BIOINFORMATICS

Vol. 18 no. 4 2002
Pages 536–545

Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees

Ying Xu, Victor Olman and Dong Xu*

Protein Informatics Group, Life Sciences Division, Oak Ridge National Laboratory, MS 6480, Oak Ridge, TN 27831-6480, USA

Received on August 28, 2001; revised on November 2, 2001; accepted on November 7, 2001

ABSTRACT

Motivation: Gene expression data clustering provides a powerful tool for studying functional relationships of genes in a biological process. Identifying correlated expression patterns of genes represents the basic challenge in this clustering problem.

Results: This paper describes a new framework for representing a set of multi-dimensional gene expression data as a Minimum Spanning Tree (MST), a concept from the graph theory. A key property of this representation is that each cluster of the expression data corresponds to one subtree of the MST, which rigorously converts a multi-dimensional clustering problem to a tree partitioning problem. We have demonstrated that though the inter-data relationship is greatly simplified in the MST representation, no essential information is lost for the purpose of clustering. Two key advantages in representing a set of multi-dimensional data as an MST are: (1) the simple structure of a tree facilitates efficient implementations of rigorous clustering algorithms, which otherwise are highly computationally challenging; and (2) as an MST-based clustering does not depend on detailed geometric shape of a cluster, it can overcome many of the problems faced by classical clustering algorithms. Based on the MST representation, we have developed a number of rigorous and efficient clustering algorithms, including two with guaranteed global optimality. We have implemented these algorithms as a computer software EXpression data Clustering Analysis and VisualizATiOn Resource (EXCAVATOR). To demonstrate its effectiveness, we have tested it on three data sets, i.e. expression data from yeast *Saccharomyces cerevisiae*, expression data in response of human fibroblasts to serum, and *Arabidopsis* expression data in response to chitin elicitation. The test results are highly encouraging.

Availability: EXCAVATOR is available on request from the authors.

*To whom correspondence should be addressed.

Contact: xyn@ornl.gov

1 INTRODUCTION

As probably the most explosively expanding tool for genome analysis, microchips of gene expression have made it possible to simultaneously monitor the expression levels of tens of thousands of genes under different experimental conditions. This provides a powerful tool for studying how genes collectively react to changes in their environments, providing hints about the structures of the involved gene networks. One of the basic problems in interpreting the observed expression data is to cluster genes with correlated expression patterns over some time series and/or under different conditions.

A number of computer algorithms/software have been developed for clustering gene expression patterns. The most prevalent approaches include: (i) hierarchical clustering (Eisen *et al.*, 1998; Wen *et al.*, 1998); (ii) K-means clustering (Herwig *et al.*, 1999); and (iii) clustering through Self-Organizing Maps (SOMs; Tamayo *et al.*, 1999). While all these approaches have clearly demonstrated their usefulness in applications (Sherlock, 2000), some basic problems remain: (1) none of these algorithms can, in general, rigorously guarantee to produce a globally optimal clustering for any non-trivial objective function; (2) both K-means and SOM heavily depend on the 'regularity' of the geometric shape of cluster boundaries; they generally do not work well when the clusters cannot be contained in some non-overlapping convex sets—just to name a few. Figure 1a shows an example that none of above algorithms generally do well.

Unlike a continuous optimization problem, finding a globally optimal solution for a combinatorial optimization problem is often possible. Consider the simple one-dimensional (1D) optimization problem in Figure 2: we want to cluster the nine data points into three groups so that the total distance between each pair of adjacent points of the same group is minimized. It should not be hard to convince ourselves that by cutting the two longest lines

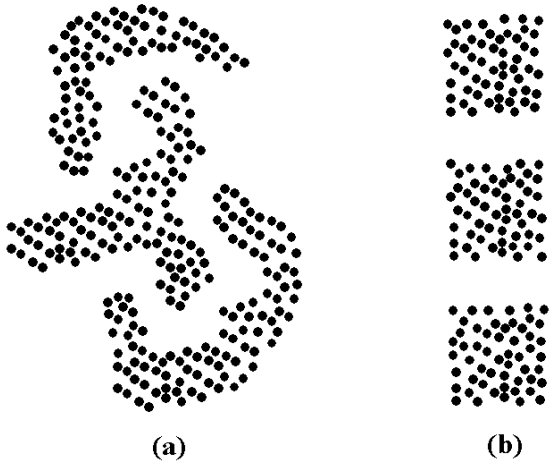


Fig. 1. Two clustering problems. (a) A data set consists of three clusters, each of which has a non-regular shape. (b) A data set of three clusters contained in three non-overlapping rectangles.

Fig. 2. An example of clustering 1D data points. A three-way cut between points 3 and 4 and between points 6 and 7 gives the optimal solution to this clustering problem.

connecting two adjacent points, we will get a globally optimal solution to this particular clustering problem. Actually, the general 1D clustering problem of grouping n data points into k clusters can be solved by finding the $k-1$ connecting lines and cutting them. Can this approach be generalized to multi-dimensional clustering problems? This is the main question we will address in this paper.

We have developed a framework for representing a set of multi-dimensional data as a Minimum Spanning Tree (MST), a concept from the graph theory. A tree is a simple structure for representing binary relationships, and any connected component of a tree is called a *subtree*. Through this MST representation, we can convert a multi-dimensional clustering problem to a tree partitioning problem, i.e. finding a particular set of tree edges ('long' edges from either local or global point of view) and then cutting them. Representing a set of multi-dimensional data points as a simple tree structure will clearly lose some of the inter-data relationship. However we have rigorously demonstrated that no essential information is lost for the purpose of clustering. This is achieved through a rigorous proof that each cluster corresponds to one subtree, which does not overlap the representing subtree of any other cluster. Hence a clustering problem is equivalent to a problem of identifying these subtrees through solving a tree partitioning problem. Because of the simplicity of a tree structure, many tree-based optimization problems can be solved efficiently in a similar but generalized fashion to that of their corresponding 1D problems. We will describe, in the following sections, a number of efficient and rigorous tree-based clustering algorithms, some of which have guaranteed global optimality.

