Volume 4, Issue 5, May 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

KEGG Analysis of De-noised Mutual Information based Microarray Data using Empirical Mode Decomposition

Sanchita Mallick
Electronics & Communication Dept,
Adamas Institute of Technology, Barasat,W.B
India

Sankhamitra Roy
Electronics & Communication Dept,
Adamas Institute of Technology, Barasat, W.B
India

*Abstract— **The expression of thousands of genes can be monitored in parallel with the help of the Microarray Technique. Clustering methods have become a key step in microarray data analysis because it can identify groups of genes or samples displaying a similar expression profile. The clustering approach used here is the Fuzzy C means method in which one gene or one piece of data can belong to more than one cluster. Clustering co-expressed genes usually requires the definition of 'distance' or 'similarity' between measured datasets, the most common choices being Pearson correlation or Euclidean distance. Here we use mutual information i.e. nonlinear distance measurement in cluster analysis and visualization of large-scale gene expression data. Noise can be present in the gene expression data. The noise in data would affect the clustering results. Noise can be removed from gene expression data by a simple denoising scheme called Emperical Mode Decomposition(EMD. FCM method is combined with EMD for clustering microarray data in order to reduce the effect of noise. . The results shows the clustering structures of de-noised data are more reasonable and genes have tighter association with their clusters. After obtaining the results i.e. with EMD and without EMD, the validation has been done based based on KEGG pathways. The aim is to determine whether the clusters produced by the different dissimilarity measures are enriched with KEGG pathways.***

Keywords— Microarray data Clustering, fuzzy C-means method, Emperical Mode Decomposition, Intrinsic mode functions, KEGG pathway

I. Introduction

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Since the work of Eisen and Colleagues, [1] cluster algorithms generally aim at grouping objects according to some notion of similarity. For microarray data [2], clustering may be applied to the genes whose expression levels are measured, with the expectation that functionally related or co–regulated genes will show similar expression patterns [3,4]. None of the existing clustering algorithms perform significantly better than the others when tested across multiple datasets as microarray datasets tend to have very diverse structures. Some even do not have well defined clustering structures due to the complex nature of biological systems [5]. Among different approaches of clustering, fuzzy C-means (FCM) method is an efficient one [6]. In this method one gene or one piece of data can belong to more than one cluster unlike in hard clustering where a gene strictly belongs to one cluster.

Clustering co-expressed genes calculates 'distance' or 'similarity' between measured datasets using linear distance measurement like Pearson correlation or Euclidean distance[7,8] . Information theory based approach, the mutual information[9], also provides a general measure of dependencies between datasets. Here this approach is applied to determine the dependencies between datasets.

However, microarray data contains noise and the noise would affect clustering results. EMD, Emperical Mode Decomposition [10] can be used to remove noise from micro array data. The key feature of EMD is to decompose a signal into intrinsic mode function. There is some difficulties in EMD algorithm. These drawbacks can be overcome with modified algorithm [11]. Most noisy IMFs are considered as noise in the signal. If these noisy IMFs are removed from the raw data, the trend component can be obtained [12]. The trend can be used as denoised data to perform clustering analysis. FCM method is combined with EMD for clustering microarray data in order to reduce the effect of noise [13]. Once the clustering results are obtained KEGG pathway enrichment analysis can be conducted to validate the clustering results.

In this paper, here we used partial dataset which is extracted from a yeast cell cycle dataset generated by Spellman et al 1998 [14]. The dataset has 500 genes at 12 different conditions. Here instead of using the entire dataset which contains 4382 genes, a partial dataset which has 500 genes is used for clustering because of the variability of the data due to noise across the different condition. This may lead to convergence problems in Fuzzy C Means clustering algorithm. We applied the modified EMD algorithm to remove the noise [11]. Then FCM method is combined with this modified EMD algorithm to get better results. To calculate 'distance' or 'similarity' we used both linear and nonlinear measure of dependencies. After getting the clustering results, KEGG (Kyoto Encyclopedia of Genes and Genomes) pathwayenrichment analysis is done to validate the clustering results. Here KEGG pathway enrichment analysis is done by using a software called EXPANDER [15,16].

II. Methods

A. Empirical Mode Decomposition:

The empirical mode decomposition method was proposed by Huang et al. [11] and it was originally designed for nonlinear and non-stationary data analysis. Here, we give a brief introduction of EMDs [10,11]. EMD decomposes a time series into components which are called Intrinsic Mode Functions (IMFs) and a final residue is called 'trend'. Here intrinsic mode functions are considered as noise in microarray data and the trend as the de-noised data which we used for clustering analysis. Suppose a signal, X(t), the IMFs are found by an iterative procedure called sifting algorithm which is composed of the following steps:

(a) Find all the local maxima and local minima

(b) Compute the corresponding interpolating signals M(t) and m(t).These signals are the upper and lower envelopes of the signal.

(c) Let e(t)= (M(t)+m(t))/2

(d) Subtract e(t) from the signal: x(t):=x(t)-e(t)

(e) Return to step (a)—stop when X(t) remains nearly unchanged.

(f) Once we obtain an IMF, f(t), remove it from the signal x(t):=x(t)-f(t) and return to (a) if x(t) has more than one extremum (neither a constant nor a trend).

The number of extrema should decrease when going from one IMF to the next, and the whole decomposition is expected to be completed with a finite number of IMFs.

An intrinsic mode function is defined by two conditions:

(i) The numbers of extreme values and zero-crossings must be either equal or differ at most by one in the whole time series.

(ii) The mean value of the envelopes defined by local maxima and local minima is zero.

If the conditions of IMF are not satisfied after one iteration of aforementioned procedure, the same procedure is applied to the residue signal until properties of IMF are satisfied.

A.1. Some attempts to get a better algorithm

A.1.1. The mean envelope removal

This is an important aspect of the algorithm [10], since we may be adding a non-existing component that can distort the actual IMF and will appear in, at least, one of the following IMFs. To attenuate this we modified step (d) by introducing a step size (0≤ α ≤1) : x(t) =x(t)-αe(t). This increases the iteration time , but the algorithm becomes more reliable.

