Applied Soft Computing 38 (2016) 587–605

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Fuzzy Correlated Association Mining: Selecting altered associations among the genes, and some possible marker genes mediating certain cancers

Anupam Ghosh [a], Rajat K. De [b,*]

[a] Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata 700152, India
[b] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700108, India

ARTICLE INFO

Article history:
Received 23 December 2014
Received in revised form 8 August 2015
Accepted 27 September 2015
Available online 17 October 2015

Keywords:
Transcriptional regulation
p-Value
Biochemical pathways
Functional enrichment

ABSTRACT

Association mining is a well explored topic applied to various fields. In this article, the associations among the genes have been identified from microarray gene expression data. Here a methodology, called Fuzzy Correlated Association Mining (FCAM), is developed for identifying the associations among the genes that have altered quite significantly from normal state to diseased state with respect to their expression patterns. This idea leads to predict the disease mediating genes along with their altered associations. The proposed methodology involves generation of fuzzy gene sets, construction of fuzzy items, computation of fuzzy support for fuzzy items and fuzzy correlation coefficient of a pair of fuzzy items, generation of associations, and identification of altered associations from normal to diseased state. The concept of finding fuzzy correlation between two groups of items, generation of altered associations among the items (groups of items) and then rank these items (groups of items) according to their importance are the novel contribution of the present article. The effectiveness of the methodology has been demonstrated on five gene expression data sets dealing with human lung cancer, colon cancer, sarcoma, breast cancer and leukemia. As a result, some possible genes, like IGFBP3, ERBB2, TP53, HBB, KRAS, PTEN, CALCA, CDKN2A, has been found as important genes that may mediate the development of various cancers considered here. For comparison, we have considered 11 existing association rule mining algorithms. The results are appropriately validated in terms of gene–gene interactions, functional enrichment, biochemical pathways, and using NCBI database.

© 2015 Published by Elsevier B.V.

1. Introduction

Exploratory data mining techniques are needed that can, roughly speaking, be considered as the search for interesting bi-sets, i.e., sets of biological situations and sets of genes, which are associated in some way. Indeed, it is interesting to look for groups of co-regulated genes, for which a reasonable assumption is that they participate in a common function within the cell [1]. Genes are grouped together according to similar expression profiles. The association among a set of co-regulated genes and its discovery pave a way to a better understanding of gene regulation.

Fuzzy set theory is capable of handling uncertainty in the gene expression values arising due to incompleteness, imprecision, noise and experimental errors. Moreover, genes have expression values that are in different intervals under two conditions (i.e., normal or diseased). Although each interval has a well-defined boundary, they are highly overlapped. Fuzzy set theory is especially suitable to model such imprecise and overlapping data. Thus incorporation of the notion of fuzzy sets in the methods enables one to handle such overlapping intervals in a better way [2,3].

The notion of fuzzy sets has been used in the domain of gene expression analysis. They include, among others, development of rule discovery procedure [4] based on knowledge extraction of gene by classification; transformation of gene expression by fuzzy heuristic rule set [5]; classifying fuzzy inference system [6]; development of a fuzzy model for gene regulatory networks [7]; measuring performance of small rule-based classifiers using fuzzy logic [8]; identification of normal and tumor patients using a fuzzy neural network model [9].

Global gene expression profiling, both at the transcript level and at the protein level, can be a valuable tool in the understanding of genes, biological networks and cellular states. As larger and larger gene expression data sets become available, data mining techniques can be applied to identify patterns of interest in the data. Association rules, used widely in the area of market basket analysis, can be applied to the analysis of expression data as well. Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression [10]. Items in gene expression data can include genes that are highly expressed or repressed, as well as relevant facts describing the cellular environment of the genes (e.g., the diagnosis of a tumor sample from which a profile was obtained).

Association rule discovery has been applied to gene expression data, searching for patterns of differential expression across tens of thousands of genes. In

* Corresponding author. Tel.: +91 3325753105; fax: +91 3325783357.
E-mail addresses: anupam.ghosh@rediffmail.com (A. Ghosh), rajat@isical.ac.in (R.K. De).

http://dx.doi.org/10.1016/j.asoc.2015.09.057
1568-4946/© 2015 Published by Elsevier B.V.

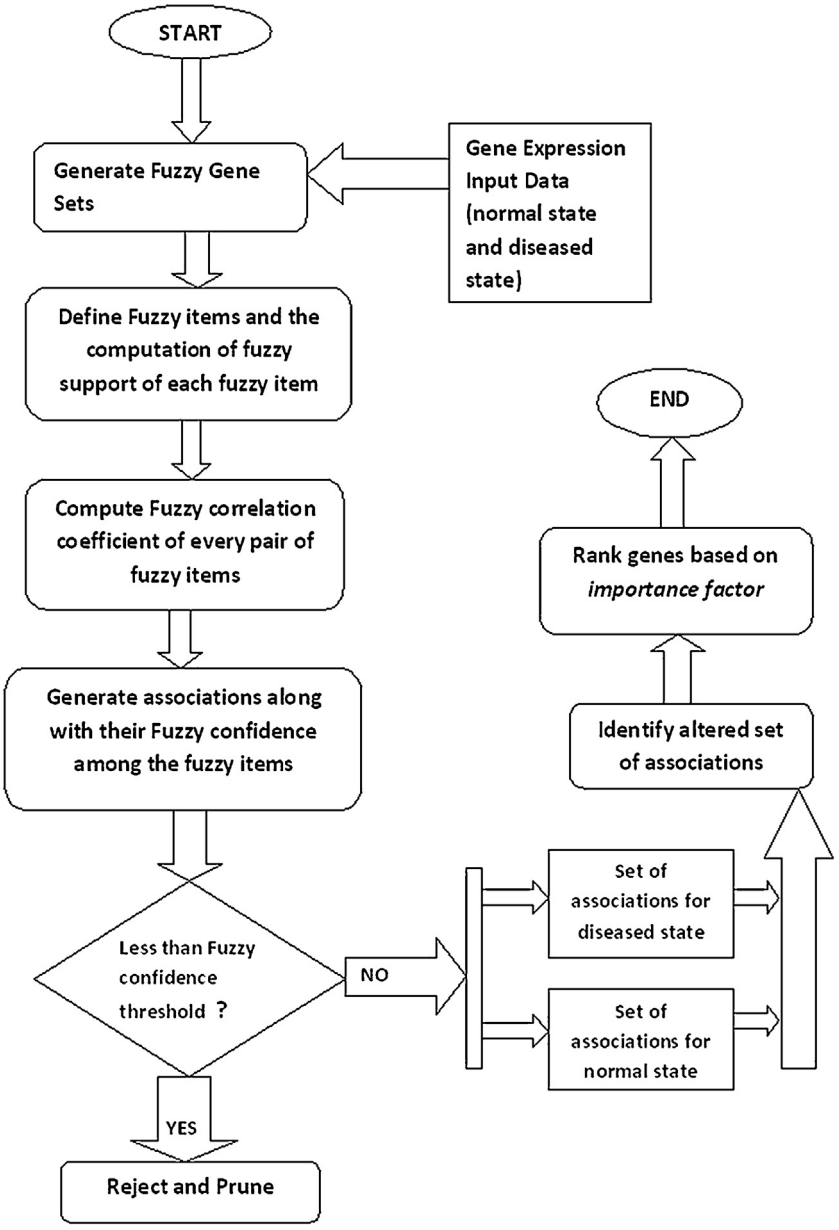

Fig. 1. Flowchart of FCAM.

life threatening diseases, such as cancer, where the effective diagnosis includes annotation, early detection, distinction, and prediction, data mining and statistical approaches offer the promise for precise, accurate, and functionally robust analysis of gene expression data [11]. The computational extraction of derived patterns from microarray gene expression is a non-trivial task that involves sophisticated algorithm design and analysis for specific domain discovery. In an earlier investigation, a model was proposed for feature extraction by first applying feature selection heuristics based on the statistical impurity measures like Gini Index, Max Minority, and the Twoing Rule and obtaining the top 100–400 genes and then analyze the associative dependencies between the genes and assign weights to the genes based on their degree of participation in the rules [12]. An analysis of some of these rules reveals numerous associations among certain genes, many of which make sense biologically, others suggesting new hypotheses that may warrant further investigations.

In this article, an association rule-mining algorithm, called fuzzy correlated association mining (FCAM), has been developed, which uses gene expression data to generate the association among the large set of genes and more importantly to identify the set of altered associations. The algorithm considers a large set of genes and determines the disease mediating genes (marker genes) (Fig. 1).

Unlike the proposed algorithm FCAM, there is no method that determines gene–gene associations that have altered from normal state to diseased one and thereby finding possible disease mediating genes. Thus, identifying the associations among the genes, discovering the altered associations from gene expression data and finally the proposal of gene ranking technique that is used to identify the importance of the genes from the set of altered associations may be considered as a novel concept. Biological data are often imprecise and noisy. Moreover, in microarray gene expression data, genes have expression values that are in different intervals under two conditions, i.e., normal or diseased. Although each interval has a well-defined boundary, they are highly overlapped. Fuzzy set theory is especially suitable to model such imprecise and overlapping data. This idea leads us to develop the methodology using the concept of fuzzy set theory. Thus incorporation of the notion of fuzzy sets in the proposed method enables one to handle such overlapping intervals in a better way.

We have applied the proposed methodology (FCAM) (Fig. 1) on gene expression data sets of 5 different human cancers (lung, colon, lymphocyte, sarcoma and breast). There exist several methods for finding gene–gene associations. These existing methods may be classified into 3 categories, *viz.*, Category 1, Category 2 and Category 3. The methods under Category 1 consider finding gene–gene associations from structural data, i.e., sequence data [13–15]. Under Category 2, the methods determine associations among gene functions and protein–protein interactions by using gene expression data as well as some other databases [16–20]. The methods, under Category 3, determine association among objects in a non-biological domain [21–26]. Since we could not find any method that works exactly the same way as FCAM does, we have considered 11 methods under Categories 2 and 3 for comparison. Although it can extract the associations among the

genes from microarray gene expression data, we can put FCAM under Category 2.

The superior performance of FCAM has been demonstrated in terms of gene–gene interactions, functional enrichment, biochemical pathways and based on some earlier information available in NCBI database. Moreover, *F*-score statistics has been used to validate the result (for details about *F*-score, one may see Appendix B).

Section 2 mainly deals with different existing methods of gene–gene interactions and its related investigations. Section 3 is the methodology, i.e., the description of FCAM (Fig. 1) consists of several subsections like Mathematical Preliminaries (Section 3.1), Algorithm for generation of associations (Section 3.2), Altered set of associations from normal to diseased state (Section 3.4) and Ranking of genes based on their importance (Section 3.5). Section 5 provides the result which consists of several subsections, namely, Description of Datasets (Section 5.1), Analysis of the results (Section 5.2) and Comparisons (Section 5.3). The comparison is done using gene–gene interactions, biochemical pathways, database and functional enrichment. In addition, some possible reasons behind superior performance of FCAM have been explained in Section 5.3. Finally, the paper concludes in Section 6. Further, the proofs of three lemmas, definition of *F*-score, and some more tables and figures are provided in Appendix.

2. Related works

In this section, we describe briefly some of the methods under Categories 2 and 3, as they have been considered for comparison. It is to be mentioned here that the methods [13–15] under Category 1 is not considered here as they deal with structural data related to genes/proteins, unlike FCAM. The algorithms in [16–20,27–31] and also FCAM fall under Category 2 as the methods determine the gene–gene interactions by using the gene expression data. Similarly, we have found a set of algorithms [21–23,26,32–39] that may be classified into Category 3 as they determine the associations among the objects in non-biological domains. Here we have considered Fuzzy FP-Growth (FFPG) [20], Ant-ARM [16] and Li-An [19] methods under Category 2, and Apriori [32], FP-Growth [33], T-Apriori [34], H-Apriori [34], CFARM [35], FCBAR [21], FWFPG [22] and FHARM [39] under Category 3, for comparison.

Both FCAM and FFPG work under fuzzy framework. Unlike FCAM, FFPG the number of requires extra processing time during the mining process as it continuously generates large conditional bases and conditional FP-tree. On the other hand, Ant-ARM is based on crisp environment with higher time complexity as the concept is based on the swarm intelligence and association rule mining. Li-An method also works under crisp environment and generates the clusters during the generation of the associations, which actually increases the time complexity with compared to FCAM.

The algorithms CFARM, FCBAR, FWFPG, FHARM and FCAM are all based on fuzzy set theoretic approach. CFARM selects the rule based on certainty factors. FCBAR constructs the dynamic tree structure for implementing the fuzzy clustering and subsequently the tree pruning for mining the rules. On the other hand, FCAM uses fuzzy correlation to generate the rules and mine the significant rules based on fuzzy confidence threshold values. FWFPG combines the concept of fuzzy weight and fuzzy partition methods in data mining, and uses FP-Growth for rule mining, which requires huge mining cost compared to FCAM. However, both FCAM and FWFPG uses triangular membership functions. FHARM uses fuzzy support and confidence values as well as correlation for interestingness measures to generate the rules where as FCAM measures it using fuzzy confidence threshold values.

3. Fuzzy Correlated Association Mining (FCAM)

In this section, we describe the proposed methodology, called Fuzzy Correlated Association Mining (FCAM) (Fig. 1). Let us consider an expression data for a set of n genes $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, for each of which m expression values are given. Let $\mathcal{G}$ be the set of n genes $\{g_1, g_2, g_3, \ldots, g_n\}$. For each gene g_i, there is an m-dimensional vector $\mathbf{x}_i$, where x_{il} ($l = 1, 2, \ldots, m$) is the lth expression value of g_i.

3.1. Mathematical preliminaries

Before describing the algorithm, some functions that are needed by FCAM, are defined.