In addition to being able to facilitate efficient clustering algorithms, an MST representation also allows us to deal with clustering problems that classical clustering algorithms have problems with. As these algorithms rely on either the idea of grouping data around some 'centers' or the idea of separating data points using some regular geometric curve like a hyperplane, they generally do not fare well when the boundaries of the clusters are very complex (see Figure 1a). An MST, on the other hand, is quite invariant to detailed geometric changes in the boundaries of clusters. For example, the MST representation of the data set of Figure 1b will be quite stable even if we stretch the rectangular-shaped clusters along different directions to make the data set look like the one in Figure 1a as long as the relative distances between clusters (versus the distances within a cluster) do not change significantly. This implies that if our MST-based clustering algorithms work well in the case of Figure 1b, they should basically do equally well in the seemingly more difficult case of Figure 1a. Hence the shape complexity of a cluster has very little effect on the performance of our MST-based clustering algorithms.

Based on the MST-representation, we have developed a number of clustering algorithms. Each of these algorithms finds the optimal clustering based on a different objective function. For a simple case like the one in Figure 1a where the inter-cluster distance is clearly larger than the intra-cluster distance, a simple algorithm by finding and cutting a set of long edges will solve the problem. For cases where boundaries between clusters may not be very clear, an objective function addressing more global properties of a cluster may be needed. Three clustering algorithms will be described. These algorithms, along with the MST representation, have been implemented as a computer program, called EXpression data Clustering Analysis and VisualizATiOn Resource (EXCAVATOR).

Because of the computational efficiency of our clustering algorithms, we can check the optimal K-clustering for many different Ks, i.e. partitioning a data set into K clusters. By examining how the clustering quality improves as we use larger and larger Ks, we should be able to detect where the improvement levels off as K further increases. This capability allows EXCAVATOR to select, for the user, the most 'natural' number of clusters for a given clustering problem. When clustering a set of expression data, the user may have some *a priori* knowledge about

which genes should or should not belong to the same clusters. EXCAVATOR allows a user to specify this type of information as clustering constraints, and it finds optimal clustering results that are consistent with the specified constraints.

MSTs have been used for data classification in the field of pattern recognition (Duda and Hart, 1973) and image processing (Gonzalez and Wintz, 1987; Xu and Uberbacher, 1997; Xu *et al.*, 1998). We have also seen some limited applications in biological data analysis (States *et al.*, 1993). One popular form of these MST applications is called the *single-linkage cluster analysis* (Gower and Ross, 1969; Aho *et al.*, 1974; Jain and Dubes, 1988; Mirkin, 1996). Our study on these methods has led us to believe that all these applications have used the MSTs in some heuristic ways, e.g. cutting long edges to separate clusters, without fully exploring their power and understanding their rich properties related to clustering. In this paper, we will provide in-depth studies for MST-based clustering. Our major contributions include a rigorous formulation for general clustering problems, the discovery of new relationship between MSTs and clustering, and novel algorithms for MST-based clustering.

We have applied EXCAVATOR to a number of expression data sets. In this paper, we will show the clustering results on three data sets, i.e. the gene expression data (a) of the budding yeast *Saccharomyces cerevisiae*, (b) in response of human fibroblasts to serum, and (c) of *Arabidopsis* in response to chitin elicitation.

2 SYSTEMS AND METHODS

2.1 Spanning tree representation of a data set

We will use a MST to represent a set of expression data and their significant inter-data relationships to facilitate fast rigorous clustering algorithms.

Let $D = \{d_i\}$ be a set of expression data with each $d_i = (e_i^1, \ldots, e_i^t)$ representing the expression levels at time 1 through time t of gene i. We define a weighted (undirected) graph $G(D) = (V, E)$ as follows. The vertex set $V = \{d_i | d_i \in D\}$ and the edge set $E = \{(d_i, d_j) | \text{ for } d_i, d_j \in D \text{ and } i \neq j\}$. Hence $G(D)$ is a complete graph. Each edge $(u, v) \in E$ has a weight that represents the distance (or dissimilarity), $\rho(u, v)$, between u and v, which could be defined as the Euclidean distance, the correlation coefficient, or some other distance measures as we will discuss in Section 2.3.

A *spanning tree* T of a (connected) weighted graph $G(D)$ is a connected subgraph of $G(D)$ such that (i) T contains every vertex of $G(D)$, and (ii) T does not contain any cycle. A MST is a spanning tree with the minimum total distance.

A MST of a weighted graph can be found by a *greedy* method, as illustrated by the following strategy used in the classical Kruskal's algorithm (see p. 222 in Aho *et al.*, 1974): *the initial solution is a singleton set containing an edge with the smallest distance, and then the current partial solution is repeatedly expanded by adding the edge with next smallest distance (from the unconsidered edges) under the constraint that no cycle is formed, until all vertices are connected by the selected edges.* A simple implementation of the Kruskal's algorithm (Kruskal, 1956) runs in $O(\|E\| \log(\|E\|))$ time, where $\|\cdot\|$ represents the number of elements in a set. Figure 3 shows an example of a MST of a 2D data set, consisting of four 'natural' clusters.