A.1.2. The stopping criterion

To state a stopping criterion in the sifting procedure, a resolution factor is defined which is the ratio between the energy of the signal at the beginning of the sifting, x(t) and the energy of average of the envelopes, e(t). If the ratio grows above the allowed resolution, then the IMF computation must stop. This criterion gives a scale independent stopping way, as opposed to criteria based on iteration count.

B. Fuzzy C-means algorithm

Fuzzy C-means Clustering allows on data to belong to two or more clusters. Any clustering algorithm develops a membership matrix that would describe the association of all the genes in the dataset with a specified number of clusters, The order of the membership matrix is K x N where K is the number of clusters and N is the number of genes.. The matrix may be represented as U=u_{kj}, where k = 1, ... , K and j = 1, ... , N. The following conditions holds good on U, $0 \le u_{kj} \le 1, \sum_{j=1}^{N} u_{jk} \le N, \sum_{k=1}^{K} u_{kj} = 1$ and $\sum_{k=1}^{K} \sum_{j=1}^{N} u_{jk} = N$. The data used for clustering is gene expression data where rows correspond to genes and the columns correspond to samples. The membership value u_{kj} close to 1 indicates that gene j is strongly associated to cluster k and u_{kj} close to 0 indicates a weak association . The clustering algorithm used here is fuzzy C-means (FCM) algorithm [13]. The final membership matrix U and the final cluster centroid vector C are obtained after the convergence of the FCM algorithm i.e. minimization of the objective function. The highest membership value of a gene tells the cluster to which it belongs to. The algorithm is as follows:

$$J(K,m)=\sum_{k=1}^{K} \sum_{j=1}^{N} u_{kj}^{m} (d_{kj}^{2}) \ldots\ldots (1)$$

Where

$$d_{kj}^{2} = ||x_j - c_k|| = (x_j - c_k)^T A(x_j - c_k) \ldots\ldots (2)$$

$$c_k = \sum_{k=1}^{K} u_{kj}^{m} x_j / \sum_{k=1}^{K} u_{kj}^{m} \ldots\ldots (3)$$

For a given gene *j*, the sum of membership values for each of the clusters should satisfy the following equation

$$\sum_{k=1}^{K} u_{jk} = 1, \quad 0 \le u_{jk} \le 1 \ldots\ldots (4)$$

The initiation of the algorithm requires the predetermination of the cluster number K the fuzziness parameter m and the initial cluster centroid. The cluster centroids c_k are calculated from (3). New membership values are calculated from

$$u_{kj} = 1/ \sum_{i=1}^{K} (d_{kj}/d_{ki})^{2/(m-1)} \ldots\ldots (5)$$

Iterative calculations based on Eqs. (2), (3), and (5) are then performed until the stop criterion is reached, i.e., minimization of objective function i.e Eq. (1).

C. Mutual Information: Detecting and evaluating dependencies between variables:

The Mutual information (MI) provides a general measure for dependencies in the data, in particular, positive, negative and nonlinear correlations. It is a well known measure in information theory that has been used to analyze gene-expression data. The used MI measure requires the expression patterns to be represented by discrete random variables.

C.1. *Kernel density estimation:*

Kernel density estimation is an effective algorithm for estimating the mutual information between two variables. With a generalized weight or kernel function K(x) the kernel density estimator f ̂(x) is given by,

$$f\hat{}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right) \text{……….(1)}$$

The parameter h is called smoothing parameter or window width and the kernel function K(x) is required to be a (normalized) probability density. It may be easily verified that in this case f ̂ itself is a probability density. Further, f ̂ will inherit all the continuity and differentiability properties of the kernel K. For simplicity we focus on the Gaussian kernel. The density estimate then reads:

$$f\hat{}(x) = \frac{1}{N}\frac{1}{h\sqrt{2\pi}}\sum_{i=1}^{N}\exp\left(-\frac{(x-x_i)^2}{2h^2}\right) \dots\dots\dots\dots (2)$$

C.2. *Estimating the mutual information*:

The mutual information I (X, Y) is a functional of probability densities. Thus an obvious way to find an estimate for I (X, Y) is to find estimates of the densities and then to substitute these into the required integral. By kernel estimation we obtain probability densities only. Remarkably enough, the discretization of the (x, y)-plane into infinitesimal bins of size ΔV = Δx Δy corresponds to the continuous form of the mutual information.

$$\hat{I}(X,Y) = \int_x \int_y \hat{f}\,(x,y)\log\frac{\hat{f}(x,y)}{\hat{f}(x)\hat{f}(y)}\,dx\,dy \quad \text{……(3)}$$

To evaluate Equation , we have to integrate over a smooth function. The choice of the integration steps Δx and Δy could thus be based entirely on standard procedures for numerical integration.

III. Results and Discussion

A. Testing

Here we used the partial dataset extracted from a yeast cell cycle dataset generated by Spellman et al 1998 [14]. The dataset has 500 genes at 12 different conditions. Modified EMD algorithm is used to remove the noise [10]. Then FCM method is combined with this modified EMD algorithim. To calculate 'distance' or 'similarity' between measured datasets we used both linear (Euclidean distance) and nonlinear (Mutual Information) measure of dependencies. After getting the clustering results with EMD and without EMD for both linear and nonlinear dependency measure, KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis is done to validate the clustering results. Here KEGG pathway enrichment analysis is done by using a software called EXPANDER [15,16]. The m value used in the clustering algorithm is 1.3 and the number of clusters is 10 in linear measure and 8 in nonlinear measure. This increases the iteration time, For the present problem the best value for α found by trial and error method is **0.6.**

B.Enrichment Analysis

B.1. Results for linear measure [without EMD]

