Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Iterative local Gaussian clustering for expressed genes identification linked to malignancy of human colorectal carcinoma

Ito Wasito[1,*], Siti Zaiton M Hashim[1] and Sri Sukmaningrum[2]

[1]Department of Software Engineering, Faculty of Computer Science and Information Systems, University Technology Malaysia, Skudai, Johor Bahru, Malaysia; [2]Faculty of Biology, University of Jenderal Soedirman Purwokerto, Central Java, Indonesia; Ito Wasito* - E-mail: ito@gmx.co.uk; *Corresponding author

received December 14, 2007; accepted December 28, 2007; published online December 30, 2007

Abstract:
Gene expression profiling plays an important role in the identification of biological and clinical properties of human solid tumors such as colorectal carcinoma. Profiling is required to reveal underlying molecular features for diagnostic and therapeutic purposes. A non-parametric density-estimation-based approach called iterative local Gaussian clustering (ILGC), was used to identify clusters of expressed genes. We used experimental data from a previous study by Muro and others consisting of 1,536 genes in 100 colorectal cancer and 11 normal tissues. In this dataset, the ILGC finds three clusters, two large and one small gene clusters, similar to their results which used Gaussian mixture clustering. The correlation of each cluster of genes and clinical properties of malignancy of human colorectal cancer was analysed for the existence of tumor or normal, the existence of distant metastasis and the existence of lymph node metastasis.

Keywords: gene expression; unsupervised clustering; Gaussian kernel; colorectal cancer

Background:
Gene expression profiling is an effective approach to extract useful information from a large number of simultaneously expressed genes within specific cell types. This approach is not only useful for investigating a known biological cell, it also can be applied to explore unknown biological cells in relation to specific gene functions **[1]**. Comprehensive profiles of mRNA levels can be obtained and used to discriminate cancer cells from normal cell, and to provide sub-classes of tumor types. The possibility of measuring thousands of simultaneously expressed genes represents a challenge in terms of analysis and interpretation. One useful application is the identification of genes whose expression levels are associated with human colorectal carcinoma where there is still limited knowledge of the biological and clinical properties of malignancy **[4]**. This solid tumor is one of the most prevalent and well-characterized human cancers, and, in spite of recent advances in diagnosis and therapeutics, is still a leading cause of death **[3]**.

Clustering is a powerful exploratory technique for the analysis of gene expression profiles. In the past decades, a number of clustering algorithms have been proposed in this context including Hierarchical Clustering **[1, 2]** and Gaussian Mixture Clustering **[3]**. The Hierarchical cluster analysis is probably the most popular and powerful method for unveiling underlying features of gene expression profiles. However, because of a lack of valid statistical evaluation methods, the results are subject to interpretation by the investigator. Gaussian Mixture Clustering is also another powerful approach. This parametric clustering method has been applied to gene expression linked to malignancy of human colorectal carcinoma with promising results **[3]**. However, this cluster analysis requires prior information about the number of clusters in the dataset, which is often not realistically possible.

In this paper, a density-based clustering method for uncovering underlying structures of gene expression data will be explored. The advantages of of this method called Iterative Local Gaussian Clustering (ILGC) includes the simplicity of the technique, no need for prior information on the number of clusters, and the requirement of only one parameter, the nearest neighbour.

Methodology:
Through density-based estimation, we try to approximate the 'true' density of genes. Basically, there are two main approaches to implement density estimation: parametric and non-parametric. The first approach was implemented by Muro *et al* **[3]** using Gaussian Mixture clustering and Bayesian framework with promising results. However, to avoid the requirement for information with regards number of clusters in advance, we used the non-parametric-based approach to determine the density of genes.

The original form of density based approach can be formulated as in equation (1) (see supplementary material)

© 2007 Biomedical Informatics Publishing Group

Due to its simplicity, the K-nearest neighbour based is one of the most popular non-parametric-based approach **[5, 6, 7]**. In this report, we extend the K-nearest neighbour (KNN) density estimation combined with Gaussian kernel function. In the proposed method, the KNN would contribute in determining the 'best' local genes iteratively for Gaussian kernel density estimation. The local best is defined as the set of neighbours genes that maximizes the Gaussian kernel function. This leads to an alternative non-parametric clustering approach that is called iterative local Gaussian clustering (ILGC).

Iterative local Gaussian clustering

Basically, Gaussian kernel function for genes clustering has basic form as in equation (2) (see supplementary material).

