BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 16 2006, pages 1971–1978
doi:10.1093/bioinformatics/btl185

Gene expression

Distance-based clustering of CGH data

Jun Liu[1,*], Jaaved Mohammed[1], James Carter[1], Sanjay Ranka[1], Tamer Kahveci[1] and Michael Baudis[2]

[1]Computer and Information Science and Engineering, University of Florida Gainesville, FL, 32611 USA and [2]Institut fuer Humangenetik, Rheinisch-Westfaelische Technische Hochschule, Aachen, Germany

Received on February 6, 2006; revised on April 20, 2006; accepted on May 10, 2006
Advance Access publication May 16, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: We consider the problem of clustering a population of Comparative Genomic Hybridization (CGH) data samples. The goal is to develop a systematic way of placing patients with similar CGH imbalance profiles into the same cluster. Our expectation is that patients with the same cancer types will generally belong to the same cluster as their underlying CGH profiles will be similar.

Results: We focus on distance-based clustering strategies. We do this in two steps. (1) Distances of all pairs of CGH samples are computed. (2) CGH samples are clustered based on this distance. We develop three pairwise distance/similarity measures, namely raw, cosine and sim. Raw measure disregards correlation between contiguous genomic intervals. It compares the aberrations in each genomic interval separately. The remaining measures assume that consecutive genomic intervals may be correlated. Cosine maps pairs of CGH samples into vectors in a high-dimensional space and measures the angle between them. Sim measures the number of independent common aberrations. We test our distance/similarity measures on three well known clustering algorithms, bottom-up, top-down and *k*-means with and without centroid shrinking. Our results show that sim consistently performs better than the remaining measures. This indicates that the correlation of neighboring genomic intervals should be considered in the structural analysis of CGH datasets. The combination of sim with top-down clustering emerged as the best approach.

Availability: All software developed in this article and all the datasets are available from the authors upon request.

Contact: juliu@cise.ufl.edu

1 INTRODUCTION

Numerical and structural chromosomal imbalances are one of the most prominent and pathogenetically relevant features of neoplastic cells (Mitelman *et al*., 1972). Over the past decades, thousands of (molecular-) cytogenetic studies of human neoplasias have searched for insights into genetic mechanisms of tumor development and the detection of targets for pharmacologic intervention. It is assumed that repetitive chromosomal aberration patterns reflect the supposed cooperation of a multitude of tumor relevant genes (Vogelstein and Kinzler, 1993) in most malignant diseases.

This material is based upon work supported by the National Science Foundation under Grant ITR 0325459. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

*To whom correspondence should be addressed.

One method for measuring genomic aberrations is Comparative Genomic Hybridization (CGH) (Kallioniemi *et al*., 1992). CGH is a molecular-cytogenetic analysis method for detecting regions with genomic imbalances (gains or losses of DNA segments). Raw data from CGH experiments is expressed as the ratio of normalized fluorescence of tumor and reference DNA. Normalized CGH ratio data surpassing predefined thresholds is considered indicative for genomic gains or losses, respectively. In contrast to array CGH, chromosomal CGH data (on which this paper is based) does not consist of a (large) number of single measurements (e.g. spot ratios), but on the ratio data measured along human metaphase chromosomes, averaged over a number of measurements (du Manoir *et al*., 1992). Because no single measurements are used for the results composition, the chromosomal CGH results are annotated in a reverse *in situ* karyotype format (Mitelman, 1995) describing imbalanced genomic regions with reference to their chromosomal location. CGH data of an individual tumor can be considered as an ordered list of status values, where each value corresponds to a genomic interval (e.g., a single chromosomal band).The status can be expressed as a real number (positive, negative or zero for gain, loss or no aberration, respectively).

In this paper, we focus on the problem of clustering the CGH data of a population of cancer patient samples. A large number of clustering methods have been developed for various types of datasets (Jain *et al*., 1999). However, these methods are not directly applicable to CGH data. Cytogenetic aberration data is structurally different from ordinary high-dimensional data since consecutive dimensions (i.e., genomic intervals) may be correlated. Regional genomic imbalances arise from the advantage of tumor cells in gaining additional copies of oncogenes (Schwab *et al*., 1984), or losing one or both copies of genes that inhibit oncogenesis [tumor suppressor genes Knudson, 1971]. The minimal change involving one relevant gene is a 'point like event' on the cytogenetic scale, beyond the spatial resolution of Metaphase-based techniques. Therefore, a point-likes genomic aberration may expand to the neighboring intervals and result in a contiguous run of gain or loss status in CGH data.

We develop novel distance-based methods that effectively exploit these correlations between consecutive genomic intervals. Our work is built in two steps. In the first step, we measure the distance/similarity between all pairs of samples. For this purpose, we develop three metrics to compute the similarity/distance between two CGH samples. The first one, raw distance, compares the value or status of each genomic interval separately. The second measure, segment-based similarity, merges contiguous aberrations of the same type

into segments. It then counts the number of common segments between the given two samples. The third measure, segment-based cosine similarity maps segments to vectors in a high dimensional space. It computes the distance between two vectors as the cosine of the angle between them. In the second step, we build clusters of samples based on pairwise similarities. We use three main clustering techniques *k*-means (MacQueen, 1967), complete-link bottom-up (King, 1967) and top-down (Steinbach *et al*., 2000). Two techniques to further improve the cluster qualities were also implemented. The first one combines each of the bottom-up and top-down clustering method with *k*-means so that the former method can provide reasonable initial cluster seeds for the *k*-means method. The second one shrinks centroid (Tibshirani *et al*., 2002) to reduce the number of features contributing to the nearest centroid computation in *k*-means. Experimental results show that segment-based similarity distance measures are better indicators of biological proximity between pairs of samples. This measure when combined with the top-down method produces the best clusters.

The rest of the paper is organized as follows. Section 2 introduces the proposed distance measures and clustering techniques. Section 3 presents the experimental results. Section 4 discusses the related work and Section 5 concludes with a brief discussion.

