Identifying Good Diagnostic Gene Groups from Gene Expression Profiles Using the Concept of Emerging Patterns

Jinyan Li

Kent Ridge Digital Labs
21, Heng Mui Keng Terrace, Singapore, 119613
jinyan@krdl.org.sg

Limsoon Wong

Kent Ridge Digital Labs
21, Heng Mui Keng Terrace, Singapore, 119613
limsoon@krdl.org.sg

Contact Author: Dr. Li Jinyan
Address: Kent Ridge Digital Labs
21, Heng Mui Keng Terrace, Singapore 119613
Emai: `jinyan@krdl.org.sg`
Phone: +65 874 8217; Fax: +65 774 8056

Running head: Discovering special gene groups from expression data.

Keywords: Gene expression profiles, emerging patterns, entropy-based discretization, classification, treatment plan.

Abstract

Motivations and Results: Gene groups that are significantly related to a disease can be detected by conducting a series of gene expression experiments. This work is aimed at discovering special types of gene groups that satisfy the following property. In each group, its member genes are found to be one-to-one contained in pre-determined intervals of gene expression level with a large frequency in one class of cells but are never found unanimously in these intervals in the other class of cells. We call these gene groups *emerging patterns*, to emphasize the patterns' frequency changes between two classes of cells. We use effective discretization and gene selection methods to obtain the most discriminatory genes. We also use efficient algorithms to derive the patterns from these genes. According to our studies on the ALL/AML dataset and the colon tumor dataset, some patterns, which consist of one or more genes, can reach a high frequency of 90%, or even 100%. In other words, they nearly or fully dominate one class of cells, even though they rarely occur in the other class. The discovered patterns are used to classify new cells with a higher accuracy than other reported methods. Based on these patterns, we also conjecture the possibility of a personalized treatment plan which converts colon tumor cells into normal cells by modulating the expression levels of a few genes.

Contact: `jinyan@krdl.org.sg`; `limsoon@krdl.org.sg`

INTRODUCTION

Current DNA chip-based technologies (Schena *et al.*, 1995; Lockhart & *et al.*, 1996; Velculescu *et al.*, 1995) can measure the expression level of thousands of genes simultaneously, generating a large amount of high-dimensional data. It is of great interest to analyze the data and to extract comprehensible patterns. Based on the concept of emerging patterns (Dong and Li, 1999), this work is aimed at discovering special gene groups from the ALL/AML dataset (Golub *et al.*, 1999) and the colon tumor dataset (Alon *et al.*, 1999). The discovered patterns consist of groups of genes that are constrained to specific intervals of gene expression levels such that these patterns only occur in one class of cells but do not occur in the other cells. In other words, these patterns correspond to the conditions on gene expression levels that dominate that class of cells. Then pharmacological intervention may be introduced based on these patterns to modulate the expression levels of some genes to treat a disease.

Emerging patterns (Dong and Li, 1999) —EPs for short—are an important concept used in this paper. EPs capture significant differences between two classes of things, such as edible mushrooms vs poisonous mushrooms, normal tissues vs cancer tissues, and so on. The significance of the differences is measured by the magnitude of the frequency-change ratio of the patterns from one class to another. The larger the frequency-change ratio, the more important the patterns. Due to the sharp change in frequency, EPs can be used to distinguish instances between different classes and to predict the class label of new instances. For example the following EP,

$$\{gene(K03001) \geq 89.20\} \text{ } and \text{ } \{gene(R76254) \geq 127.16\} \text{ } and \text{ } \{gene(D31767) \geq 63.03\}$$

discovered from the colon tumor dataset later, changes its frequency of 0% in normal tissues to a frequency of 75% in cancer tissues. Here $gene(X)$ represents the expression level of the gene X. According to this emerging pattern, if the expression levels of K03001, R76254, and D31767 in a new cell experiment are, respectively, not less than 89.20, 127.16, and 63.03, then this cell is much more likely a cancer cell.

Gene expression data are typically organized as a matrix. Assume that such a matrix has n rows and m columns. Then n usually represents the number of considered genes and m the number of experiments. The experiments are mainly categorized into two types. The first type of experiments are aimed at simultaneously monitoring the n genes m times under a series of varying conditions (Chu *et al.*, 1998; DeRisi *et al.*, 1997; Roberts *et al.*, 2000; Wen *et al.*, 1998). The second type is used to examine the n genes in a single environment but from m different cells (Alon *et al.*, 1999; Golub *et al.*, 1999; Perou *et al.*, 1999; Wang *et al.*, 1999; Zhu *et al.*, 1998). The first type of experiments are intended to detect trends and regularities of the expression of every single gene under a series of conditions. The resulting data are generally temporal. The latter type of experiments are expected to provide information for classifying the type of new cells and for identifying useful genes whose expressions are good diagnostic indicators (Alon *et al.*, 1999; Golub *et al.*, 1999). The resulting data are generally spatial.

The colon tumor dataset (Alon *et al.*, 1999) and the ALL/AML dataset (Golub *et al.*, 1999) were obtained by conducting the second type of experiments. The colon tumor dataset consists of 22 normal tissues and 40 colon tumor tissues. The ALL/AML dataset consists of 38 bone marrow samples obtained from acute leukemia patients; 27 of them are acute leukemias arising from lymphoid precursors (acute lymphoblastic leukemia, ALL), and the remaining 11 cases are acute leukemias arising from myeloid precursors (acute myeloid leukemia, AML).

We investigate the problems below: (i) Which intervals of the expression levels of a group of genes occur only in cancer tissues but not in the normal tissues and vice versa, or only in ALL patients but not in AML patients and vice versa? (ii) How to discretize a range of the expression levels of each gene into multiple intervals so that the contrasting intervals, which form our EPs, are informative and reliable? (iii) How to explain the results and then to suggest a plan for treating the disease?

