Published in final edited form as:

IEEE Trans Nanobioscience. 2010 March ; 9(1): 44–50. doi:10.1109/TNB.2009.2037745.

A Special Local Clustering Algorithm for Identifying the Genes Associated With Alzheimer's Disease

Chao-Yang Pang,

Biomedical Informatics and Cheminformatics Group, Conjugate and Medicinal Chemistry Laboratory, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115 USA, and also with the Key Software Laboratory, College of Mathematics and Software Science, Sichuan Normal University, Chengdu 610066, China (cpang@bwh.harvard.edu).

Wei Hu,

College of Mathematics and Software Science, Sichuan Normal University, Chengdu 610066, China (cyp_900@hotmail.com).

Ben-Qiong Hu,

College of Information Management, Chengdu University of Technology, Chengdu 610059, China (bqh@msn.com).

Ying Shi,

Conjugate and Medicinal Chemistry Laboratory, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115 USA (yshi@rics.bwh.harvard.edu).

Charles R. Vanderburg,

Department of Neurology and Harvard NeuroDiscovery Center, Massachusetts General Hospital and Harvard Medical School, Boston MA 02114, USA (crvanderburg@partners.org).

Jack T. Rogers, and

Neurochemistry Laboratory, Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114 USA (jtrogers@partners.org).

Xudong Huang

Conjugate and Medicinal Chemistry Laboratory, Division of Nuclear Medicine and Molecular Imaging, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115 USA (xhuang3@partners.org).

Abstract

Clustering is the grouping of similar objects into a class. Local clustering feature refers to the phenomenon whereby one group of data is separated from another, and the data from these different groups are clustered locally. A compact class is defined as one cluster in which all similar elements cluster tightly within the cluster. Herein, the essence of the local clustering feature, revealed by mathematical manipulation, results in a novel clustering algorithm termed as the special local clustering (SLC) algorithm that was used to process gene microarray data related to Alzheimer's disease (AD). SLC algorithm was able to group together genes with similar expression patterns and identify significantly varied gene expression values as isolated points. If a gene belongs to a compact class in control data and appears as an isolated point in incipient, moderate and/or severe AD gene microarray data, this gene is possibly associated with AD.

Application of a clustering algorithm in disease-associated gene identification such as in AD is rarely reported.

Index Terms

Algorithm; Alzheimer's disease (AD); compact class; local clustering (LC) feature; special local clustering (SLC)

I. INTRODUCTION

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters or classes), so that the data in each subset (ideally) share some common trait—often proximity according to some defined distance measure. Data clustering is a common technique for statistical data analysis that is used in many fields, including image analysis and bioinformatics [1].

The computational task of classifying a data set into k clusters (or classes) is often referred to as k -means clustering, which is also called LBG algorithm in signal processing or general Lloyd's algorithm (GLA) [2], [3]. The Linde-Buzo-Gray (LBG) algorithm assigns each point to the clusters whose center (i.e., centroid) is nearest until the assignment does not change. The algorithm steps are as given in the following [1].

Step 1: Choose the number of clusters k .

Step 2: Randomly generate k clusters and determine the cluster centers, or directly generate k random points as cluster centers.

Step 3: Assign each point to the nearest cluster center.

Step 4: Recompute the new cluster centers.

Step 5: Repeat Step 3 and Step 4 until some convergence criterion is met (the assignment has not changed).

In a common data distribution, an ordinary phenomenon is that one group of data is separated from another group of data evidently, and that data in these two groups cluster locally and separately. That is, the data in these two groups clusters independently. This phenomenon, illustrated in Fig. 1, is called *local clustering feature* in this paper.

The local clustering feature is useful for a clustering algorithm. During the t th iteration of LBG algorithm, the i th class has distortion that is the average distance between each point and its centroid, which is denoted by $D_i^{(t)}$ ($1 \leq i \leq m$), where m is the number of classes. Pang proves that every sequence $\{D_i^{(0)}, D_i^{(1)}, \dots, D_i^{(t)}, D_i^{(t+1)}, \dots\}$ is convergent under ideal conditions if the distribution of data has a local clustering feature [3]. That is, distortion sequences are convergent locally. According to this property, an algorithm named the local clustering (LC) algorithm is presented by Pang [3] and Chaoyang *et al.* [5]. The essential idea of LC algorithm is that, by deleting all points in the class in which distortion sequences converge firstly from the training set, the training set is comprised by residual points and calls for the LBG algorithm again. The LC algorithm is faster than the LBG algorithm by factors of 4–13 approximately [5].

Suppose the i th class is $R_i^{(t)}$ during the t th iteration of LBG algorithm. Set $R_i^{(t)}$ is a probability system and has entropy $H(R_i^{(t)})$. It was proved that entropy sequence

$\{H(R_i^{(0)}), H(R_i^{(1)}), \dots, H(R_i^{(t)}), \dots\}$ is convergent [3], [6]–[9]. That is, the termination criterion of LBG algorithm can be replaced by the convergence of entropy sequence [10], [11].

