A novel fuzzy and multiobjective evolutionary algorithm based gene assignment for clustering short time series expression data

Ashish Anand, P.N. Suganthan and Kalyanmoy Deb

***Abstract*— Conventional clustering algorithms based on Euclidean distance or Pearson correlation coefficient are not able to include order information in the distance metric and also unable to distinguish between random and real biological patterns. We present template based clustering algorithm for time series gene expression data. Template profiles are defined based on up-down regulation of genes between consecutive time points. Assignment of genes to templates is based on fuzzy membership function. Multi-objective evolutionary algorithm is used to determine compact clusters with varying number of templates. Statistical significance of each template is determined using permutation based non-parametric test. Statistically significant profiles are further tested for their biological relevance using gene ontology analysis. The algorithm was able to distinguish between real and noisy pattern when tested on artificial and real biological data. The proposed algorithm has shown better or similar performance compared to STEM and better than k-means on a real biological data.**

I. Introduction

The advent of microarray technology has made it possible to explore the dynamics of transcription on genome-wide scale in single experiment. Data from microarray experiments have provided an opportunity to understand the genomic level mechanism, i.e., relationships between genes under the particular experimental condition. On the other hand microarray technology has also generated many challenging computational problems. Some of the problems are of common nature, irrespective of the experiment design, while some other problems are specific to particular experiments. Microarray experiments can be classified into static and dynamic based on the nature of experiments [1]. In static experiments, expression of genes are measured in different conditions and analyzed for differentially expressed genes under those conditions. Examples of static microarray experiments include knock-out vs wild type studies, mutant (or treatment) vs wild type (or normal). Whereas, in dynamic experiments, gene expressions are measured in a particular order, e.g. at different time points or at different dose levels, under a certain condition. Aims of these studies are to understand the dynamic nature of genes, which might give insights into regulatory networks, transcriptional controls and other various biological phenomena. In this paper, only time series or temporal experiments are mentioned but, unless otherwise stated, the same is true with dose-response or any other order-restricted experiments. Apart from differences in experiment design and motivation of the two types of experiments, there exists differences at data analysis level. While static data can be assumed to be independent and identically distributed (iid), time series data cannot be assumed to be iid. Rather order of the data is important and it exhibits a strong auto-correlation between successive time points.

In recent years, time series microarray experiments have been performed to understand the various biological phenomena. Examples include yeast cell cycle study [2], yeast sporulation study [3], developmental studies [4], immune response to *Helicobacter pylori* infection [5] and temporal profiling during neurogenesis [6].

Based on the assumption that genes with similar expression profiles are functionally related or co-regulated [7], most of the methods proposed in the gene expression analysis literature attempt to identify groups of genes with similar expression profiles. Much of the early work used methods developed originally for static data [2], [3], [5]. Most commonly used among such methods are hierarchical clustering [7], k-means [8], self-organizing maps [9].

Evolutionary algorithms (EAs) [10] have been used for clustering [11], [12], [13], [14] mainly in pattern recognition domain.[1] In [11], Murthy and Chowdhury used EA to optimize intra cluster variation. But the user needs to predefine number of clusters and also the string representation for chromosome is not suitable for cases where number of samples in data is quite large as in the case of gene expression data. Bandyopadhyay and Maulik [12] proposed EA based clustering in the context of image classification to optimize Davies-Bouldin [12] index. Though they do not need to specify exact number of cluster but chromosome representation again may not be suitable for gene expression data. Each chromosome was of length K_{max} and made up of real numbers representing co-ordinates of centers and "don't care" symbol. The "don't care" symbol represented absence of the center and helped EA to search for varying number of clusters. Obviously for large dimensional data this representation is not appropriate. Gesu et al. [13] used EA for clustering of gene expression data. Again EA was used to optimize the intra cluster variation. GenClust [13] also required to predefine the actual number of clusters. Handl and Knowles [14] formulated clustering problem as an multi-

Ashish Anand is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Sinagpore (email: anand.ashish@pmail.ntu.edu.sg).

P. N. Suganthan is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Sinagpore (email: epnsugan@ntu.edu.sg).

Kalyanmoy Deb is with the Kanpur Genetic Algorithms Lab, Department of Mechanical Engineering, Indian Institute of Technology, Kanpur, India (email: deb@iitk.ac.in).

[1]Due to space limitations we have not discussed basics of EAs. Interested readers can look at [15], [10] as good reference point for the same.

1-4244-1340-0/07$25.00 ©2007 IEEE

objective problem [15] and proposed an algorithm named MOCK. They used compactness and connectedness as two criteria to optimize.

Although all the above discussed generic methods lead to many biologically significant results, they are not designed for time series or order-restricted data and hence overlooking characteristics of these data. All the above algorithms group genes based on some distance measures (Euclidean and correlation based distances are more common) and only look for compact clusters. These measures ignore the sequential nature of expression data and assume that data at each time point is independent of each other.