3.1.1. Generation of fuzzy gene sets

The set of expression values of a gene constitute a fuzzy set, called a fuzzy gene set. The membership function μ_{il} of a fuzzy gene set corresponding to ith gene, is defined as

$$\begin{aligned}\mu_{il} &= \max\left\{0, \frac{1}{2} + \frac{1}{2} \times \left(\frac{x_{il} - x_{\min_i}}{x_{c_i} - x_{\min_i}}\right)\right\}, \quad \text{where } 0 < x_{il} \le x_{c_i},\\ &= \max\left\{0, \frac{1}{2} + \frac{1}{2} \times \left(\frac{x_{\max_i} - x_{il}}{x_{\max_i} - x_{c_i}}\right)\right\}, \quad \text{where } x_{il} \ge x_{c_i},\\ &= 0, \qquad \text{otherwise,}\end{aligned} \tag{1}$$

where $\mu_{il} \in [0, 1]$ is the fuzzy expression level of ith gene in lth sample. The terms $x_{\min_i}$ and $x_{\max_i}$ are minimum and maximum expression values of ith gene, respectively, and $x_{c_i} = \frac{x_{\max_i} + x_{\min_i}}{2}$.

That is, a fuzzy set is considered around a gene g_i in such a way that at $x_{il} = x_{\min_i}$ or $x_{il} = x_{\max_i}$ (i.e., at the boundary), the degree of belongingness (fuzzy expression level) of gene g_i to the fuzzy set is 0.5, i.e., the situation of most uncertainty. At $x_{il} = x_{c_i}$, the degree of belongingness of gene g_i to the fuzzy set is 1 (maximum) as most of the expression values are expected to be around x_{c_i} for a particular situation (normal or diseased). Thus a gene is represented by a set of membership values corresponding to their expression values in different samples (normal or diseased). In other words, each membership degree of a gene represents the truth value of the gene to the corresponding fuzzy set with respect to an expression value.

3.1.2. Fuzzy items and fuzzy support for a fuzzy item

Let $G_r = (g_{r_1}, g_{r_2}, \ldots, g_{r_p})$ be a list of genes of length p. Here we call G_r a fuzzy item. The fuzzy support for a fuzzy item G_r is defined as

$$F_{\text{sup}}(G_r) = \frac{1}{m}\sum_{l=1}^{m} \min_{g_i \in \mathcal{G}_r} \mu_{il} \tag{2}$$

where $\mathcal{G}_r$ is the set of genes that are in the list G_r.

3.1.3. Fuzzy correlation coefficient of a pair of fuzzy items

Let $\bar{\mu}_{1l}$ and $\bar{\mu}_{2l}$ be two p and q dimensional vectors corresponding to two fuzzy items G_1 and G_2 of lengths p and q, respectively. An ith (jth) component of $\bar{\mu}_{1l}$ ($\bar{\mu}_{2l}$) is μ_{il} (μ_{jl}) (computed using Eq. (1)), where g_i (g_j) is involved in G_1 (G_2). The number of genes involved in G_1 is p and that in G_2 is q. The *fuzzy correlation coefficient* between two fuzzy items G_1 and G_2 is defined as [40,41]

$$F_{corr}(G_1, G_2) = \frac{S(G_1, G_2)}{S(G_1) \times S(G_2)} \tag{3}$$

where

$$S(G_1, G_2) = \frac{1}{(m-1)} \times \sum_{l=1}^{m} |\mu_{1l} - \bar{\mu}_1| \times |\mu_{2l} - \bar{\mu}_2|, \tag{4}$$

$$S(G_r) = \sqrt{\frac{1}{(m-1)} \times \sum_{l=1}^{m} |\mu_{rl} - \bar{\mu}_r|^2}, \quad r = 1, 2 \tag{5}$$

and

$$\bar{\mu}_r = \frac{1}{m} \times \sum_{l=1}^{m} \mu_{rl}, \quad r = 1, 2 \tag{6}$$

Once $F_{corr}(G_1, G_2)$ is computed, it needs to be tested to determine if it is significant. For this purpose, t-test has been used, where t is given by

$$t = \frac{F_{corr}(G_1, G_2) - T_{corr}}{\sqrt{\frac{(1-F^2_{corr}(G_1,G_2))}{m-2}}} \tag{7}$$

The term $T_{corr} > 0$ is a predefined threshold value for fuzzy correlation coefficient, called *minimal fuzzy correlation coefficient*. Then it is checked whether t is at least $t_{(1-\alpha),(m-2)} > 0$, where $t_{(1-\alpha),(m-2)}$ is the $(1-\alpha)$th percentile in the t distribution with degree of freedom $(m-2)$. If $t_{(1-\alpha),(m-2)} > 0$ then it is concluded that $F_{corr}(G_1, G_2)$ is greater than the predefined *minimal fuzzy correlation coefficient* T_{corr} with the level of significance of $(1-\alpha)$.

3.1.4. Fuzzy confidence of the associations

From the aforesaid fuzzy items, associations are generated. An association $\langle G_1, G_2 \rangle$ indicates that genes in G_2 will be over/under expressed if genes in G_1 are over expressed. After generating the associations, the fuzzy confidence of each association $\langle G_1, G_2 \rangle$, denoted by $F_{conf}(\langle G_1, G_2 \rangle)$, is calculated by [42],

$$F_{conf}(\langle G_1, G_2 \rangle) = \frac{F_{\sup}(G_3)}{F_{\sup}(G_1)} \tag{8}$$

where G_3 is the list of all the genes involved in either G_1 or G_2. After computation of the fuzzy confidence values for the associations, they will be compared with *minimal fuzzy confidence threshold* C_F. If the fuzzy confidence value of an association is less than C_F, then the association is rejected and pruned from the list. Otherwise, it will be in the list of associations.

3.2. Algorithm: Generation of associations

Here we describe the algorithm. It starts with computing μ_{il} (Eq. (1)), for all $i = 1, 2, \ldots, n$ and $l = 1, 2, \ldots, m$. The algorithm is an iterative one; the computations in different iterations are narrated below.

- **Begin**
- *Step 0*: Initialize S_F, α, T_{corr}, C_F.
- *Step I*: Generate terms and $F_{sup}(T_r)$ from T_r where $r = 1$.
  - *Step I.1*: Set $r = 1$ and $T_r = \{(g_i) | i = 1, 2, \ldots, n\,; g_i \in T_r\}$.
  - *Step I.2*: If $F_{sup}(T_r) \geq S_F$ then assign $L_r = T_r$.
  - *Step I.3*: If $|L_r| = \phi$, a null set, then no association is generated and the algorithm terminates. Otherwise, set $r = r + 1$ and go to Step II.
- *Step II*: Generate terms T_r, where $r \geq 2$.
  - *Step II.1*: Generate T_r from T_{r-1}, such that $T_r = T_{r-1} \bowtie T_{r-1}$. The operation '$\bowtie$' is the join operation [32].
  - *Step II.2*: If $F_{sup}(T_r) \geq S_F$ and $t(T_r) \geq t_{(1-\alpha),(m-2)}$ then assign $L_r = T_r$.
  - *Step II.3*: If $|L_r| = \phi$, a null set, then go to Step III. Otherwise, set $r = r + 1$ and go to Step II.
- *Step III*: Generate associations from L_{r-1}.
  - *Step III.1*: Compute C_{r-1}=Association(L_{r-1}).
  - *Step III.2*: If $F_{conf}(C_{r-1}) \geq C_F$ then assign $C_{FINAL} = C_{r-1}$.
- **End**

- **Begin Function ()**
  - **Function:** Association(L_j).
  - **Define:** Set $L_j = (G_j, G_{j+1}, \ldots, G_k)$ where $k = 2^{j-1}$ and Set G_j=j-term, *i.e.*, $G_j = (g_1, g_2, \ldots, g_j)$.
  - Association(L_j) = $\cup_{j=1}^{k}$ Association(G_j).
  - return.
- **End Function ()**

- **Begin Function ()**
  - **Function:** Association(G_r).
  - **Define:** Set G_r=r-term, *i.e.*, $G_r = (g_1, g_2, \ldots, g_r)$, and define G_i such that $G_i \subset G_r$ $\forall i = 1, 2, \ldots, (r-1)$.
  - Association(G_r)=$\cup_{i=1}^{r-1} < G_i, (G_r - G_i) >$.
  - return.
- **End Function ()**

3.3. Example

Input: In this example, we have considered 5 genes, *viz.*, g_1, g_2, g_3, g_4, g_5 with expression values over 5 samples. The dataset is mentioned in Table 1.

Initialization: Here, each parameter is initialized, and $S_F = 0.30$, $\alpha = 0.10$, $T_{corr} = 0.20$, $C_F = 0.95$, have been considered for the experiment. Since $m = 5$ and $\alpha = 0.10$, $t_{(1-\alpha),(m-2)} = 1.638$.

Generation of Fuzzy gene sets: Based on the aforesaid initialization and input data set, the fuzzy gene set has been determined (Table 1).

Computation in 1st iteration: Generation of terms and their fuzzy support. In the very first step, T_1 term set has been generated. T_1 is set of 1-gene set. After that the fuzzy support (F_{sup}) of T_1 has been calculated (Table 2). Hence, L_1 has been generated by pruning T_1 using fuzzy support threshold $S_F = 0.30$ (Table 2). Since, $L_1 \neq \phi$, a null set T_2 has been generated.

Computation in 2nd iteration: Generation of terms and their fuzzy supports, t values. In the 2nd iteration, T_2 has been generated from L_1. T_2 is set of 2-gene set. Now, F_{sup} and t-value (using fuzzy correlation coefficient) for T_2 has been calculated (Table 2). After that, L_2 has been generated by pruning T_2 using fuzzy support threshold $S_F = 0.30$ and predefined $t = 1.638$ (Table 2). As $L_2 \neq \phi$ then T_3 has been generated.

Beginning of 3rd iteration: Generation of terms and their fuzzy supports, t values. In the 3rd iteration, T_3 has been generated from L_2. T_3 is set of 3-gene set and 4-gene set. Here, T_3 generates 3-gene set. Now, F_{sup} and t-value (using fuzzy correlation coefficient) for T_3 has been calculated (Table 2). After that, L_3 has been generated by pruning T_3 using fuzzy support threshold $S_F = 0.30$ and predefined $t = 1.638$ (Table 2). Since $L_3 = \phi$ no further iteration has been required, and association has been generated from L_2.

Generation of associations C_2 from L_2: The associations (C_2) has been generated from L_2. The fuzzy confidence F_{conf} of C_2 has been calculated (Table 3), and final association set C_{FINAL} has been generated by pruning C_2 using minimal fuzzy confidence $C_F = 0.95$ (Table 3).

Lemma 1. *The number of elements in T_r in rth iteration is $|T_r| = \sum_{i=r}^{2^{r-1}} \binom{n}{i}$, $r \geq 1$, $n \geq 2^{r-1}$. Proof of the lemma is given in Appendix A.1. However, the algorithm terminates in iteration R, if $L_R = \phi$, $R \geq 1$, and the set of significant associations is generated from L_{R-1}. After determining the significant associations, the fuzzy confidence value is computed for each association (Eq. (8)). The set of associations finally reduced by pruning some associations based on minimal fuzzy confidence (C_F) value.*

Lemma 2. *The maximum number of iterations of FCAM is $\lceil 1 + \log_2 n \rceil$. Proof of the lemma is given in Appendix A.2.*

Table 1
Input dataset and corresponding fuzzy gene set.

	Gene name	Expression value				
		Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Input dataset	g_1	170	59.7	80	92.4	104
	g_2	69.4	18.1	26	96.9	72.8
	g_3	250.7	146.8	150	177.8	228.7
	g_4	957.1	186.8	340.2	515.8	540.8
	g_5	25.4	−7.7	−16.3	18	26
Fuzzy gene set	g_1	0.500000	0.848985	0.500000	0.500000	0.514184
	g_2	0.500000	0.500000	0.500000	0.500000	0.000000
	g_3	0.684044	0.600254	0.530799	0.699143	0.000000
	g_4	0.796464	0.500000	0.798364	0.927106	0.689125
	g_5	0.901632	0.805838	0.711742	0.959561	0.500000

Table 2
Fuzzy support, t-value of terms in different iterations and pruning.

	Terms (T_1)	F_{sup}	t-value
Fuzzy support of terms in 1st iteration	(g_1)	0.676428	
	(g_2)	0.651015	
	(g_3)	0.608181	
	(g_4)	0.717162	
	(g_5)	0.340662	
Fuzzy support of terms in 1st iteration after pruning	(g_1)	$0.676428 > S_F$	
	(g_2)	$0.651015 > S_F$	
	(g_3)	$0.608181 > S_F$	
	(g_4)	$0.717162 > S_F$	
	(g_5)	$0.340662 > S_F$	
Fuzzy support and t-value of terms in 2nd iteration	(g_1, g_2)	0.581218	−0.229618
	(g_1, g_3)	0.607801	2.129630
	(g_1, g_4)	0.676428	7.916760
	(g_1, g_5)	0.337825	0.503083
	(g_2, g_3)	0.548508	−0.613424
	(g_2, g_4)	0.581218	−0.386906
	(g_2, g_5)	0.302837	0.281405
	(g_3, g_4)	0.608181	3.279048
	(g_3, g_5)	0.337825	1.298586
	(g_4, g_5)	0.337825	0.706486
Fuzzy support and t-value of terms in 2nd iteration after pruning	(g_1, g_3)	$0.607801 > S_F$	$2.129630 > t_{0.9,3}$
	(g_1, g_4)	$0.676428 > S_F$	$7.916760 > t_{0.9,3}$
	(g_3, g_4)	$0.608181 > S_F$	$3.279048 > t_{0.9,3}$
Fuzzy support and t-value of terms in 3rd iteration	(g_3, g_1, g_4)	$0.607801 > S_F$	$0.737023 < t_{0.9,3}$
Fuzzy support and t-value of terms in 3rd iteration after pruning	Null		

Table 3
Associations C_{r-1} or C_2 with their fuzzy confidence and final associations C_{FINAL}.

	Associations (C_2)	$F_{conf}(C_2)$
Associations C_{r-1} or C_2 with their fuzzy confidence	$<g_3, g_1>$	0.999375
	$<g_1, g_3>$	0.898545
	$<g_4, g_1>$	0.943201
	$<g_1, g_4>$	1.000000
	$<g_4, g_3>$	0.848038
	$<g_3, g_4>$	1.000000
Associations C_{FINAL} after pruning based on $C_F = 0.95$	$<g_3, g_1>$	$0.999375 > C_F$
	$<g_1, g_4>$	$1.000000 > C_F$
	$<g_3, g_4>$	$1.000000 > C_F$

Lemma 3. *The number of associations in candidate set C_r in rth iteration is* $|C_r| = \sum_{i=r}^{2^{r-1}} (2^{i-1} - 1) \binom{n}{i}$, $r \geq 1$, $n \geq 2^{r-1}$. *Proof of the lemma is given in Appendix* A.3.