The following statements introduce the application of the clustering method to genome analysis whereby similar genes cluster in the same class.

A microarray is a powerful tool for genome analysis. Each microarray experiment yields at least a thousand data points. Since the volume of data grows exponentially, data analysis becomes challenging. In data analysis, the information from the raw data file is further processed to yield meaningful biological results. A lot of approaches have been developed for analyzing huge microarray data, such as clustering, classification, and genetic network analysis [12], [13]. Zhong *et al.* improved a k -means clustering algorithm for exploring local protein sequence motifs representing common structural property. The improved k -means algorithm generated high-quality clusters [14]. In addition, clustering can be combined with other optimization algorithms to generate a phylogenetic tree [15]. To predict the genes that are associated with a given disease from microarray data, a difference expression (DE), such as fold change, t -test, analysis of variance, nonparametric analysis, and regression analysis, can be used [12]. We report on the use of clustering to predict the genes associated with Alzheimer's disease (AD) from its microarray data in the following.

AD is the most common cause of dementia, afflicting 24 million people worldwide. The cause and progression of AD is not well understood. Possible causes and cures of the disease have been conjectured with varying evidence supporting each claim [16]. Few genes have proved to be related to AD, such as ApoE4, PS1, PS2 and APP. Some genes have been tested for association with late-onset sporadic AD [17], [18]. In sum, predicting genes associated with AD is difficult. The use of a clustering method to predict the genes associated with AD is rarely reported. In this paper, the authors report on a designed clustering algorithm that makes similar genes cluster tightly together to predict the genes associated with AD.

II. IMPROVED LC ALGORITHM TO GENERATE COMPACT CLASSES

A. Compact Set and Its Method of Generation

For any subset of Euclidean space R^n , every sequence in the set has a convergent subsequence, the limit point of which belongs to the set. This subset is called *compact set* [19] (http://en.wikipedia.org/wiki/Compact_set; http://en.wikipedia.org/wiki/Cauchy_sequence). The conception of the compact set (or compactness) is a topology conception. To understand it easily, compactness can be described visually as the phenomenon that many points focus very tightly in a small area, while the noncompact set are those points that cluster loosely in a big area [3], [19].

The LBG, LC, or other algorithms will partition training sets into classes. Some classes are compact, and some classes are noncompact. The most common situation is that a class contains a compact subset and some loose points, and the compact subset is around the center of the class. To extract the compact subset as a class, the following 3σ -principle is introduced.

For Gauss distribution, suppose that σ denotes the deviation of all distances between data and centroid. The 3σ -principle of statistics states that, with more than 99% probability, a point will fall into a domain with a radius of 3σ [3], [20]. If the distribution is comprised by many points, the domain with a limited radius 3σ contains many points and is possibly a

compact set. Even the domain with a radius of 3σ is not compact, the subdomain with size $3\sigma/4$ or $3\sigma/16$ and so on is compact. For Gauss distribution that is comprised by enough points, the compact domain (or subset) always exists. In general, for a class generated by clustering algorithm, all distances between points and the centroid of the class comprise a Gauss distribution approximately. Therefore, it will generate a compact subset that extracts the points by which distance to the centroid is less than a radius $3\sigma/4$ or $3\sigma/16$.

B. Subroutine 1: Local Clustering Algorithm With 3σ -Principle

Suppose the i th class is $R_i^{(t)}$ at the t th iteration of LBG or LC algorithm, where t denotes the t th iteration. With the increasing of the iteration, class sequence $\{R_i^{(0)}, R_i^{(1)}, \dots, R_i^{(t)}, R_i^{(t+1)}, \dots\}$ ($1 \leq i \leq m$) appears, where m denotes the number of classes. Let

$$D_i^{(t)} = \frac{1}{|R_i^{(t)}|} \sum_{x \in R_i^{(t)}} d(x, c_i^{(t)}) \quad (1)$$

$$\delta_i^{(t)} = \frac{1}{|R_i^{(t)}|} \sum_{x \in R_i^{(t)}} |d(x, c_i^{(t)}) - D_i^{(t)}| \quad (2)$$

$$K_i^{(t)} = \left\{ x | d(x, c_i^{(t)}) \leq \frac{1}{4^p} (D_i^{(t)} + 3\delta_i^{(t)}), x \in R_i^{(t)} \right\} \quad (p \geq 0). \quad (3)$$

Clearly, $D_i^{(t)}$ is the distortion of class $R_i^{(t)}$ and $\delta_i^{(t)}$ is the deviation of $D_i^{(t)}$. The parameter p is used to adjust the size of subset $K_i^{(t)}$ and make it always compact. Fig. 2 illustrates the 3σ -principle and compact subset $K_i^{(t)}$.

