IJBPAS, November, 2015, 4(11), Special Issue: 691-699
ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)
'A Bridge Between Laboratory and Reader'
www.ijbpas.com

APPLICATION OF APPROXIMATE FREQUENT ITEMSETS FOR FEATURE SELECTION OF PROSTATE CANCER GENE EXPRESSION DATA

HAMID ALAVI MAJD[1], AHMAD REZA BAGHESTANI[1], SEYYED MOHAMMAD TABATABAEI[2], SOODEH SHAHSAVARI[1*], MOSTAFA REZAEI TAVIRANI[3], MOHSEN HAMIDPOUR[4]

1. Biostatistics Department, Faculty of Paramedical Sciences, ShahidBeheshti University of Medical Sciences, Darband Avenue, Qods Square, Tehran, Iran.
2. Medical Informatics Department, Faculty of Paramedical Sciences, ShahidBeheshti University of Medical Sciences, Darband Avenue, Qods Square, Tehran, Iran.
3. Proteomics Research Center, Faculty of Paramedical Sciences, ShahidBeheshti University of Medical Sciences, Darband Avenue, Qods Square, Tehran, Iran.
4. Hematology Department, Faculty of Paramedical Sciences, ShahidBeheshti University of Medical Sciences, Tehran, Iran.

***Correspondence author: soodeh_shahsavari@yahoo.com**

ABSTRACT

The main goal of this study was to conduct biclustering processing of prostate cancer gene expression data, to identify subgroups of tumors with similar clinical features. One of the most important methods of data mining is association rules and discovered traits and characteristics are similar with larger data sets. Error-toletant Frequent Itemsets (ETI) is an approach that involves association pattern mining to discover error-tolerant biclusters from noisy real-valued gene-expression data that is contained 22277 genes and 4 conditions (normal and MED1 conditions). In this study the ETI algorithm was applied with different error-tolerance and range support. 12 biclusters were discovered for selected scenario, of which their minimum and maximum number of genes were 1759 and 5913 respectively. These biclusters had almost correlated structures under MED1 conditions and so genes in prostate cancer are better co-expressed in this situation. GO analysis and randomization test are shown results are reliable.

Keywords: Prostate cancer, gene expression data, biclustering algorithm

INTRODUCTION

Cancer is a term that covers a complex set of diseases(1).Carcinogenic, the transformation of a normal cell into a cancer cell, is complex process with many stages(2). Clinically, cancer refers to a large group of diseases in which difference occurs according toa wide range of conditions such asage of onset, rate of growth, cell differentiation status, ability for identification with diagnostic measures, invasion, metastasis, response to treatment and prognosis(3). In molecular and cellular biology, cancer represents a small number of diseases caused by the same molecular defects that occur in cellular activity and cause similar changes in cell genes. Finally, cancer is a disease caused by abnormal gene expression(4). Prostate cancer is the second most frequently diagnosed cancer and the sixth leading cause of cancer death in males, accounting for 14% (903,500) of the total new cancer cases and 6% (258,400) of the total cancer deaths in males in 2008(5).It is known that about 90 percent of patients with advanced prostate cancer, metastasis will progress to bone tissue(6). The global burden of prostate cancer is expected to raise 1.7 million new cases and 499,000 deaths by 2030 due to growth and aging of the worldwide population (7). Several studies have shown familial aggregation of prostate cancer. The main reason for this is that the genes involved in the development of cancer are inherited and some of these genes show high penetration while other genes are involved in polymorphism and have low penetration(8).Research in gene expression profiling of this disease is useful to identify similar clinical characteristics and to help researchers in diagnostics. In recent years, DNA microarray technology has provided simultaneous monitoring of thousands of gene expressionsfor cells under various different conditions and processes. This technology has a key role in accelerating and increasing the efficiency of gene expression studies(9). One of the most important pattern recognition techniques in analysis of gene expression data is clustering that can be applied to identify groups of genes with similar expression patterns(10).The aim of clustering is to find groups that are very different from each other but in which members of groups are very similar. But use of classical clustering techniques is limited for gene expression(11).The clustering process is based on the assumption that genes under all measurement conditions have the same behavior;while usually, under similar conditions subsets of genes may have the same behavior but in other circumstances they may have independent behaviors. This situation is particularly evident in cases