3.4. Determining altered set of associations from normal to diseased state

The algorithm is applied to both normal and diseased samples of the gene expression data. Thus two sets A_N and A_D of associations, corresponding to normal and diseased samples, have been obtained. Now the set of associations that have altered from normal state to diseased state, has been identified. The altered set of associations $A = (A_N \cup A_D) - (A_N \cap A_D)$ is identified. This set of altered associations indicates that they influence mediating the development of the disease.

3.5. Ranking of genes based on their importance

Once the set A for the data set has been identified, we consider the genes involved in the set of associations that have altered from normal state to diseased state. Here we describe a gene ranking mechanism based on their importance in the altered association set. In order to measure the importance of a gene, a parameter, called *importance factor*, is introduced. The *importance factor* of gene g_i, denoted by $IMP(g_i)$ ($\in [0, 1]$), is defined as

$$IMP(g_i) = \frac{V_{association}(g_i)}{|A|} \quad (9)$$

Here, $|A|$ is the number of associations that have altered from normal to diseased condition. $V_{association}(g_i)$ indicates the number of altered associations in which g_i is involved. Higher the value of $IMP(g_i)$, higher is the importance of g_i in mediating the disease and vice versa. Based on these *IMP*-values, the genes are ranked according to their importance in mediating a disease.

4. Computational cost

In this section, we shall derive an estimate for the cost incurred in the computation. We have done it on n, number of genes.

4.1. Cost for generating the associations

The entire method starts with generation of term sets and then associations are generated from the terms set in each iteration. The number of computations required for generating the associations in rth iteration of n genes is $\sum_{i=r}^{2^{r-1}} (2^{i-1} - 1) \binom{n}{i}$. Hence the computational complexity for generating the associations for n genes in the iteration is $O(n^c)$ where c is the constant. Thus, the computational complexity required for generating the associations in each iteration of n genes is $O(n^c)$.

4.2. Cost for generating the associations in all iterations

We can get $n \geq 2^{(r-1)}$, where r is the number of iterations. Thus, for the maximum value of $r = R$, $2^{R-1} = n$. That is, $R = 1 + log_2\, n$. If n is not a power of 2 then $R = \lceil 1 + log_2\, n \rceil$. Thus maximum number of iterations of the algorithm requires $\lceil 1 + log_2\, n \rceil$. Hence the computational cost for the algorithm is $O(1 + log_2\, n) * O(n^c)$ or $O(n^c * log_2\, n)$.

5. Results

The effectiveness of Fuzzy Correlated Association Mining (FCAM), along with comparisons, is demonstrated on five cancer gene expression datasets, *viz.*, lung (7129 genes with 10 normal lung and 86 tumor samples), breast (22,645 genes with 2 normal and 4 tumor samples), colon (6600 genes with 18 normal and 18 tumor samples), sarcoma (22,283 genes with 15 normal and 39 diseased samples) and leukemia (22,283 genes with 13 normal and 43 diseased samples). We have considered the methods, like Composite Fuzzy Association Rule Mining (CFARM), Fuzzy Cluster-Based Association Rules (FCBAR), Fuzzy Frequent Pattern Growth (FFPG), Fuzzy Weighted Frequent Pattern Growth (FWFPG), Apriori, FP-Growth, T-Apriori, Hybrid Apriori (H-Apriori), Ant-ARM, Li-An method and FHARM, for comparisons.

For every dataset, the parameters of each algorithm have been set to values in such a way that the number of selected genes obtained by each algorithm is closed to that in NCBI database [43].

5.1. Description of datasets

Here we describe briefly five microarray gene expression datasets. Table 4 shows the number of genes and expression values in these datasets.

5.1.1. Human lung expression data

Human lung gene expression data has been obtained by Affymetrix microarray experiments with tumors and normal lung samples [44]. In this data set, there are 7129 genes (more specifically, Affymetrix probe-sets) for 86 lung tumor and 10 normal lung samples. The gene expression profiles represent 86 primary lung adenocarcinomas, including 67 state I and 19 state III tumors, as well as 10 neoplastic lung samples. More details on this data set can be found in [44]. Database web link for this data is [49].

5.1.2. Human colon expression data

In human colon expression data, the expression level of 6600 genes in [45] 18 tumor and 18 normal samples have been considered. The data set is available at [43]. Samples were obtained from colon adenocarcinoma specimens snap-frozen in liquid nitrogen within 20 min of removal/collection from patients. From some of these patients, paired normal colon tissue was also obtained. The tissue was snap-frozen in liquid nitrogen within 20–30 min of harvesting and stored thereafter at −80 °C. mRNA was extracted from the bulk tissue samples and hybridized (using Affymetrix Hum600) to the array using standard procedure. The adenocarcinoma samples were specifically re-reviewed by a pathologist where the samples were obtained using paraffin-embedded tissue that was adjacent or in close proximity to the frozen sample from which the mRNA was extracted.

5.1.3. Human breast cell expression data

Human breast cell expression data constitutes expression level of 22,645 genes [46]. The array based gene expression profiling of breast cancer cell lines HCC 1954 and MDA-MB-436 were considered in reference to mammary epithelial cells. In this data set, there are 6 samples; two samples are for normal breast epithelium control replicate human mammary epithelial cells, and remaining 4 samples for breast cancer cells. Database web link for this data is [43].

5.1.4. Human soft tissue sarcoma expression data

Human soft tissue sarcoma expression data consists of expression profiling of soft tissue sarcoma samples of *Homo sapiens*. Hypoxic regions often develop in tumors as they increase in size. Results provide insight into the expression of hypoxia-related genes in sarcomas under oligonucleotide technology. In this data set, there are 22,283 genes with 15 normal samples and 39 diseased samples [47]. Among these 39 diseased samples, 7 fibrosarcoma samples, 2 GIST (gastrointestinal stromal) samples, 6 Leiomyosarcoma samples, 4 dedifferentiated liposarcoma samples, 3 pleomorphic liposarcoma samples, 9 MFH (malignant fibrous histiocytoma) samples, 4 Round cell sarcoma samples and 4 Synovial sarcoma samples are present. The data set can be available at [43].

5.1.5. Human lymphocytes and plasma cell expression data

Human lymphocytes and plasma cell expression data is related to Waldenstrom's macroglobulinemia (B lymphocytes and plasma cells) [48]. It has been used for analysis of B lymphocytes (BL) and plasma cells (PC) from patients with Waldenstrom's macroglobulinemia (WM), a B-lymphoproliferative disorder (BLPD). The entire data set consists of expression level of 22,283 genes in 56 samples. Among them, there are 13 normal samples (8 normal B lymphocytes and 5 normal plasma cells) and 43 diseased (20 Waldenstrom's macroglobulinemia, 11 chronic lymphocytic leukemia, 12 multiple myeloma) samples.

5.2. Analysis of the results

At first, FCAM is applied to normal samples of human lung expression data. In normal state, 112 associations have been generated. Here the threshold value for t has come out to be 1.39, *i.e.*, $t_{(1-\alpha),(m-2)} = t_{0.9,8} = 1.39$, where m (=10), the number of samples, is the degree of freedom. Likewise, the algorithm is applied to diseased samples of lung expression data. Here the value of m is 86. Thus, the threshold t-value has become 1.29 ($t_{(1-\alpha),(m-2)} = t_{0.9,84} = 1.29$). As a result, 97 associations have been

Table 4
Highlights of the datasets.

Sl. No.	Dataset	Number of genes	Number of normal samples	Number of diseased samples	Reference
1.	Lung	7129	10	86	[44]
2.	Colon	6600	18	18	[45]
3.	Breast	22,645	2	4	[46]
4.	Sarcoma	22,283	15	39	[47]
5.	Leukemia	22,283	13	43	[48]

determined for the diseased state. Now, these two sets of associations (one set for normal state and another set generated from diseased state) have been compared. After comparing these two sets, 55 associations that have altered from normal to diseased states of lung, have been obtained. Table C.5 in Appendix C shows these 55 altered associations, for lung expression data, along with their fuzzy confidence values. It is to be noted that 77 associations are common in both the states (i.e., normal and disease). Finally, the genes involved in these 55 associations (Table C.5 in Appendix C) have been ranked based on *IMP* values (Eq. (9)). Likewise, 50 associations for colon, 112 associations for soft tissue sarcoma, 96 associations for lymphocytes and plasma cell, and 108 associations for breast have been found, by FCAM, to have altered from normal state to carcinogenic state in the respective organs. We have not listed them to improve the conciseness of the article. The genes involved in these associations have been considered and ranked using *IMP*-values.

Table C.6 in Appendix C provides the numbers of associations obtained by all the methods implemented for all the datasets. The entries in the table correspond to the number of associations ($|A_N|$ and $|A_D|$) obtained on gene expression data under both normal and diseased states, that obtained on any of gene expression data under normal or diseased ($|A_N \cup A_D|$) and also on both $|A_N \cap A_D|$. Likewise, the numbers $|(A_N \cup A_D) - (A_N \cap A_D)|$ of altered associations obtained by all the methods for all the datasets are also depicted in the table. The table shows that FCAM has been able to identify maximum number of altered associations for most of the datasets. Table C.7 in Appendix C provides the parameter values used by the methods to obtain the sets of above associations.

5.3. Comparisons

In this section, we have made a comparative study of the results derived from gene expression datasets (lung, colon, breast, sarcoma, leukemia) using FCAM and 11 other methods (under Categories 2 and 3). We have used some criteria like gene–gene interactions and function enrichment for comparisons. The comparisons are also based on biochemical pathways and NCBI database.

5.3.1. On gene–gene interactions

Here, we compare the results obtained by FCAM with other methods, in terms of gene–gene interactions. By gene–gene interactions, we mean corresponding protein–protein interactions. For this purpose, we have mainly focussed on the fact that how correctly FCAM has discovered the associations in terms of gene–gene interaction with respect to other existing methods.

In NCBI, we get pathway related information from bio-system database. The bio-system consists of a set of genes and their interactions with other genes. We have found some cancer specific pathways from bio-system (pathway) database of NCBI [43]. In the database, we have found non-small cell lung cancer, small cell cancer, colorectal cancer, chronic myeloid and acute myeloid leukemia related pathways. These pathways are related to human lung, colon and lymphocyte and plasma cells. However, we could not find similar pathways for human sarcoma and breast cells. From this database, we have collected the information related to gene–gene interaction for leukemia, lung and colon cancers. The list of interactions for three cancers have been compared with associations generated by the 12 methods (including FCAM) for normal state of lung, colon, and lymphocyte cells.

For human lung expression dataset, 129 interactions among the genes in normal state have been identified using FCAM. The database provides 126 such interactions for human lung. We have found 98 interactions (true positive) that are common in both the sets. Likewise, we have calculated false positive and false negative. In this way, we have computed the aforesaid parameters for remaining 11 existing methods for lung. Finally, we have calculated these parameters for all the cases except breast and sarcoma. From Fig. 2, it is clearly observed that our method FCAM produces the high *F*-score value compared to the other existing methods for all the three datasets. It is to be noted that FCAM (Fig. 2) also generates less number of false positives and false negatives compared to almost all the other methods. From Fig. C.10 in Appendix C, it is clearly observed that for Category 2, the method FFPG under fuzzy frame work performs much better (higher *F*-score value) than the methods under crisp frame work (Ant-ARM and Li-An) for lung, leukemia and colon cancer datasets. Similarly, for Category 3, methods under fuzzy frame work (i.e., CFARM, FCBAR, FWFPG, FHARM) perform better than that under crisp frame work (i.e., Apriori, FP-Growth, H-Apriori, T-Apriori) for all the cancer datasets. Thus the methods under fuzzy framework performs that better than that under crisp frame work both for Categories 2 and 3. On the other hand, it is clearly observed that our proposed method FCAM (from Fig. C.9 in Appendix C) results in higher *F*-score value than all the methods under both the Categories (Fig. C.10 in Appendix C). Thus, we can say that FCAM is the best method compared to other methods considered here, in terms of gene–gene interactions.

5.3.2. Using biochemical pathways

From the aforesaid pathways, the genes (protein) involved in these pathways have been identified. Now we consider the altered associations in such a way the set of genes involved in these associations has become similar to the set of genes involved in the pathways. For lung cancer, we have found non-small cell lung cancer and small cell cancer pathways. A set of 409 genes are involved in these two pathways. For lung expression data (7129 genes), top ranked 409 genes have been considered from the results obtained by all the 12 methods. These genes have been compared with 409 genes (Table C.7 in Appendix C). FCAM has identified 415 genes that are involved in altered associations. Here 342 genes have been identified, which are common in database information and the results of FCAM. We have called these genes as *true positive* (*TP*) genes. Thus we have 73 genes that are within the set of 415 genes (obtained by FCAM) but not involved in the pathways. These 73 genes are considered as *false positive* (*FP*). Similarly, the number of *false negative* genes is 67 for FCAM. Likewise, the numbers of true positive, false positive and false negative genes have been computed for remaining 11 methods.