The local clustering algorithm with 3σ -principle is described in the following, revealing an improvement over the LBG. Its essential idea is that, using the criterion of entropy convergence to mark the stable class $R_i^{(t)}$, using 3σ -principle to extract compact subset $K_i^{(t)}$ of $R_i^{(t)}$, removing $K_i^{(t)}$ from training set T and updating T , repeating again.

Input parameters:

T : Training set.

m : The number of class.

ε : The stop threshold.

$C^{(0)} = \{c_i^{(0)} | 1 \leq i \leq m\}$: Initial codebook.

p : The parameter to adjust the size of subset $K_i^{(t)}$ ($p \geq 0$).

Output:

$\varphi(T) = \{K_1^{(t)}, K_2^{(t)}, \dots, K_i^{(t)}, \dots, K_m^{(t)}\}$ (i.e., the set of compact subset.).

$\sigma(T) = \{B_1^{(t)}, B_2^{(t)}, \dots, B_i^{(t)}, \dots, B_m^{(t)}\}$, where $B_i^{(t)} = R_i^{(t)} - K_i^{(t)}$, and it is comprised by dispersive points ($1 \leq i \leq m$).

Void Subroutine1 ($T, m, \varepsilon, C^{(0)}, p, \varphi(T), \sigma(T)$)

{

Step 1. Initialization: Let iteration number $t = 0$. Let *Counter* = m . Let $\varphi(T) = \phi$ and $\sigma(T) = \phi$. According to initial codebook $C^{(0)}$, generate initial partition

$$\varphi^{(0)} = \{R_i^{(0)} | R_i^{(0)} \subset T, 1 \leq i \leq m\}.$$

Step 2. While (*Counter* > 0) {

Step 2-1. Generate new codebook

$$C^{(t+1)} = \{c_i^{(t+1)} | 1 \leq i \leq m\}$$

and new partition $\varphi^{(t+1)} = \{R_i^{(t+1)} | 1 \leq i \leq m\}$

/*Note: Check whether entropy sequence $\{H(R_i^{(0)}), \dots, H(R_i^{(t)}), \dots\}$ is convergent. If it is convergent, let the convergent marker *StableMarker* ($R_i^{(t+1)}$) = True */

Step 2-2. For ($i = 1; i \leq \text{Counter}; i++$) {

Estimate the entropy of class $R_i^{(t+1)}$, i.e.,

$$H(R_i^{(t+1)}) = \log_2 |R_i^{(t+1)}|.$$

/*Note: Using entropy as the convergence criterion of algorithm */

$$\text{If } \left(\frac{|H(R_i^{(t+1)}) - H(R_i^{(t)})|}{H(R_i^{(t)})} < \varepsilon \right)$$

{ *StableMarker* ($R_i^{(t+1)}$) = True; }

else { *StableMarker* ($R_i^{(t+1)}$) = False }

}

/*Note: Extract the data around the centroid of class as a genuine class */

Step 2-3. For ($i = 1; i \leq \text{Counter}; i++$) {

If (*StableMarker* ($R_i^{(t+1)}$) = True) {

Calculate compact subset $K_i^{(t)}$ according to formula (3).

Calculate $B_i^{(t)}: B_i^{(t)} = R_i^{(t+1)} - K_i^{(t)}$

Let $\varphi(T) = \varphi(T) \cup \{K_i^{(t)}\}$

Let $\sigma(T) = \sigma(T) \cup \{B_i^{(t)}\}$

Update training set: $T = T - R_i^{(t)}$

Update codebook:

$$C^{(t+1)} = C^{(t)} - \{c_i^{(t)}\}$$

$Counter = Counter - 1$

}

}

$t = t + 1.$

}

}

C. Special LC Algorithm to Generate Compact Classes (SLC)

Note that subroutine1 is not a partition of the training set. Subroutine1 extracts only compact subsets of all classes, and the residual points are unclassified. The residual points can be regarded as a new training set. Possibly, some residual points cluster together tightly and form a new compact subset that has a much smaller size.

Input parameters:

T_0 : Training set.

m_0 : The initial number of classes.

ε : The stop threshold for clustering.

Output:

Num : The final number of classes.

CLS : The partition of T_0 in which each class is compact.

$unCLS$: The set of isolated points.

SLC Algorithm:

Step 1. Initialization:

Let $T = T_0$, $m = m_0$, $CLS = \phi$, and $p = 0$.

