Chapter 72
A Density-Based Clustering for Gene Expression Data Using Gene Ontology

Koyel Mandal and Rosy Sarmah

Abstract Gene expression clustering is built on the premise that similarly expressed genes are included in the same kind of biological process. Recent research has focused on the fact that incorporation of biological knowledge such as gene ontology (GO) improves the result of clustering. This paper demonstrates a *Semi-supervised Density-based Clustering* (SDC) which uses GO to detect positive and negative co-regulated patterns from the noisy gene expression data. SDC improves a previous algorithm *DenGeneClus* (DGC) which could handle only positive co-regulation and did not include GO in the clustering process. Experimental results on four real-life data show that SDC outperforms DGC based on z-score and gene ontology enrichment analysis.

Keywords Gene expression data · Semi-supervised clustering · Gene ontology
Positive and negative co-regulated gene

Introduction

High-throughput experiments such as DNA microarray technology have generated huge amount of data, analysis of which require high performing computational methods such as clustering. DNA microarrays have helped researchers to observe the expressions of enormous amount of genes at different conditions such as time, different stages of diseases, or drug applications [1]. The output of DNA microarrays after several preprocessing steps is finally obtained as a gene expression numerical matrix which is mathematically written as, $GE_{G \times C} = \{ev_{i,t} | i \in G, t \in C\}$, where

K. Mandal (✉) · R. Sarmah
Department of Computer Science and Engineering, Tezpur University,
Sonitpur 784028, Assam, India
e-mail: koyel@tezu.ernet.in

R. Sarmah
e-mail: rosy8@tezu.ernet.in

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_72

$ev_{i,t}$ denotes the expression value of ith gene at condition t, G the total gene set, and C the condition set, respectively [1].

Clustering helps to identify co-expressed, coherent, and co-regulated genes from the given expression data [2]. Co-regulated genes can be of two types: positively co-regulated and negatively co-regulated genes [3]. Genes g_1 and g_2 are said to be positively co-regulated if the expression value of g_1 increases (or decreases) from condition t_i to t_j then the expression level of g_2 also increases (or decreases) from t_i to t_j. Two genes g_1 and g_2 are said to be negatively co-regulated if the expression value of g_1 increases (or decreases) from condition t_i to t_j then the expression level of g_2 also decreases (or increases) from t_i to t_j.

Unsupervised clustering algorithms are built on the presumption that co-expressed genes are likely to have common biological functions. However it is seen that most of the algorithms miss the gene functional prediction at the time of clustering. This has motivated us to shift from unsupervised to semi-supervised clustering by incorporating gene ontology (GO) knowledge in the clustering process. GO is the fundamental database of bioinformatics that specifically gives the annotations for gene products with consistent and structured vocabularies [4].

Algorithms based on the density information give quality clusters, and when GO knowledge is incorporated into the clustering process, we get more biologically significant clusters. Therefore, in this paper, we have combined both density and GO information to get the benefits of both in our proposed algorithm.

A Semi-supervised Density-based Clustering (SDC) is being proposed, using the density information of a gene and external knowledge from GO to discover more biologically relevant clusters from noisy data. This work overcomes the drawbacks of a density-based clustering algorithm (DenGeneClus, DGC) [5] by discovering both the positively and negatively co-regulated genes.

Background

Clustering of gene expression data is widely classified into five types, viz. hierarchical, partitional, density-based, graph-theoretical, and model-based [1, 6, 7]. From the various approaches surveyed, we find that density-based algorithm is not dependent on number of clusters. DGC [5], DHC [8], OverDBC [9], and Bayesian-OverDBC [10] are the examples of density-based gene expression data clustering. Conventional clustering algorithms find sets of genes depending on their proximity (similarity or dissimilarity) measure. Most commonly used proximity measure is Euclidean distance which gives the dissimilarity between gene g_i, g_j as $Dis_{Euc}(g_i, g_j) = \sqrt{(\sum_{t=1}^{|C|}(g_{i,t}, g_{j,t})^2)}$ [1]. Expression based measures may not find the potential relationships among the genes. Therefore, it is necessary to guide the clustering process with external domain knowledge. Semantic similarity measure is the key technique to incorporate the knowledge of known genes from gene ontology and gene annotation file. Some of the well-known semantic similarities are

