



Function Clustering Self-Organization Maps (FCSOMs) for mining differentially expressed genes in *Drosophila* and its correlation with the growth medium

L.L. Liu, M.J. Liu and M. Ma

School of Mathematical Sciences, Ocean University of China, Qingdao, China

Corresponding author: M.J. Liu
E-mail: 18706480795@163.com

Genet. Mol. Res. 14 (3): 11658-11671 (2015)
Received March 4, 2015
Accepted May 29, 2015
Published September 28, 2015
DOI <http://dx.doi.org/10.4238/2015.September.28.18>

ABSTRACT. The central task of this study was to mine the gene-to-medium relationship. Adequate knowledge of this relationship could potentially improve the accuracy of differentially expressed gene mining. One of the approaches to differentially expressed gene mining uses conventional clustering algorithms to identify the gene-to-medium relationship. Compared to conventional clustering algorithms, self-organization maps (SOMs) identify the nonlinear aspects of the gene-to-medium relationships by mapping the input space into another higher dimensional feature space. However, SOMs are not suitable for huge datasets consisting of millions of samples. Therefore, a new computational model, the Function Clustering Self-Organization Maps (FCSOMs), was developed. FCSOMs take advantage of the theory of granular computing as well as advanced statistical learning methodologies, and are built specifically for each information granule (a function cluster of genes), which are intelligently partitioned by the clustering algorithm provided by the DAVID_6.7 software platform. However, only the gene functions, and not their expression values, are considered in the fuzzy clustering algorithm of DAVID. Compared to the clustering algorithm of DAVID, these experimental results show a

marked improvement in the accuracy of classification with the application of FCSOMs. FCSOMs can handle huge datasets and their complex classification problems, as each FCSOM (modeled for each function cluster) can be easily parallelized.

Key words: Function Clustering Self-Organization Maps; Granular computing; Differentially expressed gene mining

INTRODUCTION

It is axiomatic in functional genomics that genes with similar mRNA expression profiles are likely to be regulated via the same mechanisms (Altman and Raychaudhuri, 2001; Schulze and Downward, 2001). This hypothesis has formed the basis for almost all attempts at using microarray mRNA expression data in the discovery of regulatory networks (Wolfsberg et al., 1999; Wu et al., 2002; Mozhayskiy and Tagkopoulos, 2011; Patnala et al., 2013). Mining of differentially expressed genes corresponding to different functions, and the discovery of functional changes is a highly desirable objective with potential applications in biological projects. This paper aims to analyze the expression of *Drosophila* genes in different larval growth media, in order to determine the possible relations between the differentially expressed gene cluster and similar functions. The Database for Annotation Visualization and Integrated Discovery (DAVID) platform and artificial neural networks provides an opportunity to study these gene clusters in biological networks.

DAVID_6.7 is based on the functional annotation of each gene (and not the gene name), and uses a fuzzy clustering algorithm to group functionally related genes together, as the self-organization map (SOMs) learning subspace. SOM handles nonlinear classification by implicitly mapping input samples from the input feature space into another high dimensional feature space with a nonlinear driving function (Ghouila et al., 2009; Abe et al., 2014). Therefore, SOM may be more effective in revealing the nonlinear gene-to-medium relationship when compared to fuzzy clustering. Since SOM is not suitable for a large dataset (Gu et al., 2012), modeling the whole sample space containing almost one million data samples with one SOM would be impractical. Furthermore, each subspace of the whole sample space corresponds to a different function in this study. This disadvantage of SOM was the major trigger to considering the use of the theory of granular computing.

MATERIAL AND METHODS

This section, introduces the principles of granular computing and SOM; in addition, it explains the motivation for combining granular computing and SOM to create a new computational model offering significant advantages. In addition, the procedures used to train Function Clustering Self-Organization Maps (FCSOMs) modeled for different cluster groups are introduced. Finally, this section introduces a detailed mechanism for the mining of key gene function clusters of *Drosophila melanogaster*, corresponding to different growth media, by FCSOMs.

Granular computing

The basic principles of granular computing have been applied to many fields, including programming, artificial intelligence, interval computing, rough set theory, machine learning, and databases (Tang et al., 2005; Yao, 2005). Granular computing provides true and natural

representations of natural, social, and artificial systems (Zhong et al., 2007).

