A Line Graph Algorithm for Clustering Chemical Structures Based on Common Substructural Cores

John W. Raymond (john.raymond@pfizer.com)
Pfizer Global Research and Development, Ann Arbor Laboratories,
2800 Plymouth Road, Ann Arbor, Michigan 48105, USA

Peter Willett (p.willett@sheffield.ac.uk)
Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

Abstract

There is a need among chemists for the ability to cluster large numbers of chemical structures based on the presence of common substructural templates. This paper describes a simple algorithm for this task that is based on a line graph interpretation of the proximity graph and on a graph representation of 2D chemical structures. This permits the use of a graph-theoretic similarity measure based on the maximum common edge subgraph to determine the appropriate substructural template needed by the algorithm.

1. Introduction

The clustering of chemical structures based on pair-wise inter-molecular structural similarities is well studied and has become an effective research tool in chemical information management [1], with the pair-wise similarities being calculated using graph-based or feature-based measures [2]. One of the potential limitations of the pair-wise similarity approach, however, is that it is possible for collections of structures exhibiting a sufficient degree of pair-wise similarity to be clustered together, but whose commonality is far less when considered from the perspective of all of the structures in the cluster. Conversely, it is also possible to have a collection of chemical structures which would be classified together by a chemist based on a perceived substructural commonality but are classified into multiple clusters by a pair-wise similarity algorithm dependent upon the variation in the non-conserved portion of the chemical structures. What is desired is a clustering procedure that attempts to enforce collective similarity in a cluster of chemical structures by preserving a sufficient degree of substructural commonality.

In this paper, we introduce a technique based on a line graph interpretation of clustering which allows structures to be classified based on common substructural cores rather than exclusively using a pair-wise similarity score. It is suggested that this approach may more adequately mirror a chemist's notion of chemical structure similarity than do existing approaches to the calculation of structural similarity. The algorithm also allows structures exhibiting multiple classes of activity to be assigned to multiple clusters; thereby, resolving the problem of overlapping clusters without arbitrarily disconnecting related clusters which can result in a loss of important information.

2. Definitions

All graphs referred to in the following text are assumed to be labeled and undirected. For an introduction to graph related concepts and notation, the reader is referred to an introductory text on graph theory [3]. A graph G consists of a set of vertices $V(G)$ and a set of edges $E(G)$. The vertices in G are connected by an edge if there exists an edge $e_k = (v_i,v_j)$ Î $E(G)$ connecting the vertices v_i and v_j in G such that v_i Î $V(G)$ and v_j Î $V(G)$. The vertex and edge labels are denoted as $w(v_i)$ and $w(v_i,v_k)$, respectively. The set of vertices adjacent to vertex v_i is the ***neighborhood***, $N(v_i)$, of v_i.

A *line graph* $L(G)$ is a graph whose vertex set consists of the edge set of G; therefore, if (v_i, v_j) is an edge in G it is also a vertex in $L(G)$. A pair of vertices in $L(G)$ are adjacent if the two corresponding edges in G are incident on each other [4]. A ***maximum common edge subgraph*** (MCES) is a subgraph consisting of the largest number of edges common to both G_1 and G_2. Note that the MCES between two graphs is not necessarily connected or unique by definition. To illustrate these concepts, Figure 1(a) depicts the MCES between two graphs G_1 and G_2, and Figure 1(b) illustrates the line graph $L(G)$ of graph G.

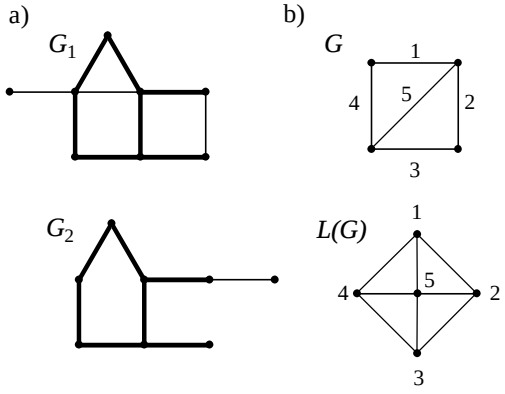

Figure 1. a) MCES between graphs G_1 and G_2 b) Line graph of graph G

3. Line Graph Formulation

As mentioned previously, the proposed clustering method addresses the question of whether objects should be clustered together by considering information in addition to a simple pair-wise measure of similarity. A convenient means with which to compare graphical objects is the MCES between each pair of graphs. It has been shown that the MCES is directly related to the edit distance between graphs [5, 6], providing a convenient description of graph similarity. Recently, an efficient MCES algorithm for the purpose of calculating graph similarity has been published [2, 7, 8]. In the clustering procedure proposed here, pairs of graph objects are clustered with other pairs, based on how similar the corresponding MCESs are to each other.

