Evolutionary Techniques for Hierarchical Clustering Applied to Microarray Data

José A. Castellanos-Garzón and Luis A. Miguel-Quintales

University of Salamanca, Department of Computer Science and Automatic, Faculty of Sciences, Plaza de los Caídos s/n, 37008 Salamanca, Spain
{jantonio,lamq}@usal.es
http://informatica.usal.es

Summary. In this paper we propose a novel hierarchical clustering method that uses a genetic algorithm based on mathematical proofs for the analysis of gene expression data, and show its effectiveness with regard to other clustering methods. The analysis of clusters with genetic algorithms has disclosed good results on biological data, and several studies have been carried out on the latter, although the majority of these researches have been focused on the partitional approach. On the other hand, the deterministic methods for hierarchical clustering generally converge to a local optimum. The method introduced here attempts to solve some of the problems faced by other hierarchical methods. The results of the experiments show that the method could be very effective in the cluster analysis on DNA microarray data.

Keywords: Hierarchical clustering, DNA microarray, evolutionary algorithm, genetic algorithm, cluster validity, combinatorial optimization.

1 Introduction

Genetic algorithms ([1, 2]) represent a powerful tool to solve complex problems where the traditional methods face difficulties in finding an optimal solution. The power of genetic algorithms (GAs) lies in its emulation of natural processes, such as adaptation, selection, reproduction and their merge with chance produces robust methods.

The application of GAs to *data mining* has significant importance in the knowledge extraction through the study of classification problems. Likewise, the analysis of clusters ([3, 4]) as an unsupervised classification is the process of classifying objects into subsets that have meaning in the context of a particular problem.

The hierarchical cluster analysis ([5, 6]) is a powerful solution for showing biological data using visual representations, where data can easily be interpreted through a spacial type of tree structures that show cluster relationships, such as the *dendrogram* representation. Overall, the clustering techniques applied to DNA microarray data ([7, 8]) have proven to be helpful in the understanding of the gene function, gene regulation, cellular processes, and subtypes of cells. The latter, it can be a tool very useful for the human health research.

J.M. Corchado et al. (Eds.): IWPACBB 2008, ASC 49, pp. 118–127, 2009.
springerlink.com

The first aspect of this research is that the problem of finding the best dendrogram on a data set, is a problem approached from the combinatorial optimization field, and it is an *NP-Complete* problem. Furthermore, it is not feasible to explore all possibilities on the dendrogram search space, and thereby arises the need of introducing methods that do not consider every solution in the search space, such as evolutionary techniques [9, 10].

The most important accomplishment of this work is the novel method definition of *hierarchical clustering based on GAs*, aimed at the search of global optimums into dendrogram search space on a data set. In addition, we present the theoretical basis of this method, as well as carry out a comparative analysis with other hierarchical clustering methods.

2 The Genetic Algorithm

In recent years, there has been an increasing interest in dealing with the problem of clustering using genetic algorithms, mainly for the partitional approach,[11, 12]. The method is aimed at the search of high quality clustering dendrograms.

The individuals (chromosomes) are dendrograms on a given data set, encoded as an ordered set of clusterings, where each clustering has an order number, called level. Initially, each dendrogram of a population is built up from an initial level to a higher level by joining two clusters chosen randomly in a level, in order to build the next one.

Length of an Individual

The length of a dendrogram can be defined as its number of levels (clusterings), but in the best case, until the half of the dendrogram levels, there will be unitary clusters[1] and that does not have a practical meaning, hence, those levels can be removed. Thus, a parameter can be introduced in order to remove the part of a dendrogram that does not give information. Therefore, we define the length of the dendrogram as follows:

Definition 1. *Dendrogram length.*
Set $\mathfrak{P}_n$ be a data set of size n and set $\mathfrak{G}$ be a dendrogram on $\mathfrak{P}_n$, then the length of $\mathfrak{G}$ is the clustering number of it and is defined as:

$$|\mathfrak{G}| = n - \lfloor n \cdot \delta \rfloor - 2, \tag{1}$$

where δ [2], is the part of $\mathfrak{G}$ to remove, assuming $\delta \geq 1/2$.

