Clustering on Gene Expression and Fold Values: Identification of Some Possible Genes Mediating Allergic Asthma

Rajat K. De and Anindya Bhattacharya

Abstract—**The present article focuses on a systematic application of clustering algorithms (Fuzzy c-means (FCM) and Partitioning Around Mediod (PAM)) on gene expression data. We show a way of applying these algorithms to select some possible genes responsible for a particular disease. The genes those are severely over or under expressed in the allergen samples are identified. Two different techniques are applied on the same gene expression datasets containing both allergen samples and control samples. First technique uses clustering algorithms on expression values, followed by determining similarity/dissimilarity among control and disease clusters, and measuring the extent of over/under expression of genes from normal to disease condition. By the second technique, we apply clustering algorithms on fold values and measure the over/under expression of genes. By these two techniques we have identified several genes those have significantly changed their expression values for asthmatic condition, and have reported in the present article. Some of these observations are supported by some earlier investigations. Others have been stayed unnoticed so far, but may play crucial role in mediating the development of asthma.**

Index Terms—**Mouse, fold value, Jaccard score, DB-index, wild type samples, IL-13 knocked out samples.**

I. Introduction

CLUSTERING is one of the most important unsupervised learning techniques that deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects, which are "similar" among themselves and are "dissimilar" to the objects belonging to other clusters. A clustering algorithm belongs to either of the three general categories: Hierarchical, Partitional and Fuzzy clustering. Clustering algorithms use different distance measures for finding similarity/dissimilarity among all pairs of patterns and thereby group them into different clusters. Euclidean distance and Maholanobis distance are commonly used distance measures.

Although clustering algorithms have already been applied to various gene expression data [11], [15], [23], [33], [16], [21], [18], [26], [20], no definite technique is there to use clustering for identification of genes those are over/under expressed in a condition with respect to another condition. Number of clusters to be created from a particular input dataset is also not well defined. One major problem with hierarchical clustering technique is that it is likely to produce one single large cluster and several singletons. This makes them inefficient for gene expression data analysis.

The application of fuzzy set theoretic clustering algorithms in the domain of gene expression data is not explored much. Fuzzy set theoretic clustering methodologies incorporate the notion of fuzzy sets, which is a generalization of classical set theory. According to classical set theory, the boundary of a set is well defined, whereas there is no well defined boundary of a fuzzy set. Notion of fuzzy sets enables one to deal with uncertainties in different tasks, arising from deficiency (eg., vagueness, incompleteness etc.) in information, in an efficient manner. Some of the attempts made for grouping gene expression data using fuzzy set theoretic clustering algorithms include, among others, fuzzy c-means and normal mixture modelling based classification methods for separating microarray data into reliable and unreliable signal intensity populations [5], and fuzzy partitional clustering method (fuzzy c-means) in finding co-regulated genes [19].

In this article, we propose two techniques based on clustering (FCM and PAM) for identifying a set of some possible genes of interest. In order to determine number of cluster to be created from input dataset, we have used DB-index. First technique uses clustering algorithms on expression values, followed by determining similarity/dissimilarity among control and disease clusters, and measuring the extent of over/under expression of genes from normal to disease condition. On the other hand, the second technique considers the application of clustering algorithms to fold values. Both these methods are applied to a oligonucleotide micro array gene expression data GDS958. GDS958 contains expression patterns of as many as 22690 genes of both normal and allergic asthma samples, to identify a set of some possible genes over/under expressed in asthma samples compared to normal ones. These genes are selected from a large set of genes using the partitioning around mediod (PAM) [14] and fuzzy c-means (FCM) [6] clustering algorithms. A set of some possible genes mediating the development of asthma has been newly identified along with those already reported by some earlier investigations [13], [30], [9].

The remaining part of this paper is organized as follows.

Rajat K. De, Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India. Anindya Bhattacharya, Department of Computer Science and Engineering, Netaji Subhash Engineering College, Kolkata, India. Emails: rajat@isical.ac.in(R. De),anindyamail@rediffmail.com(A. Bhattacharya).

Section 2 describes the dataset GDS958. Section 3 highlights the methods used in analysis. Section 4 presents results and Section 5 provides some concluding remarks.

II. Data Description

The oligonucleotide microarray gene expression data (GDS958) generated by Wills-Karp et. al. [1], [31] from lung tissue of mouse were used here. Wills-Karp et. al. used GDS958 data set to find out responsible genes in asthma mediation [31].

The data set contains samples which had undergone either of the two types of strain, and are termed as Wild Type mouse and IL-13 Knocked Out mouse samples. Each of these Wild Type and IL-13 Knocked Out mouse samples can be either an Allergen sample or a Control sample. Two different types of treatments were used for both Wild Type strain and IL-13 Knocked Out strain: (i) House Dust Mite (HDM) and (ii) Phosphate Buffered Saline (PBS). Allergen samples were obtained from HDM treated Wild Type strain and IL-13 Knocked Out strain, whereas control samples from PBS treated Wild Type strain and IL-13 Knocked Out strain.

Thus we have four different types of sample: (1) HDM treated Wild Type allergen samples, (2) HDM treated IL 13 Knocked Out allergen samples, (3) PBS treated Wild Type control samples and (4) PBS treated IL 13 Knocked Out control samples. The data set GDS958 contains expression pattern for 22690 genes obtained from six Wild Type samples and six IL-13 knocked out samples. Three samples GSM21415, GSM21418 and GSM21420 are HDM treated Wild Type allergen samples, whereas GSM21422, GSM21424 and GSM21426 are PBS treated Wild Type control samples. GSM21403, GSM21405 and GSM21407 are HDM treated IL-13 knocked out allergen samples, whereas GSM21409, GSM21411 and GSM21413 are PBS treated IL-13 knocked out control samples. Further information on this data is available at [1]. It is to be mentioned here that the data set contains expression profiles of cytokines including IL-13, IL-4, IL-5 and their receptors being known as key mediators for allergen induced immediate development of asthma. This is due to hyperreactivity of the airway and mucus overproduction in the lung as immediate response for house dust mite allergen. The data set is last updated on December 12, 2004, so it is expected to contain recent information.

