

Prediction of Treatment Response Using Gene Expression Profiles

Michael J. Korenberg*

Department of Electrical and Computer Engineering, Queen's University, Kingston, Ontario K7L 3N6, Canada,
and Cascade Genomics Corporation, Box 1847, Kingston, Ontario K7L 5J7, Canada

Received November 15, 2001

This paper concerns prediction of clinical outcome from gene expression profiles using work in a different area, nonlinear system identification. In particular, the approach can predict long-term treatment response from data of a landmark article by Golub et al. (Golub, T. R.; Slonim, D. K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J. P. et al. *Science* **1999**, *286*, 531–537) that has not previously been achieved with these data. The present paper shows that, for these data, gene expression profiles taken at time of diagnosis of acute myeloid leukemia contain information predictive of eventual response to chemotherapy. This was not evident in previous work; indeed, the Golub et al. article did not find a set of genes strongly correlated with clinical outcome. However, the present approach can accurately predict outcome class of gene expression profiles even when the genes do not have large differences in expression levels between the classes.

Keywords: clinical outcome • treatment response • gene expression • DNA chips

1. Introduction

Prediction of future clinical outcome, such as treatment response, may be a turning point in improving cancer treatment. This has previously been attempted via a statistically based technique for class prediction founded on gene expression monitoring, which showed high accuracy in distinguishing acute lymphoblastic leukemia (ALL) from acute myeloid leukemia (AML).¹ The technique involved selecting “informative genes” strongly correlated with the class distinction to be made, e.g., ALL versus AML, and found families of genes highly correlated with the latter distinction.¹ Each new tissue sample was classified on the basis of a vote total from the informative genes, provided that a “prediction strength” measure exceeded a predetermined threshold. However, the technique did not find a set of genes strongly correlated with response to chemotherapy, and class predictors of clinical outcome were less successful.

In a sample of 15 adult AML patients treated with anthracycline–cytarabine, eight failed to achieve remission while seven achieved remissions of 46–84 months. Golub et al. “found no evidence of a strong multigene expression signature correlated with clinical outcome, although this could reflect the relatively small sample size”. While no prediction results for clinical outcome were presented in the paper, they stated that such class predictors were “not highly accurate in cross-validation”.¹ Similarly, Schuster et al.² could not predict therapy response using the same data in a study of five different clustering techniques: Kohonen-clustering, fuzzy-Kohonen-network, growing cell structures, K-means-clustering, and fuzzy-K-means-clustering. They found that none of the tech-

niques clustered the patients having similar treatment response. The ALL–AML dataset¹ was one of two specified for participants in the CAMDA'00 meeting, and none reported accurate prediction of treatment response with these data.

Prediction of survival or drug response using gene expression profiles can be achieved with microarrays specialized for non-Hodgkin's lymphoma³ involving some 18 000 cDNAs or via cluster analysis of 60 cancer cell lines and correlation of drug sensitivity of the cell lines with their expression profiles.⁴ Also, using clustering, Alon et al.⁵ showed that tumor and normal classes could be distinguished even when the genes used had small average differences between the classes.

In the present paper, it is shown that a technique for modeling nonlinear systems, called parallel cascade identification (PCI),⁶ can accurately predict class from gene expression profiles even when the genes do not have large differences in expression levels between the classes. In particular, the technique is able to predict long-term treatment response to chemotherapy with anthracycline–cytarabine, which was not previously possible with the data from ref 1. The present work shows that gene expression profiles taken at time of diagnosis, and lacking a set of genes strongly correlated with clinical outcome, still enable prediction of treatment response otherwise only evident several years later.

Although the sample size of the AML treatment response group is not large, there are several reasons why the class prediction method below, and its performance over this group, is significant for interpreting both gene expression profiles and proteomics data. First, since neither the Golub et al. method¹ nor any other published to date has accurately predicted treatment response for this group of patients, it may serve as a benchmark for gauging the sensitivity of new methods. Second, the successful predictions of treatment response by

* Tel: 613-533-2931. Fax: 613-353-1729. E-mail: korenber@post.queensu.ca; genomics@attcanada.ca.