Using the terminology of Matula [9], the input to a clustering algorithm can be represented by a proximity graph (G_p) where each vertex of the proximity graph corresponds to an object being clustered, and an edge between any two vertices of the proximity graph is weighted with the pair-wise similarity value between the two objects represented by the edge's two endpoint vertices. Rather than clustering the vertices of the proximity graph, our algorithm clusters its edges. This is accomplished by performing the clustering on the line graph of the proximity graph, $L(G_p)$, rather than the proximity graph itself. Since each vertex of $L(G_p)$ corresponds to an edge in the proximity graph, it is weighted with the MCES corresponding to the edge in the proximity graph. An edge of $L(G_p)$ is weighted with the similarity between each pair of MCESs (i.e., the MCES between two MCESs).

As an example of how the line graph approach may better identify chemical series, a data set of 550 structures containing some well-defined series as well as numerous unrelated compounds was clustered in two ways: using the well-known Ward's clustering method with the Kelley validation index [10]; and using the heuristic line graph-based algorithm proposed in this paper. The Kelley validation index for a particular clustering at level $\mathbf{l}$ is calculated using:

$$Kelley = (n - 2) \times \left(\frac{\bar{d}_{\mathbf{l}} - \min\{\bar{d}\}}{\max\{\bar{d}\} - \min\{\bar{d}\}} \right) + 1 + k_{\mathbf{l}}$$

where n is the number of objects being clustered, $k_{\mathbf{l}}$ is the number of clusters, $\bar{d}_{\mathbf{l}}$ is the average similarity of all clusters, and $\min\{\bar{d}\}$ and $\max\{\bar{d}\}$ are the minimum and maximum of $\bar{d}_{\mathbf{l}}$ over all clustering levels, respectively.

The 18 melatonin structures contained in the data set were all correctly clustered into a single series by the line graph-based algorithm, whereas, the Ward's/Kelley approach

separated the series into six different clusters. In addition, the series of 10 bioflavanoid analogues and 5 steroid analogues were properly clustered into two distinct clusters by the line graph method, but the Ward's/Kelley approach clustered the bioflavanoids into three different clusters and the steroids into two clusters. The data set also contained three distinct opiate series: phenylpiperidines (9 compounds), phenylheptylamines (5 compounds), and phenanthrenes (12 compounds). The proposed line graph procedure clustered the phenylpiperidines into two clusters, the phenylheptylamines into a single cluster, and the phenanthrenes into a single cluster. The Ward's/Kelley method clustered the phenylpiperidines into six clusters, the phenylheptylamines into a single cluster, and the phenanthrenes into two clusters.

Although the proposed line graph approach to clustering may better reflect the desired description of chemical graph similarity desired by many chemists (by focusing on how chemicals are similar in addition to the magnitude of that similarity), any line graph-based formulation suffers a potentially significant limitation. Since a proximity graph consisting of N vertices can contain $O(N^2)$ edges, $L(G_p)$ can contain $O(N^2)$ vertices and $O(N^4)$ edges, with the result that the line graph approach can become computationally infeasible for large N. Hence, the proposed line graph algorithm employs several simplifying heuristics in an attempt to preserve the desirable characteristics of the line graph formulation while significantly reducing the computational burden in practice.

4. Algorithm

The output of the proposed algorithm is "fuzzy" in the sense that an object being clustered can appear in more than one cluster. This is a potentially desirable feature in a clustering algorithm when dealing with possibly overlapping clusters.