III. Methodology

The methodology we have used here are based on well known partitioning around medoid (PAM) and fuzzy c-means (FCM) clustering algorithms. Both these algorithms require an expected number k of clusters as input for their execution. For example, if we give $k = 3$ as input, the algorithms partition the entire data set into 3 clusters. Thus selecting a suitable value of k is crucial for these algorithms. In order to find the best k for a set of data points we are going to cluster, we have used a cluster validity index, namely, Davies-Bouldin index (DB-index) [10] which varies with the number of clusters k. Lower the value of DB-index, better is the clustering and vice versa. For details of the clustering algorithms and mathematical expression for DB-index, one may refer to [14], [10]. These clustering algorithms have been executed on the (i) expression values of genes given in the data set and (ii) fold values of genes define in Section 3.2.

A. Clustering on expression values

Clustering on expression values considers gene expression values in both normal (control) and disease samples. Let there be n number of normal samples and m number of disease samples. Each of the genes present in the data set is considered as a data point or a pattern. That is, each data point (*i.e.*, each gene) has dimension n for normal and m for disease sets. Each of these sets of genes were grouped into k number of clusters using the clustering algorithms. It is to be noted that we select k for which DB-index value attains a minimum. Thus for normal samples, we have (say) k_{normal} control clusters $S_1, S_2, \ldots, S_{k_{normal}}$ containing genes with different domains of expression values. Similar is the case for disease samples, where $k_{disease}$ number of clusters $T_1, T_2, \ldots, T_{k_{disease}}$ consist of genes with different domains of expression values. The genes in any of these clusters are expected to be co-expressed. On the other hand, genes in different clusters should have different domains of expression values.

Now we find the relationship (*i.e.*, similarity or dissimilarity) between the control and disease clusters. For this purpose, we use Jaccard score that is widely used in literature [17]. Let us assume that jth control cluster S_j has the highest degree of similarity with ith disease cluster T_i. Thus it is expected that the domains of expression values of genes in S_j and T_i are similar.

If T_i and S_j are found most similar using Jaccard score then the genes in the set $(T_i - (S_j \cap T_i))$ are either over expressed or under expressed in disease. If gene $g \in (T_i - (S_j \cap T_i))$ and $g \in S_k$ then g is over expressed if average expression values of genes in S_k is less than that in T_i. Otherwise g is under expressed. If there is only two control clusters then we can identify over/under expressed genes with lesser effort. In this case, we have two control clusters S_1 and S_2, and two diseased clusters T_1 and T_2.

If S_1 and T_1 contain genes in control and disease samples, respectively, with high expression values, then it is expected that most of the genes (with high expression values) in cluster S_1 are present in the cluster T_1. If a gene, say g, is present in T_1 but not in S_1 then we may infer that the gene g is over expressed in the disease samples. Thus the genes in the set $(T_1 - (S_1 \cap T_1))$ are over expressed in the disease samples. On the other hand, if both T_2 and S_2 contain genes with low expression values, the presence of gene g in T_2 but not in S_2 would lead to the inference of the gene g being under expressed in the disease samples. Thus the set $(T_2 - (S_2 \cap T_2))$ contains only the genes that are under expressed in the disease samples. The method is depicted pictorially in Fig. 1.

B. Clustering on fold values

Apart from the clustering on expression values, genes were grouped by these clustering algorithms on their fold values that are defined below. Let there be n number of control samples

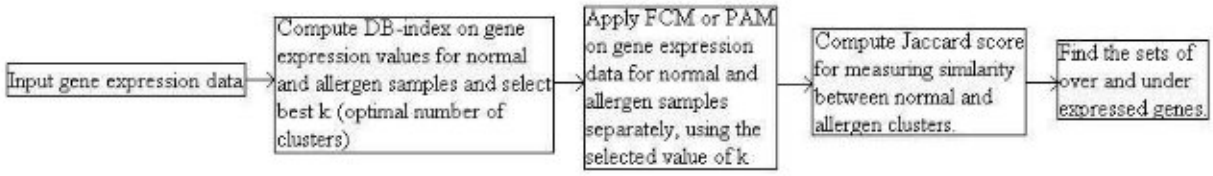

Fig. 1. Flowchart of the method based on gene expression values.

and m number of disease samples, and the expression values of a gene g in ith control sample be c_g^i and that in jth disease sample be a_g^j. Then the fold value f_g^{ij} of gene g is defined as

$$f_g^{ij} = a_g^j / c_g^i \tag{1}$$

Thus we have mn number of fold values of a gene g. We form an mn-dimensional fold vector $\mathbf{v}_g$ corresponding to gene g as

$$\mathbf{v}_g = [f_g^{11}, f_g^{12}, \ldots, f_g^{ij}, \ldots, f_g^{mn}]^T$$

If fold value f_g^{ij} of a gene g is nearly equal to one, expression values of the gene g in both ith control and jth disease samples are almost equal. Therefore, we may say that the gene g has no role in mediating the development of asthma. Fold value f_g^{ij} of gene g grater than one (or less than one) indicates that gene g gets over expressed (or under expressed) in jth disease sample compared to ith control sample. It is to be mentioned here that in an ideal situation if gene g has no role in mediating asthma, the corresponding fold vector would be closed to the vector $\mathbf{v}_0 = [1, 1, \ldots, 1]^T$, where all the components of $\mathbf{v}_0$ are 1.

In order to compute an average extent of over or under expression of genes in a disease cluster, we have measured Euclidean distance D_{avg} between the corresponding cluster center and $\mathbf{v}_0$, which is given by

$$D_{avg} = \sqrt{||\mathbf{v}_A - \mathbf{v}_0||} \tag{2}$$

Here $\mathbf{v}_A = [f_A^{11}, f_A^{12}, \ldots, f_A^{ij}, \ldots, f_A^{mn}]^T$, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, m$ is the center of an disease cluster. The higher the value of D_{avg}, the higher will be the extent of over or under expression of genes contained in the disease cluster. Moreover, in order to determine whether D_{avg} indicates an average over or under expression, we use a term $SIGN$ defined as

$$SIGN = \sum_{i=1}^{n} \sum_{j=1}^{m} (f_A^{ij} - 1) \tag{3}$$

If $SIGN > 0$ then D_{avg} indicates an average over expression. On the other hand, if $SIGN < 0$ then D_{avg} indicates an average under expression. For $SIGN > 0$, the highest D_{avg} value indicates that the genes in the disease cluster get the highest extent of over expression (on an average). Similarly, for $SIGN < 0$, the highest D_{avg} value indicates the highest extent of under expression (on an average). The method is depicted pictorially in Fig. 2

IV. Results and Analysis

Here we identify some possible genes mediating the development of asthma, by applying the aforesaid methodology based on the clustering algorithms (PAM and FCM) on the gene expression data set GDS958 containing 22690 genes. Some of these genes those are over or under expressed in the allergen samples were not known earlier. The clustering algorithms were applied both on expression values and fold values of genes.