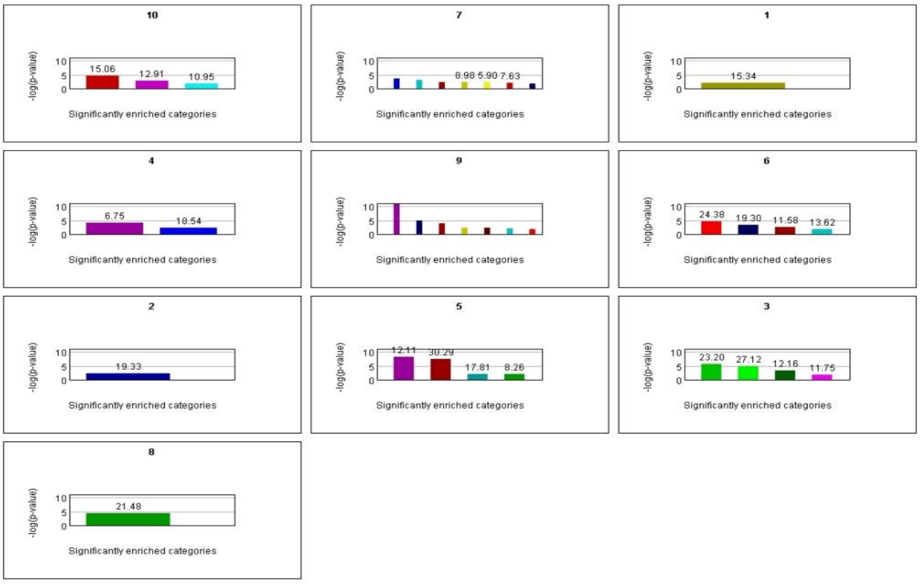
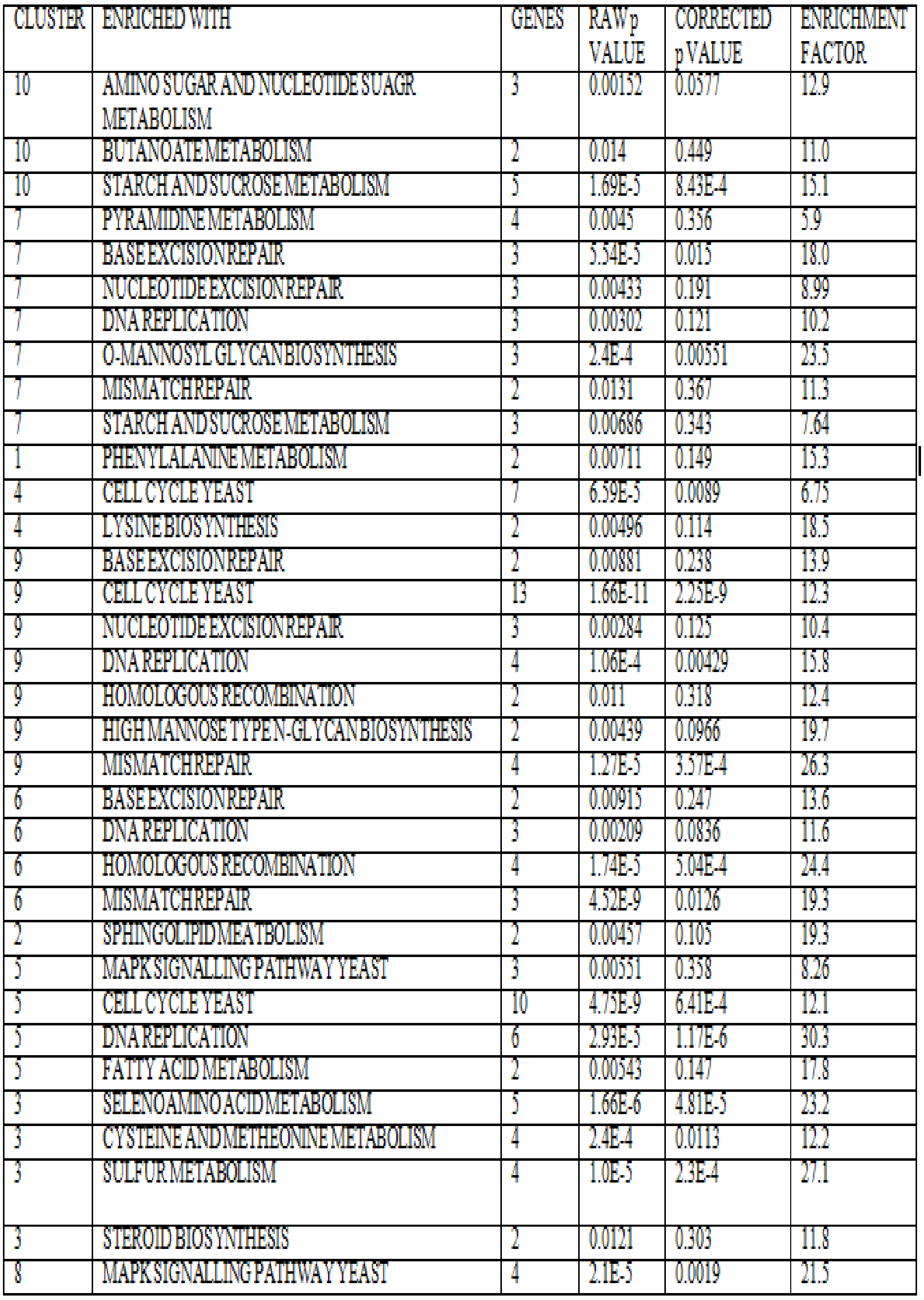


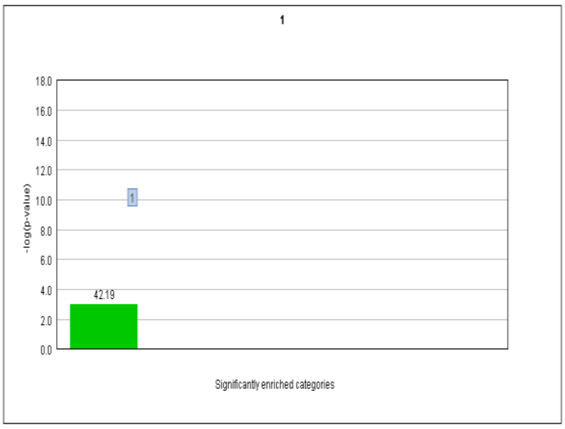
Fig. 1 Enriched pathways for different clusters [without EMD]