There are two main rules to deal with this problem of selecting the best local genes: KNN-rule and Bayesian-rule. In ordinary KNN density estimation, the KNN-rule is applied to assign a target gene to a certain cluster based on the majority of number of gene neighbours criterion. On other hand, ILGC implements a Bayesian decision rule such that the target gene will be assigned to the *c-th* cluster, if the majority of k-neighbours of the target gene maximizes the density function, $K_c(x)$. To do this, we perform the rule iteratively using the inequality illustrated in equation (3) (see supplementary material).

Note that we do not use the scale parameter term explicitly in the equation as it will be determined in k-nearest neighbour selection process. The iterative local Gaussian clustering algorithm can be summarized follows:

ILGC Algorithm (Database, k neighbours)

1) Set the number clusters to the N "informative gene" selected
2) Each gene x_i (i=1....N) with k neighbours is assigned to cluster c as in equation (4) (see supplementary material)
3) If there is no change in the cluster structure, iterations have converged. Re-index the clusters and stop. Otherwise go to step 4.
4) Re-calculate cluster membership in equation (3) (supplementary material) then go to step 2.

Data imputation

To implement ILGC algorithm, there are number of missing entries in the original datasets which we fill in. We apply the INI algorithm **[6, 7]** to impute these missing data entries. This method is based on a least squares principle. This approach minimizes the sum of squared differences between the data entries and those reconstructed via bilinear modelling which is akin to the singular value decomposition (SVD) of a data matrix. Details of INI algorithm can be obtained elsewhere **[6, 7]**.

Gene selection

Another issues addressed in our implementation of cluster analysis is the "noisy" gene which is not so informative. We use a Correlation Ratio (CR) method as illustrated in equation (5) in the supplementary material to select the informative genes **[3]**.

Discussion:

In this work, we used the informative genes selected by Muro *et al* **[3]** which consists of 341 genes out 1536 genes and 100 cancerous samples and 11 normal samples with their clinical parameters. Using ILGC with 10 number of nearest neighbour and 95% of rate convergence, three clusters were found, similar to the Gaussian Mixture Clustering results of Muro *et al* **[3]**. However, the ILGC uncovers a different structure of clusters compared to those found by the Gaussian Mixture method. The structure of clusters can be visualized in 2-D graph based on plotting the first and second component of principal component analysis (PCA) as shown in Figure 1. The results show that there are two large numbers of genes clusters and one small cluster.

