Simple Rules Underlying Gene Expression Profiles of More than Six Subtypes of Acute Lymphoblastic Leukemia (ALL) Patients

Jinyan Li*
Laboratories for Information Technology
21 Heng Mui Keng Terrace, Singapore 119613
jinyan@lit.a-star.edu.sg

Huiqing Liu
Laboratories for Information Technology
21 Heng Mui Keng Terrace, Singapore 119613
huiqing@lit.a-star.edu.sg

James R. Downing
St Jude Children's Research Hospital
Memphis, TN 38105, USA
james.downing@stjude.org

Allen Eng-Juh Yeoh
National University of Singapore
5 Lower Kent Ridge Road, Singapore 119074
paeyej@nus.edu.sg

Limsoon Wong
Laboratories for Information Technology
21 Heng Mui Keng Terrace, Singapore 119613
limsoon@lit.a-star.edu.sg

Abstract

Motivations and Results: For classifying gene expression profiles or other types of medical data, simple rules are preferable to non-linear distance or kernel functions. This is because rules may help us understand more about the application in addition to performing an accurate classification. In this paper, we discover novel rules that describe the gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. We also introduce a new classifier, named PCL, to make effective use of the rules. PCL is accurate and can handle multiple parallel classifications. We evaluate this method by classifying 327 heterogeneous ALL samples. Our test error rate is competitive to that of support vector machines, and it is 71% better than C4.5, 50% better than Naive Bayes, and 43% better than k-nearest neighbour. Experimental results on another independent dataset are also presented to show the strength of our method.
Availability: Under `http://sdmc.lit.org.sg/GEDatasets/`, click on *Discovery of Emerging Patterns.*
Keywords: Gene expression profiles, rules, emerging patterns, discretization, classification, acute lymphoblastic leukemia.

INTRODUCTION

Childhood leukemia is a heterogeneous disease containing many subtypes. Since different subtypes of the disease respond differently to the same therapy, it is important to customise treatments for patients using risk-based stratification. The problem of accurate subtype classification and outcome prediction in pediatric acute lymphoblastic leukemia (ALL) has been studied previously (Yeoh et al., 2002) using microarray gene expression profiling and supervised machine learning algorithms. In this work, we use the data from (Yeoh et al., 2002) again, but we report many novel rules that describe the expression profiles of more than six subtypes of 327 ALL samples. We also introduce a new method that makes use of the discriminating power of the discovered rules for an excellent classification.

Classification and rule induction are two closely related but different aspects of supervised learning problems. Classification can be done with high accuracy by non-linear models like neural networks and support vector machines, but it it is hard for these models to provide rules. An ideal learning model should be accurate and at the same time can induce valid, novel, and useful rules, because such rules can

*Contact address: Laboratories for Information Technology, 21 Heng Mui Keng Terrace, Singapore 119613. Tel. +65-6874-8217. Fax. +65-6774-8056.

help in deeper understanding of the problem at hand. We propose a classifier called PCL, which is short for Prediction by Collective Likelihoods. PCL is accurate and can handle multiple parallel classifications while providing valid and useful rules.

PCL is based on the concept of *emerging patterns* (Dong & Li, 1999). An emerging pattern (EP) is a pattern—in this case, the expression intervals of multiple genes—whose *frequency* increases significantly from one class of data to another class of data. For example, if a pattern has 10% frequency in one class and 0.1% frequency in another class, then it is an emerging pattern with a 100-fold frequency change between the two classes. These emerging patterns are:

- *useful* because they can be combined into reliable rules for predicting class labels,
- *understandable* because they are conjunctive conditions on the underlying feature values, and
- *non-obvious* because they involve multiple features and their correlations.

Additionally, the validity of the patterns can be conveniently assessed by cross validations and blinded samples. In this work, we focus on a special type of emerging patterns: those with *zero* frequency in one class but high frequency in another class. These patterns are naturally more interesting because they have an infinite frequency change rate.

With the discovery of emerging patterns, PCL proceeds to calculate a classification score for every class when a test sample is presented, and the class with the highest score is predicted. The classification scores are calculated by aggregating the frequencies of multiple top-ranked EPs: the "committee" of patterns and their collective discriminating power show strong strength in this study.

In one of our previous work (Li & Wong, 2002), we have used the concept of emerging patterns to discover diagnostic gene groups from a colon tumor dataset (Alon et al., 1999) and a leukemia dataset (Golub et al., 1999). The colon data have 62 samples with two classes, and the leukemia data have 38 training and 34 test samples with two classes as well. In this study, we are aiming to see the effectiveness of the concept when applied to large datasets: the data used in this study have hundreds of samples with multiple classes. Furthermore, the newly proposed PCL classifier aims to test the reliability of the patterns.

In the following, we describe the data in detail, then we discuss an entropy-based discretization method (Fayyad & Irani, 1993) for selecting the most important genes. This method is different from the widely used signal-to-noise metrics (Golub et al., 1999). While the latter is based on the expression average and deviation of the genes between two classes of cells (ie., normal and tumor cells), the former is based on the boundaries of the expression values of those cells. Our emerging patterns are discovered from the discretized data. Then we present novel rules for describing hierarchical tree-structured gene expression profiles of the ALL samples. These rules are derived directly from our emerging patterns. Then we evaluate the reliability of the rules, develop a proposal for strengthening the system, and perform an accuracy comparison with other prediction models. We finally report the PCL's multiple-class parallel classification performance on Yeoh et al.'s data and on another independent dataset (Armstrong et al., 2002).

DATA

The data are available at `http://www.stjuderesearch.org/data/ALL1`. The data consist of gene expression profiles of 327 ALL samples (Yeoh et al., 2002). The profiles were obtained by hydridization on the Affymetrix U95A GeneChip containing probes for 12558 genes. The samples contain all the known acute lymphoblastic leukemia subtypes, including T-cell (T-ALL), E2A-PBX1, TEL-AML1, MLL, BCR-ABL, and hyperdiploid (Hyperdip>50). The data were divided into a *training set* of 215 samples and a blind *testing set* of 112 samples (Yeoh et al., 2002). There are 28, 18, 52, 9, 14, and 42 training instances and 15, 9, 27, 6, 6, and 22 test samples respectively for T-ALL, E2A-PBX1, TEL-AML1, BCR-ABL, MLL, and Hyperdip>50. There are also 52 training and 27 test samples of other miscellaneous subtypes called OTHERS here.

