

Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures

John W. Raymond^{a,*}, C. John Blankley^a, Peter Willett^b

^a Pfizer Global Research and Development, Ann Arbor Laboratories, 2800 Plymouth Road, Ann Arbor, MI 48105, USA

^b Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

Received 2 August 2002; received in revised form 6 November 2002; accepted 14 November 2002

Abstract

This paper compares several published methods for clustering chemical structures, using both graph- and fingerprint-based similarity measures. The clusterings from each method were compared to determine the degree of cluster overlap. Each method was also evaluated on how well it grouped structures into clusters possessing a non-trivial substructural commonality. The methods which employ adjustable parameters were tested to determine the stability of each parameter for datasets of varying size and composition. Our experiments suggest that both graph- and fingerprint-based similarity measures can be used effectively for generating chemical clusterings; it is also suggested that the CAST and Yin–Chen methods, suggested recently for the clustering of gene expression patterns, may also prove effective for the clustering of 2D chemical structures.

© 2002 Elsevier Science Inc. All rights reserved.

Keywords: Bit-string; Chemical graph; Chemical series; Clustering method; Fingerprint; Maximum common subgraph; Molecular similarity; Similarity

1. Introduction

Cluster analysis methods are used to identify groups, or clusters, of similar objects in multivariate datasets [1]. In brief, a cluster analysis involves the following components: a set of objects, each of which is represented by one or more attributes; a measure of the similarity (or dissimilarity or distance) between pairs of objects, between an object and a cluster, or between a pair of clusters; and a clustering method that processes the similarity data to identify groups that are both internally homogeneous and distinct. The reader should note that there may be several different algorithms that can implement a particular clustering method; for example, Rohlf reviews a range of algorithms for the single linkage method [2], which is an hierarchic agglomerative method that fuses pairs of clusters on the basis of the objects, one in one cluster and one in the other, that are most similar to each other. Many other methods, conversely, are defined solely in algorithmic terms, e.g. the Jarvis–Patrick method [3] that has been extensively used in previous studies of chemical clustering and that is one of the methods considered later in this paper.

Biological taxonomy [4] provided the basis for the development of many of the clustering techniques that are in use

today, but these are now used in a wide range of application domains, with the current interest in data mining spurring the introduction of many new methods. Structural features provide an obvious source of attributes for chemical applications of clustering but early studies of the use of such features [5–8] were restricted to very small datasets. An extensive series of studies by Willett and co-workers in the early and mid-eighties (as reviewed in [9]) demonstrated the use of large-scale clustering for the selection of compounds for biological screening and for the processing of substructure search output, and highlighted the Jarvis–Patrick method as providing an appropriate combination of effectiveness and efficiency. Later work [10–12] demonstrated the greater effectiveness of Ward’s method, [13] and the availability of improved algorithms for this method [14] has allowed it to join Jarvis–Patrick for widespread use in chemical clustering applications.

The structural features that are normally used in chemical clustering are the fragment substructures encoded in a fingerprint to enhance the efficiency of 2D substructure searching. The similarity between two molecules is then computed as a function of the number of bits (and thus fragment substructures) that are common to the fingerprints representing those molecules. The Tanimoto coefficient is generally used to calculate such similarities but there are many other coefficients that can be used for this purpose. Fingerprint-based similarity

* Corresponding author. Tel.: +1-734-622-3015; fax: +1-734-622-2782.
E-mail address: john.raymond@pfizer.com (J.W. Raymond).

ties can be calculated extremely rapidly and have been found to perform reasonably well in practice, but there are many other ways in which one might seek to quantify the structural relationships between pairs of molecules [15]. One such approach uses a maximum common subgraph isomorphism algorithm to identify the largest substructure common to a pair of molecules, with the size of this maximum common substructure (MCS) being determined by some function of the numbers of constituent atoms and/or bonds. This provides a natural way of calculating the degree of similarity between a pair of molecules but the NP-complete nature of the maximum common subgraph isomorphism problem has ruled out the large-scale use of MCS-based similarities. We have recently described a new MCS algorithm, called RASCAL, that is sufficiently rapid in execution to permit graph-based similarity searching of large chemical databases [16,17] and that seems to provide a viable complement, or even an alternative, to existing, fingerprint-based approaches to virtual screening [18]. Given the close relationship that exists between similarity searching (where a single target molecule is matched against each of the molecules in a database) and clustering (where each molecule is matched against every other molecule in a database) this paper seeks to assess the suitability of graph-based similarity measures for chemical clustering and to compare their effectiveness with that of fingerprint-based measures. The natural starting point for such an evaluation is to take the current clustering methods of choice (i.e. Ward's method and the Jarvis–Patrick method, for which there is already a large body of practical experience) and to use them to process graph-based similarities, with the results from conventional fingerprint-based similarities providing a benchmark of comparison. However, we have taken the opportunity to consider several additional clustering methods, one of which has been designed specifically for use with graph-based measures of chemical similarity.