Step 2. For ($i = 0$; $i < [\log_2 m]$; $i++$)

/* $[\log_2 m]$ denotes the integer */

{

Step 2.1. Generate initial codebook

$$C^{(0)} = \{c_i^{(0)} | 1 \leq i \leq m\}.$$

Step 2.2. Call

Subroutine 1 ($T, m, \varepsilon, C^{(0)}, p, \phi(T), \sigma(T)$)

Step 2.3. $CLS = CLS \cup \phi(T)$;

Step 2.4. $T = \sigma(T)$;

Step 2.5. $m = \left\lceil \frac{m}{2} \right\rceil$; $p = p + 1$;

}

Step 3. Every residual point x in the last set $\sigma(T)$ is marked as isolated points. If there are some classes in CLS whose number of points is less than 4, every point x in these classes will be marked as isolated points too. All isolated points are selected into $setunCLS$. Let Num denote the number of classes contained in CLS .

The three outputs are CLS , $unCLS$, and Num .

III. APPLICATION OF SLC: CLASSIFYING MICROARRAY DATA OF AD AND PREDICTING SOME GENES ASSOCIATED WITH AD

A. Microarray Data of AD Has Local Clustering Feature

In this paper, the original microarray data is downloaded from GEO DataSets, NCBI [21], which includes 22 283 genes. Four cases of control, incipient, moderate, and severe data are provided in the original microarray data. The authors of this paper collected nine control samples for each gene as a matrix with 22 283 lines and nine columns. In this matrix, each line vector is a 9-D vector. All line vectors form a training set, denoted by T_{ctrl} . Seven samples of incipient for each gene are selected to form a 7-D vector, and 22 283 vectors form the training set T_{incip} . Eight samples of moderate for each gene are selected to form an 8-D vector and training set T_{moder} . Seven samples of severe for each are selected to form training set T_{Severe} . The four training sets T_{ctrl} , T_{incip} , T_{moder} , and T_{Severe} have 22 283 vectors (i.e., 22 283 genes), respectively (see Fig. 3).

To find the feature of training set T_{incip} , the LBG algorithm is applied to the training set T_{incip} , and ten classes are generated. The data distribution of set T_{incip} is shown in Fig. 4. It indicates that the AD microarray data have a local clustering feature. Therefore, the SLC algorithm presented in this paper can be applied to group AD data and to generate compact classes or isolated points.

B. Clustering of AD Data Generated by SLC

Fig. 4 shows that AD data has a local clustering feature with which the SLC algorithm fits for its processing. The data in training set T_{ctrl} is log-transformed, and the mean of each vector is subtracted in this section. The LBG algorithm is applied to the cluster data, and ten classes are generated. At the same time, the SLC algorithm is applied to process the same data, and the initial number of classes is ten. Parts of the result of the clustering are shown in Fig. 5.

The SLC makes all classes compact. Similar genes cluster together. Clearly Fig. 5 shows that the clustering quality of SLC is better than that of the LBG algorithm. The authors come to the same conclusion from the other training sets T_{incip} , T_{moder} , and T_{Severe} .

C. Apply SLC to Predict Genes Associated With AD

1) Variation of Gene Expression Values will Cause the Variation of Its Position at Clustering and Possibly Become an Isolated Point—If a gene is controllable, it will cluster with other similar genes and belong to a class of training set T_{ctrl} , where T_{ctrl} is comprised by control data. When the expression values of this gene suddenly change, they will not cluster with the previous other genes and possibly may become an isolated point. That is, this gene will possibly become an isolated point in other training set T_{incip} , T_{moder} , or T_{Severe} . The gene with the obvious variation of position is a potential gene that is possibly associated with AD. It can be detected by its obvious variation of position.

The following fact found by the authors of this paper supports the opinion about the obvious variation of the position.

Ten initial codevectors are selected randomly. Let SLC act on training set T_{ctrl} . The gene chromogranin B (secretogranin 1) is classified in a compact class that contains 2127 genes (see Fig. 6). The middle part in Fig. 6 shows that these 2127 similar genes cluster together and are similar to gene chromogranin B (secretogranin 1). Then, the authors let SLC act on training set T_{Severe} . Gene chromogranin B (secretogranin 1) does not cluster with the previous genes; it becomes an isolated point (see Fig. 6).

Fig. 6 demonstrates the obvious variation of the position of the gene expression values that are used to find genes possibly associated with AD.

2) Apply SLC to Predict the Genes that are Associated With AD—According to the opinion about the variation of the position presented in the above section, the author proposed the following simple prediction method.

First, let SLC act on the training set T_{ctrl} . Then let SLC act on T_{incip} , T_{moder} , and T_{Severe} , respectively. If there is a gene A that is classified in a compact class in T_{ctrl} , and it becomes an isolated point in T_{incip} , T_{moder} , or T_{Severe} , gene A will be possibly associated with AD.

To predict genes, 512 initial code vectors (i.e., the seeds of classes) are randomly selected from the training set. The SLC is executed 16 times. Table I lists part of the predicted genes, and the explanation of biology of these genes is listed in Table II.