Genomic Intervals	1	2	3	4	5	6	7	8	9	10	11	12
X	0	1	1	1	0	0	−1	−1	0	1	−1	−1
Y	0	0	1	1	1	0	0	0	0	1	1	1
Diff(x,y)	1	1	0	0	1	1	1	1	1	0	1	1

Fig. 1. The distance on raw CGH data. *X* and *Y* are two CGH samples. The value of each genomic interval shows the status (i.e. gain loss or no change) of that interval. The distance between *X* and *Y* is $\sum_{j=1}^{m}$ diff $(x_j, y_j) = 9$.

Genomic Intervals	1	2	3	4	5	6	7	8	9	10	11	12
X	0	1	1	1	0	0	−1	−1	0	1	−1	−1
Y	0	0	1	1	1	0	0	0	0	1	1	1

Fig. 2. *X* and *Y* are two CGH samples with the values of genomic intervals shown in the order of positions. The segments are underlined. The overlapping segments are shown with arrows. Since there are two overlapping segments; one from position 3 to 4 and the other at position 10, the similarity between *X* and *Y* is 2.

2 METHOD

Genomic aberration data from CGH experiments is usually communicated in a reverse *in situ* karyotype annotation format (Mitelman, 1995). We use this strategy and represent gain, loss and no change with +1, −1 and 0, respectively, throughout the paper.

We propose to use three different distance-based clustering methods for CGH data and survey their performance. The key problem, however, is to compute the proximity of two CGH samples. In Section 2.1, we discuss the three measures we developed for such pairwise comparison. We briefly explain the three clustering algorithms we used to cluster a population of samples in Section 2.2. Two techniques that further optimize the cluster qualities are discussed in Section 2.3.

2.1 Comparison of two samples

Let $X = x_1, x_2, \ldots, x_m$ and $Y = y_1, y_2, \ldots, y_m$ be two CGH samples. Here, x_i and y_i denote the value or status of the *i*-th genomic interval of *X* and *Y*, respectively. The proximity between *X* and *Y* can be computed in terms of distance or similarity. In this section we develop three such measures of distance/similarity.

2.1.1 Raw distance Our first measure assumes that the genomic intervals are independent of each other. This assumption is often made in existing literature to simplify the problem of computing distances (Picard *et al*., 2005b). If both samples have gain (or loss) at the same genomic interval then we consider them similar at that position. Otherwise, that genomic interval contributes to the distance between them. Also, we assume that all genomic intervals have the same importance. Thus, each genomic interval contributes the same amount to the total distance. Formally, the distance is computed as $\sum_{j=1}^{m}$ diff(x_j, y_j). Here diff$(x_j,y_j) = 1$ if $x_j \neq y_j$ or $x_j = 0$. Otherwise diff$(x_j, y_j) = 0$. The similarity is obtained by subtracting the distance from *m*, the number of genomic intervals of the CGH samples. An example is shown in Figure 1

This distance function is similar to Hamming distance in principle because it compares the genomic intervals of both samples one by one. We call this distance Raw since it is computed on raw CGH data. Raw distance between two samples is small only if the samples have gains or losses in the same positions. Raw distance ranges between [0, *m*].

2.1.2 Segment-based similarity This method takes the fact that consecutive genomic intervals are usually correlated. A contiguous block of gains (or losses) can be caused by a point-like aberration at a single genomic interval. We use the term segment to represent a contiguous block of aberrations of the same type. For example, in Figure 2, sample *X* contains four segments. The first and third segments are gain type while the second and fourth segment are loss type. We call two segments from two samples overlapping if they have at least one common genomic interval of the same type. For example, the first segment of *X* is overlapping with the first segment of *Y* in Figure 2. Also the third segment of *X* is overlapping with the second segment of *Y*. Next, we develop a segment-based similarity measure called Sim.

Given two CGH samples *X* and *Y*, Sim constructs maximal segments by combining as many contiguous aberrations of the same type as possible. Formally, the genomic intervals $x_i, x_{i+1}, \ldots, x_j$, for $1 \leq i \leq j \leq m$, define a segment if genomic intervals x_i through x_j are in the same chromosome, the values from x_i to x_j are all gains or all losses, and x_{i-1} and x_{j+1} are different than x_i. Thus, each sample translates into a sequence of segments. After this transformation, Sim assumes that the segments are independent of each other and gives the same importance to all the segments regardless of the number of genomic intervals in them. Sim computes the similarity between two CGH samples as the number of overlapping segment pairs. This is justified because each overlap may indicate a common point-like aberration in both samples which then led to the corresponding overlapping segments. An example is shown in the Figure 2. There are two important observations that follows from the definition of Sim. First, unlike the Raw distance measure, Sim considers an overlap of arbitrary number of genomic intervals as a single match. Second, although two samples have different values for the same genomic interval, Sim does not consider this as a mismatch if it is an extension of an overlap. For example, in Figure 2, the fifth genomic intervals of sample *X* and *Y* have different values, but we still consider this position a match because it could be an extension of an overlap.

2.1.3 Segment-based cosine similarity Segment-based similarity grows linearly with the number of common segments. However, the aberration patterns of some cancer types can be less complex than the others. The samples that belong to these cancer types share fewer common segments leading to small values of Sim even though the samples are almost identical.

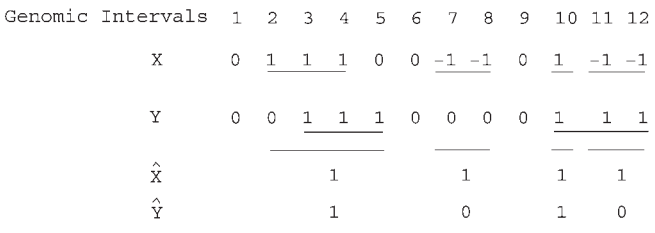

Fig. 3. This figure shows the cosineNoGaps similarity between two CGH samples. X and Y are two CGH samples with the values of genomic intervals shown in the order of positions. The segments are underlined. First, X and Y are mapped to two vectors $\hat{X}$ and $\hat{Y}$ respectively. Second, the similarity between X and Y is computed as $C(\hat{X}, \hat{Y}) = 0.7071$

Cosine similarity of two vectors normalizes the similarity by measuring the cosine of the angle between them. This measure is the most commonly used method to compute the similarity between two directional data in vector-space model (Salton, 1989). In this section, we extend the cosine similarity to measure the proximity of two CGH samples.