with a higher number of conditions. Secondly, clustering techniques often divide heterogeneous populations into a number of homogenous sub-populations to which a gene can only belong to one of the clusters so genes that they may have in common are not shown. Also, it is biologically possible that some genes become down regulated in any of the experimental conditions and therefore cannot be expressed; but in classical clustering methods it is necessary that all genes in a cluster be assigned(12).In order to overcome these limitations and to identify suitable gene expression patterns, a two-dimensional clustering concept has been proposed that has a more flexible computing framework(13).In this method, a search is done for a subset of genes with similar expression patterns in a subset of conditions. Namely, the homogeneity submatrices found in gene expression data and local patterns of the data are obtained. These submatrices are defined as biclusters and this process of identifying biclusters is called biclustering(14).

One of the most important methods of data mining is association rules and discovered traits and characteristics are similar with larger data sets(15). Perhaps this technique could be the perfect method for detecting patterns in unsupervised learning systems and can be used to find all possible patterns in a database so that all possible relationships are explored in this technique. The disadvantage of this method is that detecting and searching is very time consuming and that a user faces a large amount of information to analyze(16). The Frequent Item sets algorithm is one of the most widely used and most important methods in association rules and it presents a good alternative to biclustering algorithms. Error-toletant Frequent Itemsets (ETI) is aapproach that involves association pattern mining to discover error-tolerant biclusters from noisy real-valued gene-expression data(17). The ETI algorithm was applied in this study because with tolerating error in the biclusters, it captures the true underlying structure of the data. It is assumed that the gene expression data matrix D with dimensions n × m determined that rows in accordance with genes and columns are in accordance with the laboratory conditions. Elements of $D_{j,k}$ show gene expression level for jth gene and kth condition that

$T_0=\{t_1,\ldots,t_n\}$ j=1,…,n set of genes

$I_0=\{i_1,\ldots,i_m\}$ k=1,…,m set of conditions

The tolerances in algorithm for genes and conditions are displayedε_r and ε_c respectively that $\varepsilon_r,\varepsilon_c\in[0,1]$and the following conditions must be met for discovering of biclusters.

$$\forall j\in T,\quad \frac{1}{|I|}\sum_{j\in T} D(i,j)\geq(1-\varepsilon_r) \quad (1)$$

$$\forall k\in I,\ \frac{1}{|T|}\sum_{k\epsilon I} D(i,j) \geq (1-\varepsilon_c) \quad (2)$$

$$\forall (j\in T),\ (k\in I),\ \frac{1}{|IJ|}\sum_{j\in T, k\in I} D(i,j) \geq (1-\varepsilon) \quad (3)$$

In this algorithm, the most important criterion in the selection and the discovery of genes on a subset is defined as support and compared with the predetermined value (minsup) that is determined by the user and $minsup \in (0,1)$. If in a subset, support criterion is larger than minsup, it is called a frequent set and can be accepted as a bicluster. Another criteria is Range Support, which is calculated in all subsets and it is expected to be less than that of the predetermined value ($Rs \in \mathbb{R}$) introduced by the user(17). To implement the ETI algorithm, the Apriori Rule is used as follows:if there is a frequent pattern then all of its subsets are also frequent. In other words, if a set is not frequent then its subsets cannot be frequent. The Apriori rule is used to implement the algorithm and the name is derived from the way in which it uses the information of the previous stageat any stage thereby reducing the search space. The Apriori rule is a level search method and when searching in stage k, run out can be shifted to stage k+1 and in each stage some of the genes will be deleted according to the criteria(18).In this study a biclustering technique, the ETI algorithm was used for feature selection and to identify patterns of gene expression in prostate cancer. The main goal of this study was to conduct biclustering processing of prostate cancer gene expression data, to identify subgroups of tumors with similar clinical features.