According to Tables I and II, most of the predicted genes are associated with AD (The genes marked by “*” in Table I are associated with AD according to biology). In addition, because the initial seeds of all classes are randomly selected, the correctness of prediction and the repeating time of the same gene are affected. As part of their future work, the authors aim to improve the repeating time.

3) Comparison Between SLC and LBG Algorithm—The SLC not only has the function to predict genes, but also has many advantages compared with the LBG algorithm (the LBG cannot predict genes because it cannot extract the loose genes in the classes.). The comparison of the performance of the SLC and LBG is listed in Table III. The SLC clearly has some advantages.

IV. CONCLUSION

Clustering is the classification of objects into different classes so that the data in the same class (ideally) shares some common traits. Data clustering is used in many fields, such as image analysis and bioinformatics. The local clustering feature is the phenomenon that, one group of data is evidently separated from another group of data, and that the data in these two groups cluster locally and separately (see Fig. 1). All class sequences are convergent locally under an ideal condition. The clustering algorithm of the SLC is presented in this paper, and it fits the processing of the microarray data of genes. The compact class is the class in which all elements tightly cluster in a small area. The basic idea of the SLC is that, by removing convergent compact classes as soon as possible, isolated points are left. Compact classes make similar genes cluster tightly together, and the genes with variations in the position of their expression values are related to isolated points. If a gene belongs to a class in the control data and becomes an isolated point in the incipient, moderate, or severe data, this gene is possibly associated with some disease (see Fig. 6). Applying the SLC, some genes associated with AD are predicted, and the genes marked by “*” in Table I are

associated with AD according to biology. The advantages of the SLC are listed in Table III. Noteworthy, is that, unlike many other clustering algorithms, the SLC algorithm can identify disease-associated genes.

Acknowledgments

The authors would like to thank Prof. G.-C. Guo, Prof. Z.-F. Han, and Prof. Z.-W. Zhou of University of Science and Technology of China, and Prof. J. Zhang of Sichuan Normal University for their valuable advices. The authors would also like to thank Ms. Kimberly Lawson of the Radiology Department of Brigham and Women's Hospital for editing the manuscript.

This work was supported by National Institute of Health/National Institute on Aging under Grant 5R21AG028850 and Alzheimer's Association under Grant IIRG-07-60397, and by the Radiology Department, Brigham and Women's Hospital, Boston, MA.

REFERENCES

1. Wikipedia. Cluster analysis. 2008 Aug.. [Online]. Available: http://en.wikipedia.org/wiki/Cluster_analysis
2. Linde Y, Buzo A, Gray RM. An algorithm for vector quantization design. IEEE Trans. Commun 1980 Jan.;vol. 28(no. 1):84–95.
3. Pang, C-Y. Ph.D. dissertation. Dept. Comput. Sci. Univ. Electron. Sci. Technol. China. Chengdu, Chengdu, China: 2002. Vector Quantization and Image Compression.
4. Fritzke B. The LBG-U method for vector quantization-an improvement over LBG inspired from neural networks. Neural Process. Lett 1997 Feb.;vol. 5:35–45.
5. Chaoyang P, Shixin S, Ye P, Haiying G. A fast code-book training algorithm using local clustering. J. Electron. Inf. Technol 2002;vol. 9(no. 9):1282–1286.
6. Wikipedia. Information entropy. 2008 Aug.. [Online]. Available: http://en.wikipedia.org/wiki/Shannon_entropy
7. Wikipedia. Lebesgue integration. 2008 Jul.. [Online]. Available: http://en.wikipedia.org/wiki/Lebesgue_integral
8. Wikipedia. Riemann integral. 2008 Apr.. [Online]. Available: http://en.wikipedia.org/wiki/Riemann_integral
9. Wikipedia. Mutual information. 2008 Jul.. [Online]. Available: http://en.wikipedia.org/wiki/Mutual_information
10. Pang C-Y, Shi-xin S. Codebook training algorithm by the convergence of entropy sequence for vector quantization. Syst. Eng. Electron 2002;vol. 24(no. 1):83–85.
11. Xia L, Xuehui L, Jihong Z. Modeling of vector quantization image coding in an ant colony system. Chin. J. Electron 2004;vol. 13(no. 2):305–307.
12. Reimers M. Statistical analysis of microarray data. Addict. Biol 2005 Mar.;vol. 10:23–35. [PubMed: 15849016]
13. Liu Q, Yang J, Chen Z, Yang MQ, Sung AH, Huang X. Supervised learning-based tagSNP selection for genome-wide disease classifications. BMC Genomics 2008;vol. 9 Suppl 1, S6:1–9.
14. Zhong W, Altun G, Harrison R, Tai PC, Pan Y. Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property. IEEE Trans. Nanobiosci 2005 Sep.;vol. 4(no. 3):255–265.
15. Qin L, Chen Y, Pan Y, Chen L. A novel approach to phylogenetic tree construction using stochastic optimization and clustering. BMC Bioinf 2006;7 Suppl 4, S24 [Online]. Available: <http://www.biomedcentral.com/1471-2105/7/S4/S24>.
16. Wikipedia. Alzheimer's disease. 2008 Aug.. [Online]. Available: http://en.wikipedia.org/wiki/Alzheimer%27s_disease#cite_note-43
17. Waring SC, Rosenberg RN. Genome-wide association studies in Alzheimer disease. Arch. Neurol 2008;vol. 65:329–334. [PubMed: 18332245]
18. Li Y, Grupe A, Rowland C, Holman P, Segurado R, Abraham R, Jones L, Catanese J, Ross D, Mayo K, Martinez M, Hollingworth P, Goate A, Cairns NJ, Racette BA, Permuter JS, O'donovan