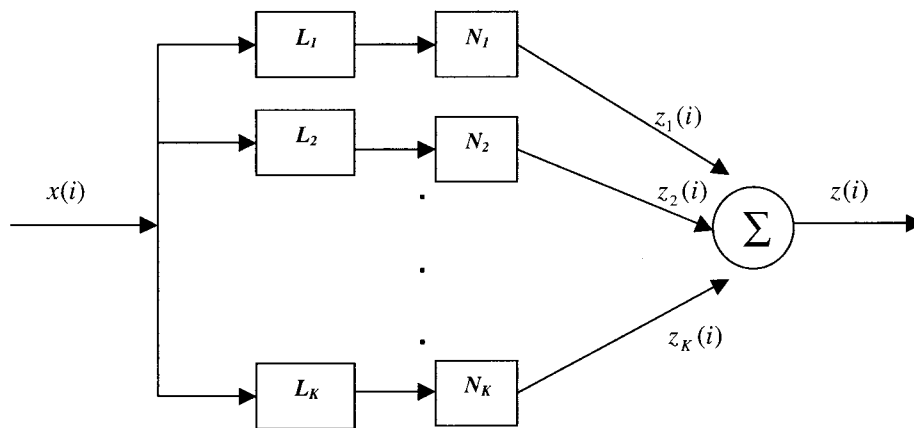
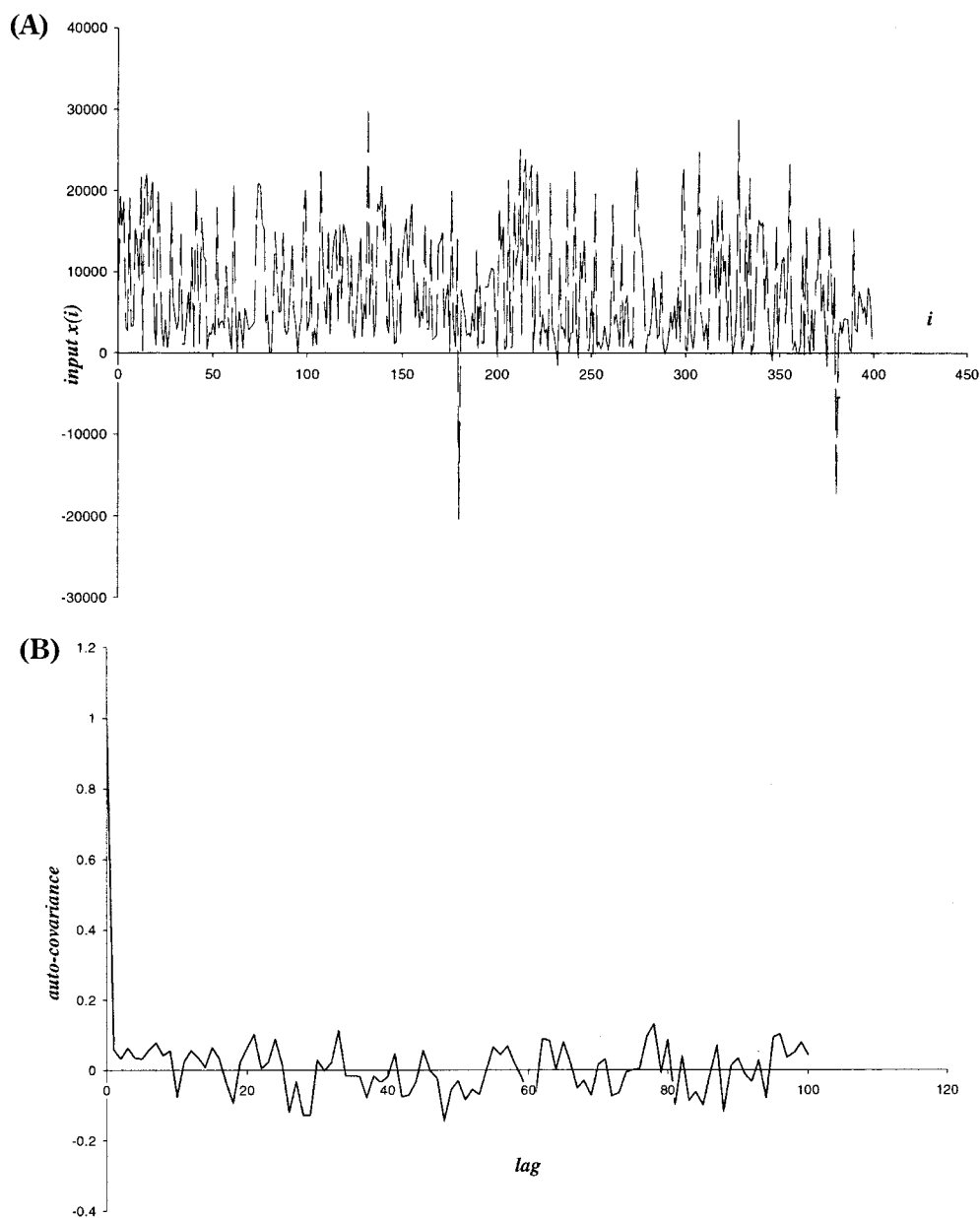


