

New Insights in Clinical Impact of Molecular Genetic Data by Knowledge-driven Data Mining

Daniel Berrar*,
Martin Granzow**+,
Werner Dubitzky

Intelligent Bioinformatics Systems
German Cancer Research Institute
Im Neuenheimer Feld 280
69120 Heidelberg, Germany.

{d.berrar, m.granzow,
w.dubitzky}@dkfz.de

Peter Lichter

Organization of Complex Genomes
German Cancer Research Institute
Im Neuenheimer Feld 280
69120 Heidelberg, Germany.

p.lichter@dkfz.de

Roland Eils⁺

Intelligent Bioinformatics Systems
German Cancer Research Institute
Im Neuenheimer Feld 280
69120 Heidelberg, Germany

r.eils@dkfz.de

and

⁺phase-it intelligent solutions AG
69115 Heidelberg, Germany

* both authors contributed equally to this paper.

ABSTRACT

Traditionally, classification of complex genetic diseases such as cancer has been performed on the basis of nonmolecular criteria such as tumor tissue type, pathological features, and clinical stage. It has been generally accepted that some patients grouped into a given category will have a certain survival prognosis and response to a particular therapy. While several studies have recently reported on the application of microarray gene expression analysis for molecular classification of cancer, attempts to integrate complex and heterogeneous molecular genetic data with clinical parameters are just at the beginning. Based on heterogeneous microarray gene expression and interphase cytogenetics data in combination with clinical patient data we have applied machine learning approaches for the classification of B-Cell chronic lymphocytic leukemia (B-CLL) patients into genetic risk groups. We believe this knowledge-driven approach is an important step forward for our capabilities to study complex functional relationships between molecular genetic and clinical data. It is bound to become a wide spread tool for future studies in clinical genomics.

1. INTRODUCTION

The identification of relevant information with biological importance has come to a new age with emerging technologies that provide the research community with vast amounts of data at comparatively short experimental time costs. Array approaches like cDNA, RNA, and protein chips accumulate information regarding gene expression levels and protein status, respectively, of different tissues including those of tumor origin that can hardly be investigated with standard biostatistical methods. As a consequence machine learning algorithms have recently been applied to process the complex data obtained by the aforementioned techniques. To name but a few, cluster algorithms like the well known Kohonen self-organizing maps [1], fuzzy clustering methods [2] for partitioning the data into subgroups of relevant classes, and artificial neural networks to address the classification of previously known groups based on the array data have been used.

Cancer classification was based on well established criteria long before array techniques were invented. These include clinical, pathomorphological, cytogenetic, and molecular cytogenetic aspects amongst others. Thus, investigating the expression status of genes in cancer tissue is an additional contribution to further examine the nature of the tumor. Furthermore, the systemic aspect of complex diseases like cancer can be addressed more sufficiently by taking all available data sources into account than by looking merely at gene expression profiles. To contribute to this, we performed an analysis of B-CLL patients that were characterized by clinical, molecular cytogenetic, and gene expression data. B-CLL is a cancer of the lymphatic tissue with a highly variable clinical course. For example, survival times show a range from a few months to more than 20 years.

Within the different forms of leukemia, B-CLL is the most frequent one in adults in the Western World. Prognostic relevance has only been shown for clinical stage and, recently, for genetic alterations (deletions of 17p13 and 11q22.3-11q23.1) [3, 4]. Two aspects of applying data mining techniques to complex data of heterogeneous sources are addressed in the following study. Firstly, a classification approach to predict the appropriate genetic risk group of the patients (for more details, see Section 2.1) has been performed on gene expression data. Secondly, a new association algorithm has been developed and applied to the data set in order to detect subsets of genes that are correlated with respect to their expression states in the genetic risk groups.

2. DATA COMPLEXITY PROBLEM

The analysis of microarray data is hampered by its characteristic complexity. In general, a typical data set is described by a $n \times m$ matrix of n patients and m gene expression levels. Typically, m is larger than n by a factor of 10 to 100, and the characterizing features are real number values. The following subsections describe the nature of the data used in this work and the analysis methods.