Granular computing decomposes information to aggregates, such as subsets, classes, and clusters of universal sets, and subsequently solves the targeted problems in each granule (Yao, 2004). Granular construction and computing are the two major tasks of granular computing (Yao, 2005). Granular computing conceptualizes the whole feature space at different granularities and has the ability to switch between these granularities (Yao, 2004). Granular computing uses the divide-and-conquer principle to divide a complex data-mining problem into a series of smaller problems that are simpler to compute. It focuses on each small problem by omitting unnecessary and irrelevant information. Consequently, granular computing increases the intelligence and flexibility of data mining algorithms.

This study, utilized a SOM to determine the relationship between the key genes (differentially expressed genes) and their media. Since different regions of the sample space may correspond to different functions, the building of one SOM in the whole sample space may not be practical. Therefore, the whole sample space was divided into multiple subspaces using an effective granulation method to model a SOM for each function. The fuzzy clustering algorithm of DAVID_6.7 for gene functions was used as the granulation method. The genes in the same subspace are closely related; therefore, SOM can be used (more efficiently) to capture inherent functional distributions for these samples.

Granulation method: Fuzzy clustering algorithm based on gene functions

Fuzzy clustering is a widely used machine learning technique in many real world applications, especially in exploratory gene functional classification (Liu and Xu, 2008; Avogadri and Valentini, 2009; Mukhopadhyay and Maulik, 2009; Maraziotis, 2012; Yan et al., 2013). Fuzzy sets, probabilistic sets, decision trees, clusters, and association rules are some of the granulation methods that come under the framework of granular computing (Yao, 2005). Since a hard partition mechanism may be not suitable for the overlapping datasets, fuzzy clustering is efficient for gene function annotation in the DAVID_6.7 database. Therefore, this study utilized fuzzy clustering as the preferred method of granulation. Genes with similar functional annotations can be grouped together when using the fuzzy clustering algorithm. Therefore, the whole sample space is partitioned intelligently into subspaces, and the complex data mining work is mapped into a series of simpler tasks that are computationally tractable. The fuzzy clustering of DAVID for gene function provides 79, 75, and 92 clusters, corresponding to the different media used. The first 25 clusters were relatively suitable for the SOM developed in this study.

SOMs

SOMs were first proposed by Kohonen, and were described as unsupervised, self-organizing, self-learning networks, widely used in a variety of clustering analyses. Compared to other clustering methods, SOM networks possess self-stability, can perform clustering processes without evaluations, demonstrate real-time learning, and can identify the most meaningful features of the vector space. This makes them especially suitable for unsupervised clustering of high dimensional data sets.

An SOM network is composed of an input layer and an output layer. The number of nodes in the input layer is equal to the dimension of the input vector; each node accepts one component of the input vector. Generally, the output layer is an array of one- or two-dimensional layers. The

topology (structure) of the two-dimensional output layer can be rectangular, hexagonal or a random connection. Each node in the input layer is connected to each node in the output layer by weight. Figure 1 shows the topology of the SOM networks.

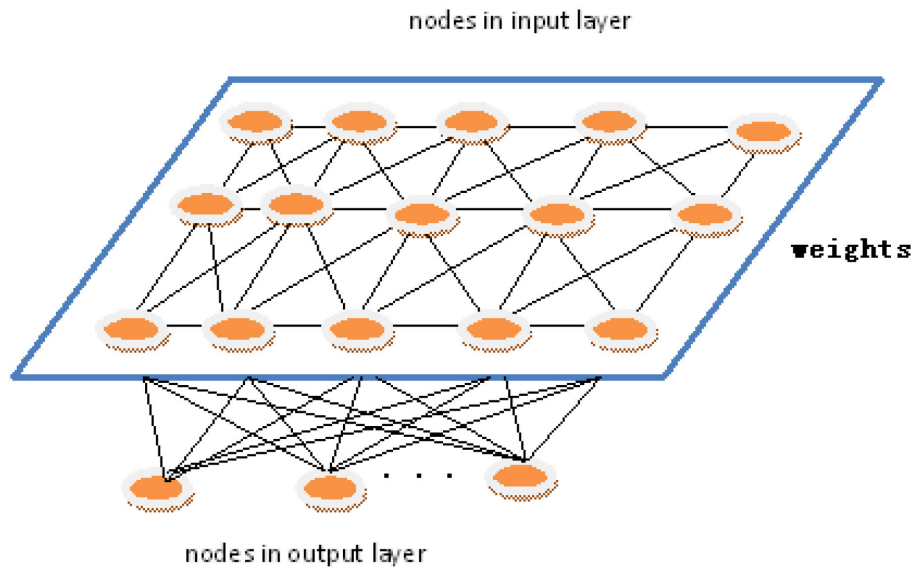


Figure 1. Topology of the SOM networks.