Differences in the nature of static and temporal data led to the development of several new algorithms specifically targeting time series experiments. Majority of these methods are model based clustering methods. Gaussian or other distribution based mixture modeling [16], hidden Markov model (HMM) [17], spline [18], [19] and auto-regressive model [20] are examples of some of the different model-based clustering methods.

Schliep et al. [17], authors presented HMM based clustering algorithm for time series expression data. Given gene expression data, goal is to partition them into K HMMs. Assignment of genes to different HMMs and parameter estimation of HMMs were performed using expectation maximization (EM) approach. Ji et al. [21] also used HMM based model to cluster the gene expression data. But all HMM based methods require larger number of time points than the number of states. This makes these methods more appropriate for large time series expression data.

There are several methods proposed where spline based modeling was used for analyzing time series expression profiles. Bar-Joseph et al. [18] used cubic spline to represent the continuous nature of temporal data. Luan and Li [19] proposed mixed-effects models using B-splines. Spline based methods require us to define a fixed number of knots between first and last time points to approximate the expression profile over time points. Even use of few knots require estimation of several parameters for each gene. This may result in overfitting of data if there are very small number of time points. Thus, this method is also not suitable for small time series. Authors in [22] suggested to fit linear splines for short time series expression data. But their method requires several replicates [22].

Ramoni et al. [20] used a model-based clustering approach, where auto-regression was used as the model. The method represented time-series data as auto-regressive equations of fixed order and used agglomerative procedure to search for the most probable set of clusters given the data. This method is quite appropriate for long temporal data but has tendency to overfit and to give poorly separated clusters for short time series data [23].

These algorithms work well for long temporal gene expression profiles, but they are not suitable for short temporal expression profiles. Short time series data are more prevalent [23] in microarray experiments due to many reasons. Cost of arrays and limited biological samples are the two most common limiting factors. Also, some of the above-mentioned methods require several replicates, which again may not be possible due to the above-discussed reasons. Motivated by this, many new algorithms are being proposed for clustering of short time series gene expression data [23], [24], [25], [26].

Though methods based on regression [26] or model-based methods [24] have been proposed but template based methods [23], [25], [27]) have dominated clustering approach for short time series expression data. In [27], authors have used sinusoid to identify cycling yeast genes. This method required the prior knowledge of shape of the curve to be fitted. In general, such a priori information is not available. Moller-Levet et al. [25] suggested a method where each gene is converted into a pattern vector. Pattern vector corresponding to each gene was then assigned to one of the predefined cluster prototype. One problem with this approach is that the conversion of original gene expression data resulted in information loss. Ernst et al. [23] predefined profiles based on change in expression levels units between consecutive time points. Their algorithm first finds a set of representative model profiles from the set of all possible profiles. Selected model templates were quite independent to data. Genes were then assigned to one of the model profiles. Statistical significance of each profile were determined and only significant profiles were selected for further analysis. These profiles can be further grouped into larger clusters.

In this paper, we propose novel template based clustering algorithms for time series gene expression data. A novel method of gene assignment to templates is proposed. Multi-objective evolutionary algorithm (MOEA) is used to get optimal number of model templates minimizing the quantization error. Statistical significance of each selected template is determined and only significant templates are further tested for biological significance. Motivation of the proposed method can be summarized as:

- An algorithm giving importance to order of data
- An algorithm giving importance to shape rather than distance or correlation measure
- An algorithm able to distinguish between random and real pattern
- An algorithm which does not require many replicates as basic requirement.

II. Approach

In this paper, we have proposed a new approach for identifying significant profiles among the set of all possible template profiles. Instead of choosing any distance measure, genes are assigned to profiles based on its membership values calculated using fuzzy membership functions. The fuzzy membership function is defined on the basis of fold change significance. MOEA is used to get a set of trade-off solutions minimizing quantization error with simultaneous minimization of number of model templates. As many of such profiles could be enriched just by random chance, we

applied permutation based statistical significance test [23] on each enriched profiles. Only significant profiles were further analyzed using Gene Ontology (GO) annotations to interpret biological information.

A. Template Profiles

Template profiles or pattern vectors are defined based on change in gene expression levels at consecutive time points. A gene can have either positive change in expression level or negative change or there may not be any change in expression level at all. Hence, we have three possible transition states for each gene from one time point to next time point. We denoted positive change as 1, negative change as 2 and no change as 0 for defining profiles. For n_t time points, total number of distinct templates can be given by $3^{(n_t-1)}$ and each template profile can be represented by $(n_t - 1)$ tuples. For illustration purpose, we have shown a profile '0110' in Figure 1. As it is shown in Figure 1, there is significant positive change in gene expression level from

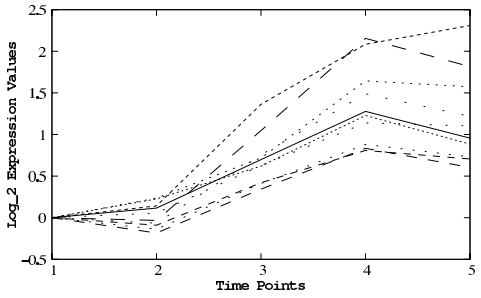


Fig. 1. Example Profile '0110'.

second to third time point and from third to fourth time point. There is almost no change in expression levels from first to second and fourth to fifth. It can be observed that it is difficult to put exact boundary between different transition states. We defined fuzzy membership function to measure belongingness of change in expression levels to different transition states.

