Computers in Biology and Medicine 43 (2013) 1120–1133

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Multi-stage filtering for improving confidence level and determining dominant clusters in clustering algorithms of gene expression data

Shahreen Kasim [a,c,*], Safaai Deris [c], Razib M. Othman [b]

[a] Software Multimedia Center, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Malaysia
[b] Laboratory of Computational Intelligence and Biotechnology, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Skudai, Malaysia
[c] Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Skudai, Malaysia

ARTICLE INFO

Article history:
Received 12 December 2010
Accepted 15 May 2013

Keywords:
Confidence level
Dominant cluster
Fuzzy clustering
Gene expression

ABSTRACT

A drastic improvement in the analysis of gene expression has lead to new discoveries in bioinformatics research. In order to analyse the gene expression data, fuzzy clustering algorithms are widely used. However, the resulting analyses from these specific types of algorithms may lead to confusion in hypotheses with regard to the suggestion of dominant function for genes of interest. Besides that, the current fuzzy clustering algorithms do not conduct a thorough analysis of genes with low membership values. Therefore, we present a novel computational framework called the “multi-stage filtering-Clustering Functional Annotation” (msf-CluFA) for clustering gene expression data. The framework consists of four components: fuzzy *c*-means clustering (msf-CluFA-0), achieving dominant cluster (msf-CluFA-1), improving confidence level (msf-CluFA-2) and combination of msf-CluFA-0, msf-CluFA-1 and msf-CluFA-2 (msf-CluFA-3). By employing double filtering in msf-CluFA-1 and *apriori* algorithms in msf-CluFA-2, our new framework is capable of determining the dominant clusters and improving the confidence level of genes with lower membership values by means of which the unknown genes can be predicted.

1. Introduction

Microarray technology has become a significant tool in functional genomics and biomedical research. This technology allows for simultaneous measurement of gene expression for thousands of genes. The resulting data can then be used in various ways, such as in diagnosing tumours [25], drug-effect profiling [35] and identification of genes that contribute to common functions by grouping genes with similar expression of patterns using either clustering or classification techniques [22,28].

In discovering similar patterns in gene expression datasets, clustering has been used extensively and this may lead to an insight into significant connections within the gene regulatory networks. In order to understand the pattern of these genes, many contributions have been made by clinical [23,30], biological [47,38], toxicological [14,27] and pharmacological [37,17] studies. There are many clustering algorithms currently used for clustering gene expression datasets such as the *k*-means [9], hierarchical clustering [48], Self-Organising Maps (SOM) [16], graph theoretical algorithms [43], Genetic Algorithms (GA) [4] and fuzzy *c*-means [26,41]. Since imprecision and uncertainty are considered to be the natural behaviour of gene expression datasets [29], the fuzzy *c*-means clustering has become an appropriate choice [24]. This is due to the ability of fuzzy *c*-means algorithms to cluster genes into more than one group thus providing a systematic, unbiased method to change precise values into several descriptors of cluster membership [32]. Furthermore, the fuzzy *c*-means clustering provides more information on the degree of similarity [11] among genes. Therefore, the fuzzy *c*-means algorithm is applied in the clustering process to partition a given gene expression dataset. However, clustering alone, without paying attention to the coherence of biological functions within the clusters, brings no meaningful results. Coherence can be seen when one gene is involved in multiple biological functions. In order to capture the coherence of biological functions in the clusters, biological knowledge is applied during the clustering process. One of the many popular sources of biological knowledge is the Gene Ontology (GO) [1]. The popularity of GO is due to its capability to provide a standard species-independent controlled vocabulary for describing genes in terms of their biological processes, cellular components and molecular functions. Related research incorporating the GO in gene expression clustering is found in numerous studies [15,34,7,21,39].

