Modular Decision System and Information Integration for Improved Disease Outcome Prediction

Matthias E. Futschik [(1)], Mike Suillivan [(2)], Anthony Reeve [(3)], and Nikola Kasabov [(4)]

[(1)] Department of Information Science, University of Otago, PO Box 56
Dunedin - New Zealand
mfutschik@infoscience.otago.ac.nz

[(2)] Department of Paediatrics, Christchurch School of Medicine and Health Sciences, PO Box 4345
Christchurch - New Zealand
mike.sullivan@otago.ac.nz

[(3)] Department of Biochemistry, University of Otago, PO Box 56
Dunedin - New Zealand
anthony.reeve@stonebow.otago.ac.nz

[(4)] KEDRI, Auckland University of Technology
Auckland - New Zealand
nkasabov@aut.ac.nz

Keywords: Disease outcome prediction, Data integration, Modular hierachical models, Lymphoma

Introduction

Prediction of clinical behaviour and treatment for cancers is based on the integration of clinical and pathologic parameters. Recent reports have demonstrated that gene expression profiling provides a powerful new approach for determining disease outcome [3, 4, 6]. Although the studies showed the ability to predict disease outcome based on gene expression data, they also revealed a limited accuracy in the derived predictions. This may not be surprising, since microarrays are restricted to measuring the RNA abundance within cells. Many important post-translational events may not be reflected in microarray measurements. To exploit the full power of microarray techniques for medical applications, it will be necessary to integrate the data from microarrays with various other information. Here we have used existing clinical information and microarray data to generate a combined prognostic model which achieves a significantly improved accuracy of outcome prediction for diffuse large B-cell lymphoma (DLBCL).

Construction of prognostic modules

If clinical and microarray data each contain independent information then it should be possible to combine these data sets to gain more accurate prognostic information. To test the hypothesis that clinical and microarray data contain independent information, we constructed two separate prognostic models for DLBCL using the clinical information and microarray data published recently by Shipp *et al.* [4]. For the clinical model, the hitherto standard International Prediction Index (IPI) for DLBCL was converted into a Bayesian classifier based on the five-year survival rates in the original IPI clinical study [5]. The clinical prognostic model achieved an overall accuracy of 73.2%.

For the microarray-based predictor, we used an evolving fuzzy neural network (EFuNN) classifier for adaptive supervised learning [2]. For the 56 (of 58 samples) for which the IPI classification was present, we used the same pre-processing, gene selection and testing procedure (leave-one-out validation method) as in Shipp *et al.* [4]. The microarray-based predictor achieved an accuracy of 78.5% which was slightly better than that of the previously used support vector machines [4].

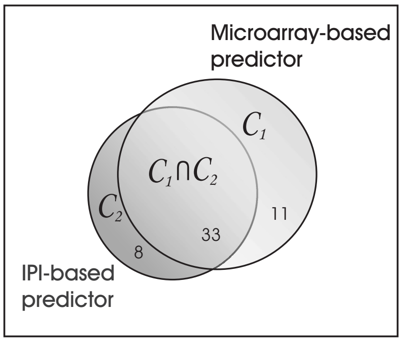

Figure 1. Venn-Diagram of overlapping sets of samples correctly predicted by the microarray-based predictor C_1 and by the IPI-based clinical predictor C_2. It shows that the predictions are complementary in 19 of 56 cases: 11 samples are correctly classified only by the microarray-based predictor, whereas 8 samples only by the clinical predictor. Altogether, 52 samples are predicted correctly by at least one predictor setting an upper threshold (92.9%) for the accuracy of the combined model.

Independence of prognostic modules

Basic set theory demonstrates that the two constructed predictors are partially complementary (Fig. 1). The independence of the predictors can be assessed by the calculation of their mutual information I:

$$I(x,y) = \sum_{x,y} R(x,y) \log_2 \frac{R(x,y)}{P(x)Q(y)} \approx 0.05$$

where x, y are predictions by the EFuNN and clinical predictor (0 for class "cured", 1 for class "fatal"), $P(x), Q(y)$ are probability distributions for the predictions and $R(x,y)$ is the joint probability of both predictors. The two variables x, y are statistically independent if the mutual information I is zero. Thus, the calculation of I shows the surprising result that the clinical prognostic model is statistically independent of the microarray-based predictor even though several of the IPI risk factors were considered biological surrogates of underlying molecular mechanisms and were expected to be correlated with gene expression in the tumor [2, 4].