Let X and Y be two CGH samples. We first map X and Y to two vectors $\hat{X}$ and $\hat{Y} \in \mathcal{R}^g$, where g is the number of dimensions of the vectors. Usually, $g \ll m$, where m is the number of genomic intervals of CGH samples. The mapping process is also based on segments and works as follows. First, we translate each sample into a sequence of segments. Let us define segment sequence G, H that corresponds to the sample X, Y respectively. Without loss of generality, we can assume that for all the genomic intervals in Y, if they belong to any segment in H, the genomic intervals in X at the same positions are also covered by the segments in G. Here, we say that a segment covers a consecutive block of genomic intervals only if for each genomic interval, either it belongs to this segment or it is of no-change status and the aberration of this segment can be extended to this genomic interval. Next, we scan the segment sequence G in the ascending order of the genomic intervals. For each segment $g_i \in G$, if there exist an overlapping segment $h_j \in H$, we add a new dimension to both vectors $\hat{X}$ and $\hat{Y}$. We then assign value 1 to this dimension of $\hat{X}$ and $\hat{Y}$, indicating that the value of this dimension are exactly the same in the two vectors. If no overlapping segment $h_j \in H$ exists, we add a new dimension to both vectors with value 1 assigned to vector $\hat{X}$ and value 0 assigned to vector $\hat{Y}$, which indicates that the values of the new dimension in two vectors are orthogonal. An example of the segmenting and mapping step for this measure is shown in Figure 3. After the two CGH samples X and Y have been mapped to two vectors, the cosine similarity between X and Y is computed as

$$C(\hat{X}, \hat{Y}) = \frac{\sum_{i=1}^{m} \hat{x}_i \cdot \hat{y}_i}{\sqrt{(\sum_{i=1}^{m} \hat{x}_i \cdot \hat{x}_i)(\sum_{i=1}^{m} \hat{y}_i \cdot \hat{y}_i)}}.$$

The majority of genomic intervals in CGH data have zero values (i.e. no aberration). We call a consecutive block of these genomic intervals gaps. We ignore the impact of gaps in the above cosine similarity measure. However, considering the overlapping gaps between two samples might contribute greatly to the similarity between them. We develop another variant of cosine similarity which takes the overlapping gaps into consideration. The new similarity measure changes the mapping step that translates the CGH data into vectors. First, it extends the definition of segments to be a consecutive block of genomic intervals that share the same status, i.e. gain, loss or no change. That means, gaps are also included in the segments in this way. Then it translates the CGH data into a sequence of segments with some of the segments representing gaps. Next, a scan is performed on the segment sequence G. For each gap in G, if there exists an overlapping gap in H, a new dimension will be added to both vectors and a pair of value 1 will be assigned to them. Other mapping steps of gain or loss segments and computation of cosine similarity remains unchanged. Compared to the previous cosine similarity measure, this measure offers a larger similarity between two CGH samples due to the impact of overlapping gaps. Thus, we use the term CosineGaps to represent it, whereas the term CosineNoGaps is used to represent the previous definition. Both of these measures produce a value within a range of [0, 1] indicating the similarity between two samples.

2.2 Clustering of samples

With one of the aforementioned distance/similarity measures between two CGH samples, we can easily apply a distance-based clustering algorithm to group similar CGH samples together. At a high-level, the problem of clustering is defined as follows. Given a set S of n samples $s_1, s_2, \ldots, s_n$, we would like to partition S into k subsets $C_1, C_2, \ldots, C_k$, such that the samples assigned to each subset are more similar to each other than the samples assigned to different subsets. Here, we assume that two samples are similar if they correspond to the same cancer type.

As we mentioned earlier, our focus in this paper is to evaluate the suitability of various distance/similarity measures together with clustering algorithms in the context of the CGH data clustering problem. In this section, we briefly introduce the three distance-based clustering algorithms we used in our experiments.

2.2.1 K-means Clustering K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. Its key step is to compute the distance/similarity between a sample data and the cluster centroid, which is not necessary a real sample. Since CGH samples are represented as an array of status values, it is not trivial to compute an accurate centroid for a set of CGH samples. Here, we develop a variant of the k-means algorithm which is more suitable for our distance/similarity measures. Compared with standard k-means, our algorithm omits the step of computing the cluster centroids, but reassigns a sample according to its average distance to all the samples in a cluster rather than the distance to the centroid of that cluster. These changes let our algorithm work for any distance/similarity measure described in Section 2.1.

We first partition the n samples into k clusters by randomly assigning each sample to one of the k clusters. This random partition forms the initial cluster seeds for our k-means algorithm. Then we scan the n samples one by one. For the i-th sample, compute its average distance to all the samples in cluster j, for $1 < j < k$, and then move it to the cluster with the minimum average distance if that cluster is different from the one it already belongs to. This scanning process is repeated until there is no movement of samples during a scan or until a maximum number of iterations is reached.

2.2.2 Complete link bottom-up clustering Complete link (King, 1967) clustering defines the distance between two clusters as the largest distance between a sample from the first cluster and a sample from the second cluster. The bottom-up clustering works by designating each sample as its own cluster initially. Next, each cluster is compared to each other cluster, and the closest clusters are merged. This process will continue until k clusters remain.