B. Fuzzy Membership Function

Here we assume that data is already normalized and transformed into log ratios. From the normalized gene expression matrix (**GEM**) we obtain $n_g \times (n_t - 1)$ matrix. Entries in the matrix corresponds to difference of gene expression values at successive time points instead of actual gene expression values. n_g is number of genes being considered for cluster analysis. For simplicity, we call this matrix *difference gene expression matrix* (**DGEM**).

Entries of gene g_i in DGEM is represented as (dg_1^i, dg_2^i, ..., dg_k^i, ..., $dg_{n_t-1}^i$) and profile p_j as (p_1^j, p_2^j, ..., p_k^j, ..., $p_{n_t-1}^j$), where $dg_k^i \in [-1:1]$ and $p_k^j \in \{1, 0, 2\}$.

Membership $mf(dg_k^i, p_k^j)$ of dg_k^i is calculated with respect to p_k^j using mf defined in equ 1.

$$mf(dg_k^i, p_k^j) = \begin{cases} 1.0/(1.0 + exp(-a * dg_k^i + b) & \text{if } p_k^j = 1, \\ 1.0/(1.0 + exp(a * dg_k^i + b) & \text{if } p_k^j = 2, \\ 1.0/(1.0 + (|dg_k^i/a|)^{2*b} & \text{otherwise.} \end{cases} \quad (1)$$

Overall membership of $m(g_i, p_j)$ of gene g_i with respect to p_j is calculated using equ 2. Gene g_i is then assigned to profile p_k for which $m(g_i, p_.)$ is maximum.

$$m(g_i, p_j) = avg_k mf(dg_k^i, p_k^j) \quad (2)$$

S-shaped and bell-shaped functions are among the mostly used functions for defining membership functions in fuzzy-based methods. We choose s-shaped function to define positive and negative change in expression levels and bell-shaped function for no change. Parameters of membership function are chosen based on fold change significance. Parameters a and b of s-shaped function are related as the ratio $\frac{b}{a}$ represents the value where the function-value reach 0.5. We decided to choose 1.5 fold change either up or down-regulated to represent this value. Parameters of bell-shaped function is chosen such that it gives maximum value at no change in expression level and a value of 0.25 corresponds to fold-change of 1.30. A typical membership function is shown in Figure 2.

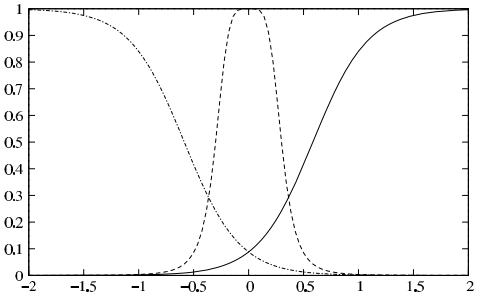


Fig. 2. Membership Function. Values at x-axis corresponds to log2 fold change.

C. Selection of Model Template Profiles

Our algorithm assign genes to template profiles based on fuzzy membership function as discussed in Section II-B. Even for small number of time points 5 or 6, there are quite large number of possible templates. Many of them are likely to be very sparsely populated by genes. Hence, there is need to select meaningful and manageable number of model templates. But knowing such a number a priori is non-trivial. We choose to minimize the quantization error to get compact clusters with simultaneous minimization of number of model templates. By quantization error we mean the overall distance between template and assigned genes to it. We have formally define the quantization error in equation (3). Since, template and gene expressions are already normalized, euclidean distance can be used. We like to emphasize here that genes are not assigned to templates

based on distance matrix, rather they are assigned based on fuzzy membership function. Euclidean distance is used to calculate the quantization error only.

$$quanterr = \sum_{j=1}^{m}(\sum_{i,i\epsilon S_j} (\sum_{k}^{n_t-1} (dg_k^i - p_k^j)^2)^{1/2}) \qquad (3)$$

where m is number of model templates, S_j is set of genes assigned to template j.

As it is obvious that the two chosen objectives are conflicting, we decided to use multi-objective evolutionary algorithms (MOEA) [15] to get set of trade-off solutions. EAs are well-suited for multiobjective optimization as being population based approach, it approximates whole Pareto front in single simulation. Solutions in Pareto front represent trade-off between quantization error and number of model templates.

