^{(0)}(i)-z^{(j)}(i))^2}$$

for each one of the remaining output signals. The query output signal was then assigned the class of the closest one of the 57 other output signals. This process was repeated until all 58 profiles had been classified.

In summary, the model was identified using only the first F and S profiles and values from the AML study [3], and so was never tested on profiles employed to obtain the model. The leave-one-out protocol was simply used to interpret the output signals of the identified model. Moreover, to show the benefit of the model, results are also provided when simple Euclidean distance was used, in a leave-one-out protocol, to classify the 200-point input signals without first obtaining corresponding model output signals.

Next, to fairly compare PCI classifiers with predictors in [1], the model architectural parameters and the number of genes to use were determined specifically for medulloblastoma outcome prediction. Genes were selected using the same criterion as employed previously [3], namely the top-ranked genes having greatest difference in raw expression levels between the first F and S profiles, and only these two profiles were used to create the training input. PCI models were tested corresponding to 400, 300, 100, and 20–30 genes. Good classification ensued when the top-ranked 22 genes were selected (Table 1), and memory length (R+1) was 4, polynomial degree was 5, two cascades were allowed in the model, and threshold was 6. Model MSE was about 4.9%. Note that here a 44-point training input was employed, and a 22-point input signal corresponded to each of the 58 test profiles. However, good classification was also observed for other numbers of genes used and other architectural parameters. In Table 1, some genes alone would not be useful to predict outcome, but are important when their expression levels are considered in combination with others.

A leave-one-out protocol was again employed in the class prediction. In place of using simple Euclidean distance, better accuracy generally resulted from calculating the correlation coefficient of a query output signal with each of the remaining 57 output signals. Since memory length was 4, the first three points of each 22-point output signal were not used. Let $z^{(0)}(i)$ and $z^{(j)}(i)$ respectively represent the query output signal and one of the remaining 57 output signals, where now $i=1,...,22$. Then, for each of the remaining output signals, calculate correlation coefficient:

$$r=\frac{\sum_{i=4}^{22}(z^{(j)}(i)-\overline{z^{(j)}})(z^{(0)}(i)-\overline{z^{(0)}})}{\left(\sum_{i=4}^{22}(z^{(j)}(i)-\overline{z^{(j)}})^2\right)^{\frac{1}{2}}\left(\sum_{i=4}^{22}(z^{(0)}(i)-\overline{z^{(0)}})^2\right)^{\frac{1}{2}}}$$

where $\overline{z^{(0)}}$ and $\overline{z^{(j)}}$ denote the average of $z^{(0)}(i)$ and $z^{(j)}(i)$ respectively over $i=4,...,22$. The query output signal was assigned the class of that one of the 57 other output signals it is most positively correlated with, i.e. the correlation coefficient is largest. Again, to show the utility of the identified PCI model, classification accuracy is also reported below when the correlation coefficient was used, in a leave-one-out protocol, to classify the 22-point input signals without first obtaining corresponding model output signals.

Finally, PCI is shown to combine well with methods considered previously [1]. The improvement is both in diminishing the asymmetry of the resulting predictor and in reducing the number of classification errors.

3. Predicting clinical outcome

3.1. Using parameters from the AML study

In this section, the PCI architectural parameters, and the

Table 1
Twenty-two genes used to predict medulloblastoma outcome

Accession no.	Position in profile (1–7129)	Description
M33197	42	AFFX-HUMGAPDH/M33197_5_at (endogenous control)
D14530	226	40S ribosomal protein S23
D79205	568	Ribosomal protein L39
HG1612-HT1612	804	Macmarcks
HG3549-HT3751	930	Wilm'S tumor-related protein
J02611	1054	APOD apolipoprotein D
J03040	1068	SPARC SPARC/osteonectin
M63379	2128	CLU clusterin (complement lysis inhibitor; testosterone-repressed prostate message 2; apolipoprotein J)
U03057	2606	Actin bundling protein mRNA
U12404	2757	HSPB1 heat shock 27 kDa protein 1
X53331	4247	MGP matrix protein gla
X67951	4463	PAGA proliferation-associated gene A (natural killer-enhancing factor A)
X70683	4504	SOX4 SRY (sex-determining region Y)-box 4
Z48950	5168	HISTONE H3.3
D86974	5507	KIAA0220 gene, partial cds
M19311	5642	CALM1 calmodulin 1 (phosphorylase kinase, delta)
L04483	6026	RPS21 ribosomal protein S21
M14483	6181	PTMA gene extracted from human prothymosin alpha mRNA
Z19554	6209	VIM vimentin
M37457	6311	Na^+,K^+-ATPase catalytic subunit alpha-III isoform gene
S54005	6388	Thymosin beta-10
X01703	6915	Alpha-tubulin mRNA

number (200) of genes used to construct the training input, were not chosen from medulloblastoma profiles but from the AML study [3]. The PCI model was obtained using only the first F and first S medulloblastoma profiles, while the remaining 58 profiles were reserved for testing. This is important because it means that the PCI model was tested here on an independent dataset from that used to obtain the model. Of the 58 test profiles, 12 of 20 F (60%) and 31 of 38 S (81.6%) were correctly classified, a 71% average. This probably underestimates PCI accuracy in predicting medulloblastoma outcome, since critical parameters were not tailored for this prediction task but came, as just noted, from a study to predict AML treatment response [3]. When the 200-point input signals corresponding to the test profiles were classified without first obtaining corresponding model output signals, the accuracy dropped to about 50% (35% on F profiles, 65.8% on S profiles), showing that the PCI model was essential.

