BMC Bioinformatics

BioMed Central

Methodology article

Open Access

NIFTI: An evolutionary approach for finding number of clusters in microarray data

Sudhakar Jonnalagadda and Rajagopalan Srinivasan*

Address: Department of Chemical and Biomolecular Engineering, National University of Singapore, 10 Kent Ridge Crescent, 119260 Singapore

E-mail: Sudhakar Jonnalagadda - sudhakar@nus.edu.sg; Rajagopalan Srinivasan* - chergs@nus.edu.sg

*Corresponding author

Published: 30 January 2009

BMC Bioinformatics 2009, **10**:40 doi: 10.1186/1471-2105-10-40

Received: 11 July 2008

Accepted: 30 January 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/40

Abstract

Background: Clustering techniques are routinely used in gene expression data analysis to organize the massive data. Clustering techniques arrange a large number of genes or assays into a few clusters while maximizing the intra-cluster similarity and inter-cluster separation. While clustering of genes facilitates learning the functions of un-characterized genes using their association with known genes, clustering of assays reveals the disease stages and subtypes. Many clustering algorithms require the user to specify the number of clusters a priori. A wrong specification of number of clusters generally leads to either failure to detect novel clusters (disease subtypes) or unnecessary splitting of natural clusters.

Results: We have developed a novel method to find the number of clusters in gene expression data. Our procedure evaluates different partitions (each with different number of clusters) from the clustering algorithm and finds the partition that best describes the data. In contrast to the existing methods that evaluate the partitions independently, our procedure considers the dynamic rearrangement of cluster members when a new cluster is added. Partition quality is measured based on a new index called Net InFormation Transfer Index (NIFTI) that measures the information change when an additional cluster is introduced. Information content of a partition increases when clusters do not intersect and decreases if they are not clearly separated. A partition with the highest Total Information Content (TIC) is selected as the optimal one. We illustrate our method using four publicly available microarray datasets.

Conclusion: In all four case studies, the proposed method correctly identified the number of clusters and performs better than other well known methods. Our method also showed invariance to the clustering techniques.

Background

Clustering is a statistical technique that partitions a large number of objects into a few clusters such that objects within the same cluster are more similar to each other than to the objects in other clusters. Clustering is widely used in gene expression data analysis to cluster genes and/or samples (assays) based on their similarity in expression patterns. Since gene clusters are often enriched with genes involving in common biological processes, identifying such clusters discloses potential roles of previously un-characterized genes and provides insights into gene regulation. Similarly, clustering of samples reveals different stages or subtypes of diseases such as cancer leading to development of customized diagnostic procedures and therapies.

Despite the widespread use of clustering algorithms in gene expression data analysis [1-6], selection of clustering parameters continues to be a challenge. In many cases, the optimal specification of number of clusters, k, is difficult especially if there is inadequate biological understanding of the system [7]. A suboptimal specification of number of clusters can generally result in misleading results – either all classes may not be identified or spurious classes may be generated [8]. While the correct number of clusters can be identified by visual inspection in some cases, in most gene expression datasets, the data dimensions are too high for effective visualization. Hence, methods that find the optimal number of clusters are essential.

Several methods have been proposed for finding the number of clusters in data. The popular methods evaluate the partition using a metric and optimize it as a function of number of clusters. Comprehensive reviews of these methods are available elsewhere [9-11]. Here we briefly describe some recent methods recommended for gene expression data analysis. Tibshirani et al. [12] proposed the gap statistic that measures the difference between within-cluster dispersion and its expected value under the null hypothesis. The k that maximizes the difference is selected. Since the gap statistic uses within-cluster sum of squares around the cluster means to evaluate the within-cluster dispersion, this method is suitable for compact, well separated clusters. Dudoit and Fridlyand [13] proposed a prediction based re-sampling method for finding the number of clusters. For each value of k, the original data is randomly divided into training and testing sets. The training data is used to build a predictor for predicting the class labels of the test set. The predicted class labels are compared to that obtained by clustering of test data using a similarity metric. This value is compared to that expected under an appropriate null distribution. The k for which the evidence of significance is the largest is selected. Ben-Hur et al. [14] proposed a similar re-sampling approach where two random subsets (possibly overlapping) are selected from the data. The two random subsets are subsequently clustered independently and the similarity between the resulting partitions is measured. The similarity is measured for multiple runs and its distribution is visualized for each k. The optimal number of clusters is selected where transition from high to low similarity occurs in the distribution. The approach of Dudoit and Fridlyand as well as Ben-Hur et al. assume that the sample subset can represent the inherent structure in the original data which may not be true for small clusters. Furthermore, the user has to manually locate the transition in Ben-Hur et al. approach.

Recently, Bolshakova and Azuaje [15] employed Silhouette [16], Generalized Dunn's index [8], and Davies-Bouldin index [17] on gene expression data. These methods use the intra- and inter-clusters distances to identify the best partition. In general, cluster validation is easier when the underlying clusters are well separated. But, most cluster validation methods lead to suboptimal results when inter- and intra-cluster distances vary largely. To illustrate this, consider the artificial dataset in Figure 1 consisting of 600 objects in three clusters (A, B, and C). Clusters B and C are closer to each other and far from Cluster A. Figure 2 shows the results of Silhouette, normalized Dunn's and normalized Davies-Bouldin indices for this dataset. For ease of visualization, all indices have been min-max re-scaled to [0 1]. For a given index value $I_k(k = 1,2,3, \ldots k_{max})$, the re-scaled index value is obtained as

$$\hat{I}_k = \frac{I_k - min(I_k)}{max(I_k) - min(I_k)} \quad (1)$$

Silhouette, Generalized Dunn's index, and Davies-Bouldin indices incorrectly identified only 2 clusters in this dataset. A partition with two clusters $\{A\}$ and $\{B \cup C\}$ is more favorable according to intra- and inter-cluster distance based methods. Gene expression data contain clusters of different sizes, shapes, and there exist smaller clusters within the larger well-separated cluster [18]. Hence, the methods for finding number of clusters based on intra- and inter-cluster distances do not perform well for gene expression data (see results). This finding motivates development of new methods that do not rely on intra- and inter-cluster distances.

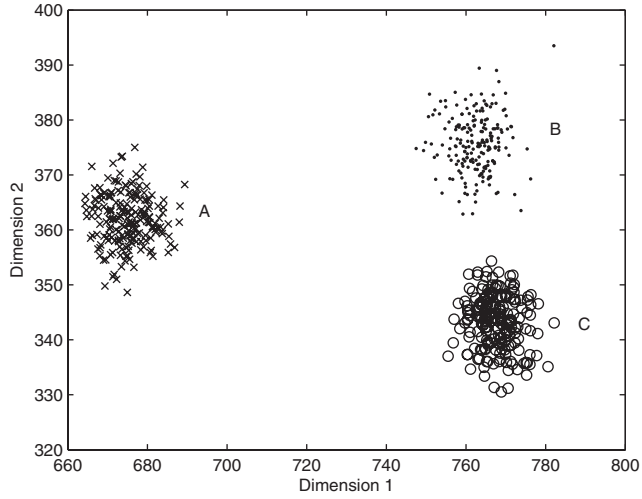

Figure 1
Two dimensional artificial dataset with 3 inherent clusters (A, B, and C). Clusters B and C are closer to each other and far from Cluster A.