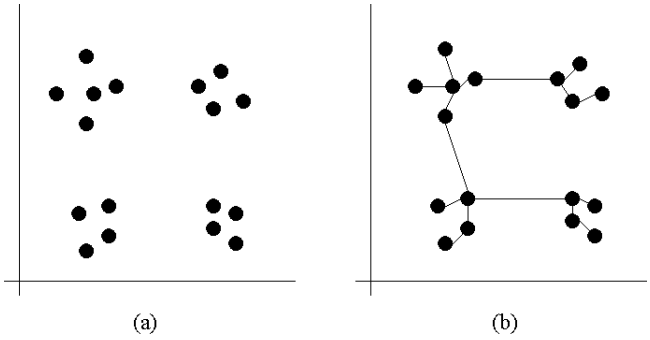


Fig. 3. An MST representation of a set of data points. (a) A set of 2D points. (b) An MST connecting all the data points, using the Euclidean distance. These data points form four natural clusters, based on their relative distances.

By examining Figure 3b, we observe that data points of the same cluster are connected with each other by short tree edges (without data points from other clusters in the middle) while long tree edges link clusters together. We found this is generally the case with an MST representation of any multi-cluster data set. To rigorously prove this, we need a formal definition of a cluster. So what constitutes a cluster in a data set? Though there is a large literature body of clustering algorithms, people have been trying to avoid giving a general definition of a cluster, possibly due to the non-trivial nature of doing so. Without such a definition, any rigorous discussion about clustering algorithms is virtually impossible. Here we provide a *necessary* condition for a subset of a set to be a *cluster*. Let D be a data set and ρ represent the distance between two data points of D.

> *$C \subseteq D$ forms a cluster in D only if for any arbitrary partition $C = C_1 \cup C_2$, the closest data point d to C_1, $d \in D - C_1$, is from C_2. Formally, this can be written as*

$$\arg \min_{d \in D - C_1} \{\min\{\rho(d, c) | c \in C_1\}\} \in C_2, \qquad (1)$$

where $D - C$ represents the subset of D by removing all points of C. We call this the *separability condition* of a cluster. In essence, by this definition, we are trying to capture our intuition about a cluster; that is distances between neighbors within a cluster should be smaller than any inter-cluster distances. Clearly, each of the four 'natural' clusters in Figure 3 satisfies this necessary condition. So does the whole data set. However the subset formed by the cluster in the up-left corner plus any proper subset of the cluster in the up-right corner does not form a cluster.

Now we can rigorously prove that any cluster, C, corresponds exactly to one subtree of its MST representation. That is

> *if c_1 and c_2 are two points of a cluster C, then all data points in the tree path, P, connecting c_1 and c_2 in the MST, must be from C.*

This statement can be proved rigorously. We only give a sketch of the proof here. Let's assume that the statement is incorrect. Hence there exists a point a in path P, which does not belong to C (see Figure 4). Without loss of generality, we assume that a is right next to c_1 on P so (c_1, a) is an edge in P. We define a data set A as follows. Initially $A = \{c_1\}$. We then repeatedly expand A using the following operation until A converges: *select the data point x from $D - A$, which is closest to A; if $x \in C$ add x to A*. Apparently when A converges, $A = C$, based on the separability condition (1) of C being a cluster. This means that there exists a path, P', from c_1 to c_2 that consists of only data points of C and all its edges have smaller distances (ρ) than $\rho(c_1, a)$ (see Figure 4b). We know that at least one edge of P' is not in the current MST. For the simplicity of discussion, we assume that exactly one edge, e, of P' is not in the current MST (the case with multiple such edges can be reduced to the case with only one edge). So $P \cup P'$ contains a cycle with one edge of P' not in the MST. By removing edge (c_1, a) and adding e, we get another spanning tree with smaller total distance. This contradicts the fact that a MST has the minimum total distance among all spanning trees. By having this contradiction, we have proved the statement.

The above statement implies that clustering (of multi-dimensional data) can be rigorously achieved through tree partitioning. So to cluster, all we have to do is to find the *right* set of edges of the MST representation of the data set and cut them; the connected subtrees will give us the desired clusters.

2.2 Comparison between two expression profiles

In our program EXCAVATOR, we provide a number of different ways of measuring the 'distance' between two expression profiles. Based on a user's selection of the distance measure, the program constructs the MST

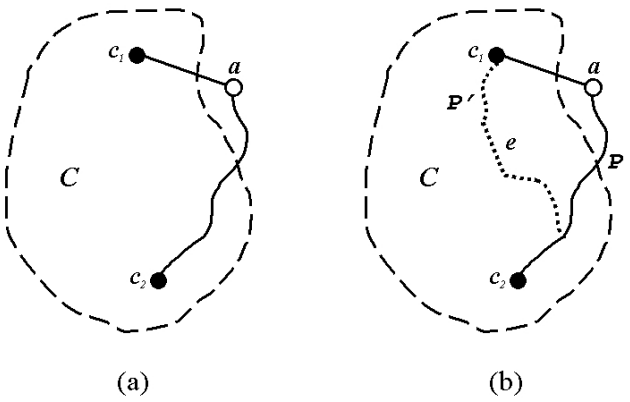


Fig. 4. (a) A path connecting two vertices c_1 and c_2 of the same cluster C (C's boundary is given by the dashed line) with one vertex a from a different cluster. (b) A schematic of the result of the *expansion* operation.

representation of the data set. Let $d_1 = (e_1^1, \ldots, e_1^t)$ and $d_2 = (e_2^1, \ldots, e_2^t)$ be two data points, where e_i^j is a log-transformed (base 2) fluorescence intensity ratio from the microarray data. The distance measures supported in EXCAVATOR include

- Euclidean distance

$$\rho(d_1, d_2) = \sum_{i=1}^{t} (d_1^i - d_2^i)^2.$$

- Correlational distance

$$\rho(d_1, d_2) = 1 - \text{correlation coefficient}(d_1, d_2).$$

- Mahalanobis distance

$$\rho(d_1, d_2) = \sum_{i,j} (d_1^i - \overline{d_1})(d_2^j - \overline{d_2}) w_{i,j},$$

where $\overline{d_i}$ represents the average of the vector d_i, $i = 1, 2$, and $\{w_{i,j}\}$ is the inverse covariance matrix of the data set D.