TABLE I

Jaccard score computing similarity between allergen and control clusters.

Allergen cluster	Control cluster	For Wild Type samples		For Knocked Out samples	
		PAM	FCM	PAM	FCM
T1	S1	0.997392	0.997013	0.997968	0.996971
T1	S2	0.002117	0.002093	0.001174	0.001829
T2	S1	0.000363	0.000652	0.000748	0.000940
T2	S2	0.952389	0.967599	0.966094	0.969175

A. On expression values

In the case of clustering based on expression values of the data set GDS958 we separately analyzed wildtype and IL-13 knocked out samples. The dimension corresponding to a gene in wildtype normal and wildtype allergen samples are three, as $n = m = 3$. For IL-13 knocked out normal and IL-13 knocked out allergen samples, this dimension is again three as $n = m = 3$. For wild type samples, we have provided $k = 2$ as input to both these algorithms as DB-index attains its minimum at $k = 2$ for both PAM and FCM (Figure 3). Similarly, for IL-13 knocked out samples we have provided $k = 2$ as input to both these algorithm as DB-index attains its minimum at $k = 2$ for both PAM and FCM (Figure 4). Thus we have obtained two control clusters (S_1 and S_2) and two allergen clusters (T_1 and T_2). We have computed Jaccard score (Table I) to determine the similarity between control and allergen clusters. We have found, from Jaccard score (Table I), that the cluster S_1 is the most similar to T_1, and S_2 to T_2. In fact, the clusters S_1 and T_1 contain genes with low expression values, and S_2 and T_2 contain those with high expression values. Thus we infer that the genes being present in S_1 but not in T_1 are over expressed in the allergen samples. Similarly, the genes in S_2 but not in T_2 are under expressed in the allergen samples.

By applying both PAM and FCM, we have found a few genes, out of the entire set of 22690 genes, which are over expressed in the allergen samples. Some of these observations were already found in [13], [32], [30], [29], [12], [22], [28], [35], [34]. In the present investigation, we have found some new genes including Gpnmb (glycoprotein (transmembrane) nmb), Serpina3n (serine (or cysteine) proteinase inhibitor, clade A, member 3N), Slc26a4 (solute carrier family 26, member 4), Igl-V1 (immunoglobulin lambda chain, variable

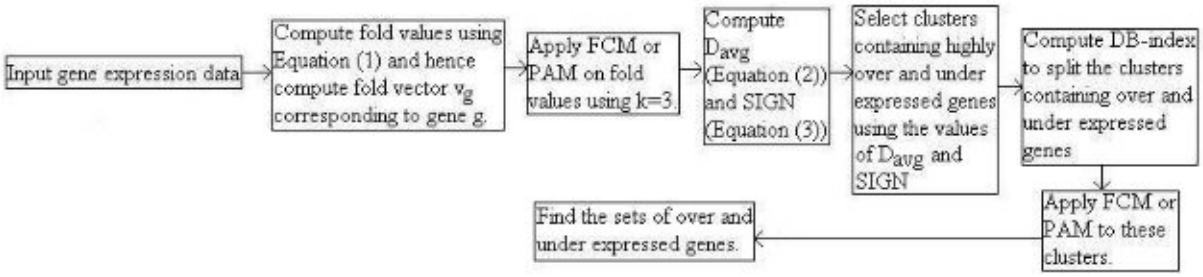

Fig. 2. Flowchart of the method based on gene fold values.

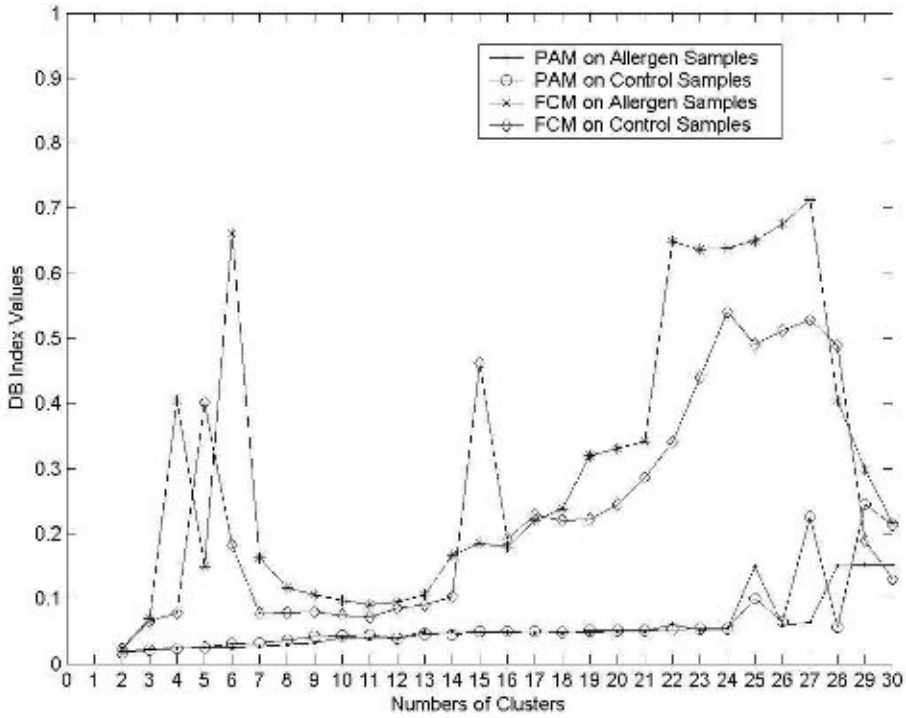

Fig. 3. Variation of DB-index computed on expression values with respect to k.

