Fuzzy J-Means and VNS Methods for Clustering Genes from Microarray Data

Belacel, Nabil; Cuperlovic-Culf, Miroslava; Laflamme, Mark; Ouellette, Rodney

Publisher's version / Version de l'éditeur:
https://doi.org/10.1093/bioinformatics/bth142
Bioinformatics, 20, 11, pp. 1690-1701, 2004-02-26

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Fuzzy J-Means and VNS Methods for Clustering Genes from Microarray Data *

Belacel, N., Cuperlovic-Culf, M., Laflamme, M., and Ouellette, R.
March 2004

* published in the International Journal of Bioinformatics, Oxford University Press. March 2004. NRC 46546.

Canada

Fuzzy J-Means and VNS methods for clustering genes from microarray data

Nabil Belacel[1‡], Miroslava Čuperlović-Culf[2‡*], Mark Laflamme[2], Rodney Ouellette[2]

‡ contributed equally to this work

1.National Research Council Canada, Institute for Information Technology-e-Health

2.Institut de recherche médicale Beauséjour

* Contact information:

Miroslava Cuperlovic-Culf
Institut de recherche médicale Beauséjour,
Hotel-Dieu Pavilion 35 Providence
Moncton, NB E1C 8X3
Telephone(506) 862-7572
Fax(506) 862-7571
e-mail: miroslavac@health.nb.ca

Abstract

Motivation: In the interpretation of gene expression data from a group of microarray experiments that include samples from either different patients or conditions, special consideration must be given to the pleiotropic and epistatic roles of genes, as observed in the variation of gene co-expression patterns. Crisp clustering methods assign each gene to one cluster, thereby omitting information about the multiple roles of genes.

Results: Here we present the application of a local search heuristic, Fuzzy J-Means, embedded into the Variable Neighborhood Search metaheuristic for the clustering of microarray gene expression data. We show that for all data sets studied this algorithm outperforms the standard Fuzzy C-Means heuristic. Different methods for the utilization of cluster membership information in determining gene co-regulation are presented. The clustering and data analyses were performed on simulated data sets as well as experimental cDNA microarray data for breast cancer and human blood from the Stanford Microarray Database.

Availability: The source code of the clustering software (C programming language) is freely available from Nabil.Belacel@nrc-cnrc.gc.ca

Contact: Miroslava Cuperlovic-Culf, e-mail: miroslavac@health.nb.ca

1 Introduction

The adaptability of cells and the diversity in cellular responses to various internal and external stimuli is accomplished through the co-operation and multifunctionality of a limited number of proteins. Depending on the cellular environment, groups of genes are often co-expressed and each group is regulated by a specific mechanism that depends on the particular cellular condition. Information regarding gene co-expression and co-operation should theoretically be accessible from various expression profiling assays, such as microarrays. Microarrays provide huge data sets that are currently primarily analyzed using various crisp clustering

techniques. While these classical algorithms can accurately identify distinct expression patterns by grouping genes with similar expression behavior, they are unable to identify genes whose expression levels are similar to multiple, distinct groups of genes, thereby hiding any information about the inter-relatedness of genes. In addition, when analyzing large gene-expression data sets collected under various conditions, where genes are likely to be co-expressed with different groups of genes under different conditions, crisp clustering methods may result in inaccurate clusters, therefore leading to incorrect conclusions about gene product behavior (Gasch and Eisen, 2002).

A number of methods have been developed to deal with the complex relationships between objects (Friedman et al. 2000, Ihmels et al. 2002, Sheng et al. 2003). Alternatives to the crisp clustering methods include the fuzzy clustering methods, which provide a systematic and unbiased way to change precise values into several descriptors of cluster memberships (Bezdek, 1981). In other words, fuzzy logic methods uncover information about the relative likelihood of each gene belonging to each of a predefined number of clusters, thus providing information regarding gene multifunctionality. In addition, fuzzy logic methods inherently account for noise in the data because they extract trends rather than the precise values

(Woolf and Wang 2000).

A fuzzy logic method, which had recently been introduced to microarray data analysis, (Dougherty *et al.* 2002; Woolf and Wang 2000; Gasch and Eisen, 2002; Dembele and Kastner, 2003), has been shown to reveal additional information concerning gene co-expression. In particular, information regarding overlapping clusters and overlapping cellular pathways has been identified from fuzzy clustering results (Gasch and Eisen, 2002). The method of choice in all applications up-to-date has been the Fuzzy C-Means algorithm (F-CM). F-CM is the fuzzy logic extension of the K-Means heuristic used for crisp clustering. The F-CM method searches for the membership degrees and centroids until there is no further improvement in the objective function value, thereby risking the possibility of remaining in a local minimum of a poor value.

An alternative fuzzy clustering method called Fuzzy J-Means (F-JM) has been recently developed (Belacel, *et al.* 2002). The F-JM method was inspired by the local search heuristic J-Means, developed for solving the minimum sum-of-squares clustering problem (Hansen and Mladenovic, 2001). J-Means has already been proven to be superior to the standard K-Means method, especially for the clustering of large data sets (Hansen and Mladenovic, 2001). In J-Means and F-JM methods, centroid moves belong to the neighborhood of the current solution

defined by all possible centroid-to-pattern relocations. In F-JM, the "integer" solution is moved to the continuous one by finding centroids and membership degrees for all patterns and clusters. Like F-CM, the F-JM is a local heuristic and can therefore determine only the closest, possibly non-optimal solution. Thus, F-JM heuristic is embedded into the Variable Neighborhood Search metaheuristic (VNS), which searches for distant, possibly more appropriate cluster arrangements (Hansen and Mladenovic, 1997). In this study the applicability of F-JM and VNS methods for microarray data analysis was investigated for the first time. The accuracy of clusters obtained using these methods in both simulated and experimental microarray data sets is compared to the results obtained using F-CM method.

2 Materials and methods

2.1 Data sets

2.1.1. *Simulated data.* Three simulated sets built around nine distinct temporal patterns over ten sample points were considered. The first set (SD/450/10) consisted of 450 genes separated in nine families with 50 genes in each. The expression level values were generated by adding independent random noise to the nine different median expression values. The second synthetic set (SD1/90/10)

consisted of data with a much larger overlap between groups. It was developed with 90 genes separated in nine groups with values determined by multiplying random integer value from the set range by the random noise making it a much harder set for classification. The third set (SD2/90/10) was based on the TIGR simulated set (Quackenbush, 2001). As in the first set the data was generated by adding independent random noise to the median expression values but with larger variations in the median expression value over different sample points. The mean expression levels with average standard deviation for all groups in all three sets are shown in Figure 1.

*** FIGURE 1 ABOUT HERE ***

2.1.2. *Human breast cancer data*. The complete breast cancer data set available from the Stanford Microarray Database, was contributed and described in detail by Sorlie *et al.* (2001). The total set contains gene expression levels information for 8102 genes measured in 85 human tissue samples (including ductal and lobular cancers, ductal carcinomas *in situ* as well as normal breast tissue samples from different individuals). For this work we selected two subsets, one with a relatively small and one with a relatively large number of genes in comparison to the number of samples. In both subsets we chose only genes that did not have missing data in any of the experiments. The first set (BC/69/85) includes a small

subsection of genes (transcription factors and kinases). The second set (BC/1022/85) includes all genes from the original set, which did not have any empty (missing) data in all the experiments.