2.1 Data Description

The original data set included expression profiles (real values) of 1559 human DNA probes of 47 patients with B-CLL analyzed

with a microarray chip made by Incyte Pharmaceuticals, Inc. (USA) [5]. Based on fluorescence in situ hybridization (FISH) data for these patients and their correlation to survival time, four different genetic risk groups could be identified: (1) *del(17p)*, (2) *del(13qSingle)*, (3) *del(11q)*, and (4) *No aberrations* [6]. Each patient has been assigned to one genetic risk group. Table 1 shows the number of patients in each group and the survival chances that are correlated with these groups:

Table 1: The number of patients per genetic risk group and the correlated survival chances (fewer stars represent a lower survival chance).

Genetic Risk Group	Number of patients	Survival chances
<i>del(13qSingle)</i>	21	****
<i>No aberrations</i>	3	***
<i>del(11q)</i>	17	**
<i>del(17p)</i>	6	*

Before the data mining techniques were applied, the expression profiles are subject to a discretization step that produces three different symbolic values representing underexpressed, balanced, and overexpressed states. Furthermore, genes showing the same expression value in all 47 cases were excluded from further analysis, as they do not carry any discriminatory information with respect to the risk groups.

2.2 Basic Methodology

The basic analysis framework of this study is characterized by three distinct phases:

- (1) *data preprocessing*: Remove control genes and discretize real values in underexpressed, balanced, and overexpressed states.
- (2) *discriminant analysis*: Apply decision tree C5.0 to infer rules for the genetic risk groups.
- (3) *association analysis*: Apply association algorithm to identify subsets of genes that are underexpressed, overexpressed, or balanced in the genetic risk groups.

3. DATA PREPROCESSING

The gene expression profiles of the original data set are represented as absolute integral-numbered expression intensities. The decision tree algorithm used in this study is in principle able to handle continuous inputs. However, it is useful to distinguish between balanced expression, underexpression, and overexpression of genes. The cut-off levels of the expression profiles are not available, so that the gene expression profiles are discretized according to the following rules: (1) missing values are replaced by zero; (2) values greater than zero and smaller than (or equal to) 0.49 are considered as underexpressed, (3) values between 0.50 and 2.00 are considered as balanced, and (4) values greater than (or equal to) 2.01 are considered as overexpressed.

The choice of these cut-off levels is based on a visual inspection of the distribution of the expression profiles. Figure 1 depicts the discretization:

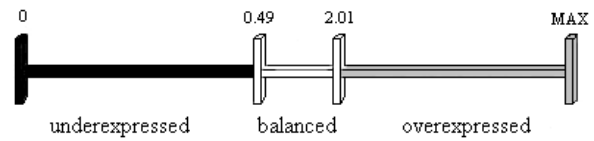


Figure 1: Discretization of the absolute gene expression profile data into underexpressed, balanced and overexpressed genes.

For all data preprocessing operations, proprietary algorithms, implemented with MATLAB 5.3 [7], have been used.

4. CLASSIFICATION

4.1 Decision Tree Algorithm

Decision trees are used for classification and prediction tasks and follow a kind of top-down, divide-and-conquer learning process. The working scheme of a decision tree algorithm can be described in the following way. The attribute that – based on an information gain measure – provides the best split of the cases with respect to the attribute to be predicted is selected as the root node of the tree. A branch for each possible value of the tree is generated from this root node, splitting the data set into subgroups. These steps are recursively repeated for each of the branches with only those cases that reach the respective branch. The algorithm stops the processing of a certain branch when all associated members were classified equally. These end nodes of a branch are hence called leaf nodes. The root node of a decision tree is regarded as the most important attribute with respect to the classification task. The importance of the following nodes is sequentially decreasing. Due to this, decision trees are capable of extracting rules by which the classification was achieved. In contrast to other widely used classification algorithms (e.g., artificial neural networks), these rules are understandable for humans.