1), A430103C15Rik (RIKEN cDNA A430103C15 gene), Igh-4 (immunoglobulin heavy chain 4 (serum IgG1)), AI324046 (expressed sequence AI324046), IgK-V1 (immunoglobulin kappa chain variable 1 (V1)), IgK-V5 (immunoglobulin kappa chain variable 5 (V5 family)), which are over expressed in both the wildtype allergen samples and IL-13 knocked out allergen samples. Fxyd4 (FXYD domaincontaining ion transport regulator 4) is over expressed in the wildtype allergen samples but not over expressed in IL-13 knocked out allergen samples. None of them were previously found responsible for asthma, but in our experiment, very large changes in expression values for all of these genes in the allergen samples suggest that they may have very important role in mediating the development asthma.

Likewise, the set of genes that are significantly under expressed include Ppp1r2 (protein phosphatase 1, regulatory (inhibitor) subunit 2), Cyp2a4 (cytochrome P450, family 2, subfamily a, polypeptide 4), Tncc (troponin C, cardiac/slow skeletal), Dusp16 (dual specificity phosphatase 16), Scnn1b (sodium channel, nonvoltage-gated 1 beta) and Slc7a10 (solute carrier family 7 (cationic amino acid transporter, y+ system), member 10) are some of the significantly under expressed genes found in both the wildtype allergen samples and IL-13 knocked out allergen samples. These genes have newly been found under expressed in the allergen samples. Some of the observations that are in accordance with earlier investigations include over expression of Clca3 (chloride channel calcium activated 3) [22], Ctsk (cathepsin k) [12], Ear11 (eosinophil-associated, ribonuclease A family, member 11) [9], Sprr2a (small proline-rich protein 2A) [34], Chi3l3 (chitinase 3-like 3) [29] and Arg1 (Arginase) [28], [35] in wildtype and IL-13 knocked out allergen samples.

B. On fold values

Clustering algorithms were executed on the fold values. Here we are looking for three clusters corresponding to (i) the genes not over/under expressed in the allergen samples, (ii) those over expressed in the allergen samples and (iii) those under expressed in the allergen samples. Thus we have considered $k = 3$ as input to these algorithms. It is to be

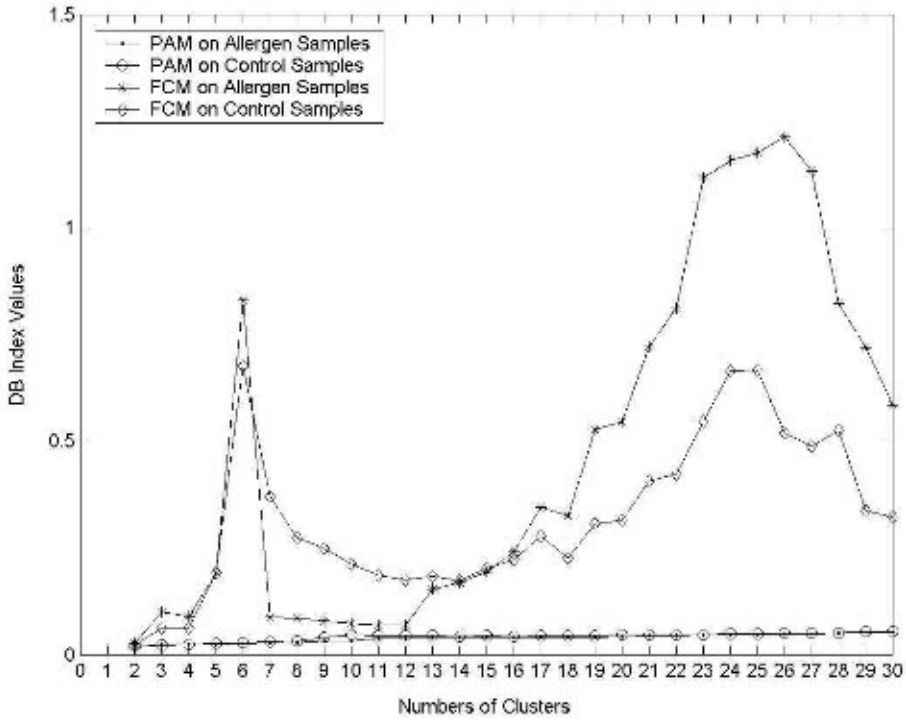

Fig. 4. Variation of DB-index computed on expression values with respect to k.

mentioned here that we have three control samples for both wild type and IL-13 knocked out, *i.e.*, $n = 3$, and three allergen samples for both wild type and IL-13 knocked out, *i.e.*, $m = 3$. So the fold vector $\mathbf{v}_g$ of a gene g has the dimension of 9 $(= mn)$ for both wildtype and IL-13 knocked out cases.

It is to be mentioned here that the clusters containing over or under expressed genes may also include those which were not (over or under) expressed much compared to those in control samples. In order to find the genes which were much over or under expressed compared to those in control samples, we separately analyzed wildtype samples and IL-13 knocked out samples. We have further divided the sets of over or under expressed genes. For this purpose, we have used DB-index to find the number of clusters into which each of these sets to be divided further.

The cluster containing the over expressed genes based on wild type samples was divided into 2 sub clusters (for both PAM and FCM) as the DB-index attains a minimum at $k = 2$ (Figure 5). The cluster containing the over expressed genes based on IL-13 knocked out samples was divided into 4 sub clusters for PAM and 2 sub clusters for FCM as the DB-index attains a minimum at $k = 4$ for PAM and $k = 2$ for FCM (Figure 6). The distances of these cluster centers from the vector $\mathbf{v}_0$ were computed. This distance shows the extent of over expression of the genes (contained in the corresponding cluster) in allergen samples as compared to those in control samples. The genes in the cluster with center farthest from the vector $\mathbf{v}_0$ have the highest degree of over expression.

Considering the cluster containing the most over expressed genes, as shown by the distance from its center to the vector $\mathbf{v}_0$, 158 genes were found to be the most over expressed in the wild type allergen samples obtained by both PAM and FCM. Out of these 158 genes, Gpnmb, Fxyd4, Serpina3n, Slc26a4, Igl-V1, A430103C15Rik, Igh-4, AI324046, IgK-V1, IgK-V5 are some examples. Gpnmb, Serpina3n, Slc26a4, Igl-V1, A430103C15Rik, Igh-4, AI324046, IgK-V1, IgK-V5 are also found most over expressed based on the IL-13 knocked out allergen samples of the genes. It is to be mentioned here that the over expressed genes Clca3 [22], Ctsk [12], Sprr2a [34], Chi3l3 [29] and Arg1 [28], [35] obtained by this method have also been found by some earlier investigations. Clca3, Ctsk, Sprr2a, Chi3l3 and Arg1 found over expressed in both wild type and IL-13 knocked out samples.