TABLE I: Pathway Enrichment table before EMD



CLUSTER	ENRICHED WITH	GENES	RAW p VALUE	CORRECTED p VALUE	ENRICHMENT FACTOR
10	AMINO SUGAR AND NUCLEOTIDE SUAGR METABOLISM	3	0.00152	0.0577	12.9
10	BUTANOATE METABOLISM	2	0.014	0.449	11.0
10	STARCH AND SUCROSE METABOLISM	5	1.69E-5	8.43E-4	15.1
7	PYRAMIDINE METABOLISM	4	0.0045	0.356	5.9
7	BASE EXCISION REPAIR	3	5.54E-5	0.015	18.0
7	NUCLEOTIDE EXCISION REPAIR	3	0.00433	0.191	8.99
7	DNA REPLICATION	3	0.00302	0.121	10.2
7	O-MANNOSYL GLYCAN BIOSYNTHESIS	3	2.4E-4	0.00551	23.5
7	MISMATCH REPAIR	2	0.0131	0.367	11.3
7	STARCH AND SUCROSE METABOLISM	3	0.00686	0.343	7.64
1	PHENYLALANINE METABOLISM	2	0.00711	0.149	15.3
4	CELL CYCLE YEAST	7	6.59E-5	0.0089	6.75
4	LYSINE BIOSYNTHESIS	2	0.00496	0.114	18.5
9	BASE EXCISION REPAIR	2	0.00881	0.238	13.9
9	CELL CYCLE YEAST	13	1.66E-11	2.25E-9	12.3
9	NUCLEOTIDE EXCISION REPAIR	3	0.00284	0.125	10.4
9	DNA REPLICATION	4	1.06E-4	0.00429	15.8
9	HOMOLOGOUS RECOMBINATION	2	0.011	0.318	12.4
9	HIGH MANNOSE TYPE N-GLYCAN BIOSYNTHESIS	2	0.00439	0.0966	19.7
9	MISMATCH REPAIR	4	1.27E-5	3.57E-4	26.3
6	BASE EXCISION REPAIR	2	0.00915	0.247	13.6
6	DNA REPLICATION	3	0.00209	0.0836	11.6
6	HOMOLOGOUS RECOMBINATION	4	1.74E-5	5.04E-4	24.4
6	MISMATCH REPAIR	3	4.52E-9	0.0126	19.3
2	SPHINGOLIPID MEATBOLISM	2	0.00457	0.105	19.3
5	MAPK SIGNALLING PATHWAY YEAST	3	0.00551	0.358	8.26
5	CELL CYCLE YEAST	10	4.75E-9	6.41E-4	12.1
5	DNA REPLICATION	6	2.93E-5	1.17E-6	30.3
5	FATTY ACID METABOLISM	2	0.00543	0.147	17.8
3	SELENOAMINO ACID METABOLISM	5	1.66E-6	4.81E-5	23.2
3	CYSTEINE AND METHEONINE METABOLISM	4	2.4E-4	0.0113	12.2
3	SULFUR METABOLISM	4	1.0E-5	2.3E-4	27.1
3	STEROID BIOSYNTHESIS	2	0.0121	0.303	11.8
8	MAPK SIGNALLING PATHWAY YEAST	4	2.1E-5	0.0019	21.5

B.2 Results for linear measure [with EMD]



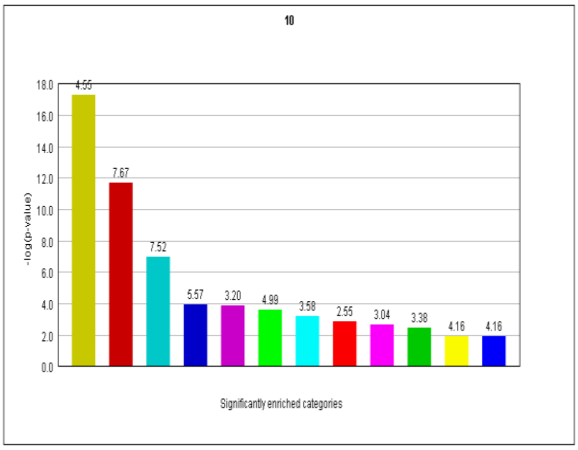
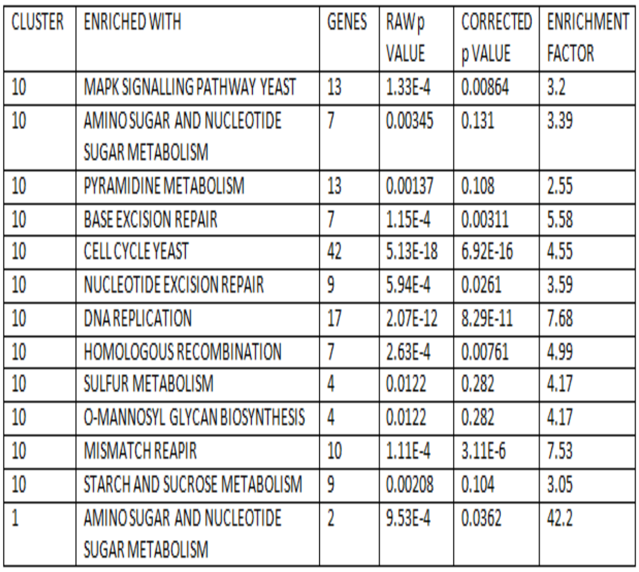


Fig. 2 Enriched pathways for different clusters [with EMD]