DISCOVERY OF EMERGING PATTERNS

We next discover emerging patterns from the training data. The discovery of emerging patterns is a two-step process (Li & Wong, 2002): first, a small number of discriminatory genes are selected from among the 12558 genes; second, emerging patterns based on the selected genes are derived by a naive method or by our efficient border-based algorithms (Dong & Li, 1999; Li et al., 2000).

Table 1: Five ideally discriminatory genes discovered from the training data for differentiating E2A-PBX1 from all other subtypes. The cut points are boundaries separating C1 and C2 samples clearly. Here C1 = E2A-PBX1 and C2 = other subtypes. The expression distribution of the test data are shown in parenthesis.

Probes	# of samples in the left interval $[0, x)$ C1 + C2	cut points (x)	# of samples in the right interval $[x, +\infty)$ C1 + C2
32063_*at*	$0 + 197$ ($0 + 103$, test)	4068.7	$18 + 0$ ($9 + 0$, test)
41146_*at*	$0 + 197$ ($1 + 102$, test)	25842.15	$18 + 0$ ($8 + 1$, test)
33355_*at*	$0 + 197$ ($0 + 103$, test)	10966	$18 + 0$ ($9 + 0$, test)
1287_*at*	$0 + 197$ ($1 + 103$, test)	34313.9	$18 + 0$ ($8 + 0$, test)
430_*at*	$0 + 197$ ($0 + 101$, test)	30246.05	$18 + 0$ ($9 + 2$, test)

Table 2: Five sub-optimal genes which can not be individually used to differentiate E2A-PBX1 from other subtypes completely, but their combinations can do. Here C1 = E2A-PBX1 and C2 = other subtypes.

Probes	# of samples in the left interval $[0, x)$ C1 + C2	cut points (x)	# of samples in the right interval $[x, +\infty)$ C1 + C2
40454_*at*	0+187 (0+96, test)	8280.25	18+10 (9+7, test)
41425_*at*	0+183 (0+93, test)	6821.75	18+14 (9+10, test)
753_*at*	2+197 (6+102, test)	7106.35	16+0 (3+1, test)
35974_*at*	0+189 (0+99, test)	43730.15	18+8 (9+4, test)
37493_*at*	1+189 (0+95, test)	4543.55	17+8 (9+8, test)

Step 1: Selecting Genes and Discretizing Genes' Expression Ranges

An entropy-based discretization method (Fayyad & Irani, 1993) is used to select the important features. This method can automatically remove about 90-95% of the whole feature space as those features exhibit random expression distributions. It can also automatically detect "ideally discriminatory genes", that contain clear boundaries separating two classes of cells.

Single genes whose expressions have a two-end distribution with a pure class of cells at each end are very useful for classification. For example, we found five such ideally discriminatory genes (Table 1) from the training data pair, E2A-PBX1 versus all other subtypes of ALL.

For example, the *cut point* of the gene 32063_*at* is 4068.7. This cut point partitions the expression range of this gene into two intervals, $[0, 4068.7)$ and $[4068.7, +\infty)$. Observe that this gene's expression of all E2A-PBX1 samples were ≥ 4068.7, falling into the right interval; however its expression of any other subtype samples was less than 4068.7, falling into the left interval. This rule was valid to the 215 training samples without any exception. This rule was also 100% valid when applied to the reserved 112 testing ALL samples (see Table 1).

However, the other genes did not exhibit the nice distribution. In contrast, about 11840 genes (94.28% of the whole feature space) had a random expression distribution without any interval covering a sufficient percentage of E2A-PBX1 or other subtypes of samples. Obviously, these genes were not relevant for differentiating E2A-PBX1 from other subtypes, and we can exclude them from our consideration.

The remaining 713 genes were sub-optimal. Their expression were not randomly distributed. But they were not ideally polarized either: there did not exist a cut point such that the resulting two intervals each contain a pure class of cells. We list 5 of them in Table 2.

The method (Fayyad & Irani, 1993) can find the "best" cut points for every feature, and then can identify the ideally discriminatory features, sub-optimal features, and those genes with random expression distributions efficiently. In this work, the method ignored most of the 12558 genes: Only about 1000 genes were considered useful in classification. The 10% selection rate provides a much easier platform to derive important rules. However, to manually examine 1000 or so genes is still tedious. We then use the *Chi-Squared* ($\mathcal{X}^2$) method (Liu & Setiono, 1995; Witten & Frank, 2000) to further narrow our search for the important genes. In this study, if there exist ideally discriminatory genes for a pair of training datasets, then we use those ideal genes for deriving emerging patterns. Otherwise we use 20 top-ranked genes by the $\mathcal{X}^2$ method.

Step 2: Generating EPs

From the data, we found very few genes that can act individually as an arbitrator by itself alone to distinguish one subtype from others. However, we found many gene groups which can be used to separate clearly one subtype from others. For example, an ideal multi-gene discriminator can be obtained if the gene 40454_*at* and the gene 41425_*at* (shown in Table 2) are combined into a group. The rule can be interpreted as:

If the expression of 40454_*at* is $\geq$ 8280.25 and the expression of 41425_*at* is $\geq$ 6821.75, then this sample was an E2A-PBX1. Otherwise it was one of other subtypes.

Mathematically, we denote the rule as the following *set* by combining the two genes together with the conditions (Li & Wong, 2002):

$$\{40454\_at@[8280.25, +\infty), 41425\_at@[6821.75, +\infty)\}.$$

Such patterns are called *emerging patterns* (Dong & Li, 1999), more precisely, a special type of emerging pattern known as "jumping" emerging patterns. As 100% of the E2A-PBX1 training samples satisfy these two conditions, we say the pattern's *frequency* in the E2A-PBX1 class is 100%. As a discriminator, the pattern's frequency in the other subtypes is 0%. That is, no other subtype sample satisfies these two conditions.