The decision tree algorithm used in the present study is the powerful SPSS' Clementine [8] implementation of Ross Quinlan's C5.0 [9], the advanced successor of the well known C4.5 [10]. One of the major advantages of C5.0 is its capability to generate trees with a varying number of branches per node unlike other decision tree algorithms like CART that provide binary splits [11]. In order to improve the accuracy of a classifier, Clementine's C5.0 implements a cross-validation method called *boosting* [12]. This method maintains a distribution of weights over the data set, where initially each case is assigned the same weight. Those cases that were misclassified in the first classification process get a higher weight and the data set is classified again. This provides an accentuation of the hard-to-classify cases resulting in (1) an elevated accuracy of the classifier and (2) more than one rule set that denotes the classifier.

4.2 Classification Results

C5.0 was applied to the data set of 47 patients with B-CLL to predict the genetic risk group of each individual case. The estimated accuracy using 3-fold boosting was 100% meaning that with a model made up of these 3 rule sets, it was possible to predict each case within the data set correctly. The extracted rule sets identified a number of genes the algorithm recognized as important for the classification into the four genetic risk groups. The result of the first rule set has been visualized in Figure 2.

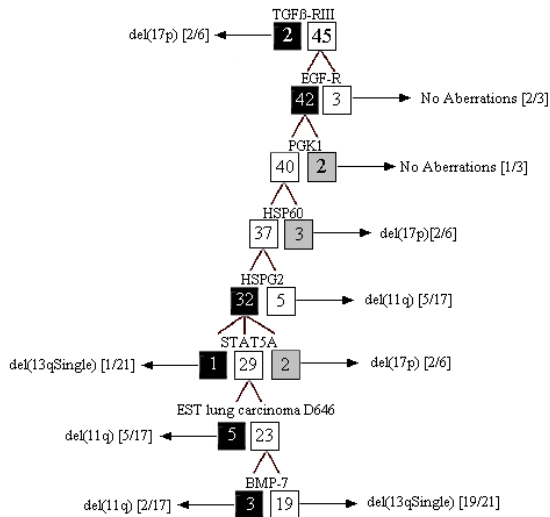


Figure 2: Visualization of the first rule set of the decision tree.

Presented is the first rule set of 3 comprising the prediction model. White boxes indicate a balanced gene expression state, black boxes underexpressed, and grey boxes overexpressed states, respectively. Abbreviations of genes are written on top of the respective boxes (TGFβ-RIII: transforming growth factor receptor type III; EGF-R: epidermal growth factor receptor; PGK-1: phosphoglycerate kinase 1; HSP60: chaperonin; HSPG2: heparansulfate proteoglycan; Stat5A: signal transducer and activator of transcription 5A; EST: estimated sequence tag; BMP-7: bone morphogenic protein 7). Numbers inside the boxes represent the number of cases that follow this rule. The numbers in brackets written behind the genetic risk groups include the number of cases of the respective group that follow this rule and the total number of cases within this group. The rule set in Figure 2 has to be read as follows. The root node TGFβ-RIII splits into balanced expression status of the gene counting 45 of the 47 cases in the whole data set (white box). The second split refers to the underexpressed status that holds 2 cases (black box). The first rule classifies 2 of the 6 cases of group *del(17p)* into this group and there is no other case where this rule applies in the whole data set. Of those cases where TGFβ-RIII is balanced, EGF-R is underexpressed in 42 cases and balanced in 3 cases. 2 of these 3 cases are covered by the rule “if TGFβ-RIII is balanced and EGF-R is balanced then classify to group *No aberrations*” which resemble 2 of all 3 cases in this genetic risk group. Thus, this very rule describes one additional case that does not belong to the group *No aberrations* but to another (which is *del(11q)*). Interestingly, 19 out of the 21 cases (90%) comprising the group *del(13qSingle)* are characterized by one rule with the root node TGFβ-RIII balanced and ending at the leaf node BMP-7 balanced.