© 2007 Biomedical Informatics Publishing Group

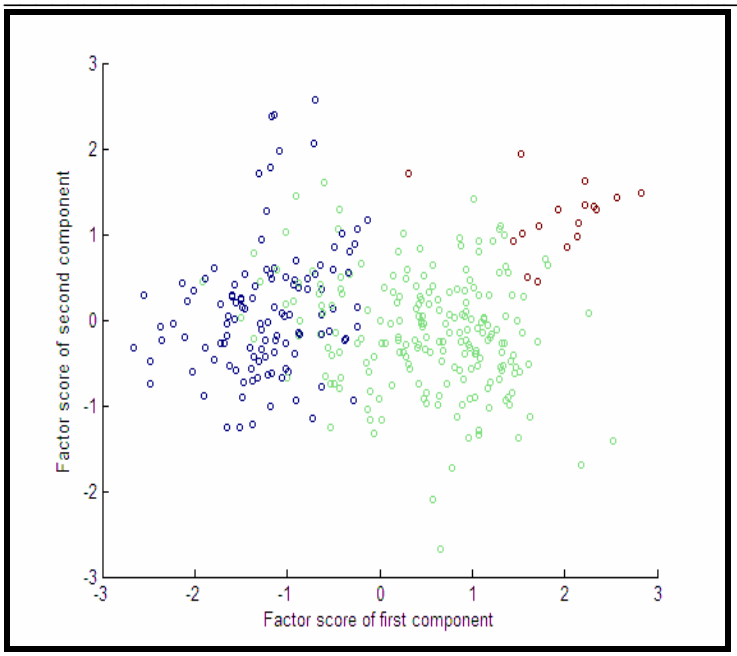


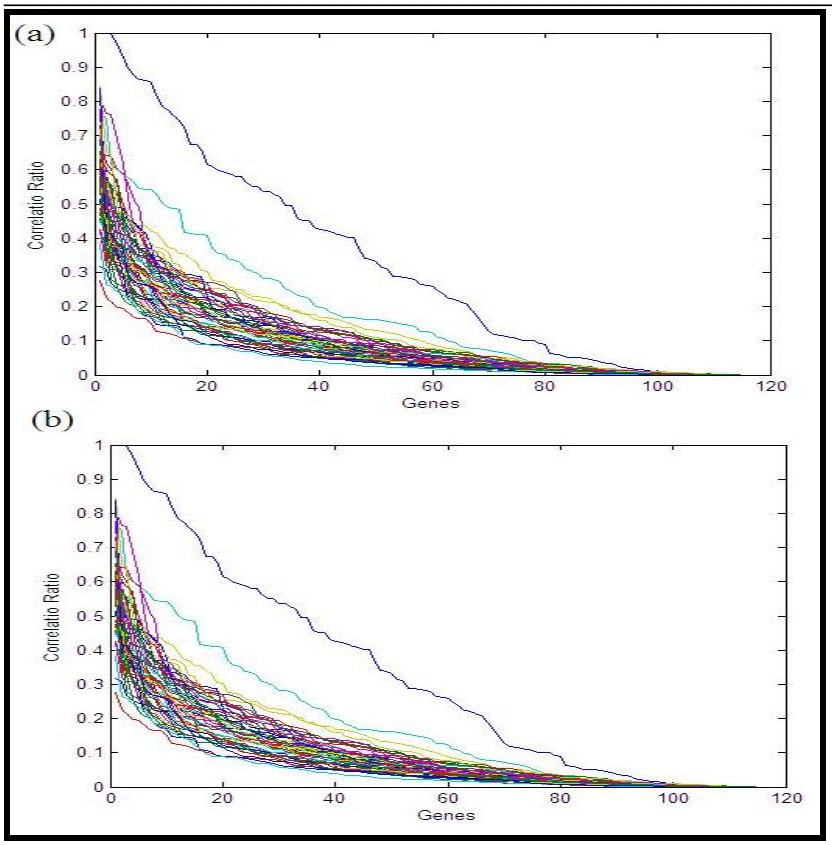
Figure 1: The structure of clusters those to be found by ILGC algorithm. Green, blue and red represent cluster I, cluster II and cluster III, respectively.

For the two large numbers of clusters, cluster I and cluster II, further analysis was carried out to detect any relationship to the cancer clinical parameters: cancerous or normal, distant metastases and lymph node metastasis. Correlation Ratio (CR) analysis was used, based on the following procedure: (a) Calculate CR value for each gene in cluster I and II; (b) Sort genes with key CR value from (a); (c) Permute sample position for each gene, then calculate CR of the permuted samples; and (d) Draw all CR values from (b) and (c).

Figure 2 shows that cluster I and II correlate to the differences between cell tissues that contain tumour or normal. Figures 3a and Figure 3b show that the cluster I and II have significant correlation with the existence of distant metastasis in cell tissues. However, cluster I and cluster II have no correlation to the existence of lymph-node metastasis in cell tissues (Figure 4a and Figure 4b).

Since cluster III contains only a small number of genes (17), we use the difference correlation analysis technique. Since this cluster contains TCL (tumor classifier) genes, this cluster appears to correlate with the existence of tumor. Figure 5 shows that when distant metastasis exists, cluster III correlates to the third colorectal clinical parameter.

© 2007 Biomedical Informatics Publishing Group



Figures 2: Cluster I and II have correlation to the differences between cell tissues that contain tumor or normal. The vertical axis represents CR-value of the differences of cell tissues which contains cancer and normal in cluster I (a) and cluster II (b). The horizontal axis represents sorted genes based on their CR-values. The top blue line represents clusters found by ILGC, others represent permuted samples.

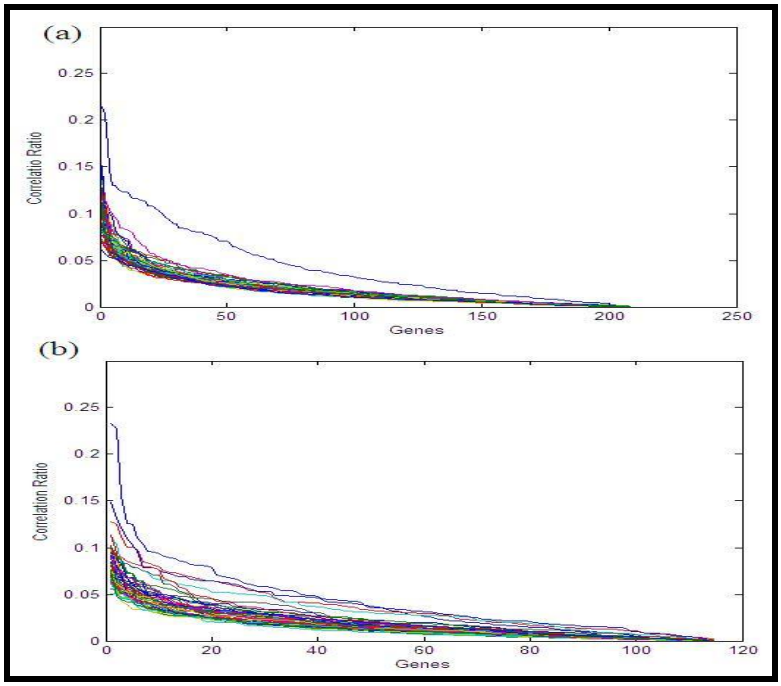


Figure 3: The vertical axis represents the CR-value of the differences of cell tissues which contain distance metastasis: cluster I (a) and cluster II (b). The horizontal axis represents sorted genes based on their CR-values. The top blue line represents clusters found by ILGC; others represent permuted samples.