For human colon expression data, we have found 62 genes that are present in a colon cancer related pathway (i.e., colorectal cancer pathway). Similarly, 355 genes have been found in leukemia related

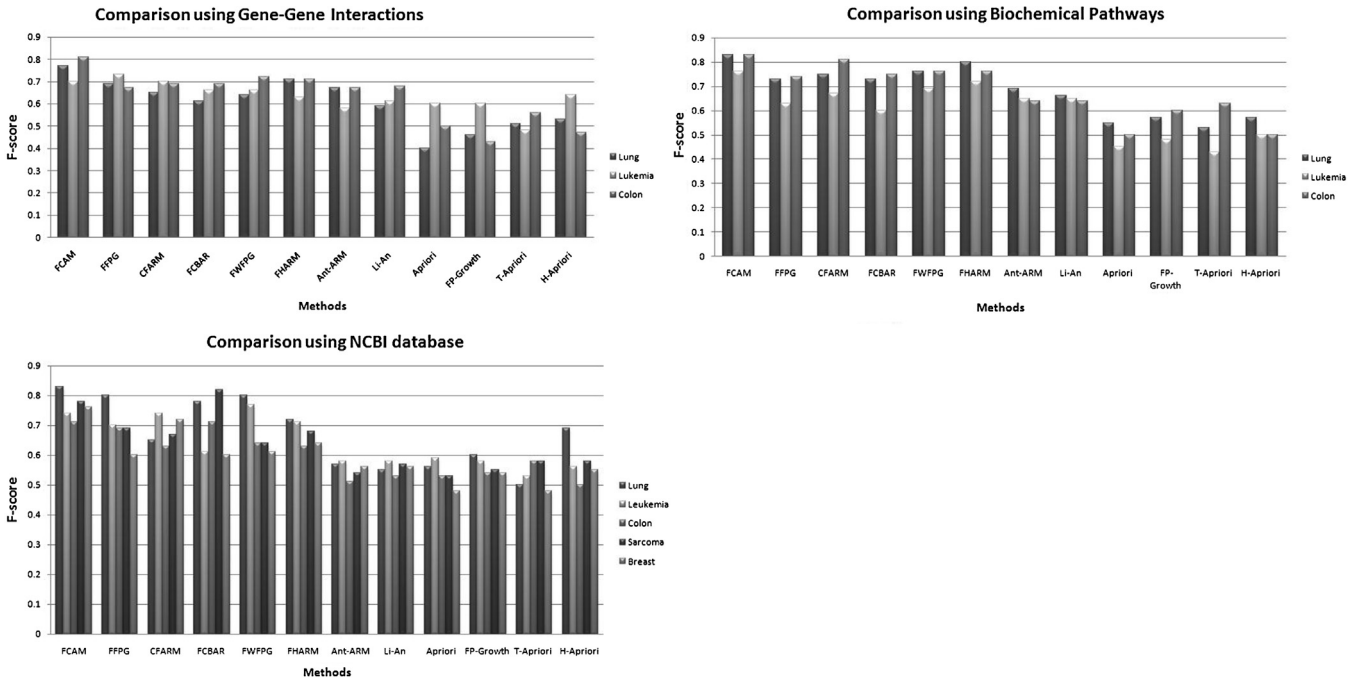

Fig. 2. Comparison of FCAM with all other methods in terms of gene–gene interactions, biological pathways, NCBI database using *F*-score.

pathways like chronic myeloid and acute myeloid leukemia. From Fig. 2, it is clear that FCAM have provided the best *F*-score result to find out true positives compared to the other existing methods for all the cases. It is to be mentioned that FCAM is capable of finding out less number of false positive and false negative genes compared to the other methods for all the cases.

Next, the results, obtained by the methods under Categories 2 and 3 in both fuzzy and crisp framework, have been compared, using *F*-score, in terms of biochemical pathways. From Fig. C.11 in Appendix C, it is clearly observed that all the methods under fuzzy framework, that Categories 2 and 3, perform better than the under crisp framework. Here FCAM results in higher *F*-score value than all the methods considered here for all the cancer datasets.

5.3.3. Using database

NCBI provides a gene database [50] where the disease mediating gene list corresponding to a specific disease can be obtained. The list is arranged in terms of relevance of the gene. We have got different sets of genes for lung cancer, colon cancer, sarcoma, breast cancer and leukemia. From each gene list, we can consider first c genes if a method results in c genes. For lung expression data (7129 genes), 55 altered associations have been identified using FCAM. A set of 90 genes are involved in these altered associations. We have compared this set of genes with first 90 genes from NCBI, which are supported by some earlier investigations. Here 75 genes have been identified, which are common in both the sets. We call these genes as *true positive* (*TP*) genes. Thus 15 (=90 − 75) genes are not in the list of 90 genes obtained from NCBI. We denote these genes as *false positives* (*FP*). Likewise, 15 (=90 − 75) genes which are in NCBI list, are not in the set of genes obtained by FCAM and are called *false negative* genes.

In this way, we have taken top ranked 110, 80, 80, 120, 75, 75, 80, 70, 100, 95 and 100 genes from NCBI corresponding to CFARM, FCBAR, FFPG, FWFPG, Apriori, FP-Growth, T-Apriori, Hybrid Apriori, Ant-ARM, Li-An and FHARM. Similarly, the results have been validated for other four expression datasets. From Fig. 2, it is clearly shown that FCAM produces the best *F*-score value compared to the other existing methods for all the five datasets. We have made a comparative study using *F*-score obtained by the methods under Categories 2 and 3 under both fuzzy and crisp framework, in terms of NCBI database. From Fig. C.12 in Appendix C, it is clearly observed that all the fuzzy association rule mining methods provide higher *F*-score values than the crisp rule mining methods. FCAM performs much better than all existing rule mining methods considered here for all the five cancer datasets. Hence, we can conclude that the proposed method FCAM is capable of identifying the significant associations among the genes in terms of gene–gene interactions, biochemical pathways and NCBI database.

5.3.4. Using functional enrichment

Since the genes have been ranked based on their occurrences in the altered associations, top ranked genes are expected to mediate the respective carcinoma, being involved in some particular biological functions [51]. Thus the functional categories, related to these biological functions, of these genes should be enriched.

In our study, the enrichment of each GO category for each of the genes has been calculated by its p-value. A low p-value indicates that the genes belonging to the enriched functional categories are biologically significant. Here only functional categories with p-value $\leq 5 \times 10^{-5}$, p-value $\leq 5 \times 10^{-7}$ and $\leq 5 \times 10^{-9}$ have been considered. Figs. C.3–C.5 in Appendix C show the number of functionally enriched attributes corresponding to FCAM for various sets of genes. Higher number of enriched attributes for a set of top ranked genes indicates that the resulting genes are belonging to the same functional categories. In other words, this group of genes are performing the same set of functions. This means, if one of the genes from the pool is responsible for cancer then the other genes may have a strong influence in mediating the disease.

In order to demonstrate the ability to identify cancer mediating genes correctly, the number of enriched attributes of the first 5, 10, 15, 20 genes has been computed for all the five datasets. From Figs. C.3–C.5 in Appendix C, it is clearly seen that gene ranking resulted in by FCAM is the best of all the other 11 methods considered in the present investigation. Similarly, we have calculated the number of enriched attributes for the last 5, 10, 15, 20 gene sets to establish the fact that how correctly the methods along with *importance factor* is capable of ranking the genes according to their

importance (Figs. C.6–C.8 in Appendix C). Figs. C.3–C.8 in Appendix C clearly depict that FCAM provides the best results for all the five datasets.

5.3.5. Some possible reasons behind superior performance of FCAM

Methods under Category 1 deal with structural data and other type of biological data except gene expression data, and have not been considered here for comparison. We have considered some methods under both Categories 2 and 3 for comparison.

Since the proposed method FCAM has been developed for gene expression data and based on the notion of fuzzy set theory, FCAM has provided better results than the methods under Category 3. Although the methods under Category 2 uses gene expression data, they ultimately determine associations among the gene functions based on other databases viz., GO, SAGE. The methods under Category 3 deal with non-biological data in different domains.

The methods Apriori, T-Apriori, H-Apriori and FP-Growth partition a continuous domain into crisp intervals while splitting a domain of an attribute into intervals. Here the problem of sharp boundary problem arises. Elements near the boundaries of a crisp set (interval) may be either ignored or overemphasized. In case of T-Apriori, the use of clustering method, in partitioning the domain of an attribute, maps the quantitative association rule mining problem into the Boolean association rule mining problem, and needs careful consideration, as the partitions may affect the generated rules directly. The algorithms, FP-Growth, FFPG and FWFPG do not produce the candidate item-sets, but it uses growth models to mine frequent patterns. Moreover, FP-growth needs transaction database twice, due to which mining process needs extra processing time and space as it continuously generates large conditional bases and conditional FP-Trees. These are some possible reasons behind the better performance of FCAM over the others.

Out of the methods under both Categories 2 and 3, those based on the notion of fuzzy sets have shown better performance in most of the cases than the non-fuzzy methods. This is shown in Figs. C.9–C.12 in Appendix C. The reason behind it may be as follows.

Biological data are often imprecise and noisy. Moreover, in microarray gene expression data, genes have expression values that are in different intervals under two conditions (i.e., normal or diseased). Although each interval has a well-defined boundary, they are highly overlapped. Fuzzy set theory is especially suitable to model such imprecise and overlapping data. Thus incorporation of the notion of fuzzy sets in the methods enables one to handle such overlapping intervals in a better way.

Finally, we have executed the fuzzy set theoretic algorithms developed based on (i) triangular membership function that has been used in FCAM, and (ii) the membership function used originally by the respective methods. Fig. C.13 in Appendix C shows the results. However, both Cases (i) and (ii) have resulted in similar performance. Case (i) has performed better for some datasets/databases than Case (ii), while it is the reverse for some others.

5.4. Some possible cancer mediating genes

Here some genes are considered, which are among the most significant top genes. Applying FCAM on human lung expression data, we have found the genes, like IGFBP3, TP53, HBB, HLA-B, SFTPA2, TNF, IGHG3, PRKACA and SORT1 are among the top 10 most important genes. Likewise, for human colon expression profile, the genes like, microtubule-associated protein 2 (MAP2), tumor necrosis factor (TNF), calcitonin (CALCA), colon carcinoma kinase-4 (CCK4), isoleucyl-tRNA synthetase (IARS), thymidylate synthase (TYMS), Hemoglobin Beta Chain (HBB), insulin-like growth factor binding protein 6 (IGFBP6) are among the top 10 most important genes. For human breast cancer dataset, we have found that genes like KRAS, HAL, FNTA, BCAN, GDI2, ERBB2, NARS are the most important genes. Similarly, genes, like PTEN, IGF1, BRCA1, TYMS, IARS, HBB are also among the top 10 most important genes for human soft tissue sarcoma expression data. Finally, we report that genes, like BAX, CALCA, ATP6V0B, NARS, CDKN2A, GDI2, SDHC and H3F3A are among the top 10 most significant genes for human lymphocytes and plasma cell expression data.

6. Conclusions

In this article, we have developed a methodology, called fuzzy correlated association mining (FCAM) that has identified associations among the genes and found out the disease mediating genes from gene expression data. The algorithm has identified the associations among the genes, which have altered quite significantly from normal state to diseased state. Moreover, a gene ranking measurement has been formulated to identify the importance of the genes that are present in the set of altered associations.

The effectiveness of FCAM, along with superior performance over 11 existing methods has been demonstrated on five gene expression datasets (lung, colon, sarcoma, leukemia and breast). The comparative study has been made in terms of gene–gene interactions, biochemical pathways, using NCBI database and functionally enriched attributes. We have found genes, like IGFBP3, ERBB2, TP53, HBB, KRAS, PTEN, CALCA, CDKN2A as some possible important genes mediating certain cancers considered in this article. It has been found that FCAM has been able to find more true positives than the existing methods considered here. We have also validated the altered associations among the genes using various pathways available in NCBI database. As a consequence, we can say that these sets of identified genes along their altered associations, have a significant role of mediating the various types of cancers.

Appendix A. Some Lemmas and their proofs

A.1. Proof of Lemma 1

Lemma 1. *The number of elements in T_r in rth iteration is* $|T_r| = \sum_{i=r}^{2^{r-1}} \binom{n}{i}$, $r \geq 1$, $n \geq 2^{r-1}$.

Proof. We now prove Lemma 1 by the method of induction. In general, if there is a set $D = \{g_1, g_2, \ldots, g_{n-1}, g_n\}$ consisting of n genes then $T_1 = \{(g_1), (g_2), \ldots, (g_{n-1}), (g_n)\}$ and $|T_1| = \binom{n}{1} = \sum_{i=1}^{2^{1-1}} \binom{n}{i}$. Similarly, the set T_2 consists of 2-terms. The number of 2-terms is $\binom{n}{2}$. Thus, $|T_2| = \binom{n}{2} = \sum_{i=2}^{2^{2-1}} \binom{n}{i}$. The set T_3 consists of both 3-terms and 4-terms. The number of 3-terms is $\binom{n}{3}$ and that of 4-terms is $\binom{n}{4}$. Thus, $|T_3| = \binom{n}{3} + \binom{n}{4} = \sum_{i=3}^{2^{3-1}} \binom{n}{i}$. Similarly, the set T_4 consists of 4-terms, 5-terms, 6-terms, 7-terms, 8-terms. Thus, $|T_4| = \binom{n}{4} + \binom{n}{5} + \binom{n}{6} + \binom{n}{7} + \binom{n}{8} = \sum_{i=4}^{2^{4-1}} \binom{n}{i}$. Therefore, the Lemma 1 is valid for $r = 1$, 2, 3 and 4. Let us assume that the lemma is valid for $r = r$. That is, $|T_r| = \sum_{i=r}^{2^{r-1}} \binom{n}{i}$. Now the set T_{r+1} consists of $(r+1)$-terms, $(r+2)$-terms, …, $2 \times 2^{r-1}$ i.e.,

2^r terms. Therefore, $|T_{r+1}| = \binom{n}{r+1} + \binom{n}{r+2} + \cdots + \binom{n}{2^r} = \sum_{i=r+1}^{2^{r+1}-1} \binom{n}{i}$. Thus, if Lemma 1 is valid for $r=1$ then it is valid for $r=2$. If Lemma 1 is true for $r=2$, it holds for $r=3$, and so on. Thus, $|T_r| = \sum_{i=r}^{2^{r-1}} \binom{n}{i}$, $r \geq 1$, $n \geq 2^{r-1}$.

A.2. Proof of Lemma 2

Lemma 2. *The maximum number of iterations of FCAM is* $\lceil 1 + \log_2 n \rceil$.

Proof. From Lemma 1, we can get $n \geq 2^{(r-1)}$, where r is the number of iterations. Thus, for the maximum value of $r = R$, $2^{R-1} = n$. That is, $R = 1 + \log_2 n$. If n is not a power of 2 then $R = \lceil 1 + \log_2 n \rceil$.

A.3. Proof of Lemma 3

Lemma 3. *The number of associations in candidate set* C_r *in rth iteration is* $|C_r| = \sum_{i=r}^{2^{r-1}} (2^{i-1} - 1) \binom{n}{i}$, $r \geq 1$, $n \geq 2^{r-1}$.