The group *del(13qSingle)* is known to be the best with respect to the survival chances. Figure 3 depicts a Kaplan-Meier survival analysis of these 19 patients vs. all other patients:

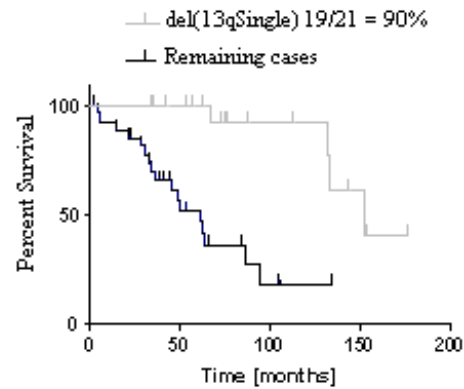


Figure 3: Kaplan-Meier survival analysis of 19 patients, following the rule with the root node TGFβ-RIII balanced and ending at the leaf node BMP-7 balanced (cf. Figure 2), vs. all other patients. (The curves are significantly different, $p = 0.0001$).

Every rule has to be read from the root node to its respective leaf node. Whenever the number in a box with an arrow pointing towards a genetic risk group is equal to the first number in brackets listed after the respective group the corresponding rule applies only to cases of this group. Furthermore, with the exception of 4 cases belonging to group *del(11q)*, every case is classified with the presented rule set. The remaining cases can be classified taking all three rule sets of the decision tree model together (data not shown).

As it is common in gene expression data sets the number of cases (in our study 47) is by far too low with respect to the attributes considered. Thus it was not suitable to split the data set into a training and a test set to which the model could have been applied in order to evaluate the strength of the rules learned from the training data. To address this limitation, we performed a 20-fold cross-validation, that divided the data set into 20 equally sized blocks according to the distribution of the cases whereby holding out a number of cases for testing. Thereafter a classifier was built upon each of the 20 reduced sets, and it was tested on the respective hold-out set. The cross-validation yielded a test accuracy of 40% (with a standard error of 6.8%).

The biological implications of decision tree results are non-trivial to interpret. On the one hand, you have to look at each of the genes that were found to be important to distinguish between the given groups. Table 2 gives a summary of genes in the three rule sets provided by C5.0. On the other hand, the genes highlighted by the classification algorithm can be seen on a more systemic view in context of the pathways they are involved in. An overlap of some pathways can be seen, e. g. genes encoding for EGF-R, GRB-2, and MAP2K2 are listed in Table 2. It has been shown that GRB-2 associates with EGF-R, and both gene products are entangled in the RAS-pathway, as is MAP2K2. Thus it is tempting to speculate whether the mentioned pathways do play a concerted role in B-CLL, which, of course, has to be verified by molecular biological experiments. This demonstrates the power of applying machine learning techniques to complex data sets, as the

results formulate hypotheses that can be validated by low-throughput biological experiments.

Table 2: Gene abbreviations, gene accession numbers (Access#), and keywords of biological role of genes found by the decision tree algorithm (PDGF-R: platelet derived growth factor receptor; n.p.: not provided).

Gene	Access#	Biological keywords
TGF β -RIII	L07594	Apoptosis
EGF-R	U48722	Apoptosis
PGK-1	n.p.	Glycolysis
HSP60	M34664	Stress factor
HSPG2	M85289	Stress factor
Stat5A	U43185	JAK/Stat pathway
BMP 7	X51801	Growth factor
AK2A	U39945	essential for maintenance and cell growth
PAFAH	n.p.	inactivates platelet-activating factor
bcl-2	M13994	Apoptosis regulator
PPP5	X89416	RNA biogenesis?
HIAP2/BIRC2	U45879	Apoptotic suppressor
GRB-2	L29511	EGF-R/PDGF-R pathway
MCP-1/SCYA2	n.p.	Chemotactic factor/augments monocyte anti-tumor activity
PDHA1	J03503	Pyruvate metabolism
PLAUR	U08839	mediates the signal transduction activation effects of urokinase plasmin
MAP2K2	U12779	Ras/Raf pathway
IGFBP4	U20982	Enhancer of apoptosis

In summary, Table 2 presents genes known to be involved in apoptosis, stress reaction, metabolism, and tumor relevant pathways despite a few not correlated to any of these categories. In addition to the study of Stratowa et al. [5] that found genes involved in lymphocyte trafficking to be of prognostic relevance in B-CLL patients using the same gene expression data set, the majority of the genes found in our study are located in tumor relevant pathways.