Bioinformation, an open access forum
© 2007 Biomedical Informatics Publishing Group

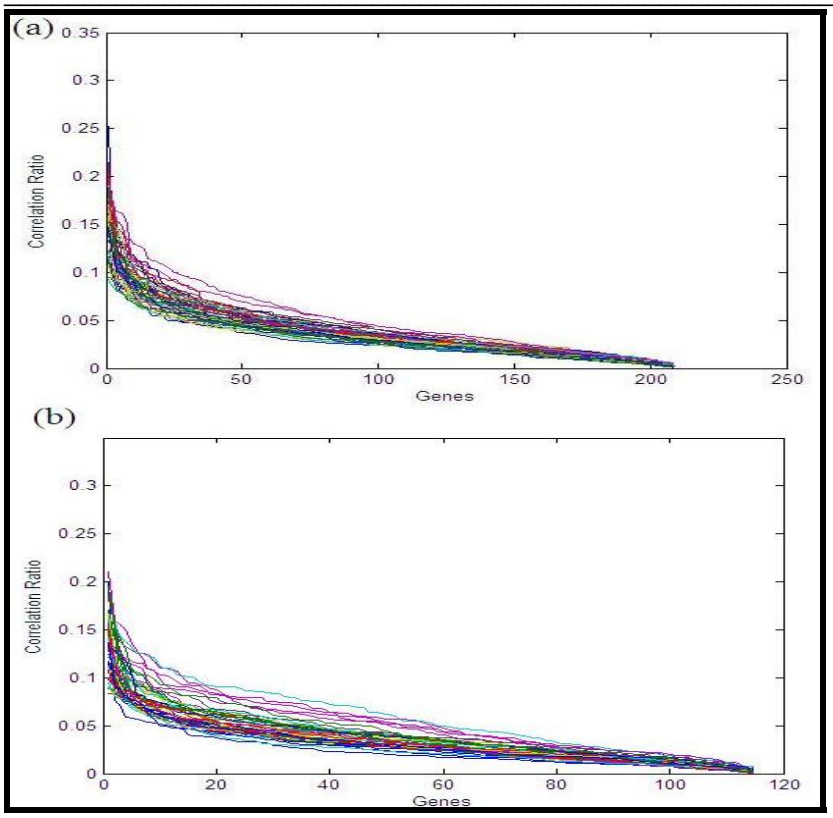


Figure 4: The vertical axis represents CR-value of the differences of cell tissues which contain lymph node metastasis: cluster I (a) and cluster II (b). Horizontal axis represents sorted genes based on their CR-values. No correlation to the existence of lymph-node metastasis in cell tissues is observed.