1) Evolutionary Algorithms: NSGA-II [28] is one of the most popular MOEAs and is used here to get the optimal front. Application of MOEA requires proper choice of

i) an appropriate chromosome representation
ii) two or more objective functions
iii) selection of crossover and mutation operator

These choices are non-trivial and the performance of the algorithm depends largely on them. We discuss below each of them one by one.

2) Chromosome Representation: We choose binary string of length equal to total template profiles considered. '1' at particular position 'j' signifies the presence of template profile of index 'j' as one of chosen model templates. Number of '1s' is kept variable for each chromosome so that variable number of templates can be considered. Also, as we are looking for model profiles, we do not want very small number of template profiles. Thus we make at least K_{min} bits on in each chromosome.

3) Objective Functions: Two objective functions are considered. First objective function is chosen to minimize the number of model templates. Minimizing the quantization error (equation (3)) is chosen as the second objective function.

4) Genetic Operators: A simple representation of chromosome makes it easy to use any of the standard crossover operators (single point, two point, uniform). We use single point crossover operator. Standard bit wise mutation operator is used. Since we want to have at least K_{min} bits on in each chromosome, but their is high chance that crossover and/or mutation operator may lead to less number of bits on in chromosome. To take care of that we employ following repair mechanism. Repair operator generates a random number K_{rand} between $(K_{min} - K)$ and $(K_{max} - K)$. Here, K is the current number of '1s' in chromosome and K_{max} is total template profiles considered. Repair operator randomly makes K_{rand} bits on, which were not already on.

D. Statistical Significance

As genes are assigned to different template profiles only on the basis of membership function, many templates can be expected to be enriched by random chance. Selecting profiles only on the basis of number of genes in them definitely would lead to many non-significant profiles. As determining the underlying distribution of test statistics is difficult, non parametric based permutation test is commonly used in gene expression data analysis [23], [29], [30]. We used the same permutation based test as discussed in [23].

Basic assumption of our analysis of temporal expression data is that gene expression at a particular time point is dependent on other time points. Based on this assumptions, authors in [23], defined null hypothesis as the probability of observing a value at certain time point was independent to other time points. Thus if an enriched profile showing significant deviation from the null hypothesis and assigned more genes than expected by random chance, we expect it should also be biologically relevant. Permutation was used to quantify the expected number of genes that would have been assigned to each profile, if values at each time point were generated independent to others.

We briefly described the exact procedure for permutation based test.

- For n_t time points, get $n_t!$ permutation of data
- For each possible permutation j,
  - Assign genes to a profile as discussed in section II-B.
  - Calculate s_i^j as number of genes assigned to profile i.
- Calculate total number of genes assigned to profile i in all permutations as $S_i = \sum_j s_i^j$.
- Calculate expected number of genes assigned to profile i if all values were generated at random as $E_i = S_i/n!$.
- Since each gene is assigned to one profile only, it can be assumed that each profile is binomially distributed with $n = n_g$ and $p = E_i/m$, where m is total number of enriched profiles.

Thus if s_i genes were assigned to profile i, p-value was calculated as $p(X \geq s_i)$, where, $X \sim Bin(n_g, p)$. Since we were testing many profiles for significance, Bonferroni multiple test correction [31] was applied.

Schematic diagram of the complete approach is shown in Figure 3.

III. Evaluation

Evaluation is performed on both simulated and biological data. We have compared the proposed algorithm to both, a general clustering algorithm (k-means) and an algorithm developed specifically for time series gene expression data (STEM [23]). Results obtained for simulated data is discussed in section III-B and section III-C discusses results obtained on real biological data. Section III-C also discusses the results obtained by STEM[23] and K-means. In both cases, simulated or real, data is filtered using STEM [32] (STEM algorithm was implemented in a software also named STEM) with default values before applying any algorithm.

A. Experimental Setup

We used $a = 4$ and $b = a * 0.585$ for s-shaped function and $a = 0.30$, $b = 2.33$ were used for bell-shaped function

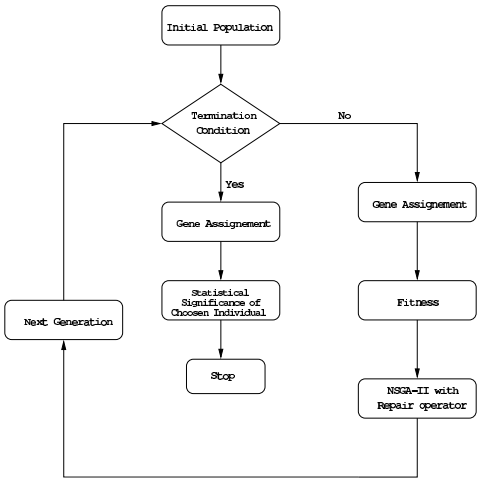


Fig. 3. Schematic Diagram of the proposed approach.

for defining membership function. 0.585 corresponds to 1.5 fold change at logarithmic scale of base 2 (as discussed in section II-B).