Resnik's, Jiang and Conraths's and Lin's [11]. These semantic similarities are built on the information theory which means how much information they commonly share. Information content of a term t represented as IC in a specific corpus is described by $IC = -log(P(t))$, where $P(t)$ represents probability of occurrence of t. In our proposed method, we have used Lin similarity between two terms say t_i and t_j and given by $Sim_{Lin}(t_i, t_j) = \frac{2 \times IC(LCA)}{IC(t_i)+IC(t_j)}$. Sim_{Lin} gives the IC between two terms by considering the IC of each individual term and the IC of lowest common ancestors (LCA). The value of Sim_{Lin} lies between 0 and 1. We combined Sim_{Lin} and Dis_{Euc} to improve the clustering result. We first convert Dis_{Euc} into a similarity measure as $Sim_{Euc} = \frac{1}{1+Dis_{Euc}}$. Then we find the combined similarity (Com_sim) given next.

$$Com\_sim = w1 * Sim_{Euc} + w2 * Sim_{Lin} \tag{72.1}$$

where, $w1 + w2 = 1, 0 \leq w1 \leq 1$ and $0 \leq w2 \leq 1$ [4]. $w1$ and $w2$ control the weights to two similarity measures. Hang et al. [12] proposed an algorithm using two information such as gene density function and biological knowledge, and the proposed one gave better result than standard algorithm. Zhou et al. [13] also proposed an algorithm incorporating density of data and gene ontology in distance-based clustering algorithm. Both the algorithms do not address the issue of identifying the positive and negative co-regulated genes. An algorithm which finds clusters comprised with co-regulated genes is being proposed by Ji and Tan [3]. To identify interesting partial negative positive co-regulated gene cluster, Koch et al. [14] proposed an algorithm which also discovers overlapping clusters.

Proposed Method

Our SDC is a density-based clustering algorithm which works in two phases (i) preprocessing and (ii) clustering phase.

Preprocessing step is initiated by normalizing (standard deviation 1 and mean 0) the gene expression data. Then, a discretization process discretizes the gene expression data, and the discretized data (GE_{disct}) is fed as input to the clustering algorithm.

In **Discretization** step, each cell $ev_{i,t}$, (where $t = 1$) of the gene expression data (GE) for the first condition is discretized by using Eq. 72.2, and for the other conditions ($C - t_1$), each cell $ev_{i,t}$ (where $t = 2, 3..., |C|$) is computed using Eq. 72.3. Each gene in GE_{disct} will now have a pattern of regulation values 0^s, 1^s, and 2^s across condition known as regulation pattern.

$$GE_{disct}(i, 1) = \begin{cases} 2 & \text{if } ev_{i,1} < 0 \\ 0 & \text{if } ev_{i,1} = 0 \\ 1 & \text{if } ev_{i,1} > 0 \end{cases} \tag{72.2}$$

$$GE_{disct}(i,t) = \begin{cases} 2 & \text{if } ev_{i,t} < ev_{i,t-1} \\ 0 & \text{if } ev_{i,t} = ev_{i,t-1} \\ 1 & \text{if } ev_{i,t} > ev_{i,t-1} \end{cases} \tag{72.3}$$

After the computation of each gene's regulation pattern, next job is to calculate the match (M) between genes g_i and g_j stated in Eq. 72.4.

Definition 1 Match: Match (M) gives the number of common regulation value according to the conditions except the first one, which signifies how similar two patterns are with respect to their expression values.

If $M = |C| - 1$, it can be said that two patterns are almost similar. The match between g_i and g_j is calculated as below.

$$Pat_t^{i,j} = \begin{cases} 1 & \text{if } GE_{disct}(ev_{i,t}) = GE_{disct}(ev_{j,t})\text{where } t = 2, ..., |C| \\ 0 & \text{otherwise} \end{cases} \tag{72.4}$$

$$M(g_i, g_j) = number\ of\ 1^s\ in\ Pat_t^{i,j} \tag{72.5}$$

Definition 2 Maximal Match: If match between g_i and g_j is equal or greater to the minimum threshold value δ, ($M(g_i, g_j) >= \delta$) and no other gene exists whose match (M) with respect to g_i is greater than g_j, then g_i has a maximal match (MM) with another g_j ($g_i \neq g_j$).