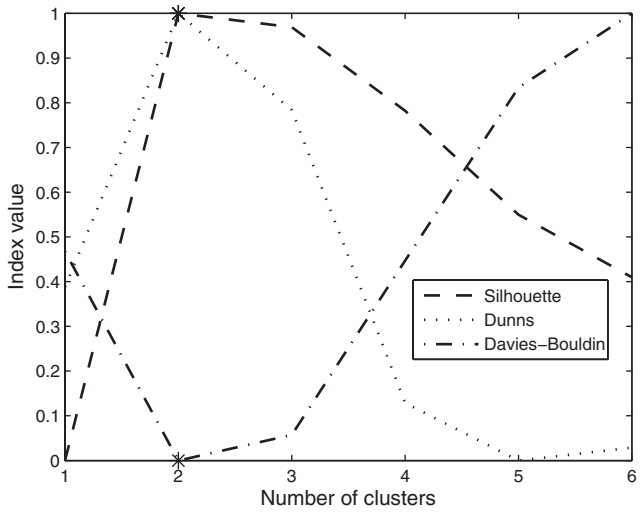

Figure 2
Cluster validation results for the artificial dataset in Figure 1. All three indices, Silhouette (dash line), Dunn's (dot line), and Davies-Bouldin (dash-dot line) incorrectly predict 2 clusters although the underlying data can be seen to have 3 clusters (* indicates the optimal number of clusters predicted by specific index).

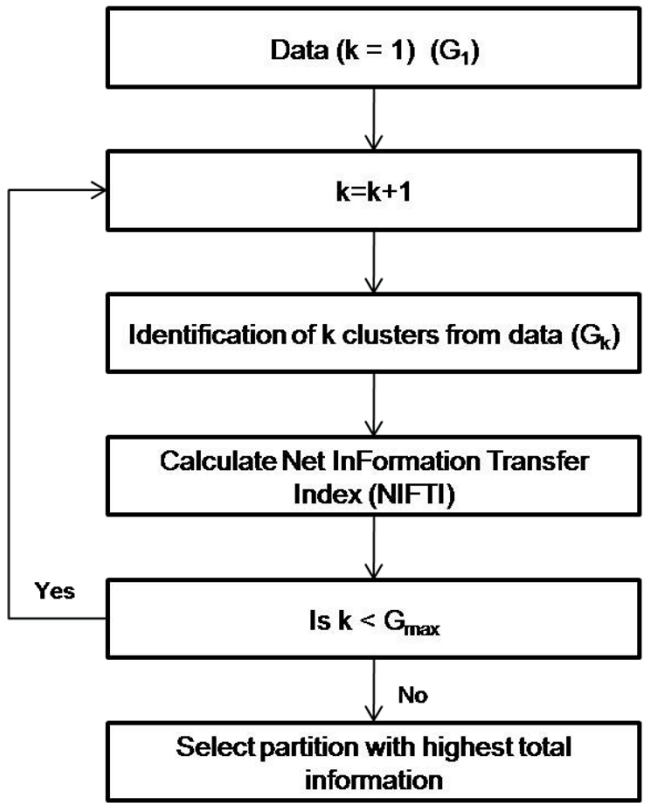

Figure 3
Proposed cluster validation procedure. The procedure starts with unclustered data (G_1). In each subsequent generation, an additional cluster is added and the data reclustered. The Net InFormation Transfer calculated based on the evolution of objects during the generation. This procedure is carried out for a predefined number of generations (G_{max}). Finally the partition with highest total information is selected as the optimal partition.

In this paper, we propose a new method to find optimal number of clusters in the data. Our approach is based on an evolutionary view of the clustering process (Figure 3). We start by considering the whole dataset as a single cluster and notate it as Generation 1 (G_1). In each subsequent generation, the number of clusters, k, is incremented by one and the data re-clustered. A generation with k clusters is notated as G_k. The net change in the information content due to the addition of a cluster is measured using Net InFormation Transfer Index (NIFTI). NIFTI includes two components – *direction* of information change and *magnitude* of information change – in its calculation. The *direction* of change indicates whether information is gained or lost during evolution. The *magnitude* indicates the extent of change. During evolution, objects from i^{th} cluster, C_k^i, in the current generation, G_k, will be distributed across several clusters in the next generation, G_{k+1}. The clusters in G_{k+1} that receive objects from C_k^i are called as offspring of parent cluster C_k^i. NIFTI considers this rearrangement of cluster members when a new cluster is added for calculating the information change. The net information change is the sum of the information change for all parent clusters. Information increases if offspring clusters are separable. We use a simple but effective procedure with statistical basis to check the separability of offspring clusters. The *magnitude* of information change is calculated using information theory. This evolutionary procedure is carried out for a predefined number of generations (G_{max}). The Total Information Content, *TIC*, of a partition is defined as the cumulative information gained till that generation. A partition with the highest *TIC* is selected as the best partition. While testing for separability of clusters, NIFTI does not give weightage for largely separated clusters or penalize marginally separated clusters, thus eliminates the problems associates with varying inter-cluster distances.

Results

Four publicly available microarray datasets are used to illustrate the performance of the proposed approach. The first two datasets are time-course datasets. In time-course datasets, genes are clustered based on their similarity in expression patterns. The other two datasets contain data from different samples.

Two different clustering techniques, namely k-means and model-based, are used for generating partition with different number of clusters. The distance metrics used

for clustering are the same as those used by the data publishers *i.e* Pearson coefficient for first, third, and fourth case studies and standard correlation coefficient for the second dataset. In all the case studies, the maximum number of generations, G_{max} is selected as [19]:

$$G_{max} \leq \sqrt{N} \quad (2)$$

where N is the number of objects to be clustered.

Case Study I : Yeast cell-cycle data

The Yeast cell-cycle dataset was generated by Cho et al. [20]. Oligonucleotide microarrays were used to monitor the expression levels of all known and predicted Yeast genes during two cell-cycles. Expression levels were measured at 17 time points with a time period of 10 min. The aim of this experiment was to identify the cell-cycle controlled genes in Yeast. Cho et al. visually observed the highly variant genes for consistent periodicity during the cell-cycle and identified 384 genes. These 384 genes were classified into five classes – early G1, late G1, S, G2, and M phases – based on their peak expression. Since the number of clusters is known for this dataset, the 384 cell-cycle genes are used to validate the proposed method.

The proposed method, NIFTI, correctly identifies five clusters in this dataset using k-means method (Figure 4). For comparison, the results for Silhouette, Dunn's, and Davies-Bouldin indices are shown in Figure 4. All three indices predict 4 clusters in this data. The reason is as follows. At k = 4, genes from S and G2 phases are combined into one cluster while those from Early G1, Late G1, and M phases are clustered correctly. These four clusters are well-separated. When the number of clusters is increased to 5, while S and G2 clusters are identified correctly, the inter-cluster distance is small. The three methods therefore identify the partition with four clusters as optimal. In contrast to these distance based methods, the proposed method gives no weightage for larger inter-cluster distances and correctly identifies 5 clusters.

The five clusters identified by k-means clustering correspond to the five phases of cell-cycle – early G1, late G1, S, G2, and M phases. For example, cluster 1 contains the cell-cycle regulated genes including PCL9, SIC1 and DNA replication genes CDC6 and CDC46 that are classified into early G1 by cho et al. [20]. The mean expression profile of this cluster shows single peak during the early stage of G1 (Figure 5). Similarly, other clusters are also enriched with genes that are classified into one of the reported clusters and their mean expression profiles peak during the corresponding stages (Figure 5).