1) NSGA-II Parameters: We tried few parameters but results were not not much dependent on parameters. We have used parameters (Table I) in all the reported results unless otherwise mentioned . In all cases, we have taken 10 independent runs and as we discuss later NSGA-II converges to same front. Thus for final analysis, we consider simulation from single run only.

TABLE I
PARAMETER SETTINGS FOR NSGA-II

Parameter	Value
Number of generations	300
Population Size	100
Crossover probability	0.9
Mutation probability	0.01
Number of independent runs	10
K_{min}	10

2) K-means: For our experiments, we use Matlab 6.5 implementation of k-means using correlation coefficient as distance measure. In this case also 10 independent runs were taken for each value of k and the one which gave minimum error, is selected for further analysis. Silhouette width [33] is used to select the final value of k. Since k-means does not assign significance to the found clusters, we choose the top several enriched clusters for biological significance analysis.

3) STEM: STEM [23] is clustering algorithm recently developed for the clustering of short time series gene expression data. All simulation of STEM algorithm was performed using STEM [32] with default values of all parameters.

B. Simulated Data

We have used two sets of simulated data containing 5000 genes with 5 time points. One of these datasets were taken from www.cs.cmu.edu/∼jernst/st/ and we generated the second dataset. In each of the simulation experiment, we have first filtered out genes using STEM [32]. For STEM, we have used all recommended default values with c=2, and 50 possible model profiles. For our algorithm all 81 possible templates (discussed in Section II-A) were considered for analysis. Significance of each profile were tested at Bonferroni corrected p-values of 0.05, which corresponds to uncorrected p-values of 0.001.

First simulated data, taken from www.cs.cmu.edu/∼jernst/st/, was totally random and was not containing any pattern. Uniform (10,100) distribution was used to generate raw expression value at each time point. Each value thus generated was random and independent from all other values. 4519 genes were selected for further analysis after filtering by STEM. In this case only 5 independent runs of NSGA-II were taken and Figure 4 shows convergence of NSGA-II. As can be seen from the figure, all runs were converged to almost the same front, we choose arbitrarily one particular run to further analyze the result. It can be also observed from Figure 4, that after around 50 templates, error curve is almost flattened. We decided to consider the individual with 50 model templates for further analysis as the same number of model templates were also considered in STEM. Statistical significance analysis of these 50 templates has shown that none of them are significant. We plotted expected versus assigned number of genes in each template profile (shown in Figure 5). The diagonal line corresponds to the number of genes expected at uncorrected significance level of 0.001. STEM also did not find any of the profiles significant.

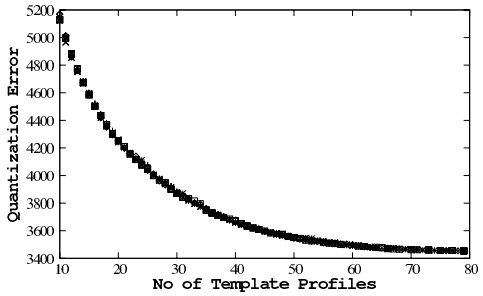


Fig. 4. Simulated Data 1: Quantization error versus number of model templates. Around 50 model templates error curve started to get flat.

We generated another set of artificial data with three profiles and 50 genes were pre-assigned to each of them. Rest of the genes were generated as stated above. The three profiles were 1101, 2102 and 1201 (as shown in Figure 6). In each case value at time point 0 was generated with uniform (10,100) distribution. The raw expression values at the other time points were generated as follows

$$x_k = \begin{cases} x_{k-1} * U(1.5,3) + U(0,1) & \text{if } p_k = 1, \\ x_{k-1} * U(0,0.6) + U(0,1) & \text{if } p_k = 2, \\ x_{k-1} + U(0,1) & \text{if } p_k = 0. \end{cases}$$

Here, x_k denotes the expression value at time point k and

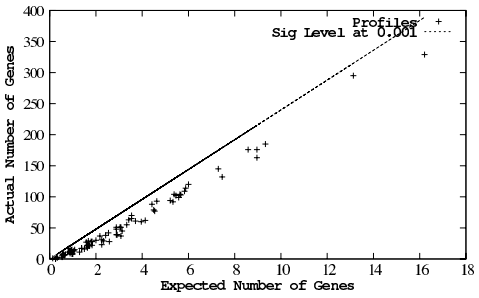


Fig. 5. Simulated Data 1: Expected versus actual number of genes assigned to each profile. Diagonal line represents uncorrected significance level of 0.001. No profiles were found significant by the proposed method.