2.2.3 Top-down clustering This algorithm (Steinbach *et al*., 2000) starts by assigning all samples into one cluster. It then bisects this cluster recursively until k clusters are produced, where k is a user defined parameter. The bisection is performed in two phases. In the first phase, two samples are randomly selected as the seeds of two clusters. Then, for each remaining sample, its similarity to these two seeds is computed and it is assigned to the cluster whose seed has a higher similarity to that sample. In the second phase, the clusters are refined. A refinement consists of a number of iterations. During each iteration, samples are visited one by one. Each sample s_i, is then moved to all of the clusters one by one, and a criterion function is computed for each positioning of s_i. The criterion function evaluates the quality of the clusters. We use the term *internal measure* to represent this criterion function. The formal definition of internal measure is addressed in Section 3.1. The sample s_i is kept in the cluster that maximizes the internal measure. This refinement process ends as soon as there is no movement of samples during

an iteration or after a predefined maximum number of iterations have been performed. In our experiments, the number of iterations were typically < 20. After the refinement is finished, the cluster with the largest number of samples is bisected similarly. Once k clusters are created, the top-down algorithm ends.

In each iteration of the refinement, O(n) time is needed to compute the change of the internal measure for each sample. This is because, the similarity between that sample and every other sample in each cluster needs to be accumulated. The time complexity of each iteration is $O(n^2)$ as there are totally n samples. Since the total number of iterations is limited by a small constant, the complexity of refinement is $O(n^2)$. The refinement is performed every time a new cluster is created. In the above described process the number of clusters increases by one in every stage until k clusters are created. Therefore, the overall time complexity of top-down clustering is $O(n^2k)$.

To reduce this time complexity, we modify the top-down clustering algorithm. Essentially, the refinement process is limited to the cluster being decomposed into smaller clusters. There are two differences between the modified and the original top-down clustering. First, only the samples in the decomposed cluster are considered for refinement. Second, a sample is relocated only to the two newly created clusters rather than all the clusters. In the best case, the clusters are decomposed in a balanced fashion. The overall time complexity in this case is $O(n^2 + 2(\frac{n}{2})^2 + \cdots + 2^{log_2 k}(\frac{n}{2^{log_2 k}})^2) \approx O(2n^2)$. In the worst case, a cluster with n samples could be decomposed into two clusters with $n-1$ samples in one cluster and 1 sample in the other. If this case happens to all the bisections, the worst case time complexity could be $O(kn^2)$. Thus, with this enhanced refinement process, the average time complexity of top-down clustering is between $O(n^2)$ and $O(kn^2)$. We generally expect the time complexity to be close to $O(n^2)$, which results in a factor of k improvement in time. We call this faster refinement process in the top-down clustering Local Refinement and the previous refinement process Global Refinement. It is worth noting that local refinement may produce lower quality clusters. Our experimental results described in Section 3 show that this deterioration is small.

2.3 Further optimization on clustering

In this paper, we use two approaches to further optimize the clusters obtained by the bottom-up or top-down algorithms. We also compare the optimized results with the non-optimized results of these algorithms in Section 3.

2.3.1 Combining k-means with bottom-up or top-down methods Similar to the standard k-means, the k-means algorithm used in this paper does not necessarily find the optimal clusters because it is significantly sensitive to the initial cluster seeds. This observation motivates our further optimization by choosing the results of bottom-up or top-down algorithms as the initial seeds for k-means. That is, after the bottom-up or top-down clustering, a k-means method will be invoked and the clusters produced by the bottom-up or top-down clustering will serve as the initial cluster seeds of k-means. The rest of the k-means clustering remains the same. This additional k-means step further refines the clusters by using the more CGH specific distance measures proposed in this paper. We use the term top-down + k-means to represent the optimization approach that combines the top-down algorithm with the k-means algorithm. Similarly, we use term bottom-up + $k$means to represent the combination of the bottom-up algorithm and the k-means algorithm.

2.3.2 Centroid shrinking The idea of centroid shrinking was first introduced by Robert *et al*. in (Tibshirani *et al*., 2002) to improve the nearest-centroid classification. The centroids of a training set are defined as the average expression of each gene. This idea shrinks the centroids of each class towards the overall centroid after normalizing by the intra-class standard deviation for each genomic interval. This normalization has the effect of assigning more weight to the genomic interval whose status is stable within samples of the same class, and thus reduces the number of features contributing to the nearest centroid calculation. We apply this idea to achieve further optimization of clustering. The centroids of initial clusters found by the different clustering methods, i.e. bottom-up, top-down, k-means, bottom-up + $k$means and top-down + $k$means, are shrunk towards the overall centroid. Then, a standard k-means using Euclidean distance is invoked to re-cluster the samples using the shrunken centroids as its initial centroids.

3 RESULTS

Experimental setup: We evaluated the quality and the performance of all the distance/similarity measures and the clustering methods discussed in this paper. For evaluation of quality we used different measures belonging to two categories, external and internal measures. We discuss these measures in detail in Section 3.1.

We implemented all four distance measures (Raw, Sim, CosineGaps, CosineNoGaps) and five clustering algorithms (k-means, top-down, bottom-up, top-down + k-means, bottom-up + k-means). Thus, we had 20 different combinations. We have also implemented the centroid shrinking strategy and applied on each combination. Note that we use local refinement strategy (see Section 2.2.3) for top-down in our experiments unless otherwise stated.

We use a dataset consisting of 5 020 CGH samples (i.e. cytogenetic imbalance profiles of tumor samples) taken from the Progenetix database (Baudis and Cleary, 2001). These samples belonged to 19 different histopathological cancer types with > 100 cases and had been coded according to the ICD-O-3 system (Fritz *et al*., 2000). The subset with the smallest number of samples consists of 110 non-neoplastic cases, while the one with largest number of samples, Adenocarcinoma, NOS (ICD-O 8140/3), contains 1054 cases. Each sample in the dataset consists of 862 ordered genomic intervals extracted from 24 chromosomes. Each interval is associated with one of the three values −1, 1 or 0, indicating loss, gain or no change status of that interval. In principle, our CGH an dataset can be mapped to an integer matrix of size 5 020 × 862. We also use a small dataset with 2 510 samples by randomly selecting 50% of the entire dataset. This small dataset is generated each time an experiment is running over it.

Our experimental simulations were run on a system with dual 2.59 GHz AMD Opteron Processors, 8 Gb of RAM, and an Linux operating system.