Multi-gene discriminators (the EPs) can be discovered by a level-wise approach in a naive manner. Firstly, combine all possible pairs of genes into two-gene groups. For a group, assign each gene with an interval to form a candidate pattern. So, a gene group can spawn off at least 4 candidate patterns in the form of $\{gene\_a@[0, x), gene\_b@[y, 45000]\}$. Then, calculate the percentage of a *class* of samples that satisfy the two conditions of the pattern. If the frequency in one class is non-zero, but in the other class is 0, then this pattern is an EP (also a multi-gene discriminator). The next level of candidate patterns are produced by adding one gene into the previous two-gene groups, except those two-gene discriminators. Repeat this process until the candidate set is empty.

We overcome the heavy computational cost of the naive method described above by (i) focusing on top 20 ranked genes and by (ii) using border-based algorithms (Dong & Li, 1999; Li et al., 2000) to discover EPs. The algorithms do not need to generate all candidate patterns. Instead, they directly produce EPs in an efficient manner.

DERIVING RULES FROM A TREE-STRUCTURED LEUKEMIA DATASET

In a previous study, a tree-structured decision system was used to classify ALL samples. See Figure 19 in Supplementary Information of (Yeoh et al., 2002) in the web site `http://www.stjuderesearch.org/data/ALL1`. By using the tree, a sample is first tested to be either a T-ALL or a sample of other subtypes. If it is classified as T-ALL, then the process is terminated. Otherwise, the process is moved to level 2 to see whether the sample can be classified as E2A-PBX1 or the *remaining* other subtypes. Similarly, the system can demine whether the test sample is one of TEL-AML1, BCR-ABL, MLL, Hyperdip>50, or OTHERS.

In accordance with this hierarchical decision system, six pairs of training and test subsets, one for each level of the tree, were generated from the original training and test data (Yeoh et al., 2002). Our rules are discovered from each level's training data; they are then applied to the test data and the error rates of our PCL classifier are computed.

This tree-structured classification scheme was suggested by the doctors of our team (Yeoh and Downing) who have extensive domain knowledge. Alternative perspectives are possible: in later sections of this paper, we will show how to classify the 327 samples in a parallel manner instead of hierarchically. In that case, the original training and test data do not need to be subdivided and layered.

Rules for T-ALL vs OTHERS1

For the first pair of training datasets, T-ALL vs OTHERS1, we found one and only one ideally discriminatory gene, 38319_*at*. Here, OTHERS1 ={E2A-PBX1, TEL-AML1, BCR-ABL, Hyperdip>50, MLL, OTHERS}. The discretization method partitioned the expression range of this gene into two intervals:

$(-\infty, 15975.6)$ and $[15975.6, +\infty)$. Using the EP discovery algorithm, we derived two EPs:

$$\{gene_{\text{-}(38319\_at)}@(-\infty, 15975.6)\} \text{ and } \{gene_{\text{-}(38319\_at)}@[15975.6, +\infty)\}.$$

The former has a 100% frequency in the T-ALL class but a zero frequency in the OTHERS1 class. The latter has a zero frequency in the T-ALL class, but a 100% frequency in the OTHERS1 class. Therefore, we have the following rule:

If the expression of 38319_*at* is less than 15975.6, then this ALL sample must be a T-ALL. Otherwise it must be a subtype in OTHERS1.

This simple rule regulates the 215 (28 T-ALL plus 187 OTHERS1) ALL training samples without any exception.

Rules for E2A-PBX1 vs OTHERS2

We also have a simple rule for E2A-PBX1 vs OTHERS2. Here, OTHERS2 = {TEL-AML1, BCR-ABL, Hyperdip>50, MLL, OTHERS}. Our method picked one gene, 33355_*at*, and discretized it into two intervals: $(-\infty, 10966)$ and $[10966, +\infty)$. Then $\{gene_{\text{-}(33355\_at)}@(-\infty, 10966)\}$ and $\{gene_{\text{-}(33355\_at)}@[10966, +\infty)\}$ were found to be EPs with 100% frequency in E2A-PBX1 and OTHERS2 respectively. So, a rule for these 187 (18 E2A-PBX1 plus 169 OTHERS2) ALL samples would be:

If the expression of 33355_*at* is less than 10966, then this ALL sample must be a E2A-PBX1. Otherwise it must be a subtype in OTHERS2.

Rules through Level 3 to Level 6

For the remaining four pairs of datasets, we did not find any ideally discriminatory gene. So, we used the $\mathcal{X}^2$ method to select 20 top-ranked genes for each of the four pairs of datasets. Table 3, Table 4, Table 5, and Table 6 list the names of the selected genes, their partitions, and our index to the intervals for the four pairs of datasets. As the index matches and joins the genes' name and their intervals, it is very convenient for us to write and read EPs using the index numbers. (These probes' accession numbers and descriptions have been reported in Supplementary Information of (Yeoh et al., 2002), but the cut points have not been published anywhere.)

After discretizing the selected genes, we discovered two groups of EPs for each of the four pairs of datasets. Table 7 shows the numbers of our discovered emerging patterns. The fourth column of Table 7 shows that the number of the discovered EPs is relatively large. We use another four tables, Table 8, Table 9, Table 10, and Table 11 to list 10 top EPs according to their frequency. The frequency of these top-10 EPs can reach 98.94% and most of them are around 80%. Even though a top-ranked EP may not cover an entire class of samples, it dominates the whole class. Their absence in the counterpart classes demonstrates that top-ranked emerging patterns can capture the nature of a class.

Where are the rules? In the following, we interpret EPs into rules. For example, the first EP of the TEL-AML1 class is $\{2, 33\}$. According to the index in Table 3, the reference number 2 in this EP matches the right interval of the gene 38652_*at*, and stands for the condition that: the expression of 38652_*at* is larger than or equal to 8997.35. Similarly, the reference number 33 matches the left interval of the gene 36937_*s_at*, and stands for the condition that: the expression of 36937_*s_at* is less than 13617.05. Thus the pattern $\{2, 33\}$ says that 92.31% of the TEL-AML1 class (48 out of the 52 samples) satisfy the two conditions above, but no single sample from OTHERS3 satisfies both of these conditions.