Similarly, the cluster containing under expressed genes based on wild type samples was divided into 2 sub clusters for PAM and 3 sub clusters for FCM, as DB-index attains its minimum at $k = 2$ for PAM and at $k = 3$ for FCM (Figure 7). The cluster containing under expressed genes based on IL-13 knocked out samples was divided into 2 sub clusters for both PAM and FCM, as DB-index attains its minimum at $k = 2$ for both PAM and FCM (Figure 8). As in the case of over expressed genes, distances of all the cluster centers from the vector $\mathbf{v}_0$ is computed to check the extent of under expression. The cluster containing the genes having the highest extent of under expression in allergen samples compared to that in control samples is selected based on these distance values. Using the above process based on the distance measure, the farthest cluster is identified. As in the case of clustering based on expression values, Ppp1r2, Cyp2a4, Tncc, Dusp16, Scnn1b and Slc7a10 are some of the possible examples of under expressed genes in both wild type and IL-13 knocked out samples. In order to restrict the size of the article, we have not reported all these over/under expressed genes here.

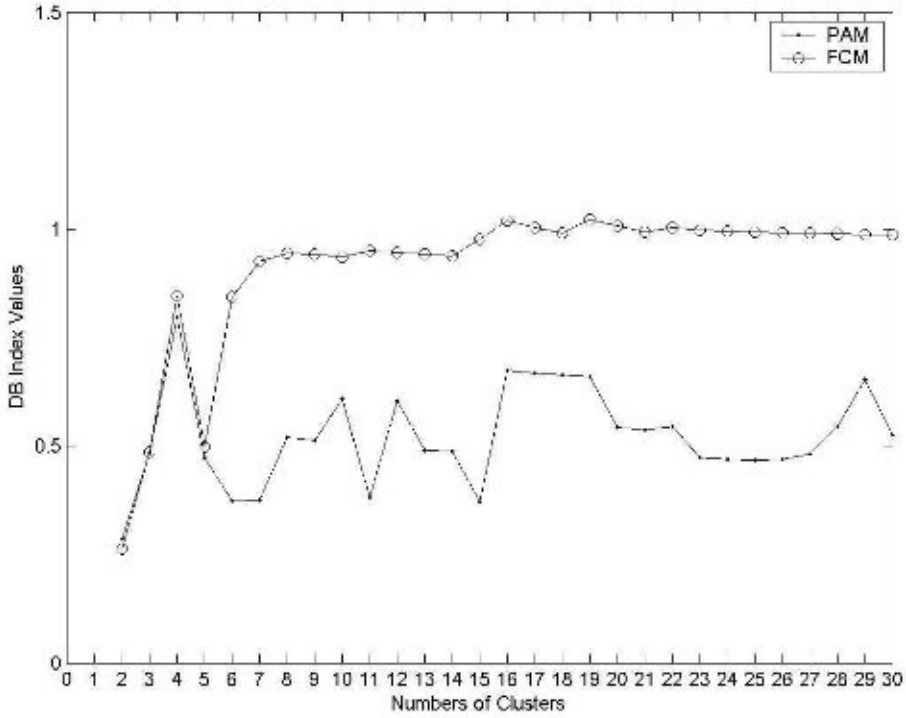

Fig. 5. Variation of DB-index computed on fold values of over expressed genes with respect to k.

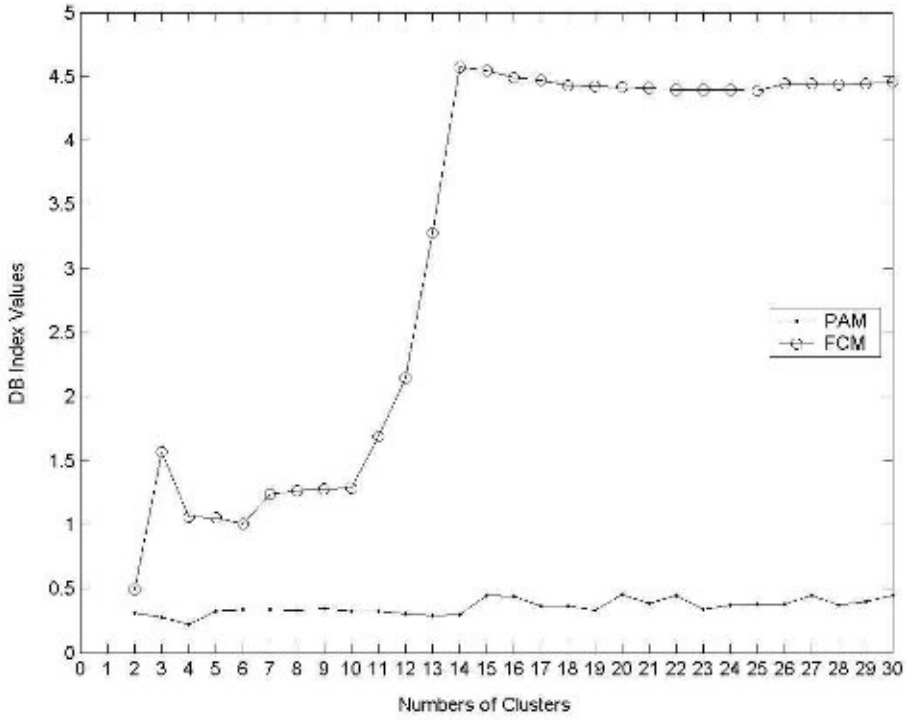

Fig. 6. Variation of DB-index computed on fold values of over expressed genes with respect to k.