In conclusion, the consequences arising from the fact that the studied data set comprised only 47 patients have to lead to additional investigations with a higher number of patients involved. This would facilitate the learning process of the algorithm, and the model could be tested with unseen data. On the other hand, it can be hypothesized that those genes found by the decision tree algorithm may play a pivotal role in B-CLL.

5. ASSOCIATION

5.1 Maximum Association Algorithm

The goal of mining association rules in a data space is to derive multi-feature correlations between the attributes. Association algorithms associate a particular conclusion with a set of

conditions. In commercial applications, association rules can be used to determine what items are often purchased together by customers, and use that information to arrange, e.g., store layout. A typical rule in this domain is given by the following expression: “80% of the customers that purchase product X also purchase product Y .” Association rules differ from classification rules in that they can be used to predict any attribute and not just a class [13]. Furthermore, classification rules are intended to be used as a set. Association rules, on the other hand, express different intrinsic regularities in the data set, so that they can be used separately. The two most important measures of interest for association rules are the *coverage* (also called *support*) and the *accuracy* (also called *confidence*). The coverage of an association rule is the number of cases in which it is applicable (i.e. in which the antecedent – the *if*-clause – of the rule holds). The accuracy is the number of cases that the rule predicts correctly, expressed as a proportion of all cases it applies to (i.e. the number of cases in which the rule is correct relative to the number of cases in which it is applicable). Table 3 shows an example for association rules in a hypothetical gene expression data set:

Table 3: An example for association rules in a gene expression data set.

Patient ID	Genetic Risk Group	Gene_X	Gene_Y	Gene_Z
1	A	1	1	1
2	A	1	1	-1
3	A	0	1	0
4	B	1	1	0
5	B	-1	0	0

One association rule that can be derived from this data set is given by the following expression:

if Gene_X = 1 and Gene_Y = 1 then Genetic Risk Group = A
(coverage: 3 (0.6), accuracy: 2/3).

The *if*-clause of the rule applies three times, for the case #1, #2, and #4. Therefore, the coverage is 3 (or, relative to the number of all cases of the data set, 0.6). For case #1 and #2, the *then*-clause is correct, but for case #4, it is not. Consequently, the accuracy is 2/3. This example clearly illustrates that even from a tiny data set, a large amount of association rules can be derived. Therefore, only the “most interesting” rules, based on their coverage and accuracy, should be capitalized.

In our analysis, we were not mainly interested such association *rules*, but rather in associations of genes that have different expression states in the different genetic risk groups. For the gene expression data set, such an association could consist of the following statement: “In the genetic risk group *del(17p)*, Gene_X, Gene_Y, and Gene_Z are underexpressed in 100% of the cases, but in the group *del(13qSingle)*, they are overexpressed in 100% of the cases.” If a gene is over- or underexpressed in 100% of the cases of a genetic risk group A , we call this gene “totally overexpressed in A ”, respectively “totally underexpressed in A ”.

The advantage of association rule algorithms over decision tree algorithms is that associations can exist between any of the attributes. A decision tree algorithm will only build rules with a single conclusion, whereas association algorithms attempt to find

many rules, each with a different conclusion. On the other hand, associations may exist between a plethora of attributes, so that the search space for association algorithms can be very large. Therefore, association algorithms can require orders of magnitude more time to run than a decision tree algorithm. The *Apriori* algorithm [14], e.g., cannot reveal all possible associations because of the complexity of the search space. Therefore, we developed an alternative algorithm, called the *maximum association algorithm*, that is able to reveal all sets of associations that apply for 100% of the cases in one genetic risk group. This algorithm operates in four steps, each of them yielding interesting results.