The proposed algorithm is given in the following series of steps:

Step 1: Calculate Proximity Graph Similarity Values

Calculate the MCES-based similarity between each pair of chemical graphs being clustered [2, 8] and establish a minimum similarity threshold, S_{G_p}, for which the edge corresponding to any pair-wise similarity value not meeting the threshold value is deleted from the proximity graph (G_p). This has the effect of significantly reducing the number of edges in G_p (i.e., vertices in $L(G_p)$) and should not detrimentally affect the clustering results as chemical structures exhibiting similar biological activity tend to exhibit similar graph-based similarity [2, 11]. Each edge in the proximity graph is therefore weighted with the

corresponding MCES-based similarity value if it exceeds the threshold value; otherwise, no edge exists.

Step 2: Determine Connected Components

Next, G_p is separated into connected components (i.e., disconnected subgraphs) since deleting edges not exceeding threshold similarity S_{G_p} may disconnect the proximity graph. This is a simple $O(N^2)$ operation [12].

Step 3: Generate Sub-Cluster Sequences

This step is performed for each connected component of G_p. Generate a sub-cluster sequence for each connected component by separating the neighborhood of each vertex v_i in each component using the following procedure: For each vertex $v_i \hat{I} V(G_p^m)$, where G_p^m denotes the m^{th} component of the proximity graph G_p, separate the edges of the neighborhood of v_i present in the m^{th} component, denoted $N^m(v_i)$, into sub-clusters by calculating the similarity between pairs of edges in $N^m(v_i)$ and dividing the set of edges based on these calculated similarities. The n^{th} sub-cluster generated for the m^{th} component will be denoted by C_n^m.

Since the similarity between each neighborhood edge is defined using the MCES between two MCESs, the choice of similarity coefficient is important. Suppose an edge in $N^m(v_i)$ corresponds to an MCES between two chemical graphs which are almost identical and another edge in $N^m(v_i)$ corresponds to an MCES between one of these two chemical graphs and a third chemical graph. The two MCESs represented by the two edges in $N^m(v_i)$ may in fact both contain the same substructural template characterizing the perceived pharmacological activity, but since the MCES between the two almost identical chemical graphs can be substantially larger than the other MCES, a similarity coefficient which considers the size of each MCES equally may not adequately describe the desired description of similarity between the two pairs of compounds.

To avoid this potential limitation, it is suggested that the asymmetric similarity coefficient be used to calculate the similarity between neighborhood edges. This is

$$S_{ij,ik} = |G_{ij,\,ik}| / \min\{|G_{ij}|, |G_{ik}|\},$$

where $|G_{ij}|$ and $|G_{ik}|$ are the sizes of the MCESs between the pairs of chemical graphs (G_i, G_j) and (G_i, G_k) in the proximity graph, respectively, and $|G_{ij,ik}|$ is the size of the MCES between the MCESs G_{ij} and G_{ik}. Two edges in $N^m(v_i)$ are assigned to the same sub-cluster using a

greedy procedure if the value of $S_{ij,ik}$ exceeds a specified intra-cluster similarity value, S_a. Given M connected components in the proximity graph, this process will result in M distinct sub-cluster sequences being generated.