Proof. The number of associations generated from each r-term is $\left(\binom{r}{(r-1)} + \binom{r}{(r-2)} + \ldots + \binom{r}{1} \right) / 2$, i.e., $(2^{r-1} - 1)$ where $r \geq 1$. Thus, the number of associations generated from all r-terms becomes $(2^{r-1} - 1) \binom{n}{r}$. Similarly, number of associations generated from all $(r+1)$-terms is $(2^r - 1) \binom{n}{(r+1)}$. In this way, we can get $(2^{(2^{r-1}-1)} - 1) \binom{n}{2^{r-1}}$ associations generated by all 2^{r-1}-terms. All these associations are included in the set C_r. Therefore, total number of associations in C_r, $r \geq 1$, in rth iteration is $|C_r| = (2^{r-1} - 1) \binom{n}{r} + (2^r - 1) \binom{n}{(r+1)} \cdots + (2^{(2^{r-1}-1)} - 1) \binom{n}{2^{r-1}} = \sum_{i=r}^{2^{r-1}} (2^{i-1} - 1) \binom{n}{i}$ It is to be noted that L_r is generated from C_r in rth iteration based on fuzzy support and fuzzy correlation coefficient using corresponding threshold values.

Appendix B. *F*-score

F-score ($\in [0, 1]$) [52] is defined as

$$\text{F-score} = \frac{(\beta^2 + 1) * Precision * Sensitivity}{\beta^2 * Precision + Sensitivity} \tag{10}$$

where

$$Sensitivity = \frac{tp}{(tp + fn)} \tag{11}$$

$$Precision = \frac{tp}{(tp + fp)} \tag{12}$$

The terms tp, fp, fn and β stand for true positive, false positive, false negative and balance factor, respectively. All the above three measures distinguish the correct classification of labels within different classes. They concentrate on one class (positive examples). Sensitivity or Recall is a function of its correctly classified examples (true positives) and its misclassified examples (false negatives). Precision is a function of true positives and examples misclassified as positives (false positives). The *F*-score [52] is evenly balanced when $\beta = 1$. It favors precision when $\beta > 1$, and sensitivity or recall otherwise. In our experiment, we have chosen $\beta = 1$.

Appendix C. Some tables and figures

See Appendix Figs. C.3–C.13 and Tables C.5–C.7 .

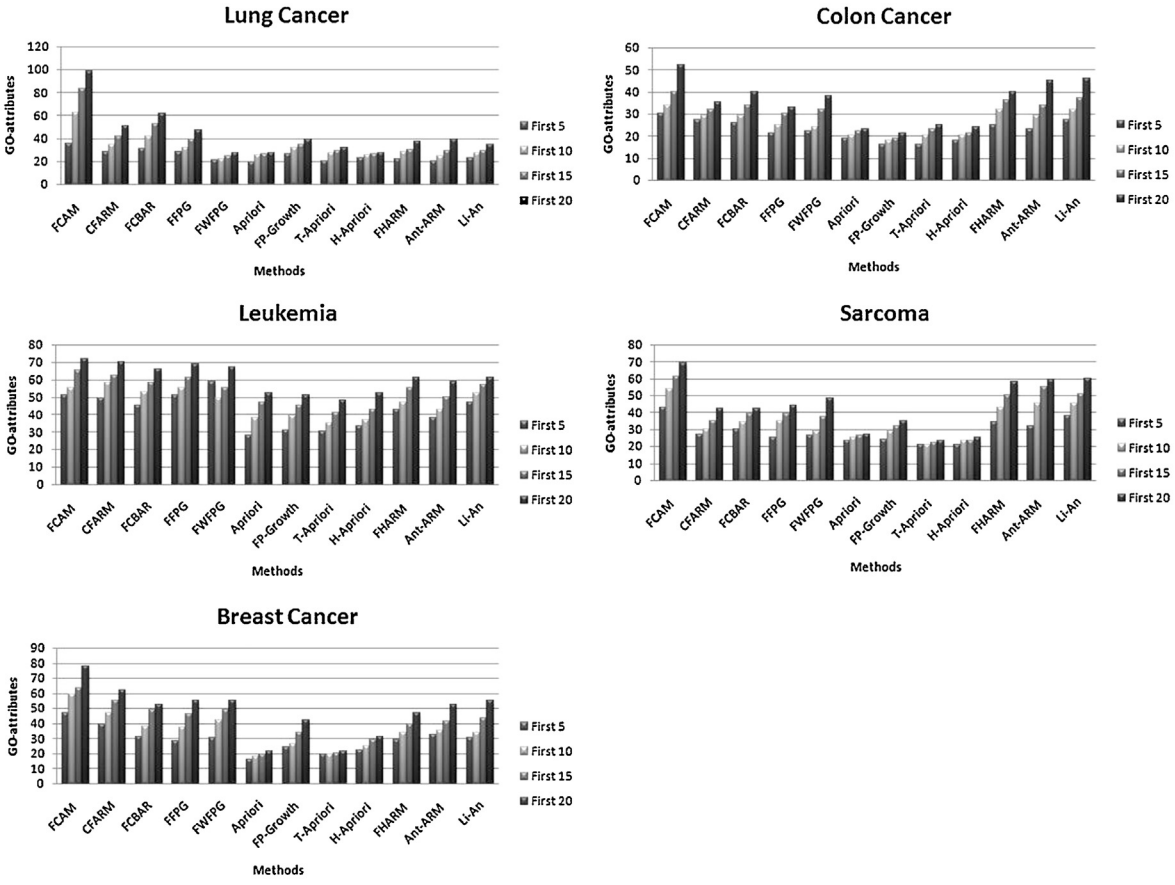

Figure C.3. Comparative results on number of enriched attributes of first 5 to first 20 gene sets with p-val $\leq 5 \times 10^{-5}$.

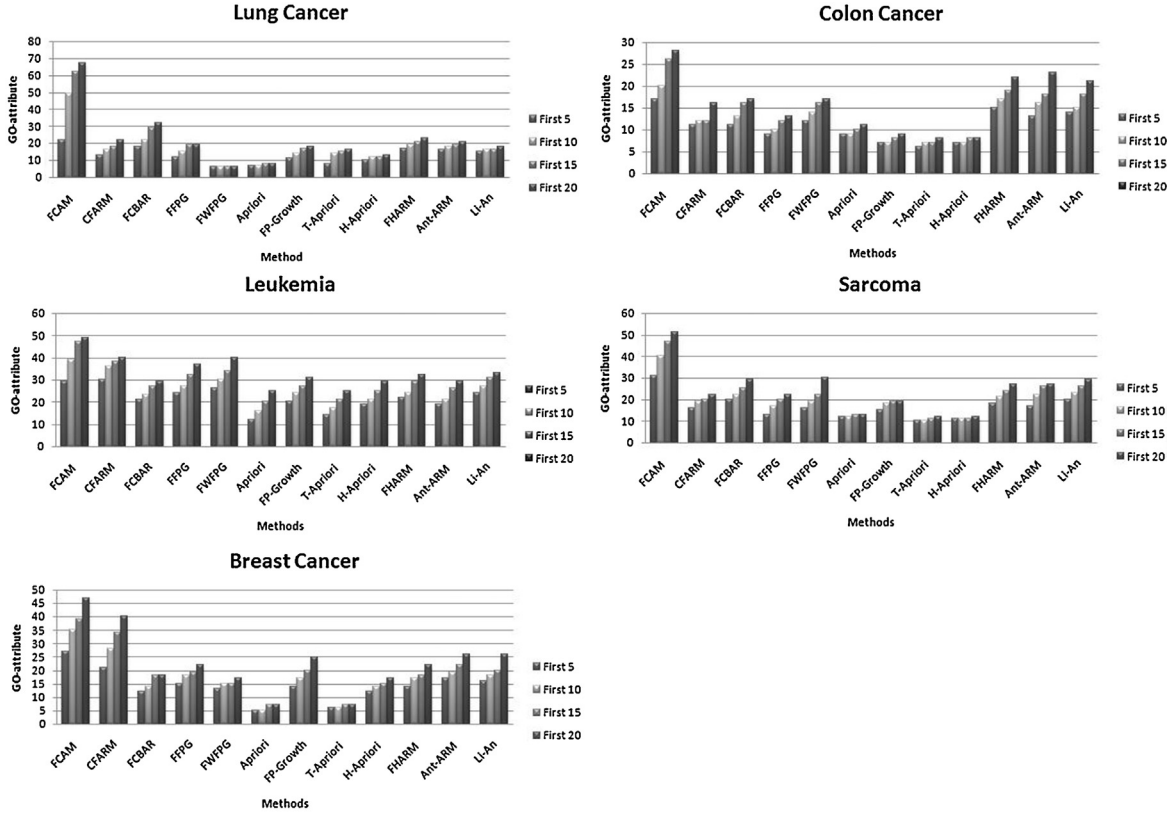

Figure C.4. Comparative results on number of enriched attributes of first 5 to first 20 gene sets with p-val $\leq 5 \times 10^{-7}$.

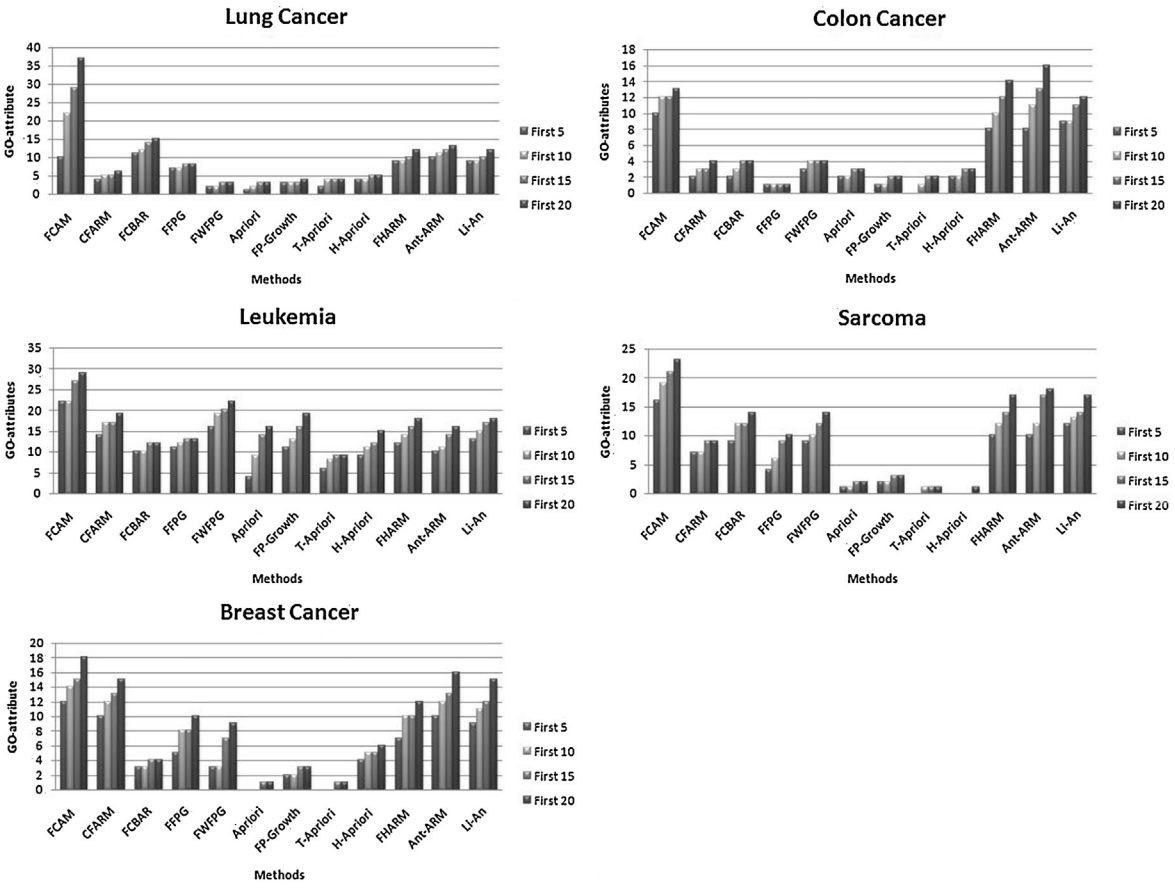

Figure C.5. Comparative results on number of enriched attributes of first 5 to first 20 gene sets with p-val $\leq 5 \times 10^{-9}$.

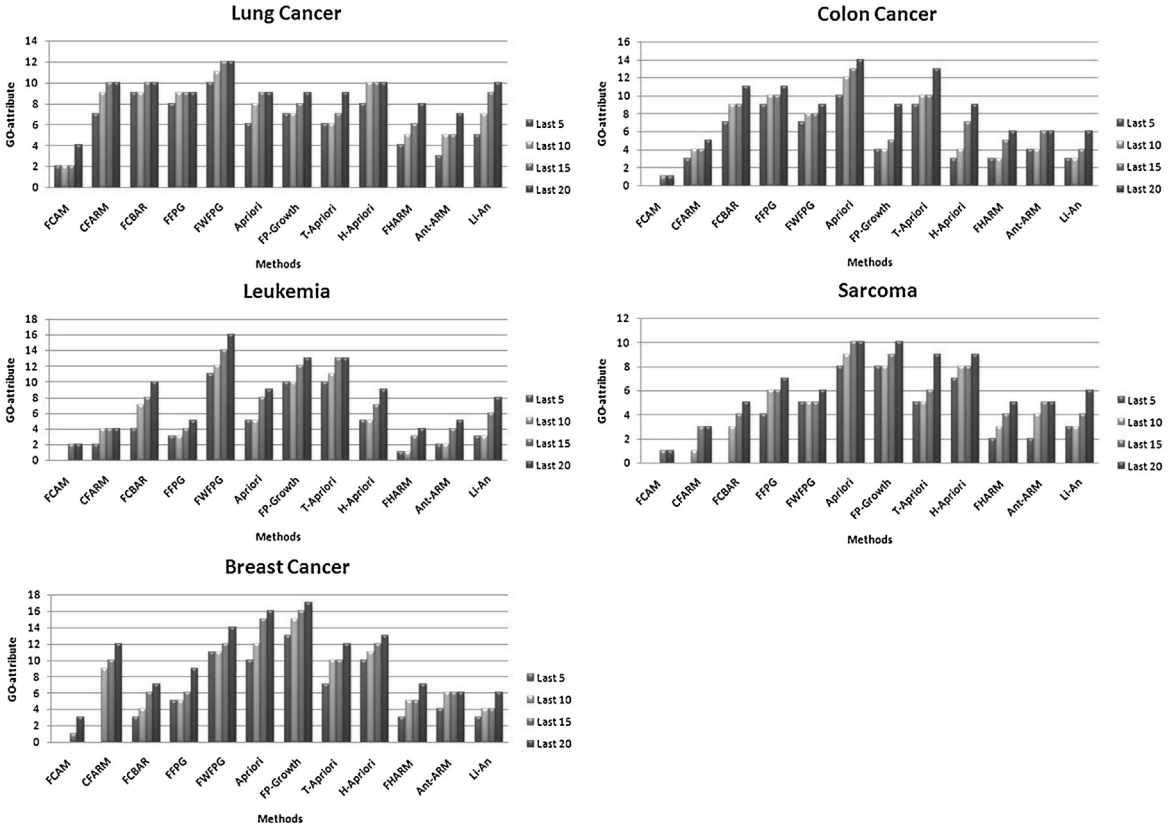

Figure C.6. Comparative results on number of enriched attributes of last 5 to last 20 gene sets with p-val $\leq 5 \times 10^{-5}$.