2.1.3.*Human blood data*. This data set, also downloaded from the Stanford Microarray Database, was contributed and described by Whitney, et al. (2003). Human blood data set comprises gene expression patterns of approximately 18,000 genes measured in blood samples of 82 healthy donors (including data for total RNA from the whole blood and for the peripheral blood mononuclear cells for a total of 147 experiments). As for the breast cancer set we again selected two subsets from this set, one with relatively small and one with relatively large number of genes in comparison to the number of samples. In both subsets we chose only genes that did not have missing data in any of the experiments. The first set (HB/43/147) includes a small subsection of genes (transcription factors and kinases). The second set (HB/2197/147) includes all genes from the set without missing data points in all experiments.

The expression values are represented as experiment mean normalized ratios of expression levels.

2.2 Algorithms

The crisp clustering methods assign each object (gene) to one cluster only. In fuzzy clustering methods, an indicator variable showing whether an object is a member of a given group/cluster is extended to a weighting factor called membership (w). The membership has values ranging from 0 to 1, where membership values close to 1 indicate strong association to the cluster, and values close to 0 indicate weak or absent association to the cluster. The goal of fuzzy clustering of genes is to assign a gene, according to the results from several experiments, to a given number of clusters, such that any gene may belong to more than one cluster, with different degrees of membership.

2.2.1. Fuzzy C-Means clustering method

F-CM is a fuzzy logic extension of the classic, crisp, K-Means method (Bezdek 1981, Dunn 1974, Ruspini 1969). In terms of classification, the results of a microarray experiment can be presented in terms of an $n \times N$ matrix where n is the number of genes and N is the number of experiments.

$$\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \dots \\ \boldsymbol{X}_n \end{bmatrix} = \begin{bmatrix} X_{11}\, X_{12} \dots\dots\dots X_{1N} \\ X_{21}\, X_{22} \dots\dots\dots X_{2N} \\ \dots\dots\dots \\ X_{n1}\, X_{n2} \dots\dots\dots X_{nN} \end{bmatrix} \quad (1)$$

Here, each x_{ij} represents the background subtracted, normalized,

expression level or $\log_2$ of the expression level (absolute or relative depending on the type of experiment) of a gene $i = 1,...,n$, in experiment $j = 1,...,N$.

Then, for a chosen number of clusters, c, and for an $n \times c$ matrix $W = [w_{ik}]$, where w_{ik} is the membership degree for gene i, $i = 1,...,n$ to cluster k, $k = 1,...,c$, the F-CM clustering problem can be represented as:

$$(\min_{\boldsymbol{W},\boldsymbol{V}}) J_m(W,V) = \sum_{i=1}^{n}\sum_{k=1}^{c} w_{ik}^{m} \|\boldsymbol{x}_i - \boldsymbol{v}_k\|^2 \quad (2)$$

where:

- $J_m(W,V)$ is the objectivity function defining the quality of the result obtained for centroids V and memberships W;
- m is the fuzziness parameter which regulates the degree of fuzziness in the clustering process; for m=1 the problem is the classical minimum sum of squares clustering and the partition is crisp;
- $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_c] = \begin{bmatrix} V_{11}\, V_{12} ..... V_{1c} \\ V_{21}\, V_{22} ..... V_{2c} \\ ..... \\ V_{N1}\, V_{N2} ..... V_{Nc} \end{bmatrix}$ gives a set of c centroids or prototypes, *i.e.* positions of cluster centres;
- $\|\boldsymbol{x}_i - \boldsymbol{v}_k\|^2 = \langle \boldsymbol{x}_i - \boldsymbol{v}_k | \boldsymbol{x}_i - \boldsymbol{v}_k \rangle = \sum_{j=1}^{N} |X_{ij} - V_{jk}|^2$ is the Euclidean norm determining distances between expression level vectors and centroids;
- membership degrees w_{ij} are defined such that: $0 \leq w_{ik} \leq 1$ and

$$\sum_{k=1}^{c} w_{ik} = 1 \; \forall \; i = 1, ..., n \quad .$$

The algorithm describing the F-CM procedure is briefly outlined in Figure 2a. Detailed equations for the calculation of membership factors and centroids are explained elsewhere (Belacel *et al.* 2002; Dembele and Kastner 2003).

2.2.2. Fuzzy J-Means clustering method

As the fuzziness parameter $m = 1$ defines a crisp clustering, the m parameter for fuzzy logic applications has to be $m > 1$. Equation (2) can be therefore reformulated to (Hathaway and Bezdek1995):

$$(\min_{\boldsymbol{V}}) R_m(\boldsymbol{V}) = \sum_{i=1}^{n} \left[\sum_{k=1}^{c} \| \boldsymbol{x}_i - \boldsymbol{v}_k \|^{2(1-m)} \right]^{(1-m)} \qquad (3)$$

where $R_m(V)$ is the reformulated objectivity function dependent only on the centroid positions. Therefore, centroid positions can be obtained directly by minimizing the equation (3). The obtained centroids can be used to calculate the membership values, and the results can subsequently be iteratively improved. The F-JM method recently introduced by Belacel *et al.* (2002) uses all possible centroid-to-pattern relocations in order to construct move-defined neighborhoods. The algorithm describing F-JM is briefly outlined in Figure 2b. Membership values and centroids are calculated in the same

way as in F-CM (Belacel *et al.* 2002).

2.2.3. Variable Neighborhood Search Method

Both F-CM and F-JM are local heuristics, *i.e.* they search only for the clustering solution closest to the starting centroid values. Application of these methods can not guarantee that the final result is the overall optimal clustering solution, even when using several different starting points. When fuzzy methods are used on large data sets, and with a large number of clusters as is characteristic of microarray applications, it is possibile to obtain only the closest solution instead of the global one, ideally or at least an improved, more distant local one. This problem is alleviated by using the VNS method. The VNS is a previously developed metaheuristic for solving combinatorial and global optimization problems (Hansen and Mladenovic 1997). The basic goal of the method is to proceed to a systematic change of neighborhood within a local search algorithm. The algorithm remains in the same locally optimal solution exploring increasingly distant neighborhoods by random generation of a point and descent, until another solution, better than the incumbent, is found. The algorithm then jumps to the new solution and continues the search from there. The neighborhood centroid structures are obtained by replacing, at random, a predetermined number k of existing

centroids of clusters with k randomly chosen patterns, *i.e.* genes. The set of neighborhood structures is denoted $N_k, (k = 1, ..., k_{max})$ and the set of solutions forming neighborhood N_k of a current centroid solution V is $N_k(V)$. A brief algorithmic description of the VNS procedure is given in Figure 2c. The stopping criterion may be set either to the maximum CPU time or a maximum number of iterations allowed.

*** FIGURE 2 ABOUT HERE ***

3 Results and Discussion

Initially, the optimal fuzziness parameter (m) was determined, and this value was used to compare the objectivity factors for clusters obtained using F-CM, F-JM and VNS methods; this was repeated for all four data sets and for various number of clusters. Finally, we investigated several procedures for the determination of genetic properties of samples from membership values.