2.3 MST-based clustering algorithms

Apparently, different clustering problems may need different objective functions, in order to achieve the best clustering results. In this section, we describe three objective functions and their corresponding clustering algorithms. All algorithms presented here are for partitioning a tree into K subtrees, for some specified integer $K > 0$. It is worth mentioning that these algorithms become possible only because of the MST framework that we are using.

2.3.1 Clustering through removing long MST-edges. One simple objective function is to partition an MST into K subtrees so that the total edge-distance of all the K subtrees is minimized. This objective function intends to capture the intuition that two data points with a short edge-distance should belong to the same cluster (subtree) and data points with a long edge-distance should belong to different clusters and hence be cut. It is not hard to rigorously prove that by finding the $K-1$ longest MST-edges and cutting them, we get a K-clustering that achieves the global optimality of the above objective function. This simple algorithm works very well as long as the inter-cluster (subtree) edge-distances are clearly larger than the intra-cluster edge-distances. However, when clusters are not connected through long edges rather by a series of short edges or when there are many un-clustered data points, i.e. data points that do not belong to any apparent clusters, this simple algorithm may fail to produce the best clustering results.

To determine automatically how many clusters there should be, the algorithm examines the optimal K-clustering for all $K = 1, 2, \ldots$, up to some large number to see how much improvement we can get as K goes up. Typically after K reaches the 'correct' number (of clusters), the quality improvement levels off, as we can see in Figure 6a. By locating the transition point, our program can automatically choose the number of the clusters for the user.

2.3.2 An iterative clustering algorithm. We now give another clustering algorithm that attempts to partition the MST T into K subtrees, $\{T_i\}_{i=1}^{K}$, to optimize a more general objective function than the previous one:

$$\sum_{i=1}^{K} \sum_{d \in T_i} \rho(d,\ \text{center}(T_i)), \tag{2}$$

that is to optimize the K-clustering so that the total distance between the 'center' of each cluster and its data points is minimized—this is a typical objective function for data clustering. When different distance measure is used, the *center* has a different value. For example, when the Euclidean distance is used, the average of a cluster is its center; when the correlational distance is used, it can be shown that the center of a cluster C is

$$\text{center}(C) = \sum_{c_i \in C} \frac{c_i}{\sigma_i},$$

where σ_i is the standard deviation of $c_i \in C$.

Our iterative algorithm starts with an arbitrary K-partitioning of the tree (selecting $K-1$ edges and removing them gives a K-partitioning). Then it repeatedly does the following operation until the process converges: *For each pair of adjacent clusters* (*connected by a tree edge*)*, go through all tree edges within the merged cluster of the two to find the edge to cut, which globally optimizes the 2-partitioning of the merged cluster, measured by the objective function* (2). Our experience with this iterative algorithm indicates that the algorithm converges to a local minimum very quickly.

2.3.3 A globally optimal clustering algorithm. We now present an algorithm that rigorously finds the globally optimal solution of the clustering problem defined as follows. We use a slightly different objective function than the objective function (2). In the previous one, we want to group data points around the center of each cluster (to be clustered). Here we want to group data points around the 'best' representatives from our data set. The representatives are not pre-selected but rather they are the results of the optimization process, i.e. our optimization algorithm attempts to partition the tree into K subtrees and simultaneously to select K representatives in such way to optimize the objective function (3). More formally, for a given MST T, we want to partition T into K subtrees, $\{T_1, \ldots, T_K\}$, and to find a set of data points $d_1, \ldots, d_K \in D$ such that the following objective function is minimized.

$$\sum_{i=1}^{K} \sum_{d \in T_i} \rho(d, d_i) \tag{3}$$

where $\rho(\ )$ is the distance function used. The rationale for using a 'representative' rather than the 'center' is that a center may not belong to, or even be close to, the data points of its cluster when the shape of the cluster boundary is not convex, which may result in biologically less meaningful clustering results. The representative-based scheme provides an alternative when center-based clustering does not generate desired results. When clusters have shapes close to being convex sets, the selected representative is generally the closest data point to the center, and hence the two objective functions should lead to similar results (assuming that the clustering algorithm can find the global minimum solution). A good property of the representative-based objective function is that it facilitates an efficient global optimization algorithm.

The very basic idea of our algorithm can be explained as follows. It first converts the MST into a *rooted* tree (Aho *et al.*, 1974) by arbitrarily selecting a tree vertex as the root. Now the *parent–child* relationship is defined among all tree vertices. At each tree vertex v, we define the following: $S(v, k, d)$ *is defined to be the minimum value of the objective function* (3) *on the subtree rooted at vertex* v*, under the constraint that the subtree is partitioned into* k *subtrees and the representative of the subtree rooted at* v *is* d. By definition, the following gives the global

minimum of objective function (3):

$$\min_{d \in D} S(\text{root}, K, d). \tag{4}$$

Our algorithm uses a Dynamic Programming (DP) approach (Aho *et al.*, 1974) to calculate the $S(\,)$ values at each tree vertex v, based on the $S(\,)$ values of v's children in the rooted MST. The core of the algorithm is a set of DP recurrences relating these $S(\,)$ values. The *boundary conditions* of this DP system are given as follows: if a tree vertex v does not have any child, then

$$S(v, k, d) = \begin{cases} +\infty, & \text{for } k > 1, \\ \rho(v, d), & \text{for } k = 1. \end{cases} \tag{5}$$

For each v with children, $S(\,)$ of v is calculated as follows

$$S(v, k, d) = \min_{X \subseteq C_v} \min_{\sum_{i=1}^{\|C_v\|} k_i = k + \|X\| - 1, k_i > 0} \left(\sum_{v_j \in C_v - X} S(v_j, k_j, \overline{d}) + \sum_{v_j \in X} S(v_j, k_j, d) + \rho(v, d) \right), \tag{6}$$

where

$$S(v_j, k_j, \overline{d}) = \min_{x \in D, x \neq d} S(v_j, k_j, x),$$

and C_v represents the set of all children of vertex v. Our algorithm calculates the $S(v, k, d)$ values for all combinations of $v \in T$, $k \in [1, K]$, and $d \in D$.