TABLE II: Pathway Enrichment table after EMD



CLUSTER	ENRICHED WITH	GENES	RAW p VALUE	CORRECTED p VALUE	ENRICHMENT FACTOR
10	MAPK SIGNALLING PATHWAY YEAST	13	1.33E-4	0.00864	3.2
10	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	7	0.00345	0.131	3.39
10	PYRAMIDINE METABOLISM	13	0.00137	0.108	2.55
10	BASE EXCISION REPAIR	7	1.15E-4	0.00311	5.58
10	CELL CYCLE YEAST	42	5.13E-18	6.92E-16	4.55
10	NUCLEOTIDE EXCISION REPAIR	9	5.94E-4	0.0261	3.59
10	DNA REPLICATION	17	2.07E-12	8.29E-11	7.68
10	HOMOLOGOUS RECOMBINATION	7	2.63E-4	0.00761	4.99
10	SULFUR METABOLISM	4	0.0122	0.282	4.17
10	O-MANNOSYL GLYCAN BIOSYNTHESIS	4	0.0122	0.282	4.17
10	MISMATCH REAPIR	10	1.11E-4	3.11E-6	7.53
10	STARCH AND SUCROSE METABOLISM	9	0.00208	0.104	3.05
1	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	2	9.53E-4	0.0362	42.2

B.3 Results for nonlinear measure [without EMD]

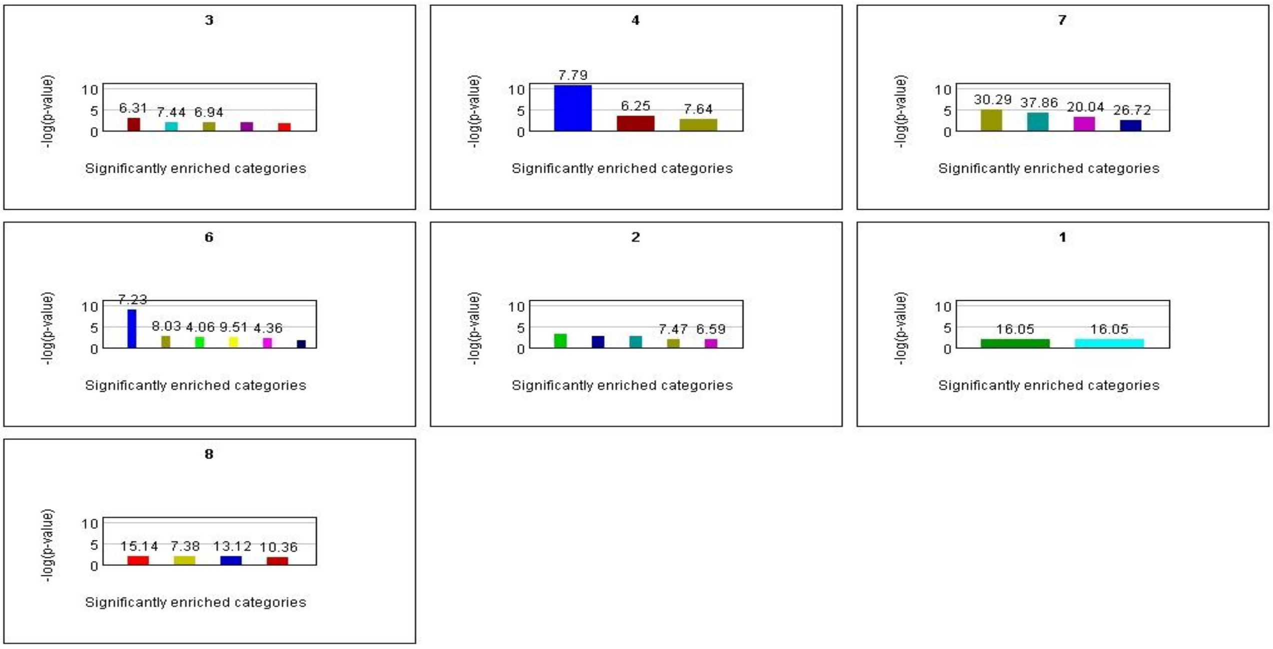
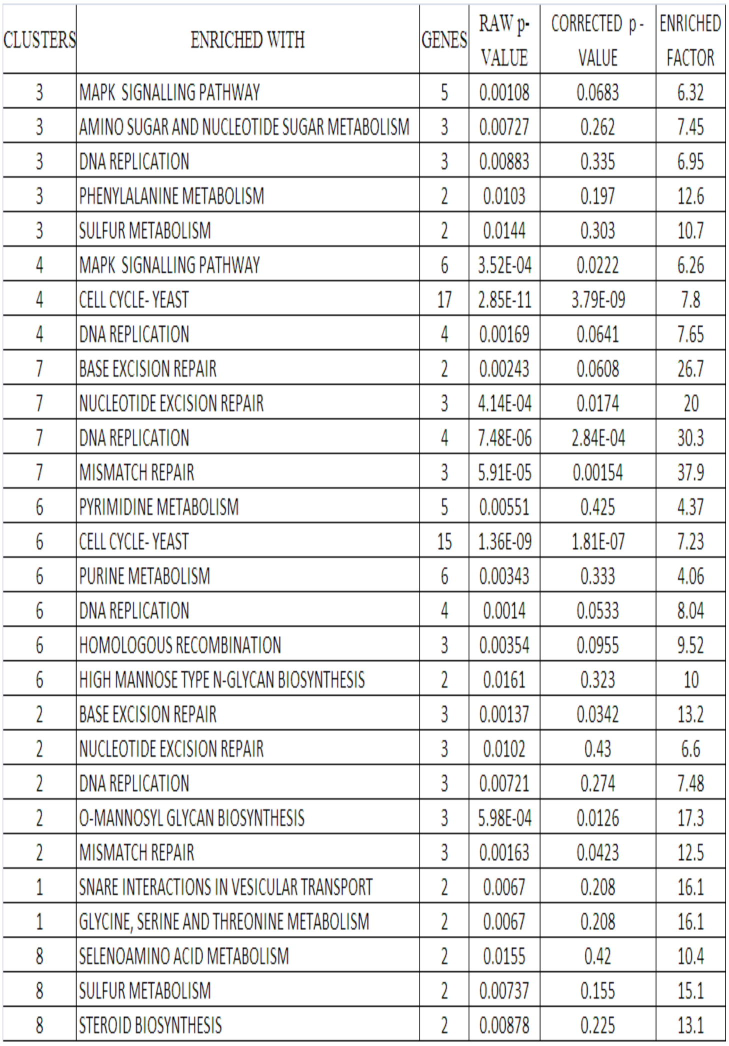


Fig. 3 Enriched pathways for different clusters [without EMD]

