Translation of a robust, biology-driven, prognostic classifier of cancer patients' outcome into clinically relevant rules

D. Cangelosi[1] ✉, M. Muselli[2], F. Blengio[1], R. Versteeg[3], A. Eggert[4], A. Schramm[4], A. Garaventa[1], C. Gambini[1], L. Varesio[1]

[1]Laboratory of Molecular Biology, G. Gaslini Institute, Genoa, Italy
[2]Institute of Electronics, Computer and Telecommunication Engineering, Italian National Research Council, Genoa, Italy
[3]Department of Human Genetics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands
[4]Department of Pediatric Oncology and Hematology, University Children's Hospital Essen, Essen, Germany

Motivations

Cancer patient's outcome is written in part in the gene expression profile of the tumor. Clinical bioinformatic is instrumental in extracting the relevant gene clusters (signatures), the prediction models and the rules for patients stratification and clinical decision making. Hypoxia, a condition of low oxygen tension occurring in poorly vascularized tissues, has a profound effects on tumor growth and resistance to therapy. We utilized a novel biology driven approach coupled with appropriate feature selection [1] to identify the signature of hypoxic neuroblastoma cells (NB-Hypo) that stratifies neuroblastoma patients in good and poor outcome [2]. In the present work, we develop and validate a robust classifier to predict neuroblastoma patients' outcome on the bases of tumor hypoxia and we identify the most informative clinically relevant rules that combine classical risk factors and NB-hypo prognostic signature.

Methods

Gene expression profiles of 182 neuroblastoma tumors were used to develop and validate our prediction models. Validation was performed either by leave one out cross validation or 66% split of the dataset in training and testing. We utilized a Multi Layer Perceptron (MLP), a feedforward Artificial Neural Network classifier, to predict patients' outcome (alive or dead 5 years after diagnosis). Integrated analysis of classical risk factors and prognostic NB-hypo signature utilized either C4.5, an efficient algorithm used to generate decision tree or Rulex 2.0, a software suite capable of building Intelligible Learning Machines (ILM) through Shadow Clustering (SC) [3]. The classifiers were trained and tested on the expression values of the 62 probsets constituting NB-hypo signature. Furthermore, in some instances NB-hypo was condensed into a single binary attribute by means of an unsupervised k-means clustering of the patients' cohort based on the NB-hypo, 62 probsets expressionvalues.

Results

The NB-hypo classifier based on MLP predicted the outcome with an accuracy of 87% and was able to correctly predict poor outcome of all patients in the low-intermediate risk classes. The accuracy increased when NB-hypo classifier predicted the hypoxic state of neuroblastoma tumors even when it was not associated with poor outcome. Thus, NB-hypo classifier, while probing the hypoxic status of the tumor, is a new and robust predictor of neuroblastoma patient's outcome with very low errorrate that decreases to negligible levels in localized tumors. Clinicians utilize established risk factors (tumor stage, amplification of the MYCN oncogene or age at diagnosis) for prognosis and treatment choice. We investigated whether NB-hypo could be successfully added to the classical decision process improving risk definition and whether the relevant rules could be expressed in a clinically applicable form. We choose to utilize decision tree and ILM algorithms to facilitate the extraction of clinically applicable rules. Decision tree analysis revealed interesting relationships among risk factors and pointed to NB-hypo as the key element in identifying poor prognosis patients in stage 3 neuroblastoma. However, the divide and conquer approach employed by decision tree was unsatisfactory for a global study of these variables because of the excessive fragmentation of the database into small, poorly indicative groups of patients. In contrast, the covering algorithm adopted by SC identified a set of 11 rules each with a coverage greater than 30% and with less than 0.1% error. NB-hypo was integral part of these rules either as representative probsets or as a single binary attribute. Interestingly, the algo-

rithm divided the expression values of individual probsets in broad (generally two), statistically different, categories corresponding to clearly identifiable low and high expression levels. This conversion was critical for counteracting the variability microarray experiment data.

The importance of NB-hypo for outcome prediction of low risk neuroblastoma patients was confirmed. Moreover, we established rules for outcome prediction of stage 4 neuroblastoma patients that are very heterogeneous and difficult to stratify. In summary, we found that biology-based gene expression signatures and machine learning lead to patients outcome prediction and that appropriate clinical bioinformaticapproaches can extract relevant rules translatable to the clinical setting. This approach was developed for neuroblastoma tumors but the rationale and methodology can be applied successfully to other types of cancer.

References

1. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Varesio L. The l1-l2 regularization framework unmasks the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines. BMC Genomics. 2009 Oct 15;10:474.
2. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Versteeg R, Caron HN, Molenaar JJ, Ora I, Eva A, Puppo M, Varesio L. A biology-driven approach identifies the hypoxia gene signature as a predictor of the outcome of neuroblastoma patients. Mol Cancer. 2010 Jul 12;9:185.
3. Muselli M and Ferrari E. Coupling Logical Analysis of Data and Shadow Clustering for Partially Defined Positive Boolean Function Reconstruction. IEEE Trans. on Knowl. and Data Eng. 23, 1, 2011.