2.1 Fitness Function

In every GA it is necessary to measure the goodness of the candidate solutions. In this problem, the fitness of a dendrogram must be evaluated, hence we based

[1] One-element clusters.
[2] Is the fraction of $\mathfrak{G}$ that does not give information.

on one of the given definitions of cluster in [3], that is, *The objects inside of a cluster are very similar, whereas the objects located in distinct clusters are very different.* Thereby, the fitness function will be defined according to the concepts of both, *homogeneity* and *separation*, introduced in [13].

We begin by defining cluster homogeneity and afterwards defining more complex structures until reaching the dendrogram structure.

Definition 2. *Cluster homogeneity.*
If $\mathfrak{D} = [d(i,j)]$ is the proximity matrix on the $\mathfrak{P}_n$ data set, being d the defined metrics on this data set, $\mathfrak{C}$ a clustering of objects in $\mathfrak{P}_n$, C a cluster into $\mathfrak{C}$ and $m = |C|$, then the homogeneity of C is:

$$h(C) = \frac{2}{m \cdot (m-1)} \sum_{i \neq j}^{m \cdot (m-1)/2} d(i,j), (\forall i, j \in C). \tag{2}$$

Definition 3. *Clustering homogeneity.*
Set $\mathfrak{C}$ be a clustering of $\mathfrak{P}_n$, being $k = |\mathfrak{C}|$, then the homogeneity of $\mathfrak{C}$ is:

$$\mathcal{H}(\mathfrak{C}) = \frac{1}{k} \sum_{i=1}^{k} h(C_i). \tag{3}$$

Definition 4. *Distance between two clusters.*
Set $\mathfrak{C}$ be a clustering of $\mathfrak{P}_n$, set C_1 and C_2 be two clusters of $\mathfrak{C}$, then the distance d_m between these clusters is defined as:

$$d_m(C_1, C_2) = \min\{d(i,j)/i \in C_1, j \in C_2\}. \tag{4}$$

Definition 5. *Clustering separation.*
Set $\mathfrak{C}$ be a clustering of $\mathfrak{P}_n$, set C_1 and C_2 be two clusters of $\mathfrak{C}$, $k = |\mathfrak{C}|$, then the $\mathfrak{C}$ separation is:

$$\mathcal{S}(\mathfrak{C}) = \frac{2}{k \cdot (k-1)} \sum_{i \neq j}^{k \cdot (k-1)/2} d_m(C_i, C_j), \quad (\forall i, j \in [1, k]). \tag{5}$$

Definition 6. *Clustering fitness function.*
Set $\mathfrak{C}$ and $\mathfrak{D}$ be a clustering of objects in $\mathfrak{P}_n$ and the proximity matrix of $\mathfrak{P}_n$ respectively, then the fitness function of $\mathfrak{C}$ is defined as:

$$f_c(\mathfrak{C}) = \max \mathfrak{D} + \mathcal{S}(\mathfrak{C}) - \mathcal{H}(\mathfrak{C}). \tag{6}$$

Definition 7. *Dendrogram fitness function.*
Set $\mathfrak{G}$ and $\mathfrak{C}_i$ be a dendrogram on $\mathfrak{P}_n$ and a clustering of $\mathfrak{G}$ respectively, then the fitness function of $\mathfrak{G}$ is:

$$f_d(\mathfrak{G}) = \frac{1}{|\mathfrak{G}|} \sum_{i=1}^{|\mathfrak{G}|} f_c(\mathfrak{C}_i). \tag{7}$$

Based on the previous definition, an *ac agglomerative coefficient* can be used in order to estimate the level into a dendrogram $\mathfrak{G}$, where a cut can be carried out, that is:

Definition 8. *Agglomerative coefficient.*
Set $\mathfrak{G}$ and $\mathfrak{C}_i$ be a dendrogram on $\mathfrak{P}_n$ and a clustering of $\mathfrak{G}$, respectively. The agglomerative coefficient of $\mathfrak{G}$ is defined as:

$$ac(\mathfrak{G}) = \arg_{i \in [1,|\mathfrak{G}|]} \max f_c(\mathfrak{C}_i), \tag{8}$$

the level i whose clustering has the maximum fitness of the whole dendrogram.

