Tumor Identification by Gene Expression Profiles: A Comparison of Five Different Clustering Methods

A. Schuster[1], W. Dubitzky[2], F. J. Azuaje[2], M. Granzow[2], D. Berrar[2], R. Eils[2*]

[1]University of Ulster at Jordanstown, Northern Ireland; [2]Division "Intelligent Bioinformatics Systems" (H0900), German Cancer Research Center, 69120 Heidelberg, Germany; [3]Department of Computer Science, Oriel House, Trinity College, Dublin, Ireland

*corresponding author: r.eils@dkfz-heidelberg.de

Background: Tumors are generally classified by means of classical parameters such as clinical course, morphology and pathohistological characteristics. Nevertheless, the classification criteria obtained with these methods are not sufficient in every case. For example, it creates classes of cancer with significantly differing clinical courses or treatment response. As advanced molecular techniques are being established, more information about tumors is accumulated. One of these techniques, cDNA microarrays, is profiling the expression of up to many thousand genes in one single experiment of a tissue sample, e.g. a tumor. The derived data may contribute to a more precise tumor classification and prediction of clinical parameters such as prognosis or therapy response.

Method: The gene expression profiles of 72 patients diagnosed as either acute myeloid leukemia (AML) or acute lymphatic leukemia (ALL) [1] were taken to compare five different clustering methods in respect of their ability to divide this data set in clusters of corresponding cases. We applied the following machine learning methods to the expression data (except controls) without any processing:

1. Kohonen-clustering 4. K-means-clustering
2. Fuzzy-Kohonen-network [2] 5. Fuzzy-K-means-clustering [3]
3. Growing cell structures

We aimed to compare the possibilities of the different clustering methods
A) to reproduce the classification into the classes given in the data set (AML/ALL, sublasses of ALL, etc.)
B) to find further subclasses within the given groups
C) and to predict therapy response.

Results: The five clustering methods produced between 2 and 16 clusters, but only fuzzy-kohonen-network was successful in dividing the data set according to the respective gene expression profiles into clusters corresponding to biological classes. Best matches concerning the two classes AML and ALL was performed by clustering into 9 clusters. Here, 5 clusters contained solely ALL cases, one only AML cases, and in the remaining clusters there was only one mismatch (either AML or ALL). Concerning subclasses of ALL (B-cell or T-cell ALL) fuzzy-kohonen was able to generate 3 clusters of either B-cell ALL or T-cell ALL, in 4 clusters only one case mismatched, in the remaining there were 2 cases not corresponding. Further subclasses of the groups were not found. Due to the small number of cases none of the methods succeeded in clustering patients with similar treatment response.

Conclusions: Comparing five different methods of clustering biological data showed only one machine learning approach to be successful in reflecting the biological background available for the data set. Fuzzy-kohonen-clustering provided a quite accurate division of the data set into corresponding classes. After providing classes by clustering the next step is to identify the genes responsible for the clustering output, to deduct network dependencies between those genes and

to relate these networks to molecular genetic pathways. The results of this approach will be described in the final paper.

References:

[1] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439): 531-537, 1999.
[2] Huntsberger T.L. and Aijimarangsee P., 1992. Parallel self-organising feature maps for unsupervised pattern recognition. In: Bezdek J.C. and Pal N.R, Editors: Fuzzy models for pattern recognition, pp 483-495. IEEE Press, New York.
[3] Bezdek J.C., 1981. Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York, London.