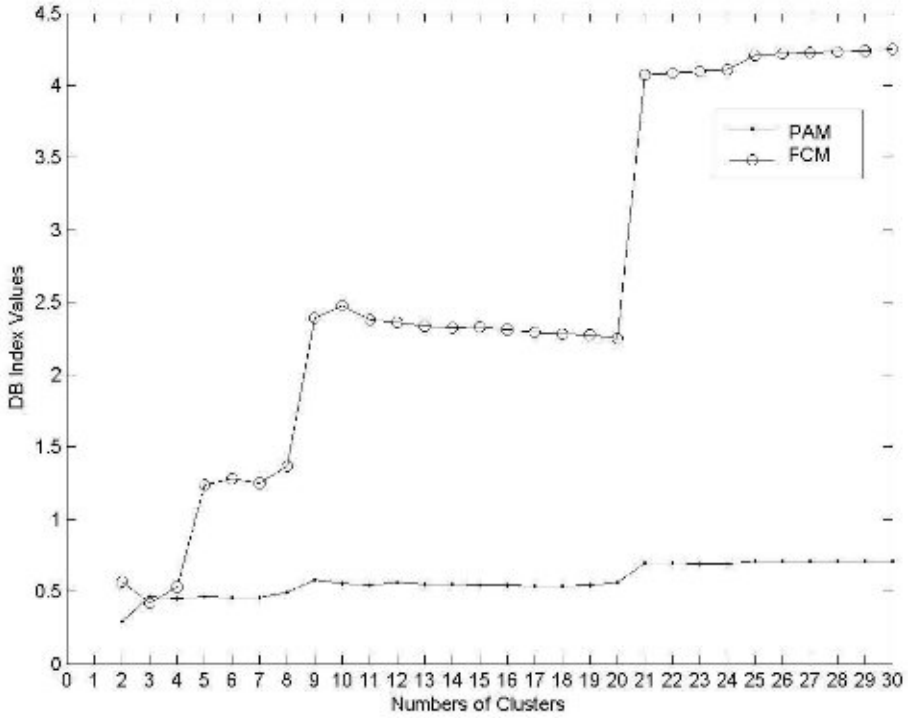

Fig. 7. Variation of DB-index computed on fold values of under expressed genes with respect to k.

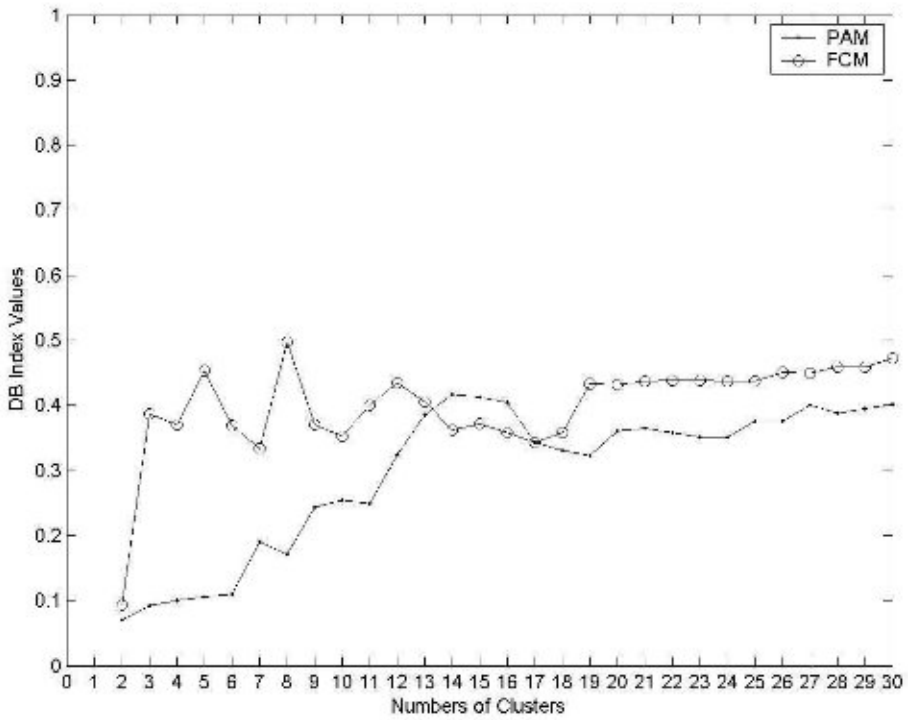

Fig. 8. Variation of DB-index computed on fold values of under expressed genes with respect to k.

TABLE II
LIST OF GENES NEWLY FOUND OVER EXPRESSED IN ALLERGEN SAMPLES.

GenBank Accession Number	Gene Title	Gene Symbol
NM_053110	glycoprotein (transmembrane) nmb	Gpnmb
NM_033648	FXYD domain-containing ion transport regulator 4	Fxyd4
NM_009252	serine (or cysteine) proteinase inhibitor, clade A, member 3N	Serpina3n
NM_011867	solute carrier family 26, member 4	Slc26a4
AK008145	immunoglobulin lambda chain, variable 1	Igl-V1
BC024402	RIKEN cDNA A430103C15 gene	A430103C15Rik
BC008237	immunoglobulin heavy chain 4 (serum IgG1)	Igh-4
AK007826	immunoglobulin heavy chain (J558 family)	Igh-VJ558
BE687919	expressed sequence AI324046	AI324046
295478	immunoglobulin kappa chain variable 1 (V1)	IgK-V1
U29768	immunoglobulin kappa chain variable 5 (V5 family)	IgK-V5
BM250782	tumor necrosis factor receptor sPAM, FCMer family, member 9	Tnfrsf9
BB113173	cDNA sequence BC032204	BC032204

TABLE III
LIST OF GENES NEWLY FOUND UNDER EXPRESSED IN ALLERGEN SAMPLES.

GenBank Accession Number	Gene Title	Gene Symbol
NM_025800	protein phosphatase 1, regulatory (inhibitor) subunit 2	Ppp1r2
NM_007812	cytochrome P450, family 2, subfamily a, polypeptide 4	Cyp2a4
NM_009393	troponin C, cardiac/slow skeletal	Tncc
NM_130447	dual specificity phosphatase 16	Dusp16
NM_011325	sodium channel, nonvoltage-gated 1 beta	Scnn1b
NM_017394	solute carrier family 7 (cationic amino acid transporter, y+ system), member 10	Slc7a10

Thus we have found some possible new genes including Gpnmb, Fxyd4, Serpina3n, Slc26a4, Igl-V1, A430103C15Rik, Igh-4, AI324046, IgK-V1, IgK-V5 which were found over expressed in the allergen samples, by applying both PAM and FCM, executed on both expression and fold values (Table II). Similarly, Ppp1r2, Cyp2a4, Tncc, Dusp16, Scnn1b and Slc7a10 were found to be under expressed in allergen samples (Table III). We may infer from this fact that these genes may play important role in mediating the development of asthma.