In the first step, the algorithm screens the matrix of discretized expression data and identifies those genes that are either totally under- or totally overexpressed in one specific genetic risk group. To achieve this, the algorithm slides a window over all genes and all genetic risk groups. The following figure illustrates the procedure for the group *del(13qSingle)* and the gene #1. (Note that this is only a simplified example to illustrate the concept of the algorithm; the expression values in this example do not correspond to the real values in the data set of this study.)

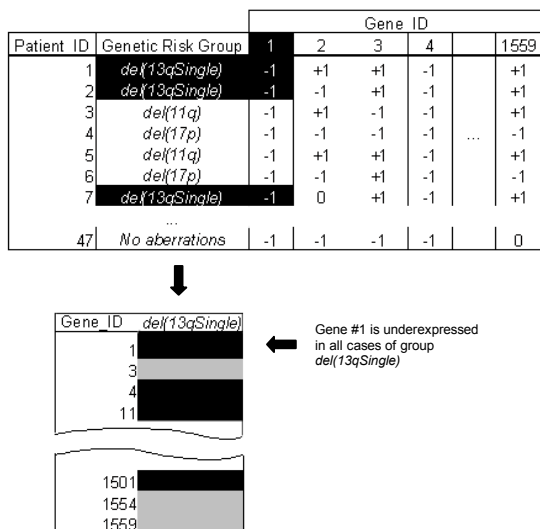


Figure 4: Selection of totally under- or overexpressed genes per genetic risk group. Underexpression is encoded by -1, balanced expression is encoded by 0, and overexpression is encoded by +1. Totally underexpressed genes are colored in black; totally overexpressed genes are colored in grey.

The sets of under- or overexpressed genes of one group are of course not necessarily disjoint with the sets of another group, for a specific gene can be underexpressed for all patients of a genetic risk group *A* and also for all patients of a group *B*.

The results of the first step of the maximum association algorithm have been stored in a cytogenetics database that has been developed for data mining purposes [15]. Via user-friendly graphical interfaces, a remote access to these results is possible, and even complex queries can be easily formulated. One example for such a query is the following: “Select all genes that are totally overexpressed in the genetic risk group *del(17p)*, totally underexpressed in the group *del(13qSingle)*, and neither totally expressed in *No aberrations* nor in *del(11q)*.”

In the second step, the algorithm eliminates those genes that are equally expressed in all genetic risk groups. If a specific gene is equally expressed in all groups, it has no discriminatory function, and hence it is removed. Figure 5 illustrates the elimination process. The arrows indicate which genes will be removed; here, gene #1, #4, #6, and #1555 will be excluded from further analysis:



Figure 5: Elimination of genes that are equally expressed in all genetic risk groups. Totally underexpressed genes are colored in black, and totally overexpressed genes are colored in grey. Genes that are neither totally under- nor totally overexpressed are colored in white.

In the third step, the algorithm operates as follows: if a specific gene is totally under- or totally overexpressed in a genetic risk group *A* but not in a group *B*, then the algorithm counts the number of cases in *B* for which this gene is balanced, the number of cases for which it is underexpressed, and the number of cases for which it is overexpressed. The expression state of this gene for the group *B* is then determined based on a majority vote: (1) if the number of cases for which this gene is underexpressed exceeds both the number of cases where the same gene is overexpressed and the number of cases where this gene is balanced, then this gene will be regarded as *underexpressed by the majority*; (2) if the number of cases for which this gene is overexpressed exceeds both the number of cases where the same gene is underexpressed and the number of cases where this gene is balanced, then this gene will be regarded as *overexpressed by the majority*; (3) if this gene is balanced in at least 50% of the cases, then it will be regarded as *balanced by the majority*.