MC, Morris JC, Brayne C, Rubinsztein DC, Lovestone S, Thal LJ, Owen MJ, Williams J. Evidence that common variation in NEDD9 is associated with susceptibility to late-onset Alzheimer's and Parkinson's disease. *Hum. Mol. Genet* 2008;vol. 17:759–767. [PubMed: 18063669]

19. Wikipedia. Compact space. 2008 Aug.. [Online]. Available: http://en.wikipedia.org/wiki/Compact_set
20. Wikipedia. Normal distribution. 2008 Aug.. [Online]. Available: http://en.wikipedia.org/wiki/Gauss_distribution
21. Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *PNAS* 2004 Feb.;vol. 101:2173–2178. [PubMed: 14769913]

Biographies

Chao-Yang Pang received the B.S. degree in pure mathematics from the Sichuan Normal University, Chengdu, China, in 1993, the M.S. degree in applied mathematics from the University of Electronic Science and Technology of China (UESTC), Chengdu, in 1996, and the Ph.D. degree in computer science from the UESTC, in 2002.

He was a Postdoctoral Fellow in quantum computation and quantum information and image compression at the University of Science and Technology of China, Hefei, China. He has been a Research Fellow in biomedical informatics in the Department of Radiology, Brigham and Women's Hospital Boston, MA, a teaching affiliate of Harvard Medical School. He is currently a Professor with the Key Software Laboratory, College of Mathematics and Software Science, Sichuan Normal University. His research interests include the theory and applications of quantum computing and bionics algorithm in bioinformatics.

Wei Hu received the M.S. degree in computational mathematics from the Sichuan Normal University, Chengdu, China, in 2009.

He is currently with the College of Mathematics and Software Science, Sichuan Normal University. His research interest includes algorithm analysis and design.

Ben-Qiong Hu received the M.S. degree in applied mathematics from the University of Electronic Science and Technology of China, Chengdu, China.

She is currently an Instructor at the College of Information Management, Chengdu University of Technology, Chengdu. Her research interests include algorithm analysis and design.

Ying Shi received the M.D. degree from Peking University, Beijing, China, in 1989 and the M.S. degree in biology from Concordia University, Montreal, QC, Canada, in 2006.

She is currently a Research Fellow in the Department of Radiology, Brigham and Women's Hospital, Harvard Medical School. Her research interests include molecular biology and neuroscience, mechanism of nanoneurotoxicity and its links to neurodegenerative diseases such as Alzheimer's disease. Her research is also focused on characterizing potentially theranostic agents targeting human diseases.

Charles R. Vanderburg received the B.S. degree in biochemistry from the University of Pittsburgh, Pittsburgh, PA, in 1978, the M.S. degree in immunology from Seton Hall University, South Orange, NJ, in 1983, and the Ph.D. degree in cell and molecular biology from the University of Medicine and Dentistry of New Jersey, Stratford, in 1989.

He was a Research Fellow at Harvard Medical School, Boston, MA. He is currently Director of the Department of Neurology and Harvard NeuroDiscovery Center, Massachusetts General Hospital and Harvard Medical School, Boston. He is the author or coauthor on a wide range of neurological disorders including amyotrophic lateral sclerosis, Parkinson's disease, Alzheimer's disease, autism, and bipolar disorder. His research interests include the microgenomic and microproteomic profiling of human neuropathological tissue specimens and their derivatives including miRNAs and exosomes, and with a focus on developing artifact-control methodologies for such samples, his laboratory is actively investigating novel bioinformatic strategies and technologies.

Jack T. Rogers received the Ph.D. degree in biochemistry from St. Mary's Medical School, University of London, London, U.K., in 1983.

He was a Postdoctoral Fellow at Massachusetts Institute of Technology, Cambridge, where he was engaged in research on molecular biology.