The correctness of these DP recurrences can be proved based on the observation that $S(v, k, d)$ can be decomposed as the sum of some combination of its children's $S(\,)$ values and that the above DP recurrences covers all possible such combinations. We omit the detailed proof.

The computational time of this algorithm can be estimated as follows. It is not hard to see that for each tree vertex v, computing its DP recurrences takes

$$O\left(2^{\|C_v\|} \binom{K + \|C_v\| - 1}{\|C_v\| - 1} \|C_v\|\right)$$

time, where $\binom{X}{Y}$ denotes the number of possible ways of selecting Y elements out of X elements. Hence the total time, $\mathcal{T}$, for computing all the DP recurrences for the whole tree T is

$$\mathcal{T} \leqslant O\left(\sum_{v \in T} 2^{\|C_v\|} \binom{K + \|C_v\| - 1}{\|C_v\| - 1} \|C_v\|\right).$$

Since

$$\binom{K + \|C_v\| - 1}{\|C_v\| - 1} \leqslant (K + 1)^{s-1},$$

we have

$$\mathcal{T} \leqslant 2^s K^s \sum_v \|C_v\|,$$

where s is the maximum number of children of any tree vertex. Since $\sum_{v \in T} \|C_v\| = n - 1$, we have shown that it takes $O(n(2K)^s)$ time to compute all the $S(\,)$ values, where n is the number of data points in our data set and K is the maximum number of clusters we want to consider. To get the actual clustering that achieves the global minimum value, we need some simple bookkeeping to trace back which tree edges are cut. This can be done within the computational time needed for calculating the $S(\,)$ values. We omit further discussions.

This algorithm runs in exponential time only in the maximum number of children, s, of a tree vertex. To get a sense about how large s could be for a typical application, we have done a number of simulations to estimate s. In the simulation, we have randomly generated a set of 60-dimensional (60 is chosen arbitrarily) data points, and constructed an MST representation of the set. Then we count the number of children of each vertex in this MST. Figure 5 summarizes these counts. This study shows that this global optimization algorithm runs efficiently for a typical clustering problem with a few hundred data points consisting of a dozen or so clusters.

Note that our algorithm finds the optimal k-clustering for all ks simultaneously, $k \leqslant K$, for some pre-selected K. A user can choose the value of K in EXCAVATOR, and the default value of K is a third of the total number of genes. For a particular application, if we set K to, say, 30 or to certain percentage of the total number of vertices, we will get the optimal objective values for any $k = 1, 2, \ldots, K$. By comparing these values, we can automatically select the number of clusters that is most 'natural' as we will discuss in Section 3.1.

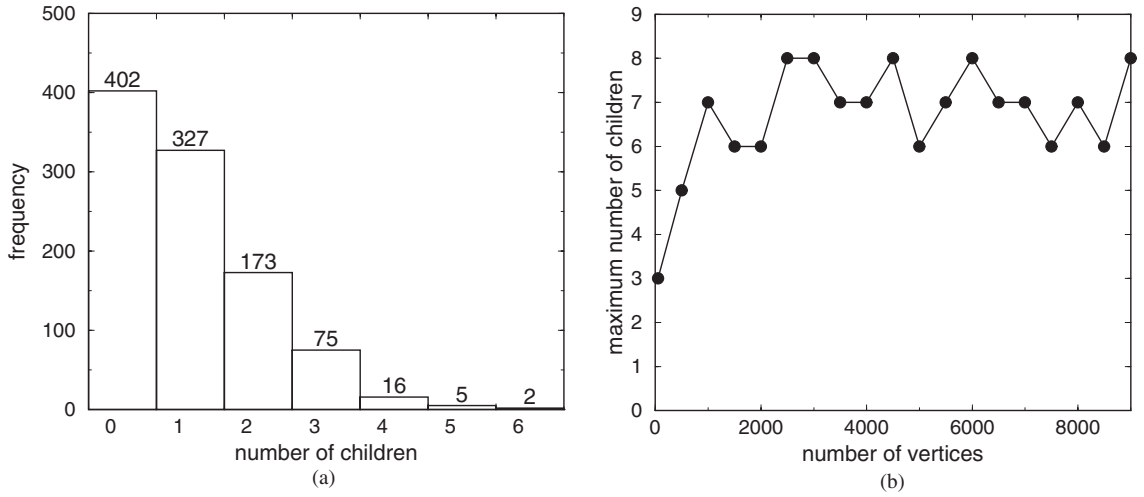
3 RESULTS

3.1 Key features of EXCAVATOR

The core of the EXCAVATOR program is a set of MST-based clustering algorithms. While detailed description of EXCAVATOR will be discussed elsewhere (manuscript in preparation), we now highlight a few key and unique features of the EXCAVATOR program, in addition to the MST-based rigorous and efficient clustering algorithms that we have described above.