2. Graph- and fingerprint-based clustering

2.1. Terms and definitions

All graphs referred to in the following text are assumed to be simple, undirected graphs. For an introduction to graph-related concepts and notation, the reader is referred to a standard text on graph theory such as the recent book by Diestel [19]. A graph G consists of a set of vertices $V(G)$ and a set of edges $E(G)$ representing lines connecting all or some of the vertices in $V(G)$. A *subgraph* of G is a graph whose vertices and edges are subsets of G preserving the connectivity between the vertices and edges. A maximum common edge subgraph (MCES) is a subgraph common to two or more graphs consisting of the largest number of edges possible. Fig. 1 illustrates the MCES G_{12} between two molecular graphs G_1 and G_2 .

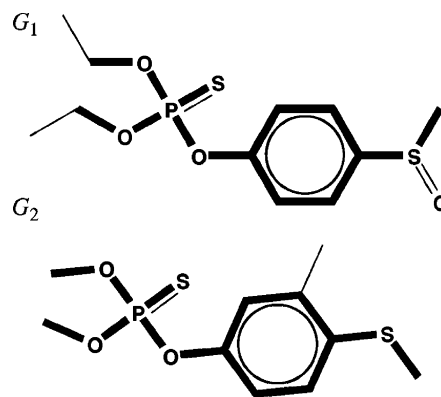


Fig. 1. Example of MCES.

In this paper, two different types of similarity measures are investigated, *feature-* and *cost-based* measures, these corresponding to the use of fingerprints and structure diagrams (i.e. 2D chemical graphs), respectively. In feature-based measures, a set of features or invariants is established from a structural description of a graph, and these features are then used in a vector representation to which various distance or similarity coefficients can be applied. Similarity coefficients obtained using the feature-based approach are functions of the relative number of bit positions that are set in each fingerprint (as reviewed by Willett et al. [15]). For instance, the well-known Tanimoto coefficient is given as $c/(a + b - c)$, where a and b are the number of features present in the two structures being compared and c the number of features in common between the two structures. In our experiments, the feature-based measures are calculated using Daylight fingerprints, which have been shown previously to be effective in chemical database studies [20].

In cost-based measures, the similarity between two compounds reflects the number of edit operations that are required in order to transform one structural graph into the other. Recently, an efficient cost-based method based on the MCES called RASCAL has been published [16,17]. RASCAL can be used with the same similarity coefficient formulae as are used with the feature-based methods [18], the difference being that the size of each graph is used to replace the number of features representing each structure. Therefore, the size of the MCES graph G_{12} replaces the number of features in common, and the sizes of the two molecular graphs being compared replace the number of bits set in each respective fingerprint. For example, the Tanimoto coefficient is given as $|G_{12}|/(|G_1| + |G_2| - |G_{12}|)$.

In its simplest form, the graph size is determined by treating atoms and bond pairs equally (i.e. $|G_{12}| = |V(G_{12})| + |E(G_{12})|$). However, it has been found that the RASCAL approach better approximates a chemical notion of similarity using

$$|V(G_{12})| + \beta(1 - \alpha(n(p, G_{12}) - 1))|E(G_{12})|$$

for $|G_{12}|$, and

$$|V(G_1)| + \beta |E(G_1)| \quad \text{and} \quad |V(G_2)| + \beta |E(G_2)|$$

for $|G_1|$ and $|G_2|$, respectively [18]. The function $n(p, G_{12})$ represents the number of unconnected subgraph components in the MCES (G_{12}) containing p or more edges: if all subgraphs have fewer than p edges, then $n(p, G_{12})$ will be assumed to be the total number of subgraph components. The constant β 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_{12} . Values of $p = 3$, $\alpha = 0.05$, and $\beta = 2.0$ were found to be effective in discerning chemical similarity, and are used in all of the experiments reported here. The current analysis also uses strict atom and bond typing so that only atoms and bonds of the same type can be matched, e.g. chlorine cannot match to fluorine. It may be possible to improve the results of a graph-based method by allowing some “fuzziness” in the compatibility between the various atom and bond types, but this prospect is not investigated here.