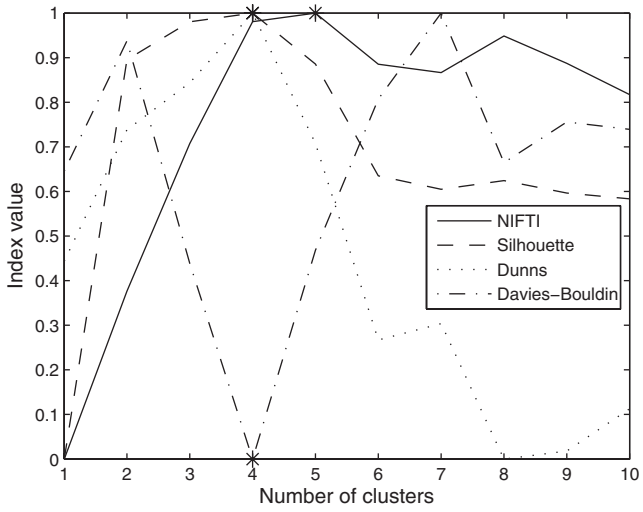

Figure 4
Results for Yeast cell-cycle dataset using k-means clustering. NIFTI (solid line) correctly finds 5 clusters in this dataset. Silhouette (dash line), Dunn's (dot line), and Davies-Bouldin (dash-dot line) indices predict only 4 clusters.

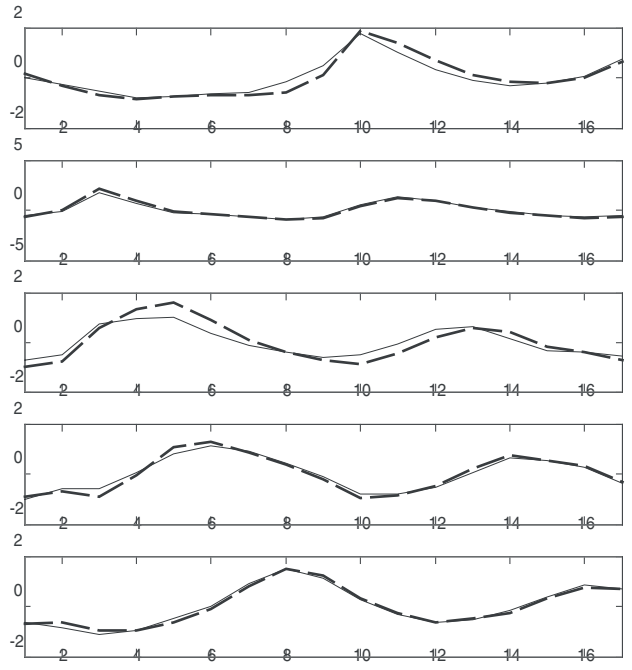

Figure 5
Mean expression levels of Yeast cell-cycle clusters. Solid line represents the mean expression profile of clusters reported by [20] and dash line corresponds to the optimal clusters from NIFTI. A strong similarity between the two can be observed.

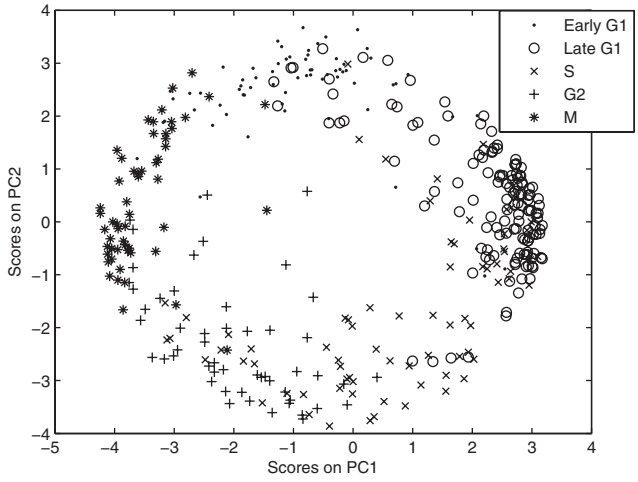

Figure 6
Scores plot of Yeast cell-cycle dataset. The first two PCs capture 65% variance.

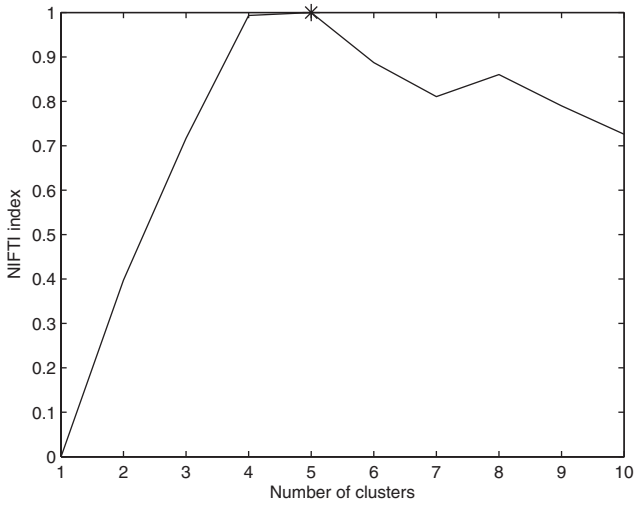

Figure 7
Results for Yeast cell-cycle dataset using model-based clustering. NIFTI correctly finds 5 clusters in this dataset.

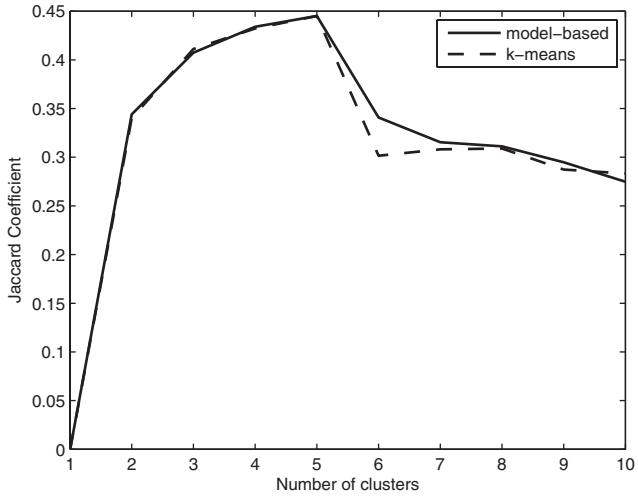

Figure 8
Jaccard Coefficient for Yeast cell-cycle dataset. The JC has a maximum at k = 5 indicating that there are 5 clusters.

However, some of the genes especially S phase genes are found to be 'mis-classified' by k-means clustering algorithm. To understand the discrepancy, we used Principal Component Analysis and plotted the scores with the first two dominant Principal Components (Figure 6). From Figure 6, it is clear that some of the genes from reported classes, especially S phase genes, are distributed to other classes. The k-means algorithm put those genes in appropriate classes which explains the mismatch between the reported and k-means partitions [see additional file 1].