CLASSIFICATION BY PCL ON THE TREE-STRUCTURED DATA

In the previous section, we have found many novel non-obvious rules that can well describe the gene expression profiles of more than six subtypes of ALL samples. An important methodology to test the reliability of the rules is to apply them to previously unseen samples (the so-called blind testing samples). Recall that we have previously reserved 112 blind testing samples. Our testing results are highlighted as follows: At level 1, all the 15 T-ALL samples are correctly predicted as T-ALL; all the 97 OTHERS1 samples are correctly predicted as OTHERS1. At level 2, all the 9 E2A-PBX1 samples are correctly predicted as E2A-PBX1; all the 88 OTHERS2 samples are correctly predicted as OTHERS2. For levels 3 to 6, we misclassified only 4-7 samples depending on the number of EPs used. With more EPs, the error rate decreased. Our method PCL is proposed as follows.

Table 3: The top 20 genes selected by the $\mathcal{X}^2$ method from TEL-AML1 vs OTHERS3. Here, OTHERS3 = {BCR-ABL, Hyperdip>50, MLL, OTHERS}. The intervals produced by the entropy method and our index to the intervals are listed in columns 2 and 3.

Probes	Intervals	Index to Intervals
38652_*at*	$(-\infty, 8997.35)$, $[8997.35, +\infty)$	1,2
36239_*at*	$(-\infty, 14045.5)$, $[14045.5, 16328.55)$, $[16328.55, +\infty)$	3,4,5
41442_*at*	$(-\infty, 15114.1)$, $[15114.1, 26083.95)$, $[26083.95, +\infty)$	6,7,8
37780_*at*	$(-\infty, 2396.3)$, $[2396.3, 5140.5)$, $[5140.5, +\infty)$	9,10,11
36985_*at*	$(-\infty, 19499.6)$, $[19499.6, 26571.05)$, $[26571.05, +\infty)$	12,13,14
38578_*at*	$(-\infty, 7788.95)$, $[7788.95, +\infty)$	15,16
38203_*at*	$(-\infty, 3721.3)$, $[3721.3, +\infty)$	17,18
35614_*at*	$(-\infty, 9930.15)$, $[9930.15, +\infty)$	19,20
32224_*at*	$(-\infty, 5740.45)$, $[5740.45, +\infty)$	21,22
32730_*at*	$(-\infty, 2864.85)$, $[2864.85, +\infty)$	23,24
35665_*at*	$(-\infty, 5699.35)$, $[5699.35, +\infty)$	25,26
1077_*at*	$(-\infty, 22027.55)$, $[22027.55, +\infty)$	27,28
36524_*at*	$(-\infty, 1070.65)$, $[1070.65, +\infty)$	29,30
34194_*at*	$(-\infty, 1375.85)$, $[1375.85, +\infty)$	31,32
36937_*s_at*	$(-\infty, 13617.05)$, $[13617.05, +\infty)$	33,34
36008_*at*	$(-\infty, 11675.35)$, $[11675.35, +\infty)$	35,36
1299_*at*	$(-\infty, 3647.7)$, $[3647.7, 9136.35)$, $[9136.35, +\infty)$	37,38,39
41814_*at*	$(-\infty, 6873.85)$, $[6873.85, +\infty)$	40,41
41200_*at*	$(-\infty, 11030.5)$, $[11030.5, +\infty)$	42,43
35238_*at*	$(-\infty, 4774.85)$, $[4774.85, 7720.4)$, $[7720.4, +\infty)$	44,45,46

Table 4: The top 20 genes selected by the $\mathcal{X}^2$ method from the data pair BCR-ABL vs OTHERS4. Here, OTHERS4 = {Hyperdip>50, MLL, OTHERS}.

Probes	Intervals	Index to Intervals
1637_*at*	$(-\infty, 5242.15)$, $[5242.15, +\infty)$	1,2
36650_*at*	$(-\infty, 13402)$, $[13402, +\infty)$	3,4
40196_*at*	$(-\infty, 2424.4)$, $[2424.4, +\infty)$	5,6
1635_*at*	$(-\infty, 5279.3)$, $[5279.3, +\infty)$	7,8
33775_*s_at*	$(-\infty, 1130.75)$, $[1130.75, +\infty)$	9,10
1636_g_*at*	$(-\infty, 11112.9)$, $[11112.9, +\infty)$	11,12
41295_*at*	$(-\infty, 33488.7)$, $[33488.7, +\infty)$	13,14
37600_*at*	$(-\infty, 24168.95)$, $[24168.95, +\infty)$	15,16
37012_*at*	$(-\infty, 18127.7)$, $[18127.7, +\infty)$	17,18
39225_*at*	$(-\infty, 14137.25)$, $[14137.25, +\infty)$	19,20
1326_*at*	$(-\infty, 3273.55)$, $[3273.55, +\infty)$	21,22
34362_*at*	$(-\infty, 13254.9)$, $[13254.9, +\infty)$	23,24
33150_*at*	$(-\infty, +\infty)$	25
40051_*at*	$(-\infty, +\infty)$	26
39061_*at*	$(-\infty, +\infty)$	27
33172_*at*	$(-\infty, +\infty)$	28
37399_*at*	$(-\infty, +\infty)$	29
317_*at*	$(-\infty, +\infty)$	30
40953_*at*	$(-\infty, 2569.55)$, $[2569.55, +\infty)$	31,32
330_*s_at*	$(-\infty, 6237.5)$, $[6237.5, +\infty)$	33,34

Table 5: The top 20 genes selected by the $\mathcal{X}^2$ method from MLL vs OTHERS5. Here, OTHERS5 = {Hyperdip>50, OTHERS}.