3. Clustering methods

Five different clustering methods were evaluated in this study. Ward [13], and Jarvis and Patrick [3] are well-known within the chemical information community and have previously proven effective for the clustering of chemical structures [9]. The other three have been selected from the clustering literature as being new to the clustering of chemical structure databases and hence appropriate for evaluation. Each method is described briefly below: the reader is referred to the original publications for details of the various procedures. In addition, the algorithm of Umesh [21] was also investigated, but in preliminary investigations it proved inferior to the others tested and was therefore omitted from further consideration.

3.1. Ben-Dor et al. (CAST)

The CAST method is based on an approximate clique-finding algorithm that avoids much of the costly enumeration necessary in traditional algorithms, and that uses a threshold parameter t to establish cliques of mutually similar objects. Adjustable parameter(s): t [22].

3.2. Jarvis–Patrick

This clustering method uses a table of k nearest neighbors for each object being clustered, and then sequentially merges clusters which have at least k_t nearest neighbors in common. Adjustable parameter(s): k and k_t [3].

3.3. Raymond–Willett

This method is based on a greedy algorithm that establishes clusters using a technique involving line graphs. It is a

fuzzy clustering procedure in that it allows for the possibility of overlapping clusters using three adjustable similarity thresholds. Adjustable parameter(s): S , S_a and S_b [23].

3.4. Ward

This is Ward’s well-known hierarchical clustering method, which establishes a hierarchy of clusterings whereby each level in the hierarchy represents a unique clustering. A representative clustering is typically selected using a rapidly calculable cluster validation index. The use of cluster validation indices in conjunction with Ward’s algorithm has been studied in detail, and it was found that Kelley’s validation index [24] was among the best of those tested [20]. In our studies, Ward’s method will be used in conjunction with the Kelley index. Adjustable parameter(s): *none* [13].

3.5. Yin–Chen

This approach is basically a two phase threshold method. It uses a built-in constant for thresholding as published, but we have found that converting this constant to an adjustable parameter y_t affords significantly greater flexibility (a value of y_t equal to 0.5 is equivalent to the originally published method). Adjustable parameter(s): y_t [25].

4. Parameter optimization

4.1. Methods

To evaluate the relative quality of the clusterings resulting from the various methods, we compare each calculated clustering with a reference clustering of the same data. In our experiments, we have used the seven datasets used in a previous evaluation of cluster validation indices [20] as well as two additional datasets created specifically for these trials. The characteristics of each dataset are summarized in Table 1. Each dataset was manually clustered in order to establish an ‘ideal’ clustering. This procedure is obviously subjective to a certain degree but, we believe, represents a reasonable partitioning of the structures. Four of the datasets were taken directly from the NCI anti-HIV database. Three are taken from the Pfizer corporate compound collection, and the two final datasets are a subset of the ID-alert database. These datasets represent various possible scenarios that may arise in a practical application.

NCI-A and NCI-B contain multiple distinct, but similar, structures, and NCI-C and NCI-D contain a more random assortment of structures. PD-X is a diverse set of compounds determined to be active in a high-throughput screening assay, and PD-Y and PD-Z are combinatorially synthesized compounds derived from a single scaffold for a single project. The clusters contained in ID-1 and ID-2 are diverse in that some of the clusters contain closely related compounds while others contain more loosely related compounds. Some

Table 1
Dataset characteristics

| Dataset | N | C ₁ | S | MW | R | %L | Source | Comments |
|---------|-----|----------------|-----|-----|------|-----|-------------------------------|--|
| NCI-A | 55 | 7 | 4 | 306 | 2.5 | 98 | NCI anti-HIV database | Distinct but similar clusters |
| NCI-B | 79 | 5 | 2 | 424 | 10.1 | 78 | NCI anti-HIV database | Distinct but similar clusters |
| NCI-C | 564 | 45 | 198 | 439 | 6.9 | 77 | NCI anti-HIV database | More subjective clusters than for NCI-A and NCI-B |
| NCI-D | 194 | 21 | 73 | 421 | 7.0 | 78 | NCI anti-HIV database | Random subset of NCI-C |
| PD-X | 305 | 29 | 43 | 361 | 5.3 | 92 | Pfizer compound library | Diverse set found active in HTS |
| PD-Y | 345 | 13 | 7 | 389 | 6.0 | 96 | Pfizer combinatorial compound | Derived from a single scaffold |
| PD-Z | 538 | 68 | 18 | 441 | 10.5 | 97 | Pfizer combinatorial compound | Derived from a single scaffold |
| ID-1 | 358 | 68 | 41 | 373 | 6.5 | 100 | ID-alert database | Collection of closely related and loosely related clusters |
| ID-2 | 262 | 43 | 24 | 367 | 5.4 | 100 | ID-alert database | Collection of closely related and loosely related clusters |