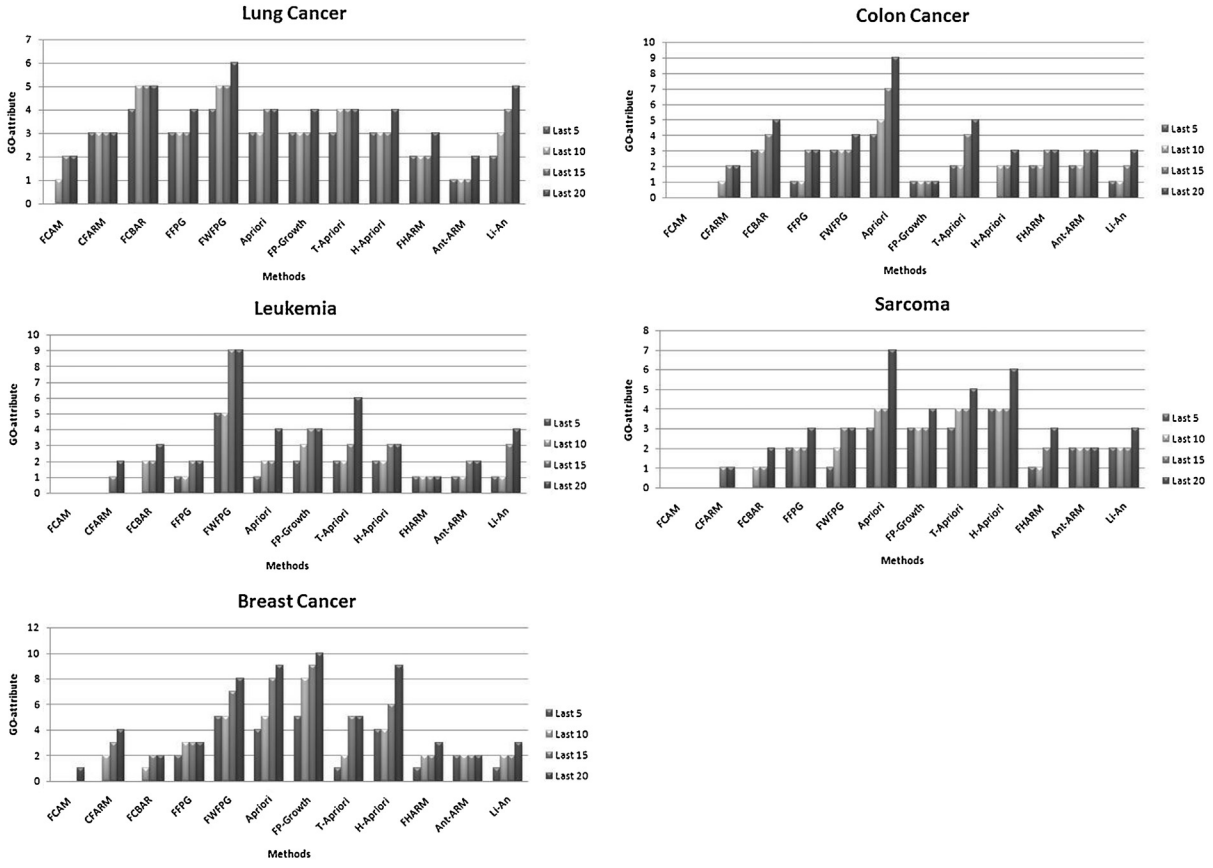

Figure C.7. Comparative results on number of enriched attributes of last 5 to last 20 gene sets with p-val $\leq 5 \times 10^{-7}$.

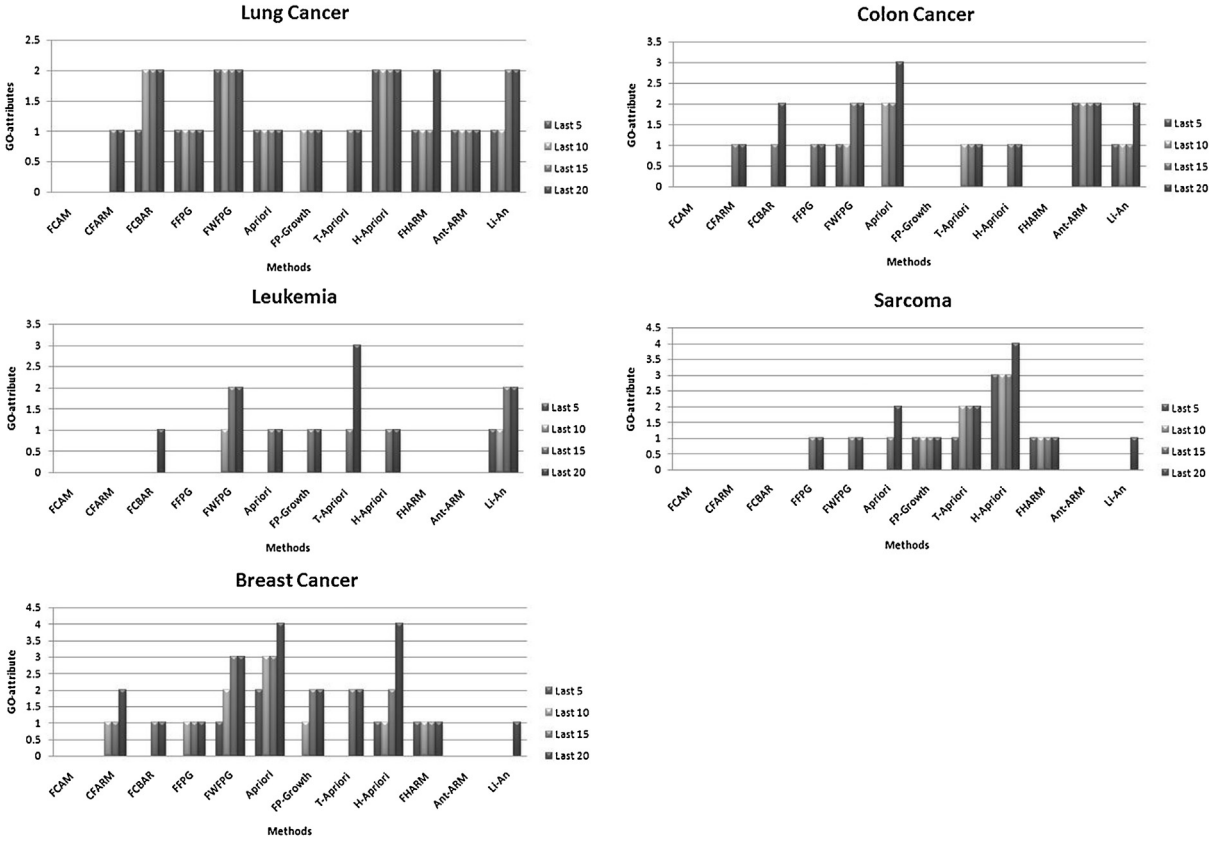

Figure C.8. Comparative results on number of enriched attributes of last 5 to last 20 gene sets with p-val $\leq 5 \times 10^{-9}$.

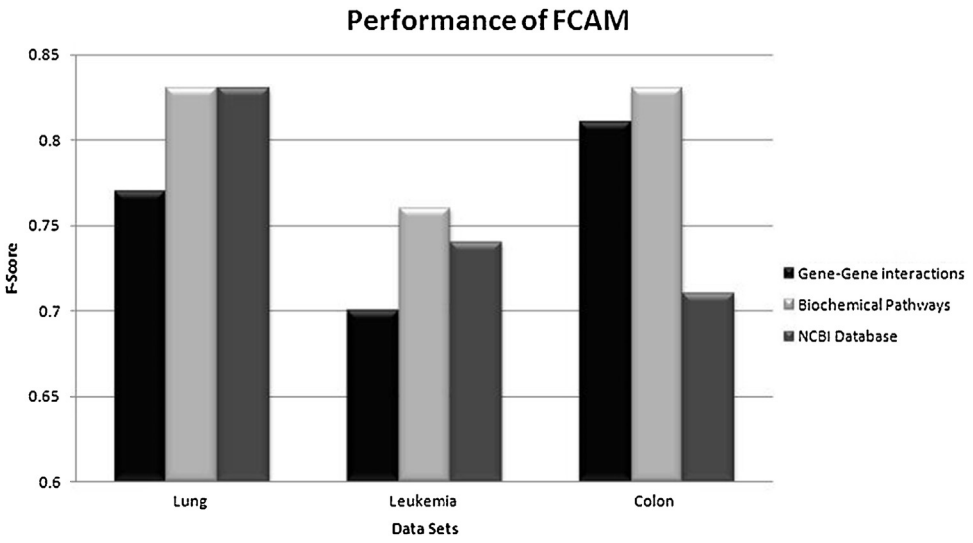

Figure C.9. *F*-scores obtained by FCAM.

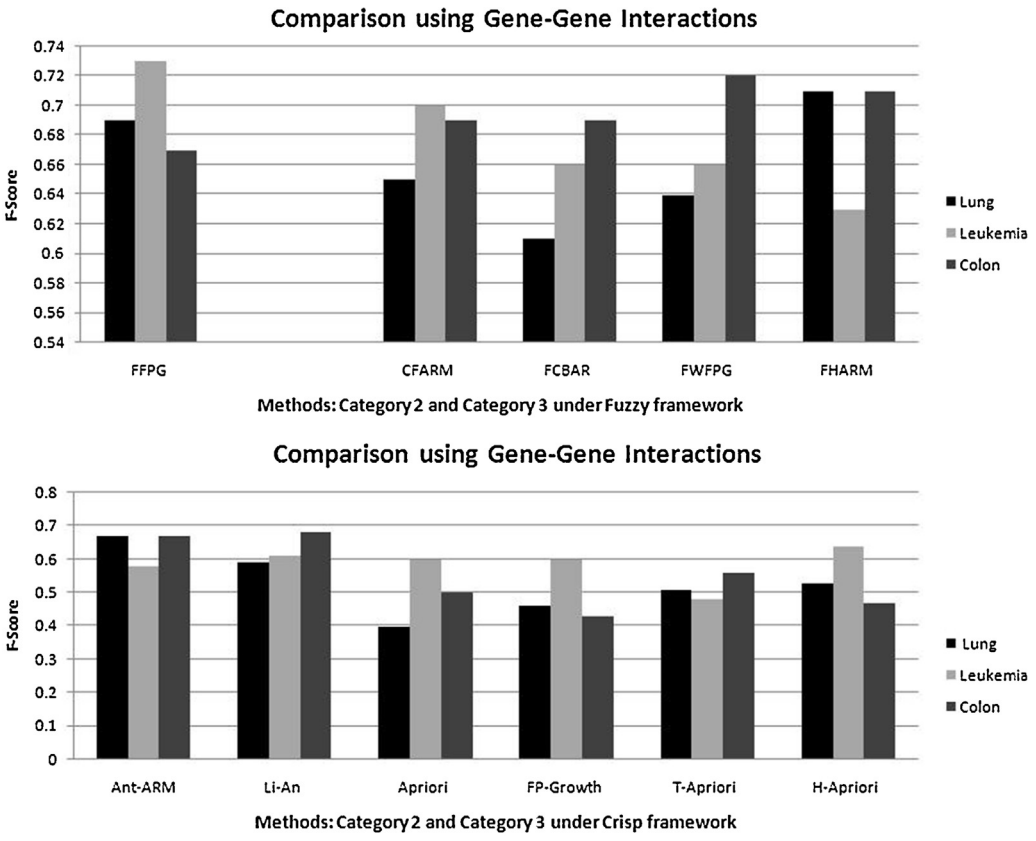

Figure C.10. Comparison using *F*-scores on gene–gene interactions for lung, leukemia and colon cancer datasets.

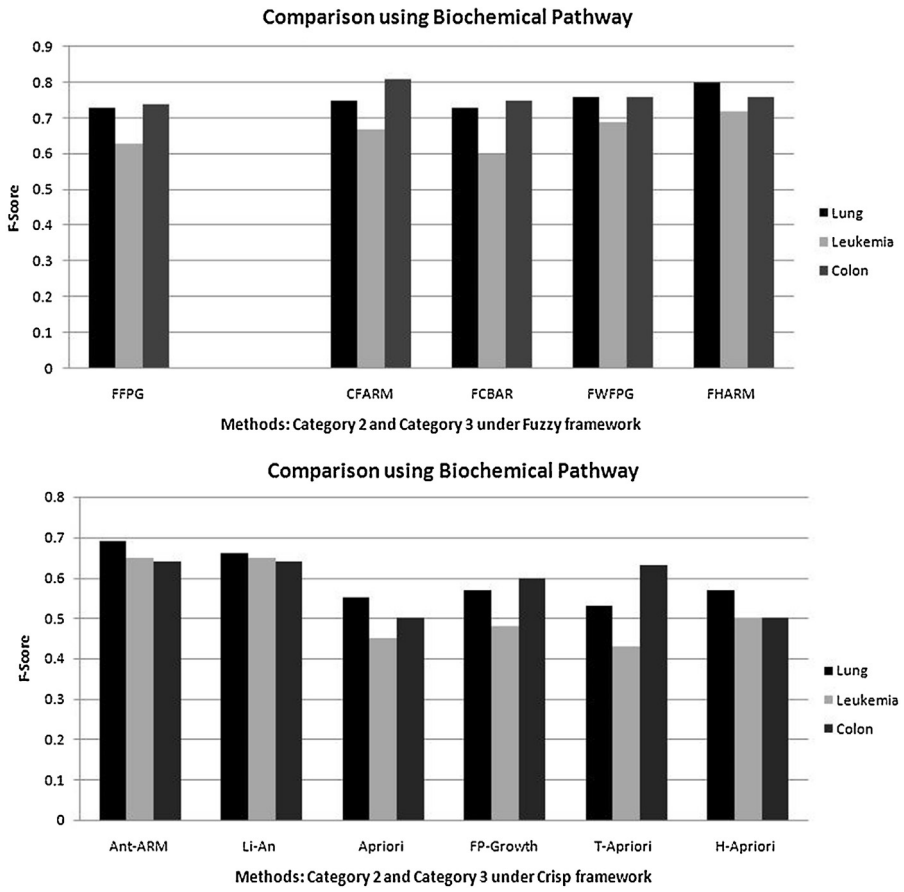

Figure C.11. Comparison using *F*-scores on biochemical pathways for lung, leukemia and colon cancer datasets.