3.1 Quality analysis measures

In this paper, we hope to identify disease-related signatures of CGH data by clustering a large number of samples. We assume that samples belonging to the same cancer type are homogeneous and should be clustered together. There are a range of different cluster validation techniques that can be grouped into two categories, external measure and internal measure (Handl *et al*., 2005). We use both measures to evaluate the quality of the clusters. An external measure evaluates how well the clusters separate samples that belong to different cancer types. Thus external measure can compare clusters based on different distance/similarity measure. On the other hand, an internal measure evaluates how good the clustering algorithm operates on a given distance/similarity measure. This measure ignores the cancer types of the input samples. Compared with internal measures, external measures are more reasonable in reflecting the quality of clusters as they take the cancer types into consideration. Note that internal measure is a better indicator of

quality for cancer types that have multiple aberration patterns that differ significantly.

External measure: An external measure takes a value in [0, 1] interval. Higher values of this function represent better clustering quality. An important note is that this measure is independent of the underlying distance/similarity measure. Thus, the results of different distance measures can be compared using external measure.

We use three external measures to evaluate the cluster quality. Let n, m and k denote the total number of samples, the number of different cancer types and the number of clusters respectively. Let $a_1, a_2, \ldots, a_m$ denote the number of samples that belong to each cancer type. Similarly, let $b_1, b_2, \ldots, b_k$ be the number of samples that belong to each cluster. Let $c_{i,j}$, $\forall i, j$, $1 \leq i \leq m$ and $1 \leq j \leq k$, denote the number of samples in j-th cluster that belong to the i-th cancer type. The first external measure used, known as the Normalized Mutual Information (NMI) (Zhong and Ghosh, 2005) function is computed as:

$$\mathrm{NMI} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{k} c_{i,j} log(\frac{n \cdot c_{i,j}}{a_i b_j})}{\sqrt{(\sum_i a_i log \frac{a_i}{n})(\sum_j b_j log \frac{b_j}{n})}}.$$

The second external measure is F_1-measure (Tan *et al*., 2005). It is defined as

$$F_1 = \frac{1}{n} \sum_{i=1}^{m} a_i \max_j \frac{c_{i,j}}{a_i + b_j}.$$

The third external measure is known as Rand Index (Tan *et al*., 2005). In order to compute the Rand Index measure for a given clustering, two values are calculated.

- f_{00} = the number of pairs of samples that have different cancer types and belong to different clusters.
- f_{11} = the number of pairs of samples that have the same cancer type and belong same cluster.

The Rand Index is then computed as

$$\text{Rand Index} = \frac{f_{00} + f_{11}}{[n(n-1)/2]}.$$

Unlike other external measures, NMI was computed based on mutual information $I(X; Y)$ between a random variable X, governing the cluster labels and a random variable Y, governing the cancer types. It has been argued that the mutual information is a superior measure than purity or entropy (Strehl and Ghosh, 2002). Moreover, NMI is quite impartial to the number of clusters (Zhong and Ghosh, 2005).

Internal measure: Unlike the external measure, the value of internal measure depends on the distance/similarity measure. Thus, the internal measure of different clusterings obtained by different similarity measures are not comparable. Instead, we use this measure to compare the clusters obtained by applying different clustering methods with same similarity function. In this paper, we implement two internal measures. One is the internal measure based on compactness (cohesion) (Tan *et al*., 2005), the other is the internal measure based on separation.

Let k denote the total number of clusters. Let $b_1, b_2, \ldots, b_k$ be the number of samples that belong to each cluster. We use s_i and C_r to represent i-th sample and the r-th cluster respectively. Let $S(s_i, s_j)$ be the function that evaluates the similarity between the i-th and j-th sample. The internal measure based on compactness is computed as

$$\mathcal{IC} = \sum_{r=1}^{k} \frac{\sum_{i<j,\, s_i, s_j \in C_r} \mathcal{S}(s_i, s_j)}{b_r}.$$

The internal measure based on separation is computed as:

$$\mathcal{IS} = \frac{\sum_{r=1}^{k} \sum_{q=1,\, q \neq r}^{k} \sum_{s_i \in C_r,\, s_j \in C_q} \mathcal{S}(s_i, s_j)}{\sum_{r=1}^{k} \sum_{q=1,\, q \neq r}^{k} b_r \cdot b_q}$$

Since both internal measures are computed with pairwise similarity, higher values of $\mathcal{IC}$ and lower values of $\mathcal{IS}$ represent better clustering quality respectively.

Table 1. The highest value of external measures for different distance/similarity measure

	Sim	CosineNoGaps	CosineGaps	Raw
NMI	0.368	0.265	0.228	0.239
F_1-measure	0.34	0.258	0.215	0.235
Rand Index	0.903	0.899	0.898	0.896

All numbers here are the medians of 100 results.

3.2 Experimental evaluation

In this section, we applied the combinations of four distance/similarity measures and five clustering methods over the entire dataset and the small dataset. We compared each combination according to the qualities of clusters. The cluster results are evaluated using different external measures. Owing to the space limit, we mainly report the results using NMI and F_1-measure in the paper unless otherwise stated. For the small dataset that are randomly generated each time, we apply our experiments 100 times and report the results between fifth and ninety-fifth percentile as the error bar.

Evaluation of distance measures. The purpose of this experiment is to compare the distance/similarity measures discussed in this paper, namely Raw, Sim, CosineNoGaps, and CosineGaps. In the experiment, we randomly select 50% of the entire dataset as a small dataset with 2 510 samples. For each distance/similarity measure, we created 2, 4, 8, 16, 32 and 64 clusters using five clustering methods: top-down, bottom-up, k-means, top-down + k-means, and bottom-up + k-means. This resulted in $6 \times 5 = 30$ sets of clusters per measure. We report the highest value of external measure of all these 30 sets as the best quality of a measure. We repeat this experiment for 100 times.