SOM is (essentially) a nonlinear map from a free-dimension discrete or continuous space “V” to a free-dimension discrete space “A”. The application of the vector map helps determine the best matching unit “S” in the output space “A” for a specified vector belonging to “V”. The weight vector “W” of “S” can be regarded as the coordinate of “S” projected onto the input space. The driving function of neurons is generally a linear function; therefore, the output depends on the Euclidean distance between input vectors “X”:

$$X = (x_1, x_2, \dots, x_N)^T$$

And the weight vector “W”

$$W = (w_1, w_2, \dots, w_N)^T$$

the value is given as

$$\sum_{i=1}^N (x_i - w_i)^2$$

When the neuron “C” is the winner, it meets the conditions:

$$\|X - W_0\| = \min_i \|X - W_i\|, i = 1, 2, \dots, N$$

Here, $\| \square \|$ presents the Euclidean distance. The output space "A" represents the input space "V", by approximately adjusting the weight vector "W".

The algorithm is summarized as follows: 1) Initialization: Random values were chosen for the initial weight vector $W_i(0)$. 2) Sampling: Sample X was drawn from the input space with a certain probability; the vector X represents the activation pattern that is applied to the lattice. The dimension of vector X is equal to N. 3) Similarity matching: The best-matching (winning) neuron was found.

$$j(X) = \arg \min_i \|X - W_i\|, i = 1, 2, \dots, N$$

4) Updating: The weight vector of the neighborhood neurons of $j(X)$ was adjusted using the following update formula:

$$w_i(n+1) = w_i(n) + \eta(n)h_{ij}(x)(x(n) - w_i(n))$$

where, $\eta(n)$ is the learning-rate parameter, and $h_{ij}(x)$ is the neighborhood function centered on the winning neuron $j(X)$; both $\eta(n)$ and $h_{ij}(x)$ are dynamically varied during learning, for best results.

5) Continuation: Step 2 was continued with until no noticeable changes in the feature map were observed.

FCSOMs

In this experiment, a new computational model, called FCSOMs, was introduced. FCSOMs creatively take advantage of granule computing and the statistical learning theory to provide a new model for solving complex classification problems.

Advantages of FCSOMs

FCSOMs are built from information granules, which are intelligently partitioned using clustering algorithms. The inherent function distribution of the genes can be performed by intelligent partitioning of the fuzzy clustering algorithms. A complex classification problem can be converted to multiple smaller problems via function partitioning, in order to increase the specificity and efficiency of the learning tasks for each FCSOM. Each FCSOM concentrates on highly correlated samples in each function subspace, without being distracted by noise from other function clusters. Therefore, the key genes in each function cluster can be potentially mined by the FCSOMs.

Since granulation by fuzzy clustering may introduce irrelevant information into each granule without considering the expressional values of the gene, machine learning techniques are required to identify the strength of correspondence between key genes and the medium for each function belonging to the same function. FCSOM can filter out potentially unreliable genes and can select potentially reliable differentially expressed genes for each function after learning the relationship between key genes and a specific function.

Modeling of FCSOMs for each function cluster

The distribution patterns for the expression data of the genes in each cluster are quite different. This requires customization of the functionality and modeling of FCSOMs to identify the differentially expressed genes in each cluster. (Please refer to the section explaining the setup for the definition of differentially expressed genes). The FCSOM for each function cluster was designed to identify differentially expressed genes than could be reliably predicted to be highly correlated with the media. The most effective classification principle is to select the simplest model; therefore, a two-neuron (each neuron represents one cluster center) SOM network structure was adopted.

The procedure is summarized as follows:

1) Sampling: Calculate the Euclidean distances of data vectors of each gene between the standard group and tomato group, between the standard group and ethanol group and between the tomato and ethanol groups.

$$d_i = \|X_i - Y_i\|.$$

Here, X_i and Y_i separately represent the vector of the gene “i” in two different groups. The former 3000 genes with larger Euclidean distances were considered to be the important genes in each contrastive group.

2) Clustering: The important genes in each contrastive group were clustered according to their functions using DAVID_6.7. Following this, the first 25 significant gene clusters were reserved.

3) Classification: If one cluster consists of n genes, there would be n elements in 32 sample vectors. The data samples of one functional cluster were divided into two classes (based on the type media) by SOM.

