Emerging Patterns and Gene Expression Data

Jinyan Li[1] **Limsoon Wong**[1]
jinyan@krdl.org.sg limsoon@krdl.org.sg

[1] Kent Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore, 119613.

Abstract

One important purpose of conducting gene expression experiments is to understand the correlation of gene expression profiles to disease states. Based on the notion of emerging patterns and an entropy-oriented discretization method, we discover groups of genes that are correlated to disease states in a significant way. In each group, every member gene constrained by a specific expression interval, unanimously occurs only in one type of cells with a maximally large frequency, but never unanimously happens in the other types of cells. According to our studies on the colon tumor dataset, such gene groups (also called patterns) can reach a frequency of 90%, providing good insight into the correlation of gene expression profiles to disease states. The patterns can be used to correctly predict whether a new cell is normal or cancerous.

Keywords: Gene expression data, emerging patterns, entropy-based discretization, intervals, frequency, classification, colon tumor prediction.

1 Introduction

The notion of *emerging patterns* [4], EPs for short, has been proposed to capture significant differences between two classes of things (e.g., between edible and poisonous mushrooms, between normal tissues and cancer tissues, between genes coding for ribosomal proteins and genes not coding for ribosomal proteins, and so on). The significance of the differences is measured by the magnitude of the frequency-change-ratio of the patterns over one class to another. The larger the frequency-change-ratio is, the more important the patterns are. 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,

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

whose discovery is detailed in Section 5, changes its frequency of 0% in normal tissues to a frequency of 75% in cancer tissues. Here $gene(X)$ represents the expression value of the gene X. According to this emerging pattern, in a new cell experiment if the gene K03001's value is not less than 89.20 and the gene R76254's is not less than 127.16 and the gene D31767's is not less than 63.03, then this cell would be much more likely a cancerous cell.

Gene expression data, obtained by highly parallel experiments using technologies like microarrays [16], oligonucleotide 'chips' [13], and SAGE [17], records expression levels of genes under specific experimental conditions. 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 represents the number of experiments. The experiments are mainly categorized into two types. The first type of experiments is aimed at simultaneously monitoring the n genes m times under a series of varying conditions [2, 3, 15, 19]. The second type is used to examine the n genes in a single environment but from m different cells [1, 8, 14, 18, 20]. Briefly speaking, the first type of experiments is intended to provide any possible trends or regularities of every single gene under a series of conditions. The resulting data is generally temporal. The latter type of experiments is expected to provide information for classifying the type of new cells and for the identification of useful genes whose expressions are good diagnostic indicators [1, 8]. The resulting data is generally spatial.

In this paper, we focus our studies on the colon tumor dataset [1], which is a dataset obtained by one of the second type of experiments. This dataset consists of 22 normal tissues and 40 colon tumor tissues. In this work we primarily investigate the following problems:

1. Which intervals of the expression values of a gene or which combinations of multiple genes' intervals only occur in the cancer tissues but not in the normal tissues, or only occur in the normal tissues but not in the cancer tissues?

2. How to discretize a range of the expression values of a gene into multiple intervals so that the above mentioned contrasting intervals or interval combinations, all our EPs, are informative and reliable?

3. Can the discovered patterns be used to perform classification tasks, i.e. predicting whether a new cell is normal or cancerous after conducting the same type of expression experiment?

We solve these problems using several unique techniques. First of all, to discretize a range of gene expression values, we use an entropy-based discretization method [6]. The basic idea of this method is to partition a range of real values into a number of disjoint intervals such that the entropy of the intervals is minimal. The selection of the cut points in this discretization process is crucial. With the minimal entropy idea, the intervals are "maximally" and reliably discriminatory between expression values from normal cells and expression values from cancerous cells. This method can automatically ignore those ranges which contain relatively uniformly mixed normal and cancerous cells' expression values. For the colon cancer dataset [1], of its 2000 genes, only 35 relevant genes are discretized into 2 intervals while the remaining 1965 genes are ignored by the method. This result is very important since most of the genes have been viewed as "trivial" ones, resulting in an easy platform where a small number of good diagnostic indicators are concentrated.