Probes	Intervals	Index to Intervals
34306_at	$(-\infty, 12080.7)$, $[12080.7, +\infty)$	1,2
40797_at	$(-\infty, 5331.15)$, $[5331.15, +\infty)$	3,4
33412_at	$(-\infty, 29321.15)$, $[29321.15, +\infty)$	5,6
39338_at	$(-\infty, 5813.1)$, $[5813.1, +\infty)$	7,8
2062_at	$(-\infty, 10476.05)$, $[10476.05, +\infty)$	9,10
32193_at	$(-\infty, 2605.6)$, $[2605.6, +\infty)$	11,12
40518_at	$(-\infty, 23228.2)$, $[23228.2, +\infty)$	13,14
36777_at	$(-\infty, 5873.9)$, $[5873.9, +\infty)$	15,16
32207_at	$(-\infty, 7238.8)$, $[7238.8, +\infty)$	17,18
33859_at	$(-\infty, 23053.2)$, $[23053.2, 24674.9)$, $[24674.9, +\infty)$	19,20,21
38391_at	$(-\infty, 16251.65)$, $[16251.65, +\infty)$	22,23
40763_at	$(-\infty, 3301.3)$, $[3301.3, +\infty)$	24,25
1126_s_at	$(-\infty, 6667.6)$, $[6667.6, +\infty)$	26,27
34721_at	$(-\infty, 8743.05)$, $[8743.05, +\infty)$	28,29
37809_at	$(-\infty, 2705.75)$, $[2705.75, +\infty)$	30,31
34861_at	$(-\infty, 4780)$, $[4780, 5075.05)$, $[5075.05, +\infty)$	32,33,34
38194_s_at	$(-\infty, 859.2)$, $[859.2, 6860.6)$, $[6860.6, +\infty)$	35,36,37
657_at	$(-\infty, 8829.8)$, $[8829.8, +\infty)$	38,39
36918_at	$(-\infty, 5321.15)$, $[5321.15, +\infty)$	40,41
32215_i_at	$(-\infty, 2464.1)$, $[2464.1, +\infty)$	42,43

Table 6: The top 20 genes selected by the $\mathcal{X}^2$ method from the data pair Hyperdip>50 vs OTHERS.

Probes	Intervals	Index to Intervals
36620_at	$(-\infty, 16113.1)$, $[16113.1, +\infty)$	1,2
37350_at	$(-\infty, 10351.95)$, $[10351.95, +\infty)$	3,4
171_at	$(-\infty, 6499.25)$, $[6499.25, +\infty)$	5,6
37677_at	$(-\infty, 41926.9)$, $[41926.9, +\infty)$	7,8
41724_at	$(-\infty, 20685.45)$, $[20685.45, +\infty)$	9,10
32207_at	$(-\infty, 15242.9)$, $[15242.9, +\infty)$	11,12
38738_at	$(-\infty, 15517.2)$, $[15517.2, +\infty)$	13,14
40480_s_at	$(-\infty, 4591.95)$, $[4591.95, +\infty)$	15,16
38518_at	$(-\infty, 13840)$, $[13840, +\infty)$	17,18
41132_r_at	$(-\infty, 10490.95)$, $[10490.95, +\infty)$	19,20
31492_at	$(-\infty, 17667.05)$, $[17667.05, +\infty)$	21,22
38317_at	$(-\infty, 4982.05)$, $[4982.05, +\infty)$	23,24
40998_at	$(-\infty, 11962.6)$, $[11962.6, +\infty)$	25,26
35688_g_at	$(-\infty, 3340.55)$, $[3340.55, +\infty)$	27,28
40903_at	$(-\infty, 3660.4)$, $[3660.4, +\infty)$	29,30
36489_at	$(-\infty, 6841.95)$, $[6841.95, +\infty)$	31,32
1520_s_at	$(-\infty, 10334.05)$, $[10334.05, +\infty)$	33,34
35939_s_at	$(-\infty, 9821.95)$, $[9821.95, +\infty)$	35,36
38604_at	$(-\infty, 13569.7)$, $[13569.7, +\infty)$	37,38
31863_at	$(-\infty, 8057.7)$, $[8057.7, +\infty)$	39,40

Table 7: Total number of EPs discovered from the four pairs of training datasets.

Dataset pair ($\mathcal{D}_1$ vs $\mathcal{D}_2$)	Number of EPs in $\mathcal{D}_1$	Number of EPs in $\mathcal{D}_2$	Total
TEL-AML1 vs OTHERS3	2178	943	3121
BCR-ABL vs OTHERS4	101	230	313
MLL vs OTHERS5	155	597	752
Hyperdip>50 vs OTHERS	2213	2158	4371

Table 8: Ten most frequent EPs in the TEL-AML and OTHERS3 classes.

EPs { index numbers }	% frequency in TEL-AML1	% frequency in OTHERS3	EPs { index numbers }	% frequency in TEL-AML1	% frequency in OTHERS3
$\{2, 33\}$	92.31	0.00	$\{1, 23, 40\}$	0.00	88.89
$\{16, 22, 33\}$	90.38	0.00	$\{17, 29\}$	0.00	88.89
$\{20, 22, 33\}$	88.46	0.00	$\{1, 17, 40\}$	0.00	88.03
$\{5, 33\}$	86.54	0.00	$\{1, 9, 40\}$	0.00	88.03
$\{22, 28, 33\}$	84.62	0.00	$\{15, 17\}$	0.00	88.03
$\{16, 33, 43\}$	82.69	0.00	$\{1, 23, 29\}$	0.00	87.18
$\{22, 30, 33\}$	82.69	0.00	$\{17, 25, 40\}$	0.00	87.18
$\{2, 36\}$	82.69	0.00	$\{17, 23, 40\}$	0.00	87.18
$\{20, 43\}$	82.69	0.00	$\{9, 17, 40\}$	0.00	87.18
$\{22, 36\}$	82.69	0.00	$\{1, 9, 29\}$	0.00	87.18

Table 9: Ten most frequent EPs in the BCR-ABL and OTHERS4 classes.

EPs { index numbers }	% frequency in BCR-ABL	% frequency in OTHERS4	EPs { index numbers }	% frequency in BCR-ABL	% frequency in OTHERS4
$\{22, 32, 34\}$	77.78	0.00	$\{3, 5, 9\}$	0.00	95.37
$\{8, 12\}$	77.78	0.00	$\{3, 9, 19\}$	0.00	95.37
$\{4, 8, 34\}$	66.67	0.00	$\{3, 15\}$	0.00	95.37
$\{4, 8, 22\}$	66.67	0.00	$\{3, 13\}$	0.00	95.37
$\{6, 34\}$	66.67	0.00	$\{3, 5, 23\}$	0.00	93.52
$\{8, 24\}$	66.67	0.00	$\{11, 17, 19\}$	0.00	93.52
$\{24, 32\}$	66.67	0.00	$\{3, 19, 23\}$	0.00	93.52
$\{4, 12\}$	66.67	0.00	$\{7, 19\}$	0.00	93.52
$\{8, 32\}$	66.67	0.00	$\{11, 15\}$	0.00	93.52
$\{12, 34\}$	66.67	0.00	$\{5, 11\}$	0.00	93.52

Table 10: Ten most frequent EPs in the MLL and OTHERS5 classes.