|N|: number of compounds, |C₁|: number of clusters as determined by manual review, |S|: number of singletons, |MW|: average molecular weight, |R|: average number of rotatable bonds, %L: percent of compounds satisfying the Lipinski Rule of Five.

clusters are structurally related to other clusters, and other clusters are distinct from all other clusters. Each dataset was evaluated for self-similarity by calculating the average nearest neighbor (ANN), average farthest neighbor (AFN), and the overall average similarity for all neighbors (AAN). The results are listed in Table 2. As previously mentioned, the combinatorial sets display a marked degree of self-similarity.

In this paper, we use two separate methods to evaluate the clusterings resulting from each method by comparing them with a reference clustering. The first comparison measure is the well-known Jaccard formula [26] given as

$$J(C_1, C_2) = \frac{c}{a + b - c}$$

where c is the number of pairs of structures that share a common cluster in both respective clusterings (C_1 and C_2), a the number of pairs of structures that share a common cluster in the first clustering C_1 , and b the number of pairs of structures that share a common cluster in the second clustering C_2 . The Jaccard measure ranges from 0 to 1, where 0 indicates a perfect mismatch and 1 indicates a perfect match. In our studies, C_1 will indicate the reference clustering for a particular dataset, and C_2 will represent the calculated clustering resulting from each method.

Table 2
Dataset similarities

| Data set | Self-similarity | | |
|----------|-----------------|-------|-------|
| | ANN | AFN | AAN |
| NCI-A | 0.818 | 0.077 | 0.242 |
| NCI-B | 0.916 | 0.133 | 0.401 |
| NCI-C | 0.693 | 0.027 | 0.171 |
| NCI-D | 0.616 | 0.055 | 0.169 |
| PD-X | 0.739 | 0.082 | 0.197 |
| PD-Y | 0.940 | 0.425 | 0.623 |
| PD-Z | 0.922 | 0.223 | 0.439 |
| ID-1 | 0.708 | 0.059 | 0.194 |
| ID-2 | 0.714 | 0.076 | 0.201 |

ANN: average nearest neighbor, AFN: average farthest neighbor, AAN: average of all neighbors.

The second comparison measure is based on the distance between two clusterings using an assignment procedure, where the distance can be regarded as the number of misclassified structures when a calculated clustering is compared to the reference clustering. Gusfield [27] has proposed a method whereby the distance between two clusterings C_1 and C_2 is calculated using $D(C_1, C_2) = |N| - A(C_1, C_2)$, where N is the set of structures in the reference clustering C_1 and $A(C_1, C_2)$ is the value of the assignment of the clusters from clustering C_1 to C_2 .

The value of $A(C_1, C_2)$ is calculated by first constructing an assignment matrix where each row i corresponds to a unique cluster in C_1 and each column j corresponds to a unique cluster in C_2 . The value of each element (i, j) in the assignment matrix consists of the number of structures that cluster i and cluster j have in common. The value of $A(C_1, C_2)$ then corresponds to the value of the linear assignment of the assignment matrix. A linear assignment is a subset of elements (i, j) in the assignment matrix whose sum is the maximum possible subject to the constraint that no two selected elements can be located in the same row or the same column in the matrix. Efficient algorithms exist for this procedure [28,29].

5. Results

With the exception of Ward's method, all of the clustering methods considered in this study involve the use of adjustable parameters. This presents a problem for the general application of these methods since, in order for a particular clustering method to be useful to the general practitioner, the user must have some idea of what parameter values to use with each method for a given problem. This raises two questions. What is a good 'rule of thumb' value to use for each adjustable parameter for a given problem? Are these values consistent from one problem to another? To be an effective general purpose method, it must be possible to determine a representative value for each parameter for a given clustering method, and these representative values must be consistent across similar problems. The less variable a

clustering method's parameters are, the less difficult it becomes for non-expert users of the method to apply it in practice.

To determine the most appropriate values for each method's adjustable parameters, we have run several optimization experiments. These used the scatter search optimization procedure [30,31], with the objective functions to be minimized being $-J(C_1, C_2)$ and $D(C_1, C_2)$, respectively, where C_1 is the manual reference clustering and C_2 is the calculated clustering.