We approach these problems using several techniques. First, we use an entropy-based discretization method (Fayyad and Irani, 1993) to discretize gene expression levels into suitable intervals. This method partitions a range of real values into a number of disjoint intervals such that the entropy of the partition is minimal. The selection of *cut points*, breaking points in a range, in this discretization process is crucial. With the minimal entropy idea, the intervals are "maximally" discriminatory between expression values from normal cells and expression values from cancer cells. This method automatically ignores those ranges which contain uniformly mixed normal and cancer cells' expression values. For the colon cancer dataset, of its 2000 genes, only 35 genes are discretized into 2 intervals. The remaining 1965 genes are ignored by the method. This result is very important because it implies that most of the genes are viewed as irrelevant by the method. It gives us an easy platform where a small number of good diagnostic indicators are concentrated. Second, we use efficient algorithms (Dong and Li, 1999; Li, 2001; Li *et al.*, 2000b) to discover EPs based on the discretized data. The EPs are combinations of intervals of gene expression levels of these relevant genes. The EPs that are discovered are maximally frequent in one class of data but do not occur in the other class. The discovered EPs always contain only a small number of the relevant genes. This result reveals interesting conditions on the expression of these genes that differentiate between two classes of data. Third, each EP with high frequency is considered as a common property of a class of cells and, based on this idea, we propose a strategy for treating colon tumors by adjusting the expression level of some improperly expressed genes. We show later that almost all "adjusted" cells are predicted as normal cells by a number of good classifiers that were trained to distinguished between normal vs colon tumor cells.

The remainder of this paper is organized as follows. We begin with a brief introduction to the concept of emerging patterns (Dong and Li, 1999). Then we describe a method (Fayyad and Irani, 1993) to discretize continuous features (attributes). Then we present our main results on the colon tumor dataset, including the discretization results, the EP discovery results, the classification accuracy, and a treatment plan for curing the disease. Then we provide the discovered patterns from the ALL/AML dataset. Finally we conclude this paper.

EMERGING PATTERNS

We now describe some basic terms pertaining to the concept of emerging patterns. Given a gene, denoted as $gene_j$, its expression levels, under a series of varying conditions or under a single condition but from different types of cells, form a range of real values. Suppose this range is $[x, y]$ and an interval $[a, b]$ (or in other forms, $(a, b]$, $[a, b)$, or (a, b)) is contained in $[x, y]$. We call $gene_j@[a, b]$ an *item*, meaning that the expression levels of $gene_j$ are limited inclusively between a and b. A set of items, denoted as $\{gene_{i_1}@[a_{i_1}, b_{i_1}], \cdots, gene_{i_k}@[a_{i_k}, b_{i_k}]\}$, where $i_t \neq i_s, 1 \leq t, s \leq k$ if $k >= 1$, is called a *pattern*.

A pattern always has a frequency in a dataset. We use an example to show how to calculate the frequency of a pattern. Table 1 consists of the expression levels of four genes in six cells, three of which are normal and three of which are cancerous. Each of the six columns is called an *instance*. The pattern

Table 1: A simple gene expression dataset.

Cell Type	normal	normal	normal	cancer	cancer	cancer
gene_1	0.1	0.2	0.3	0.4	0.5	0.6
gene_2	1.2	1.1	1.3	1.4	1.0	1.1
gene_3	-0.70	-0.83	-0.75	-1.21	-0.78	-0.32
gene_4	3.25	4.37	5.21	0.41	0.75	0.82

$\{gene_1@[0.1, 0.3]\}$ has a frequency of 50% in the dataset because the expression levels of $gene_1$ in the first three instances are in the interval $[0.1, 0.3]$. The pattern $\{gene_1@[0.1, 0.3], gene_3@[0.30, 1.21]\}$ has a 0% frequency because no one instance satisfies the two conditions that (i) $gene_1$'s value must be in $[0.1, 0.3]$ and (ii) $gene_3$'s value must be in $[0.30, 1.21]$. However, the pattern $\{gene_1@[0.4, 0.6], gene_4@[0.41, 0.82]\}$ has a frequency of 50%.

We now turn to describe the notion of emerging patterns. Let the dataset of Table 1 be divided into two small sub-datasets, one consisting of values of the three normal cells and the other of values of the three cancer cells. Then for the same pattern, its frequency can change from one sub-dataset to another sub-dataset. Emerging patterns are those patterns whose frequency is *significantly* changed. The pattern $\{gene_1@[0.1, 0.3]\}$ is an emerging pattern as it has a frequency of 100% in the sub-dataset with normal cells but it has a 0% frequency in the sub-dataset with cancer cells. The pattern $\{gene_1@[0.4, 0.6], gene_4@[0.41, 0.82]\}$ is also an emerging pattern. Observe that it has a 0% frequency in the sub-dataset with normal cells.

We focus on a special type of EPs that are called *left-boundary EPs*. These emerging patterns are *maximally frequent* in one dataset but have a 0% frequency in the other dataset. That is, they have a frequency-change ratio of ∞. The left-boundary EPs are also required to satisfy that any proper subset of a left-boundary EP is not an EP (Dong and Li, 1999; Li, 2001; Li *et al.*, 2000b).

AN ENTROPY-BASED DISCRETIZATION METHOD

One challenge of gene expression data to data mining algorithms is the large number of genes involved. This work introduces a discretization method (Fayyad and Irani, 1993) that makes use of the heuristic of entropy minimization. As discussed earlier, the method automatically removes many noisy features (genes) and effectively explores the remaining most discriminatory features.

