Generation of explicit rules predicting neuroblastoma patients' outcome

Davide Cangelosi[1]✉, Fabiola Blengio, Rogier Versteeg[2], Angelika Eggert[3], Alberto Garaventa[4], Claudio Gambini[5], Massimo Conte[4], Alessandra Eva[1], Luigi Varesio[1]

[1]Laboratory of Molecular Biology, Gaslini Institute, Genoa, Italy
[2]Department of Human Genetics, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands
[3]Department of Pediatric Oncology and Hematology, University Children's Hospital Essen, Essen
[4]Department of Hematology-Oncology, Gaslini Institute, Genoa, Italy
[5]Departments of Pediatric Pathology, Gaslini Institute, Genoa, Italy

Motivation and Objectives

Neuroblastoma (NB) is the most common pediatric solid tumor characterized by clinical and molecular risk factors. The mortality is about fifty percent and this makes exploration of new and more effective risk factors for improving stratification mandatory. Hypoxia is a condition of low oxygen tension occurring in poorly vascularized areas of the tumor associated with poor prognosis. We had previously defined a robust gene expression signature measuring the hypoxic component of NB tumors (NB-hypo) that is a novel, independent risk factor (Fardin et al., 2010). Integrating classical risk factors with NB-hypo could improve the stratification of NB patients. We wanted to develop a prognostic classifier of NB patients' outcome blending existing knowledge on clinical and molecular risk factors with the prognostic NB-hypo signature. Furthermore, we were interested in the decision tree classifier that outputs explicit rules easily translated into the clinical setting.

Methods

A total of 182 NB patients were enrolled on the bases of availability of gene expression profile by Affymetrix GeneChip HG-U133plus2.0, clinical and molecular information. NB tumor stage was defined according to the International NB Staging System (INSS). Age at diagnosis was dichotomized as greater or equal than 1 year and less than one year. MYCN status was amplified or normal. Good and poor outcome were defined as the patient's status alive or dead 5 years after diagnosis respectively. The risk group was assigned according to the International Neuroblastoma Risk Group (INRG) Consensus Pretreatment Classification Schema. The 182 NB patients cohort was clustered in High and low hypoxia by k-means analysis of the 62 probsets constituting the NB-hypo signature previously described to measure tumor hypoxia (Fardin et al., 2009). We utilized the k-means algorithm implemented in the WEKA software (Hall et al., 2009) setting up number of clusters to 2, 500 iterations, preserving instances order and using Manhattan distance.

The classification was performed by induction of decision trees. We utilized the Weka J48 implementation of the popular C4.5 algorithm (Kotsiantis, 2007; Murthy, 1998) and we set up the following options: pruning parameter was 0.25, pruning method was sub-tree raising and minimum number of instances per leaf was 2. Each leaf of the decision tree classifier identifies a non overlapping group of patients and each decision node identifies a branch which splits the dataset. We utilized Fisher's exact test to measure the statistical significance of groups and branches. Fisher's test was utilized in the context of decision trees to design a top-down approach to prune out non statistically significant branches (Liu et al., 2010). For each leaf we counted the number of correctly classified (named n) and the number of incorrectly classified instances (named m). We considered the marginal totals y and ¬y which represent poor and good outcome patients respectively. We designed a 2x2 contingency table of the two possible outcomes (Good or Poor) against the number of instances included in a give leaf and the remaining instances. Application of the Fisher exact test to this table generates a p value giving the probability of observing 2x2 table, or more extreme tables, knowing the marginal totals (y, ¬y) and assuming independency among the patients in a specific leaf and those in other leaves.

Results and Discussion

Patients were stratified in good and poor outcome on the bases of the following risk factors: Age at diagnosis, INSS stage, MYCN status and NB-hypo. The algorithm generated a decision tree classifier composed by 3 decision nodes

and 7 leaves covering 87% of non overlapping good outcome patients and 100% of non overlapping poor outcome patients. Each path from the root to a leaf utilizes some, but not all, considered risk factors. Interestingly NB-hypo was included in the decision node that stratified stage 3 tumors demonstrating its usefulness in NB patients' stratification.

The leaves classifying good outcome patients had the very low error of 2% indicating a good performance of the algorithm in predicting this class. In contrast, the classification of poor outcome patients produced the leaf with the highest error of 13%. This leaf includes stage 4 tumors that are traditionally difficult to stratify by any known risk factor.

To test statistical significance of the splits performed by the algorithm and the groups of patients, we utilized the Fisher's exact test with a confidence set at 95%. The results showed that the groups obtained at each split were statistically significant. Furthermore, analysis of single groups of patients identified by leaves demonstrated that some, but no all reached the significance threshold. The significant leaves included the NB-hypo among the utilized risk factors. These results further strengthen the value of NB-hypo in predicting patients' outcome. Lack of significance was often associated with a rather low number of patients in the leaf. This is a limitation of the divide-and-conquer approach of the algorithm, applied to relatively small patients' cohorts, that recurrently splits the dataset.

We then assessed the concordance of the predictions with INRG risk assessment. High Risk patients were correctly included in the leaves classifying poor outcome patients and Low Risk patients mapped correctly in the good outcome leaves. Interestingly, NB-hypo generated a leaf identifying a new group of poor outcome patients, sharing the high NB-hypo, whose characteristic fell into both High and Intermediate Risk.

We collected and analyzed the results of 1000 10-fold cross validations and we observed that most classifiers had 7 leaves and only 5 out of 10^4 deviated from this pattern. The recurrence of 7 leaves demonstrated the high stability of the decision tree classifier that we generated. Analysis of the pruning parameters revealed optimal performance in the range of 0.1-0.3 in line with what used in this study.

The path to reach each leaf can be easily transposed into a "if…than…" rule that, in turn provides an easy readout of the classifier, precious for translating the classification into the clinical setting. In conclusion, we demonstrated that the decision tree algorithm C4.5 can derive explicit rules for NB patients stratification if classical risk factors are blended with the NB hypo signature. These rules are statistically significant and quite stable and suitable to be conveyed to the clinic to design new therapies perhaps taking hypoxia into consideration as a potential target.

Acknowledgements

The work was supported by the Fondazione Italiana per la Lotta al Neuroblastoma, the Associazione Italiana per la Ricerca sul Cancro, the Società Italiana Glicogenosi, the Fondazione Umberto Veronesi and the Ministero della Salute Italiano. Davide Cangelosi and Fabiola Blengio are recipients of a fellowship from the Fondazione Italiana per la Lotta al Neuroblastoma.

References

1. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Varesio L. (2009) The l1-l2 regularization framework unmasks the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines. BMC Genomics, 10:474. doi:10.1186/1471-2164-10-474
2. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, et al. (2010) A biology-driven approach identifies the hypoxia gene signature as a predictor of the outcome of neuroblastoma patients, Journal of Molecular Cancer 9:1. 185. doi:10.1186/1476-4598-9-185
3. Hall M, Eibe F, Holmes G, Pfahringer B, Reutemann P, Witten IH. (2009) The WEKA Data Mining Software: An Update: SIGKDD Explorations. doi: 10.1145/1656274.1656278
4. Kotsiantis S.B. (2007) Supervised Machine Learning: A Review of Classification Techniques.
5. Informatica 31:249-268.
6. Liu W, Chawla S, Cieslak D, Chawla N. (2010) A Robust Decision Tree Algorithm for Imbalanced Data Sets. In: SDM. 766-777.
7. Murthy S.K. (1998) Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, Data Mining Knowledge Discovery 2:4. 345-389. doi: 10.1023/A:1009744630224