We use efficient algorithms [4, 10, 11, 12] to discover emerging patterns based on the discretized data. We discover those emerging patterns which are maximally frequent in one class of data (normal tissues or cancerous tissues), but totally do not occur in the other class. Having this type of emerging patterns, we can easily derive other emerging patterns. The discovered emerging patterns always contain a few number of genes. This result not only allows users to focus on a few number of good diagnostic indicators, but more importantly it reveals some interactions of the genes which are originated in the combination of the genes' intervals and the frequency of the combinations. The discovered emerging patterns can be used to predict the properties of a new cell. According to the classification results on the same dataset, our method performs much better than a SVM method [7] and clustering method [1].

The remainder of this paper is organized as follows. Section 2 briefly describes the concept of emerging patterns [4]. Section 3 presents a method [6] to discretize continuous features (attributes). Section 4 outlines the colon cancer dataset [1]. Section 5 presents our main results, including the discretization results and the EP discovery results. Section 6 shows the usefulness of the discovered patterns in classification. Finally, Section 7 concludes this paper.

2 Brief Description of Emerging Patterns

Firstly, we introduce the notion of patterns. Gene expression values are continuous. Given a gene, denote $gene_j$, its expression values, under a series of varying conditions or under a single condition but from different types of cells, forms a range of real values. Suppose this range is $[a, b]$ and an interval $[c, d]$ is contained in $[a, b]$. We call $gene_j@[c, d]$ an *item*, meaning the values of $gene_j$ is limited

Table 1: A simple gene expression dataset.

Cell Type	normal	normal	normal	cancerous	cancerous	cancerous
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

inclusively between c and d. A set of one single item is called a *pattern*. A set of several items which come from different genes is also called a *pattern*. So, a pattern looks like:

$$\{gene_{i_1}@[a_{i_1}, b_{i_1}], \cdots, gene_{i_k}@[a_{i_k}, b_{i_k}]\}$$

where $i_t \neq i_s, 1 \leq t, s \leq k$ if $k > 1$.

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 four genes' expression values of six cells (three normal and three cancerous). We call each of the six columns an *instance*. For the pattern $\{gene_1@[0.1, 0.3]\}$, it has a frequency of 50% in the dataset as the $gene_1$'s values of the first three instances are in the interval $[0.1, 0.3]$. For another pattern $\{gene_1@[0.1, 0.3], gene_3@[0.30, 1.21]\}$, it has a 0% frequency. This is because no single 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%.

Next we describe the notion of emerging patterns [4]. If the dataset of Table 1 is divided into two small sub-datasets: one consists of the values of the three normal cells, the other of the values of the three cancerous 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 cancerous 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. In the current work, we put our effort on discovering those emerging patterns which are maximally frequent in one dataset, but have a 0% frequency in the other dataset, having a frequency-change-ratio of ∞.

Usually, a subset of an EP need not alos be an EP. However, in our experiments, we did find some EPs whose non-empty-subsets are all emerging patterns. We call this type of EPs *strong* EPs. Though the number of strong EPs may be small, they are important as they tend to be more robust than other EPs when one or more new instances are added into training data. We also discover strong k-EPs: An EP is called a strong k-EP if every subset of cardinality at least k is also an EP. As will been seen, strong EPs play an important role in classification.

3 An Entropy-based Discretization Method

Many data mining tasks need continuous features to be discretized. We describe here a discretization method [6] which makes use of the entropy minimization heuristic. This method can automatically remove many noisy features and effectively explores the remaining discriminatory features.

We follow the notations presented in [5, 6]. Let T partition the set S of examples into the subsets S_1 and S_2. Let there be k classes $C_1, \cdots, C_k$ and 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 [6]. 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 [6]. 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.

This method has been implemented by MLC++ techniques [9]. The executable codes are available at `http://www.sgi.com/tech/mlc/`.

4 The Colon Tumor Dataset

Using the Affymetrix Hum6000 array, Alon et al have obtained a dataset consisting of the expression values on about 6500 genes of 40 tumor and 22 normal colon tissue samples [1]. In order to reduce the size of the dataset, only 2000 of the 6500 genes were chosen according to their minimal intensity across the samples, those genes with lower minimal intensity were ignored. The reduced dataset is publicly available at `http://microarray.princeton.edu/oncology/affydata/index.html`.