For example, let gene #2 be underexpressed for 2 cases of the group *del(13qSingle)*, and let this gene be overexpressed in the remaining 19 cases. Then for this group, gene #2 will be regarded

as *overexpressed by the majority*. Figure 6 illustrates this operation:

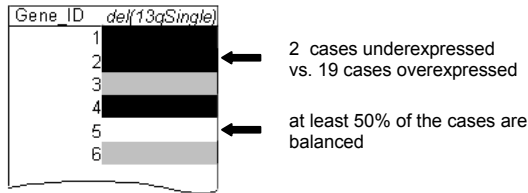


Figure 6: Determination of the gene expression state based on the majority vote.

After the operation in the third step, some genes can be equally expressed in all genetic risk groups. These genes are removed in the fourth step. This procedure is analogous to the operation described in the second step.

The maximum association algorithm has been developed with MATLAB 5.3 [7]. The analysis has been carried out on a standard PC in a very reasonable time.

5.2 Association Results

Table 4 summarizes the results of the maximum association algorithm after step 4:

Table 4: Results of the maximum association algorithm. Genes that are totally under- or overexpressed in one group are labeled explicitly. Genes balanced by majority (>50%) are colored in white. The genetic risk groups are encoded as follows: *A* = *del(13qSingle)*, *B* = *No aberrations*, *C* = *del(11q)*, and *D* = *del(17p)*. (n.p.: not provided).

Gene	Access#	A	B	C	D
epidermal growth factor receptor	n.p.	100%			100%
tyrosine phosphatase (Ch-1PTPase)	D64053		100%		
mitochondrial DNA	n.p.		100%		
ATPase coupling factor 6 subunit mitochondrial (ATP5A)	n.p.		100%		
Na,K-ATPase alpha-1 subunit	n.p.		100%		
guanidinoacetate N-methyltransferase	n.p.		100%		
<i>laminin B2 chain</i>	<i>J03202</i>		100%		100%
<i>oncoprotein 18 (p18, stathmin)</i>	<i>M31303</i>				100%
angiogenin	n.p.				100%
serum amyloid A SAA1 beta	M10906				100%
granulocyte colony-stimulating factor receptor (G-CSFR-1)	M59818				100%
15-hydroxyprostaglandin dehydrogenase (PDGH)	U63296				100%
laminin B1 chain	M61916				100%
CD40 ligand receptor	X60592				100%

In total, 14 genes “survived” the selective operations of the maximum association algorithm. The two most interesting genes are highlighted in Table 4. In the genetic risk groups *del(17p)* and in the group *No aberrations*, the gene with the accession number **J03202** is totally overexpressed, whereas it is overexpressed by the majority in the group *del(13qSingle)* and balanced by the majority in *del(11q)*. The gene identified by the accession number **M31303** is totally underexpressed in the group *del(17p)*, while it is balanced by the majority in all other groups.

6. DISCUSSION

When the number of features exceeds the number of observed cases, decision trees are prone to overfitting, i.e. the decision tree tends to encode the idiosyncrasies of the specific data set instead of inferring generalized rules. In this study, the number of attributes (1559 human DNA probes) exceeds by far the number of cases (47 patients). Consequently, it was not possible to improve the decision tree’s ability to generalize by splitting the data set into a training set and a test set. Therefore, we decided to perform a 20-fold cross-validation, that divided the data set into 20 equally sized blocks. In each cross-validation fold, a number of cases have been hold out for training, and another number of cases for testing. In the first cross-validation fold, each case had the same probability to fall into the training set or the test set. To those cases that have been misclassified in the *n*-th cross-validation fold was assigned a higher probability to fall into the training set of the (*n* + 1)-th fold. This procedure called *boosting* provides an accentuation of the hard-to-classify cases and results in a more precise and reliable classifier. The resulting model is fully satisfactory with a test accuracy of 40% (standard deviation of 6.8%).