We follow the notation presented in (Dougherty *et al.*, 1995; Fayyad and Irani, 1993). Let T partition the set S of examples into the subsets S_1 and S_2. Let there be k classes $C_1, \cdots, C_k$. Let $P(C_i, S_j)$ be the proportion of examples in S_j that have class C_i. The *class entropy* of a subset $S_j, j = 1, 2$ is defined as:

$$Ent(S_j) = -\sum_{i=1}^{k} P(C_i, S_j) log(P(C_i, S_j)).$$

Suppose the subsets S_1 and S_2 are induced by partitioning a feature A at point T. Then, the *class information entropy* of the partition, denoted $E(A, T; S)$, is given by:

$$E(A, T; S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2).$$

A binary discretization for A is determined by selecting the cut point T_A for which $E(A,T;S)$ is minimal amongst all the candidate cut point (Fayyad and Irani, 1993). The same process can be applied recursively to S_1 and S_2 until some stopping criteria is reached.

The *Minimal Description Length Principle* is used to stop partitioning (Fayyad and Irani, 1993). Recursive partitioning within a set of values S stops iff

$$Gain(A,T;S) < \frac{log_2(N-1)}{N} + \frac{\delta(A,T;S)}{N},$$

where N is the number of values in the set S, $Gain(A,T;S) = Ent(S) - E(A,T;S)$, $\delta(A,T;S) = log_2(3^k - 2) - [k \cdot Ent(S) - k_1 \cdot Ent(S_1) - k_2 \cdot Ent(S_2)]$, and k_i is the number of class labels represented in the set S_i.

Note that this method is resistant to the effect of data normalization (e.g., taking logarithm) provided that the normalization does not change the order of the original values. The method has been implemented by MLC++ techniques (Kohavi *et al.*, 1994). The executable codes are available at `http://www.sgi.com/tech/mlc/`.

RESULTS ON THE COLON TUMOR DATASET

We next report the discretization results on the colon tumor dataset to see which genes are selected and which genes are discarded. Then, we present our important results—the emerging patterns together with their frequency. We then discuss how the expression scale of every single gene shifts between normal and cancer tissues. Based on the discovered EPs, we also suggest a therapy for treating the colon tumor patients.

Discretization Results

The discretization method partitions 35 of the 2000 genes each into two disjoint intervals, while there is no cut point in the remaining 1965 genes. This indicates that only $35/2000 = 1.75\%$ of the genes are considered as the most discriminatory genes and the others can be considered as irrelevant indicators. By deriving a small number of good diagnostic indicator genes, the discretization method lays down a foundation for us to efficiently discover reliable emerging patterns. Otherwise, a large number of noisy patterns would be generated.

The discretization result is summarized in Table 2. The first column is our index of the genes, the second column shows the gene accession numbers, the third column presents the intervals of gene expression levels, followed by the gene's name at column 4. Note that there are a total of 70 intervals. Accordingly, there are 70 items involved. Recall that an item is a pair where a gene is linked with an interval. We index the 70 items. The first gene's two intervals are indexed as the 1st and 2nd items, $\cdots$, the ith gene's two intervals as the $(i*2-1)$th and $(i*2)$th items, $\cdots$, the 35th gene's two intervals as the 69th and 70th items. This index is convenient to read and write emerging patterns; for example, the pattern $\{2\}$ represents $\{gene_{T51560}@[101.3719, +\infty)\}$.

Discovered Emerging Patterns

The EPs are discovered by two efficient border-based algorithms, BORDER-DIFF and JEPPRODUCER (Dong and Li, 1999; Li, 2001; Li *et al.*, 2000b). They do not need to conduct an exhaustive enumeration of EP candidates. With borders, a concise representation tool, the algorithms do not need to output all EPs but only the most frequent ones from which the others can be derived. According to our experimental studies, the algorithms are linearly scalable to the size of samples, and linearly scalable to the number of features approximately.

A total of 19501 EPs that have a non-zero frequency in the normal tissues of the colon tumor dataset were discovered. A total of 2165 EPs that have a non-zero frequency in the cancer tissues were derived by our algorithms.

Table 3 and Table 4 present, in terms of frequencies, the top 20 EPs which occur in the 22 normal tissues, and the top 20 EPs which occur in the 40 cancer tissues. Column 1 shows the emerging patterns. The numbers in the patterns, for example 16, 58, and 62 in the pattern {16 58 62}, stand for the items just discussed and indexed.

Computationally, we explain the emerging patterns as follows:

- Some of the emerging patterns are very interesting because they contain a relatively large number of genes. For example, the pattern {2 3 6 7 13 17 33} combines 7 genes together and still has a very large frequency (90.91%) in the normal tissues. Almost every normal cell's expression values satisfy all of the conditions implied by these 7 items. However, no single cancer cell satisfies all the conditions. By definition (see the section of EMERGING PATTERNS), all of the proper sub-patterns of the pattern {2 3 6 7 13 17 33}, including singletons and the combinations of six items, must have a non-zero frequency in both of normal and cancer tissues. This means that there must exist at least one cell from both of normal and cancer tissues satisfying the conditions implied by any sub-patterns of {2 3 6 7 13 17 33}.
- The frequency of a singleton emerging pattern is not necessarily larger than emerging patterns containing more than one item. For example the pattern {5} is an emerging pattern in cancer cells with a frequency of 32.5%. Comparing with the frequency (75%) of the pattern {16 58 62}, the frequency of {5} is about 2.3 times less. This indicates that, for the analysis of gene expression data, groups of genes and their correlations are better and more important than single genes.
- Without the discretization method and the border-based EP discovery algorithms, it is very hard to discover these reliable emerging patterns with large frequencies. Assume the 1965 genes are each partitioned into two intervals as well, then there are $C_{20}^{7} * 2^{7}$ possible patterns having a length of 7. The enumeration of so huge a number of patterns and the calculation of their frequencies is definitely impossible. Even with the discretization method, the naive enumeration of $C_{35}^{7} * 2^{7}$ patterns is still too expensive for discovering the pattern {2 3 6 7 13 17 33}. Furthermore, some of the discovered EPs (not listed here) contain more than 7 genes. Through the use of the two border-based algorithms, only those EPs whose proper subsets are not emerging patterns are discovered. We can derive other EPs using the discovered EPs. Generally, any proper superset of a discovered EP is also an emerging pattern. For example, using the EPs with the count of 20 (shown in Table 3), we can derive a very long emerging pattern, {2 3 6 7 9 11 13 17 23 29 33 35}, consisting of 12 genes, with the same count of 20. Furthermore, the border-based algorithms are guaranteed to discover all the emerging patterns.