4) Calculating for accuracy: If the calculated accuracy was 90% upon completion of the classification, the data was directed towards step 6 (all genes in the function clusters were regarded as differentially expressed genes); otherwise, step 5 was performed.

5) The accuracy (p) was calculated according to the following formula; here, the $number_F$ denotes the number of correctly classified genes in cluster F by SOM, while $total\ number_F$ indicates the total number of genes in cluster F.

$$p = \frac{number_F}{total\ number_F}$$

6) Gene knock-out algorithm: The gene knock-out algorithm is summarized in Figure 2.

7) Continuation: Step 3 was continued with until the former 25 function clusters of each contrastive group were completed.

Assuming n genes to be in one cluster and i goes from 1 to n, the algorithm terminates when the accuracy stops rising. In case of more than one classification with the highest accuracy in the algorithm, the distance score was used to evaluate the quality of the classifications with the highest accuracy; in addition, the gene with the highest distance score was retained.

The distance score was calculated using the following formula. Upon the completion of classification of data samples in each cluster by SOM, n-dimension weight vectors of the two winning neurons can be thought of as characteristics data representing different media. The classification

levels for the two neurons of the weight vector were measured from the Euclidean distance. Differences can arise as the clusters have different numbers of genes and the dimensions of the weight vectors of the two winning neurons are different, when data samples from different clusters are classified. This problem was avoided by dividing the Euclidean distance by the dimensions of the weight vectors to measure the classification level (or the distance score):

$$l = \frac{\|w_1 - w_2\|}{n}$$

Here, w_1 and w_2 separately represent the weight vector of neuron 1 and 2, respectively; n indicates the dimension of weight vectors.

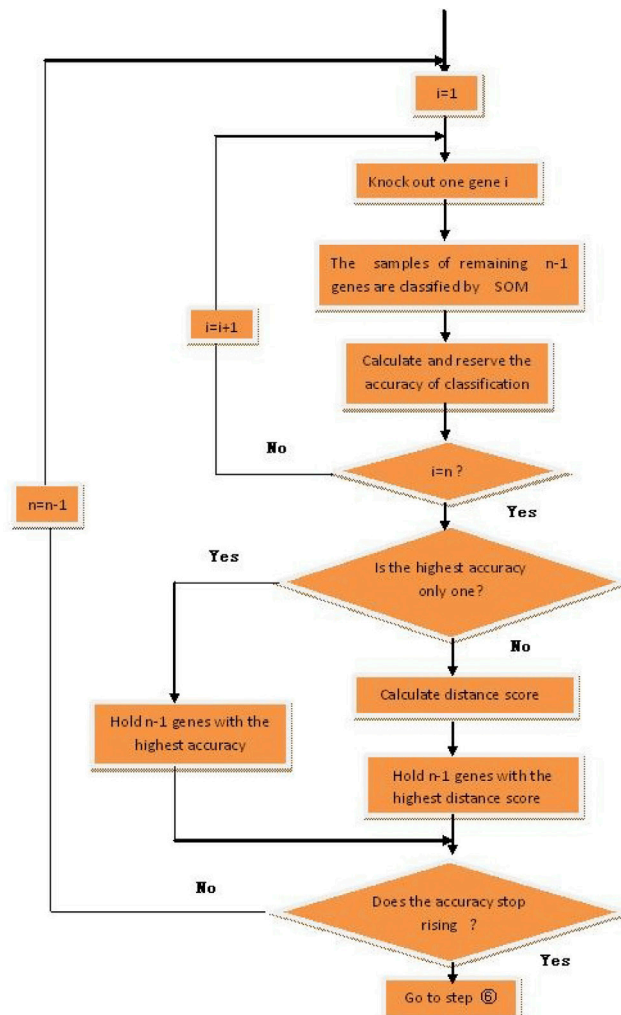


Figure 2. Flow diagram of gene knock-out algorithm.

RESULTS AND DISCUSSION

The purpose of this study was to explore the effects of different larval growth medium on the expression of adult *D. melanogaster* genes, and to mine the differentially expressed genes.

First, the former 3000 genes with the greatest differences in expression were selected from among the standard-tomato, standard-ethanol, and tomato-ethanol contrastive groups for data preprocessing by DAVID_6.7; subsequently, three working databases were obtained for use in the next section.