3.1 Determination of the fuzziness parameter

Recent work by Dembele and Kastner (2003) has shown that it is not appropriate for the fuzziness parameter, m, to be set to a typical value of 2 when the F-CM method is applied to microarray data

analysis. Similarly to the work of Dembele and Kastner (2003), for data sets tested here *m* values close to 2 resulted in membership values of: $w_{ik} \to \frac{1}{c}, \forall i = 1,\ldots, n; \forall k = 1,\ldots, c$, thus failing to extract any useful clustering information. Thus, an initial stage in any further application of fuzzy methods is the determination of an optimal value of *m* for the studied data sets. The box plot representations of the memberships values for each gene, in the decreasing order for several values of *m* are shown in Figure 3a (breast cancer data sets) and Figure 3b (human blood data sets).

**** FIGURE 3 ABOUT HERE *****

Figure 3 shows that the fuzziness obtained for a given value of *m* depends strongly on the type of data and less strongly on the number of genes in the data set. From membership values for all four data sets and all investigated values of *m,* we calculated median of the top membership values for all genes, μ_T, and the overall median membership value μ_d (Table 1).

The distribution of the top two membership values is further observed using the scatter plots of the two largest memberships for each gene for the same *m* values (Figure 4).

***** FIGURE 4 ABOUT HERE ******

Our results show that, for *m* value of 1.15 (or less) in all

studied data sets, all genes have very high top membership values and thereby high top membership median values, and obtained clusters are thus almost crisp (Figures 3,4; Table 1). Also, for all data sets, *m* values greater than 1.75 result in the overall median approaching 1/c (data not shown), again resulting in the loss of any useful information about membership values. Therefore, the optimal fuzziness parameter for all data sets is greater than 1.15 and smaller than 1.75. The more precise value of *m* was estimated empirically from Figure 3,4 and Table 1. Our empirical rules for the determination of optimal *m* were:

1. median of the top membership values greater than or equal to 0.5 (prevents the results from being overly fuzzy);
2. median of all membership values greater than 0 (prevents the results from becoming crisp).

The optimal *m* values determined in this way are different for each data set (Table 1). Thus, for any application of fuzzy methods in microarray data analysis for different samples or different data set sizes the *m* factor has to be determined independently. But, since membership values do not represent absolute probabilities of a gene belonging to a cluster but rather a relative membership in a cluster with respect to other clusters, minor errors caused by the empirical nature of the method should not cause any errors in the final

application of membership values.

3.2. Comparative Analysis of Fuzzy C-Means, Fuzzy J-Means and VNS Methods

The J-Means and VNS methods have been shown to give accurate results for data sets of any size. For large data sets (n>1000), the VNS method has been shown to have an average error over 60 times smaller than K-Means (Hansen and Mladenovic, 2001). From the results on crisp methods, fuzzy logic methods based on J-Means and VNS can be expected to outperform F-CM. The three heuristics: F-CM, F-JM and VNS, were compared using the three simulated data sets as well as the four experimental data sets with m=1.25. Here we compare for the three methods the objectivity function R. For simulated sets we also compared the Jaccard coefficients, which represent the quality of determined clusters in comparison to the correct results (Everitt, 1993). Relative values of objectivity function R determined for three methods are independent of the membership values. Therefore differences of the optimal m value among some data sets from the used value of 1.25 are irrelevant for the comparison method used here. All heuristics were coded in C and run on a DELL Latitude c840, Pentium 4 computer with CPU = 1.60 GHZ and 261.56KB RAM. These codes were compiled using an optimizing option (C++ -O4). For the determination

of the F-CM cut off point (Figure 2a) we used the constant $\epsilon = 0.001$. Table 2 summarizes the results of the comparison of the three methods for a different number of clusters with random, equal initial solutions, for a maximum of 1000 iterations for all F-JM heuristic. The best known solution was determined as the lowest obtained objectivity function R_{best}, which represents the best centroid positions obtained using F-CM, F-JM or VNS. Columns 3-5 in Table 2 show the percent deviation of the objectivity function R for the methods in comparison to R_{best} calculated as: $[(R - R_{best})/R_{best}] \times 100$. Finally, the last three columns show the computer time needed to obtain the best solution of the heuristic.

Several conclusions can be drawn from Table 2. VNS and F-JM have, in all instances, a better performance than F-CM. Overall the VNS is the best method. Several instances in the largest data sets when the VNS objectivity function is slightly larger than the F-JM are caused by the small number of allowed loops (1000). In most cases, F-CM is the fastest method but the quality of the results worsened for larger data sets and larger number of clusters. Therefore, F-CM is not the most appropriate fuzzy method for microarray classification. Even though the VNS method takes the longest time to classify data, the greater accuracy of the obtained results makes it an ideal method for large data sets with many

clusters, especially for the application of microarray methodology in pathway and gene multifunctionality analysis.

Simulated data

For the simulated sets the exact number of clusters was known (k=9). For the set SD/450/10 VNS and F-JM were able to pick all nine patterns perfectly without any fuzzines in the cluster memberships, which is expected due to the large separation of mean expression levels between groups in comparison to their standard deviation (Figure 1). For the set SD1/90/10 membership values obtained show the large overlaps in the data. For SD2/90/10 set VNS and F-JM resulted in perfect clusters unlike F-CM. In addition to the objectivity functions comparison of clustering results for simulated data, where the correct partitions are known, it was possible to perform comparison of accuracy of F-CM, F-JM and VNS methods using Jaccard coefficients. Jaccard coefficients are one of the standard ways for the comparison of two partitions (Jain and Dubes, 1988; Yeung, et al. 2000; Dudoit and Fridlyand, 2002). The procedure used for determination of Jaccard coefficients as well as their values for all three simulated data sets and for F-CM, F-JM and VNS results are shown in Figure 5.

***** FIGURE 5 ABOUT HERE ******

The Jaccard coefficients show that clusters obtained using the F-JM

and VNS method are in significantly better agreement with correct clusters than the ones obtained using F-CM even for these relatively small data sets.

3.3. Final Cluster Results

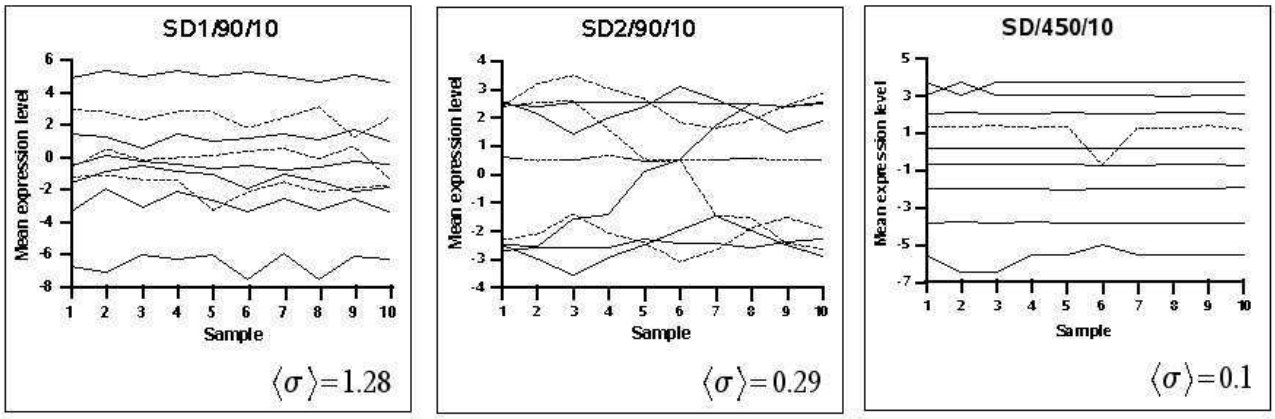
One of the most significant advantages of fuzzy clustering methods is that genes can belong to more than one group revealing distinct aspects of their functionality and regulation (Gasch and Eisen, 2002). The membership values obtained from the fuzzy methods can be applied on several levels. On the basic level, the top membership values can be used to assign each gene to one cluster, resulting in a solution equivalent to the crisp clustering methods. Also, using membership values, it is possible to identify genes most tightly assigned to one cluster and therefore most likely to be part of only one pathway in all studied cases. A good cut-off point for the tightly clustered genes is the median of the top membership values as recently suggested (Dembele and Kastner, 2003). By decreasing the membership cut-off point and by looking at all membership values, rather than just the highest ones, an increasing number of genes will be assigned to each cluster. These new clusters will highlight genes that are members of several groups or pathways in the studied examples, thereby extracting information about