Modular prediction system

To exploit the complementary nature of the two prognostic models, we constructed a hierarchical modular system combining both predictors (Fig. 2). The model parameters $(\alpha, \beta_1, \beta_2)$ were optimized by error-backpropagation, and the accuracy recorded using the leave-one-out method. Our combined model achieved an improved accuracy of 87.5%. An analysis of the combined model showed that both microarray data and clinical information were required for improved prediction accuracy. Stratification of the samples into clinical subgroups according to their IPI value demonstrated that the two predictor modules differed in their 'areas of expertise'. The IPI risk groups 'Low', 'Low-Intermediate' and 'Intermediate-High' were weighted towards the microarray-based predictor, while the 'High' risk group was weighted towards the clinical predictor for the final outcome prediction of the combined model. Interestingly, the SVM approach used by Shipp *et al.* as well as our neural network method classified the samples with IPI 'High' incorrectly. This indicates that the molecular basis of DLBCL for this group of patients might differ from patients in other IPI risk groups.

Conclusions

Integration of information from a variety of sources and the construction of complex models in molecular biology and medicine are major challenges in the advancing post-genomic era. Although considerable research has been undertaken on tumor classification based on microarray data or on clinical data, this is the first study

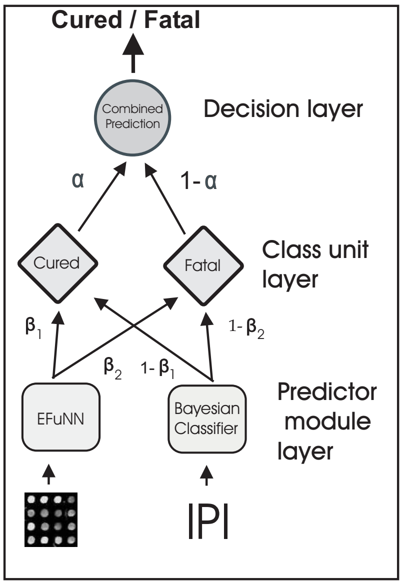

Figure 2. Three-layered hierarchical model for combination of microarray-based and clinical predictors: The first layer (*Predictor module layer*) consists of independently trained predictor modules; the second layer (*Class unit layer*) integrates the weighted predictions of both predictors for classes 'cured'/'fatal'; the third layer (*Decision layer*) produces the final prediction based on the weighted sum of the outputs of class units. Model parameter α balances possible class bias, whereas β_1 weights the predictions from both modules for class 'cured' and β_2 for class 'fatal', respectively.

focusing on the combination of clinical and microarray prediction systems. Our analysis demonstrates that the integration of microarray data with previously established clinical parameters can considerably improve disease outcome prediction. This result may lead to new possibilities of incorporating microarray techniques into clinical practise. The concept of combining predictors based on different sources of information is not restricted to the data analysed here, but generic. The inclusion of other types of clinical and molecular data is possible and may be favourable for personalizing patient care.

Acknowledgements

Further details of the analysis presented here can be found in ref.[1]. MF was supported by a PhD scholarship and bridging grant of the University of Otago. We would like to thank the reviewers for their constructive comments and Bronwyn Carlisle for proof-reading. Commercial use of techniques described and applied in this paper is subject to obtaining a license from Pacific Edge Biotechnology (*http://www.peblnz.com*).

References

[1] M. Futschik et al. Prediction of clinical behaviour and treatment for cancers. *OMJ Applied Bioinformatics*, (accepted for publication).
[2] N. Kasabov. On-line learning, reasoning, rule extraction and aggregation in locally optimised evolving fuzzy neural networks. *Neurocomputation*, 41(1-4):25–45, 2001.
[3] S. Pomeroy et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002.
[4] M. Shipp et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine*, 8:68–74, 2002.
[5] The International Non-Hodgkin's Lymphoma Prognostic Project. A predictive model for aggressive non-Hodgkin's lymphoma. *New England Journal of Medicine*, 329:987–994, 1993.
[6] L. van 't Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.