Results for this dataset using model-based clustering are shown in Figure 7. NIFTI correctly identifies 5 clusters using model-based clustering as well. Since the 'true' (reported) partition is available for this dataset, we compare the clustering results using k-means and model-based clustering with reported partition using Jaccard Coefficient (JC) which measures the similarity between two partitions. Let C be the partition from the clustering algorithm and P be the reported solution. The JC measures the extent to which C matches with P

$$JC = \frac{n_{11}}{n_{11}+n_{10}+n_{01}} \quad (3)$$

where n_{11} is the number of pairs of objects that are in the same cluster in both C and P, n_{10} is the number of pairs of objects that are in the same cluster in C but not in P, and n_{01} is the number of pairs of objects that are in the same cluster in P but not in C. JC takes a value between 0 (complete mismatch) and 1 (perfect match). The better the agreement between identified and the 'true' solution, the higher the value of JC. Figure 8 shows the JC for Yeast cell-cycle five phase criterion data as a function of number of clusters using k-means and model-based algorithm. The JC takes a maximum value of 0.445 at k = 5 indicating that in the given range of k the extracted partition best matches with the reported one. This clearly shows that the 5 clusters identified using proposed method are correct.

Case Study 2 : Serum data

The Serum gene expression dataset is reported by Iyer et al. [21]. In this study, the response of human fibroblasts

to serum was measured using microarrays containing around 8000 probes. Iyer et al. employed filtering techniques and shortlisted 517 most variant genes. They used hierarchical clustering and identified 10 clusters in this dataset using visualization tools. We use these 517 genes in this case study.

NIFTI identifies 6 clusters in this dataset using k-means clustering (Figure 9). This result is supported by an other independent study using a graph-theoretical clustering algorithm [6]. The Silhouette, Dunn's and Davies-Bouldin indices identify only 2 clusters in the dataset (Figure 9). This dataset is more complex than the previous one. It contains two large clusters – one with up-regulated genes and another with down-regulated genes. All the other clusters are embedded in these large clusters. The ratio of difference between the intra- and inter-clusters distances is highest at $k = 2$. So any distance based method will generally identify only two clusters in this dataset. Multiple peaks were observed for NIFTI index for this dataset with the highest peak at $k = 9$ when model-based clustering is used for generation partitions (Figure 10). However, the Jaccard Coefficient between the partitions from model-based clustering and the reported partition has the highest value at $k = 6$ (Figure 11) indicating 6 clusters in this dataset.

In the next two case studies, the datasets contain gene expression data from different cancer samples. In these datasets, samples are clustered based on their similarity in expression patterns. Model-based clustering is not suitable for these datasets as it uses covariance matrix in its computation. The estimation of covariance matrix is inaccurate for sample clustering as the number of samples in each cluster are very small. So results are given for only k-means clustering.

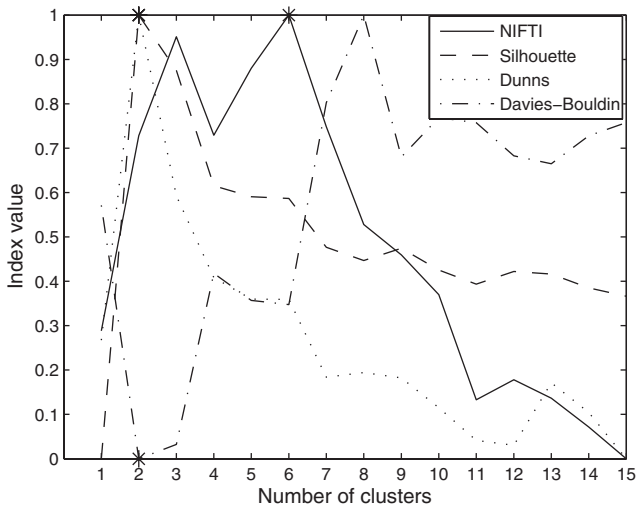

Figure 9
Results for Serum dataset using k-means clustering. NIFTI (solid line) predicts 6 clusters. Silhouette (dash line), Dunn's (dot line), and Davies-Bouldin (dash-dot line) estimate only 2 clusters.

Case Study 3 : Lymphoma data

The lymphoma dataset was reported by Alizadeh et al. [22]. In this experiment, cDNA microarrays were used to

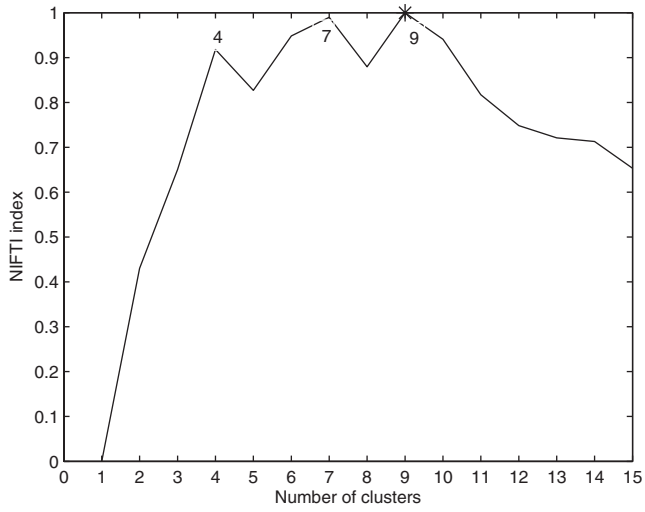

Figure 10
Results for Serum dataset using model-based clustering. NIFTI index has multiple peaks with a maximum peak ak $k = 9$. However, the Jaccard coefficient between the partition from model-based clustering and expert partition has maximum at $k = 6$ (Figure 11).

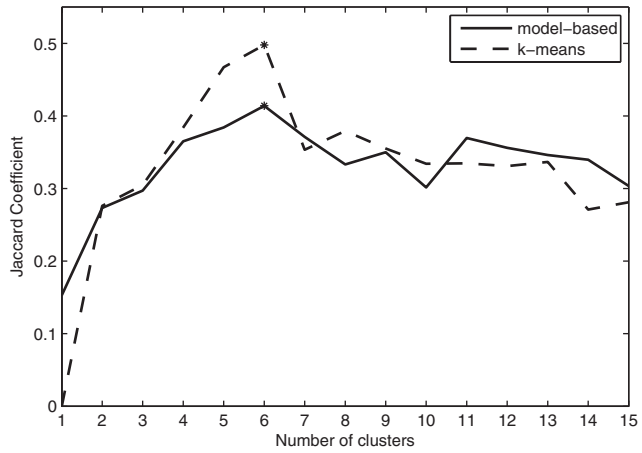

Figure 11
Jaccard Coefficient for Serum dataset. The Jaccard Coefficient for Serum dataset has maximum at number of clusters $k = 6$ indicating that identifying 6 clusters is correct.

characterize gene expression patterns in adult lymphoid malignancies. After filtering, the final data contain 4026 genes whose expression levels were measured using 96 arrays. The dataset comprises samples from three prevalent adult lymphoid malignancies – Diffuse Large B-cell Lymphoma (DLBCL), Follicular Lymphoma (FL), and Chronic Lymphocytic lymphoma (CLL). For comparison, the normal lymphocyte subpopulation under a variety of conditions is also included. The objective of the study was to identify if the presence of malignancy and its type can be identified from gene expression patterns. Alizadeh et al. used hierarchical clustering for clustering the samples and identified two distinct subtypes of DLBCL-Germinal Center B-like DLBCL and Activated B-like DLBCL.