EPs { index numbers }	% frequency in MLL	% frequency in OTHERS5	EPs { index numbers }	% frequency in MLL	% frequency in OTHERS5
$\{2, 14\}$	85.71	0.00	$\{5, 24\}$	0.00	98.94
$\{12, 14\}$	71.43	0.00	$\{5, 22, 38\}$	0.00	96.81
$\{2, 39\}$	64.29	0.00	$\{24, 28, 42\}$	0.00	96.81
$\{14, 16\}$	64.29	0.00	$\{5, 28, 30\}$	0.00	96.81
$\{16, 17\}$	64.29	0.00	$\{5, 7, 30\}$	0.00	96.81
$\{4, 36\}$	64.29	0.00	$\{24, 26, 42\}$	0.00	96.81
$\{4, 8\}$	64.29	0.00	$\{7, 15, 24\}$	0.00	96.81
$\{14, 36\}$	64.29	0.00	$\{15, 24, 26\}$	0.00	96.81
$\{8, 36\}$	57.14	0.00	$\{15, 24, 28\}$	0.00	96.81
$\{2, 31\}$	57.14	0.00	$\{7, 24, 42\}$	0.00	96.81

Table 11: Ten most frequent EPs in the Hyperdip>50 and OTHERS classes.

EPs { index numbers }	% frequency in Hyperdip>50	% frequency in OTHERS	EPs { index numbers }	% frequency in Hyperdip>50	% frequency in OTHERS
$\{14, 24\}$	78.57	0.00	$\{15, 17, 25\}$	0.00	78.85
$\{2, 12, 14\}$	71.43	0.00	$\{7, 15\}$	0.00	76.92
$\{12, 14, 38\}$	71.43	0.00	$\{5, 15\}$	0.00	76.92
$\{4, 14\}$	71.43	0.00	$\{1, 15\}$	0.00	76.92
$\{12, 14, 34\}$	69.05	0.00	$\{15, 33\}$	0.00	76.92
$\{12, 14, 16\}$	69.05	0.00	$\{3, 15\}$	0.00	76.92
$\{2, 8, 14\}$	69.05	0.00	$\{15, 17, 31\}$	0.00	75.00
$\{14, 32\}$	69.05	0.00	$\{15, 17, 19\}$	0.00	75.00
$\{10, 21, 24\}$	66.67	0.00	$\{15, 17, 27\}$	0.00	75.00
$\{12, 21, 24\}$	66.67	0.00	$\{15, 39\}$	0.00	75.00

PCL: Prediction by Collective Likelihood

At level 1 and level 2 of the hierarchical tree-structured datasets, we discovered one rule each. So, there was no ambiguity in using these two rules. However, a large number of EPs were found for the remaining levels of the tree. A testing sample may contain not only EPs from its own class, but also EPs from its counterpart class. This makes the prediction a bit more complicated. Naturally, a testing sample should contain many top-ranked EPs from its own class and contain a few low-ranked, preferably no, EPs from its opposite class. However, according to our observations, a testing sample can sometimes, though rarely, contain 10 or 20 top-ranked EPs from its counterpart class. To make reliable predictions, it is reasonable to use multiple highly frequent EPs of the "home" class to avoid the confusing signals from counterpart EPs.

Given two training datasets $\mathcal{D}_P$ and $\mathcal{D}_N$ and a testing sample T, the first phase of the PCL classifier is to discover EPs from $\mathcal{D}_P$ and $\mathcal{D}_N$. Denote the ranked EPs of $\mathcal{D}_P$ as,

$$EP_1^{(P)}, EP_2^{(P)}, \cdots, EP_i^{(P)},$$

in descending order of their frequency. Similarly, denote the ranked EPs of $\mathcal{D}_N$ as

$$EP_1^{(N)}, EP_2^{(N)}, \cdots, EP_j^{(N)}$$

also in descending order of their frequency. Suppose T contains the following EPs of $\mathcal{D}_P$:

$$EP_{i_1}^{(P)}, EP_{i_2}^{(P)}, \cdots, EP_{i_x}^{(P)},$$

where $i_1 < i_2 < \cdots < i_x \leq i$, and the following EPs of $\mathcal{D}_N$:

$$EP_{j_1}^{(N)}, EP_{j_2}^{(N)}, \cdots, EP_{j_y}^{(N)},$$

where $j_1 < j_2 < \cdots < j_y \leq j$.

The next step is to calculate two scores for predicting the class label of T. Suppose we use k ($k \ll i$ and $k \ll j$) top-ranked EPs of $\mathcal{D}_P$ and $\mathcal{D}_N$. Then we define the score of T in the $\mathcal{D}_P$ class as

$$score(T)\_\mathcal{D}_P = \sum_{m=1}^{k} \frac{frequency(EP_{i_m}^{(P)})}{frequency(EP_m^{(P)})},$$

and similarly the score in the $\mathcal{D}_N$ class as

$$score(T)\_\mathcal{D}_N = \sum_{m=1}^{k} \frac{frequency(EP_{j_m}^{(N)})}{frequency(EP_m^{(N)})}.$$

If $score(T)\_\mathcal{D}_P > score(T)\_\mathcal{D}_N$, then T is predicted as the class of $\mathcal{D}_P$. Otherwise it is predicted as the class of $\mathcal{D}_N$. We use the size of $\mathcal{D}_P$ and $\mathcal{D}_N$ to break tie.

The spirit of our proposal is to measure how far the top k EPs contained in T are away from the top k EPs of a class. Assume $k = 1$, then $score(T)\_\mathcal{D}_P$ indicates whether the number one EP contained in T is far from the most frequent EP of $\mathcal{D}_P$. If the score is the maximum value 1, then the "distance" is very close, namely the most common property of $\mathcal{D}_P$ is also present in this testing sample. With smaller scores, the distance becomes further. Thus the likelihood of T belonging to the class of $\mathcal{D}_P$ becomes weaker. Using more than one top-ranked EPs, we intend to utilize a "collective" likelihood for more reliable predictions.