Definition 3 Maximally Matched Regulation Pattern: For genes g_i and g_j, let g_i be maximally matched with g_j, then the Maximally Matched Regulation Pattern ($MMRP$) is computed using Eq. 72.6 by considering the subset (two gene profiles may not match throughout $|C| - 1$ conditions) of conditions where they maximally matched based on δ.

$$MMRP(g_{i,t}) = MMRP(g_{j,t}) = \begin{cases} 2 & \text{if } GE_{disct}(g_{i,t}) = 2 = GE_{disct}(g_{j,t}) \\ 0 & \text{if } GE_{disct}(g_{i,t}) = 0 = GE_{disct}(g_{j,t}) \\ 1 & \text{if } GE_{disct}(g_{i,t}) = 1 = GE_{disct}(g_{j,t}) \\ x & \text{otherwise} \end{cases} \tag{72.6}$$

where, $t = 2, 3, ..., |C|$. Therefore, for the whole set of t conditions, we obtain an $MMRP$ pattern of 0^s, 1^s, 2^s and x^s.

Definition 4 Negative Maximally Matched Regulation Pattern: The Negative Maximally Matched Regulation Pattern ($NMMRP$) of g_j is determined by comparing the $MMRP$ of g_i as stated in Eq. 72.7.

$$NMMRP(g_{j,t}) = \begin{cases} 2 & \text{if } MMRP(g_{i,t}) = 1 \\ 1 & \text{if } MMRP(g_{i,t}) = 2 \\ 0 & \text{if } MMRP(g_{i,t}) = 0 \\ x & \text{if } MMRP(g_{i,t}) = x \end{cases} \tag{72.7}$$

Therefore, we obtain a *NMMRP* pattern for t conditions ($t = 2, 3, ..., |C|$).

Definition 5 Rank: *Rank* gives the ascending order of expression levels of a gene across conditions.

Rank is measured by giving a ranked value starting from 1 to all the expression values in the MMRP pattern except for those conditions having a x value. The working example of the computation of *M*, *MM*, *MMRP*, *NMMRP*, and *Rank* is available in http://agnigarh.tezu.ernet.in/~rosy8/workingexampleSDC.pdf.

The second phase, **Clustering** of SDC is based on some of the fundamental concepts of density-based clustering. The following definitions are trivial to the clustering process.

Definition 6 ϵ-neighbor: ϵ-neighbors with respect to $g_i \in G$ are those genes $g_k \in G$, which have more similarity than the user defined threshold (ϵ). Here, we have used combined similarity which is mentioned in Eq. 72.1.

$$\epsilon - neighbors(g_i) = \{g_k | where\; g_k \in G\; and\; com\_sim(g_i, g_k) >= \epsilon\} \tag{72.8}$$

Definition 7 Core-neighbors: Core-neighbors of a gene $g_i \in G$ is described by a set of genes $G^* \in G$ and should satisfy the following four criteria. A gene, say g_i is considered as core gene if.

1. $\forall g_y \in G^*, g_y \in \epsilon - neighbors(g_i)$.
2. $MMRP(g_y) \approx MMRP(g_i)$.
3. $Rank(g_y) \approx Rank(g_i)$.
4. $|G^*| >= min\_points$ (a user defined threshold).

To compute the core-neighbors of a particular gene g_i, we check the above-mentioned four criteria for all the $C - 1$ dimensions (except condition 1). If we do not get the core-neighbors, we will go on checking the criteria by reducing the search space one condition at a time. At first we reduce the condition set by $C - \{t_1, t_l\}$, where t_1 and t_l are the first and the last condition respectively, i.e., $|C| - 1 - 1 = |C| - 2$. If we still do not find the core-neighbors of g_i, we further reduce the search space by the second last condition i.e., $C - \{t_1, t_{l-1}, t_l\}$, where t_1, t_{l-1} and t_l are the first, second last and the last condition respectively. In other words the condition set is reduced by $|C| - 2, |C| - 3, |C| - 4 \ldots$ and so on.

Definition 8 Direct density reachable: g_i is direct density reachable with respect to g_j if it fulfills three basic principles.

1. g_j must be a core gene or g_j must have core-neighbors.
2. $g_i \in \epsilon - neighbors(g_j)$.
3. $MMRP(g_i) \approx MMRP(g_j)$.

In case of pairs of core genes, direct density reachable relation holds symmetric relation.

Definition 9 Density reachable: Gene g_q is density reachable from g_p provided there is a chain of genes $g_1, g_2, g_3, ..., g_n$ such that $g_1 = g_p$ and $g_n = g_q$ and every g_{i+1} gene is directly density reachable from g_i^{th} gene.