Note that any one of the 62 tissues must match at least one emerging pattern from its own class, but never contain any EPs from the other class. So, our system has well learned the whole data.

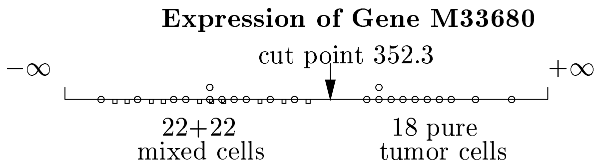

Figure 1: The distribution of the expression profile of gene M33680 in the 62 cells.

Gene Expression Trends

Next, we answer the question of which genes change their expression level significantly between normal and tumor tissues. We first look for those genes whose expression in normal and cancer tissues are two-end polarized. However, this ideal case does not happen to the colon tumor dataset. So, for any single gene, there does not exist any cut point such that the resulting two intervals contain pure-class points. (Observe that if an interval contains pure-class points, then it must be an EP.) Sub-optimally, we search for genes which have two intervals containing *almost* pure-class points.

Using Table 5, we report the cut point for every gene's expression level range and the coverage of the intervals. By the coverage of an interval, we mean the number of samples from both normal and cancer tissues that are covered by the interval. In the 8th column of Table 5, we provide the entropy of every gene's range when a cut point is inserted. By definition, the smaller the entropy is, the more discriminatory the gene is. For the ideal case where two intervals cover only pure-class points, the gene's entropy is 0.

From Table 5, we see that the 15th gene, M33680, has the lowest entropy among the total 35 genes when the cut point of 352.3133 is set. The gene's left interval, $(-\infty, 352.3133)$, covers the expression of 22 normal cells and 22 cancer cells. The gene's right interval, $[352.3133, +\infty)$, covers the expression levels of 18 cancer cells. See Figure 1. This indicates that the expression level of gene M33680 in normal cells is less than 352.3133. Furthermore, if the expression level of gene M33680 in a cell is larger than 352.3133, then this cell is cancerous. Therefore, it is understood that 45% (18/40), nearly a half, of the cancer cells shift the expression level of gene M33680 from its normal range to an abnormal range.

Similarly, let us examine the 22nd gene, X12671, which has the second lowest entropy among the 35 genes. Neither its left interval nor its right interval covers pure-class expression levels. However, the gene X12671's expression levels in 86% (19/22) of the normal cells fall into the left interval, and its expression in 70% of the cancer cells are higher, falling into the right interval $[149.4719, +\infty)$. It can be seen that the gene X12671's expression levels in nearly three quarters of the cancer cells were increased during cancer progression.

In summary, there is no single gene that acts as an arbitrator to distinguish the normal and cancer cells. However, genes with lower entropies can be considered as good diagnostic markers. We next consider gene groups and use those highly frequent emerging patterns for planning a treatment strategy.

A Treatment Plan

The basic idea of the treatment plan is to increase or decrease the expression levels of some particular genes in a cancer cell, so that it has the common properties of normal cells. As a result, instead of killing it, the cancer cell is "converted" into a normal one.

As shown in Table 3, the frequency of emerging patterns can reach a very high level such as 90%. The conditions implied by a highly frequent EP form a common property of one class of cells. Using the EP {2 3 6 7 13 17 33} again, we have understood that each of the about 91% of normal cells expresses the seven genes (T51560, T49941, M62994, R34701, L02426, U20428, and R10707) in such a way that the seven expression levels are one-to-one contained in the seven intervals (the 2nd, 3rd, 6th, 7th, 13th, 17th, and 33rd intervals as shown in Table 2). Although a cancer cell may express some of the seven genes in a similar manner as normal cells do, according to the dataset, a cancer cell can never express all of the seven genes in the same way as normal cells do. So, if the expression levels of those improperly expressed genes can be adjusted, the cancer cell can be made to have one common property that normal cells exhibit. A cancer cell can then be iteratively converted into a normal one. As there usually exist some genes of a cancer cell which express in a similar way as their counterparts in normal cells, less than 35 genes' expression levels are required to be changed. The most important issue is to determine which genes need an adjustment. Our EPs can be used to address this issue.

We use a cancer cell (T1) of the colon tumor dataset as an example to show how a tumor cell is converted into a normal one. Firstly, we list the expression profile of 7 genes of T1:

Gene	T51560	T49941	M62994	R34701	L02426	U20428	R10707
Expression	1148.16	364.34	467.61	290.80	140.64	197.32	178.15

Note that about 91% (20 out of 22 cases) of the normal cells have the following expression intervals for these 7 genes:

Gene	T51560	T49941	M62994	R34701	L02426	U20428	R10707
Intervals	$[101.37, +\infty)$	$(-\infty, 272.54)$	$[94.40, +\infty)$	$(-\infty, 446.03)$	$(-\infty, 390.61)$	$(-\infty, 207.80)$	$(-\infty, 378.80)$

Comparing T1's gene expression levels with the intervals of normal cells, we see that only T49941 of the 7 genes of the cancer cell T1 behaves in a different way from those the 22 normal cells commonly express. Namely, 364.34 is beyond the interval $(-\infty, 272.54)$. However, the rest of the 7 genes of T1 are in the same expression range as most of the normal cells. So, if the gene T49941 of T1 can be down regulated to scale below the gate point 272.54, then this adjusted cancer cell will have a common property of normal cells. This is because {2 3 6 7 13 17 33} is an emerging pattern which does not occur in cancer cells. This idea is at the core of our suggestion for a treatment plan.