The median of 100 highest values for Sim, CosineNoGaps, CosineGaps and Raw are shown in Table 1. The results of both NMI and F_1-measure show that Sim produces the highest quality compared with other distance measures. Sim obtains this quality with top-down clustering method. CosineNoGaps gives slightly better quality than the other two measures, Raw and CosineGaps. We conclude that Sim is the most suitable distance/similarity measure for clustering CGH data.

Evaluation of clustering methods and optimizations. The purpose of these experiment is to compare the quality of clustering algorithms with a fixed distance/similarity measure. We create 8, 16, 32 and 64 clusters using different clustering methods with and

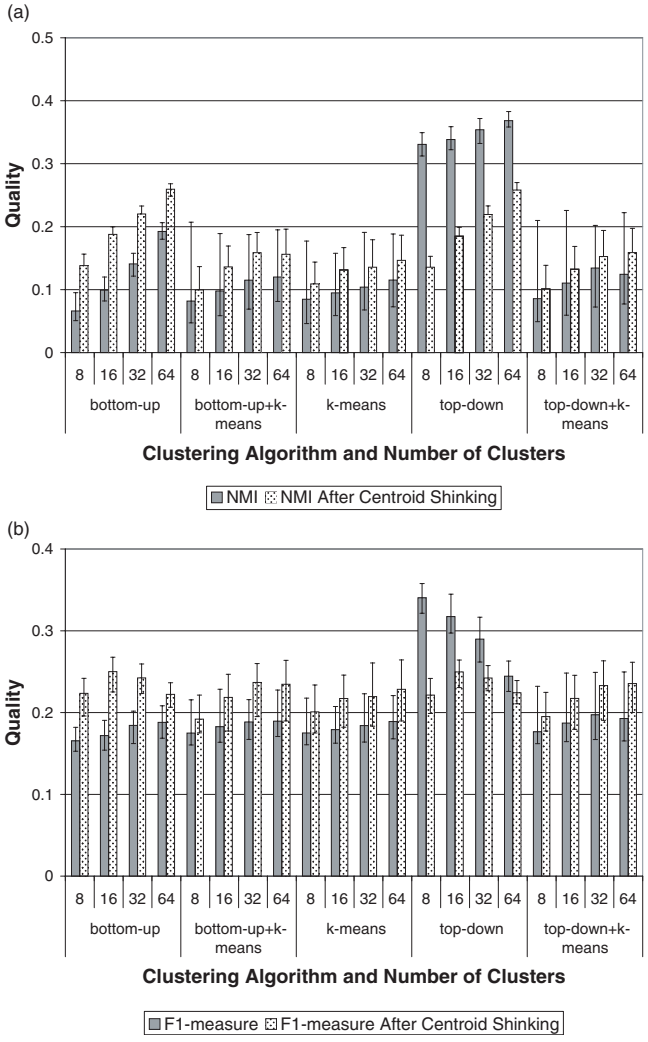

Fig. 4. Cluster qualities after applying different clustering methods with Sim measure using (**a**) NMI and (**b**) F_1-measure respectively as the quality measure. The fifth and the ninety-fifth percentile of the results are reported as the error bar.

without centroid shrinking strategy. We only report the results for Sim due to the space limitations and because Sim gives the best external measure values among all distance/similarity measures.

We randomly select 50 % of the entire dataset (i.e. 2 510 samples) and cluster it. We then compute the external measure for the underlying clusters. We repeat this process 100 times and compute the error bar for the external measure. The error bar indicates the interval where 5–95 % of the results lie. Figure 4a and b show the NMI and F_1-measure respectively. Top-down clustering method without centroid shrinking gives the best quality consistently in both figures. The additional k-means step in top-down + k-means method deteriorates the qualities. Centroid shrinking improves the results when the quality of the clustering method is low. It hurts the quality when the quality is high, especially when top-down method is used. This can be explained as follows. The clustering quality is low when the patients with different cancer types are clustered together. This usually indicates that different samples in the same cluster can contain gain, loss, and no-change status for the same genomic interval. Such genomic intervals can be considered as noise. Centroid shrinking filters them out. However, centroid shrinking has the limitation that its results can be followed by a standard k-means clustering using Euclidean distance. Therefore, the underlying similarity measure (i.e. Sim) cannot be used after shrinking the centroid. Thus, we conclude that top-down method works best in conjunction with the Sim measure. At the same time, centroid shrinking strategy does not help the clustering using this combination. The error bars confirm that the top-down clustering without centroid shrinking works best for Sim measure. The error bars show that the top-down and the bottom-up methods are more stable than the k-means method. This is because k-means is significantly sensitive to the initial seeds that are randomly generated. The NMI value of the top-down method increases as the number of clusters increase from 8 to 64 in Figure 4a. On the other hand, the F_1-measure drops in Figure 4b. This is because F_1-measure favors coarser clustering and is biased towards small number of clusters while NMI is quite impartial to the number of clusters (Zhong nad Ghosh, 2005). We do not see the same effect for other clustering methods because the large variance in the results of other methods, except bottom-up, hides this effect. For bottom-up method with or without centroid shrinking, we can see that the increase in the quality gets flattened when the number of clusters increases.

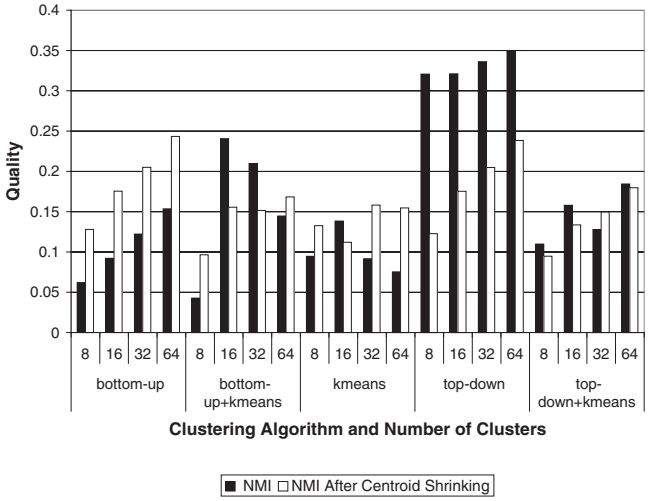

Fig. 5. Cluster qualities of applying different clustering methods with Sim measure over the entire dataset. The cluster qualities are evaluated using NMI.