Figure 1. Parallel cascade model used to predict treatment response and leukemia class. Each L is a dynamic linear element; each N is a polynomial static nonlinearity.

the method described below are readily shown to be statistically significant (up to the 0.01 level) on well-known tests that

examined two different aspects of the prediction. Third, the same method is also shown to perform well on the ALL vs AML



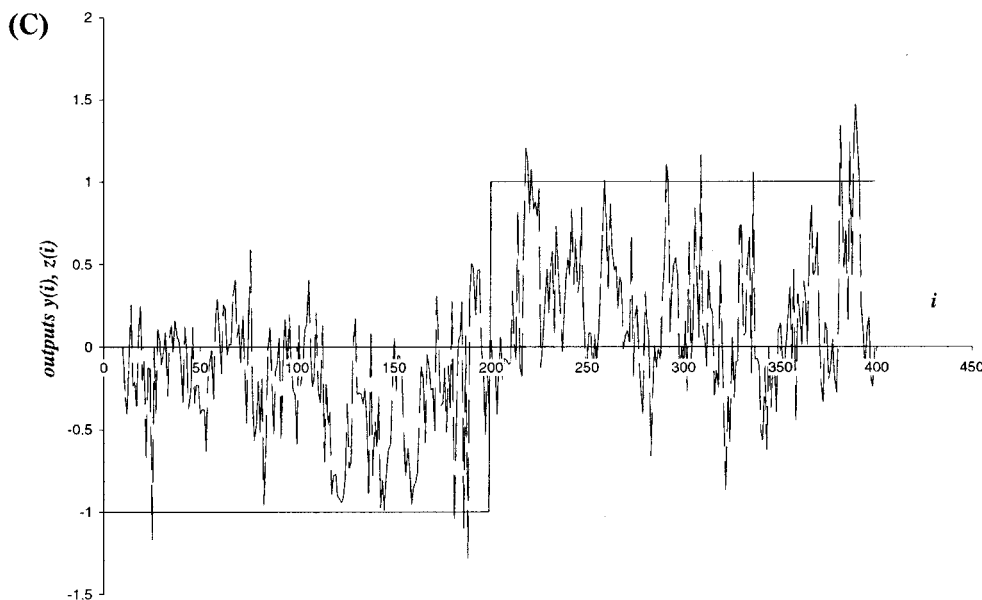


Figure 2. (A) Training input $x(i)$ formed by splicing together the raw expression levels of genes from the first “failed treatment” profile no. 28 and first “successful treatment” profile no. 34. The genes used were the 200 having greatest difference in expression levels between the two profiles. (B) Order used to append the expression levels of the 200 genes caused the auto-covariance of the training input to be nearly a δ function, indicating that the training input was approximately white. (C) Training output $y(i)$ (solid line) defined as -1 over the “failed treatment” portion of the training input and 1 over the “successful treatment” portion. The training input and output were used to identify a parallel cascade model of the form in Figure 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 over the “failed treatment” portion and positive over the “successful treatment” portion of the training input.

task. While the latter class distinction is much easier, the resulting performance indicates that the proposed method for class prediction has more general applicability for interpreting gene expression profiles. Indeed, the performance observed here is consistent with that in classifying protein sequences,⁷ where PCI classifiers have been successfully tested on thousands of novel sequences.⁸ Moreover, the approach described below is directly applicable with two-dimensional electrophoresis (2-DE) gels: the 2-DE images can similarly be interpreted using this method, e.g., for diagnosis of disease or prediction of clinical outcome.