C. Biological validation of results on over expressed genes

In our results we found Igl-V1, IgK-V1, IgK-V5, Igh-4, Igh-VJ558 and AI324046 significantly over expressed in asthma samples. These genes have been found to take part in construction of Immunoglobulin [2]. So over expression of these genes may have significant role to play for allergic asthma by increasing production of Immunoglobulins.

Other genes found significantly over expressed in asthmatic samples compared to normal samples include Gpnmb, Tnfrsf9, BC032204, Serpina3n, Fxyd4 and SLC26A4. The protein encoded by Gpnmb gene is a type I transmembrane glycoprotein. Another type I transmembrane glycoprotein gene CD44 already known to have a role in the pathogenesis of asthma [24]. Gpnmb may have similar important role in the pathogenesis of asthma. Tnfrsf9 gene has been found responsible negative regulation of cell proliferation and to play a role in immune response also. The GEArray S Series Mouse Autoimmune and Inflammatory Response Gene Array [3] is designed to profile the expression of genes known to be involved in and important to immune responses such as autoimmunity and inflammation. It places gene Tnfrsf9 to functional gene group Cytokines & Receptors along with all known cytokines (IL-13, IL-4, ...) responsible for allergic asthma.

BC032204 gene found responsible for cell adhesion. Previous findings on expression of cell adhesion molecules in asthma [7] suggest that BC032204 may have very important role to play in mediation of asthma. Serpina3n is expressed at high levels in brain, heart, liver, lung, spleen, testis and thymus, and at low levels in bone marrow, kidney and skeletal muscle. Serpina3n gene provides instructions for making a protein called Serine protease inhibitor A3N. Genetic Association Database [4] shows that Serpina3n is associated with asthma. Fxyd4 gene provides instructions for making a protein called FXYD domain-containing ion transport regulator 4. Potentially this is a type I membrane protein. Fxyd4, originally named CHIF for channel-inducing factor, has been shown to modulate the properties of the Na, K-ATPase [25]. A previous study [8] shows that the activities of Na, K-ATPase were decreased in patients suffering from asthma and these abnormalities may modulate the clinical severity of asthma.

SLC26A4 gene provides instructions for making a protein called pendrin. This protein transports negatively charged particles (particularly chloride, iodide and bicarbonate) across cell membranes. Previous results [27] confirm the involvement of the CFTR (cystic fibrosis transmembrane conductance regulator, ATP-binding cassette (sub-family C, member 7)) gene in asthma. CFTR markedly activates Cl and OH/HCO3 transport by pendrin. Increased expression level of SLC26A4 gene may have a role to play in asthma by increasing the level of pendrin. Thus biological information about all newly found over expressed genes shows that these genes may play important roles in mediating the development of asthma.

D. Biological validation of results on under expressed genes

We have found several genes including Ppp1r2, Tncc, Dusp16, Scnn1b, Slc7a10 that are heavily under expressed in asthmatic condition. Biological information about Ppp1r2, Tncc, Dusp16, Scnn1b and Slc7a10 genes and their functionlaities are available in the web site [2]. The site shows that Ppp1r2 gene is an inhibitor of protein-phosphatase 1. Troponin (Tn) is the central regulatory protein of striated muscle contraction, which consists of three components, *viz.*, Tn-I (inhibitor of actomyosin ATPase), Tn-T (containing the binding site for tropomyosin and Tn-C) and Tn-C (inactivating the inhibitory action of Tn on actin families on binding calcium to it). It is to be noted that Tn-C protein is produced by Tncc gene. DUSP16 gene is involved in the inactivation of MAP kinases. Scnn1b belongs to the amiloride-sensitive sodium channel family. Scnn1b controls the reabsorption of sodium in kidney, colon, lung and sweat glands. Slc7a10 may play a role in the modulation of glutamatergic transmission through mobilization of D-serine at the glutamatergic synapse.

V. CONCLUSIONS

The present article has identified several possible over and under expressed genes, using clustering based techniques, mediating the development of asthma. It is to be mentioned here that although clustering algorithms have been applied on various gene expression data, those based on fuzzy set theory have not been considered much, specifically in the domain asthma gene expression samples. The methodology involves the application of clustering algorithms on expression and fold values of genes, followed by determining similarity/dissimilarity among various clusters and measuring the

extent of over and under expression of genes contained in different clusters.

On applying the said methodology, we have identified Clca3, Ctsk, Earl1, Sprr2a, Chi3l3 and Arg1 genes that are greatly over expressed in both wild type and IL-13 knocked out asthma samples. Previous work and biological evidences validate some results. We have identified several possible new genes, by the techniques applied to both expression and fold values, including Gpnmb, Fxyd4, Serpina3n, Slc26a4, Igl-V1, A430103C15Rik, Igh-4, AI324046, IgK-V1, IgK-V5 that are severely over expressed in allergic asthma condition. Similarly, we have found several genes including Ppp1r2, Cyp2a4, Tncc, Dusp16, Scnn1b, Slc7a10 that are heavily under expressed in asthmatic condition. Therefore, these genes may be considered as mediators for asthma.