multifunctional genes (Gasch and Eisen 2002). In addition, some groups of genes can be expected to be involved in several pathways and thus assigned to several clusters but still be co-expressed and co-acting under all conditions. For those co-acting genes one can expect to find a similar pattern of membership values. Finally, if the focus is more on the genetic behavior, the membership values can be used to identify genes that are tightly clustered to one, two or several groups. All of these methods are briefly illustrated with the BC/1022/85 breast cancer data set classified using 10 clusters and with $m=1.25$. The results of these analyses are shown in Figures 6 and 7.

***** FIGURE 6 ABOUT HERE ******

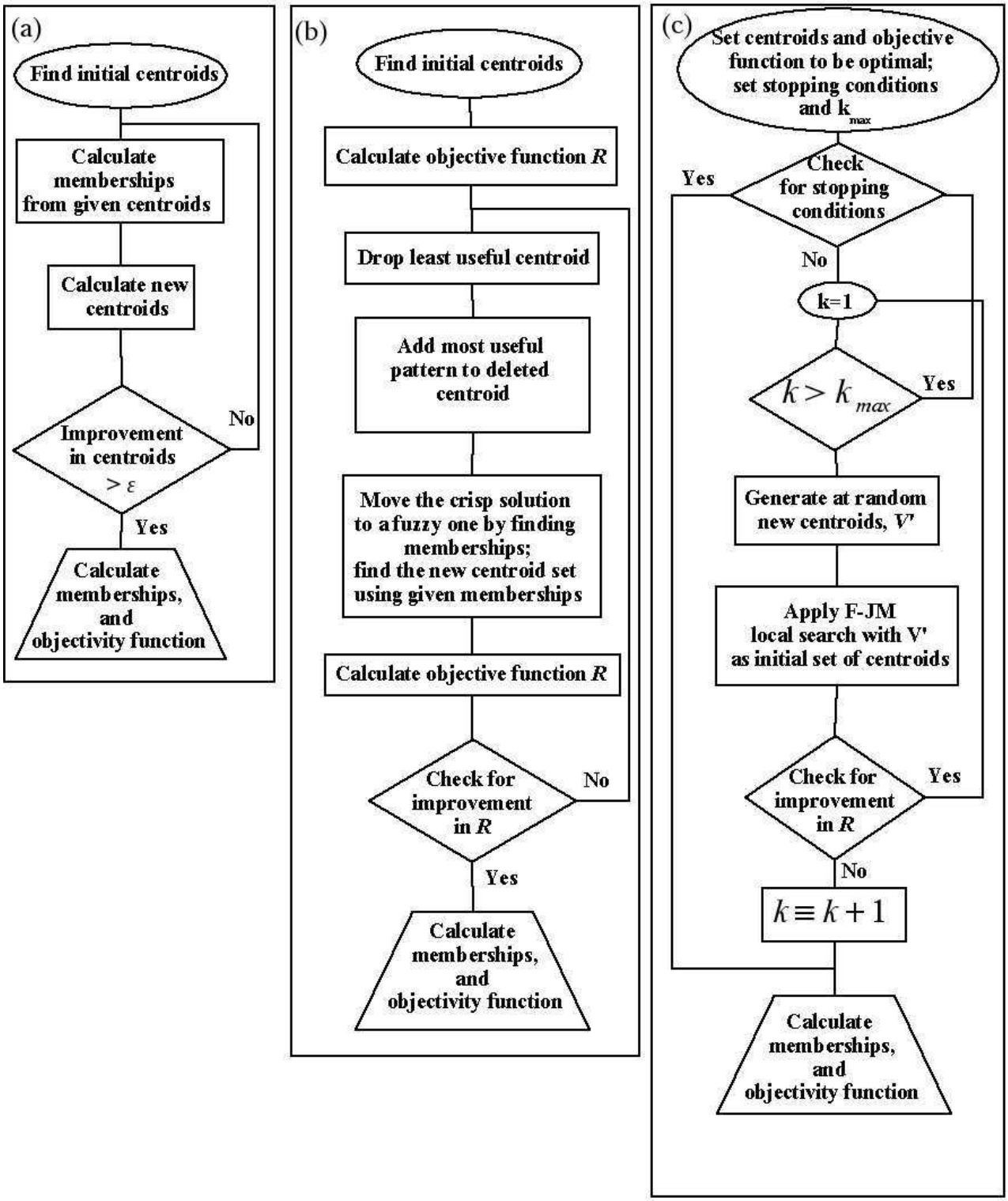
The clustering arrangement based on the highest membership values, where each gene is assigned to only one cluster is shown in Figure 6a. We subsequently selected a sub-section of the most tightly clustered genes (Figure 6b). The clusters obtained for membership values greater than the top membership median value (0.5 in the shown example) are presented. The tight clusters reveal the genes that are most likely to be involved in the same cellular pathways and processes in all experiments. There are 505 tightly clustered genes in the breast cancer data set BC/1022/85. In Figure 6c, we show a subsection of genes with second highest membership value that is

greater than the median value (0.2619). This family of genes have similar top two membership values due to the parallel involvement with two groups of genes. Also, we can separate genes which are assigned to more than two groups with a similar relative probability (Figure 6d). These genes are expected to display a more diverse behavior. Finally, using membership values, it is possible to determine which genes are always co-expressed and co-clustered, *i.e.* which genes work together even in different pathways. This information will be crucial in determining proteins which interact strongly and function only in collaboration. Figure 7 lists examples of genes coding subunits of the same final protein product. For these genes, one expects that they will be co-expressed under all conditions and in all samples, since the final protein can function only if it includes all subunit parts. Indeed, the examples listed in Figure 7 confirm this hypothesis, especially since most of the genes shown were weakly associated to several clusters. In most examples the subunit genes were co-clustered on all membership value levels and followed the same order of membership values. The experimental results in Figure 7 show that the deviations in the clustering patterns of some of the subunit genes relative to their complementary genes (PSMD5, SDHC, pointed with red arrow in Figure 7) or to the other spot for the same gene (CCT4, PSMD11, pointed with blue arrow