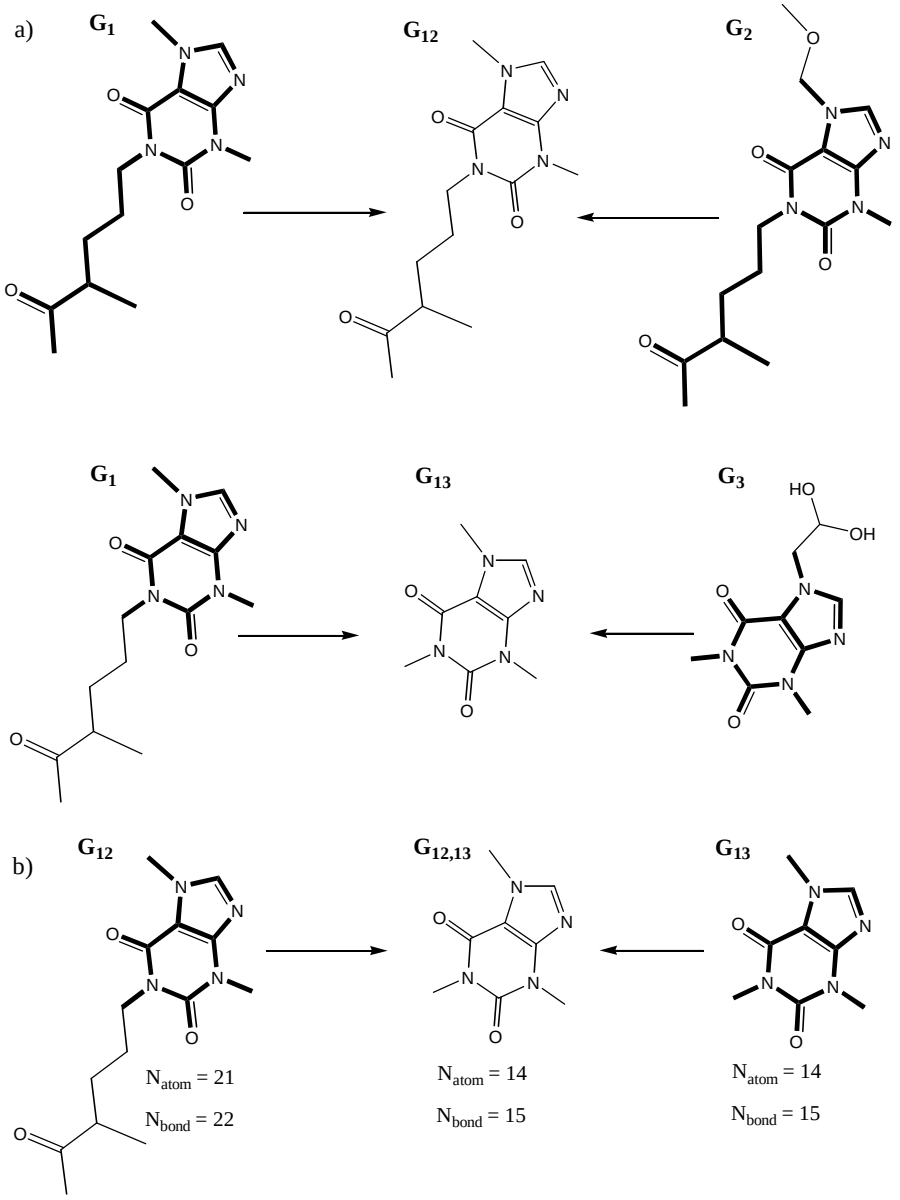

Figure 2. MCES Similarity Example

Figure 2 illustrates this scenario. First we define $|G|$ as $|V(G)| + b \times (1 - a \times (n_p^G - 1)) \times |E(G)|$ where $|V(G)|$ and $|E(G)|$ are the number of atoms and bonds in the chemical graph, respectively. The variable n_p^G represents the number of unconnected subgraph components in graph G containing p or more edges. If all subgraphs have fewer than p edges, then n_p^G will be assumed to be the total number of subgraph components. The constant b reflects the additional weight assigned to matched bond pairs with respect to

compatible atoms, and the constant α is a penalty score for each unconnected component present in *G*. In previous studies, we have found values of p=3, $\alpha = 0.05$, and $\beta = 2.0$ seem to be effective in discerning chemical similarity [2].

It can be seen in Figure 2(a) that all of the graphs (G_1, G_2, and G_3) are related by a xanthine substructural moiety. Both G_1 and G_2 are very similar, and the MCES (G_{12}) is hence very large. However, when G_1 is compared to G_3, the G_{13} is much smaller than G_{12} even though all three chemical graphs under consideration are xanthine-based compounds.

Figure 2(b) demonstrates the comparison between MCES graphs G_{12} and G_{13}. Using the asymmetric coefficient to compute the degree of similarity based on $G_{12,13}$ yields $S_{12,13}$=44/min{65,44}=1, strongly indicating that G_{12} and G_{13} should be grouped together (indicating, indeed, that G_{13} is a subgraph of G_{12}). However, using the Tanimoto coefficient which is given as $S_{ij,ik} = \left|G_{ij,ik}\right| / \left(\left|G_{ij}\right| + \left|G_{ik}\right| - \left|G_{ij,ik}\right|\right)$ yields $S_{12,13} = 44/(65 + 44 - 44) = 0.68$ which is significantly lower.