TABLE III: Pathway Enrichment table before EMD



CLUSTERS	ENRICHED WITH	GENES	RAW p-VALUE	CORRECTED p-VALUE	ENRICHED FACTOR
3	MAPK SIGNALLING PATHWAY	5	0.00108	0.0683	6.32
3	AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	3	0.00727	0.262	7.45
3	DNA REPLICATION	3	0.00883	0.335	6.95
3	PHENYLALANINE METABOLISM	2	0.0103	0.197	12.6
3	SULFUR METABOLISM	2	0.0144	0.303	10.7
4	MAPK SIGNALLING PATHWAY	6	3.52E-04	0.0222	6.26
4	CELL CYCLE- YEAST	17	2.85E-11	3.79E-09	7.8
4	DNA REPLICATION	4	0.00169	0.0641	7.65
7	BASE EXCISION REPAIR	2	0.00243	0.0608	26.7
7	NUCLEOTIDE EXCISION REPAIR	3	4.14E-04	0.0174	20
7	DNA REPLICATION	4	7.48E-06	2.84E-04	30.3
7	MISMATCH REPAIR	3	5.91E-05	0.00154	37.9
6	PYRIMIDINE METABOLISM	5	0.00551	0.425	4.37
6	CELL CYCLE- YEAST	15	1.36E-09	1.81E-07	7.23
6	PURINE METABOLISM	6	0.00343	0.333	4.06
6	DNA REPLICATION	4	0.0014	0.0533	8.04
6	HOMOLOGOUS RECOMBINATION	3	0.00354	0.0955	9.52
6	HIGH MANNOSE TYPE N-GLYCAN BIOSYNTHESIS	2	0.0161	0.323	10
2	BASE EXCISION REPAIR	3	0.00137	0.0342	13.2
2	NUCLEOTIDE EXCISION REPAIR	3	0.0102	0.43	6.6
2	DNA REPLICATION	3	0.00721	0.274	7.48
2	O-MANNOSYL GLYCAN BIOSYNTHESIS	3	5.98E-04	0.0126	17.3
2	MISMATCH REPAIR	3	0.00163	0.0423	12.5
1	SNARE INTERACTIONS IN VESICULAR TRANSPORT	2	0.0067	0.208	16.1
1	GLYCINE, SERINE AND THREONINE METABOLISM	2	0.0067	0.208	16.1
8	SELENOAMINO ACID METABOLISM	2	0.0155	0.42	10.4
8	SULFUR METABOLISM	2	0.00737	0.155	15.1
8	STEROID BIOSYNTHESIS	2	0.00878	0.225	13.1

B.4 Results for nonlinear measure [with EMD]

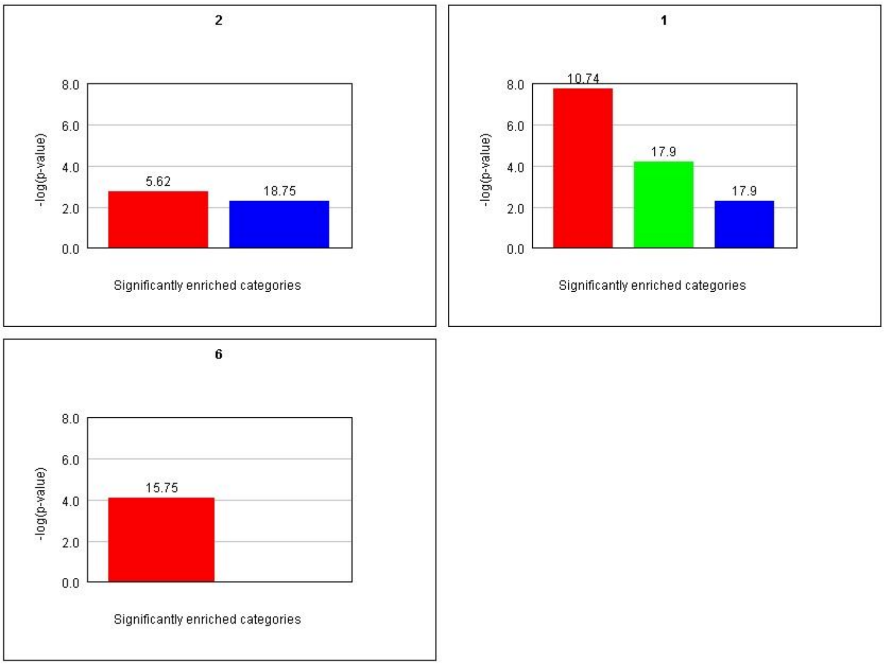
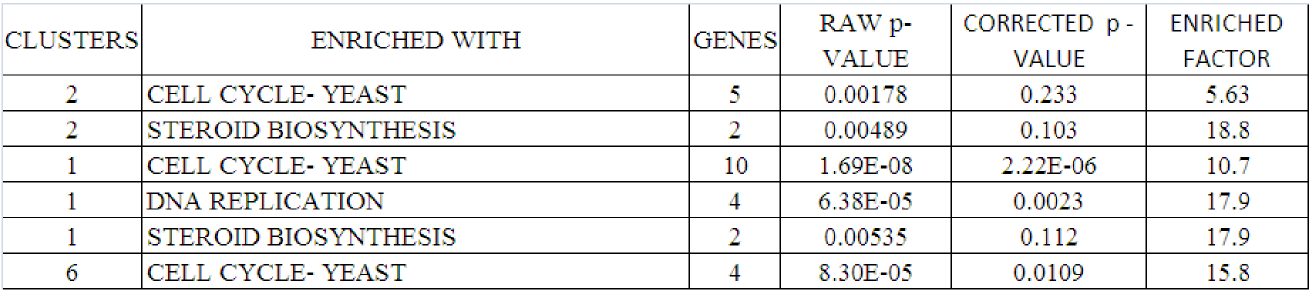


Fig. 4 Enriched pathways for different clusters [with EMD]