NIFTI finds 4 clusters in this dataset using k-means clustering algorithm with Pearson correlation as the distance measure (Figure 12). Not surprisingly, these four clusters correspond to the four distinct branches of the dendrogram reported in [22]. Two of these clusters contain the samples from two subtypes of DLBCL namely germinal center B-like DLBCL and activated B-like DLBCL. The third cluster contains all FL and CLL samples along with the resting blood samples. Most of the cell-cycle control genes, checkpoint genes and DNA synthesis genes that are defined as 'proliferation signature' by [22] are under expressed in these samples. This makes these samples distinct from DLBCL samples in which the proliferation signature genes are up-regulated. The fourth cluster comprises the remaining normal lymphocyte subpopulation under different activation conditions. However, the transformed cell line samples which are grouped with other normal sub-populations by [22] are clustered with DLBCL samples by k-means. The over-expression of proliferation signature genes in these samples might be the reason that they appear 'closer' to DLBCL samples to k-means. Nevertheless, k-means clustering correctly clustered two out of the three DLBCL samples that were incorrectly clustered by the hierarchical clustering.

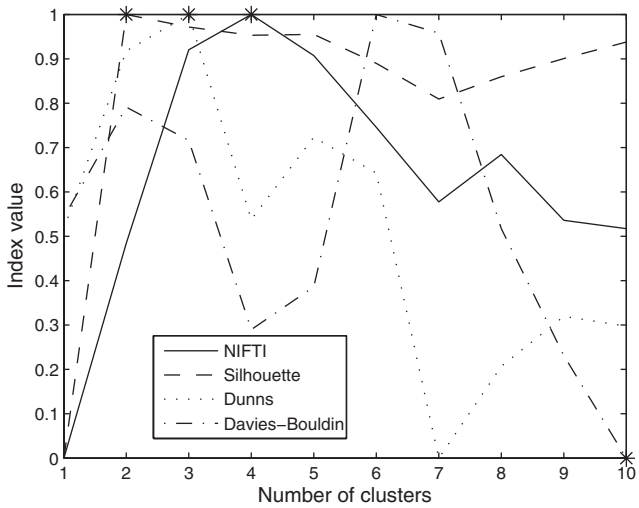

Figure 12
Results for Lymphoma dataset. NIFTI (solid line) finds 4 clusters in this dataset. Silhouette (dash line) identifies 2 clusters. Dunn's (dot line) predicts 3 clusters. Davies-Bouldin (dash-dot line) predicts 4 clusters.

The Silhouette, Dunn's and Davies-Bouldin indices for this dataset are also shown in Figure 12. The Silhouette index estimates only 2 clusters and Dunn's index predicts 3. The lowest value of Davies-Bouldin index occurred at k = 10 in the range of k values tested (it continued to decrease further with increase of k). However, Davies-Bouldin index has a local minima at k = 4 indicating four clusters in this dataset. At k = 2, all DLBCL samples are grouped into one cluster and all other samples (FL, CLL, and normal) are lumped into other. At k = 3, the latter is split and normal samples are identified as the third cluster. This indicates that at k = 2 and k = 3 subclasses of DLBCL cannot be identified. Only at k = 4, the two subclasses of DLBCL are identified. This clearly shows the usefulness of proposed method to identify correct number of clusters that aids discovering novel sub-types of diseases.

Case Study 4 : Pancreas data

The Pancreas dataset used in this study was reported by [23]. In this study, cDNA microarrays were used to analyze gene expression patterns in 14 pancreatic cell lines, 17 resected infiltrating pancreatic cancer tissues (two sub types), and 5 normal pancreases. The final filtered dataset consists of 1493 genes and 36 samples.

As shown in Figure 13, Silhouette, Dunn's, and Davies-Bouldin indices estimate 2 clusters for this dataset. A partition with two clusters lumps together the normal and pancreatic cancer tissues into a single cluster. The second cluster contains all the pancreatic cancer cell lines. NIFTI estimates four clusters in this data. A partition with four clusters describes this data well: all cancer cell line samples are accurately placed in one cluster, all normal samples are grouped together, and two different cancer tissues are well separated into two clusters. Only one sample was found to be mis-clustered. This partition with four clusters also exactly matches the dendrogram reported in [23].

Discussion and Conclusion

The use of clustering techniques in gene expression data analysis is increasing rapidly. To obtain the best results

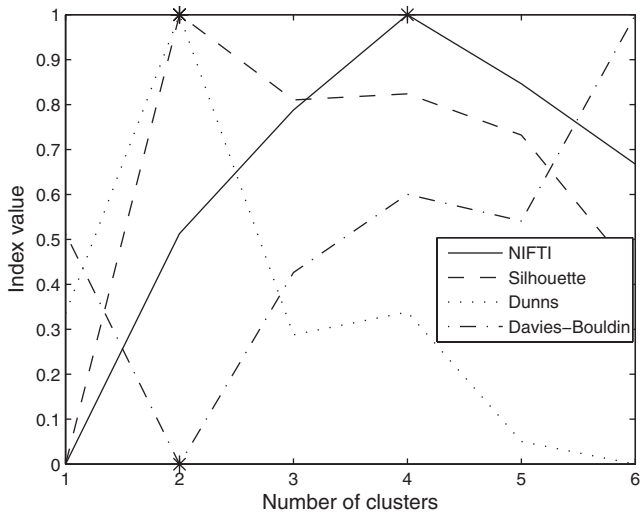

Figure 13
Results for Pancreas dataset. NIFTI (solid line) finds 4 clusters in this dataset. Silhouette (dash line), Dunn's (dot line), and Davies-Bouldin (dash-dot line) indices predict only 2 clusters.

from these clustering techniques, optimal specification of the number of clusters is essential. Hence, methods that automatically identify the number of clusters in high-dimensional gene expression data have been proposed. Methods for finding the number of clusters in a dataset can be classified as global or local methods [24]. Global methods evaluate clustering results by calculating some measure over the entire dataset whereas local methods consider pairs of clusters and test whether they should be amalgamated. The disadvantage of the global methods is that there is no definition for the measure for $k = 1$, *i.e.*, the global methods do not provide any clue whether the data should be clustered or not. Since local methods consider pairs of clusters, they can be used to decide if data should be clustered. The disadvantage of local methods is that they need a threshold value or significance level to decide whether the clusters should be amalgamated. The proposed approach combines both local and global approaches. At the local level, offspring clusters are checked for overlap and this information is converted into a global index.

The well-known methods for finding the number of clusters use within-cluster dispersion and/or inter-cluster distances. These 'distance' based methods are generally suitable when clusters are compact and well-separated but fail when sub-clusters exist. Our approach overcomes this limitation by giving no extra weightage for larger inter-cluster distances. In our approach, clusters lose or gain information based on intersection with other clusters. The actual distance between the clusters is not taken into consideration. Furthermore, the cumulative way of measuring information content of a partition ensures that information increase as long as a non-intersecting cluster can be identified.