METHOD

Overview of the ETI Algorithm

The first step was to takea scan of each condition and this was done one at a time. For this purpose, value of range support (max-min)/min was measured and compared with RS. The conditions that satisfied this criterion remained for analysis. The ETI algorithm sequentially generates (k+1)-item sets from k-itemsets. So in each step the supporting threshold can be changed and selection of the k-item set was determined as follows

$$minsup^k = minsup.\left[1-\frac{k\varepsilon_r}{1+\lfloor k+\varepsilon_c\rfloor}\right]$$

In addition, for k_{th} step if $[(k+1)\varepsilon_r] > [k\varepsilon_r]$then error-extension level is satisfied and biclusters of (k+1)-item sets can have some genes that have big expression levels, more than k-itemsets and else nonerror-extension level performed that support remain unchanged. Basis on apriori rule, union of biclusters in k-levels is applied for defining biclusters in error-extension for (K+1)-level, or the intersect of biclusters is used. In this algorithm, besides the requirement of an error stage and minsup, it

is also necessary to satisfy the thresholds denoted below

$$|T| \leq n.minsup^k$$

$$\forall j \in T,\ E(D_{j,I}) \leq \varepsilon_r.|I|$$

$$\forall k \in I,\ \left(|T| - E(D_{T,k})\right) \geq n.minsup^k$$

Real Dataset

A real-valued prostate Cancer gene-expression data set was used, taken from Affymetrix Human Genome U133A 2.0 Array. This dataset contained 22277 genes and 4 conditions (normal and MED1 conditions). MED1 is a co-activator of the androgen receptor and other signal-activated transcription factors.

Evaluation

In this study, evaluation of discovered biclusters was performed by test statistics conducted for evaluation of significance of the biclusters. This was performed by constructing 1000 random sample permutations of the gene expression dataset, size 100*8 and then compared to the number of known biclusters identified by the method with the number of discovered biclusters in the random set and calculation of the mont-carlo p-value. Also, for evaluating the result of the algorithm biologically, gene ontology biological process could be the function that measures similarities of expression levels in each biclusterand covers three domains; cellular component, molecular function and biological process. GO Enrichment Validation is a hyper-geometric test for GO enrichment. This statistical test is significant if genes in the biclusters are annotated with GO terms and are not specified by chance.

RESULT

This aim of this section was discovery of subsets of genes that are co-expressed genes in prostate cancer gene expression data. Firstly, the data set is normalized using the median approach and then missing values are imputed by the knn[1] method. R3.2.1 software was used for analysis of this dataset and implementation of the ETI algorithm. Figure 3.1 and 3.2 provides a general overview of all the biclusters obtained by the ETI algorithm on real-valued gene-expression data set using various parameter settings.

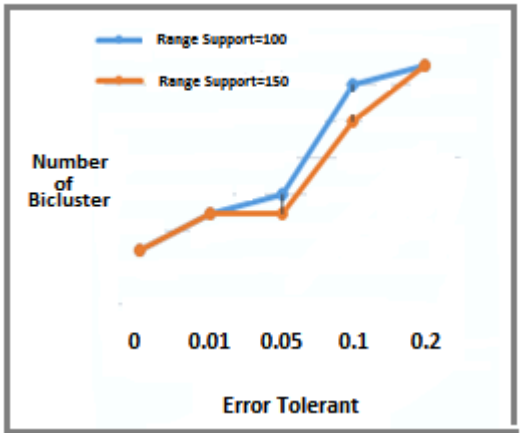

Figure 3.1- number of discovered Bicluster by ETI algorithm for prostate cancer gene expression data

[1] K Nearest Neighborhood

In these figures, the ETI algorithm was applied for different error-tolerance and range support parameters and to represent a number of biclusters and genes discovered. It can be clearly seen that introducing an error-tolerance of 20% substantially increased the total number of biclusters and the number of discovered biclusters increased 3-fold. Also, with an increase of error-tolerance, there was an increase in the number of genes covered. In all situations of error-tolerance, when the range of support was increased, the number of biclusters and genes covered showed a decrease.