Interestingly, the expression change of the gene T49941 in T1 leads to a chain of other changes. These include the change that two extra top-ten EPs of normal cells are contained in the adjusted T1. So a total of three top-ten EPs of normal cells are contained in T1 if the gene T49941's expression level is adjusted. Similarly, after scaling the gene X62153's expression level to below 395.25 and the gene T72403's expression level below 296.57, T1 will contain 9 top-ten EPs of the normal cells. As the average number of top-ten EPs contained in normal cells is 9, the changed T1 cell will now be considered as a cell that has the most important features of normal cells. Note that we have adjusted only three genes' expression level so far.

We also need to eliminate those common properties of cancer cells that are contained in T1. By adjusting the expression level of another six genes (K03001, T49732, U29171, R76254, D31767, L40992), the top-ten EPs of cancer cells all disappear from T1. (Typically, the average number of top-ten EPs of cancer cells contained in a cancer cell is 6.) Therefore, T1 is converted into a normal cell as it is now holding the common properties of normal cells and does not include the common properties of cancer cells.

By this method, all the other 39 cancer cells can be converted into normal ones after adjusting the expression levels of 10 genes or so, possibly different genes from person to person. We conjecture that this personalized treatment plan is effective if the expression of some particular genes can be modulated by suitable means.

We next discuss a validation of this idea. The "adjustments" we made to the 40 colon tumour cells were based on the emerging patterns in the manner described above. If these adjustments had indeed converted the colon tumour cells into normal cells, then any good classifier that could distinguish normal vs colon tumour cells on the basis of gene expression profiles would classify our adjusted cells as normal cells. The widely-used Support Vector Machine (SVM) is a powerful tool for prediction tasks. We thus established a SVM model using original data from the 62 (22 normal plus 40 cancer) cells as training data. (The code for constructing a SVM model is available at `http://www.cs.waikato.ac.nz/ml/weka`.) Then the emerging pattern-based method described earlier was used to adjust the expression levels of particular genes in the cancer cells. Then, the 40 adjusted cancer cells were fed to the model. The outcome was that all of the adjusted cells were predicted as normal cells. Though our "therapy" was not applied to the real treatment of a patient, the prediction result by the SVM model partially demonstrates the potential biological significance of our proposal.

There are other good classification algorithms that could be obtained at `http://www.cs.waikato.ac.nz/ml/weka`. We also tried C4.5, HyperPipes (HP), and Voting Feature Intervals (VFI) to predict the class of the converted cells. C4.5 predicted 32 of the 40 converted cells as normal and the remaining 8 still as cancer cells. The HP classifier and the VFI classifier both predicted 39 converted cancer cells as normal and the remaining one cell still as cancer. These results are very close to that of the SVM model. Note that the SVM model, the HP classifier, and the VFI classifier all have good performance on the training data (6, 5, and 3 mistakes respectively).

Usefulness of the EPs in Classification

In this section, we test the usefulness of EPs in classification by conducting a *Leave-One-Out-Cross-Validation* (LOOCV) application. By LOOCV, we pick up the first instance of the 62 tissues as a test instance, and the remaining 61 instances as training data. Repeating through the first instance to the 62nd one, we can finally get an accuracy, the percent of the instances which are correctly predicted.

For a given test instance, denoted $tInstance$, and its corresponding training data $\mathcal{D}$, our method consists of the following steps for predicting the class of $tInstance$:

1. Divide $\mathcal{D}$ into two sub-datasets, denoted $\mathcal{D}_n$ and $\mathcal{D}_c$, respectively consisting of the normal training tissues and the cancer training tissues.

2. Discover the EPs in $\mathcal{D}_n$, and similarly discover the EPs in $\mathcal{D}_c$.

3. According to the frequency and the *length* (the number of items in a pattern), sort the EPs (from both $\mathcal{D}_c$ and $\mathcal{D}_n$) into a descending order. The ranking criteria are:

   (a) Given two EPs X_i and X_j, if the frequency of X_i is larger than X_j, then X_i is prior to X_j in the list.

(b) When the frequency of X_i is equal to X_j, if the length of X_i is longer than X_j, then X_i is prior to X_j in the list.

(c) We treat the two patterns equally when their frequency and length are both identical.

Denote the ranked EP list as $orderedEPs$.

4. Put the first EP of $orderedEPs$ back into $finalEPs$.

5. If the first EP is from $\mathcal{D}_n$ (or $\mathcal{D}_c$), establish a new $\mathcal{D}_n$ (or a new $\mathcal{D}_c$) such that it consists of those instances of $\mathcal{D}_n$ (of $\mathcal{D}_c$) which do not contain the EP.

6. Repeat from Step 2 to Step 5 until a new $\mathcal{D}_n$ or a new $\mathcal{D}_c$ is empty.

7. Find the first EP in the $finalEPs$ which is contained, or one whose immediate proper EP subsets is contained, in $tInstance$. If the EP is from normal class, we predict the test instance as a normal cell. Otherwise the test instance is classified as a cancerous one.

We correctly predict 57 of the 62 tissues. Only three normal tissues (N1, N2, and N39) were wrongly classified as cancer tissues, and two cancer tissues (T28 and T33) were wrongly predicted as normal tissues. We compare this result with a result in the literature. Furey *et al* (2000) mis-classified six tissues (T30, T33, T36, N8, N34, and N36), using 1000 genes and a SVM approach. Interestingly our mis-classified examples almost differ from those mis-classified by the SVM method. (T33 was commonly mis-classified.) It can be seen that the classification performance of our method is better than the SVM method (Furey *et al.*, 2000) though we used a much smaller number of genes. Another advantage is that our method can provide easily understandable patterns rather than a "black box" embedded in the SVM model.