Table 12 shows the testing accuracy when varing k, the number of rules we want to use. From the results, it is can be seen that multiple highly frequent EPs (or multiple strong rules) can provide a compact and powerful prediction likelihood. With k of 20, 25, and 30, we made a total of 4 mis-classifications, namely, 94-0359-U95A, 89-0142-U95A, 91-0697-U95A, and 96-0379-U95A.

Strengthening the Prediction System

At levels 1 and 2, we used only one gene for the classification and prediction. To overcome some possible rarely-occured human errors on recording data and machine errors by the DNA-chips, we propose to use more than one genes to strengthen our system.

The previously selected one gene 38319_*at* at level 1 was an ideally discriminatory gene. We observed that there were no other such genes. Then we decided to select top 20 genes ranked by the $\mathcal{X}^2$ method

Table 12: The number of EPs used to calculate the scores can slightly effect the prediction accuracy. Error rate, $x : y$, means that x number of samples in the left-side class are misclassified, and y number of samples in the right-side class misclassified.

Testing Data	Error rate when varing k					
	5	10	15	20	25	30
TEL-AML1 vs OTHERS3	2:0	2:0	2:0	1:0	1:0	1:0
BCR-ABL vs OTHERS4	3:0	2:0	2:0	2:0	2:0	2:0
MLL vs OTHERS5	1:0	0:0	0:0	0:0	0:0	0:0
Hyperdip>50 vs OTHERS	0:1	0:1	0:1	0:1	0:1	0:1

Table 13: Error rates comparison of our method with k-NN, C4.5, NB, and SVM on the testing data.

Testing Data	Error rate of different models				
	k-NN	C4.5	SVM	NB	PCL ($k = 20, 25, 30$)
T-ALL vs OTHERS1	0:0	0:1	0:0	0:0	0:0
E2A-PBX1 vs OTHERS2	0:0	0:0	0:0	0:0	0:0
TEL-AML1 vs OTHERS3	0:2	1:1	0:1	0:1	1:0
BCR-ABL vs OTHERS4	4:0	2:0	3:0	1:4	2:0
MLL vs OTHERS5	0:0	0:1	0:0	0:0	0:0
Hyperdip>50 vs OTHERS	0:1	2:6	0:2	0:2	0:1
Total Errors	7	13	6	8	4

to classify the T-ALL and OTHERS1 testing samples. We discovered 96 EPs in the T-ALL class and 146 EPs in the OTHERS1 class. Using our prediction method, we got the same perfect accuracy 100% on those blind testing samples as the one gene was used.

At level 2 there were a total of five ideally discriminatory genes. The name of the five genes are: 430_*at*, 1287_*at*, 33355_*at*, 41146_*at*, and 32063_*at*. Note that 33355_*at* was our previously selected one gene. All of the five genes are partitioned into two intervals with the following cut points respectively: 30246.05, 34313.9, 10966, 25842.15, and 4068.7. There were five EPs in the E2A-PBX1 class and also five EPs in the OTHERS2 class with 100% frequency. Using our prediction method, we correctly classified all the testing samples (at level 2) without any mistakes, achieving the perfect 100% accuracy again.

Comparison with Other Methods

In the field of machine learning, there are many good prediction methods including k-nearest neighbours (k-NN) (Cover & Hart, 1967), C4.5 (Quinlan, 1993), Support Vector Machines (SVM) (Burges, 1998), Naive Bayes (NB) (Langley et al., 1992), etc. We compare PCL's prediction accuracy with the accuracies of k-NN, C4.5, NB, and SVM. Note that the best accuracy of k-NN and SVM reported in (Yeoh et al., 2002) were achieved by using chi-square selected top 20 genes. For a fair comparison, we also used the the same genes, and the same training and testing samples to obtain the accuracies of PCL, C4.5, and NB. The accuracies are summarized in Table 13. All these experiments were conducted, using the Weka machine learning software package with its standard settings (Witten & Frank, 2000). We found that our method reduced the misclassifications by 71% from C4.5's 14, by 50% from NB's 8, by 43% from k-NN's 7, and by 33% from SVM's 6. From the medical treatment point of view, this error reduction can benefit patients greatly.

How about the accuracy if the original data with the entire set of 12558 genes are applied to the prediction methods? SVM, k-NN, NB, and C4.5 made respectively 23, 23, 63, and 26 mis-classifications on the blind testing samples. These results were much worse than the error rates when the reduced data were applied. So, feature selection has played an important role for the classification models to obtain high accuracy.

Next we discuss more about these classification models. C4.5 is a widely used learning algorithm that induces from training data rules that are easy to comprehend. However, it may not have good performance if real decision boundary underlying the data is not linear. The Naive Bayes model uses

Bayesian rules to compute a probabilistic summary for each class. When given a testing sample, NB uses an evaluation function to rank the classes based on their probabilistic summary, and assigns the sample to the highest scoring class. An important assumption used in NB is that the underlying features are statistically independent. However, this is not appropriate for gene expression data analysis as genes involved in an expression profile are often closely related and appear not to be independent.

The k-nearest neighbour method assigns a testing sample the class of its nearest training sample in terms of some non-linear distance functions. Even though the k-NN is intuitive and has good performance, it is not helpful for understanding complex cases in depth. The support vector machines use non-linear kernel functions to construct a complicated mapping between samples and their class labels. SVM has good performance, but it functions as a black box.

An obvious advantage of PCL over SVM, NB, and k-NN is that we can derive meaningful and reliable patterns and rules. Those emerging patterns can provide novel insight into the correlation and interaction of the genes and can help understand the samples more than a mere classification. Though C4.5 can generate similar rules, as it sometimes perform badly (e.g. at level 6), its rules are not very reliable.

GENERALIZED PCL USED IN PARALLEL CLASSIFICATION OF MULTI-CLASS DATA

The hierarchical tree-structured strategy for the subtype classification of ALL samples was initially suggested by the doctors with domain knowledge (Yeoh et al., 2002). Our results above have shown the strength of the structure for classification. However, in the community of machine learning, parallel classification of heterogeneous data is commonly adopted. The parallel classification of multi-class data has the advantage that it does not need any prior knowledge to construct a special classification structure.