Definition 10 Connected: Gene g_i is connected to g_j with respect to ϵ, provided g_i and g_j are reachable from a common gene say g_k.

This relation holds symmetric property.

Definition 11 Cluster: A cluster CL ($|CL| >= min\_points$) is a collection of reachable and connected genes. Say, a gene $g_i \in CL$ and the gene g_j is found to be reachable from g_i, then g_j must be in cluster CL. Similarly, if a gene $g_i \in CL$ and g_j is connected to g_i then g_j will be in the same CL cluster.

Definition 12 Noise: A noise gene is a gene which does not belong to any cluster.

The steps of SDC is given next. At first, all genes are not clustered.

```
Step 1   Start with an random unclustered gene say $g_i$.
Step 2   Find the $MMRP(g_i)$ and $Rank(g_i)$.
Step 3   Find the core-neighbors of $g_i$ using Definition 7.
Step 4   For each core-neighbors of $g_i$.
    Step 4.1   Identify all connected and reachable genes with respect to each core-neighbors.
    Step 4.2   Give the same $cluster\_id$ for all these genes.
Step 5   End of step 4.
Step 6   Find the $NMMRP$ from $MMRP$ of the newly formed $cluster\_id$ .
Step 7   Find the unclustered genes which matches the $NMMRP$.
Step 8   For each gene $g_j$ with matched $NMMRP$.
Step 9   Find the core-neighbors of gene $g_j$ and all reachable and connected genes from it.
Step 10   Assign another cluster_id to all the reachable and connected genes of $g_j$.
Step 11   Repeat step 1 to 10 with the next unclustered gene.
Step 12   All the unclustered genes are marked as noise.
Step 13   End
```

Table 72.1 A brief description about the datasets

Serial no.	Name of dataset	Genes/samples	Source
D1	Yeast cell cycle	384/17	http://anirbanmukhopadhyay.50webs.com/data.html
D2	Yeast sporulation	474/7	http://cmgm.stanford.edu/pbrown/sporulation/
D3	Yeast diauxic shift	614/7	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28
D4	Prostate cancer	339/102	http://archive.broadinstitute.org/mpr/publications/projects/Cancer_Susceptibility/references_and_URLS_of_datasets.html

Experimental Result

Implementation of SDC and DGC was done in MATLAB 2015 platform and experimented over four publicly available real-life gene expression datasets. Table 72.1 gives a description about the used datasets in the experiment.
To compute Sim_{Lin}, we have downloaded the most recent gene ontology file (released on 2016-09-10) and annotation files (*Saccharomyces Genome Database* and *Homo Sapience*) from www.geneontology.org. To compare both the DGC and SDC clustering results, we use Dis_{Euc} for DGC and *Com_sim* for SDC. The parameter settings highly influence the clustering results. We keep the value of δ as minimum ($\delta = 3$) as possible. To determine the value of ϵ and *min_points* (= 4), we follow the method mentioned in [15]. As we want to give the more weightage on proximity measure than semantic similarity measure, we kept the value of $w1 = 0.6$ and $w2 = 0.4$. The ϵ for DGC and SDC changes from one dataset to another. The ϵ of DGC is 3 and 1.2 for D1 and D2; and for D3 and D4, it is 1 and 10, respectively. The ϵ of SDC is 0.3 for D1, 0.4 for D2, 0.3 for D3 and 0.3 for D4, respectively.

To assess the biological significance of clusters, we eventually investigated the clusters generated by DGC and SDC by functional enrichment analysis. A cluster is called enriched, if at least one of the GO term of a particular cluster from the Biological Process is below the level of significance. P value is being computed using FuncAssociate 3.0 with 5% level of significance [16]. We then analyze the functional category Biological Process (BP) using web (http://www.ebi.ac.uk/QuickGO/) based on GO annotation database. It can be observed from Fig. 72.1 that SDC finds more enriched clusters than DGC.