RESULTS ON THE ALL/AML DATASET

Next, we apply our method to the ALL/AML dataset. Using the discretization method, 866 of the 7129 features in the 38 training data are partitioned into two or three intervals; there is no cut point in the rest of the features. So, there are 6263 discarded features. Generally, the availability of such a large number of multi-interval features implies that it is easy to separate the two classes of samples.

We examined the 866 features and found that gene Zyxin has an entropy of 0 when the cut point is set as 994. This indicates that the expression of Zyxin in all the 27 ALL samples is less than 994, but the expression of Zyxin in all the 11 AML samples is equal to or larger than 994. So the gene Zyxin can be used solely by itself as an arbitrator to make a clear distinction between the 38 ALL and AML samples. We note that the two intervals associated with Zyxin form two EPs, each having 100% frequency in the two classes. We also found many other EPs with 100% frequency. All of these EPs with 100% frequency are arbitrators to distinguish ALL and AML samples.

To show a further potential of our EPs in classification, we again use the one gene, Zyxin, to predict the class of the 34 *testing* samples (Golub *et al.*, 1999). The rule is that if the expression of Zyxin of a cell is less than 994, then this cell is an ALL sample. Otherwise it is an AML sample. We correctly classified 31 samples, but mis-classified 3 (one AML, the 31st sample; two ALL, the 15th and 17th samples). This result is better than or comparable to several results reported in the literature (Golub *et al.*, 1999; Furey *et al.*, 2000) where at least dozens even thousands of genes were used. This comparison is summarized in Table 6.

DISCUSSION

The essence of the new concept of emerging patterns is to distinguish two classes of things. This concept has shown great potential to discover differences from the expression profile of two types of cells, and it can be extended to more than two types of cells. Particularly interesting patterns are those with six, seven, or even 12 genes and with a very large frequency. Even with only one or two of the patterns, the normal vs colon tumour tissues or the ALL vs AML samples can be totally distinguished. In this study, the EPs are ranked based on their frequency. We plan to investigate in future the interestingness of EPs based on their neighbourhood frequency fluctuation (Dong and Li, 1998) or other types of measurements (Golub *et al.*, 1999).

Gene expression values are always continuous and the number of genes involved is very large. An important problem in the discovery of EPs is how to discretize data and how to select a few number of diagnostic features. The entropy-based discretization method can automatically ignore most of the genes if their expression are randomly distributed in a range. In the field of machine learning, there are many other optional methods for feature selection. It is interesting to see whether those methods work for expression data.

In addition, commonly in real life only as few as 3 or 4 patients samples are available per class. Many prediction models are greatly affected by small numbers of training samples. Fortunately, the dataset used in this study is relatively large and our classification model works well. We would like to in future study how robust our method is in terms of the size of training data.

The proposed treatment plan is a personalized therapy. The idea is to modify the expression of some specific genes such that a cancer cell can hold many common properties of normal cells but contain few EPs of cancer cells. Consequently, the cancer cell is converted into a normal one. Eventually, this plan should be confirmed by biological and medical experiments or treatments. However, an effective means for identifying the mechanisms and pathways through which to modulate the expression of selected genes needs to be developed first.

ACKNOWLEDGEMENT

We thank Vladimir Bajic, See Kiong Ng, and Huiqing Liu for their help in this study. We also greatly thank the two reviewers for their excellent comments and suggestions.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 6745–6750.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O. and Herskowitz, I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

DeRisi, J. L., Iyer, V. R. and Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Dong, G. and Li, J. (1998) Interestingness of discovered association rules in terms of neighborhood-based unexpectedness. *Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 72-86). Springer-Verlag.

Dong, G. and 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.

Dougherty, J., Kohavi, R. and Sahami, M. (1995) Supervised and unsupervised discretization of continuous features. *Proceedings of the Twelfth International Conference on Machine Learning* (pp. 94–202). Morgan Kaufmann.

Fayyad, U. and Irani, K. (1993) Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence* (pp. 1022–1029). Morgan Kaufmann.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. and Haussler, D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

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. and Lander, E. S. (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Kohavi, R., John, G., Long, R., Manley, D. and Pfleger, K. (1994) MLC++: A machine learning library in C++. *Tools with artificial intelligence* (pp. 740 – 743).

Li, J. (2001) *Mining emerging patterns to construct accurate and efficient classifiers.* Department of Computer Science and Software Engineering, The University of Melbourne, Australia: Ph.D. Thesis.

Li, J., Ramamohanarao, K. and Dong, G. (2000b) 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.

Lockhart, D. and *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotech.*, **14**, 1675–1680.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O. and Botstein, D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 9212–9217.

Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., Tyers, M., Boone, C. and Friend, S. H. (2000) Signaling and circuitry of multiple mapk pathways revealed by a matrix of global gene expression profiles. *Science*, **287**, 873–880.

Schena, M., Shalon, D., Davis, R. and Brown, P. (1995) Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, **270**, 467–470.

Velculescu, V., Zhang, L., Vogelstein, B. and Kinzler, K. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.

Wang, K., Gan, L., Jefferey, E., Gayle, M., Gown, A., Skelly, M., Nelson, P., Ng, W., Schummer, M., Hood, L. and Mulligan, J. (1999) Monitoring gene expression profile changes in ovarian carcinomas using cdna micoroarray. *Gene*, **229**, 101–108.

Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. and Somogyi, R. (1998) Neurobiology large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. U.S.A.*, **95(1)**, 334–339.