- For a user-selected objective function and an integer value K, EXCAVATOR calculates the optimal k-clustering for all $k \in [1, K]$, and then compares these values, as shown in Figure 6. Let $Q(k)$ represent the objective value for the optimal k-clustering for our selected objective function. It selects the $k \in [2, K-1]$ with the highest following value (see Figure 6b) as the most 'natural' number of clusters:

$$\frac{Q(k-1) - Q(k)}{Q(k) - Q(k+1)}. \tag{7}$$



(a) (b)

Fig. 5. (a) The distribution of the number of children of the MST representing a data set with 1000 random data points in 60-dimensional Euclidean space. (b) The maximum number of children versus the total number of data points ranging from 50 to 9000.

This function defines a *transition profile* of $Q(\ )$.

- EXCAVATOR allows a user to specify if any genes should (or should not) belong to the same cluster, based on the user's *a priori* knowledge, and finds the optimal clustering that is consistent with the specified constraints. This feature is implemented as follows. If data points are specified to belong to the same cluster, the algorithm marks the whole MST-path connecting the two points as 'cannot be cut' when doing the clustering. So every data points on this path will be assigned to the same cluster of these two points. Similar is done for two genes that should belong to different clusters.
- EXCAVATOR provides different distance measures and different clustering algorithms. For different clustering results, the program has a capability for measuring the similarity of two clustering results, for comparison purposes. We derived a quantitative measurement, using an approach similar to others (Jain and Dubes, 1988; Mirkin, 1996). Let $\mathcal{D}_1 = \{D_1^1, D_2^1, \ldots, D_N^1\}$ and $\mathcal{D}_2 = \{D_1^2, D_2^2, \ldots, D_M^2\}$ be two clusterings of data set D, one with N clusters and the other with M clusters. We define the measure of similarity between these two clusterings as

$$P_{\text{diff}}(\mathcal{D}_1, \mathcal{D}_2) = \sum_{i,j} \frac{\|D_i^1 \bigcap D_j^2\|}{\|D_i^1 \bigcup D_j^2\|}[\|D_i^1\| + \|D_j^2\|]. \tag{8}$$

It can be proved that P_{diff} has the following upper and lower bounds,

$$P_{\min} \leqslant P_{\text{diff}}(\mathcal{D}_1, \mathcal{D}_2) \leqslant P_{\max}, \tag{9}$$

where

$$P_{\min} = \|D\| + \min \left(\sum_i \frac{\|D_i^1\|^2}{(M-1)\|D_i^1\| + \|D\|}, \sum_j \frac{\|D_j^2\|^2}{(N-1)\|D_j^2\| + \|D\|)} \right); \tag{10}$$

$$P_{\max} = 2\|D\|. \tag{11}$$

The following quantity, which ranges from 0 to 1, gives a good measurement on the (dis)similarity between the two clustering results $\mathcal{D}_1$ and $\mathcal{D}_2$,

$$\frac{P_{\text{diff}}(\mathcal{D}_1, \mathcal{D}_2) - P_{\min}}{P_{\max} - P_{\min}}. \tag{12}$$

The value is 1 if and only if the two partition results are the same. The closer the value is to 0, the more dissimilar the two partition results are.

3.2 Application results

We now describe the application results to three data sets.

3.2.1 Yeast data. Our first application is on a set of gene expression data in the budding yeast *S. cerevisiae* (Eisen *et al.*, 1998), with each gene having 79 data points (or 79 dimensions). We selected four clusters (68 genes in total) determined in the paper (Eisen *et al.*, 1998). These are (1) protein degradation (cluster C), (2) glycolysis (cluster E), (3) protein synthesis (cluster F), and chromatin (cluster H). Genes in each of these four cluster share similar expression patterns and are annotated to be in the same biological pathway. The goal of this application is to compare our clustering results with known cluster information.

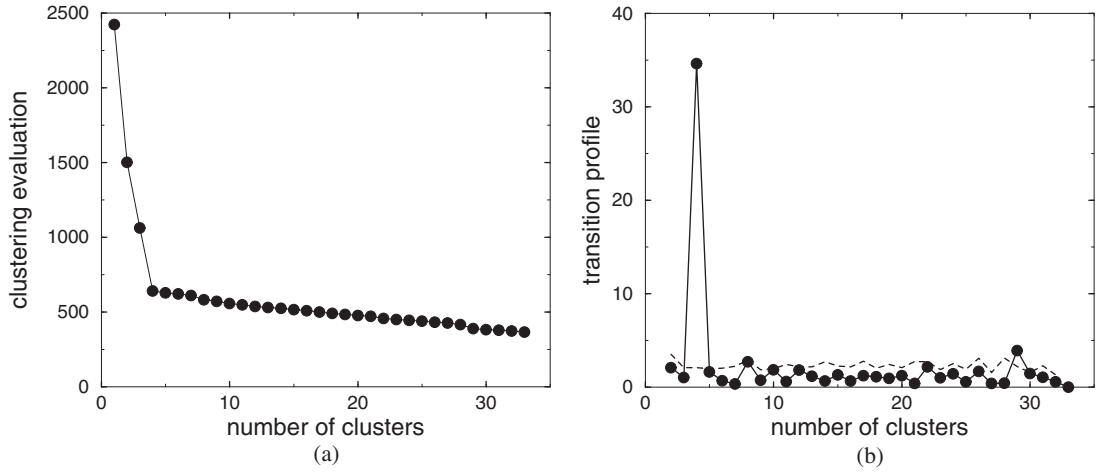


Fig. 6. (a) Objective function values versus the number of clusters. (b) The transition profile value, calculated by function (7), versus the number of clusters. The dashed line shows the transition profile for a set of random data.