5.1. Fingerprint-based clustering

The optimization procedure was performed for CAST, Jarvis–Patrick, and the Yin–Chen methods for all nine datasets using Daylight fingerprints and the Tanimoto coefficient. The Raymond–Willett algorithm has not been included in this analysis because it proved to be ineffective for use with fingerprint-based similarity coefficients. This is hardly surprising as it has been designed specifically for the processing of graph-based similarity measures. In addition, Ward's (using the Kelley level selection index) algorithm was included to serve as the benchmark method due to its success in previous analyses [20]. The results of the optimization study are presented in Table 3, which lists the optimal value for each adjustable parameter as well as the corresponding Jaccard and Gusfield score for the resulting calculated clustering.

Table 3 shows that the CAST, Jarvis–Patrick, and Yin–Chen algorithms all performed significantly better than the benchmark Ward/Kelley method. The resultant objective function values for the Jaccard and Gusfield measures for the optimal clusters were also relatively consistent between algorithms. Further inspection of the data, however, reveals that the adjustable parameter values corresponding to the optimal clusterings for Jarvis–Patrick display considerable variability with respect to the various datasets, as well as between the Jaccard and Gusfield objective functions. This indicates that while the Jarvis–Patrick method is capable of producing high quality clusterings, it does not appear that there exists a predictable range for each adjustable parameter that would provide an optimal or near-optimal clustering with any degree of confidence. The reason for this is that the other tested methods operate on the values of the pair-wise similarity coefficients directly, whereas Jarvis–Patrick operates on the ordered list of nearest-neighbor rankings. If the relative sizes of the clusters present in the data vary widely, then a particular nearest-neighbor ranking cut-off that works well for a cluster of particular size may not work well for another cluster of a significantly different size. For this reason, Jarvis–Patrick will tend to perform best when the clusters are approximately the same size.

This contrasts with the CAST and Yin–Chen methods, where we consider the datasets containing diverse sets of compounds separately from the combinatorially generated compounds possessing a common scaffold. The CAST al-

gorithm demonstrates relatively narrow ranges of values for which its adjustable parameter (t) provides optimal or near-optimal clusterings. For instance, the value of t ranges from 0.713 to 0.766 with an average of 0.740 for the combinatorial sets (PD-Y and PD-Z) and from 0.285 to 0.486 with an average of 0.384 for the diverse datasets. Like the CAST method, the Yin–Chen method also exhibits a substantial degree of consistency between the optimal clusterings resulting from the Jaccard and Gusfield objective functions with the exception of data set NCI-A. It is not obvious from the similarity statistics listed in Table 2 why adjustable parameter values associated with the data set NCI-A were notably different from those resulting from the other diverse data sets; however, it is conjectured that differences in adjustable parameter consistency between the CAST and Yin–Chen algorithms for this data set are due to the distribution of similarity values and the way each algorithm considers them. The CAST algorithm essentially constructs clusters based on the average intra-cluster similarity values; whereas, the Yin–Chen algorithm assembles clusters by using an inter-cluster similarity threshold.

While it is not possible to claim that one clustering method is the 'best', especially when evaluated on a limited number of datasets, it appears that the CAST and possibly the Yin–Chen method can be used by non-experts with a reasonable degree of confidence that the resulting clusters will represent a reasonable facsimile of a chemist's notion of a chemical series. Based on the data presented in Table 3, it is suggested that a 'rule of thumb' value for CAST's adjustable parameter t when used in conjunction with Daylight fingerprints and the Tanimoto coefficient is approximately 0.38 for diverse sets of compounds and 0.74 for combinatorial sets possessing a common scaffold. The value of 0.38 for diverse sets of compounds is an interesting discovery considering that the threshold parameter t for the CAST algorithm is simply an average similarity threshold. CAST iteratively increases the size of a cluster by adding a compound to an existing cluster if the average similarity between the compound and all other compounds in the cluster is greater than t . The value of this cluster similarity threshold value is in marked contrast to those established for similarity searching [18]. This is due primarily to the fact that clustering uses all pair-wise similarities between objects in a cluster, which tends to mitigate the presence of inappropriate pair-wise similarity values; whereas, similarity searching only considers the pair-wise similarity values between the query compound and the database of compounds, ignoring the potentially compensating information contained in the similarities between all of the compounds in the database.