Zhu, H., Cong, J.-P., Mamtora, G., Gingeras, T. and Shenk, T. (1998) Cellular gene expression altered by human cytomegalovirus: Global monitoring with oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 14470–14475.

Table 2: The 35 genes that are discretized by the entropy-based method into more than one intervals.

Our list	accession number	Intervals	Name
1	T51560	$(-\infty, 101.3719)$, $[101.3719, +\infty)$	40S Ribosomal Protein S16 (Human)
2	T49941	$(-\infty, 272.5444)$, $[272.5444, +\infty)$	Putative Insulin-like Growth Factor II Associated (Human)
3	M62994	$(-\infty, 94.39874)$, $[94.39874, +\infty)$	Homo sapiens thyroid autoantigen (truncated actin-binding protein) mRNA, complete cds
4	R34701	$(-\infty, 446.0319)$, $[446.0319, +\infty)$	Trans-acting Transcriptional Protein ICP4 (Varicella-zoster virus)
5	X62153	$(-\infty, 395.2505)$, $[395.2505, +\infty)$	H.sapiens mRNA for P1 protein (P1.h)
6	T72403	$(-\infty, 296.5696)$, $[296.5696, +\infty)$	HLA CLASS II HISTOCOMPATIBILITY ANTIGEN, DQ(3) ALPHA CHAIN PRECURSOR (Homo sapiens)
7	L02426	$(-\infty, 390.6063)$, $[390.6063, +\infty)$	Human 26S protease (S4) regulatory subunit mRNA, complete cds
8	K03001	$(-\infty, 89.19624)$, $[89.19624, +\infty)$	Human aldehyde dehydrogenase 2 mRNA
9	U20428	$(-\infty, 207.8004)$, $[207.8004, +\infty)$	Human unknown protein (SNC19) mRNA, partial cds
10	R53936	$(-\infty, 206.2879)$, $[206.2879, +\infty)$	PROTEIN PHOSPHATASE 2C HOMOLOG 2 (S.pombe)
11	H11650	$(-\infty, 211.6081)$, $[211.6081, +\infty)$	ADP-RIBOSYLATION FACTOR 4 (Homo sapiens)
12	R59097	$(-\infty, 402.66)$, $[402.66, +\infty)$	Tyrosine-Protein Kinase Receptor TIE-1 Precursor (M.musculus)
13	T49732	$(-\infty, 119.7312)$, $[119.7312, +\infty)$	Human SnRNP core protein Sm D2 mRNA, complete cds
14	J04182	$(-\infty, 159.04)$, $[159.04, +\infty)$	Lysosome-associated Membrane Glycoprotein 1 Precursor (Human)
15	M33680	$(-\infty, 352.3133)$, $[352.3133, +\infty)$	Human 26-kDa cell surface protein TAPA-1 mRNA, complete cds
16	R09400	$(-\infty, 219.7038)$, $[219.7038, +\infty)$	S39423 PROTEIN I-5111, INTERFERON-GAMMA-INDUCED
17	R10707	$(-\infty, 378.7988)$, $[378.7988, +\infty)$	Translational Initiation Factor 2 Alpha Subunit (H.sapiens)
18	D23672	$(-\infty, 466.8373)$, $[466.8373, +\infty)$	Human mRNA for biotin-[propionyl-CoA-carboxylase (ATP-hydrolysing)] ligase, complete cds
19	R54818	$(-\infty, 153.1559)$, $[153.1559, +\infty)$	Human eukaryotic initiation factor 2B-epsilon mRNA, partial cds
20	J03075	$(-\infty, 218.1981)$, $[218.1981, +\infty)$	PROTEIN KINASE C SUBSTRATE, 80 KD PROTEIN, HEAVY CHAIN (HUMAN); contains TAR1 repetitive element
21	T51250	$(-\infty, 212.137)$, $[212.137, +\infty)$	Cytochrome C Oxidase polypeptide VIII-Liver/Heart (Human)
22	X12671	$(-\infty, 149.4719)$, $[149.4719, +\infty)$	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1
23	T49703	$(-\infty, 342.1025)$, $[342.1025, +\infty)$	60S Acidic Ribosomal Protein P1 (Polyorchis penicillatus)
24	U03865	$(-\infty, 76.86501)$, $[76.86501, +\infty)$	Human adrenergic alpha-1b receptor protein mRNA, complete cds
25	X16316	$(-\infty, 65.27499)$, $[65.27499, +\infty)$	VAV ONCOGENE (HUMAN)
26	U29171	$(-\infty, 181.9562)$, $[181.9562, +\infty)$	Human casein kinase I delta mRNA, complete cds
27	H89983	$(-\infty, 200.727)$, $[200.727, +\infty)$	METALLOPAN-STIMULIN 1 (Homo sapiens)
28	T52003	$(-\infty, 180.0342)$, $[180.0342, +\infty)$	CCAAT/ENHANCER BINDING PROTEIN ALPHA (R.norvegicus)
29	R76254	$(-\infty, 127.1584)$, $[127.1584, +\infty)$	ELONGATION FACTOR 1-GAMMA (Homo sapiens)
30	M95627	$(-\infty, 65.27499)$, $[65.27499, +\infty)$	H.sapiens angio-associated migratory cell protein (AAMP) mRNA, complete cds
31	D31767	$(-\infty, 63.03381)$, $[63.03381, +\infty)$	Human mRNA (KIAA0058) for ORF (novel protein), complete cds
32	R43914	$(-\infty, 65.27499)$, $[65.27499, +\infty)$	CREB-BINDING PROTEIN (Mus musculus)
33	M37721	$(-\infty, 963.0405)$, $[963.0405, +\infty)$	PEPTIDYL-GLYCINE ALPHA-AMIDATING MONOOXYGENASE PRECURSOR (HUMAN); contains Alu repetitive element
34	L40992	$(-\infty, 64.85062)$, $[64.85062, +\infty)$	Homo sapiens (clone PEBP2aA1) core-binding factor, runt domain, alpha subunit 1 (CBFA1) mRNA, 3' end of cds
35	H15662	$(-\infty, 894.9052)$, $[894.9052, +\infty)$	GLUTAMATE (Mus musculus)

Table 3: The top 20 emerging patterns, in a descending order, sorted by their frequency in the 22 normal tissues.