After downloading the original data from the above website, for discretization, we re-organized the data in accordance with the format required by the utilities of MLC++ [9]. Basically, the re-organized dataset is diagonally symmetrical to the original dataset.

5 Results

One challenge of gene expression data to data mining algorithms is the huge number of genes involved. The entropy-based discretization method [6] can automatically ignore most of the genes and select a few most discriminatory genes. Therefore, many noisy data and noisy patterns can be effectively eliminated. In this section, we first present the discretization results to see which genes are selected and which genes are discarded. Then, we present our important results: the emerging patterns in the colon tumor dataset together with their frequency.

5.1 The 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 most discriminatory genes and the others can be negligible. Deriving a small number of good diagnostic genes, the discretization method lays down a foundation for us to efficiently discover reliable emerging patterns, otherwise huge number of noisy patterns would be generated.

We summarize the discretization results in Table 2 (presented at the second last page). The first column is our list of the genes, the second column shows the gene numbers, the intervals are presented at column 3, followed by the gene's sequence and name at columns 4 and 5.

Observe that there are a total number of 70 intervals. Accordingly, there are 70 items involved. Recall that an item is a pair where a gene is linked with an interval (see Section 2). 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 as shown in the next subsection. For example, the pattern $\{2\}$ represents $\{gene_{T51560}@[101.3719, +\infty)\}$.

5.2 Emerging Patterns

We discover emerging patterns using two efficient border-based algorithms, BORDER-DIFF and JEP-PRODUCER [4, 10, 12]. The algorithms can derive those EPs which occur in one class of data with a maximally large frequency, but never occur in the other class. A total of 19501 EPs, which have a non-zero frequency in the normal tissues of the colon tumor dataset [1], were discovered. And a total of 2165 EPs, which have a non-zero frequency in the cancerous tissues, were derived by our algorithms.

According to the frequency, Table 3 (presented at the last page) and Table 4 list the top 20 EPs and strong EPs which occur in the 22 normal tissues, and the top 20 EPs and strong EPs which occur in the 40 cancerous 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 discussed and indexed in the last subsection.

We summarize our understandings on the emerging patterns as follows:

- Some of the emerging patterns are surprisingly interesting, particularly for those containing a relatively large number of genes. For example, the pattern $\{2\ 3\ 6\ 7\ 13\ 17\ 33\}$ combines 7 genes together, it can still have a very large frequency (90.91%) in the normal tissues, namely almost every normal cell's expression values satisfy all of the conditions implied by the 7 items. However, no single cancerous cell satisfies all the conditions. Observe that 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 the normal and cancerous tissues. This means that there must exist at least one cell from both of the normal and cancerous tissues satisfying the conditions implied by any sub-patterns of $\{2\ 3\ 6\ 7\ 13\ 17\ 33\}$.

- The frequency of a singleton emerging patterns is not necessarily larger than emerging patterns containing more than one items. For example the pattern $\{5\}$ is an emerging pattern in the cancerous tissues 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 gene's.

- Without the discretization method and the border-based EP discovery algorithms, it is very hard to discover those reliable emerging patterns with large frequencies. Assume the 1965 genes are each partitioned into two intervals as well, then there are $C_{2000}^{7} * 2^7$ possible patterns having a length of 7. The enumeration of so huge 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. Interestingly, 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.

- Our border-based algorithm is guaranteed to discover all the emerging patterns.

Note that for any of the 62 tissues, it 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. Next we use emerging patterns to perform a classification task to see how useful of the patterns in predicting whether a new cell is normal or cancerous.

6 The Usefulness of EPs in Classification

As shown in Table 3 and Table 4, the frequency of the EPs is very large. Obviously, the groups of genes are good indicators for classifying new tissues. In this section, we test the usefulness of the patterns by conducting a *Leave-One-Out-Cross-Validation* (LOOCV) classification task. 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 cancerous 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), sorting the EPs (from both $\mathcal{D}_c$ and $\mathcal{D}_n$) into a descending order. The ranking criteria is that

   (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 and X_j is identical, 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.

   Denoted 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 untill 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 of 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 cancerous tissues, and two cancerous tissues (T28 and T33) were wrongly predicted as normal tissues. We compare this result with a result in the literature. Furey et al. [7] 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 [7].