5.2. Graph-based clusterings

The optimization procedure described above was then applied to the RASCAL-derived, graph-based similarities, as detailed in Table 4. The RASCAL similarity measure requires the use of a minimum similarity index threshold,

Table 3
Optimal Daylight fingerprint clusterings

| | | PD-X | | PD-Y | | PD-Z | | NCI-A | | NCI-B | | NCI-C | | NCI-D | | ID-1 | | ID-2 | |
|----------------|-------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|
| Ward/Kelley | | | | | | | | | | | | | | | | | | | |
| Jaccard | | 17.4 | | 22.8 | | 38.6 | | 66.4 | | 30.9 | | 12.8 | | 10.9 | | 4.6 | | 7.2 | |
| Gusfield | | 200 | | 212 | | 246 | | 14 | | 41 | | 329 | | 115 | | 256 | | 175 | |
| CAST | | | | | | | | | | | | | | | | | | | |
| Jaccard | t | 0.344 | 53.0 | 0.715 | 48.5 | 0.765 | 54.7 | 0.355 | 96.0 | 0.379 | 100 | 0.407 | 81.6 | 0.353 | 77.1 | 0.486 | 28.7 | 0.410 | 43.9 |
| Gusfield | | 0.285 | 103 | 0.713 | 104 | 0.766 | 147 | 0.355 | 5 | 0.379 | 2 | 0.347 | 226 | 0.331 | 77 | 0.472 | 148 | 0.473 | 85 |
| Yin–Chen | | | | | | | | | | | | | | | | | | | |
| Jaccard | y_t | 0.617 | 47.0 | 0.902 | 47.1 | 0.946 | 53.6 | 0.373 | 96.1 | 0.633 | 100 | 0.692 | 84.8 | 0.683 | 86.7 | 0.799 | 20.4 | 0.720 | 58.3 |
| Gusfield | | 0.609 | 121 | 0.945 | 140 | 0.937 | 188 | 0.375 | 5 | 0.639 | 2 | 0.633 | 222 | 0.659 | 79 | 0.774 | 186 | 0.720 | 78 |
| Jarvis–Patrick | | | | | | | | | | | | | | | | | | | |
| Jaccard | k | 19 | 46.3 | 11 | 46.0 | 10 | 61.3 | 11 | 91.2 | 37 | 88.4 | 30 | 72.2 | 13 | 71.9 | 10 | 26.9 | 32 | 41.6 |
| Jaccard | k_t | 12 | 46.3 | 2 | 46.0 | 4 | 61.3 | 6 | 91.2 | 31 | 88.4 | 20 | 72.2 | 8 | 71.9 | 5 | 26.9 | 22 | 41.6 |
| Gusfield | k | 26 | 129 | 37 | 132 | 17 | 172 | 15 | 5 | 35 | 10 | 28 | 201 | 8 | 89 | 7 | 167 | 14 | 124 |
| Gusfield | k_t | 16 | 129 | 23 | 132 | 9 | 172 | 11 | 5 | 25 | 10 | 13 | 201 | 2 | 89 | 3 | 167 | 9 | 124 |

The objective function scores are highlighted in italics next to the corresponding adjustable parameter value. Jaccard coefficients have been multiplied by 100 so that they range from 0 to 100.