We have compared the performance NIFTI with four other methods – Silhouette, Dunn's, Davies-Bouldin, and Gap statistic – in terms of percentage of correct prediction of actual number of clusters in artificial datasets. The synthetic datasets are generated with number of dimensions d = 2, 3 and 5 and number of clusters k = 3, 5, and 8. For each combination of d and k, 100 artificial datasets are generated and k-means clustering is used for generation of partitions. The results are given in additional file 2. For a given d, the performance of Silhouette, Dunns and Davies-Bouldin indices decreased significantly with increasing k. For example, for 2-dimensional datasets, the percentage success of these methods dropped from 70% to 20% as k increased from 2 to 8. This is mainly due to decrease in inter-cluster distance with increase in number of clusters. Similar trend of decreasing performance is observed with Gap statistic as well. Also, its performance is very poor (< 20%) with large number of clusters. In all the case studies, NIFTI performed better compared to the other methods. The performance of NIFTI is largely independent of the number of clusters and number of dimensions. This study clearly indicates the efficacy of NIFTI in predicting the number of clusters in data.

However, the proposed method has a limitation. It models clusters as hyper-spheres. Even though modeling clusters as hyper-spheres simplifies the task of finding cluster intersections, it may lead to incorrect results in case the clusters do not have a spherical shape. Nevertheless, this procedure consistently identified the 'correct' number of clusters suggesting, in part, the spherical shape of gene clusters. An independent study also reported that normalization techniques used in gene expression data analysis make the clusters spherical [4].

In this paper, the proposed method is evaluated using k-means clustering algorithm with Pearson correlation as distance measure for the Yeast cell-cycle and lymphoma datasets. The standard correlation coefficient (dot product of normalized vectors) is used for the Serum dataset. These two measures are bounded: the minimum and maximum distances are 0 and 2 respectively. On the other hand metrics such as Euclidian distance and Manhattan distance are unbounded. Hence, the affect of outliers will be high while estimating the cluster radii. This may lead to incorrect estimation of number of clusters. This can be overcome by suitable normalizing the data or selecting other ways to find cluster radius that

are less sensitive to outliers. Further study using various distance metrics and clustering techniques is needed to further evaluate the method.

Generally computational time is an important issue in determining the number of clusters. In this study, we used 100 replicates of k-means algorithm for all datasets. The time required for finding number of clusters is less than 10 minutes for all datasets on a *Pentium* 4 with 2.8 GHz processor.

Methods

Let Z_{Nxm} be the dataset to be clustered containing N objects on which m features are measured. In gene expression data analysis, N is number of genes and m is number of assays. We use a clustering algorithm to generate a series of partitions from G_1 through G_{max} with an increment of one cluster in each generation. The migration of the objects during evolution from parent clusters in G_k to their offspring in G_{k+1} forms the basis for evaluating the quality of partition in G_{k+1}. Consider the migration of objects among clusters during evolution from G_k to G_{k+1} shown in Figure 14. Three scenarios are possible during evolution:

1. All objects in C_k^i may continue to be clustered together as a single cluster in G_{k+1}. We call this phenomenon as *cluster conservation*. Example: The cluster C_k^1 is conserved as C_{k+1}^1 with all objects intact.

2. Most members of C_k^i may stay together as a single cluster in G_{k+1}, but a few escape to other clusters. This phenomenon is termed as *cluster leakage*. Example: Out of 400 objects in cluster C_k^2 most stay together in C_{k+1}^2, 15 leak to C_{k+1}^3.

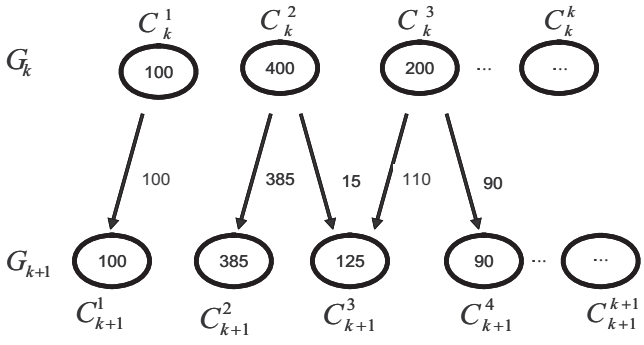

Figure 14
Behavior of cluster members during evolution. A few clusters in G_k continue as single clusters in G_{k+1} while others disassociate or undergo leakage.

3. Members of C_k^i migrate to a small number ≥ 2 of clusters in G_{k+1} such that each recipient cluster receives a significant fraction of objects. This is called as *cluster disassociation*. Example: Cluster C_k^3 disassociates to C_{k+1}^3 and C_{k+1}^4.

During evolution from G_k to G_{k+1}, some clusters are conserved, some disassociated, and others undergo leakage. The quality of the partition is measured in terms information transferred from G_k to G_{k+1} using the Net InFormation Transfer Index (NIFTI). The TIC of partition is calculated for each generation as the sum of cumulative information transferred till that generation. The partition with the largest TIC is selected as the optimal one. The TIC for a partition at $(k + 1)^{th}$ generation is given by:

$$TIC_{k+1} = TIC_k + NIFTI_{G_k \to G_{k+1}} \quad (4)$$

where $TIC_1 = 0$.

The optimal number of clusters is given by:

$$k_{optimal} = \arg\max_{1 \le k \le k_{max}} TIC_k \quad (5)$$

Net InFormation Transfer Index (NIFTI)

The Net InFormation Transfer Index during evolution from G_k to G_{k+1} is defined as the sum of the information changes of all parent clusters weighted by the fraction of total objects they contain.

$$NIFTI_{G_k \to G_{k+1}} = \sum_i^k \frac{N_k^i}{N} \times g_k^i \quad (6)$$

where N_k^i is the number of objects in i^{th} parent cluster and g_k^i is its change in information as it evolves from G_k to G_{k+1}. Equation 6 is similar to the one used by Li et al. [25] for calculating the information content of a partition.

The change in information of a parent cluster C_k^i is given by:

$$g_k^i = D_k^i \times M_k^i \quad (7)$$

D_k^i is the direction (gain or loss) and M_k^i the magnitude of information change arising from i^{th} parent cluster.

The objective of clustering is to identify clusters where objects within a cluster are more similar to each other compared to objects within other clusters. Geometrically, this means that clusters should be distant and

separable from each other in the m dimensional feature space. Here, we propose a statistical test to check whether offspring clusters are separable or not. If the offspring of parent cluster are separable from other sibling, information is deemed to have been gained during transfer and D_k^i takes +1. In contrast, if offspring are not separable, information is deemed to be lost during transfer and D_k^i is -1. In contrast to other methods, the NIFTI is not weighted as per the inter- and intra-cluster distances.

The magnitude of information change, M_k^i, is calculated using Shannon entropy given by:

$$M_k^i = \sum_{j=1}^{r} -p_k^{ij} \ln p_k^{ij} \tag{8}$$

where r is the number of offspring and p^{ij} (j = 1,2, ..., r) is the fraction of objects that j^{th} offspring inherits.

As described before, during evolution from G_k to G_{k+1}, some clusters are conserved, some disassociated, and others undergo leakage. Consequently M_k^i is 0 for conservation, small for leakage, and large for cluster disassociation. Offspring clusters are tested using a separability test and NIFTI increases if they are separable and decreases otherwise. We propose a simple but effective test for separability of clusters. The cluster separability test is described below.

Test for separability of offspring

Though a parent cluster can result in many offspring, in practice it is observed that most members of a parent cluster migrate to a few proximal offspring. This is not a surprise since only one additional cluster is added at each step. Therefore, the incremental reorganization that takes place during evolution is minimal. We term those offspring which inherit large fractions of objects from a parent as the dominant offspring. The information transferred for a parent cluster can be approximated by considering only the dominant offspring. The information change arising from the other offspring (non-dominated) is very small and can be neglected. Hence, r in Equation 8 is set to 2 for all parent clusters.

