Leukemia Prediction from Gene Expression Data---A Rough Set Approach

Jianwen Fang, Jerzy W. Grzymala-Busse, Information and Telecommunication Technology Center, University of Kansas, Lawrence, KS 66045. 2006 Annual Kansas City Area Life Sciences Research Day.

Leukemia Prediction from Gene Expression Data---A Rough Set Approach Jianwen Fang1,2 and Jerzy W. Grzymala-Busse1,3 1Bioinformatics Core Facility, 2Information and Telecommunication Technology Center, 3Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045 The development of MicroArray technology has motivated interest of its use in clinical diagnosis and drug discovery. The key step of these types of applications is to identify subsets of genes, often referred to as “biomarkers”, which distinguish cases with different labels, e.g., different tumor types, cancer versus non-cancer, response to therapy. In addition, these biomarkers are potential drug targets for treatment since they are relevant to the disease under study. Here we present our results on the prediction of leukemia from microarray data. Our methodology was based on data mining (rule induction) using rough set theory. The dataset used in the study was retrieved from public domain. There are 72 leukemia cases including 47 acute myeloid leukemia (AML) and 25 patients with lymphoblastic leukemia (ALL). It contains the expression levels of 6817 human genes measured by affymetrix high-density genechips. The data set was split into 38 training cases (27 ALL and 11 AML) and 34 testing cases (20 ALL and 14 AML). Distinguishing ALL from AML is critical for successful treatment of leukemia patients since the two types require chemotherapy to concentrate on different regimens. We used a novel methodology based on rule generations and cumulative rule sets. The final rule set contained only eight rules, using some combinations of eight genes. All cases from the training data set and all but one cases from the testing data set were correctly classified. Moreover, six out of eight genes found by us are well known in the literature as relevant to leukemia. Thus it is reasonable to believe these genes are potential biomarkers and worthy to further investigation. This work was partially Supported by the K-INBRE Bioinformatics Core, NIH grant number P20 RR016475.