Table 4
Optimal RASCAL clusterings

| | | PD-X | | PD-Y | | PD-Z | | NCI-A | | NCI-B | | NCI-C | | NCI-D | | ID-1 | | ID-2 | |
|-----------------|-------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|
| CAST | | | | | | | | | | | | | | | | | | | |
| Jaccard | t | 0.547 | 53.7 | 0.752 | 37.0 | 0.767 | 32.5 | 0.571 | 86.6 | 0.511 | 99.5 | 0.518 | 50.1 | 0.513 | 60.2 | 0.584 | 58.2 | 0.537 | 59.6 |
| Gusfield | | 0.548 | 112 | 0.752 | 156 | 0.803 | 246 | 0.571 | 9 | 0.511 | 4 | 0.516 | 282 | 0.514 | 95 | 0.578 | 85 | 0.540 | 67 |
| Yin–Chen | | | | | | | | | | | | | | | | | | | |
| Jaccard | y_t | 0.879 | 47.7 | 0.937 | 31.6 | 0.966 | 23.7 | 0.633 | 78.7 | 0.633 | 99.5 | 0.828 | 51.0 | 0.826 | 52.7 | 0.874 | 52.9 | 0.871 | 84.4 |
| Gusfield | | 0.879 | 138 | 0.935 | 206 | 0.941 | 307 | 0.633 | 11 | 0.633 | 4 | 0.810 | 308 | 0.835 | 104 | 0.875 | 110 | 0.866 | 59 |
| Jarvis–Patrick | | | | | | | | | | | | | | | | | | | |
| Jaccard | k | 14 | 50.8 | 15 | 53.1 | 14 | 39.1 | 14 | 94.0 | 38 | 79.2 | 33 | 48.7 | 15 | 50.0 | 28 | 42.3 | 22 | 42.3 |
| Jaccard | k_t | 5 | 50.8 | 6 | 53.1 | 7 | 39.1 | 3 | 94.0 | 7 | 79.2 | 9 | 48.7 | 4 | 50.0 | 6 | 42.3 | 6 | 42.3 |
| Gusfield | k | 15 | 158 | 15 | 136 | 16 | 250 | 14 | 12 | 38 | 18 | 9 | 359 | 7 | 130 | 25 | 170 | 18 | 139 |
| Gusfield | k_t | 5 | 158 | 6 | 136 | 8 | 250 | 3 | 12 | 7 | 18 | 2 | 359 | 2 | 130 | 4 | 170 | 2 | 139 |
| Raymond–Willett | | | | | | | | | | | | | | | | | | | |
| Jaccard | S_a | 0.909 | 52.5 | 0.925 | 39.1 | 0.906 | 24.9 | 0.902 | 86.6 | 0.870 | 81.1 | 0.910 | 45.7 | 0.965 | 57.1 | 0.881 | 63.6 | 0.859 | 66.6 |
| Jaccard | S_b | 0.555 | 52.5 | 0.620 | 39.1 | 0.598 | 24.9 | 0.533 | 86.6 | 0.540 | 81.1 | 0.528 | 45.7 | 0.491 | 57.1 | 0.610 | 63.6 | 0.479 | 66.6 |
| Gusfield | S_a | 0.909 | 120 | 0.967 | 143 | 0.920 | 224 | 0.902 | 9 | 0.870 | 10 | 0.904 | 305 | 0.960 | 102 | 0.833 | 69 | 0.896 | 60 |
| Gusfield | S_b | 0.430 | 120 | 0.455 | 143 | 0.496 | 224 | 0.533 | 9 | 0.540 | 10 | 0.638 | 305 | 0.489 | 102 | 0.597 | 69 | 0.482 | 60 |

The objective function scores are highlighted in italics next to the corresponding adjustable parameter value. Jaccard coefficients have been multiplied by 100 so that they range from 0 to 100.

MSI, for which a value of 0.6 (for the Wallis coefficient, which is the graph form of the Tanimoto coefficient [18]) was used for experiments involving the CAST, Jarvis–Patrick, and Yin–Chen methods. The value of 0.6 was found to be low enough so as not to affect the results of these methods. In contrast, the Raymond–Willett method is dependent upon the selected value of MSI, where it is used in lieu of the adjustable parameter S . It was found that an MSI value of 0.7 and 0.85 worked well for the diverse sets and combinatorial sets (PD-Y and PD-Z), respectively. As a note, since the use of the MSI threshold in RASCAL omits pair-wise similarities less than the MSI threshold from further consideration, it was assumed for the purposes of the CAST method that any missing similarity values had a value of 0.5. The value of 0.5 was assumed based on the observation that the RASCAL similarity values for a sufficiently large data set tend to form a Gaussian distribution from 0 to 1 with the value of 0.5 being the most populated. A more sophisticated approach could be used where a value is assumed based on the observed pair-wise similarity distribution of the data set being considered. The use of an MSI enables very substantial reductions in computation [17] but is a limitation when used in a clustering context especially when, as here, many different types of data sets need to be processed.

The results of the experiments with the RASCAL similarities show many of the same patterns observed with the fingerprint similarities. Although the Jarvis–Patrick method demonstrates more consistency with respect to the adjustable parameters when used in conjunction with RASCAL rather than fingerprints, it still exhibits more variability than is desirable for a general purpose procedure. The CAST, Raymond–Willett, and Yin–Chen methods demonstrate similar objective function results for each of the datasets; however, the CAST and Yin–Chen methods appear to show the most consistency in their adjustable parameters. The CAST method parameter ranges from 0.752 to 0.803 with an average and standard deviation of 0.769 and 0.024, respectively, for the combinatorial sets (PD-Y and PD-Z) and from 0.511 to 0.584 with an average and standard deviation of 0.540 and 0.027, respectively, for the diverse datasets (all others). The Yin–Chen parameter ranges from 0.935 to 0.966 with an average and standard deviation of 0.945 and 0.014, respectively, for the combinatorial sets and from 0.633 to 0.879 with an average and standard deviation of 0.791 and 0.106, respectively, for the diverse datasets.