Let X and Y be the two dominant offspring of a parent cluster given by:

$$X = \arg\max_{j} p^{ij} \tag{9}$$

$$Y = \arg\max_{j \neq X} p^{ij} \tag{10}$$

where p^{ij} is the fraction of objects migrated from i^{th} parent cluster, C_k^i to the j^{th} offspring cluster, C_{k+1}^j.

We use inter- and intra-cluster distances to identify whether X and Y are separable or not. X and Y are said to be separable if the distance between their centroid, δ_{XY}, is larger than the sum of their radii (Δ_X and Δ_Y). A variety of methods can be used to measure the cluster radius [8]. Here, the mean distance between the cluster centroid to all members of that cluster is used for this purpose.

Radius of cluster X:

$$\Delta_X = \frac{1}{|X|} \sum_{x \in X} d(x, \overline{v}_X) \tag{11}$$

where $|X|$ is the number of objects in X, x represents the object in cluster X, d is the distance metric used for clustering, and $\overline{v}_X$ the centroid of the cluster. Similarly the radius of cluster Y is given by:

$$\Delta_Y = \frac{1}{|Y|} \sum_{x \in Y} d(x, \overline{v}_Y) \tag{12}$$

The centroid distance between X and Y is the distance between their centroids given as:

$$\delta_{XY} = d(\overline{v}_X, \overline{v}_Y) \tag{13}$$

Hence, the separability of offspring of C_k^i notated as D_k^i is given by:

$$D_k^i = \begin{cases} +1 & if \quad \delta_{XY} \geq (\Delta_X + \Delta_Y) \\ -1 & if \quad \delta_{XY} < (\Delta_X + \Delta_Y) \end{cases} \tag{14}$$

Geometrically, the proposed procedure for finding the separability of clusters is equal to modeling each offspring clusters as a hyper-spheres with radii (Δ_X and Δ_Y)and check whether the hyper-spheres overlap. Statistically, this procedure is a hypothesis test with the following null and alternative hypotheses:

H_0 = Offspring clusters are part of single cluster

H_1 = Offspring clusters are not part of single cluster (*i.e.* different clusters)

The equations for hypothesis testing are derived considering the situation where a single cluster is artificially broken into two clusters. Let us consider a single cluster C containing n objects. Assume that the data is drawn from Gaussian distribution with mean μ and covariance matrix Σ. Without loss of generality, we can assume that the mean is at origin and covariance matrix has only diagonal elements and off-diagonal elements are all zero (if the original covariance matrix contains non-zero off diagonal elements it can be converted to diagonal matrix by principal axis rotation). Suppose, now that we

partition *C* into two clusters (offspring), we can reject the null hypothesis using the distribution functions of both centroid distance and radii of offspring clusters. There are two cases:

1. Same variance in all dimensions *i.e.*

$$\Sigma = \begin{vmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \dots \\ 0 & 0 & \dots & \sigma_m^2 \end{vmatrix}$$

and $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2$.

2. The σ_{2i}s of Σ are different.

We derive the equations for proposed test of separability of offspring for case 1 and show how it can be extended to case 2.

Case 1

Geometrically, this means that the cluster of objects form a spheroid in *m*-dimensional space. Application of any clustering algorithm to partition this cluster into two offspring results in optimal (based on the objective function used for clustering) partition. If we know the analytical solution for that optimal partitioning, we could determine the distribution functions for centroid distance and radii of clusters. Lacking the analytical solution for the optimal partitioning, we cannot derive the actual sampling distributions. However, approximate estimates can be obtained by considering the suboptimal partition provided by a hyperplane through the centroid of parent cluster [26]. This hyperplane approximation is schematically described in Figure 15 for two dimensional data. The data contains 1000 samples drawn from 2 dimensional Gaussian distribution with mean at origin and covariance matrix [1 0;0 1]. k-means clustering algorithm is used to generate the two partitions.

Because of the hyperplane, the centroids for individual offspring clusters will be same as centroid of original parent cluster except in one dimension (the dimension ⊥ to hyperplane). Let the dimension ⊥ to hyperplane is denoted as *f*. Then *f* follows half-normal distribution with mean $\sqrt{2/\pi\sigma}$ (Figure 15). So, the centroid distance between the two offspring is $2\sqrt{2/\pi\sigma}$. Considering the sample size, *n*, the squared centroid distance between the two offspring cluster follows Gaussian distribution with mean as $((n\text{-}1)/n)8/\pi\sigma^2$ and variance $2((n\text{-}1)/n^2)(64/\pi^2)\sigma^4$. The squared radius of cluster Δ^2 also follows a Gaussian distribution with mean $((m\text{-}2)/\pi)\sigma^2$ and variance $4((m\text{-}8)/\pi^2)\sigma^4$ [26].

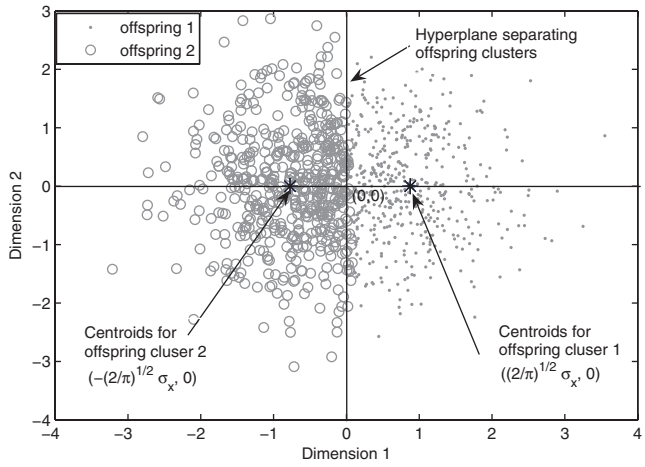

Figure 15
Artificial partitioning of natural cluster. A single natural cluster drawn from Gaussian distribution with mean at origin and identity covariance matrix. k-means clustering partitions this single cluster into two clusters separated by hyperplane.

Now consider the Equation 14 for testing the separability of offspring clusters.

$$\delta_{XY} \geq (\Delta_X + \Delta_Y) \tag{15}$$

Squaring both sides

$$\delta_{XY}^2 \geq (\Delta_X + \Delta_Y)^2 \tag{16}$$

Since the clusters are separated by a hyperplane passing through the origin, the two offspring clusters approximately contain same number of samples and hence ($\Delta_X \approx \Delta_Y = \Delta$).

Hence the test of separability of offspring clusters reduces to

$$\delta^2 \geq 4 \times \Delta^2 \tag{17}$$

where the subscripts *X* and *Y* have been removed for convenience. Hence, the offspring clusters are deemed to be separable if

$$h \geq 0 \tag{18}$$

where $h = \delta^2 - 4 \times \Delta^2$

Using the distributions for δ^2 and Δ^2 derived above the distribution for above equations can be obtained. This distribution refers to the null distribution for the proposed hypothesis test as this derivation is through artificial portioning of a single cluster. Hence, the null

Table 1: False discovery rate of cluster separability test

Sample size(n)	m = 2	m = 3	m = 4	m = 5
25	0.0068 (0.199)	8.53E-7 (0.021)	2.99E-11 (0.001)	6.66E-16 (0.002)
50	1.84E-4 (0.008)	2.44E-12 (0.002)	0 (0)	0 (0)
100	1.81E-7 (0)	0 (0)	0 (0)	0 (0)

hypothesis can be rejected considering the distribution of above equation. Since, both δ^2 and Δ^2 follows Gaussian distribution, h follows a Gaussian distribution with mean as $4\frac{(n-1)}{4}(4/\pi - m)\sigma^2$ and variance $\frac{2}{n}\left[\frac{64}{\pi^2} + 8\left(m - 8/\pi^2\right)\right]\sigma^4$.