Intelligent data analysis and data mining methods are extremely important for the present and future developments of systems biology. Molecular biologists are currently engaged in some of the most impressive data collection projects, for example, genome sequencing, gene expression profiling, and protein interaction analysis. These projects are generating an enormous amount of data related to structure, function, behaviour, and control of biological systems. The analysis and interpretation of this wealth of data will deeply affect and improve our understanding of biological systems and their underlying mechanisms. However, the elicitation and the representation of biological knowledge are extremely challenging tasks, which are demanding powerful and sophisticated data mining methodologies. Most widely used data mining software do not address the specific requirements of life science applications. The new association algorithm presented in this paper has been tailored for association mining in large data sets of gene expression data where even sophisticated methods like the *Apriori* algorithm would fail due to the complexity of the data.

7. ADDITIONAL AUTHORS

Stephan Stilgenbauer (Department of Internal Medicine III, University Hospital Ulm, Robert-Koch-Str. 3, 89081 Ulm, Germany. Email: stephan.stilgenbauer@medizin.uni-ulm.de). Klaus Wilgenbus (Department of Exploratory Research, Boehringer Ingelheim Austria, Dr. Boehringer Gasse 5-7, 1121 Vienna, Austria. Email: klaus.wilgenbus@vie.boehringer-ingelheim.com). Hartmut Döhner (Department of Internal Medicine III, University Hospital Ulm, Robert-Koch-Str. 3, 89081 Ulm, Germany. Email: hartmut.doehner@medizin.uni-ulm.de).

8. REFERENCES

- [1] Kohonen T. Self-organized formation of topologically correct feature maps. *Biol Cybern*, 43:59-69, 1982.
- [2] Granzow M., Berrar D., Dubitzky W., Schuster A., Azuaje F.J., Eils, R. Tumor Classification by Gene Expression Profiling: Comparison and Validation fo Five Clustering Methods. *ACM SIGBIO Newsletter*, vol. 21, no. 1: 16-22, April 2001.
- [3] Zwiebel J.A, Cheson B.D. Chronic lymphocytic leukemia: staging and prognostic factors. *Semin. Oncol.* 25, 42-59 (1998).
- [4] Julius G., Merup M. Cytogenetics in chronic lymphocyte leukemia. *Semin. Oncol.* 25, 19-26 (1998).
- [5] Stratowa C., Löffler G., Lichter P., Stilgenbauer S., Haberl P., Schweifer N., Döhner H., Wilgenbus, K.K. cDNA Microarray gene expression analysis of B-cell chronic lymphocytic leukemia proposes potential new prognostic markers involved in lymphocyte trafficking. *J Cancer* 91: 474-480, 2001.
- [6] Döhner H., Stilgenbauer S., Benner A., Leupolt E., Krober A., Bullinger L., Döhner K., Bentz M., Lichter P. Genomic aberrations and survival in chronic lymphocytic leukemia. *N Engl J Med* 2000 Dec 28;343(26):1910-6.
- [7] Mathworks MATLAB <http://www.mathworks.com/>.
- [8] SPSS Clementine. <http://www.spss.com/clementine>.
- [9] RuleQuest Research Data Mining Tools. <http://www.rulequest.com>
- [10] Quinlan J.R.. C4.5 : Programs for machine learning. Morgan Kaufmann, San Francisco, 1993.
- [11] Berry M.J., Linoff G. Data Mining Techniques For Marketing, Sales and Customer Support, John Wiley & Sons, Inc., New York, 1997.
- [12] Freund Y., Schapire R.E. A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Science*, 55(1): 119-139; 1997]
- [13] Witten I.H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Pub., San Francisco, 1999.
- [14] Agrawal R., Ramakrishnan S. Fast Algorithms for Mining Association Rules. *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, 1995.
- [15] Berrar D., Dubitzky W., Solinas-Toldo S., Bulashevsk S., Granzow M., Conrad C, Kalla K., Lichter P., Eils R. A Database for Comparative Genomic Hybridization Analysis. *IEEE Eng Med Biol Mag.* 20(4): 75-83, 2001.