TABLE IV: Pathway Enrichment table after EMD

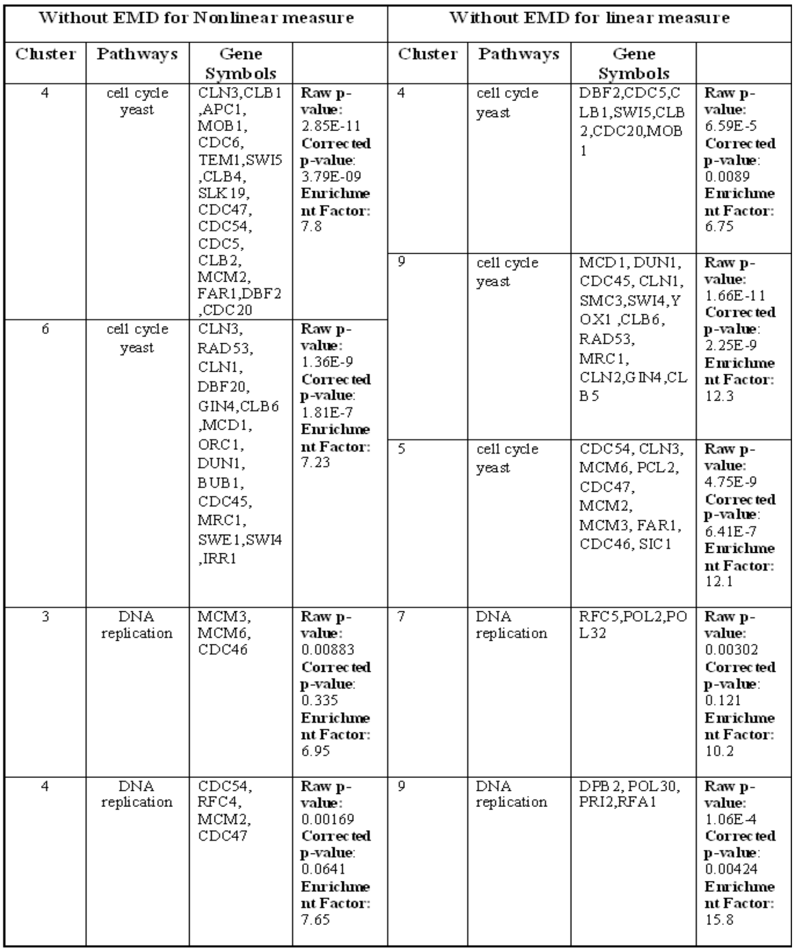


CLUSTERS	ENRICHED WITH	GENES	RAW p-VALUE	CORRECTED p -VALUE	ENRICHED FACTOR
2	CELL CYCLE- YEAST	5	0.00178	0.233	5.63
2	STEROID BIOSYNTHESIS	2	0.00489	0.103	18.8
1	CELL CYCLE- YEAST	10	1.69E-08	2.22E-06	10.7
1	DNA REPLICATION	4	6.38E-05	0.0023	17.9
1	STEROID BIOSYNTHESIS	2	0.00535	0.112	17.9
6	CELL CYCLE- YEAST	4	8.30E-05	0.0109	15.8

B. Discussion

Higher the number of genes associated to a pathway better is the enrichment of that pathway. Secondly, a lower value of corrected ‘p’ value ensures a much better enrichment. From the tables I and II (pathway enrichment table without EMD and pathway enrichment table with EMD) it can be observed that the number of genes that are associated to the pathway Cell cycle yeast is 42 in cluster 10 in table II (i.e with EMD) with a corrected p value of 6.92E-16 where it is only 13 in cluster 9 with a corrected p value 2.25E-9,10 in cluster 5 with a corrected p value of 6.41E-7 and 7 in cluster 4 with a corrected p value of 0.0089 in table I (i.e without EMD).This shows an excellent enrichment in the pathway cell cycle yeast after EMD. From the table I and table II we can say that other pathways are excellently enriched. But in nonlinear case we are getting better result in case of without EMD. There is very small decrease in p-value in case of ‘with EMD’. For DNA replication p-value is 0.335 [without EMD] and 0.0023[with EMD]. Only in case of Steroid Biosynthesis no of genes are increasing in ‘with EMD’ case. Many pathways are not enriched in case of ‘with EMD’ as we can see from KEGG pathway analysis [Table III and Table IV]. As we are de-noising the data, linear dependencies between the genes are increasing rather than nonlinear dependencies. Therefore we are getting better clustering results in linear measure for both with EMD and without EMD rather than nonlinear measure for ‘with EMD’ and ‘without EMD’. A comparative study between linear and nonlinear measure [without EMD] are given below

TABLE V: Comparative Study for linear & nonlinear[without EMD]



Without EMD for Nonlinear measure				Without EMD for linear measure			
Cluster	Pathways	Gene Symbols		Cluster	Pathways	Gene Symbols	
4	cell cycle yeast	CLN3,CLB1,APC1, MOB1, CDC6, TEM1,SWI5,CLB4, SLK19, CDC47, CDC54, CDC5, CLB2, MCM2, FAR1,DBF2,CDC20	**Raw p-value:** 2.85E-11 **Corrected p-value**: 3.79E-09 **Enrichment Factor:** 7.8	4	cell cycle yeast	DBF2,CDC5,CLB1,SWI5,CLB2,CDC20,MOB1	**Raw p-value:** 6.59E-5 **Corrected p-value**: 0.0089 **Enrichment Factor:** 6.75
6	cell cycle yeast	CLN3, RAD53, CLN1, DBF20, GIN4,CLB6,MCD1, ORC1, DUN1, BUB1, CDC45, MRC1, SWE1,SWI4,IRR1	**Raw p-value:** 1.36E-9 **Corrected p-value**: 1.81E-7 **Enrichment Factor:** 7.23	9	cell cycle yeast	MCD1, DUN1, CDC45, CLN1, SMC3,SWI4,YOX1 ,CLB6, RAD53, MRC1, CLN2,GIN4,CLB5	**Raw p-value:** 1.66E-11 **Corrected p-value**: 2.25E-9 **Enrichment Factor:** 12.3
				5	cell cycle yeast	CDC54, CLN3, MCM6, PCL2, CDC47, MCM2, MCM3, FAR1, CDC46, SIC1	**Raw p-value:** 4.75E-9 **Corrected p-value**: 6.41E-7 **Enrichment Factor:** 12.1
3	DNA replication	MCM3, MCM6, CDC46	**Raw p-value:** 0.00883 **Corrected p-value**: 0.335 **Enrichment Factor:** 6.95	7	DNA replication	RFC5,POL2,POL32	**Raw p-value:** 0.00302 **Corrected p-value**: 0.121 **Enrichment Factor:** 10.2
4	DNA replication	CDC54, RFC4, MCM2, CDC47	**Raw p-value:** 0.00169 **Corrected p-value**: 0.0641 **Enrichment Factor:** 7.65	9	DNA replication	DPB2, POL30, PRI2,RFA1	**Raw p-value:** 1.06E-4 **Corrected p-value**: 0.00424 **Enrichment Factor:** 15.8