The false discovery rate for rejecting the single cluster hypothesis can be calculated using the distribution of h. The false discovery rate is the probability of $h > 0$. The false discovery rate indicates the probability that a offspring of a single parent cluster are incorrectly deemed as two separable clusters. Table 1 shows the false discovery rate for different sample sizes, n, and number of dimensions, m. The values given in parenthesis are the false discovery rates obtained by computational study with 1000 datasets with mean at origin and $\sigma^2 = 2$. The false discovery rates are very low even for small samples sizes. It clearly shows that the proposed cluster separability test able to correctly identify the artificial break of natural clusters. When a natural clusters is artificially broken, NIFTI decreases. So, selecting a partition with highest NIFTI gives number of natural clusters in the data.

Case 2

Geometrically this means that the cluster form a ellipsoid in m-dimensional space. An Analytical solution is difficult for this case. However, it is possible to show that $\delta^2 - 4\Delta^2 \geq 0$ for many situations. Assuming that the hyperplane separating the two offspring cluster is $\perp$ to the dimension of largest variance, the δ^2 is given by: $8/\pi\sigma^2_{max}$. Similarly, Δ^2 is given $\sum_{i=1,i\neq j}^{m}\sigma_i^2 + (1-2/\pi)\sigma^2_{max}$ where j corresponds to the dimension of largest variance. Hence, the separability test δ^2-$4\Delta^2$ is given by: $4\sigma^2_{max}[4/\pi - 1] - \sum_{i=1,i\neq j}^{m}\sigma_i^2$. This means the artificial partition of single cluster is detected by proposed separability criteria whenever the sum of variances in all directions (except the variance of largest direction) has value at least $0.275 \times \sigma^2_{max}$. Since this criteria is satisfied in most of the cases, the proposed test for separability works well even in this case. To check the performance of proposed separability test, we generated 1000 random datasets with 1000 samples each in 3-dimensional space with the largest variance as $\sigma^2_{max} = 3$ and other variances are 0.75. In all the datasets the proposed method correctly identified the partition of a single cluster.

Authors' contributions

Both SJ and RS contributed to the concept and methodology development. SJ implemented the methodology and conducted the data analysis and biological interpretation. RS supervised the study and assisted in implementation. SJ drafted the manuscript. Both authors read and approved the final manuscript.

Additional material

Additional File 1

Scores plot of k-means results for Yeast cell-cycle dataset. *The first two PCs capture 65% variance. All the five clusters are homogenous and distinct.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-10-40-S1.png]

Additional File 2

Supplementary Material. *Comparison of methods for finding number of clusters using artificial datasets.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-10-40-S2.doc]

Acknowledgements

This work was supported by a grant from the National University of Singapore.

References

1. Eisen MB, Spellman PT, Brown PO and Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863–14868.
2. Tavazoie S, Huges JD, Campbell MJ, Cho RJ and Church GM: **Systematic determination of genetic network architecture.** *Nature Genetics* 1999, **22:**281–285.
3. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES and Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Nat Acad Sci U S A* 1999, **96(6):**2907–2912.
4. Yeung KY, Fraley C, Murua A, Raftery AE and Ruzzo WL: **Model-based clustering and data transformations for gene expression data.** *Bioinformatics* 2001, **17:**977–987.
5. Dembele D and Kastner P: **Fuzzy C-means method for clustering microarray data.** *Bioinformatics* 2003, **19:**973–980.

6. Sharan R, Moron-Katz A and Shamir R: **CLICK and EXPANDER: a system for clustering and visualizing gene expression data.** *Bioinformatics* 2003, **19:**1787–1799.
7. Jiang D, Tang C and Zhang A: **Cluster analysis for gene expression data: A Survey.** *IEEE Transactions on Knowledge and Data Engineering* 2004, **16:**1370–1386.
8. Bezdek JC and Pal NR: **Some new indexes of cluster validity.** *IEEE Trans Syst Man Cybern B Cybern* 1998, **28(3):**301–315.
9. Milligan GW and Cooper MC: **An examination of procedures for determining the number of clusters in a data set.** *Psychometrika* 1985, **50:**159–179.
10. Halkidi M, Batistakis Y and Vazirgiannis M: **On clustering validation techniques.** *Journal of Intelligent Information Systems* 2001, **17:**107–145.
11. Handl J, Knowles J and Kell DB: **Computational cluster validation in post-genomic data analysis.** *Bioinformatics* 2005, **21:**3201–3212.
12. Tibshirani R, Walther G and Hastie T: **Estimating the number of clusters in a dataset via gap statistic.** *Journal of Royal Statistical Society B* 2001, **63:**411–423.
13. Dudoit S and Fridlyand J: **A prediction-based resampling method to estimate the number of clusters in a dataset.** *Genome Biology* 2002, **3:**RESEARCH0036.
14. Ben-Hur A, Elisieeff A and Guyon I: **A stability based method for discovering structure in clustered data.** *Pac Symp Biocomput* 2002, 6–17.
15. Bolshakova N and Azuaje F: **Cluster validation techniques for genome expression data.** *Signal Processing* 2003, **83:**825–833.
16. Rousseeuw PJ: **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.** *Journal of Computational and Applied Mathematics* 1987, **20:**53–65.
17. Davies DL and Bouldin DW: **A cluster separation measure.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1979, **1:**224–227.
18. Jiang D, Pei J and Zhang A: **DHC: A Density-based hierarchical clustering method for time-Series gene expression data.** *Proceedings of Third IEEE Symposium on Bioinformatics and Bioengineering* 2003, 393–400.
19. Pal NR and Bezdek JC: **On cluster validity for fuzzy c-means model.** *IEEE Transactions on Fuzzy Systems* 1995, **3:**370–379.
20. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ and Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mole Cell* 1998, **2(1):**65–73.
21. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson JJ, Boguski MS, Lashkari D, Shalon D, Botstein D and Brown PO: **The transcriptional program in the response of human fibroblasts to serum.** *Science* 1999, **283:**83–87.
22. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Bird JC, Botstein D, Brown PO and Staudt M: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503–511.
23. Iacobuzio-Donahue C, Maitra A, Olsen M, Lowe AW, Van Heek NT, Rosty C, Walter K, Sato N, Parker A, Ashfaq R, Jaffee E, Ryu B, Jones J, Eshleman JR, Yeo CJ, Cam-eron JL, Kern SE, Hruban RH, Brown PO and Goggins M: **Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays.** *American Journal of Pathology* 2003, **162:**1151–1162.
24. Gordon AD: *Classification* Boca Raton: Chapman and Hall/CRC; 1999.
25. Li H, Zhang K and Jiang T: **Minimum entropy clustering and applications to gene expression analyses.** *Proce IEEE Comput Syst Bioinforma Conf* 2004, 142–151.
26. Duda RO and Hart MP: *Pattern classification and scene analysis* NY: Wiley; 1973.