We stress that the colon tumor dataset is very complex. Normally and ideally, for a test normal (or cancerous) tissue, it should contain a large number of EPs from the normal (or cancerous) training tissues, and a few number of EPs from the other type of tissues. However, based on our experiments, it can contain many EPs, even the top-ranked highly frequent EPs, from the both classes of tissues.

We next make use of strong EPs to see whether our system can be made more accurate. The steps are as follows:

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 cancerous training tissues.

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

3. According to frequency, sorting each of the two lists of EPs into a descending order. Denoted the ordered EP lists as $orderedEPs_n$ and $orderedEPs_c$ respectively for the strong EPs in $\mathcal{D}_n$ and $\mathcal{D}_c$.

4. Find the top EPs from $orderedEPs_n$ such that they must be contained in $tInstance$, denoted them as $EP_{n1}, \cdots, EP_{nk}$. Similarly, find the top EPs from $orderedEPs_c$ such that they must be contained in $tInstance$, denoted them as $EP_{c1}, \cdots, EP_{cj}$.

5. Compare the frequency of EP_{n1} with the frequency of EP_{c1}, the former is larger, we predict the test instance as a normal cell. Otherwise if the latter is larger, the test instance is classified as a cancerous one. We break tie situations using strong 2-EPs.

Using this method, we correctly predict 58 of the 62 tissues. Four normal tissues (N1, N12, N27, and N39) were wrongly classified as cancerous tissues. The result becomes better.

7 Conclusion

In this paper, we have studied the problem of how to discover emerging patterns from gene expression data. Emerging patterns are defined as those patterns which occur in one type of cells with a maximal frequency, but never occur in the other type of cells, sharply discriminating the two classes.

As gene expression values are always continuous and the number of genes involved is very large, we have presented an entropy-based discretization method to partition the ranges of the expressions. Based on the minimal entropy idea, the discretization method can automatically ignore most of the genes as there is no good cut point in them. The remaining genes, which have been partitioned into two intervals, are good diagnostic indicators.

We have ranked the emerging patterns according to their frequency and length. The surprisingly interesting patterns are those with six, seven, or even 12 genes and with a very large frequency (e.g. larger than 70%). Even with only one or two patterns, the normal and cancerous tisses can be totally distinguished.

We have conducted a LOOCV classification on the colon tumor dataset. The performance of only five mis-classified tissues shows that our method is better than the other methods described in the literature, including a SVM-based and a clustering-based method. As a future work, we will further examine the correlation between the frequency and the length of an EP, and use it for a better classification.

References

[1] 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.

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

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

[4] 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.

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

[6] Fayyad, U., & 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.

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

[8] 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.

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

[10] 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.

[11] Li, J., Dong, G., & Ramamohanarao, K. (2000a). Making use of the most expressive jumping emerging patterns for classification. *Proceedings of the Fourth Pacific-Asia Conference on Knowledge Discovery and Data Mining. An expanded version of the paper appears in Knowledge and Information Systems: An International Journal, 2001* (pp. 220–232). Kyoto, Japan: Springer-Verlag.

[12] Li, J., Ramamohanarao, K., & 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.

[13] Lockhart, D., & et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, *14*, 1675–1680.

[14] 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., & Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of National Academy of Sciences of the United States of American*, *96*, 9212–9217.

[15] 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., & Friend, S. H. (2000). Signaling and circuitry of multiple mapk pathways revealed by a matrix of global gene expression profiles. *Science*, *287*, 873–880.

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

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

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

[19] Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L., & Somogyi, R. (1998). Neurobiology large-scale temporal gene expression mapping of central nervous system development. *Proceedings of National Academy of Sciences of the United States of American*, *95(1)*, 334–339.

[20] Zhu, H., Cong, J.-P., Mamtora, G., Gingeras, T., & Shenk, T. (1998). Cellular gene expression altered by human cytomegalovirus: Global monitoring with oligonucleotide arrays. *Proceedings of National Academy of Sciences of the United States of American*, *95*, 14470–14475.