In fact, using the parallel approach, we can discover lists of ranked genes and multi-gene discriminators for differentiating one subtype from all other subtypes. The distinction is global as it is one subtype against *all* others. In contrast, under the hierarchical tree-structured classification strategy, the differentiation is somewhat local since the rules are discovered in terms of one subtype against *the remaining subtypes below it*.

In this section, we generalize the PCL method to handle data with more than two classes. The generalized PCL is then applied to Yeoh et al.'s data and to another independent dataset (Armstrong et al., 2002).

Suppose we have c ($c \geq 2$) classes of data, denoted $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_c$. In the first phase, the generalized PCL discovers c groups of EPs. The nth ($1 \leq n \leq c$) group is for $\mathcal{D}_n$ (versus $\cup_{i \neq n} \mathcal{D}_i$). The feature selection and discretization can be done as the same as dealing with typical two-class data. Denote the ranked EPs of $\mathcal{D}_n$ as,

$$EP_1^{(n)}, EP_2^{(n)}, \cdots, EP_{i_n}^{(n)},$$

in a descending order of their frequency.

Suppose a test sample T contain the following EPs of $\mathcal{D}_n$:

$$EP_{j_1}^{(n)}, EP_{j_2}^{(n)}, \cdots, EP_{j_x}^{(n)},$$

where $j_1 < j_2 < \cdots < j_x \leq i_n$. The next step is to calculate c scores for predicting the class label of T. Suppose we use k ($k \ll i_n$) top-ranked EPs. Then the score of T in the $\mathcal{D}_n$ class is defined as

$$score(T)\_\mathcal{D}_n = \sum_{m=1}^{k} \frac{frequency(EP_{j_m}^{(n)})}{frequency(EP_m^{(n)})},$$

and similarly scores for other classes.

The class with the highest score is predicted as the class of T. We use the sizes of $\mathcal{D}_n$, $1 \leq n \leq c$, to break a tie.

Next, we take an example to demonstrate the scores used by PCL. A BCR-ABL test sample contained almost all of the top-20 BCR-ABL discriminators. So, a score of 19.6 was assigned to it. Several top-20 OTHERS discriminators together with some beyond the top-20 list were also contained in this test sample. So, another score of 6.97 was assigned. This test sample did not contain any discriminators of E2A-PBX1, Hyperdip>50, or T-ALL. So, the scores are as follows:

subtype	BCR-ABL	E2A-PBX1	Hyperdip>50	T-ALL	MLL	TEL-AML1	OTHERS
score	19.63	0.00	0.00	0.00	0.71	2.96	6.97

Therefore, this BCR-ABL sample was correctly predicted as BCR-ABL with very high confidence.

By this method, we made only 6 to 8 misclassifications for the total 112 testing samples when varying k from 15 to 35. However, C4.5, SVM, NB, and 3-NN made 27, 26, 29, and 11 mistakes respectively.

Armstrong et al. (2002)'s data have only three subtypes (AML, ALL, MLL). There are a total 57 training samples (20, 17, and 20 respectively for ALL, MLL, and AML) and 15 test samples (4, 3, and 8 respectively for ALL, MLL, and AML). PCL, C4.5, SVM, NB, and k-NN made 0, 3, 1, 0, and 1 mistakes respectively for the 15 test samples.

DISCUSSION

We have already shown that the selected top-1 genes or top-20 genes are very useful for classifying the subtypes of childhood leukemia. We describe here a method to support the quality of our selected genes by examining the accuracy difference when 20 genes or 1 gene is *randomly* picked from the training data instead of selecting top-1 or top-20 genes.

The procedure is: (a) Randomly select one gene at level 1 and level 2, and randomly select 20 genes at each of the four remaining levels. (b) Run SVM and k-NN, obtain their accuracy on the testing samples of each level. (c) Repeat (a) and (b) a hundred times, calculate averages and other statistics. We found that big accuracy differences existed between these two feature selection methods. Therefore, good gene selection methods can work effectively with the prediction methods. So, good feature selections are important preliminary steps before reliable and accurate prediction models are established.

All above results was obtained by reserving roughly 2/3 of the data for training and the remaining 1/3 for blind testing. In fact, we also found that PCL made much improvement in error rate over other methods in a 10-fold cross validation test on the training data.

In this work, we have compared the performance of different classification models under the same reduced training and test data. However, it is well known that the performance of a classifier may vary if the number of selected genes is changed. It is also well known that there is no theoretical estimation of the optimal number of selected genes even for a specific classifier on a particular application. It will be interesting to see which gene number and classification model combination is the best for classifying the subtype of childhood leukemia. For future work, we plan to conduct a systematic evaluation of gene selection methods. We also want to investigate refinements of our method, particularly in techniques for measuring the "interestingness" of individual EPs. We are also planning to consider automatic ways for determining optimal numbers of the most discriminatory genes used for EP discovery.

Acknowledgments

We thank See Kiong Ng, Philip Long, and Vladmir Bajic for their helpful comments on a draft of the paper. We also thank the referees for their critical and useful comments.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of National Academy of Sciences of the United States of American*, *96*, 6745–675.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., & Korsmeyer, S. J. (2002). MLL translocations specify a distict gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, *30*, 41–47.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*, 121–167.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27.

Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 43–52). San Diego, CA: ACM Press.

Fayyad, U., & Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (pp. 1022–1029). Morgan Kaufmann.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, *286*, 531–537.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifier. *Proceedings of the Tenth National Conference on Artificial Intelligence* (pp. 223 – 228). AAAI Press.

Li, J., Ramamohanarao, K., & Dong, G. (2000). The space of jumping emerging patterns and its incremental maintenance algorithms. *Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, USA* (pp. 551–558). San Francisco: Morgan Kaufmann.

Li, J., & Wong, L. (2002). Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics*, *18*, 725–734.

Liu, H., & Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. *Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence* (pp. 338 – 391).

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Witten, H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with java implementation*. San Mateo, CA: Morgan Kaufmann.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.-H., Evans, W. E., Naeve, C., Wong, L., & Downing, J. R. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, *1*, 133–143.