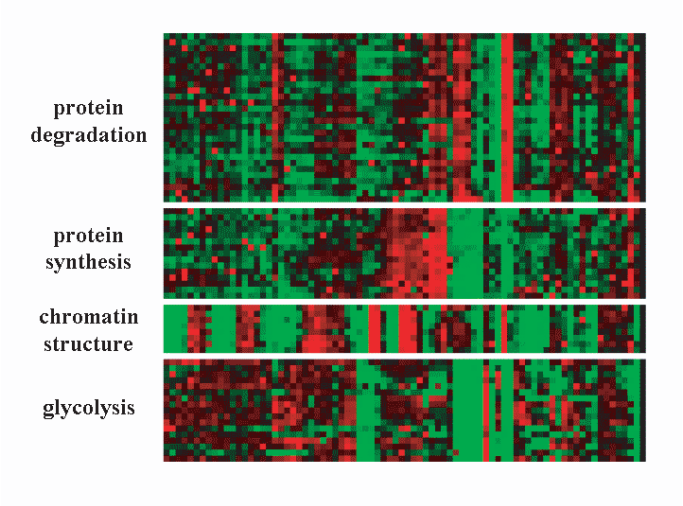


Fig. 7. Expression profiles and clustering results of the yeast data. Red indicates high expression and green indicates low expression.

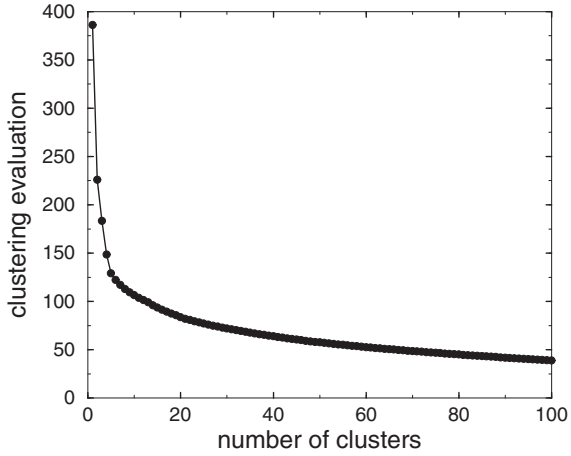


Fig. 8. Objective function values versus the number of clusters, for the gene expression data in response of human fibroblasts to serum.

For this application, we have applied all three clustering algorithms, using both the Euclidean distance and the correlation distance as the distance measure. The computering time on a PC for all the calculation from 2 clusters to 34 clusters was less than 1 s for clustering through removing long MST-edges, less than 7 s for the iterative algorithm, and less than 20 s for the globally optimal algorithm. We have achieved virtually identical clustering results, using any combination of these algorithms and functions. Here we show the clustering result obtained, using our first clustering algorithm with the Euclidean distance as the distance measure. Figure 6 shows how the objective function values improve as the number of clusters increases. This provides a profile similar to the 'Scree Test' (Cattell, 1966). Based on the transition profile in Figure 6b, the program decides a 4-clustering gives the most 'natural' number of clusters for this problem. Figure 7 gives the 4-clustering results, which is 100% in agreement with the annotated results in Eisen *et al.* (1998).

3.2.2 Human serum data. The second application is on a set of temporal gene expression data in response of human fibroblasts to serum (Iyer *et al.*, 1999). The data set consists of 517 genes, and each gene has 18 data points. We have used both the first and the second of our clustering algorithms for this problem, with the Euclidean distance as the distance measure. Figure 8 shows the optimal k-clustering quality values versus the number of clusters, k, by our second algorithm. We can see that a significant improvement in clustering quality is being made when increasing k from 1 to 5,

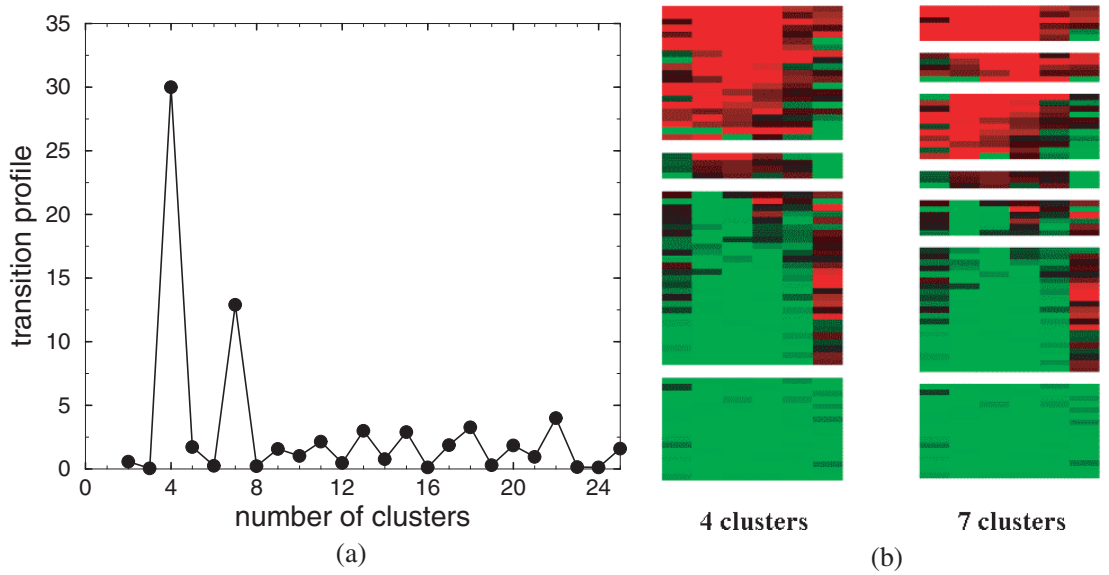


Fig. 9. Clustering results for the *Arabidopsis* data. (a) The transition profile versus the number of clusters; (b) clustering results for optimal 4-clustering and optimal 7-clustering.

and then the rate of improvements drops. Our program suggests that an optimal 5-clustering gives the initial coarse-grained clustering of the whole data set. Different levels of finer-grained clusterings are then made. Detailed clustering results can be found at http://compbio.ornl.gov/structure/clustering/pnas-paper/. Verifications of our clustering results will be done in a follow-up study. Not surprisingly, our first clustering algorithm did not do nearly as well as the second algorithm, based on our visual inspection of the similarities of 517 gene expression profiles, since unlike our first test case, cluster boundaries here are far from being apparent.