Next, we ran all the mentioned clustering methods for the entire CGH dataset (i.e. 5 020 samples). Figure 5 shows the NMI for Sim. The results confirm the experiments in Figure 4a: (1) Top-down clustering produces the best clusters. (2) The centroid shrinking strategy does not have a significant impact. (3) Most of the results on the entire dataset remain within the error intervals. The best clustering quality was obtained when 64 clusters were created. The average cluster size, i.e. number of samples in the cluster, is 78.44 and the SD is 51.03.

In our experiments on the same dataset using Rand Index, we obtained slightly better results with top-down method. The two described internal measures (compactness and separation) support this conclusion that top-down clustering is the better choice (results omitted owing space limitation).

Table 2. Comparison of average quality (i.e. internal measure $\mathcal{IC}$) and running time of top-down methods with global and local refinement. (Here L and G indicate local and global refinement respectively.)

		Number of clusters					
		2	4	8	16	32	64
Quality	L	703	797	892	927	947	904
	G	730	839	936	983	1017	971
Time [Sec]	L	0.1	0.3	1.7	3.1	6.5	9.8
	G	3.4	22.9	129.7	329.4	1151.2	2018.2

Performance issues of top-down clustering: In Section 2.2.3, we discussed two types of top-down methods, top-down method with global refinement and top-down method with local refinement. Here, we evaluate the quality and running time of these two strategies. We restrict the similarity measure to Sim as it gives the highest quality. Using each strategy, we created 2, 4, 8, 16, 32, and 64 clusters for each of the 19 cancer types. We compute the average internal measure based on compactness of all the cancer types as the quality of the clusters. We also compute the average time to create clusters as the running time.

Table 2 shows the average quality and running time of two different top-down methods. The first part of the table indicates that local refinement gives slightly worse qualities than the global refinement. However, the quality difference is negligible. The quality of the clusters increases as the number of clusters increases up to 32. The quality starts to plateau or drop after this point. This indicates that, in general, as the number of clusters increases, the clusters are more compact and the intra-similarity of clusters increases. However, when the number of clusters becomes too large compared with size of dataset, some closely similar samples will be forced into different clusters, which, instead, reduce the intra-similarity of clusters. The second part of the table indicates that the average running time for global refinement is much higher than local refinement. This observation is consistent with our analysis of time complexity in Section 2.2.3. Considering that local refinement gives only slightly worse qualities but runs much faster than global refinement, we use the former method throughout this paper.

4 RELATED WORK

The molecular cytogenetic techniques of CGH (Kallioniemi *et al*., 1992) and array- or matrix-CGH (Solinas-Toldo; 1997 Pinkel *et al*., 1998; Pollack *et al*., 1996) have previously been used to describe genomic aberration hot spots in cancer entities (Gray *et al*., 1994; Bentz *et al*., 1996), for the delineation of disease subsets according to their cytogenetic aberration patterns (Mattfeldt *et al*., 2001; Joos *et al*., 2002) and for the construction of genomic aberration trees from chromosomal imbalance data (Desper *et al*., 1999).

In contrast to Metaphase analysis, CGH techniques are not limited to dividing tumor cells which frequently do not represent the predominant clone in the original tumor. Also, CGH is not hampered by incomplete identification of chromosomal segments, which for Metaphase analysis only recently has been addressed by SKY (Spectral Karyotyping) (Veldman *et al*., 1997) and MFISH (Multiplex Fluorescent *In Situ* Hybridization) (Speicher *et al*., 1996) techniques. According to our own survey, chromosomal and array CGH now account for the majority of published analyses in cancer cytogenetics.

With > 12 000 cases (Baudis, 2006), the largest resource for published CGH data can be found in the Progenetix database, developed by one of the authors (Baudis and Cleary, 2001) (http://www.progenetix.net). Recently, the Progenetix database and the software tools developed for the project have shown its usefulness for the delineation of genomic aberration patterns with clear prognostic relevance in neuroblastomas (Vandesompele *et al*., 2005) and for producing tumor type specific imbalance maps (Mao *et al*., 2005, 2006).

Different strategies for structural analysis of CGH data have been applied previously. Most of these analysis were aimed at the description of pseudo-temporal relationships of cytogenetic events (Desper *et al*., 1999; Hoglund *et al*., 2005) or at the correlation of disease subsets with clinical parameters (Mattfeldt *et al*., 2001; Vandesompele *et al*., 2005). Other CGH related data analysis have been aimed at the the spatial coherence of genomic segments with different copy number levels. Picard *et al*. used a segmentation methods with a Gaussian based model to detect homogeneous regions that share the same relative copy number on average (Picard *et al*., 2005b). Further, he proposed a segmentation-clustering approach combining with a Gaussian mixture model to assess the biological status to the detected segments (Picard *et al*., 2005a). Fridlyand *et al*. used an unsupervised hidden Markov models approach to partition the genomic intervals into regions with the same underlying copy number (Fridlyand *et al*., 2004). Pei *et al*. built a hierarchical clustering tree based on similarity between clusters (Wang *et al*., 2005), and then select the interesting clusters at a certain level. Willenbrock et al. made a comparison study on three popular segmentation methods and demonstrated that smoothed (segmented) CGH data are adapted to downstream analyses such as classification (Willenbrock and Fridlyand, 2005). Rouveirol et al. proposed a method to identify regions with recurrent genomic alterations from more than a few tens of profiles (Rouveirol *et al*., 2006). However, so far there has been very limited study on interval-based structural analysis of large (> 1000), heterogeneous sets of smoothed CGH data.