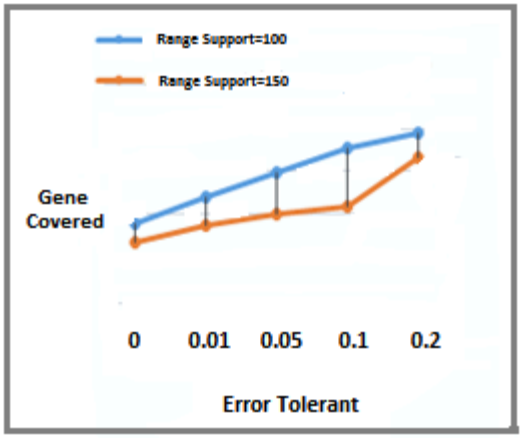

Figure 3.2- Number gene covered of Biclustering result for prostate cancer gene expression data

Table 3.1- result of ETI algorithm for error=0.2, RS=100

Bicluster	Gene number	MSR
1	1759	168.25
2	3406	234.63
3	1819	219.84
4	1794	129.89
5	1843	579.58

In this study, the scenario with error=0.2 and range support=100wasselected for analysis and the result of this scenario was used for gene ontology. For this scenario, Table 3.1showsthe number of genes for five selected biclusters that had top support compared to the others; a 25% overlap is shown between them.

Table 3.2 shows the significant GO terms for the set of genes that was discovered by each bicluster result along with their relative p-value. The web tool David was used to evaluate the discovered biclusters. For each bicluster, the first step was to denote numbers of GO terms and then to evaluate significance of the functions. Table 3.2 shows the significant GO terms for the set of genes discovered by each bicluster along with their relative p-values. The web tool David was used to evaluate enrichment analysis of the discovered biclusters. For each bicluster, the first step

was to denote numbers of GO terms and then calculate the percentage of significance of these GO terms. It is clear from the above analysis that the biomarkers obtained from ETI algorithm were indeed biologically meaningful and the percentage of significance of p-values for all of GO terms was above 75%indicating that the discovered genes in each bicluster had been performed correctly.

Table 3.2- Biological significant of biclusters result

Bicluster	Ontology	Number of GO term	Percent of significant pvalue
			P<0.05
1	Biological Process		96.3
	Molecular Function	6496	92.3
	Cellular Component		89.9
2	Biological Process		95.2
	Molecular Function	8133	94.4
	Cellular Component		92.5
	Biological Process		89.7
3	Molecular Function	7082	91.5
	Cellular Component		85.2
	Biological Process		83.6
4	Molecular Function	6972	77.9
	Cellular Component		74.9
5	Biological Process		81.6
	Molecular Function	6895	76.5
	Cellular Component		76.2

In this study a permutation test was used to evaluate biclusters discovered by ETI and 1000 sampling. For each permuted sample, the number of biclusters was calculated and the number of genes was determined by the algorithm and by random selection. If the number of genes in a random sample was at least at 80 % of the number of the genes discovered by the ETI algorithm then it was accepted that the technique identified biclusters by chance. In this study, the number of identified biclusters that were determined by chance was 13 and p-value=0.013, so the hypothesis of selection by chance was rejected.

DISCUSSION

Various methods have been used in clustering studies to find the optimal subset of genes with the highest similarity. Frequent Item sets is one such technique that is considered reliable. Recently, this technique has been extended to include error-tolerance in implementation and so it is considered a proper technique for finding patterns in gene expression data(19). The aim of this study was to discover patterns of many genes that could be used as a predictive test in patients as well as to exclude genes with low and differential expression. The real-value prostate cancer gene-expression data sets were used which were taken from Affymetrix platform

HGU133A and normalized. This dataset is included 22277genes and 4 conditions. The first step was to identify biclusters using the ETI algorithm with different scenarios for different error parameters and support. Then the scenario with $\varepsilon = 0.2$ and RS=100 was selected for analysis because this scenario produced biclusters with the top support and size. With this scenario, 12 biclusters were discovered, of which their minimum and maximum number of genes were 1759 and 5913 respectively. This study has shown that the discovered biclusters had almost correlated structures under MED1 conditions and so genes in prostate cancer are better expressed in this situation.MSR criterion shows that the genes in a bicluster are similar. Also, randomization method and gene ontology analysis have shown that these results for biclusters were appropriate.