Step 4: Merge Sub-Clusters

The final clustering of G_p is achieved by merging each sub-cluster sequence into full cluster(s). In this procedure, each of the *m* sub-cluster sequences is considered individually. The size of each sub-cluster, $\left|C_n^m\right|$, in a particular sub-cluster sequence is determined by summing the number of distinct vertices preserved when the edges contained in each sub-cluster are projected onto the proximity graph (i.e., the number of unique chemical graphs represented in each sub-cluster). The sub-clusters in each sequence are then sorted in order of decreasing value of $\left|C_n^m\right|$.

A greedy procedure is then used to merge the sub-clusters in each sequence using the property that the current cluster and a sub-cluster are merged into a larger cluster if the similarity value based on the number of structures in common exceeds a threshold value, S_b. For instance, if the cluster and the sub-cluster contain 6 and 4 unique structures, respectively, and they have 3 structures in common, then the asymmetric coefficient yields a similarity S=3/min{6,4}=0.75. If 0.75 > S_b, then the cluster and sub-cluster would be merged into a single cluster. The number of clusters resulting from the merging procedure will be greater than or equal to *M*.

To illustrate how ordering the sub-clusters in order of non-increasing values of $\left|C_n^m\right|$, a simple test was performed on a set of 358 compounds of various activities. The threshold values used in the analysis for S_{GP}, S_a, and S_b were 0.7, 0.9, and 0.6, respectively, with S_{GP}

being determined using the Tanimoto similarity coefficient. Two different clustering simulations were performed. One was run using the suggested ordering of sub-clusters, and the other was run using random selection. Each resultant clustering was then compared to a manually constructed clustering of the same data set using the Jaccard cluster similarity coefficient [13] which ranges from 0 to 1 with 1 indicating the two clustering are identical. The suggested ordering resulted in a Jaccard coefficient value of 0.61, whereas, the random selection resulted in a Jaccard value of 0.55, indicating a slight advantage for the suggested ordering.

5. Pseudo-Code

The algorithm is given more succinctly in pseudo-code below.

```
Input: Set of N graphs, similarity thresholds S_Gp, S_a, and S_b
Output: Set of final clusters A
Procedure Line Graph Cluster()
{
   Generate the proximity graph (G_p) using an MCES similarity.
   Prune the edges in G_p not exceeding S_Gp.
   Separate G_p into M connected components.
   Generate Sub-Cluster Sequences.
   Merge Sub-Clusters into cluster subgraphs.
}

Input: Set of M connected components in G_p
Output: Set of M sub-cluster sequences C^m
Procedure Generate Sub-Cluster Sequences()
{
   For each connected component G_p^m Î G_p
   {
      Set n:=1.
      For each vertex v_i Î G_p^m
      {
         Identify N^m(v_i), the neighborhood of v_i in G_p^m.
         Set P:=E(N^m(v_i)).
         Sort the edges in P in order of non-increasing similarity.
         While P ¹ Æ do
         {
            Set C_n^m :=Æ.
            Select the first unclustered edge e_k in P.
            Assign e_k to sub-cluster C_n^m (i.e., C_n^m :=C_n^m U e_k).
            Remove e_k from P (i.e., P:=P\e_k).
            While ($e_j)[e_j Î P and S_jk ³ S_a] do
            {
               Select the first unclustered edge e_j in P with an MCES
               asymmetric similarity S_jk ³ S_a.
               Set C_n^m :=C_n^m U e_j.
               Set P:=P\e_j.
            }
            Set n:=n+1.
```

```
         }
      }
   }
}

Input: Set of M sub-cluster sequences C^m
Output: Set of final clusters A
Procedure Merge Sub-Clusters()
{
   Set i:=1.
   For each sub-cluster sequence C^m
   {
      Sort the sub-clusters C_n^m of C^m in order of decreasing value of |C_n^m|.
      While C^m ¹ Æ do
      {
         Set A_i:=Æ.
         Select first unclustered sub-cluster C_n^m in C^m.
         Assign C_n^m to cluster A_i (i.e., A_i:=A_i U C_n^m).
         Remove C_n^m from C^m (i.e., C^m:=C^m \C_n^m).
         While($C_n^m)[C_n^m Î C^m | S_{A_iC_n^m} ³ S_b] do
         {
            Select the first unclustered sub-cluster C_n^m in C^m |
            S_{A_iC_n^m} ³ S_b.
            Set A_i:= A_i U C_n^m.
            Set C^m:=C^m \C_n^m.
         }
         Set i:=i+1.
      }
      Set i:=i+1.
   }
}
```

The algorithmic complexity of the proposed algorithm in the average case is difficult to determine. In practice, it has been found that the MCES comparison is the time-limiting step, and the time for clustering is dominated by the number of necessary MCES comparisons rather than the number of clustering specific operations.