The NCBI database provided the probe names and gene expression data for *D. melanogaster* in the GPL1322 platform. The first group was maintained in standard medium (standard), the second group was maintained in tomato medium (tomato), and the third group was maintained in ethanol-supplemented medium (ethanol). Each gene had a 16-dimensional vector

$$x_i = (x_{i1}, x_{i2}, \dots, x_{i16})$$

This was composed of the specific gene expression data in 16 samples (corresponding to a specific gene probe). The 16-dimensional vectors of all genes in the standard, tomato, and ethanol groups were used to populate the three original databases.

The effect of the differences in the original gene records to model were removed by normalization of each gene record to [0, 1].

Comparison of accuracy and gene number between clustering algorithm and FCSOM model in standard-ethanol group

The samples containing the former 3000 genes with larger Euclidean distances in the standard-ethanol group were clustered into 79 function clusters. The algorithm given in the FCSOM modeling section for each function cluster was used for the former 25 clusters, the results of which are detailed as Figures 3 and 4.

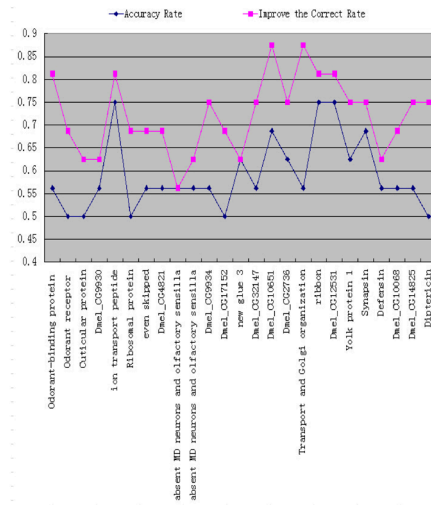


Figure 3. Comparison of classifications accuracy between the clustering algorithms and the FCSOM models in standard-ethanol group.

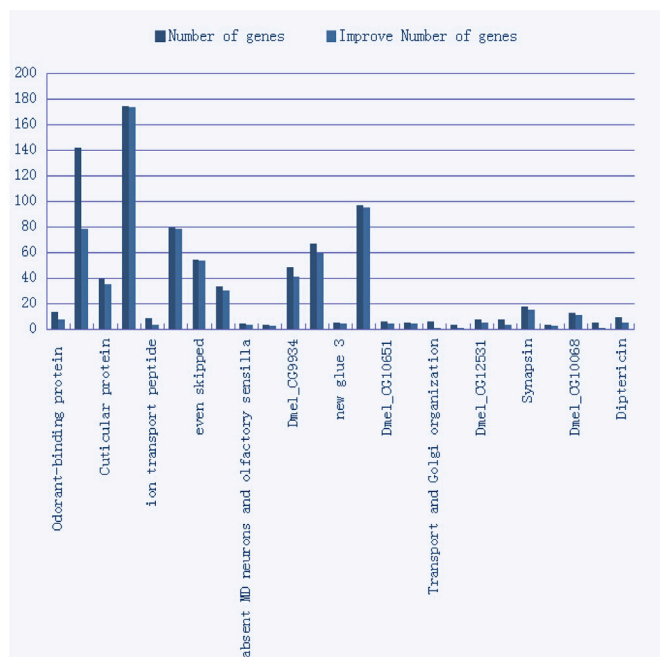


Figure 4. Comparison of numbers of genes between the clustering algorithms and the FCSOM models in standard-ethanol group.

The data presented in Figure 3 compares the accuracy of the classification between the clustering algorithms in DAVID_6.7 and the FCSOM models in standard-ethanol group. The horizontal axis displays the functional clusters arranged left to right as given by the DAVID_6.7. The vertical axis denotes the accuracy. The accuracy for the 23 function clusters improved significantly, with the average accuracy increasing from 59 to 72.25%. The cluster absent MD neurons and olfactory sensilla (5 genes) and new glue 3 (6 genes) had too few genes; therefore, the gene-knockout algorithm was not effective in these cases, and the accuracy did not increase. Despite this, the FCSOM method worked well overall.

Figure 4 displays the number of genes between the clustering algorithm and the FCSOM model. As in Figure 3, the function clusters were presented along the horizontal axis and the vertical axis displayed the number of genes. Some clusters showed obvious change, while some did not. A total of 130 genes were knocked out; the number of differentially expressed genes altered from 868 to 738. The odorant receptor cluster showed 63 genes and 79 knocked out differentially expressed genes, respectively. This indicated that the ethanol-based larval medium affects the gustatory development of *D. melanogaster* (Inoshita and Tanimura, 2006; Bai et al., 2009; Adachi et al., 2012).