3.2.3 Arabidopsis data. Our third application is on a set of gene expression data of *Arabidopsis* in response to chitin elicitation (Ramonel *et al.*, 2001). The data was averaged over two experiments. Each gene had six data points (collected at 10 min, 30 min, 1 h, 3 h, 6 h, and 24 h). 68 genes were selected for clustering, each containing at least one data point with a 3-fold change of expression level by chitin elicitation. We used both the second and third algorithms for this problem. Here we present the clustering results by the third algorithm, with the Euclidean distance as the distance measure. From Figure 9a, we can see there are two high peaks in the transition profile, indicating that there are at least two levels of clustering, one with four clusters and one further dividing the four clusters into seven clusters. Figure 9b shows the clustering results for both the optimal 4-clustering and optimal 7-clustering.

Through searching the regulatory regions of these genes, we found that a known *cis*-acting element of chitin-responsive genes, i.e. the W-box hexamer, was over-represented in genes of one of 7 clusters. This suggests that these genes are not only co-expressed, but also co-regulated through the W-box motif (Ramonel *et al.*, 2001).

4 DISCUSSION AND CONCLUSION

In this paper, we have described a rigorous computational framework for clustering multi-dimensional gene-expression profiles. To the best of our knowledge, the *separability condition* that we proposed here is the first rigorous mathematical formulation for the general clustering problem, although various mathematical models for particular clustering problems have been given previously by other researchers. Under such a formulation, we have rigorously proved that an MST representation captures all the essential information of a multi-dimensional data set for the purpose of clustering. Hence a multi-dimensional data clustering problem can be solved as a tree partitioning problem. This realization has led to the discovery of a number of both rigorous and efficient clustering algorithms, particularly the ones with guaranteed global optimality for some general objective functions.

Based on this new data-representation framework and the MST-based clustering algorithms, we have developed a computer program EXCAVATOR for gene expression data clustering. We believe that the various unique features of EXCAVATOR will make the program a highly useful tool in mining the large-scale gene expression data, in a reliable and meaningful way.

Our computational framework of clustering provides a mathematical foundation, which, we believe, will open the door for many clustering/classification problems. Research is underway to apply our framework to various biological data analysis problems, including phylogenetic classification, motif recognition in biological sequences, protein family classification, etc.

ACKNOWLEDGEMENTS

We thank Dr Li Wang and Chen Yu of the Protein Informatics Group of Oak Ridge National Laboratory for helpful discussions related to this work. We also thank Dr Shauna Somerville for providing the expression data in *Arabidopsis*. This work is supported by the Office of Biological and Environmental Research, US Department of Energy, under Contract DE-AC05-00OR22725, managed by UT-Battelle, LLC.

REFERENCES

Aho,A.V., Hopcroft,J.E. and Ullman,J.D. (1974) *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, MA.

Cattell,R. (1966) The scree test for the number of factors. *Multivariate Behavioral Res.*, **1**, 245–276.

Duda,R.O. and Hart,P.E. (1973) *Pattern Classification and Scene Analysis*. Wiley–Interscience, New York.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Gonzalez,R.C. and Wintz,P. (1987) *Digital Image Processing*, 2nd edn, Addison-Wesley, Reading, MA.

Gower,J.C. and Ross,G.J.S. (1969) Minimum spanning trees and single linkage analysis. *Appl. Stat.*, **18**, 54–64.

Herwig,R., Poustka,A.J., Müller,C., Bull,C., Lehrach,H. and O'Brien,J. (1999) Large-scale clustering of cDNA-fingerprinting data. *Genome Res.*, **9**, 1093–1105.

Iyer,V.R., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C.F., Trent,J.M., Staudt,L.M., Hudson,J. Jr, Boguski,M.S., Lashkari,D., Shalon,D., Botstein,D. and Brown,P.O. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.

Jain,A.K. and Dubes,R.C. (1988) *Algorithms for Clustering Data*. Prentice-Hall, New Jersey.

Kruskal,J.B. Jr (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.*, **7**, 48–50.

Mirkin,B. (1996) *Mathematical Classification and Clustering*. DIMACS, Rutgers University, Piscataway, NJ.

Ramonel,K.M., Zhang,B., Ewing,R., Chen,Y., Xu,D., Gollub,J., Stacey,G. and Somerville,S. (2001) Microarray analysis of chitin elicitation in *Arabidopsis thaliana*, submitted.

Sherlock,G. (2000) Analysis of large-scale gene expression data. *Curr. Opin. Immunol.*, **12**, 201–205.

States,D.J., Harris,N.L. and Hunter,L. (1993) Computationally efficient cluster representation in molecular sequence megaclassification. *Ismb*, **1**, 387–394.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Wen,X., Fuhrman,S., Carr,G.S.M.D.B., Smith,S., Barker,J.L. and Somogyi,R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA*, **95**, 334–339.

Xu,Y., Olman,V. and Uberbacher,E.C. (1998) A segmentation algorithm for noisy images: design and evaluation. *Patt. Recogn. Lett.*, **19**, 1213–1224.

Xu,Y. and Uberbacher,E.C. (1997) 2D image segmentation using minimum spanning trees. *Image Vis. Comput.*, **15**, 47–57.