2.2 Improving the Fitness Function Cost

Due to the computation complexity of the fitness function defined in (7), the need of decreasing its computation time has arisen. From the theoretical outlook, the above is verified in the following proposition:

Proposition 1. *Algorithmic complexity of f_c.*
Set $\mathfrak{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ be a clustering of a dendrogram $\mathfrak{G}$ on $\mathfrak{P}_n$ and set f_c be the fitness function defined in (6), then the order of $f_c(\mathfrak{C})$ is $O(k^2m^2)$ $(\Omega(kn^2))$, where $m = \max\{|\mathcal{C}_i|\}$, $i \in [1, k]$.

The fitness function defined in (7) can be transformed in an equivalent one but more efficient. This is shown in the following lemmas:

Lemma 1. *Recurrent homogeneity.*
Set $\mathfrak{C}_i$ be a clustering of level i and set $\mathfrak{C}_{i+1}$ be a clustering of level $i+1$, both in the dendrogram $\mathfrak{G}$; C_j and C_l, the two clusters of level i such that its join forms a new clustering $\mathfrak{C}_{i+1}$ of level $i+1$, then the homogeneity of $\mathfrak{C}_{i+1}$ $(\mathcal{H}_1(\mathfrak{C}_{i+1}))$ is computed in the following expression:

for $i = 1$, that is, for the first clustering, $\mathcal{H}_1(\mathfrak{C}_1) := \mathcal{H}(\mathfrak{C}_1)$ and for $i > 1$,

$$\begin{aligned}(k-1) \cdot \mathcal{H}_1(\mathfrak{C}_{i+1}) = k \cdot \mathcal{H}_1(\mathfrak{C}_i) - h(C_j) - h(C_l) + \tfrac{1}{k_3}[k_1 \cdot h(C_j) + \\ + k_2 \cdot h(C_l) + l_1 \cdot l_2 \cdot d_p(C_j, C_l)],\end{aligned} \tag{9}$$

where $k = |\mathfrak{C}_i|, l_1 = |C_j|, l_2 = |C_l|, k_1 = \binom{l_1}{2}, k_2 = \binom{l_2}{2}$, $k_3 = \binom{l_1 \cdot l_2}{2}$, and d_p is the average distance between the clusters C_j and C_l.

Lemma 2. *Recurrent separation.*
Keeping the same conditions of the previous lemma, one can obtain the recurrent separation of a clustering $\mathfrak{C}_{i+1}$ ($\mathcal{S}_1(\mathfrak{C}_{i+1})$):

for $i = 1, S_1(\mathfrak{C}_1) := S(\mathfrak{C}_1)$ and for $i > 1$,

$$(g-k+1)\cdot S_1(\mathfrak{C}_{i+1}) = g\cdot S_1(\mathfrak{C}_i) - d_m(C_j,C_l) - \sum_{t\neq j\wedge t\neq l}^{k-2}[d_m(C_j,C_t) + \\ + d_m(C_l,C_t) - \min\{d_m(C_j,C_t), d_m(C_l,C_t)\}], \tag{10}$$

where $k = |\mathfrak{C}_i|, g = \binom{k}{2}$, being g the number of distances among the clusters of $\mathfrak{C}_i$.

Definition 9. *Clustering recurrent fitness.*
The fitness function of a clustering $\mathfrak{C}_{i+1}$ of $\mathfrak{G}$, according to $\mathcal{H}_1$ and S_1, is defined as:

$$g_c(\mathfrak{C}_{i+1}) = \max\mathfrak{D} + S_1(\mathfrak{C}_{i+1}) - \mathcal{H}_1(\mathfrak{C}_{i+1}), \tag{11}$$

known $\mathcal{H}(\mathfrak{C}_i)$ and $S(\mathfrak{C}_i)$.

Definition 10. *Dendrogram recurrent fitness.*
The fitness function of a dendrogram $\mathfrak{G}$, being $\mathfrak{C}_i$ a clustering of it is:

$$g_d(\mathfrak{G}) = \frac{1}{|\mathfrak{G}|-1}\sum_{i=1}^{|\mathfrak{G}|-1} g_c(\mathfrak{C}_i). \tag{12}$$

Once defined the recurrences, it is possible to verify that the cost of the fitness function defined in (11) is less than the cost of this one defined in (6).