6. Conclusion

In this paper, we have addressed the clustering of chemical structures represented as graphs based on the concept of a common substructural core using a novel line graph approach. The technique has been presented in terms of a graph-based similarity measure involving the MCES between two structures represented as chemical graphs although it is equally applicable for use with a feature-based similarity method such as chemical fingerprints where the bits in common between the two fingerprints are used in lieu of the MCES. Since a naïve implementation of the line graph approach is computationally demanding, a heuristic algorithm has been proposed that employs three user-specified

similarity threshold parameters to reduce the number of comparisons necessary to form the final clustering.

In preliminary testing of the proposed algorithm, it has been found that values of S_{Gp}, S_a and S_b in the ranges (0.7-0.75), (0.8-0.85), and (0.6-0.7), respectively, seem to work well, although further testing is required to establish whether the optimal values fall within these ranges. The S_{Gp} values assume that the Tanimoto coefficient is used and that p=3, $a = 0.05$, and $b = 2.0$. The S_a range is based on the asymmetric coefficient with a equal to zero (i.e., no fragmentation penalty), and the S_b range is also based on the asymmetric coefficient.

Initial time comparisons indicate that the proposed algorithm is approximately 20% to 50% slower than Ward's/Kelley clustering on data sets ranging from a few hundred to over a thousand compounds with the time difference decreasing as the data set size increases for MCES-based similarity calculations. The time difference is due to the sub-cluster sequence generation procedure used in the proposed algorithm. Having described this algorithm in detail, it now remains to compare its effectiveness for the clustering of chemical structures when compared with existing approaches and to establish the optimal values for the threshold parameters: this work will be reported shortly.

Aside from the proposed clustering algorithm, the line graph interpretation of clustering introduced in this paper may prove to be useful in future clustering applications using existing or specifically tailored algorithms.

Acknowledgments

We thank the following: Pfizer (Ann Arbor) for funding; John Blankley, Alain Calvet, Eric Gifford, Christine Humblet, and Sherry Marcy for helpful advice and support. The Krebs Institute for Biomolecular Research is a designated centre of the Biotechnology and Biological Sciences Research Council.

References

1. P. Willett, Similarity and Clustering in Chemical Information Systems, Research Studies Press, (1987).
2. J. Raymond and P. Willett, Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases, J. Comput.-Aided Mol. Des., in the press.
3. R. Diestel, Graph Theory, Springer-Verlag, (2000).
4. A. van Rooij and H. Wilf, The Interchange Graph of a Finite Graph, Acta Math. Hungar., 16 (1965), 263-269.

5. H. Bunke, On a Relation Between Graph Edit Distance and Maximum Common Subgraph, Patt. Recog. Lett., 18 (1997), 689-694.
6. G. Chartrand, F. Saba and H. Zou, Edge Rotation and Distance Between Graphs, Cas. Pest. Mat., 110 (1985), 87-91.
7. J. Raymond, E. Gardiner and P. Willett, Heuristics for Rapid Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm, J. Chem. Inf. Comput. Sci., 42 (2002), 305-316.
8. J. Raymond, E. Gardiner and P. Willett, RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs, Comput. J., in the press.
9. D.W. Matula, Graph Theoretic Techniques for Cluster Analysis Algorithms, in: *Classification and Clustering*, J. Van Ryzin, Ed., Academic Press (1977), 95-129.
10. D.J. Wild and C.J. Blankley, Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering, J. Chem. Inf. Comput. Sci., 40 (2000), 155-162.
11. M. Johnson, Relating Metrics, Lines and Variables Defined on Graphs to Problems in Medicinal Chemistry, in: *Graph Theory and Its Applications to Algorithms and Computer Science*, Y. Alavi, *et al.*, Ed., J. Wiley & Sons (1985), 457-470.
12. E. Allburn, Graph Decomposition: Imposing Order on Chaos, Dr. Dobbs J., 16 (1991), 88,90-2,94-6,118-20,122,124.
13. G.W. Milligan, A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis, Psychometrika, 46 (1980), 187-199.