2. Method

Development of a means for predicting clinical outcome from gene expression profiles began by viewing the problem as one of nonlinear system identification, using a “black box” approach. Here, the system to be identified was defined by one or more inputs and one or more outputs; the problem was to build a model whose input/output relation approximated that of the system, with no a priori knowledge of the system’s structure. A training input was constructed by splicing together the expression levels of genes from profiles known to correspond to failed and to successful treatment outcomes. The training output was defined as -1 over input segments corresponding to failed outcomes and 1 over segments corresponding to successful outcomes. The nonlinear system having this input/output relation would clearly function as a classifier, at least for the profiles used in forming the training input. A model was then identified to approximate the defined input/output behavior and could subsequently be used to predict the class of new expression profiles. Below, only the first failed and first successful outcome profiles were used to construct the training input; the remaining seven failed and six successful outcome

profiles served as tests. The same data were used as in ref 1. All samples had been obtained at time of leukemia diagnosis. Each profile contained the expression levels of 6817 human genes,¹ but because of duplicates and additional probes in the Affymetrix microarray, in total 7129 gene expression levels were present in the profile.

Nonlinear system identification has already been used for protein family prediction,^{7,8} and a useful feature of PCI⁶ is that effective classifiers may be created using very few training data. For example, one exemplar from each of the globin, calcium-binding, and kinase families sufficed to build parallel cascade two-way classifiers that outperformed,⁸ on over 16 000 test sequences, state-of-the-art hidden Markov models trained with the same exemplars. The parallel cascade method and its use in protein sequence classification are reviewed in ref 9.

While input $x(i)$ to a parallel cascade model will represent the expression levels of genes, both input and output of the model will be treated as if they were time-series data. The rationality of considering the gene expression values as a time series is now justified, in view of the fact that genes in a profile are not ordered sequentially. In fact, lack of actual time dependency causes no problem: PCI simply looks for a pattern in the data. This point is illustrated by the type of training data that could be used to identify the protein sequence classifiers, e.g., Figure 2a of ref 8. There, five-digit binary codes were employed to represent each amino acid in a protein sequence, resulting in subtly different plaid-like patterns for different protein families. Though these patterns were artificial in that they depended upon the five-digit codes used, parallel cascade models could be trained to distinguish between the patterns and thus classify novel protein sequences.

For the approach to work, it is necessary that (1) training exemplars from different classes produce different patterns and

(2) the memory length of the nonlinear system to be identified is of appropriate size to “capture” the pattern for each class. Analogously, to learn how to distinguish between two wood grains, say mahogany and cherry, given one table of each, a much smaller sampling window than an entire tabletop would suffice to make a decision. Moreover, sliding the sampling window over both tabletops would provide multiple training examples of each grain.

3. Model Identification

The parallel cascade (Figure 1) classifier to be constructed comprises a sum of cascades of dynamic linear (L) and static nonlinear (N) elements. “Dynamic” signifies that the element L possesses memory: its output at a given instant i depends not only upon its input x at instant i but upon past input values at instants $i - 1, \dots, i - R$ (memory length = $R + 1$). Here, every nonlinear element N is a polynomial, so that each cascade output $z_k(i)$, and hence the overall model output $z(i)$, reflect high-order nonlinear interactions of gene expression values. This parallel LN model is related to a parallel LNL structure proposed by Palm¹⁰ for modeling discrete-time Volterra systems. In Palm’s structure, the static nonlinearities were exponential and logarithmic functions, rather than the polynomials used in the present paper.