Proposition 2. *Algorithmic complexity of g_c.*
Set $\mathfrak{C}_i, \mathfrak{C}_{i+1}$ be two clusterings (levels i and $i+1$) of a dendrogram $\mathfrak{G}$ on $\mathfrak{P}_n$, $k = |\mathfrak{C}_i|$ and $m = \max\{|C_j|\}, j \in [1,k]$, then the temporal complexity of $g_c(\mathfrak{C}_{i+1})$ is $O(km^2)$ ($\Omega(n^2)$).

2.3 Mutation Operator

The mutation of a dendrogram is performed according to the following steps:

1. We will consider two parameters τ and ϵ for each dendrogram $\mathfrak{G}$ where:
   - τ is the percentage of choosing cluster pairs into level i to build the following level $i+1$;
   - ϵ is a small value that represents the similarity between two clusters, according to the homogeneity measure.
2. A random number $i \in [1, |\mathfrak{G}|]$ is generated, it is the level where the mutation of $\mathfrak{G}$ is carried out.
3. For the clustering of the level i of the previous step, one of the following conditions is chosen:

- the most homogeneous join of cluster pairs of $\tau\%$ of random cluster pairs is chosen;
- the cluster pair with a difference from the cluster pair chosen in the above condition less or equal than ϵ is chosen.

4. The cluster pair chosen in the previous step is joined in order to form a new cluster so that the clustering of the next level $i+1$ can be built.
5. The steps 3 and 4 are repeated on the new level, until i reaches the level $|\mathfrak{G}|$.

2.4 Crossover Operator

The crossover is carried out on two dendrograms to obtain a child dendrogram, and is based on the idea of [14], that is:

1. Given two dendrograms $\mathfrak{G}_1$ and $\mathfrak{G}_2$ (parents), a random number i in $[1, |\mathfrak{G}_1|]$ is generated to choose the level where one can carry out the crossover between both dendrograms.
2. Through a strategy of *greedy algorithm*, the best $\lfloor k/2 \rfloor$ clusters[3] of level i of both dendrograms of the above step are chosen, being k the number of clusters of the level i. A new clustering is formed by repairing the chosen clusters [14, 15].
3. As soon as the new clustering for the level i is built, one can build up the new dendrogram:
   - the higher levels to the level i are built using the MO;
   - the lower levels to the level i are built in a divisible way, that is, for each level less than i, the less homogenous cluster is chosen to be split in two; This process is repeated until reaching the first level.
4. The parent of the less fitness value is replaced by the child dendrogram of the step 3.

3 Experiments on Gene Expression Data

In this section we study the behavior of the GA on a simulated data set of gene expression data and compare the results with other methods according to some cluster validity measures [16, 13, 4]. This data set is considered as benchmark data to prove different clustering algorithms and it was used in [17], published in `http://faculty.washington.edu/kayee/cluster`. The expression matrix of this one is composed of 384 genes evaluated on 17 conditions, labeled into 5 clusters of genes (*ground truth*) and it was normalized with mean 0 and variance 1. The method was implemented on the *R language* (R Development Core Team, [18]).

3.1 Goodness of the Individuals

In this subsection, the curves described by the fitness values of the clusterings in a dendrogram, of the initial and final population, are shown using graphs,

[3] The most homogenous clusters of both dendrograms.

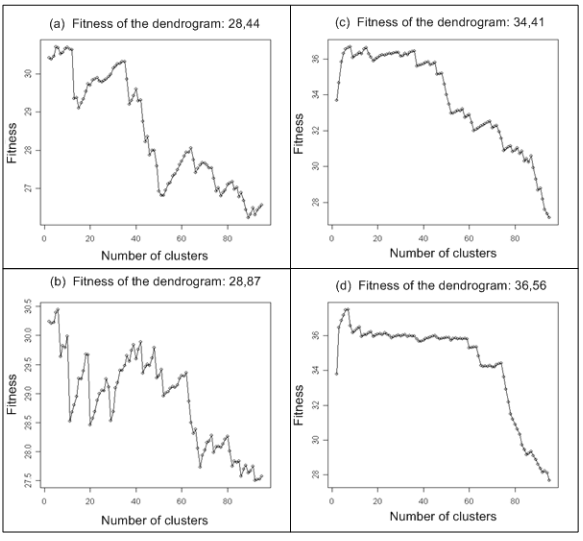

Fig. 1. Graphs of four dendrograms (from a to d), a and b belong to the initial population, while c and d are in the last population. The fitness values of each clustering are shown for each dendrogram.