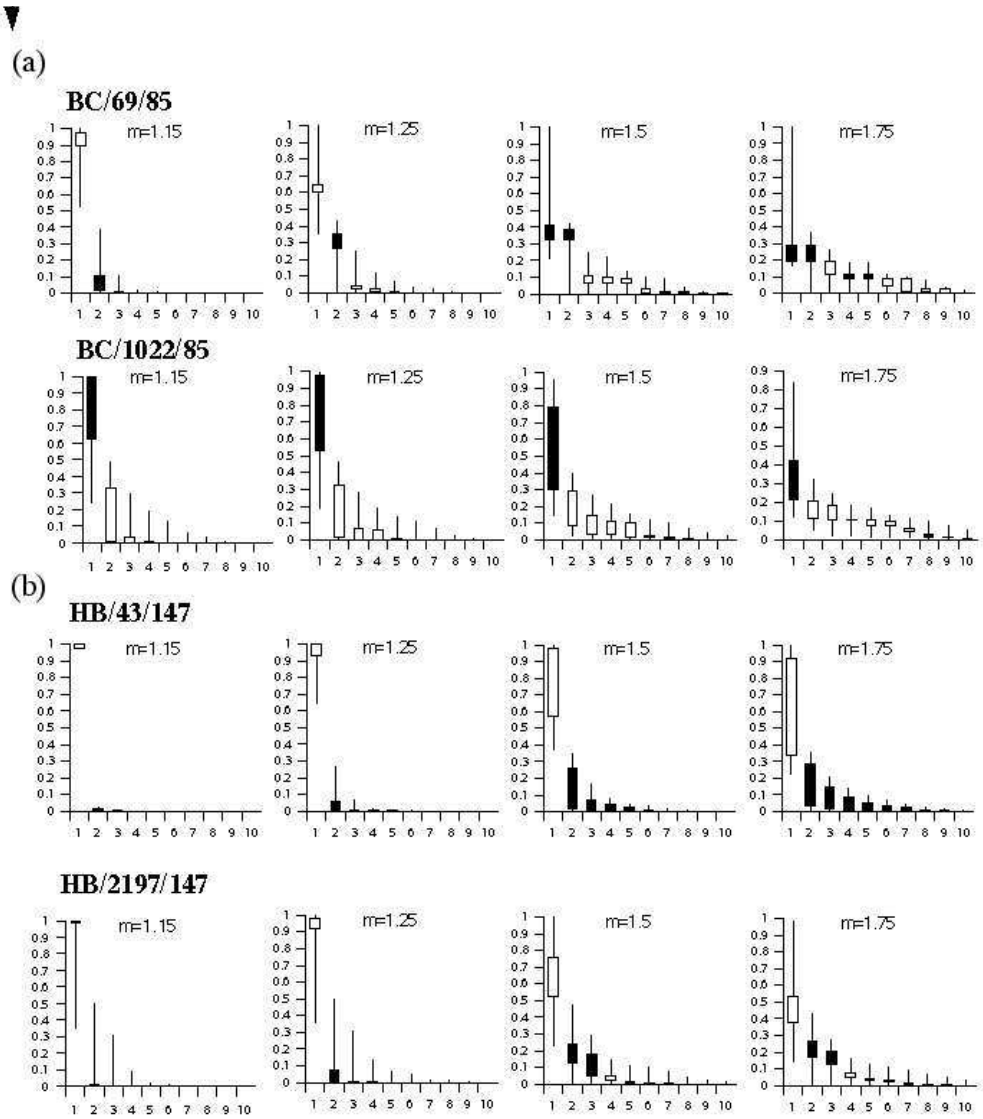
Belacel et al. Fuzzy J-Means and VNS methods for clustering genes from microarray data

Figure 1



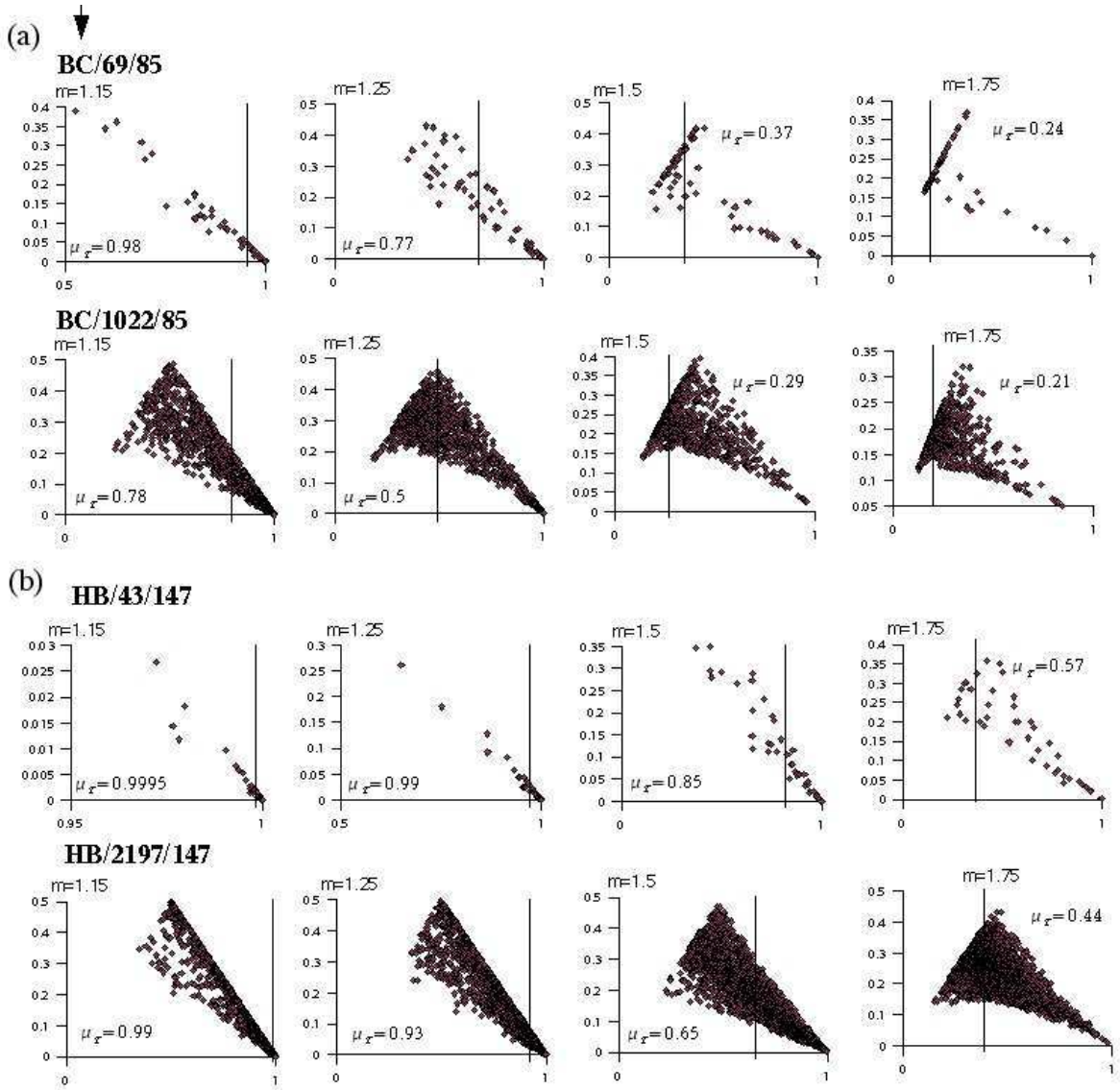
Belacel et al.Fuzzy J-Means and VNS methods for clustering genes from microarray data