The set of failed outcomes was represented by profile nos. 28–33, 50, and 51 of data from ref 1 and the set of successful outcomes by profile nos. 34–38, 52, and 53. Raw expression levels of selected genes from the first “failed treatment” profile no. 28 and first “successful treatment” profile no. 34 were concatenated to form training input $x(i)$ (Figure 2A). Order of appending the selected genes resulted in an almost white input (Figure 2B), which is typically advantageous for nonlinear system identification, including PCI. (The selected genes had the same relative ordering as in the original profiles, and this ordering caused the input autocovariance to be closest to a δ function, out of several orderings tried.) The corresponding training output $y(i)$ was defined as -1 over the “failed treatment” segment of the input and 1 over the “successful treatment” segment (solid line, Figure 2C). For this input and output, a model was identified using PCI.⁶ The identified model clearly functions as a “predictor” of treatment response, at least for expression profile nos. 28 and 34. Indeed, when training input $x(i)$ is fed through the parallel cascade model, the resulting output $z(i)$ is predominately negative (average value: -0.238) over the “failed treatment” segment and predominately positive (average value: 0.238) over the “successful treatment” segment of the input (dashed line, Figure 2C). The identified model had a mean-square error (MSE) of about 74.8%, expressed relative to the variance of the output signal.

Care was taken to ensure that the test sequences were treated independently from the training data. First, the two profiles used to form the training input were never used as test profiles. Second, the set used to determine a few parameters chiefly relating to model architecture never included the profile on which the resulting model was tested. Thus, a model was never trained nor selected as the best of competing models, using data that included the test profile.

To identify the parallel cascade model, four parameters relating mostly to its structure had to be pre-specified. These were as follows: (i) the memory length of the dynamic linear element L that began each cascade, (ii) the degree of the polynomial static nonlinearity N that followed, (iii) the maximum number of cascades allowed in the model, and (iv) a

threshold concerning the minimum reduction in MSE required before a candidate cascade could be accepted into the model.⁶ How these parameter settings were determined is explained next.

As noted, only two profiles were used to construct the training input, which left 13 profiles for testing. Each time, 12 of these 13 were used to determine values for the above parameters, and then the parallel cascade model having the chosen parameter settings was tested on the remaining (“held out”) profile. This process was repeated until each of the 13 profiles had been tested. The 12 profiles used to determine the parameter values will be referred to as the evaluation set, which never included the profile held out for testing.

The parameter values were determined each time by finding the choice of memory length, polynomial degree, maximum number of cascades allowed, and threshold that resulted in fewest errors in classifying the 12 profiles. The limit on the number of cascades allowed actually depended on the values of the memory length and polynomial degree in a trial. The limit was set to ensure that the number of variables introduced into the model was significantly less than the number of output points used in the identification. Effective combinations of parameter values did not occur sporadically. Rather, there were ranges of the parameters, e.g., of memory length and threshold values, for which the corresponding models were effective classifiers. When the fewest errors could be achieved by more than one combination of parameter values, then the combination was selected that introduced fewest variables into the model. If there was still more than one such combination, then the combination of values where each was nearest the middle of the effective range for the parameter was chosen. Each time it was found that it was possible to achieve the best performance (two to three errors depending on the 12 profiles in the evaluation set) and employ fewest cascade variables, if the same four parameter values were used. This meant that the identical parallel cascade model was in fact chosen for each classification of the held out profile. This model had seven cascades in total, each beginning with a linear element having memory length of 12, followed by a seventh degree polynomial static nonlinearity. Figure 3A shows the impulse response functions of the linear elements in the second, fourth, and sixth cascades, and Figure 3B shows the corresponding polynomial static nonlinearities that followed.

Each time, the profile held out for testing was classified by appending, in the same order as used above, the raw expression levels of genes in the profile to form an input signal. This input was then fed through the identified model, and its mean output was used to classify the profile. If the mean output was negative, the profile was classified as “failed treatment”, and if positive as “successful treatment”. This decision criterion was taken from the earlier protein classification study.⁷

4. Predicting Treatment Outcome

The parallel cascade model correctly classified five of the seven “failed treatment” (F) test profiles and five of the six “successful treatment” (S) test profiles. The corresponding Matthews’ correlation coefficient¹¹ was 0.5476. Two different aspects of the parallel cascade prediction of treatment response were tested, and both times reached statistical significance. First, the relative ordering of profiles from the two outcome types by their model *mean* outputs was tested by the Mann–Whitney test, a nonparametric test to determine whether the model detected differences between the two profile types. The