where the cluster number ($x - axis$) of each clustering of a dendrogram vs. the fitness value ($y - axis$) of each of its clustering is plotted. The graphs of two dendrograms in the initial population (a and b) and two dendrograms of the final population (c and d) are shown in figure 1, $\delta = 3/4$. As it is shown in this figure, the curve described by the dendograms of the initial population presents many oscillations and there are many large differences between the fitness values of two consecutive clusterings. However, the above individuals are improved by execution of the GA with the following parameters: the values of the crossover and mutation likelihood were assigned around 60% and 5% respectively, the generation number in $[10^3, 10^6]$, $\tau \in [15\%, 40\%]$, $\epsilon = 3\%$, $x = 90\%$ and the *euclidian distance* was used on the data set.

One can observe that the fitness of the individuals in the last generation (c and d) was improved after executing of the GA. Hence, the problems presented on the individuals of the initial population have been reduced after applying the genetic operators.

3.2 Homogeneity, Separation and Agreement with the Reference Partition

In this subsection we are going to carry out a cluster validity process to compare the results of the GA. Therefore, we have focused on the quality of clusters in terms of homogeneity (Homog), separation (Separ) [13], silhouette width (SilhoW, [4]), the *Jaccard coefficient* (JC), and the *Minkowski measure* (MM) [13],

Table 1. Cluster validity of the GA vs. five hierarchical clustering methods

Method	f_d	Cluster (ac)	f_c	Homog	Separ	SilhoW	JC	MM
Ground truth	-	5	31.10	6.13	6.60	36.54	-	-
Agnes	32.03	30	37.50	2.02	7.64	36.80	0.12	1.20
Diana	35.41	9	37.47	2.76	9.07	36.96	0.16	1.40
Eisen	23.54	3	39.74	5.80	22.63	37.29	**0.23**	1.82
HybridHclust	37.96	52	39.68	**1.45**	6.72	36.72	0.06	**1.01**
Tsvq	37.32	**5**	40.05	3.86	12.57	37.16	0.15	1.37
GA:								
1.-	32.68	22	35.28	6.82	6.44	36.13	0.09	**1.12**
2.-	33.71	8	36.68	6.70	5.20	36.15	0.21	1.73
3.-	34.41	7	37.49	6.59	5.28	36.09	**0.23**	1.79
4.-	36.56	3	**44.14**	6.36	10.37	36.40	**0.23**	1.81
5.-	37.13	2	39.34	6.27	**28.87**	**37.36**	**0.23**	1.83
6.-	37.55	4	43.11	6.16	11.57	36.32	0.21	1.79
7.-	37.74	3	43.40	6.28	17.02	36.33	**0.23**	1.81
8.-	**39.20**	4	43.51	6.15	15.99	36.29	0.21	1.80

for five methods of hierarchical clustering with *mean link* as a type of distance; *Agnes* and *Diana*[4], *Eisen*[5], *HybridHclust* [19] and *Tsvq*[20].

The GA was initialized with a population of 10 individuals, δ in $\{3/4, 4/5, 5/6, 12/13\}$ and the other parameters were assigned as in the above section. The best 8 outputs were extracted to make comparisons, such as listed in table 1, where the GA is compared with five methods according to eight measures: the *Cluster* column is the number of the cluster of the best clustering in a dendrogram (using ac coefficient) then, for that same clustering the other measures located in the right side of the *Cluster* column of the table were computed. The *Ground truth* row contains the evaluations on the 5 pre-classified clusters of the data set and the best values of the method-measure are highlighted in that table.