CONCLUSION

The purpose behind this study was to evaluate ETI Algorithm base on different degrees of noise and overlap in prostate cancer gene expression data. Results show that the algorithm can correctly find biclusters when implements with error tolerant and increase of tolerance is concluded increase of number of biclusters and number of genes covered. GO enrichment analysis showed that biological significance of each biclusters is high, especially when the size of biclusters is big. So the ETI algorithm was discovered biclusters properly and could be perform well to find genes and features.

ACKNOELEGEMENT

The authors would like to thank the Academic & Research Affairs of Para Medical Faculty of Shahid Beheshti University of Medical Sciences for financial Support.

REFERENCES

1. Johnson C, Warmoes MO, Shen X, Locasale JW. Epigenetics and cancer metabolism. Cancer Letters. 2013;
2. Bishak YK, Payahoo L, Osatdrahimi A, Nourazarian A. Mechanisms of Cadmium Carcinogenicity in the Gastrointestinal Tract. Asian Pacific J Cancer Prev. 2015;16(1):9–21.
3. Kye SY, Yoo J, Lee MH, Jun JK. Effects of a Cancer Prevention Advertisement on Beliefs and Knowledge about Cancer Prevention. Asian Pacific J cancer Prev APJCP. 2014;16(14):5793–800.
4. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet. Nature Publishing Group; 2000;24(3):227–35.
5. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer

statistics. CA Cancer J Clin. Wiley Online Library; 2011;61(2):69–90.

6. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, et al. Cancer statistics, 2008. CA Cancer J Clin. Wiley Online Library; 2008;58(2):71–96.
7. Jemal A, Center MM, DeSantis C, Ward EM. Global patterns of cancer incidence and mortality rates and trends. Cancer Epidemiol Biomarkers Prev. AACR; 2010;19(8):1893–907.
8. Kalish LA, McDougal WS, McKinlay JB. Family history and the risk of prostate cancer. Urology. Elsevier; 2000;56(5):803–6.
9. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics. Oxford Univ Press; 2002;18(suppl 1):S136–44.
10. Liu X, Wang L. Computing the maximum similarity bi-clusters of gene expression data. Bioinformatics. Oxford Univ Press; 2007;23(1):50–6.
11. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinforma. IEEE Computer Society Press; 2004;1(1):24–45.
12. Gan X, Liew AWC, Yan H. Discovering biclusters in gene expression data based on high-dimensional linear geometries. BMC Bioinformatics. BioMed Central Ltd; 2008;9(1):209.
13. Hartigan JA. Direct clustering of a data matrix. J Am Stat Assoc. Taylor & Francis Group; 1972;67(337):123–9.
14. Cheng Y, Church GM. Biclustering of expression data. Ismb. 2000. p. 93–103.
15. Hahsler M, Grün B, Hornik K, Buchta C. Introduction to arules–A computational environment for mining association rules and frequent item sets. Compr R Arch Netw. Citeseer; 2009;
16. Lopez FJ, Blanco A, Garcia F, Cano C, Marin A. Fuzzy association rules for biological data analysis: a case study on yeast. BMC Bioinformatics. BioMed Central Ltd; 2008;9(1):107.
17. Liu J, Paulsen S, Sun X, Wang W, Nobel A, Prins J. Mining approximate frequent itemsets in the presence of noise: Algorithm and analysis. Society for Industrial and Applied Mathematics Proceedings of the SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics; 2006. p. 407.
18. Han J, Kamber M, Pei J. Data mining: concepts and techniques: concepts and techniques. Elsevier; 2011.
19. Gupta R, Rao N, Kumar V. Discovery of error-tolerant biclusters from noisy gene expression data. BMC Bioinformatics. BioMed Central Ltd; 2011;12(Suppl 12):S1.