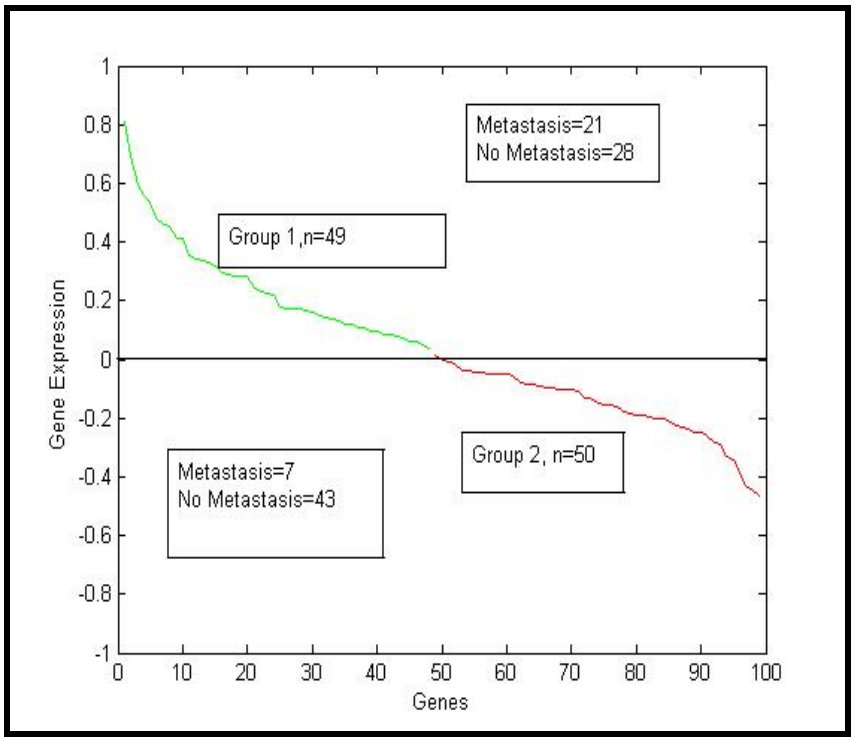


Figure 5: Linkage of the clusters of expressed genes to the existence of distant metastasis in cluster III using the difference correlation analysis technique.

Bioinformation, an open access forum
© 2007 Biomedical Informatics Publishing Group

Conclusion:

In this paper, we explored a non-parametric density based clustering technique which is called iterative local Gaussian clustering (ILGC). The advantages of ILGC includes: the simplicity of the technique, no requirement for prior information on the number of clusters and the use of a single parameter, the nearest neighbour.

ILGC algorithm has been tested on the colorectal carcinoma database of Muro *et al.*, 2003 **[3]**. The results show that the proposed method produced the same number of clusters as those found by Muro *et al*. In addition, the clusters found by ILGC were able to be linked to malignancy of human colorectal carcinoma which include the existence of tumor and distant metastasis.

Further work is needed to compare ILGC experimentally with other existing clustering techniques such as Hierarchical clustering, Gaussian Mixture clustering and K-Means for identification of other cancers.

Acknowledgement:

This work was supported by a research grant from the Higher Education Directorate General, Ministry of Education of Indonesia, 2005. The authors gratefully acknowledge many helpful comments by reviewers that have been very helpful in improving the publication.

References:

[01] P. O. Brown & D. Botsein, *Nature Genetics Supplement.,* 21: 33 (1999)

[02] M. B. Eisen, *et al.*, *Proc. Natl. Acad. Sci.,* 95: 14863 (1995) [PMID: 9915498]

[03] S. Muro, *et al*., *Genome Biology*, 4: R21 (2003) [PMID: 12620106]

[04] D. A. Notterman, *et al.*, *Cancer Res.*, 61: 3124 (1999)

[05] N. Tranh, *et al.*, *Computational Statistics and Data Analysis*, 51: 513 (2006)

[06] I. Wasito & B. Mirkin, *Information Sciences,* 1: 1 (2005)

[07] I. Wasito & B. Mirkin, *Computational Statistics and Data Analysis*, 50: 926 (2006)

Edited by O. Miotto, T. W. Tan & S. Ranganathan

Citation: Wasito *et al.*, Bioinformation 2(5): 175-181 (2007)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

© 2007 Biomedical Informatics Publishing Group

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Supplementary material

Equations

$$\hat{p}(x) = \frac{1}{V}\sum_{i=1}^{n} K_i(x) \quad \rightarrow \quad (1)$$

Where $\hat{p}$, V and K represent density estimation, volume of data and kernel function, respectively.

$$K_c(x) = \frac{1}{n}\sum_{i=1}^{n} e^{1/2(x-x_i)^2} \quad \rightarrow \quad (2)$$

where x_i, n and K_c are *i-th* gene, number of genes in *c-th* cluster and density function of the *c-th* cluster. The main issue of this framework is the selection of the best local genes to maximize $K_c(x)$.

$$\sum_{x_i \in c_i} e^{-1/2(x-x_i)^2} > \sum_{x_i \in c_j} e^{-1/2(x-x_i)^2} \quad \rightarrow \quad (3)$$

where x_i and c_i are target genes and the *i-th* cluster , respectively where $i \neq j$.

$$Clus_c = Max(\hat{p}(x_i \mid c)) \quad \rightarrow \quad (4)$$

where $\hat{p}(x_i \mid c)$ is a class-conditional density function at x_i for each cluster c.

$$(CR_i)^2 = \frac{\sum_{c=1}^{C} n_c((\sum_{j \in J_c} x_{i,j})/n_c - \bar{x}_i)^2}{\sum_{j=1}^{M}(x_{i,j} - \bar{x}_i)^2} \quad \rightarrow \quad (5)$$

Correlation Ratio (CR) in a particular class Jc; x_{ij} is the expression level of gene i in sample j; and $\bar{x}_i$ is the average expression level of gene i.

Bioinformation, an open access forum
© 2007 Biomedical Informatics Publishing Group