Figure 2



Belacel et al.Fuzzy J-Means and VNS methods for clustering genes from microarray data

Figure 3



Belacel et al.Fuzzy J-Means and VNS methods for clustering genes from microarray data

Figure 4

(a)

	S_1	S_2	...	S_N	
C_1	n_{11}	n_{12}	...	n_{1N}	n_{1o}
C_2	n_{21}	n_{22}	...	n_{2N}	n_{2o}
...	...	...	...	...	...
C_N	n_{N1}	n_{N2}	...	n_{NN}	n_{No}
	n_{o1}	n_{o2}	...	n_{oN}	

(b)

$$n_{io} = \sum_{j=1} n_{ij} \quad n_{oj} = \sum_{i=1} n_{ij}$$

$$Z = \sum_{i=1} \sum_{j=1} n_{ij}^2$$

$$n = \sum_{i=1} \sum_{j=1} n_{ij}$$

$$Jaccard = \frac{Z - n}{\sum_{i=1} n_{io}^2 + \sum_{j=1} n_{oj}^2 - Z - n}$$

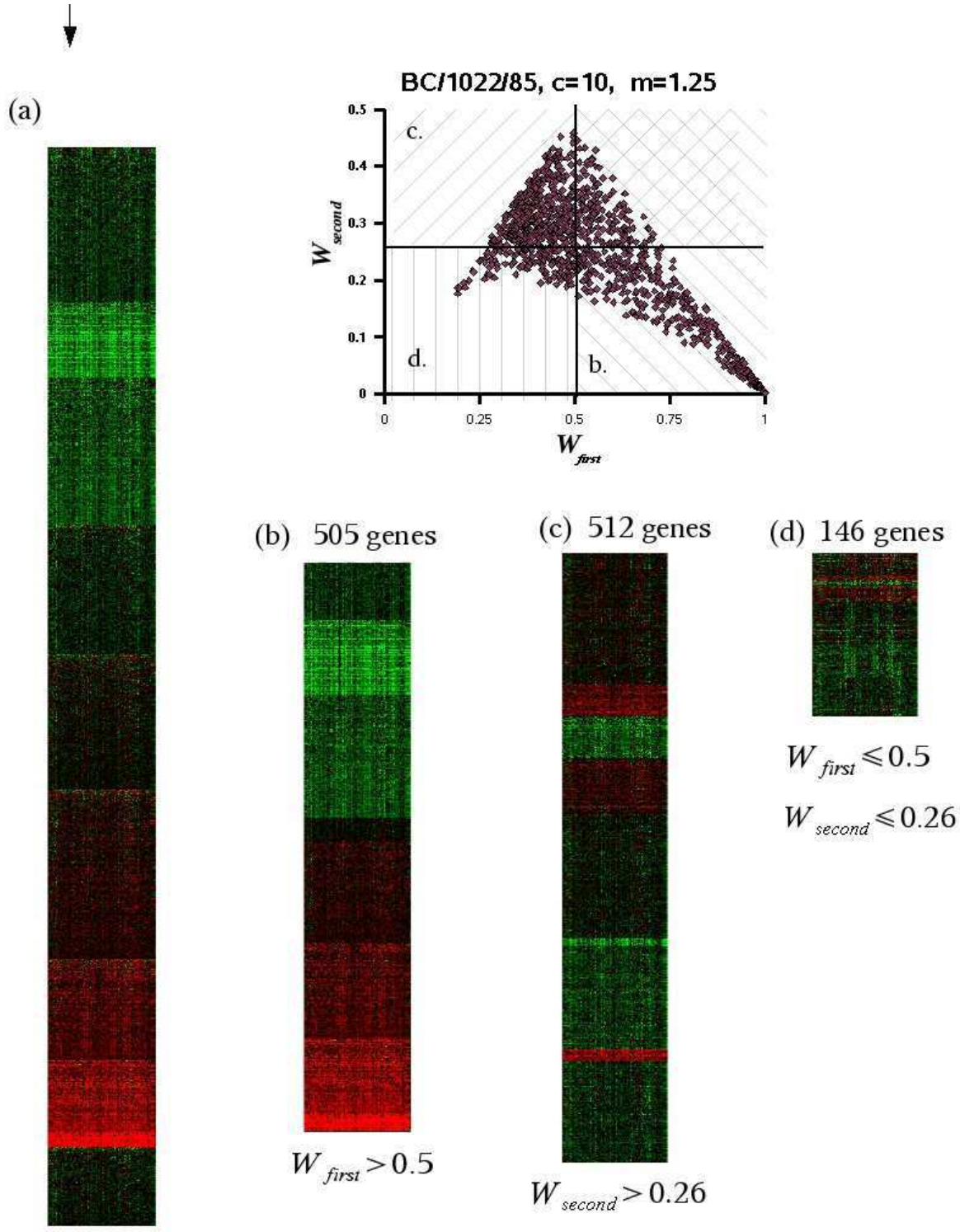
(c)

SD/450/10			SD1/90/10			SD2/90/10		
F-CM	F-JM	VNS	F-CM	F-JM	VNS	F-CM	F-JM	VNS
0.61	1	1	0.35	0.35	0.36	0.75	1	1

Belacel et al. Fuzzy J-Means and VNS methods for clustering genes from microarray data

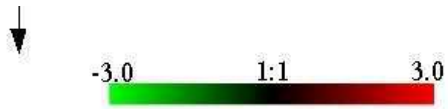
Figure 5

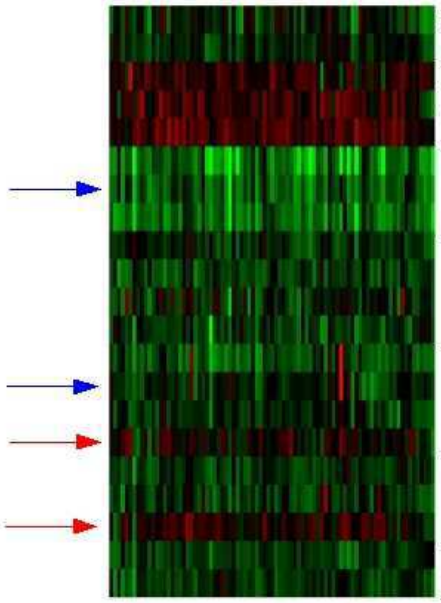
in Figure 7), result from differences in experimentally-determined relative expression levels and not from errors in the method.