3.2. Using parameters tailored to medulloblastoma

Here, 14 of 20 F (70%) and 32 of 38 S (84.2%) profiles were correctly classified, so PCI accuracy averaged 77% (misclassified F profiles: nos. 5, 8, 13, 14, 20, 21; misclassified S profiles: nos. 24, 33, 35, 44, 54, 58). When instead the 22-point input signals corresponding to the test profiles were classified without first obtaining corresponding model output signals, the accuracy averaged 63% (50% on F profiles, 76.3% on S profiles).

Fisher's exact test probabilities and average accuracy overall were respectively $P < 0.0002$ and 71% (k-NN), and $P < 0.000063$ and 77% (PCI). Even if the k-NN P-value is regarded as the critical level for significance, and one divides this by the number of results (2) for the Bonferroni correction for multiple hypothesis testing, the PCI P-value is less than the adjusted level. Moreover, Fisher's exact test probabilities and average accuracy for a localized-disease subset [1] were $P < 0.00851$ and 69.5% (k-NN), and $P < 0.001423$ and 78% (PCI). For the latter case, the breakdown was 45.5% on F and 93.5% on S profiles (k-NN), and 72.7% on F and 83.3% on S profiles (PCI). Thus, for both cases, PCI provided improved prediction.

3.3. Combining PCI with other methods

As noted earlier, combining predictors using majority voting achieved the most accurate predictions in [1]: 12 errors in total, correct on 61.9% of F and 89.7% of S profiles, averaging 76%, Fisher's exact test probability of $P < 0.0000461$. It is now shown that the PCI classification described in Section 3.2 combines well with various predictors in [1]. In each case, a majority vote decided the prediction. Combinations considered were:

1. PCI with metastatic staging, k-NN, TRKC, SVM. This combination achieved minimum total errors and highest average correct classification rate. Recognizing 14 of 20 (70%) test F profiles, and 35 of 38 (92.1%) test S profiles, it averaged 81%, making nine errors. Misclassified F profiles were nos. 2, 5, 13, 14, 20, 21; misclassified S profiles were 26, 33, 36. Fisher's exact test probability was $P < 0.000001712$, less than 1/25 of the P-value for the combination prediction in [1], suggesting that PCI would be a useful component to combine with the predictors from that study.
2. PCI with metastatic staging, TRKC. This combination achieved best symmetry in predicting F and S outcomes. Correctly classifying 16 of 20 (80%) test F profiles and 30 of 38 (78.9%) test S profiles, it averaged 79%.
3. PCI with metastatic staging, SVM. This combination ranked second in average correct classification and total errors. Recognizing 14 of 20 (70%) test F and 34 of 38 (89.5%) test S profiles, it averaged 80%, making 10 errors.

4. Discussion and future applications

In this paper, the identified parallel cascade model was essentially used as a filter through which input signals representative of the profiles were passed in order to produce output signals. Nearest neighbor was used to classify the output signals here, but many other classification algorithms, such as SVM, artificial neural networks, or PCI can also be applied to classify the output signals.

PCI combines well with methods considered by Pomeroy et al. [1] for medulloblastoma outcome prediction. A future paper will combine PCI with other techniques for interpreting gene expression profiles such as aggregative-hierarchical-clustering [8], self-organizing maps [9], and k-means-clustering [10]. Recently, microarrays were used to predict disease outcome of breast cancer [11]. Combining other methods with PCI here may enhance prediction of recurrence, and assist in selection of treatment regimen. Finally, PCI can be used to classify many other biologic profiles, e.g. proteomics data, or profiles representative of DNA methylation over large regions of the genome. PCI has been demonstrated by itself [3,4], and in combination with other methods, to be a valuable tool in predictive medicine.

References

[1] Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., McLaughlin, M.E., Kim, J.Y.H., Goumneroval, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J.A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S. and Golub, T.R. (2002) Nature 415, 436–442. Supplementary information and datasets: http://www.genome.wi.mit.edu/MPR/CNS.
[2] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Science 286, 531–537.
[3] Korenberg, M.J. (2002) J. Proteome Res. 1, 55–61.
[4] Kirkpatrick, P. (2002) Nat. Rev. Drug Discov. 1 (5), 334.
[5] Korenberg, M.J. (1991) Ann. Biomed. Eng. 19, 429–455.
[6] Korenberg, M.J., Solomon, J.E. and Regelson, M.E. (2000) Biol. Cybern. 82, 15–21.
[7] Palm, G. (1979) Biol. Cybern. 34, 49–52.
[8] Eisen, M., Spellman, P.T., Botstein, D. and Brown, P.O. (1998) Proc. Natl. Acad. Sci. USA 95, 14863–14867.
[9] 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.
[10] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Nat. Genet. 22, 281–285.
[11] Van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002) Nature 415, 530–536.