However, there are still some other issues associated with the acquisition and analysis of gene expression datasets which can impose a profound influence on the interpretation of the results. One of these problems is the degrading performance of clustering results due to certain situations in which a gene can have multiple

[*] Corresponding author at: Software Multimedia Center, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Malaysia. Tel.: +60 7 453 3776; fax: +60 7 453 6023.
E-mail address: shahreen@uthm.edu.my (S. Kasim).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.05.011

functions. It has been observed that whenever a gene belongs to multiple functions, it will create confusions in choosing the dominant function of that particular gene. Another issue that frequently occurs is the ability to assign with high confidence genes that have low membership values in the cluster. This issue arises when some of the genes are considered as being clustered with high confidence due to their high membership values. Concurrently, some other genes that belong to the same cluster but have lower membership values are assigned with low confidence. This low confidence resulted from the presence of genes that are near the cluster border line or are slightly far from the centroid.

Based upon the observations stated above, we propose for an improved fuzzy *c*-means algorithm named the msf-CluFA (*m*ulti-*s*tage *f*iltering–*Clu*stering *F*unctional *A*nnotation), as shown in Fig. 1, which not only takes into account the observation of genes with high membership values, but also handles genes with low membership values with more serious consideration. Our msf-CluFA has three main stages, referred to as the fuzzy *c*-means clustering stage (msf-CluFA-0), the achieving dominant cluster stage (msf-CluFA-1) and the improving confidence level stage (msf-CluFA-2). In msf-CluFA-0, the enhancement of traditional fuzzy *c*-means took place as the GO and Saccharomyces Genome Database (SGD: [10]) functional annotation databases were incorporated into the fuzzy *c*-means algorithm. Next, there were two filtering stages in which the genes with high membership values were initially filtered into the first stage filtering (msf-CluFA-1). At this stage, genes were assigned into their dominant cluster by filtering based on the membership values and degree of specificity. Concurrently, the genes with low membership values were filtered into the second stage filtering (msf-CluFA-2) with the intention of improving their confidence level and ability to be included in the cluster with confidence. This was done by applying an *apriori* algorithm [3] to detect the co-occurrences of annotations. From the co-occurrences, the genes were then filtered based on the ranking system. The details of our msf-CluFA are explained in the next section. The difference of ms-CLuFA compared to the method proposed by Bandyopadhyay et al. [4] is that their work used the genetic algorithm in the first stage clustering. Furthermore, a multi-objective genetic clustering together with the nearest neighbour criterion has been used in their second stage clustering.

2. Materials and methods

2.1. Datasets

The yeast gene expression datasets from Eisen et al. [13] and Gasch et al. [19] were used in order to test the new framework of our msf-CluFA. There were 6221 expression profiles corresponding to four experiments on cell cycle, sporulation, temperature shock and diauxic shift processes in the Eisen dataset. On the other hand, in the Gasch dataset, there were 6152 expression profiles gathered over 173 of various experiments tested.

In order to cluster and identify the similarity between our msf-CluFA and the GO, we downloaded the GO slim yeast data and GO terms data from an updated version from September 2005. There were 76 terms in the GO slim data and 19,458 terms in the MySQL GO term data. The SGD compiled in September 2005,