In table 1 it can emphasize on different results referring to the GA, where the convergence is proven in the f_d column, since the fitness values of the solutions can be improved. Furthermore, the values of the *Cluster* column, could be employed to determine the optimal number of the cluster by applying some statistical indicator on this list of values. Due to the above results, the method reached the best values for f_d, f_c, separation and silhouette width indicator. For the MM and JC coefficient, it can be emphasized that four executions of the GA and the *Eisen* method reached the best results on JC. In contrast, one of the executions of the GA and the *HybridHclust* method reached the best results on MM.

4 Conclusion

The main goal of this paper has been to present and discuss the theoretical results of a novel evolutionary approach for hierarchical clustering, leading to

the search of global optimums in the dendrogram space. In order to show the effectiveness of this approach with regards to other methods, we have used a simulated data set of gene expression, published as a benchmark.

The introduced method achieved good experiments results in relation to other methods on the DNA microarray data. Therefore, this method can be very important in the process of knowledge discovery as well as in the analysis of gene expression data. Moreover, the most natural way of genetic algorithm application lies precisely in the study of biological processes. Finally, the most important outcomes are:

1. Two fundamental lemmas for improving the temporal complexity of the fitness function. Moreover, the complexity of any other fitness function can be reduced, based on the proof given in those lemmas;
2. The flexibility of the GA to change the genetic operators or add other heuristics, is a strong tool for clustering;
3. The method performed well respect to both, the definition of clustering, that is, homogeneity and separation; and a reference partition of the chosen gene expression data set.

References

1. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley Longman, Inc. (1989)
2. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs, 3rd edn. Springer, Heidelberg (1999)
3. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs (1998)
4. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data. An Introduction to Clustering Analysis. John Wiley & Sons, Inc., Hoboken (2005)
5. Eisen, M., Spellman, T., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, USA 95, 14863–14868 (1998)
6. Jiang, D., Pei, J., Zhang, A.: DHC: A density-based hierarchical clustering method for time series gene expression data. In: Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE) (2003)
7. Berrar, D.P., Dubitzky, W., Granzow, M.: A Practical Approach to Microarray Data Analysis. Kluwer Academic Publishers, New York (2003)
8. Speed, T.: Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC Press LLC (2003)
9. De-Jong, K.A., Spears, W.M.: Using Genetic Algorithms to Solve NP-Complete Problems. In: Proceedings of the Third International Conference on Genetic Algorithms (1989)
10. Godefriud, P., Khurshid, S.: Exploring very large state spaces using genetic algorithms. In: Katoen, J.-P., Stevens, P. (eds.) TACAS 2002. LNCS, vol. 2280, pp. 266–280. Springer, Heidelberg (2002)
11. Chu, P.C., Beasley, J.E.: A genetic algorithm for the set partitioning problem. Technical report, Imperial College, The Management School, London, England, 481–487 (1995)

12. Maulik, U., Bandyopadhyay, S.: Genetic algorithms-based clustering technique. The Journal of the Pattern Recognition Society 33, 1455–1465 (2000)
13. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: A survey. IEEE Transactions on Knowledge and Data Engineering 16(11), 1370–1386 (2004)
14. Greene, W.A.: Unsupervised hierarchical clustering via a genetic algorithm. In: IEEE Congress on Evolutionary Computation, CEC 2003, vol. 2, pp. 998–1005 (2003)
15. Castellanos-Garzón, J.A., Miguel-Quintales, L.A.: Algoritmos genáticos para clustering de datos de expresión génica. Master's thesis, Computer Science and Automatic Department, University of Salamanca, Spain (2006)
16. Handl, J., Knowles, J., Kell, D.B.: Computational cluster validation in post-genomic data analysis 21, 3201–3212 (2005)
17. Yee-Yeung, K.: Clustering Analysis of Gene Expression Data. PhD thesis, University of Washintong (2001)
18. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2006) ISBN 3-900051-07-0
19. Chipman, H., Tibshirani, R.: Hybrid hierarchical clustering with applications to microarray data. Biostatistics 7, 302–317 (2006)
20. Macnaughton-Smith, P., Williams, W.T., Dale, M.B., Mockett, L.G.: Dissimilarity analysis: a new technique of hierarchical subdivision. Nature 202, 1034–1035 (1965)