Belacel et al.Fuzzy J-Means and VNS methods for clustering genes from microarray data

Figure 6





Gene	Clusters
ARPC1B || actin related protein 2/3 complex, subunit 1B, 41kDa	10, 7, 5
ARPC2 || actin related protein 2/3 complex, subunit 2, 34kDa	10, 7, 5
ATP6V0A1 || ATPase, H+ transporting, lysosomal V0 subunit a isoform 1	4, 1, 2
ATP6V0B || ATPase, H+ transporting, lysosomal 21kDa, V0 subunit c''	4, 2, 1
ATP6V0C || ATPase, H+ transporting, lysosomal 16kDa, V0 subunit c	2, 4
CCT4 || chaperonin containing TCP1, subunit 4 (delta)	3,
CCT4 || chaperonin containing TCP1, subunit 4 (delta)	6, 3
CCT7 || chaperonin containing TCP1, subunit 7 (eta)	3,
EIF3S1 || eukaryotic translation initiation factor 3, subunit 1 alpha, 35kDa	7, 10
EIF3S4 || eukaryotic translation initiation factor 3, subunit 4 delta, 44kDa	7, 10, 3
EIF3S6 || eukaryotic translation initiation factor 3, subunit 6 48kDa	10, 7, 5
EIF3S6IP || eukaryotic translation initiation factor 3, subunit 6 interacting protein	7, 10, 3
PSMD11 || proteasome (prosome, macropain) 26S subunit, non-ATPase, 11	10, 5, 7, 1
PSMD11 || proteasome (prosome, macropain) 26S subunit, non-ATPase, 11	3, 7, 10
PSMD2 || proteasome (prosome, macropain) 26S subunit, non-ATPase, 2	10, 7
PSMD5 || proteasome (prosome, macropain) 26S subunit, non-ATPase, 5	1, 5
SDHA || succinate dehydrogenase complex, subunit A, flavoprotein (Fp)	7, 10
SDHB || succinate dehydrogenase complex, subunit B, iron sulfur (Ip)	7, 10
SDHC || succinate dehydrogenase complex, subunit C, integral membrane protein, 15kDa	4, 1
SF3A1 || splicing factor 3a, subunit 1, 120kDa	7, 10
SF3B2 || splicing factor 3b, subunit 2, 145kDa	7,10

Belacel et al.Fuzzy J-Means and VNS methods for clustering genes from microarray data

Figure 7

Therefore, the differences in clustering patterns observed here, point out either to subunit genes that are part of more than one final product or to errors in the experimental results.

***** FIGURE 7 ABOUT HERE ******

4 Conclusion

We have presented the application of the novel fuzzy clustering method, Fuzzy J-Means, embedded in the global search metaheuristic VNS (VNS+F-JM). The objectivity factor and the Jaccob coefficients comparison shows that VNS+F-JM method gives superior cluster quality and accuracy in all data sets studied. Fuzzy methods in general and the metaheuristics like VNS+F-JM in particular will allow simultaneous determination of: a) genes which are strongly associated to one group; b) genes which are correlated only under some conditions; and c) genes which are always correlated. The former information will help in determining cellular pathways that are largely independent on the environment of the particular cell type - for example pathways which are important for all breast cancer cells regardless of the type or developmental phase of the cancer. The determination of more loosely correlated genes will help to describe the multifunctionality of genes as well as overlapping cellular pathways. Also, the use of accurate methods for fuzzy clustering will provide more accurate gene clustering from more diverse and larger

groups of experiments.

Further work is under way on the extraction of genetic and cellular pathway information from the previously published and our microarray data, as well as the exploration of clustering of samples using the VNS method with some very promising results.

Acknowledgements

We are indebted to Drs. D. Richard (IRMB), J. Nait Ajjou (IRMB) and C. Vaillancourt (Université de Moncton), for critically reviewing the manuscript and for many helpful comments.

References

Belacel, N., Hansen, P. and Mladenovic, N.(2002) Fuzzy J-Means: a new heuristic for fuzzy clustering, *Pattern Recognit.* **35,** 2193-2200.

Bezdek, J.C. (1981) *Pattern recognition with fuzzy objective function algorithms.* Plenum Press, New York.

Dembele, D. and Kastner, P. (2003) Fuzzy C-means method for clustering microarray data, *Bioinformatics* **19**, 973-980.

Dougherty, E.R. *et al.*(2002) Inference from clustering with application to gene-expression microarrays, *J. Comput. Biol.* **9,** 105-126.

Dudoit, S. and Fridlyand, J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset, *Genome Biol.* **3(7)**, *1*-21.

Dunn, J.C. (1974) A fuzzy relative ISODATA process and its use in detecting compact well-separated clusters, *J. Cybern.* **3,** 32-57.

Everitt, B. (1993) *Cluster analysis*. Edward Arnold, London.

Friedman, N., Linial, M., Nachman, I., Pe'er, D. (2000) Using

Bayesian networks to analyze expression data, *J. Comput. Biol.* **7**, 601–620.

Gasch, A.U. and Eisen, M.B. (2002) Exploring the conditional co-regulation of yeast gene expression through fuzzy K-means clustering, *Genome Biol.* **3(11)**, 1-22.

Hansen, P. and Mladenovic, N. (1997) Variable Neighborhood Search: principles and applications, *Eur. J. Oper. Res.* **130,** 449–467.

Hansen, P. and Mladenovic, N. (2001) J-Means: a new local search heuristic for minimum sum-of-squares clustering, *Pattern Recognit.* **34,** 405–413.

Hathaway, R.J. and Bezdek, J.C. (1995) Optimization of clustering criteria by reformulation. *IEEE Trans. Fuzzy Systems* **3,** 241–245.

Ihmels, J., Friedlander, G., Bergman, S., Sarig, O., Ziv, Y., Barkai, N. (2002) Revealing modular organization in the yeast transrcriptional network, *Nat. Genet.* **31**, 370–377.

Jain, A.K. and Dubes, R.C. (1988) *Algorithms for clustering data.* Prentice-Hall, Englewood Cliffs.

Quackenbush, J. (2001) Computational analysis of microarray data, *Nat. Rev. Genet*. **2**, 418–427.

Ruspini, E.H. (1969) A new approach to clustering, *Inf. Control* **15,** 22–32.

Sheng, Q., Moreau, Y., De Moor, B. (2003) Biclustering microarray data by Gibbs sampling. *Bioinformatics* **19 Suppl 2**, II196–II205.

Sorlie, T. et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proc. Natl Acad. Sci.* **98,** 10869–10874.

Stanford Microarray Database: http://genome-www5.stanford.edu/MicroArray/SMD/index.shtml

Whitney, A.R. et al. (2003) Individuality and variation in gene

expression patterns in human blood, *Proc. Natl Acad. Sci* , **100,** 1896-1901.

Woolf, P.J. and Wang, Y. (2000) A fuzzy logic approach to analyzing gene expression data, *Physiol. Genomics.* **3,** 9-15.

Yeung, K.Y., Haynor, D.R., Ruzzo, W.L. (2000) Validating clusters for gene expression data, Technical Report UW-CSE-00-01-01, University of Washington.

Tables

Table 1 Dependence of membership values on fuzziness parameter *m;* μ_T - median of the top membership values; μ_d - overall median membership value.

Cancer	BC/69/85				BC/1022/85			
m	**1.15**	**1.25**	**1.5**	**1.75**	**1.15**	**1.25**	**1.5**	**1.75**
μ_T	0.98	0.77	0.37	0.24	0.78	0.5	0.29	0.21
μ_d	0	0	0.02	0.07	0	0.01	0.05	0.09
$m_{optimal}$~	1.4				1.25			

Blood	HB/43/147				HB/2197/147			
m	**1.15**	**1.25**	**1.5**	**1.75**	**1.15**	**1.25**	**1.5**	**1.75**
μ_T	1	0.99	0.85	0.57	0.99	0.93	0.65	0.44
μ_d	0	0	0	0.02	0	0	0.01	0.03
$m_{optimal}$~	1.75				1.5			

Table 2 Comparison of F-CM, F-JM and VNS methods using the objectivity function and CPU time.