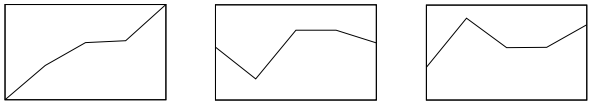


Fig. 6. Three random profiles were generated. 50 genes were planted to these three profiles.

p_k represents the k^{th} index of the pattern p. Again, all genes were filtered using default criteria of STEM. In this case also, 5 independent runs of NSGA-II were taken and again we observe the similar behavior as in the previous case. We took 50 model templates for statistical significance analysis. Our method correctly identified all three profiles as significant and none of the other profile was selected as significant (Figure 7). Whereas, STEM identified only two out of three planted profiles as statistically significant profiles (Figure 8). It is noteworthy to mention that several profiles were more enriched than the identified significant profiles, still our method was able to distinguish real pattern from random patterns. This is definitely an advantage over the conventional clustering algorithms.

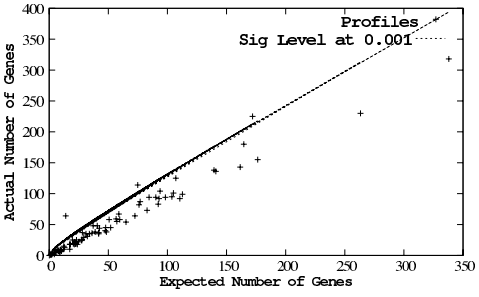


Fig. 7. Simulated Data 2: As can be seen three profiles were found significant by our method.

C. Biological Data

To further evaluate our algorithm we analyzed the time series gene expression data on immune response to *Helicobacter Pylori* infection from [5]. [5] used human cDNA microarrays to investigate the temporal behavior of gastric

Fig. 8. Simulated Data 2: Results from STEM. Only two profiles were identified as significant.

epithelial cells infected with H. pylori strain G27 and various mutants. We used the temporal data obtained from infection of G27 strain. Array was containing total 24,192 genes and time series measurement were taken at 5 time points, 0, 0.5, 3, 6 and 12 hrs. Once again genes were filtered with default values of STEM and 2137 genes were selected for further analysis.

In this case also, default values were used for STEM and all 81 possible templates (Section II-A) were considered for our method. In this case, 10 independent runs of NSGA-II were taken to get optimal number of model templates. Convergence of NSGA-II is shown in figure 9. Once again

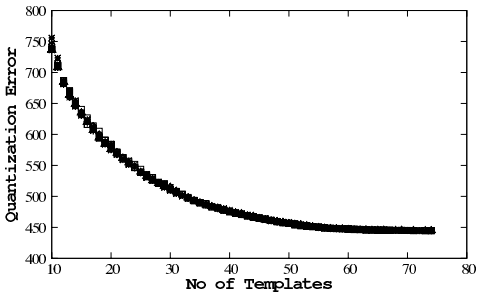
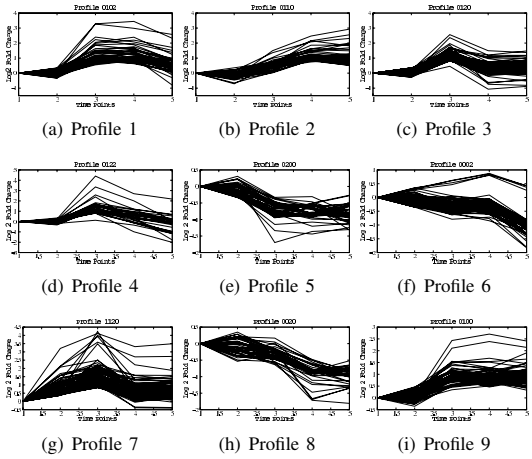


Fig. 9. Biological Data: Quantization error versus number of model templates. Around 50 model templates error curve started to get flat.

50 model templates were considered for further statistical and biological significance analysis. 9 template profiles were found statistically significant. All these profiles are shown in Figure 10. All significant profiles were analyzed for GO categories enrichment. We used EASE [34] for GO categories enrichment analysis. Four out of the 9 significant model templates were found significantly enriched for GO categories with EASE score [34] of 0.005 or less. We observe that there are many unannotated genes present in the array. This could explain why not all profiles were significantly enriched with GO categories. Number of genes assigned to each of the significant templates are shown in the Table II. Below we describe some of profiles enriched for GO categories.



(a) Profile 1 (b) Profile 2 (c) Profile 3

(d) Profile 4 (e) Profile 5 (f) Profile 6

(g) Profile 7 (h) Profile 8 (i) Profile 9

Fig. 10. All 9 statistically significant profiles found by the proposed method.