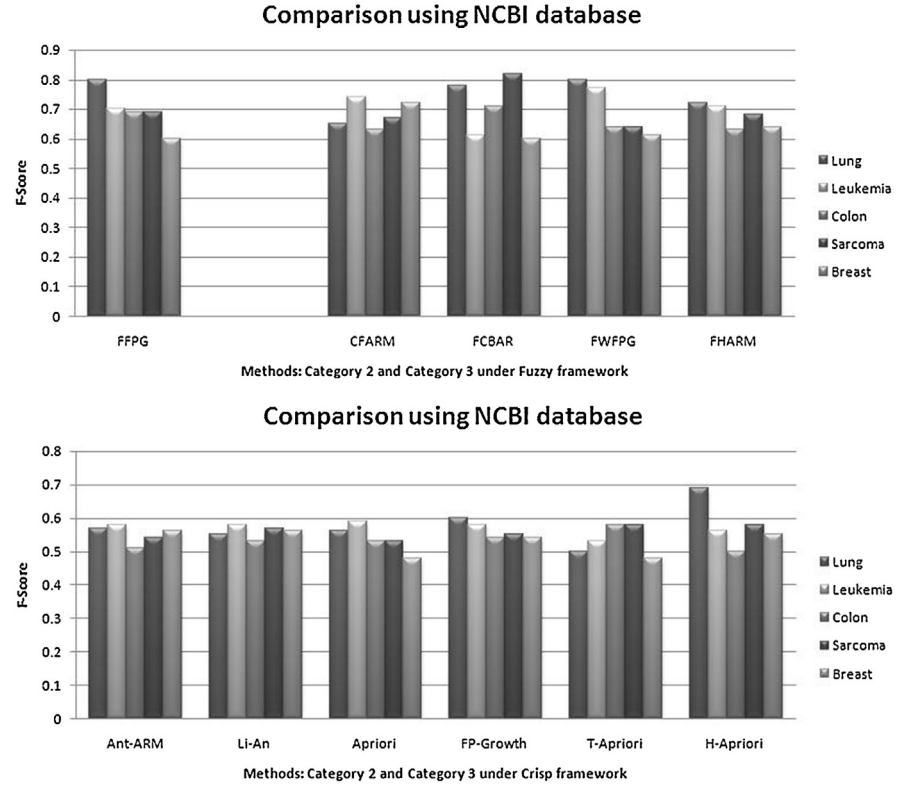

Figure C.12. Comparison using *F*-scores on NCBI database for lung, leukemia, colon, sarcoma and breast cancer datasets.

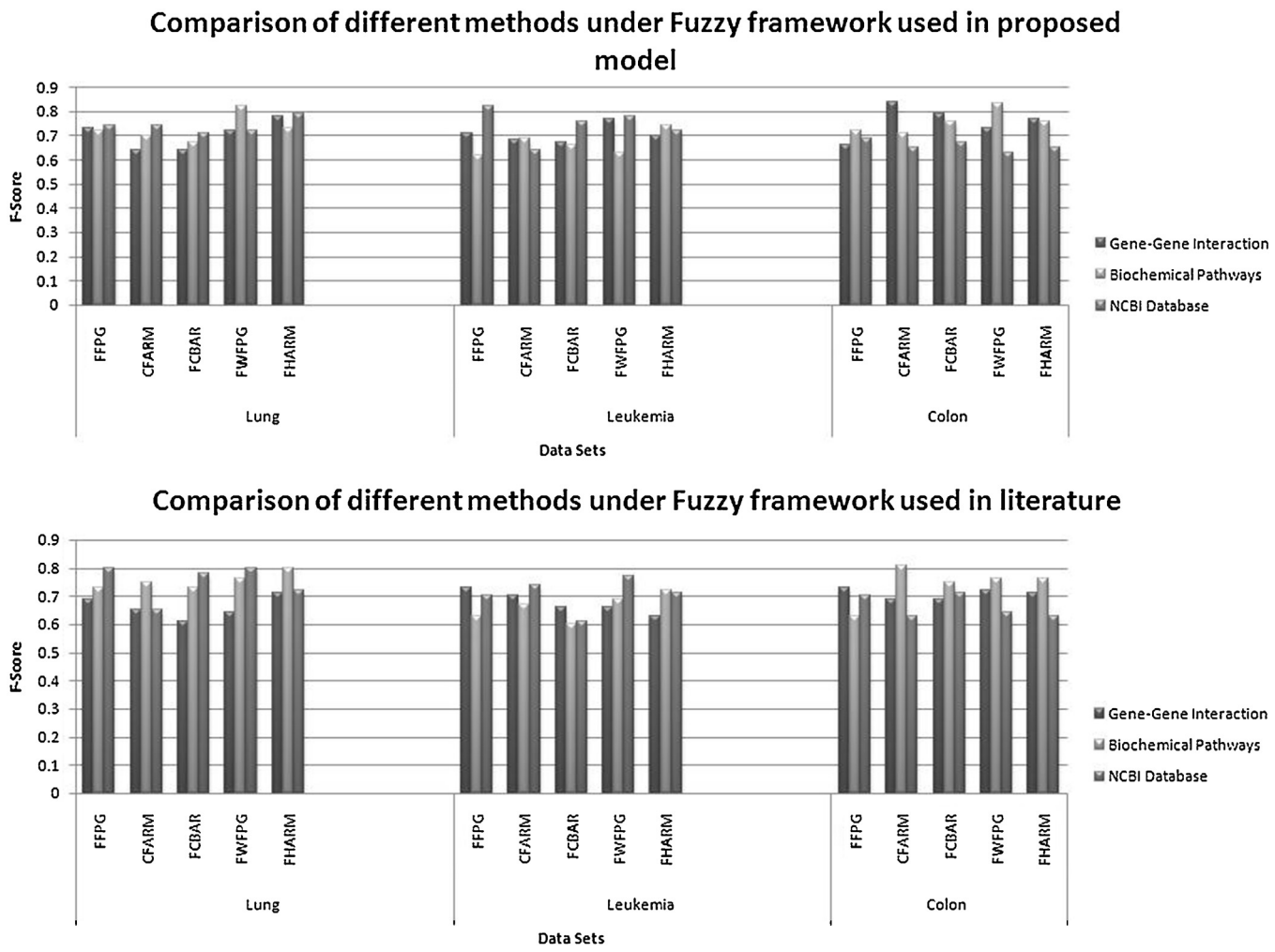

Figure C.13. Comparison between two different fuzzy frameworks.

Table C.5
List of associations that have altered from normal to tumor samples for lung expression dataset using FCAM.

Rule No.	List of rules	Fuzzy confidence value
1.	⟨(TNF,TP53),(EGFR,KRAS,BRAF,CASP8)⟩	0.82
2.	⟨(PTEN,TGFB1,TNF),(VEGFA,AKT1,CDKN2A,CDH1)⟩	0.85
3.	⟨(BCL2,IL6,HLA-B,PIK3CA),(TP53,PTEN,EGFR)⟩	0.81
4.	⟨(TNF,EGFR,STAT3),(ERBB2,HIF1A,ESR1,BRCA1)⟩	0.87
5.	⟨(TNF,EGFR,TP53),(BAX,LEP,IGF1,IGFBP3)⟩	0.89
6.	⟨(TNFRSF1A,BRAF,TP53),(PTEN,TNF)⟩	0.82
7.	⟨(EGF,IGF1R,CYP2D6,FLT1),(STAT3,BAX)⟩	0.91
8.	⟨(CASP3,KRT19,ITGB3),(IGFBP3)⟩	0.87
9.	⟨(AGER,PIK3CG,SOX4,TNF),(TP53,EGFR)⟩	0.92
10.	⟨(FAS,MYC,ACE,BRCA1),(FOXP3,BIRC2)⟩	0.90
11.	⟨(TNF),(IGFBP3,HLA-B)⟩	0.88
12.	⟨(BAX,TNF,BRAF),(TP53,EGFR)⟩	0.91
13.	⟨(PRKACA,TP53),(HBB,SFTPA1,SFTPA2)⟩	0.86
14.	⟨(EGFR),(MMP3,MMP2,MMP1,MMP9)⟩	0.88
15.	⟨(HLA-B,KRT19),(VEGFA,HLA-G)⟩	0.84
16.	⟨(TP53,BRCA1),(TNF)⟩	0.83
17.	⟨(SFTPA1,SFTPA2),(HBB)⟩	0.84
18.	⟨(HBB),(ERBB2,ERCC2)⟩	0.85
19.	⟨(APOE,IL10,IL8),(IL6)⟩	0.90
20.	⟨(IGFBP3,CDKN2A),(CDKN1B,CDKN1A)⟩	0.92
21.	⟨(MMP7),(MMP9)⟩	0.80
22.	⟨(MMP3,MMP2),(MMP7)⟩	0.86
23.	⟨(TP53),(APOA1,LPA,PTGS2)⟩	0.86
24.	⟨(TNF),(MAPK1,MAPK14,MAPK3)⟩	0.85
25.	⟨(VEGFA),(VHL,VEGFC)⟩	0.84

Table C.5 (*Continued*)

Rule No.	List of rules	Fuzzy confidence value
26.	⟨(PTEN,BRCA1),(BIRC5,CCND1,ESR2,FLT4)⟩	0.89
27.	⟨(TP53),(MAPK8)⟩	0.90
28.	⟨(MMP1),(MMP8)⟩	0.91
29.	⟨(BRCA1,TNF),(DAPK1,BCL2L1,BCL2)⟩	0.86
30.	⟨(PTEN),(PRKCD,FHIT,LTF)⟩	0.87
31.	⟨(TNF,TP53),(FHIT,BCL2L1,FLT4)⟩	0.84
32.	⟨(TNF),(PTEN,STAT3,CDKN2A)⟩	0.83
33.	⟨(TNF),(EGF,BRAF,CASP3)⟩	0.87
34.	⟨(TNF,TP53,IGFBP3),(BAX,KRAS,CDH1,KRT19)⟩	0.90
35.	⟨(TP53),(DAPK1,TRAF4,RHOA)⟩	0.83
36.	⟨(EGFR,CDKN2A),(BRAF)⟩	0.82
37.	⟨(MYC,HES1),(RIMS2,BRAF,PTGS2)⟩	0.80
38.	⟨(TP53,MYC),(BRAF)⟩	0.82
39.	⟨(JUN,RELA,HIF1A),(LPA,TNF)⟩	0.80
40.	⟨(FHIT,TGFB1,TLR4),(TNF)⟩	0.84
41.	⟨(CYP2A6,TP53,TNF),(TP73,IGFBP3)⟩	0.81
42.	⟨(BCL2,EGFR),(MAPK14,MAPK1)⟩	0.84
43.	⟨(TP73,EGR1),(HRAS)⟩	0.80
44.	⟨(TP53,SOX9),(CREB1,CEBPB)⟩	0.82
45.	⟨(TNF,IL6,IL10),(LEP,STAT3)⟩	0.81
46.	⟨(E2F1,TNF,TP53),(EGR1)⟩	0.84
47.	⟨(TNF,E2F1,CREB1,STAT1),(EGF,EGR1)⟩	0.80
48.	⟨(RELA,STAT1),(IRF1,IRF2,IRF8)⟩	0.83
49.	⟨(PTEN,E2F1,MYC),(PIK3CG)⟩	0.81
50.	⟨(CDKN2A,IL6,E2F1),(CDH1)⟩	0.86
51.	⟨(FGF2,MYC,HIF1A),(VEGFA)⟩	0.84
52.	⟨(TNF,TP53,EGF),(PTEN,TGFB1)⟩	0.80
53.	⟨(BCL2,PIK3CG),(HRAS,KRAS,HGF)⟩	0.83
54.	⟨(EGFR),(PRKACA,KRT8)⟩	0.81
55.	⟨(IGF1R,TP53,VEGFA,LEP),(IGF1,IGFBP3,IGF)⟩	0.86

Table C.6
Altered associations generated by different methods.