REFERENCES

[1] http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=958.
[2] http://biostatpub2.mdanderson.org/genecards/.
[3] http://www.eurogentec.com/module/FileLib/AR-SAMM-602.3.pdf.
[4] http://geneticassociationdb.nih.gov.
[5] M. H. Asyali, M. Alci, Reliability analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods, Bioinformatics 21 (5) (2005) 644–649.
[6] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[7] B. S. Bochner (ed.), Adhesion Molecules in Allergic Disease, Marcel Dekker Ltd, 1997.
[8] S. K. Chhabra, A. Khanduja, D. Jain, Increased intracellular calcium and decreased activities of leucocyte na+,k+-atpase and ca2+-atpase in asthma, Genomics 97 (1999) 595–601.
[9] S. A. Cormier, S. Yuan, J. R. Crosby, C.A.Protheroe, D.M.Dimina, E.M.Hines, N.A.Lee, J.J.Lee, Th2-mediated pulmonary inflammation leads to the differential expression of ribonuclease genes by alveolar macrophages, American Journal of Respiratory Cell and Molecular Biology 27 (2002) 679–687.
[10] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (2) (1979) 224–227.
[11] M. J. Edick, C. Cheng, W. Yang, M. Cheok, M. R. Wilkinson, D. Pei, W. E. Evans, L. E. Kun, C. H. Pui, M. V. Relling, Lymphoid gene expression as a predictor of risk of secondary brain tumors, Genes Chromosomes Cancer 42 (2) (2005) 107–116.
[12] E. Godat, F. Lecaille, C. Desmazes, S. Duchene, E. Weidauer, P. Saftig, D. Bromme, C. Vandier, G. Lalmanach, Cathepsin k: a cysteine protease with unique kinin-degrading properties, Biochemical Journal 383 (Pt 3) (2004) 501–506.
[13] G. Grunig, M. Warnock, A. E. Wakil, R. Venkayya, F. Brombacher, D. M. Rennick, D. Sheppard, M. Mohrs, D. D. Donaldson, R. M. Locksley, D. B. Corry, Requirement for il-13 independently of il-4 in experimental asthma, Science 282 (1998) 2261–2263.
[14] J. Han, M.Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, CA, USA, 2001.
[15] H. S. Heo, J. H. Kim, Y. J. Lee, S. H. Kim, Y. S. Cho, C. G. Kim, Microarray profiling of genes differentially expressed during erythroid differentiation of murine erythroleukemia cells, Mol Cells 20 (1) (2005) 57–68.
[16] E. R. Hruschka, R. J. Campello, L. de Castro, Evolving clusters in gene-expression data, Information Sciences 176 (12) (2006) 1898–1927.
[17] A. K. Jain, M. N. Murty, P. J. Flyn, Data clustering: A review, ACM Computing Surveys 31 (3) (1999) 264–323.
[18] P. Jonsson, K. Laurio, Z. Lubovac, B. Olsson, M. L. Andersson, Using functional annotation to improve clusterings of gene expression patterns, Information Sciences 145 (3-4) (2002) 183–194.
[19] S. Y. Kim, J. W. Lee, J. S. Bae, Effect of data normalization on fuzzy clustering of dna microarray data, BMC Bioinformatics 7 (2006) 134.
[20] A. Lindlf, B. Olsson, Could correlation-based methods be used to derive genetic association networks?, Information Sciences 146 (1-4) (2002) 103–113.
[21] M. Molla, P. Andreae, J. Glasner, F. Blattner, J. Shavlik, Interpreting microarray expression data using text annotating the genes, Information Sciences 146 (1-4) (2002) 75–88.
[22] A. Nakanishi, S. Morita, Role of gob-5 in mucus overproduction and airway hyperresponsiveness in asthma, in: Proceedings of the National Academy of Sciences of the United States of America, USA, 2001.
[23] R. B. Roth, P. Hevezi, L. Lee, D. Willihite, S. M. Lechner, A. C. Foster, A. Zlothik, Gene expression analyses reveal molecular relationships among 20 regions of the human cns, Neurogenetics 7 (2) (2006) 67–80.
[24] M. E. Rothenberg, Cd44 a sticky target for asthma, J. Clin. Invest. 111 (2003) 1460–1462.
[25] K. J. Sweadner, E. Rael, The fxyd gene family of small ion transport regulators or channels: cdna sequence, protein signature sequence, and expression, Genomics 68(1) (2000) 41–56.
[26] K. Torkkola, R. M. Gardner, T. Kaysser-Kranich, C. Ma, Self-organizing maps in mining gene expression data, Information Sciences 139 (1-2) (2001) 79–96.
[27] M. Tzetis, A. Efthymiadou, S. Strofalis, P. Psychou, A. Dimakou, E. Pouliou, S. Doudounakis, E. Kanavakis, Cftr gene mutations - including three novel nucleotide substitutions - and haplotype background in patients with asthma, disseminated bronchiectasis and chronic obstructive pulmonary disease, Human Genetics 108(3) (2001) 216–221.
[28] D. Vercelli, Arginase: marker, effector, or candidate gene for asthma?, The Journal of Clinical Investigation 3 (2003) 1815–1817.
[29] J. S. Welch, L. Escoubet-Lozach, D. B. Sykes, K. Liddiard, D. R. Greaves, C. K. Glass, Th2 cytokines and allergic challenge induce ym1 expression in macrophages by a stat6-dependent mechanism, The Journal of Biological Chemistry 277 (2002) 42821–42829.
[30] M. Wills-Karp, Interleukin-13 in asthma pathogenesis, Current Allergy and Asthma Reports 4 (2004) 123–131.
[31] M. Wills-Karp, S. L. Ewart, Time to draw breath: asthma-susceptibility genes are identified, Nat Rev Genet. 5 (2004) 376–87.
[32] M. Wills-Karp, J. Luyimbazi, X. Xu, B. Schofield, T. Y. Neben, C. L. Karp, D. D. Donaldson, Interleukin-13: Central mediator of allergic asthma, Science 282 (1998) 2258–2261.
[33] S. Zhou, Y. Liu, H. Bo, X. Bian, X. Xia, C. Lin, V. W. Wongs, Z. Lu, Expression profilings of 39 genes selected by anova could separate precursors of murine dendritic cells and macrophages, Biochem Biophys Res Commun 344 (1) (2006) 438–445.
[34] N. Zimmermann, M. Doepker, D. Witte, K. Stringer, P. Fulkerson, S. Pope, E. Brandt, A. Mishra, N. King, N. Nikolaidis, M. Wills-Karp, F. Finkelman, M. Rothenberg, Expression and regulation of small proline-rich protein 2 in allergic inflammation, American Journal of Respiratory Cell and Molecular Biology 32 (5) (2005) 428–435.
[35] N. Zimmermann, N. E. King, J. Laporte, M. Yang, A. Mishra, S. M. Pope, E. E. Muntel, D. P. Witte, A. A. Pegg, P. S. Foster, Q. Hamid, M. E. Rothenberg, Dissection of experimental asthma with dna microarray analysis identifies arginase in asthma pathogenesis, The Journal of Clinical Investigation 3 (2003) 1863–1874.