Emerging patterns	Counts	Frequency (%) in normal tissues	Frequency (%) in cancer tissues
{ 2 3 6 7 13 17 33 }	20	90.91	0
{ 2 3 11 17 23 35 }	20	90.91	0
{ 2 3 11 17 33 35 }	20	90.91	0
{ 2 3 7 11 17 33 }	20	90.91	0
{ 2 3 7 11 17 23 }	20	90.91	0
{ 2 3 6 7 13 17 23 }	20	90.91	0
{ 2 3 6 7 9 17 33 }	20	90.91	0
{ 2 3 6 7 9 17 23 }	20	90.91	0
{ 2 3 6 17 23 35 }	20	90.91	0
{ 2 3 6 17 33 35 }	20	90.91	0
{ 2 6 7 13 39 41 }	19	86.36	0
{ 2 3 6 7 13 41 }	19	86.36	0
{ 2 6 35 39 41 45 }	19	86.36	0
{ 2 3 6 7 9 31 33 }	19	86.36	0
{ 2 6 7 39 41 45 }	19	86.36	0
{ 2 3 6 7 41 45 }	19	86.36	0
{ 2 6 9 35 39 41 }	19	86.36	0
{ 2 3 17 21 23 35 }	19	86.36	0
{ 2 3 6 7 11 23 31 }	19	86.36	0
{ 2 3 6 7 13 23 31 }	19	86.36	0

Table 4: The top 20 emerging patterns, in a descending order, sorted by their frequency in the 40 cancer tissues.

Emerging patterns	Counts	Frequency (%) in normal tissues	Frequency (%) in cancer tissues
{ 16 58 62 }	30	0	75.00
{ 26 58 62 }	26	0	65.00
{ 28 58 }	25	0	62.50
{ 26 52 62 64 }	25	0	62.50
{ 26 52 68 }	25	0	62.50
{ 16 38 58 }	24	0	60.00
{ 16 42 62 }	24	0	60.00
{ 16 26 52 62 }	24	0	60.00
{ 16 42 68 }	24	0	60.00
{ 26 28 52 }	23	0	57.50
{ 16 38 52 68 }	23	0	57.50
{ 16 38 52 62 }	23	0	57.50
{ 26 52 54 }	22	0	55.00
{ 26 32 }	22	0	55.00
{ 16 54 58 }	22	0	55.00
{ 16 56 58 }	22	0	55.00
{ 26 38 58 }	22	0	55.00
{ 32 58 }	22	0	55.00
{ 16 52 58 }	22	0	55.00
{ 22 26 62 }	22	0	55.00

Table 5: The cut points in the expression ranges of the 35 genes and the entropy of the partitions.

Our list	Gene number	Left Interval's Coverage normal	Left Interval's Coverage tumor	Cut Point	Right Interval's Coverage normal	Right Interval's Coverage tumor	Entropy
1	T51560	0	13	101.3719	22	27	0.784376
2	T49941	21	21	272.5444	1	19	0.769805
3	M62994	0	13	94.39874	22	27	0.784376
4	R34701	22	27	446.0319	0	13	0.784376
5	X62153	22	25	395.2505	0	15	0.755835
6	T72403	21	21	296.5696	1	19	0.769805
7	L02426	22	24	390.6063	0	16	0.740923
8	K03001	11	2	89.19624	11	38	0.737061
9	U20428	21	18	207.8004	1	22	0.722061
10	R53936	19	15	206.2879	3	25	0.764748
11	H11650	19	14	211.6081	3	26	0.747847
12	R59097	22	25	402.66	0	15	0.755835
13	T49732	14	5	119.7312	8	35	0.735524
14	J04182	18	10	159.04	4	30	0.71121
15	M33680	22	22	352.3133	0	18	0.709677
16	R09400	20	15	219.7038	2	25	0.722073
17	R10707	22	26	378.7988	0	14	0.770311
18	D23672	22	27	466.8373	0	13	0.784376
19	R54818	18	12	153.1559	4	28	0.750364
20	J03075	20	18	218.1981	2	22	0.771865
21	T51250	19	15	212.137	3	25	0.764748
22	X12671	19	12	149.4719	3	28	0.710793
23	T49703	21	21	342.1025	1	19	0.769805
24	U03865	10	2	76.86501	12	38	0.766972
25	X16316	6	0	65.27499	16	40	0.779593
26	U29171	18	12	181.9562	4	28	0.750364
27	H89983	18	11	200.727	4	29	0.731494
28	T52003	19	13	180.0342	3	27	0.729896
29	R76254	16	9	127.1584	6	31	0.761726
30	M95627	6	0	65.27499	16	40	0.779593
31	D31767	6	0	63.03381	16	40	0.779593
32	R43914	6	0	65.27499	16	40	0.779593
33	M37721	16	40	963.0405	6	0	0.779593
34	L40992	7	0	64.85062	15	40	0.749908
35	H15662	16	40	894.9052	6	0	0.779593

Table 6: The accuracy comparison of our method vs a SVM method (Furey *et al.*, 2000) vs a clustering method (Golub *et al.*, 1999) on the 34 ALL/AML testing samples, where the "accuracy" means the number of samples that are correctly classified.

Method	# of Used Genes	Accuracy
our method	1	31
SVM	25 – 1000	30 – 32
clustering	50	29