Set m=1.25	*c*	*Best known solution*	*% Dev F-CM*	*% Dev F-JM*	*% Dev VNS*	*CPU F-CM (s)*	*CPU F-JM (s)*	*CPU VNS (s)*
SD1/90/10	9	1452.7	4.73	4.73	0	0.01	2.52	4.73
SD2/90/10 (*m=1.5*)	9	69.12	15.42	0	0	0.02	0.14	0.15
SD/450/10 (*m=1.5*)	9	56.62	733.33	0	0	0.05	2.23	21.7
BC/69/85	3	3429.15	0	0	0	0.03	0.06	0.68
	4	3003.34	1.55	1.55	0	0.02	0.1	1.45
	5	2747.83	0	0	0	0.04	0.05	0.81
	6	2531.22	2.04	2.04	0	0.04	0.05	1.6
	7	2372.35	0	0	0	0.07	0.04	1.57
	8	2232.43	3.53	1.12	0	0.05	0.22	3.51
	9	2102.88	4.46	2.75	0	0.19	0.27	7.91
	10	2002.23	14.58	0.14	0	0.07	0.14	6.71
	11	1911.12	16.7	0	0	0.06	0.27	5.76
	12	1786.74	18.13	3.4	0	0.07	0.16	10.23
	13	1759.26	14.91	1.96	0	0.07	0.4	7.65
	14	1649.03	19.28	5.89	0	0.09	0.18	9.53
	15	1590.94	15.93	2.24	0	0.06	0.59	21.51
	16	1528.51	17.06	3.32	0	0.06	0.17	10.72
	17	1492.18	17.1	2.03	0	0.2	0.16	4.66
	18	1435.15	16.43	2.42	0	0.18	0.1	12.03
	19	1363.11	18.48	3.73	0	0.07	0.11	8.89
	20	1349.15	15.02	0.26	0	0.08	0.12	5.63
Average			**16.9**	**2.52**	**0**			
BC/1022/85	10	37447.1	0	0	0	3.29	8.89	370.67
	20	31008.4	1.79	0.02	0	3.21	39.94	972.45
	30	27743.9	6.81	0	0.11	1.35	86.23	2062.02
	40	25525.8	4.44	1.2	0	2.66	82.45	92.8027
	50	23973.6	0.81	0.13	0	24.84	65.81	453.84
	60	22596.3	2.45	0.33	0	12.05	51.88	1521.84
	70	21592.9	1.36	0.35	0	23.4	49.97	1025.94
	80	20704.9	0.58	0	0.03	72.67	73.14	1053.56
	90	19833.0	0.73	0	0.12	106.02	100.55	1722.07
	100	19055.0	1.98	0.45	0	76.81	28.79	1840.54
Average			**2.09**	**0.25**	**0.03**			
HB/43/147	3	2023.4	0	0	0	0.02	0.04	0.35
	4	1512.2	18.38	18.38	0	0.01	0.03	0.79
	5	1307.8	0	0	0	0.02	0.02	0.46
	6	1153.9	9.49	0	0	0.02	0.11	0.78
	7	1034.1	6.62	1.24	0	0.07	0.43	1.87
	8	904.5	10.7	8.65	0	0.02	0.48	3.5
	9	819.0	9.87	7.42	0	0.02	1.13	6.56
	10	749.3	14.78	0.19	0	0.03	0.23	2.5
Average			**8.73**	**4.48**	**0**			

Set m=1.25	*c*	*Best known solution*	*% Dev F-CM*	*% Dev F-JM*	*% Dev VNS*	*CPU F-CM (s)*	*CPU F-JM (s)*	*CPU VNS (s)*
HB/2197/147	10	79764.1	0.35	0.35	0	6.66	41.12	1801.25
	20	59047.8	0.22	0.22	0	21.05	41.55	1809.99
	30	50660.6	0.40	0.4	0	36.2	42.01	299.315
	40	45703.8	1.06	0.23	0	96.3	298.34	480.958
	50	42261.1	11.38	0.08	0	6.84	596.86	264.672
	60	39704.9	11.49	0.00	0.13	9.3	1441.76	393.061
	70	37718.3	14.75	0.35	0	7.08	597.36	1759.84
	80	36090.1	14.23	0.27	0	9.52	1538.24	1732.63
	90	34744.5	17.80	0.01	0	8.24	1430.06	1184.22
	100	33605.3	18.28	0.00	0.06	9.4	1022.62	1240.71
Average			**9.00**	**0.19**	**0.02**			

Figure Legends

Figure 1 Mean expression profiles for nine families in three simulated data sets. In all three graphs mean expression levels for each of nine groups of genes, at each sample point are presented. For clearer separation expression profiles of some groups are presented as dashed lines. For each data set <σ> represents the average standard deviation calculated for each group and each sample point, averaged over the whole set.

Figure 2 Schematic algorithm representing F-CM (a), F-JM (b) and VNS (c) methods.

Figure 3 Box plots of sorted membership values for four experimental data sets, with 10 clusters and with four different values of *m.* On the *x*-axis are the sorted membership values for each gene (1 largest, 2 second largest, *etc.*) and on the *y*-axis is the membership value.

a. Breast cancer data sets

b. Human blood data sets.

Figure 4 Scattered plots of two top membership values for each gene (*x*-axis value of the largest and y-axis value of the second largest). Vertical line represents the median value of the top memberships

a. Breast cancer data subsets

b. Human blood data subsets.

Figure 5 Calculation of Jaccard coefficients.

a. Contingency table for two partitions of n objects, into N groups, with entry n_{ij} denoting the number of objects that are both in clusters s_i and c_j with **C** and **S** representing here the calculated and correct partitions.

b. Equations used for calculation of Jaccard coefficients from the contingency table (Jain and Dubes, 1988; Yeung, et al. 2000; Dudoit and Fridlyand, 2002)

c. Values of Jaccard coefficients for clusters obtained using F-CM, F-JM and VNS for the three simulated data sets. Cluster assignments were determined from top memberships and compared to the correct clusters in simulated sets.

Figure 6 Clusters for breast cancer data set BC/1022/85 with m=1.25 and c=10. W_{first} are the top membership values, W_{second} are the second highest membership values; 0.5 is the median of the top membership values for this data set; 0.26 is the median of second highest membership values. Gene groups shown in the graphs correspond to groups shown in figures b) – d).

a) Total clustering of all genes;

b) Tight clusters of genes with top membership values greater than top membership value median;

c) Clusters of "Double degenerate genes" with similar two top membership values;

d). Clusters of "Multiply degenerate genes" with several similar membership values.

Figure 7 Comparison of experimental gene expression values and clustering results for several gene subunits. Association to clusters is given in the decreasing order of membership values. Blue arrows points to genes with expression level results and the membership values different than the results for the same gene replicate; red arrows show subunit genes with variations from the complementary subunits.