5 CONCLUSION

We considered the problem of clustering CGH data of a population of cancer patient samples. We developed a systematic way of placing patients with same cancer types in the same cluster based on their CGH patterns. We focused on distance-based clustering strategies. We developed three pairwise distance/similarity measures, namely raw, cosine and sim. Raw measure disregards correlation between contiguous genomic intervals. It compares the aberrations in each genomic interval separately. The remaining measures assume that consecutive genomic intervals may be correlated, Cosine maps pairs of CGH samples into vectors in a high-dimensional space and measures the angle between them. Sim measure counts the number of independent common aberrations. We employed our distance/similarity measures on three well-known clustering algorithms, bottom-up, top-down and *k*-means with and without centroid shrinking.

In our experiments using classified disease entities from the Progenetix database, the highest clustering quality was achieved using Sim as the similarity measure and top-down as the clustering

strategy. This observation fits with the theory that contiguous runs of genomic aberrations arise around a point-like target (e.g. oncogene), and that consecutive genomic intervals can not be considered as independent of each other.

Conflict of Interest: none declared.

REFERENCES

Baudis,M. and Cleary,M.L. (2001) Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics*, **17**, 1228–1229.

Baudis,M. (2006) An online database and bioinformatics toolbox to support data mining in cancer cytogenetics. *Biotechniques*, **40**, 269–270, 272.

Bentz,M. *et al*. (1996) High incidence of chromosomal imbalances and gene amplifications in the classical follicular variant of follicle center lymphoma. *Blood*, **88**, 1437–1444.

Desper,R. *et al*. (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comput. Biol*., **6**, 37–52.

du Manoir,S. *et al*. (1995) Quantitative analysis of comparative genomic hybridization. *Cytometry*, **19**, 27–41.

Fridlyand,J. *et al*. (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivariate Anal*., **90**, 132–153.

Fritz,A. *et al*. (2000) *International Classification of Diseases for Oncology (ICD-O), Third edn,* World Health Organization, Geneva.

Gray,J.W. *et al*. (1994) Molecular cytogenetics of human breast cancer. *Cold Spring Harb. Symp. Quant. Biol*., **59**, 645–652.

Handl,J. *et al*. (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201–3212.

Hoglund,M. *et al*. (2005) Statistical behavior of complex cancer karyotypes. *Genes Chromosomes Cancer*, **42**, 327–341.

Jain,A.K. *et al*. (1999) Data clustering: a review. *ACM Comput. Surv*., **31**, 264–323.

Joos,S. *et al*. (2002) Classical hodgkin lymphoma is characterized by recurrent copy number gains of the short arm of chromosome 2. *Blood*, **99**, 1381–1387.

Kallioniemi,A. *et al*. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

King,B. (1967) Step-wise clustering procedures. *J. Am. Stat. Assoc*., **69**, 86–101.

Knudson,A.J. (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA*, **4**, 820–823.

MacQueen,J.B. (1967) Some Methods for Classification and Analysis of Multivariate Observations. In Le Cam,L.M. and Neyman,J. (eds), *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 281–297.

Mao,M. *et al*. (2006) Allele-specific loss of heterozygosity in multiple colorectal adenomas: towards the integrated molecular cytogenetic map II. *Cancer Genet. Cytogenet*, **167**, 1–14.

Mao,X. *et al*. (2005) Genetic losses in breast cancer: toward an integrated molecular cytogenetic map. *Cancer Genet. Cytogenet*., **160**, 141–151.

Mattfeldt,T. *et al*. (2001) Cluster analysis of comparative genomic hybridization (CGH) data using self-organizing maps: application to prostate carcinomas. *Anal. Cell. Pathol*., **23**, 29–37.

Mitelman,F. *et al*. (1972) Tumor etiology and chromosome pattern. *Science*, **176**, 1340–1341.

Mitelman,F. (1995) *International System for Cytogenetic Nomenclature*. Karger, Basel.

Picard,F. A segmentation-clustering problem for the analysis of array CGH data, International Symposium on Applied Stochastic Models and Data Analysis, (Mai 2005) Brest, France.

Picard,F. *et al*. (2005) A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**, 27.

Pinkel,D. *et al*. (1998) High resolution analysis of DNA copy number variation using comparative genomic Hybridization to Microarrays. *Nat. Genet*., **20**, 207–211.

Pollack,J. *et al*. (1999) Genome-wide analysis of DNA copy-number changes using CDNA microarrays. *Nat. Genet*., **23**, 41–46.

Rouveirol,C. *et al*. (2006) Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 849–856.

Salton,G. (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Schwab,M. *et al*. (1984) Enhanced expression of the human gene N-myc consequent to amplification of DNA may contribute to malignant progression of neuroblastoma. *Proc. Natl Acad. Sci. USA*, **15**, 4940–4944.

Solinas-Toldo,S. *et al*. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399–407.

Speicher,M. *et al*. (1996) Karyotyping human chromosomes by combinatorial multifluor fish. *Nat. Genet*., **12**, 368–375.

Steinbach,M. *et al*. (2000) A comparison of document clustering techniques. *KDD Workshop on Text Mining,* Boston, MA, USA.

Strehl,A. and Ghosh,J. (2002) Cluster ensembles—a knowledge reuse framework for combining partitionings. In *Proceedings of AAAI 2002,* AAAI, Edmonton, Canada, 93–98.

Tan,P.-N. *et al*. (2005) *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc, Boston, MA, USA.

Tibshirani,R. *et al*. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Vandesompele,J. *et al*. (2005) Unequivocal delineation of clinicogenetic subgroups and development of a new model for improved outcome prediction in neuroblastoma. *J. Clin. Oncol*., **23**, 2280–2299.

Veldman,T. *et al*. (1997) Hidden chromosome abnormalities in haematological malignancies detected by multicolour spectral karyotyping. *Nat. Genet*., **15**, 406–410.

Vogelstein,B. and Kinzler,K. (1993) The multistep nature of cancer. *Trends Genet*., **9**, 138–141.

Wang,P. *et al*. (2005) A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 45–58.

Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.

Zhong,S. and Ghosh,J. (2005) Generative model-based document clustering: a comparative study. *Knowl. Inf. Syst*., **8**, 374–384.