From table V we can see that for cell cycle yeast, we are getting better corrected p-value (3.79E-09) for without EMD in nonlinear case. It decreased than that of the linear case (0.0089). The group of genes which are related to cluster 4 in nonlinear case (without EMD) are not associated with the cluster 4 in linear case (without EMD). And the no of genes are also greater in nonlinear case (without EMD). So some extent we are getting better enrichment in the pathway cell cycle yeast.

IV. CONCLUSIONS

Fuzzy C Means clustering algorithm which incorporates the concept of Emperical Mode Decomposition for noise removal is explained. We introduced the EMD method here to remove noise in microarray data [1] . However, the number of times for this noise removal is still uncertain. When the signal becomes smooth, we consider noise has been removed, but this may not be sufficiently precise. Another problem is that the more times we denoise the more information we would lose in microarray data. So determination of the number of times of denoising is still a problem.

Clustering results of raw data and of denoised data are validated using KEGG pathway enrichment studies and the studies propose that the clustering results of denoised data are better when compared to the clustering results of raw data for linear measure. Clustering results of denoised data after EMD are not so better than that of the clustering results of raw data in non linear measure. As we are denoising the data, linear dependencies between the genes are increasing rather than nonlinear dependencies. Therefore we are getting better clustering results in linear measure for both with EMD and without EMD.

REFERENCES

[1] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, *Cluster analysis and display of genome-wid expression patterns. Proc. Natl Acad. Sci.USA*, 95:14863-14868, 1998.

[2] *"Analysis of microarray gene expression data",* by Wolfgang Huber, Anja von Heydebreck, Martin Vingron, April 2, 2003

[3] Sergio Ortiz-Gama,L. Enrique Sucar ,Andrés F. Rodríguez, *"Clustering Gene Expression Data: an Experimental Analysis*", Proceedings of the Fifth Mexican International Conference in Computer Science (ENC'04), 0-7695-2160-6/04, 2004 IEEE

[4] Daxin Jiang, Chun Tang, and Aidong Zhang*" Cluster Analysis for Gene Expression Data: A Survey",* IEEE transactions on knowledge and data engineering, vol. 16, no. 11, november 2004

[5] L. Fu and E. Medico, FLAME: a novel *fuzzy clustering method for the analysis of DNA microarray data, BMC Bioinformatics*, 8:3, 2007.

[6] " *Fuzzy C-means method with empirical mode decomposition for clustering microarray data" by Yan-Fei Wang, Zu-Guo Yu and Vo Anh, 978-1-4244-8305, 2010 IEEE 192 ,International Conference on Bioinformatics and Biomedicine*

[7] "*Evaluation of Gene-Expression Clustering Via Mutual Information distance measure",*by Ido Priness**,** Oded Maimon and Irad Ben-Gal **,** *BMC Bioinformatics*2007,**8**:111 doi:10.1186/1471-2105-8-111, The electronic version of this article is the complete one and can be found online at: http://www.biomedcentral.com/1471-2105/8/111

[8] The mutual information*: Detecting and evaluating dependencies between variables* R. Steuer , J. Kurths , C. O. Daub, J. Weise and J. Selbig,*Received on April 8, 2002; accepted on June 15, 2002, Vol. 18 Suppl. 2 2002 Pages S231–S240*

[9] Moon,Y., Rajagopalan,B. and Lall,U. (1995*) Estimation of mutual information using kernel density estimators*. *Phys. Rev. E*, **52**, 2318–2321.

[10] *A.O. Boudraa, J.C. Cexus, and Z. Saidi" EMD-Based Signal Noise Reduction", International Journal of Information and Communication Engineering 1:1 2005*

[11] *On the HHT, its problems, and some solutions by* R.T**.** Rato, M.D. Ortigueira,1, A.G. Batista Campus da FCT da UNL, Quinta da Torre, 2825 - 114 Monte da Caparica, Portugal Received 15 December 2006; accepted 30 November 2007

[12] N.E. Huang, Z. Shen, S.R. Long, M.L. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung, H.H. Liu, *The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis*, Proceedings of the Royal Society of London A 454 (1998) 903–995.

[13] " *Fuzzy C-means method with empirical mode decomposition for clustering microarray data"* by Yan-Fei Wang, Zu-Guo Yu and Vo Anh, 978-1-4244-8305, 2010 IEEE 192 ,International Conference on Bioinformatics and Biomedicine

[14] Spellman, P. T., Sherlock, G., et al.*" Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* "Mol Biol Cell 9(12), 3273-97, 1998.

[15] Ron Shamir, Adi Maron-Katz, Amos Tanay, Chaim Linhart, Israel Steinfeld, Roded Sharan, Yosef Shiloh and Ran Elkon"*EXPANDER – an integrative program suite for microarray data Analysis*", BMC Bioinformatics 2005, 6:232 doi:10.1186/1471-2105-6-232

[16] Roded Sharan, Adi Maron-Katz and Ron Shamir" *CLICK and EXPANDER: a system for clustering and visualizing gene expression data",* Vol. 19 no. 14 2003, pages 1787–1799 DOI: 10.1093/bioinformatics/btg232, Received on October 30, 2002; revised on January 28, 2003; accepted on March 28, 2003