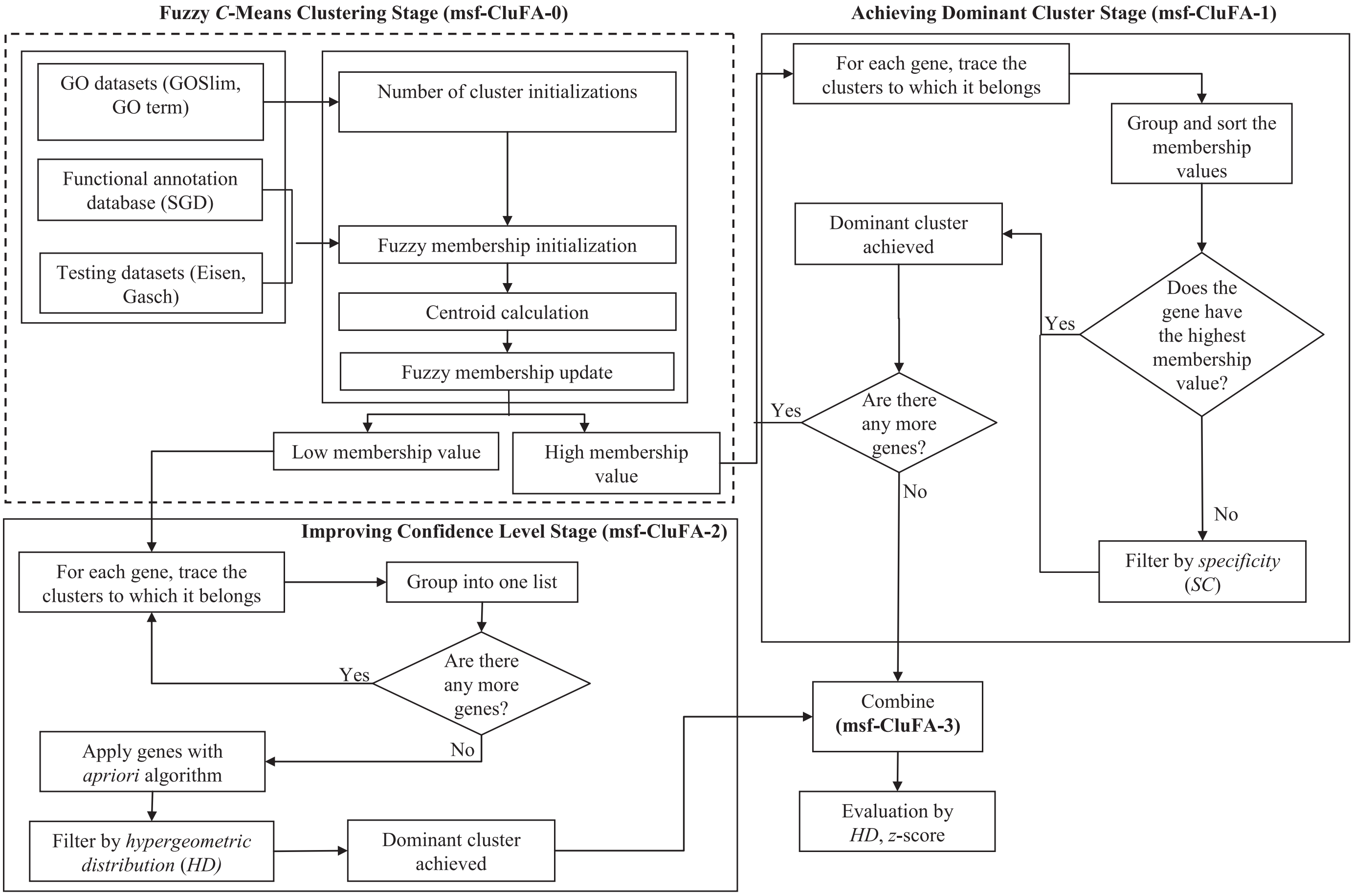
Fuzzy *C*-Means Clustering Stage (msf-CluFA-0)
GO datasets (GOSlim, GO term)
Functional annotation database (SGD)
Testing datasets (Eisen, Gasch)
Number of cluster initializations
Fuzzy membership initialization
Centroid calculation
Fuzzy membership update
Low membership value
High membership value
Achieving Dominant Cluster Stage (msf-CluFA-1)
For each gene, trace the clusters to which it belongs
Group and sort the membership values
Does the gene have the highest membership value?
Yes
No
Dominant cluster achieved
Filter by *specificity* (*SC*)
Are there any more genes?
Yes
No
Improving Confidence Level Stage (msf-CluFA-2)
For each gene, trace the clusters to which it belongs
Group into one list
Are there any more genes?
Yes
No
Apply genes with *apriori* algorithm
Filter by *hypergeometric distribution* (*HD*)
Dominant cluster achieved
Combine (msf-CluFA-3)
Evaluation by *HD*, *z*-score

Fig. 1. The flowchart of msf-CluFA.