TABLE II

NUMBER OF GENES ASSIGNED TO EACH PROFILE

Profile	Total Genes	Annotated Genes
0002	80	53
0020	41	20
0100	76	41
0102	46	24
0110	37	33
0120	138	31
0122	30	6
0200	45	29
1120	142	34

Profile 2 (0002) contained 80 genes in all and 53 of them are annotated and hence considered for GO categories enrichment analysis. This profile was significantly enriched for cell cycle, cell proliferation and DNA replication. Profile 18 (0200) contained 45 genes and only 29 of them are annotated. The most significant GO category for this profile was transferase activity. Profile 17 (0122) is the most interesting among all with two reasons. First one it contained only 30 genes and yet our algorithm was able to pick it as statistically significant. The second is its biological relevance. Response to pest/pathogen/parasite, immune response were most significant enriched GO category in this case. As the actual experiment involved pathogen infection, these categories are quite expected [5]. This profile contained many other categories, like humoral immune response, response to external stimulus etc, but due to very few annotated genes, were not found significant.

To evaluate the performance of the proposed method, we compare it to k-means and STEM, a recently developed clustering algorithm specifically designed for short time series gene expression data. We run k-means algorithm for $k = 20, 25, 30, 35, 40, 45, 50$ and choose $k = 25$ as silhouette width was maximum for this value of k. Since k-means does not give any statistical significance to the found clusters, we choose 10 most enriched clusters for comparison analysis. Performance of the three algorithms is compared based on the enrichment analysis of GO categories. STEM identified 10 statistically significant profiles. Cell cycle, immune response were important common enriched categories identified by STEM and the proposed method. K-means also identified as cell cycle as significant but fail to identify more specific category immune response or any other similar categories, relevant to the experiment[5].

All three algorithms were able to pick common relevant categories but only the proposed method and STEM were able to pick the more biologically relevant categories. This definitely shows the advantage of specially designed algorithm over the generic algorithms like k-means. There are clear advantages over k-means or any other distance based generic clustering methods, i.e., able to pick statistically significant profiles which are also more biologically consistent with the considered experiment and data. It is not very clear whether STEM or the proposed method is better than the other. This require more comparative studies on other biological data accompanied with experimental verification of the results obtained from the two methods. There are significant differences in approach between STEM and the proposed method. Though the two methods use similar way to define templates profiles, but selection of model templates and gene assignment to them are completely different. STEM identifies model templates without considering the actual data, our method do consider data while selecting model templates. In STEM, gene assignment is based on distance whereas we proposed a novel gene assignment method independent of any distance matrix and totally dependent on fold change between consecutive time points.

IV. CONCLUSION

The type and the specific purpose of experiments have to be considered in order to choose the most suitable algorithm. Conventional clustering algorithms based on the Euclidean distance or correlation coefficient are not able to properly reflect the inherent ordered information embedded in time series or any dose-response microarray experiments. In this paper, we have proposed a novel approach of gene assignment to different predefine profiles. Genes were assigned to different predefine profiles using fuzzy membership function. Fuzzy membership function was defined on transitional change in expression levels of gene during consecutive time points. As there are quite large number of template profiles, there is need to select manageable and meaningful number of templates. Since deciding such a number is non-trivial, we use a MOEA, NSGA-II to get a set of trade-off solutions with varying number of template profiles and corresponding quantization error. Based on that selected template profiles were tested for statistical significance. Statistical significance test gives the proposed method ability to distinguish between real and random patterns.

The algorithm was tested on both simulated and biological data. It was shown that our method was able to identify small number of planted genes from large random data and correctly assign them into respective correct profiles. When applied on real biological data, algorithm was again able

to identify patterns significantly enriched with biological patterns. Our method was compared with STEM on both simulated and biological data. In case of second simulated data, STEM could identify only 2 profiles, whereas our method was able to correctly identify all three planted profiles. But for biological data, it was shown that the statistical significant profiles selected by STEM and our method were similar and comparable. Our method was able to identify profiles containing genes more biological relevant to the experiment than k-means.

This approach can be extended to exploit the functional or other available biological information into gene assignment. A gene can belong to many functional categories and similarly it can be assigned to several profiles simultaneously by our method. One way of doing this could be, instead of assigning genes to a profile solely based on their membership values, we can put them into a category where more biologically homogeneous genes are assigned.

ACKNOWLEDGEMENT

Authors acknowledge the financial support offered by the A*Star (Agency for Science, Technology and Research) under the grant # 052 101 0020 to conduct this research.

REFERENCES

[1] Z. Bar-Joseph, "Analyzing time series gene expression data," *Bioinformatics*, vol. 20, no. 16, pp. 2493–2503, 2004.

[2] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle regulated genes of yeast saccharamyces cerevisiae microarray hybridization," *Molecular Biology Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[3] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. Brown, and I. Herskowitz, "The transcriptional program of sporulation in budding yeast," *Science*, vol. 282, pp. 699–705, 1998.

[4] M. N. Arbeitman, E. E. Furlong, F. Imam, E. Johnson, B. Null, B. Baker, M. Krasnow, M. Scott, R. Davis, and K. White, "Gene expression during the life cycle of drosophila melanogaster," *Science*, vol. 298, pp. 2270–2275, 2002.