Dataset	Method	$\lvert A_N\rvert$	$\lvert A_D\rvert$	$\lvert A_N \cup A_D\rvert$	$\lvert A_N \cap A_D\rvert$	$\lvert A\rvert = \lvert(A_N \cup A_D) - (A_N \cap A_D)\rvert$
Lung	FCAM	112	97	132	77	55
	CFARM	122	86	125	83	42
	FCBAR	98	74	111	61	50
	FFPG	83	91	101	73	28
	FWFPG	105	98	146	57	89
	Apriori	90	73	102	61	41
	FP-Growth	102	79	112	69	43
	T-Apriori	95	81	111	65	46
	H-Apriori	115	104	131	88	43
	FHARM	102	89	108	70	38
	Ant-ARM	110	92	115	85	30
	Li-An-Method	97	92	121	83	38
Colon	FCAM	76	58	92	42	50
	CFARM	83	64	92	55	37
	FCBAR	72	60	86	46	40
	FFPG	63	41	70	34	36
	FWFPG	87	58	91	54	37
	Apriori	72	53	82	43	39
	FP-Growth	67	49	80	36	44
	T-Apriori	61	45	69	37	32
	H-Apriori	77	59	85	51	34
	FHARM	102	89	108	70	38
	Ant-ARM	110	92	115	85	30
	Li-An-Method	97	92	121	83	38
Sarcoma	FCAM	170	128	205	93	112
	CFARM	172	109	198	83	115
	FCBAR	156	98	180	76	104
	FFPG	145	121	180	84	96
	FWFPG	112	102	149	65	84
	Apriori	136	109	176	69	107
	FP-Growth	157	92	166	83	83
	T-Apriori	169	115	198	86	112
	H-Apriori	121	77	148	50	98
	FHARM	92	85	118	77	41
	Ant-ARM	101	90	119	85	34
	Li-An-Method	112	98	125	83	42

Table C.6 (*Continued*)

Dataset	Method	$\|A_N\|$	$\|A_D\|$	$\|A_N \cup A_D\|$	$\|A_N \cap A_D\|$	$\|A\| = \|(A_N \cup A_D) - (A_N \cap A_D)\|$
Lymphocytes	FCAM	164	112	186	90	96
	CFARM	152	98	161	89	72
	FCBAR	126	90	143	73	70
	FFPG	145	121	180	84	96
	FWFPG	134	109	163	80	83
	Apriori	116	87	133	70	63
	FP-Growth	140	78	156	62	94
	T-Apriori	107	89	139	57	82
	H-Apriori	146	117	184	79	105
	FHARM	82	98	110	77	33
	Ant-ARM	115	98	131	96	35
	Li-An-Method	99	90	102	69	33
Breast	FCAM	164	112	192	84	108
	CFARM	115	88	138	65	73
	FCBAR	123	80	143	73	70
	FFPG	145	121	180	84	96
	FWFPG	134	109	163	80	83
	Apriori	116	87	135	68	67
	FP-Growth	140	78	154	64	90
	T-Apriori	107	89	139	57	82
	H-Apriori	146	117	182	81	101
	FHARM	105	98	118	82	36
	Ant-ARM	107	90	112	85	27
	Li-An-Method	109	116	127	92	35

Table C.7
Comparative results on the number of genes (proteins) involved in the pathways and those obtained by the respective method. The results of breast and sarcoma have not been included as the pathway information of these data sets are not available. Here *TP*, *FP* and *FN* indicate true positive, false positive and false negative, respectively. The column "Parameters" provide different parameter values used for various methods.

Datasets	Methods	TP	FP	FN	Parameters
Lung	FCAM	342	73	67	$S_F = 0.20$, $C_F = 0.55$, $T_{corr} = 0.15$
	CFARM	312	115	97	$Cert - factor = 0.40$ $minsup = 0.30$
	FCBAR	302	120	107	$Minsup = 0.15$
	FFPG	290	95	119	$Minsup = 0.10$
	FWFPG	315	103	94	$MinSup = 0.30$
	FHARM	333	92	76	$Minsup = 0.25$
	Apriori	224	186	185	$Minsup = 0.10$ $Minconf = 0.50$
	FP-Growth	237	188	172	$Minsup = 0.22$
	T-Apriori	220	205	189	$Minsup = 0.40$
	H-Apriori	243	197	166	$Minsup = 0.025$
	Ant-ARM	290	140	119	$Minsup = 0.15$
	Li-An-Method	275	155	134	$Minsup = 0.45$
Leukemia	FCAM	278	97	77	$S_F = 0.20$, $C_F = 0.60$, $T_{corr} = 0.15$
	CFARM	243	123	112	$Cert - factor = 0.45$ $minsup = 0.33$
	FCBAR	205	120	150	$Minsup = 0.20$
	FFPG	233	146	122	$Minsup = 0.17$
	FWFPG	253	123	102	$MinSup = 0.26$
	FHARM	333	92	76	$Minsup = 0.30$
	Apriori	167	213	188	$Minsup = 0.10$ $Minconf = 0.50$
	FP-Growth	170	190	185	$Minsup = 0.20$
	T-Apriori	154	206	201	$Minsup = 0.50$
	H-Apriori	182	188	173	$Minsup = 0.05$
	Ant-ARM	235	135	120	$Minsup = 0.27$
	Li-An-Method	235	135	120	$Minsup = 0.38$
Colon	FCAM	55	15	7	$S_F = 0.30$, $C_F = 0.75$, $T_{corr} = 0.20$
	CFARM	52	15	10	$Cert - factor = 0.50$ $minsup = 0.48$
	FCBAR	48	18	14	$Minsup = 0.29$
	FFPG	45	14	17	$Minsup = 0.28$
	FWFPG	50	20	12	$MinSup = 0.47$
	FHARM	50	20	12	$Minsup = 0.20$
	Apriori	34	41	28	$Minsup = 0.20$ $Minconf = 0.65$
	FP-Growth	38	27	24	$Minsup = 0.35$
	T-Apriori	40	25	22	$Minsup = 0.65$
	H-Apriori	33	37	29	$Minsup = 0.10$
	Ant-ARM	42	28	20	$Minsup = 0.29$
	Li-An-Method	42	28	20	$Minsup = 0.29$

References

[1] V. Velculescu, L. Zhang, B. Vogelstein, K. Kinzler, Serial analysis of gene expression, Science 270 (1995) 484–487.
[2] L.A. Zadeh, The concept of linguistic variable and its applications to approximate reasoning-II, Inf. Sci. 8 (1975) 301–307.
[3] L.A. Zadeh, Precisiated natural language – toward a radical enlargement of the role of natural languages in information processing, decision and control, in: 9th International Conference on Neural Information Processing, 2002, pp. 1–3.

[4] H. Zhang, C.Y. Yu, B. Singer, M. Xiong, Recursive partitioning for tumor classification with gene expression microarray data, Proc. Natl. Acad. Sci. 8 (2001) 6730–6735.
[5] P.J. Woolf, Y. Wang, A fuzzy logic approach to analyzing gene expression data, Physiol. Genomics 3 (2000) 9–15.
[6] L. Machado, S. Vinterbo, G. Weber, Classification of gene expression data using fuzzy logic, J. Intell. Fuzzy Syst. 12 (2002) 19–24.
[7] R. Ram, M. Chetty, T.I. Dix, Fuzzy model for gene regulatory network, IEEE Congr. Evolut. Comput. 56 (2006) 1450–1455.
[8] S.A. Vinterbo, E.Y. Kim, L. Machado, Small, fuzzy and interpretable gene expression based classifiers, Bioinformatics 21 (2005) 1964–1970.
[9] F. Azuaje, A computational neural approach to support the discovery of gene function and classes of cancer, IEEE Trans. Biomed. Eng. 48 (2001) 332–339.
[10] C. Creighton, S. Hanash, Mining gene expression databases for association rules, Bioinformatics 19 (2003) 79–86.
[11] W.S. Bush, T.A. Thornton-Wells, M.D. Ritchie, Association rule discovery has the ability to model complex genetic effects, IEEE Symp. Comput. Intell. Data Min. 19 (2007) 624–629.
[12] P. Sethi, S. Alagiriswamy, Association rule based similarity measures for the clustering of gene expression data, Open Med. Inf. J. 4 (2010) 63–73.
[13] H. Mi, A. Muruganujan, P.D. Thomas, PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees, Nucleic Acids Res. 41 (2013) D377–D386.
[14] F. Lopez, Fuzzy association rules for biological data analysis: a case study on yeast, BMC Bioinform. 9 (2008) 1–18.
[15] C. Becquet, Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data, Genome Biol. 3 (2002) 1–16.
[16] Y. He, S.C. Hui, Exploring ant-based algorithms for gene expression data analysis, Artif. Intell. Med. 47 (2009) 105–119.
[17] J. Xia, M.J. Benner, R.E.W. Hancock, Networkanalyst – integrative approaches for protein–protein interaction network analysis and visual exploration, Nucleic Acids Res. 1 (2014) 1–8, http://dx.doi.org/10.1093/nar/gku443.
[18] Y. Chen, J. Hao, W. Jiang, T. He, X. Zhang, T. Jiang, R. Jiang, Identifying potential cancer driver genes by genomic data integration, Sci. Rep. 3 (2013) 1–9, http://dx.doi.org/10.1038/srep03538.
[19] A. Li, Mining association rules among gene functions in clusters of similar gene expression maps, in: IEEE International Conference on Bioinformatics and Biomedicine, 2009, pp. 254–259.
[20] S. Barik, D. Mishra, S. Mishra, S.K. Satapathy, A.K. Rath, M. Acharya, Pattern discovery using fuzzy FP-growth algorithm from gene expression data, Int. J. Adv. Comput. Sci. Appl. 5 (2010) 50–55.
[21] R. Sheibani, A. Ebrahimzadeh, An algorithm for mining fuzzy association rules, International MultiConference of Engineers and Computer Scientists 1 (2008) 1–5.
[22] C.H. Wang, C.T. Pang, Finding fuzzy association rules using FWFP-growth with linguistic supports and confidences, World Acad. Sci. Eng. Technol. 53 (2009) 1139–1147.
[23] A. Jaiswal, G. Dubey, Identifying best association rules and their optimization using genetic algorithm, Int. J. Emerg. Sci. Eng. 18 (2013) 91–96.
[24] F.A.G. Pach, J. Abonyi, Compact fuzzy association rule-based classifier, Expert Syst. Appl. 34 (2008) 2406–2416.
[25] T. Hong, C. Kuo, S. Chi, Mining association rules from quantitative data, Intell. Data Anal. 3 (1999) 363–376.
[26] Z. Farzanyar, M. Kangavari, Efficient mining of fuzzy association rules from the pre-processed dataset, Comput. Inf. 31 (2014) 331–347.
[27] C. Olsen, K. Fleming, N. Prendergast, R. Rubio, F. Emmert-Streib, G. Bontempi, B. Haibe-Kains, J. Quackenbush, Inference and validation of predictive gene networks from biomedical literature and gene expression data, Genomics 103 (2014) 329–336.
[28] D. Wong, C. Sweetman, C. Ford, Annotation of gene function in citrus using gene expression information and co-expression networks, BMC Plant Biol. 14 (2014) 1–17.
[29] I. Ponzoni, M.J. Nueda, S. Tarazona, S. Götz, D. Montaner, J.S. Dussaut, J. Dopazo, A. Conesa, Pathway network inference from gene expression data, BMC Syst. Biol. 8 (2014) 1–17.
[30] Y. Liu, C. Cheng, V.S. Tseng, Discovering relational-based association rules with multiple minimum supports on microarray datasets, Bioinformatics 27 (2011) 3142–3148.
[31] Genesense: a new approach for human gene annotation integrated with protein–protein interaction networks, Scientific Reports 4 (2014) 1–6. doi:10.1038/srep04474.
[32] R. Agrawal, T. Imielinski, A. Swami, Database mining: a performance perspective, IEEE Trans. Knowl. Data Eng. 5 (1993) 914–925.
[33] J. Han, J. Pei, Y. Yen, R. Mao, Mining frequent patterns without candidate generation: a frequent-pattern tree approach, Data Min. Knowl. Discov. 8 (2004) 53–87.
[34] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: International Conference on Very Large Databases, 1994, pp. 487–499.
[35] M.S. Khan, M. Muyeba, F. Coenen, D. Reid, H. Tawfik, Mining fuzzy association rules from composite items, Int. J. Data Warehous. Min. 7 (2011) 1–29.
[36] B.M. Al-Maqaleh, Discovering interesting association rules: a multi-objective genetic algorithm approach, Int. J. Appl. Inf. Syst. 5 (2013) 47–52.
[37] C. Chai, B. Li, A novel association rules method based on genetic algorithm and fuzzy set strategy for web mining, J. Comput. 5 (2010) 1448–1455.
[38] R. Kumari, J. Vashishtha, Discovery of fuzzy hierarchical association rules, Int. J. Comput. Appl. 98 (2014) 20–26.
[39] F. Coenen, P. Leng, G. Goulbourne, Tree structures for mining association rules, J. Data Min. Knowl. Discov. 15 (2004) 391–398.
[40] D.A. Chiang, N.P. Lin, Correlation of fuzzy sets, Fuzzy Sets Syst. 102 (1999) 221–226.
[41] N. Lin, H. Chen, H. Cheuh, W. Hao, C. Chang, A fuzzy statistics based method for mining fuzzy correlation rules, WSEAS Trans. Math. 6 (2007) 852–855.
[42] M.D. Cock, C. Cornelis, E.E. Kerre, Elicitation of fuzzy association rules from positive and negative examples, Fuzzy Sets Syst. 149 (2005) 73–85.
[43] http://ncbi.nlm.nih.gov/projects/geo/.
[44] G.D. Beer, Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat. Med. 8 (2002) 816–823.
[45] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. 96 (1999) 6745–6750.
[46] B.H. Mecham, G.T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D.Z. Wetmore, T.J. Mariani, I.S. Kohane, Z. Szallasil, Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements, Nucleic Acids Res. 32 (2004) e74.
[47] K.Y. Detwiller, N.T. Fernando, N.H. Segal, S.W. Ryeom, P.A. D'Amore, S.S. Yoon, Analysis of hypoxia-related gene expression in sarcomas and effect of hypoxia on RNA interference of vascular endothelial cell growth factor A, Cancer Res. 65 (2005) 5881–5889.
[48] N.C. Gutierrez, E.M. Ocio, J.D. Rivas, P. Maiso, M. Delgado, E. Ferminan, M.J. Arcos, M.L. Sanchez, J.M. Hernandez, J.F.S. Miguel, Gene expression profiling of B lymphocytes and plasma cells from Waldenstrom's macroglobulinemia: comparison with expression patterns of the same cell counterparts from chronic lymphocytic leukemia, multiple myeloma and normal individuals, Leukemia 21 (2007) 541–549.
[49] ftp://ftp.camda.duke.edu/CAMDA03-DATASETS/michigan-publication.zip.
[50] http://www.ncbi.nlm.nih.gov/Database.
[51] O. Podlaha, M. Riester, S. De, F. Michor, Evolution of the cancer genome, Adv. Artif. Intell. 28 (2012) 155–163.
[52] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, F-score and roc: a family of discriminant measures for performance evaluation, Adv. Artif. Intell. 4304 (2006) 1015–1021.