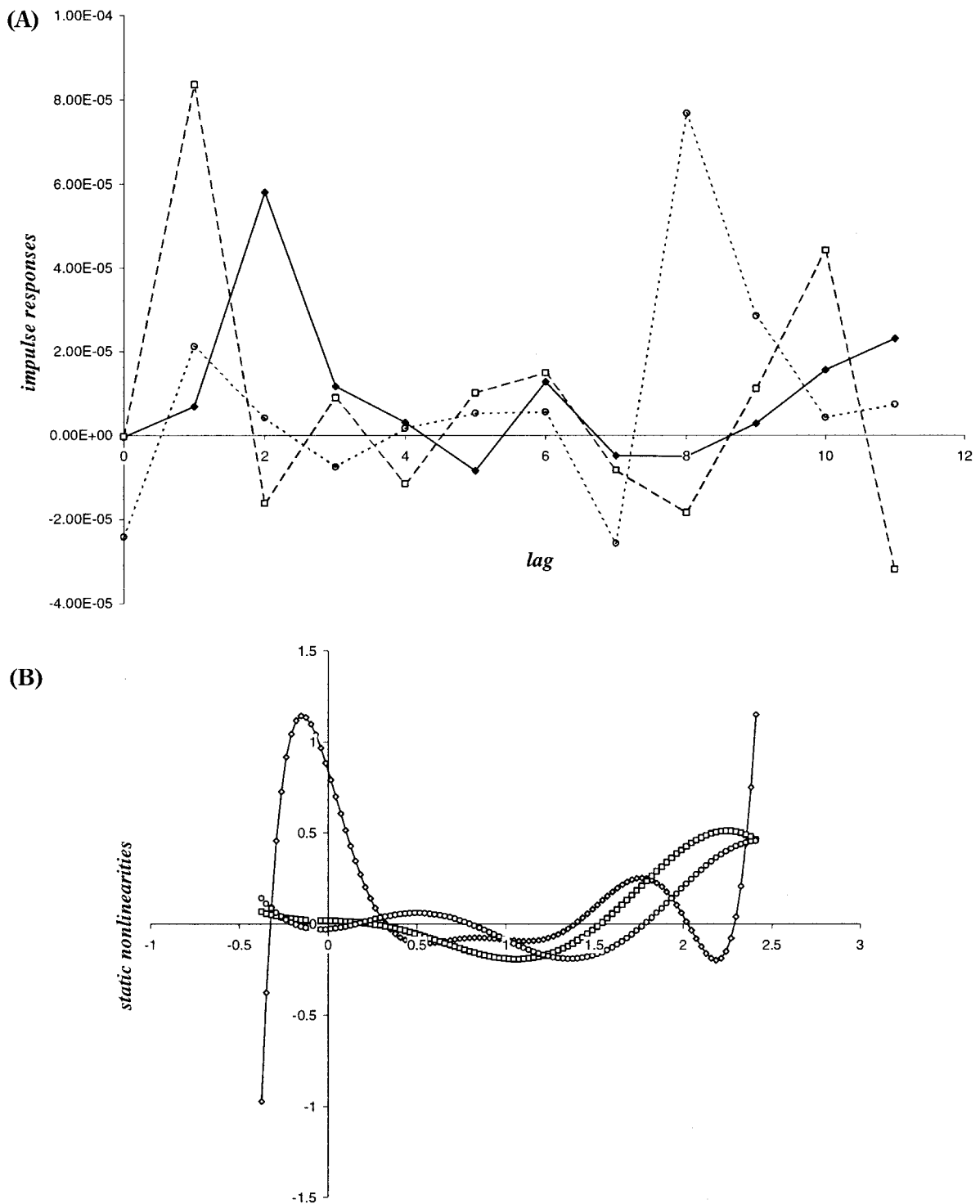


Figure 3. (A) Impulse response functions of linear elements L_2 (solid line), L_4 (dashed line), and L_6 (dotted line) in the second, fourth, and sixth cascades of the identified model. (B) Corresponding polynomial static nonlinearities N_2 (diamonds), N_4 (squares), and N_6 (circles) in the identified model; input to static nonlinearity (horizontal axis) vs output of static nonlinearity (vertical axis).

second aspect of the PCI prediction concerned how well the individual values of the classifier output for the seven F and six S test profiles correlated with the class distinction.