[5] K. Guillemin, N. Salama, L. Tompkins, and S. Falkow, "Cag pathogenicity island-specific responses of gastric epithelial cells to helicobacter pylori infection," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 23, pp. 15 136–15 141, 2002.

[6] T. V. Getchell, H. Liu, R. A. Vaishnav, K. Kwong, A. J. Stromberg, and M. L. Getchell, "Temporal profiling of gene expression during neurogenesis and remodeling in the olfactory epithelium at short intervals after target ablation," *Journal of Neuroscience Research*, vol. 80, pp. 309–329, 2005.

[7] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 14 863–14 868, 1998.

[8] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature Genetics*, vol. 22, pp. 281–285, 1999.

[9] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," *PNAS*, vol. 96, no. 6, pp. 2907–2912, 1999.

[10] D. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1989.

[11] C. Murthy and N. Chowdhury, "In search of optimal clusters using genetic algorithms," *Pattern Recognition Letters*, vol. 17, no. 8, pp. 825–832, 1996.

[12] S. Bandyopadhyay and U. Maulik, "Genetic clustering for automatic evolution of clusters and application to image classification," *Pattern Recognition*, vol. 35, no. 6, pp. 1197–1208, 2002.

[13] G. Di Gesú Vito, L. Giosué, R. Alessandra, and S. Davide, "GenClust: A genetic algorithm for clustering gene expression data," *BMC Bioinformatics*, vol. 6.

[14] J. Handl and J. Knowels, "An evolutionary approach to multiobjective clustering," *IEEE Trans. Evol. Comput.*, vol. 11, no. 1, 2007.

[15] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons, 2001.

[16] G. J. McLachlan, R. W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics*, vol. 18, pp. 413–422, 2002.

[17] A. Schliep, A. Schonhuth, and C. Steinhoff, "Using hidden markov models to analyze gene expression time course data," *Bioinformatics*, vol. 19, no. Suppl.1, pp. i255–263, 2003.

[18] Z. Bar-Joseph, G. Greber, D. K. Gifford, T. S. Jaakkola, and I. Simon, "A new approach to analyzing gene expression time series data," 2002.

[19] Y. Luan and H. Li, "Clustering of time-course gene expression data using a mixed-effects model with b-splines," *Bioinformatics*, vol. 19, no. 4, pp. 474–482, 2003.

[20] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 14, pp. 9121–9126, 2002.

[21] X. Ji, J. Li-Ling, and Z. Sun, "Mining gene expression data using a novel approach based on hidden markov models," *Febs Letters*, vol. 542, pp. 125–131, 2003.

[22] M. J. L. de Hoon, S. Imoto, and S. Miyano, "Statistical analysis of a small set of time-ordered gene expression data using linear splines," *Bioinformatics*, vol. 18, no. 11, pp. 1477–1485, 2002.

[23] J. Ernst, G. J. Nau, and Z. Bar-Joseph, "Clustering short time series gene expression data," *Bioinformatics*, vol. 21, no. Suppl.1, pp. i159–i168, 2005.

[24] L. Wang, M. Ramoni, and P. Sebastiani, "Clustering short gene expression profiles," 2006.

[25] C. S. Moller-Levet, K.-H. Cho, and O. Wolkenhauer, "Microarray data clustering based on temporal variation: Fcv with tsd preclustering," *Applied Bioinformatics*, vol. 2, no. 1, pp. 35–45, 2003.

[26] H. Liu, S. Tarima, A. S. Borders, T. V. Getchell, M. L. Getchell, and A. J. Stromberg, "Quadratic regression analysis for gene discovery and pattern recognition for non-cyclic short time-course microarray experiments," *BMC Bioinformatics*, vol. 6, 2005.

[27] L. P. Zhao, R. Prentice, and L. Breeden, "Statistical modeling of large microarray data sets to identify stimulus-response profiles," *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 5631–5636, 2001.

[28] K. Deb, S. Agarwal, A. Pratap, and T. Meyarivan, "A fast and elitist multi-objective genetic algorithm : NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, pp. 182–197, 2002.

[29] S. Dudoit, Y. H. Yang, Y. H. Callow, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments," *Statistica Sinica*, vol. 12, pp. 111–139, 2002.

[30] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Natl. Acad. Sci. USA*, vol. 98, no. 9, pp. 5116–5121, 2001.

[31] X. Cui and G. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biol*, vol. 4, no. 4, p. 210, 2003.

[32] J. Ernst and Z. Bar-Joseph, "Stem: a tool for the analysis of short time series gene expression data," *BMC Bioinformatics*, vol. 7, 2006.

[33] P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, 1987.

[34] D. Hosack, G. J. Dennis, B. T. Sherman, H. C. Lane, and R. Lempicki, "Identifying biological themes within lists of genes with ease," *Genome Biology*, vol. 4, no. 10, 2003.