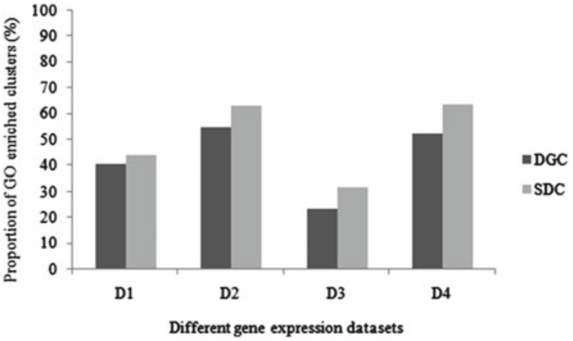

Fig. 72.1 Proportion of enriched clusters of DGC and SDC for different datasets

Table 72.2 Comparison of z-score between DGC and SDC on various datasets

	DGC		SDC	
Datasets	No. clusters	z-score	No. clusters	z-score
D1	27	2.494	25	5.042
D2	31	9.621	27	12.59
D3	47	4.121	41	4.718

To judge the quality of clusters, we have used the Web-based tool cluster judge [17]. The comparison of DGC and SDC based on z-score for different yeast datasets is shown in Table 72.2. Table 72.2 suggests that the clusters generated by SDC have higher z-score value than DGC which proves that the cluster quality is better for SDC.

Conclusion

We have proposed an algorithm incorporating gene ontology in a density-based clustering algorithm. It is being observed that external domain knowledge gives reliable clusters. The drawback of this algorithm is that it finds disjoint clusters and cannot find overlapping clusters. Biologically it is proven that one gene may participate in many biological pathways, and this allows it to belong to multiple clusters. Detecting overlapping clusters is a crucial task and will be incorporated in our future work.

References

1. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: a survey. IEEE Transactions on knowledge and data engineering **16**(11) (2004) 1370–1386

2. Jiang, D., Pei, J., Zhang, A.: Gpx: interactive mining of gene expression data. In: Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, VLDB Endowment (2004) 1249–1252
3. Ji, L., Tan, K.L.: Mining gene expression data for positive and negative co-regulated gene clusters. Bioinformatics **20**(16) (2004) 2711–2718
4. Lee, W.P., Lin, C.H.: Combining expression data and knowledge ontology for gene clustering and network reconstruction. Cognitive Computation **8**(2) (2016) 217–227
5. Das, R., Bhattacharyya, D., Kalita, J.: Clustering gene expression data using an effective dissimilarity measure. International Journal of Computational BioScience (Special Issue) **1**(1) (2010) 55–68
6. Kerr, G., Ruskin, H.J., Crane, M., Doolan, P.: Techniques for clustering gene expression data. Computers in biology and medicine **38**(3) (2008) 283–293
7. Pirim, H., Ekşioğlu, B., Perkins, A.D., Yüceer, Ç.: Clustering of high throughput gene expression data. Computers & operations research **39**(12) (2012) 3046–3061
8. Jiang, D., Pei, J., Zhang, A.: Dhc: a density-based hierarchical clustering method for time series gene expression data. In: Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on, IEEE (2003) 393–400
9. Mirzaie, M., Barani, A., Nematbakkhsh, N., Beigi, M.: Overdbc: A new density-based clustering method with the ability of detecting overlapped clusters from gene expression data. Intelligent Data Analysis **19**(6) (2015) 1311–1321
10. Mirzaie, M., Barani, A., Nematbakkhsh, N., Mohammad-Beigi, M.: Bayesian-overdbc: A bayesian density-based approach for modeling overlapping clusters. Mathematical Problems in Engineering **2015** (2015)
11. Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto, F.M.: Semantic similarity in biomedical ontologies. PLoS comput biol **5**(7) (2009) e1000443
12. Hang, S., You, Z., Chun, L.Y.: Incorporating biological knowledge into density-based clustering analysis of gene expression data. In: Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on. Volume 5., IEEE (2009) 52–56
13. Zhou, X., Sun, H., Wang, D.P., Zhang, Y., Zhou, Y.: Analysis of gene expression data based on density and biological knowledge. In: 2010 Fifth International Conference on Frontier of Computer Science and Technology, IEEE (2010) 448–453
14. Xu, X., Lu, Y., Tung, A.K., Wang, W.: Mining shifting-and-scaling co-regulation patterns on gene expression profiles. In: 22nd International Conference on Data Engineering (ICDE'06), IEEE (2006) 89–89
15. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. Volume 96. (1996) 226–231
16. Hochberg, Y., Benjamini, Y.: More powerful procedures for multiple significance testing. Statistics in medicine **9**(7) (1990) 811–818
17. Gibbons, F.D., Roth, F.P.: Judging the quality of gene expression-based clustering methods using gene annotation. Genome research **12**(10) (2002) 1574–1581