How did the model mean outputs order the test profiles from the two outcome types? If the parallel cascade model did not detect differences between the two types, then the relative ordering of F and S profiles by value of mean output should be random. The parallel cascade mean outputs were ranked as shown in Table 1, which indeed demonstrates that mean

outputs for F test profiles tend to be less than those for S test profiles. The corresponding Mann–Whitney test statistics were $U = 8$, $U = 34$. This meant that the way the parallel cascade model distinguished between F and S profiles was significant at the 0.0367 level on a one-tailed test, for $n_1 = 7$, $n_2 = 6$. A one-tailed test was used because, due to the way the model had been trained, the mean output was expected to be less for a failed than for a successful outcome profile.

Next, the identified parallel cascade model was found to act

Table 1. Parallel Cascade Ranking of Test Expression Profiles^a

rank	mean output	actual outcome	profile no.
1	-1.17	F	31
2	-0.863	F	32
3	-0.757	F	33
4	-0.408	S	37
5	-0.298	F	50
6	-0.0046	F	30
7	0.0273	S	53
8	0.078	S	38
9	0.110	F	51
10	0.148	F	29
11	0.194	S	52
12	0.267	S	36
13	16.82	S	35

^aF = failed treatment, S = successful treatment. The complete set of profiles is found in ref 1, and "profile no." follows the same numbering scheme.

as a transformation that converts input sequences of 200 gene expression values each into output signals whose individual values correlate with the F vs S class distinction. Because the model had a memory length of 12, the first 11 points of each output signal were excluded to allow the model to "settle", so that the 13 test profiles gave rise to $13 \times 189 = 2457$ output values. Of these, 1323 output values corresponded to the seven F test profiles and 1134 to the six S test profiles. For the S profiles, the proportion of output values that were positive was 109% of the corresponding proportion for F profiles, while the S profiles' proportion of output values that were negative was 90% that for F profiles. Indeed, with a Yates-corrected χ -square of 5.13 ($P < 0.0235$), output values corresponding to test S profiles tended to be positive more often, and negative less often, than those corresponding to test F profiles. Finally, the actual values of the outputs also correlated with the F vs S distinction. The 1323 output values corresponding to the test F profiles totaled -535.93 , while the 1134 output values for test S profiles totaled 3209.14. Indeed, point biserial correlation showed that there was a highly significant relation between the actual values of the output signals for the test profiles and the F vs S class distinction ($P < 0.0102$). Recall that the model's memory length was 12. Hence, limiting the calculation of point biserial correlation to every 12th output value would avoid any overlap in gene expression levels used to obtain such output values. The relation of these 208 output values to the F vs S distinction is still highly significant ($P < 0.0155$).

5. Predicting Leukemia Class

Parallel cascade identification was also used to distinguish between ALL and AML classes. Since this problem is simpler, only a few results are mentioned here; further details and results will be provided in a future paper. The first 15 ALL and all 11 AML profiles in the original set of ref 1 were used to construct a training input. The training output was again defined as -1 over input segments from the first class and 1 over segments from the second class. For trial values of the four parameters mentioned above, parallel cascade models were identified from the training input and output. Resulting classification performance over the entire original set of 27 ALL and 11 AML profiles was used to select the final model.

In this way, it was possible to build a model capable of classifying, with only two errors, the 34 profiles in an independent set from ref 1. While the whole original set had been employed to compare the performance of the different models,

only the 26 profiles noted above had been used in creating the training input. However, all of the original set could be used to construct the training input to increase the resulting classifier's generalization to the independent set. Golub et al.¹ used the entire original set to train their ALL-AML predictor, which then was able to classify correctly all of the independent set except for five where the prediction strength was too low for a decision. These results are comparable. Indeed, because of the simplicity of the ALL/AML distinction, minor differences in the number of profiles correctly classified would not indicate how different methods would compare on more difficult problems such as prediction of treatment response. However, it is notable that, when only the first ALL profile and AML profile from the original set were used to form the training input, a parallel cascade model was identified that averaged 83% correct in classifying remaining profiles in the set. Finally, as will be shown in a future paper, there is a strategy for using the profiles in training so that 11 of each class from the original set suffice for correct classification of all profiles in the independent set.