Comparison of accuracy and gene number between clustering algorithms and FCSOM models in the standard-tomato group

The samples containing the former 3000 genes with larger Euclidean distances in the standard-tomato group were clustered into 75 function clusters. The algorithm presented in the

FCSOM modeling section (for each function cluster) was used for the former 25 clusters and the results are detailed in Figures 5 and 6.

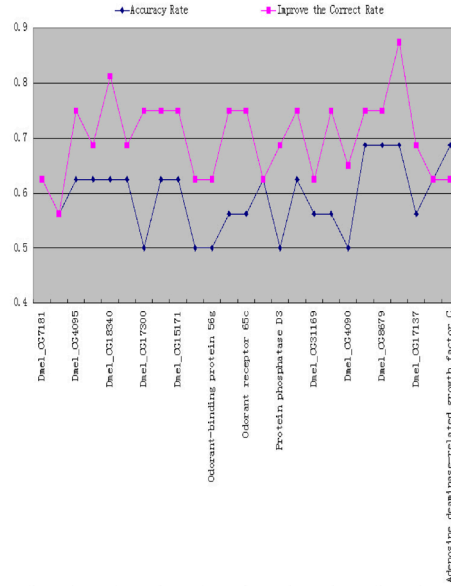


Figure 5. Comparison of classifications accuracy between the clustering algorithms and the FCSOM models in the standard-tomato group.

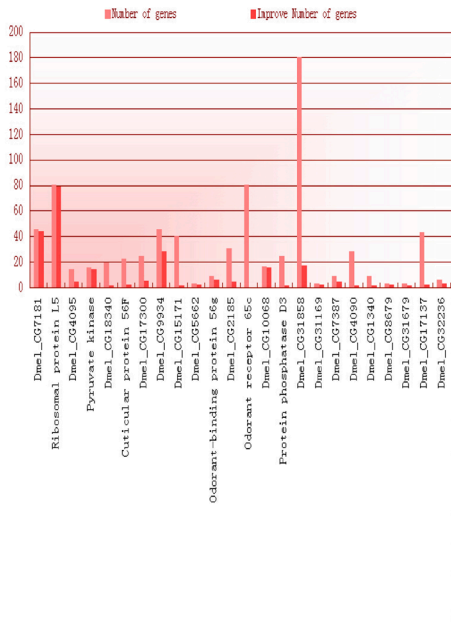


Figure 6. Comparison of numbers of genes between the clustering algorithm and the FCSOMs model in the standard-tomato group.

The data presented in Figure 5 compares the accuracy of the classification between the clustering algorithms in DAVID_6.7 and the FCSOM models in the standard-tomato group. The horizontal axis denotes the function clusters arranged left to right, as given by DAVID_6.7. The vertical axis indicates the accuracy. The accuracy for the 20 function clusters improved significantly, with the average accuracy increasing from 59.5 to 71.3%. The accuracy of the Odorant-binding protein 56g and Odorant receptor 65c proteins showed significant improvements. This showed the effectiveness of the FCSOM modeling.

Figure 6 shows the number of genes between the clustering algorithm and FCSOM model in the standard-tomato group. Some clusters showed obvious changes, while the others did not. A total of 511 genes were knocked out, while the number of differentially expressed genes altered from 778 to 267. Eighty genes were knocked out in the Odorant receptor 65c cluster, while 3 were knocked out in the Odorant-binding protein 56g cluster. The remaining genes were regarded as being differentially expressed. This also affects indicated that the tomato-based larval medium also affects the gustatory development of *D. melanogaster*.

Comparison of accuracy and gene number between clustering algorithms and the FCSOM models in the tomato-ethanol group

The samples containing the former 3000 genes with larger Euclidean distances in the tomato-ethanol group were clustered into 92 function clusters. The algorithm presented in the FCSOM modeling section (for each function cluster) was used for the former 25 clusters and the results are presented in Figures 7 and 8.