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

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

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

Emerging patterns	Counts	Frequency (%) in normal tissues	Frequency in tumor tissues	Strong EPs	Counts	Frequency in normal tissues
{ 2 3 6 7 13 17 33 }	20	90.9091	0%	{ 67 }	7	31.8182
{ 2 3 11 17 23 35 }	20	90.9091	0%	{ 59 }	6	27.2727
{ 2 3 11 17 33 35 }	20	90.9091	0%	{ 61 }	6	27.2727
{ 2 3 7 11 17 33 }	20	90.9091	0%	{ 70 }	6	27.2727
{ 2 3 7 11 17 23 }	20	90.9091	0%	{ 49 }	6	27.2727
{ 2 3 6 7 13 17 23 }	20	90.9091	0%	{ 66 }	6	27.2727
{ 2 3 6 7 9 17 33 }	20	90.9091	0%	{ 63 }	6	27.2727
{ 2 3 6 7 9 17 23 }	20	90.9091	0%	{ 49 66 }	4	18.1818
{ 2 3 6 17 23 35 }	20	90.9091	0%	{ 49 66 }	4	18.1818
{ 2 3 6 17 33 35 }	20	90.9091	0%	{ 59 63 }	4	18.1818
{ 2 6 7 13 39 41 }	19	86.3636	0%	{ 59 70 }	4	18.1818
{ 2 3 6 7 13 41 }	19	86.3636	0%	{ 59 63 }	4	18.1818
{ 2 6 35 39 41 45 }	19	86.3636	0%	{ 59 70 }	4	18.1818
{ 2 3 6 7 9 31 33 }	19	86.3636	0%	{ 49 59 66 }	3	13.6364
{ 2 6 7 39 41 45 }	19	86.3636	0%	{ 49 59 66 }	3	13.6364
{ 2 3 6 7 41 45 }	19	86.3636	0%	{ 59 61 63 }	3	13.6364
{ 2 6 9 35 39 41 }	19	86.3636	0%	{ 59 63 70 }	3	13.6364
{ 2 3 17 21 23 35 }	19	86.3636	0%	{ 59 61 63 }	3	13.6364
{ 2 3 6 7 11 23 31 }	19	86.3636	0%	{ 59 63 70 }	3	13.6364
{ 2 3 6 7 13 23 31 }	19	86.3636	0%	{ 49 59 66 }	3	13.6364

Table 4: The top 20 EPs and the top 20 strong EPs, in a descending order, sorted by their frequency in the 40 cancerous tissues.

Emerging patterns	Counts	Frequency normal tissues	Frequency (%) in tumor tissues	Strong EPs	Counts	Frequency (%) in tumor tissues
{ 16 58 62 }	30	0%	75.0000	{ 30 }	18	45.0000
{ 26 58 62 }	26	0%	65.0000	{ 14 }	16	40.0000
{ 28 58 }	25	0%	62.5000	{ 10 }	15	37.5000
{ 26 52 62 64 }	25	0%	62.5000	{ 24 }	15	37.5000
{ 26 52 68 }	25	0%	62.5000	{ 34 }	14	35.0000
{ 16 38 58 }	24	0%	60.0000	{ 36 }	13	32.5000
{ 16 42 62 }	24	0%	60.0000	{ 1 }	13	32.5000
{ 16 26 52 62 }	24	0%	60.0000	{ 5 }	13	32.5000
{ 16 42 68 }	24	0%	60.0000	{ 8 }	13	32.5000
{ 26 28 52 }	23	0%	57.5000	{ 24 30 }	11	27.5000
{ 16 38 52 68 }	23	0%	57.5000	{ 30 34 }	11	27.5000
{ 16 38 52 62 }	23	0%	57.5000	{ 24 30 }	11	27.5000
{ 26 52 54 }	22	0%	55.0000	{ 30 34 }	11	27.5000
{ 26 32 }	22	0%	55.0000	{ 10 14 }	10	25.0000
{ 16 54 58 }	22	0%	55.0000	{ 10 14 }	10	25.0000
{ 16 56 58 }	22	0%	55.0000	{ 24 34 }	9	22.5000
{ 26 38 58 }	22	0%	55.0000	{ 14 24 }	9	22.5000
{ 32 58 }	22	0%	55.0000	{ 8 10 }	9	22.5000
{ 16 52 58 }	22	0%	55.0000	{ 10 24 }	9	22.5000
{ 22 26 62 }	22	0%	55.0000	{ 8 10 }	9	22.5000