6. Discussion and Future Applications

PCI is only one approach to predicting treatment response, and other methods can certainly be applied. Competing techniques for interpreting patterns of gene expression include aggregative hierarchical clustering,¹² self-organizing maps,¹³ K-means-clustering^{2,14} referred to above, K-nearest neighbors,¹⁵ support vector machines,¹⁵ and artificial neural networks.¹⁶ Importantly, it has been shown in the present paper to be possible to predict long-term response of AML patients to chemotherapy using the Golub et al. data, and now wider implications can be considered. For example, the method for predicting clinical outcome described here may have broader use in cancer treatment and patient care. In a clinical situation, a set of expression profiles taken at time of diagnosis, for which the future clinical outcomes were known, could be used to determine the PCI classifier and its required parameter values. The resulting classifier could then be employed to predict the outcome for newly diagnosed patients. In particular, it has recently been shown that there are significant differences in the gene expression profiles of tumors with BRCA1 mutations, tumors with BRCA2 mutations, and sporadic tumors.¹⁷ Class prediction may be applied to distinguish the gene expression profiles of these tumor classes, predict recurrence, and assist in selection of treatment regimen.

References

- (1) Golub, T. R.; Slonim, D. K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J. P. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 531-537. Datasets: http://www-genome.wi.mit.edu/MPR/data_set_ALL_AML.html.
- (2) Schuster, A.; Dubitzky, W.; Azuaje, F. J.; Granzow, M.; Berrar, D.; Eils, R. Tumor Identification by Gene Expression Profiles: A Comparison of Five Different Clustering Methods. Critical Assessment of Microarray Data Analysis Techniques. CAMDA'00, **2000**; <http://bioinformatics.duke.edu/camda/CAMDA00/Abstracts/Schuster.asp>.
- (3) Alizadeh, A. A.; Eisen, M. B.; Davis, R. E.; Ma, C.; Lossos, I. S.; Rosenwald, A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **2000**, *403*, 503-511.
- (4) Scherf, U.; Ross, D. T.; Waltham, M.; Smith, L. H.; Lee, J. K.; Tanabe, L. et al. A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **2000**, *24*, 236-244.

- (5) Alon, U.; Barkai, N.; Notterman, D. A.; Gish, K.; Ybarra, S.; Mack, D. et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 6745–6750.
- (6) Korenberg, M. J. Parallel cascade identification and kernel estimation for nonlinear systems. *Ann. Biomed. Eng.* **1991**, *19*, 429–455.
- (7) Korenberg, M. J.; Solomon, J. E.; Regelson, M. E. Parallel cascade identification as a means for automatically classifying protein sequences into structure/function groups. *Biol. Cybern.* **2000**, *82*, 15–21.
- (8) Korenberg, M. J.; David, R.; Hunter, I. W.; Solomon, J. E. Automatic classification of protein sequences into structure/function groups via parallel cascade identification: a feasibility study. *Ann. Biomed. Eng.* **2000**, *28*, 803–811.
- (9) Korenberg, M. J.; David, R.; Hunter, I. W.; Solomon, J. E. Parallel cascade identification and its application to protein family prediction. *J. Biotechnol.* **2001**, *91*, 35–47.
- (10) Palm, G. On representation and approximation of nonlinear systems. Part II: Discrete time. *Biol. Cybern.* **1979**, *34*, 49–52.
- (11) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.
- (12) Eisen, M.; Spellman, P. T.; Botstein, D.; Brown, P. O. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 14863–14867.
- (13) Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2907–2912.
- (14) Tavazoie, S.; Hughes, J. D.; Campbell, M. J.; Cho, R. J.; Church, G. M. Systematic determination of genetic network architecture. *Nature Genet.* **1999**, *22*, 281–285.
- (15) Yeang, C.-H.; Ramaswamy, S.; Tamayo, P.; Mukherjee, S.; Rifkin, R. M.; Angelo, M. et al. Molecular classification of multiple tumor types. *Bioinformatics* **2001**, *17*, Suppl. 1, S316–S322.
- (16) Khan, J.; Wei, J. S.; Ringnér, M.; Saal, L. H.; Ladanyi, M.; Westermann, F. et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.* **2001**, *7*, 673–679.
- (17) Hedenfalk, I.; Duggan, D.; Chen, Y.; Radmacher, M.; Bittner, M.; Simon, R. et al. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **2001**, *344*, 539–548.

PR015510M