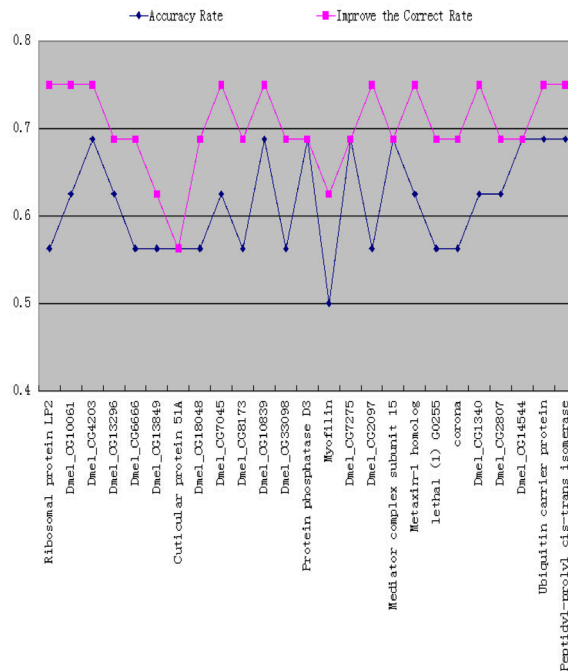


Figure 7. Comparison of classifications accuracy between the clustering algorithm and FCSOMs model in the tomato-ethanol group.

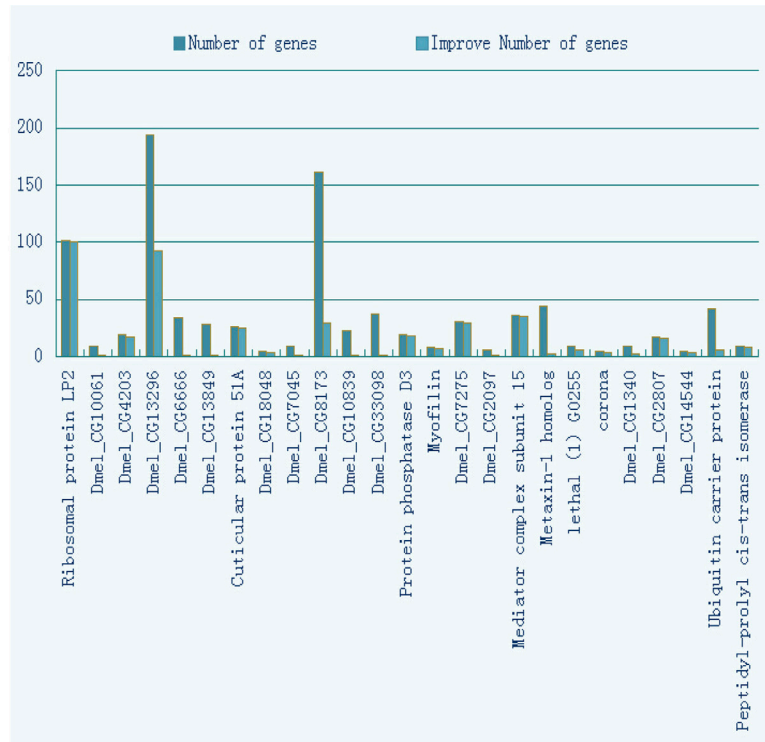


Figure 8. Comparison of numbers of genes between the clustering algorithm and FCSOMs model in the tomato-ethanol group.

The data present in Figure 7 compares the accuracy of the classification between the clustering algorithms in DAVID_6.7 and the FCSOM models in the tomato-ethanol group. The horizontal axis denotes the function clusters arranged left to right, as given by DAVID_6.7. The vertical axis denotes the accuracy. The accuracy improved significantly for 20 function clusters; while the average accuracy increasing from 61.5 to 70.25%. This indicated the effectiveness of FCSOM.

Figure 8 shows the number of genes between the clustering algorithm and the FCSOM model. A total of 468 genes were knocked out, and differentially expressed genes altered from 888 to 420. The former 25 groups did not contain the olfactory and gustatory function clusters. This implied that the use of both the tomato and ethanol media for larval affects the gustatory development of *D. melanogaster* (Inoshita and Tanimura, 2006; Bai et al., 2009; Miller and Carlson, 2010; Adachi et al., 2012; Martin et al., 2013), and the differences between these did not appear to be significant.

These numerical experiments show that the differentially expressed genes were mined successfully using FCSOM. The most important 25 function clusters with differentially expressed genes in each contrasting group were accepted. The analyses revealed the advantages of the new model.

Previous studies utilized conventional clustering algorithms to capture the gene-to-medium relationship. However, the nonlinear complex relationship may not be effectively explored by the clustering membership functions. A new model called FCSOMs was proposed to solve this problem. Each FCSOM was customized to map the data samples to two groups in each functional cluster.

This strategy has increased the generalization capability of FCSOM. The superior performance of FCSOM provides a new approach to combine granular computing and advanced statistical learning algorithms.