He is currently the Director of Neurochemistry Laboratory, Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston. He is also an Associate Professor at the Harvard Medical School. His research has been focused on the role that RNA in the maintenance of iron homeostasis related to disease processes in the blood (anemia) and in Alzheimer's disease, and more recently in the etiology of Parkinson's disease. He is the author or coauthor in established scientific journals such as *Journal of Biology and Chemistry* and *Proceedings of the National Academy of Sciences*. His numerous articles and peer review publications are on the subjects of iron metabolism, and translational control-related disease progression, now including Parkinson's disease. Recently, he has focused energy towards developing a program in drug discovery to find therapies for people with neurodegenerative diseases. An integrative program was instigated to use RNA based therapeutic strategies to ameliorate the progression of Alzheimer's disease during 2002–2007, and currently for Parkinson's disease in partnership with the Michael J-Fox Foundation in 2008. He is an excellent mentor for graduate students and postdoctoral fellows, and a scientist with a creative approach to gene base therapies and basic sciences. He and his family live in Arlington, MA.

Dr. Rogers is on several Scientific Advisory Boards.

Xudong Huang received the Ph.D. degree in nuclear science and engineering from Massachusetts Institute of Technology, Cambridge, in 1995.

He was a Postdoctoral Fellow at Massachusetts General Hospital (MGH)/Harvard Medical School, Boston. He is currently an Assistant Professor of Radiology, Harvard Medical School, and the Director of Conjugate and Medicinal Chemistry Laboratory, Division of Nuclear Medicine and Molecular Imaging, Department of Radiology, Brigham and Women's Hospital, Boston, MA. His research interests include high-throughput screen and design, synthesize, and characterize targeted theranostic compounds for human diseases such as neurological disorders, cancer, diabetes, and cardiovascular diseases, etc. In addition, his laboratory is exploring applications of biomedical informatics (such as medical image analysis and bioinformatics) and cheminformatics (such as structure-based rational drug design) in systems biology that is related to etiopathogenesis of various human diseases.

Dr. Huang serves on scientific advisory and editorial boards for several research foundations and scientific journals.

Fig. 1. Illustration of the local clustering feature. The datum in A and in B clusters locally and separately. This common phenomenon is called the local clustering feature in this paper. The figure is taken from Ref. [4].

Fig. 2.

Applying 3σ — principle to generate compact kernel subset $K_i^{(t)}$ (i.e., the shadow part): In a class, most of points locate around their centroid, and few points are far away from the centroid. Almost all points lie within three standard deviations in general. Using criterion $d(x, c_i^{(t)}) \leq (1/4^p)(D_i^{(t)} + 3\sigma_i^{(t)})$, the compact subset $K_i^{(t)}$ defined in (3) is extracted. If a class has only small compact subset, parameter p will be increased to extract it.

Fig. 3.

Structure of training sets T_{ctrl} , T_{incip} , T_{moder} , and T_{severe} . The rows of four tables are the expression data of genes, and the data of each column is a sample. Four tables contain control, incipient, moderate, and the data are downloaded from GEO DataSets, NCBI [21]. The data of each row can be regarded as a vector. All vectors contained in the first table comprise a training set, and it is denoted by symbol T_{ctrl} . The second, third, and fourth table are training sets too that are denoted by symbols T_{incip} , T_{moder} , and T_{severe} , respectively. In addition, the shown data are log-transformed data, and the mean value of every column is subtracted from the data in this column.

Fig. 4.

Microarray data of AD hide local clustering feature. All data comes from GEO DataSets, NCBI [21]. They are incipient data and comprise training set T_{incip} . They are also original data that are not log-transformed. All data in training set T_{incip} are grouped by the LBG algorithm, and ten classes are generated. To visually show a vector, its components are used to fit curve, and the Y axis of this figure represents the gene expression data. If a set is compact, all curves will be close to each other and their values at Y axis focus in a small interval. In this figure, four classes cluster locally in the interval $[0, 12067]$, while another four classes cluster locally in the different interval $[12067, 36202]$. Two intervals separate clearly, just like the data distribution of Fig. 1. That is, training set T_{incip} has a local clustering feature. In addition, some classes are compact, while some other classes are incompact. The other training sets T_{ctrl} , T_{moder} , and T_{Severe} have the local clustering feature were also subjected to the same test method. To make every class compact, the SLC algorithm presented in this paper will be applied to process AD microarray data.

Fig. 5.

Comparison of the clustering quality between the SLC and LBG algorithm. Every row in the figure denotes a gene, and every column denotes a control sample of a gene. The nearer the color is, the more similar the genes are. That is, the clustering quality will be better if the color is more regular. Clearly, the clustering quality of SLC is better than LBG. (Note: The display software is Treeview downloaded from Eisen Lab <http://rana.lbl.gov/EisenSoftware.htm>. The data processed by the two algorithms is log-transformed data, and the mean value of every column is subtracted from the data in this column, see Fig. 3).