SOM is not suitable for use in large datasets because of the large training time complexity. This could be overcome by FCSOM, which had special characteristics, such as use the parallelization of the training tasks for each FCSOM, which could allow data mining in large datasets. The satisfactory experimental results revealed the utility of this new computational model in solving complex classification problems in huge datasets.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors would like to thank all of the researchers who made publicly available the data used in this study. Research supported by the University Basic Research Foundation (#201362031) and the National Natural Science Foundation of China (#61303145).

REFERENCES

- Abe T, Hamano Y and Ikemura T (2014). Visualization of genome signatures of eukaryote genomes by batch-learning self-organizing map with a special emphasis on *Drosophila* genomes. *Biomed. Res. Int.* 2014: 985706.
- Adachi R, Sasaki Y, Morita H, Komai M, et al. (2012). Behavioral analysis of *Drosophila* transformants expressing human taste receptor genes in the gustatory receptor neurons. *J. Neurogenet.* 26: 198-205.
- Altman RB and Raychaudhuri S (2001). Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.* 11: 340-347.
- Avogadri R and Valentini G (2009). Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. *Artif. Intell. Med.* 45: 173-183.
- Bai L, Goldman AL and Carlson JR (2009). Positive and negative regulation of odor receptor gene choice in *Drosophila* by *acj6*. *J. Neurosci.* 29: 12940-12947.
- Ghouila A, Yahia SB, Malouche D, Jmel H, et al. (2009). Application of multi-SOM clustering approach to macrophage gene expression analysis. *Infect. Genet. Evol.* 9: 328-336.
- Gu B, Wang JD, Yu YC, Zheng GS, et al. (2012). Accurate on-line-support vector learning. *Neural Netw.* 27: 51-59.
- Inoshita T and Tanimura T (2006). Cellular identification of water gustatory receptor neurons and their central projection pattern in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 103: 1094-1099.
- Liu J and Xu M (2008). Kernelized fuzzy attribute C-means clustering algorithm. *Fuzzy Set. Sys.* 159: 2428-2445.
- Maraziotis IA (2012). A semi-supervised fuzzy clustering algorithm applied to gene expression data. *Pattern Recogn.* 45: 637-648.
- Martin F, Boto T, Gomez-Diaz C and Alcorta E (2013). Elements of olfactory reception in adult *Drosophila melanogaster*. *Anat. Rec.* 296: 1477-1488.
- Miller CJ and Carlson JR (2010). Regulation of odor receptor genes in trichoid sensilla of the *Drosophila antenna*. *Genetics* 186: 79-95.
- Mozhayskiy V and Tagkopoulos I (2011). *In silico* evolution of multi-scale microbial systems in the presence of mobile genetic elements and horizontal gene transfer. *Int. J. Bioinf. Res. Appl.* 6674: 262-273.
- Mukhopadhyay A and Maulik U (2009). Towards improving fuzzy clustering using support vector machine: Application to gene expression data. *Pattern Recogn.* 42: 2744-2763.
- Patnala R, Clements J and Batra J (2013). Candidate gene association studies: a comprehensive guide to useful *in silico* tools. *BMC Genet.* 14: 39.
- Schulze A and Downward J (2001). Navigating gene expression using microarrays - a technology review. *Nat. Cell Biol.* 3: E190-E195.
- Tang YC, Jin B and Zhang YQ (2005). Granular support vector machines with association rules mining for protein homology

- prediction. In: Computational intelligence techniques in bioinformatics. *Artif. Intell. Med.* 35: 121-134.
- Wolfsberg TG, Gabrielian AE, Camp bell MJ, Cho RJ, et al. (1999). Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Res.* 9: 775-792.
- Wu LF, Hughes TR, Davierwala AP, Robinson MD, et al. (2002). Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31: 255-265.
- Yan Y, Chen L and Tjhi WC (2013). Fuzzy semi-supervised co-clustering for text documents. *Fuzzy Set. Syst.* 215: 74-89.
- Yao YY (2004). Granular computing. In: Computer science. Proceedings of the 4th Chinese National Conference on Rough Sets and Soft Computing (Suan J and Xue JK, eds.). 31: 1-5.
- Yao YY (2005). Perspectives of granular computing. In: Proceedings of 2005 IEEE International Conference on Granular Computing, 85-90.
- Zhong W, He J, Harrison R, Tai PC, et al. (2007). Clustering support vector machines for protein local structure prediction. *Expert Syst. Appl.* 32: 518-526.