Fig. 6.

Illustration of the obvious variation of position of the gene expression values. The variation is detected by the SLC algorithm presented in this paper. The gene with the variation is associated with AD.

TABLE I

Predicted Genes by The SLC Algorithm

No.	Gene Name	Probe Set ID	T	Set
1 [*]	glycine receptor, alpha 3 (GlyR $\alpha 3$)	207928_s_at	5	T_{incip}
2 [*]	distal-less homeobox 2 (Dlx2)	215685_s_at	5	T_{incip}
3 [*]	protease, serine, 16 (thymus)	208165_s_at	5	T_{moder}
4 [*]	Sin3A-associated protein, 18kDa (SAP18)	208741_at	4	T_{moder}
5	glycophorin A (MNS blood group) (GPA)	205837_s_at	5	T_{serve}
6	N-acetyltransferase 8B (gene/pseudogene) (NAT8B)	206964_at	5	T_{serve}
7 [*]	low density lipoprotein receptor-related protein 3	204381_at	4	T_{serve}
8 [*]	Kruppel-like factor 15(KLF15)	221302_at	4	T_{serve}

“T” denotes the times of appearance of a gene in totally 16 times of test;

* represents gene is associated with AD according to biology.

TABLE II

Biological Functions of the Predicted Genes

No.	Explanation of biology
1	The inhibitory glycine receptor (GlyR) is a ligand-gated ion channel that mediates post-synaptic inhibition in spinal cord and other regions of the vertebrate central nervous system. GlyR alpha 3 subunit mRNA was expressed at low levels in the cerebellum, olfactory bulb, and hippocampus. $\alpha 3$ mRNAs encoding adult strychnine-sensitive GlyR protein were found here in brain regions associated with sensory and motor processing, e.g. olfactory bulb, superior colliculus, vestibular nuclei, cerebellum, and spinal cord.
2	<i>Dlx1</i> and its closely linked relative <i>Dlx2</i> are homologues of the <i>Drosophila</i> homeobox gene <i>distal-less</i> , and their function is to down-regulate the Notch signaling pathway in neuronal specification and differentiation of the telencephalon. Notch signaling pathways are closely related to Alzheimer's disease biology.
3	The gene encoding thymus-specific serine protease (PRSS16) maps to the extended HLA complex that harbors several genes that predispose for autoimmune diseases. The data suggest that it is unlikely that polymorphisms within the PRSS16 gene are involved in the predisposition to diabetes type 1. Diabetes increases Alzheimer's risk.
4	It plays a key role in the gene-specific recruitment of the HDAC complex by a number of transcription factors including Gli, GAGA, and Bicoid. This protein directly interacts with SIN3 and enhances SIN3-mediated transcriptional repression when tethered to the promoter.
5	Glycophorins A (GYPA) and B (GYPB) are major sialoglycoproteins of the human erythrocyte membrane that bear the antigenic determinants for the MN and Ss blood groups.
6	N-acetyltransferase 8 (NAT8) expressed in embryonic and the adult kidney are associated with nephrotoxicity response.
7	Low density lipoprotein receptor-related protein (LRP) is a receptor for protease complexes and may function in cell growth and repair and in tumor invasiveness. The LRP C766T polymorphism is protective against Alzheimer's disease, perhaps through alteration of LRP expression.
8	KLF15 is part of a heretofore unrecognized pathway regulating the cardiac response to hemodynamic stress. KLF15 binds to multiple 9 bp consensus sites in the Rhodospin and IRBP promoters including the CRS-1 and G-rich repressor elements. An <i>in vivo</i> role for KLF15 in repressing photoreceptor-specific gene expression in the inner retina is hypothesized.

TABLE III

Comparison Between SLC and LBG Algorithms

	SLC	LBG
Compactness (Fig.2 and 4)	All classes of output are compact	Some are in compact in general
Similarity of Genes (Fig.5)	Similar genes cluster tightly together	Non-similar genes may cluster together
Capability to Predict Genes (Fig.6, TABLE I)	Can be applied to predict genes	Cannot predict genes. Is rare that the clustering algorithm can directly predict genes.
Running Time	The SLC is very fast and fits to process huge amount of data	A lot of redundancy exists in running time.

The reason of fast computation of the SLC: The convergent classes will be removed from the training set immediately. In addition, the computation of entropy is a fast log-transform. Entropy convergence as the terminate criterion of algorithm cuts down running time drastically.

The redundancy of running time of LBG: All data of classes, even the data of convergent class, are involved in computing until the end of the algorithm. Running time is wasted. In addition, the computation of the average distortion that is used as the terminate criterion by many clustering algorithm wastes a lot of running time.