In comparison with the graph-based clusterings, the fingerprints scored consistently higher on the combinatorial sets possessing a common scaffold (PD-Y and PD-Z) as well as two of the NCI datasets (NCI-C and NCI-D). It is interesting to note that these sets possess the most subjective clusterings as the differences between the clusters in the combinatorial sets are subtle and the clusters in NCI-C and NCI-D tend to be more loosely related (not necessarily direct structural analogues). For the ID-alert datasets (ID-1 and ID-2), the graph-based clusterings outperformed the fingerprint-based

clusterings: these clusters tended to constitute structural analogues with the characteristics that some clusters are distinct from all other clusters, but there are also clusters that are structurally related to other clusters in the dataset.

6. Manual inspection of clusterings

6.1. Methods

Although the quantitative comparisons described above give an objective assessment of behavior in somewhat contrived situations, we wanted to compare the methods in a situation more closely related to a practical task. One obvious application of such methods is to generate groupings that might be designated as “series” in a medicinal chemist’s perception. This corresponds to a common task in the conduct of HTS protocols where typically an initial large and diverse set of primary hits must be organized for analysis. To simulate this situation, a customized collection of 1325 diverse drugs and drug candidates covering a broad spectrum of therapeutic classes and chemical types was used as a dataset. Each method was applied using the optimized parameter settings determined above to partition the dataset. Several known classes were then examined with the following questions in mind: how effectively were the compounds grouped; were there situations where one method was superior to another; and was there any evidence for complementarities between methods in difficult cases? Since the performance of the graph-based methods was of special interest in this work, groups were chosen for examination where an MCES-based approach might be expected to perform particularly well; in addition, cases were sought where differences might be expected from a fingerprint method.

It bears repeating at the outset of this part of the discussion that, although the group selections were driven by the commonality of their biological effects, the methods being used here for partitioning use only topological chemical information. Thus, it is the effectiveness of grouping by chemical class that is most central; if biological commonality is also observed then that is a fortuitous, but not critical, factor in the determination of effectiveness. The following classes were chosen for illustration: tetracycline antibiotics, angiotensin antagonist antihypertensives, calcium antagonist dihydropyridines, antifungal agents, β -lactam antibiotics, ace inhibitors and opiate analgesics. For each of the groups, the cluster membership and frequency are given for Ward/Kelley, Jarvis–Patrick/6:10, and CAST using Daylight fingerprints, and CAST using RASCAL similarities with the parameter settings derived from the first part of the study. What is of interest here for a determination of effectiveness is the number of clusters required to include all members of the series (and number of singleton members) and the degree to which each series is cleanly discriminated from other compounds in the dataset; i.e. the number of “non-series” compounds in class clusters. Further, could subsets be perceived

Table 5
Cluster inspection results

| Class | N | ANN | Number of clusters/singletons | | | | Number in largest cluster | | | | Purity of largest cluster (%) | | | | Comments |
|---------------------------|----|-------|-------------------------------|------------------|-------------------|-----------------|---------------------------|------------------|-------------------|-----------------|-------------------------------|------------------|-------------------|-----------------|--|
| | | | Ward/ Kelley | JarPat (6:10) | CAST/ Daylight | CAST/ RASCAL | Ward/ Kelley | JarPat (6:10) | CAST/ Daylight | CAST/ RASCAL | Ward/ Kelley | JarPat (6:10) | CAST/ Daylight | CAST/ RASCAL | |
| Tetracyclines | 8 | 0.950 | 1/0 | 1/0 | 1/0 | 1/0 | 8 | 8 | 8 | 8 | 100 | 100 | 100 | 100 | – |
| Angiotensin-2 antagonists | 5 | 0.553 | 0/5 | 1/2 | 2/0 | 1/0 | 1 | 3 | 3 | 5 | – | 43 | 75 | 100 | – |
| Dihydropyridines | 13 | 0.777 | 1/1 | 1/1 | 1/1 | 1/2 | 12 | 12 | 12 | 11 | 100 | 100 | 100 | 100 | One compound is singleton in all methods |
| Antifungals | 11 | 0.690 | 2/4 | 2/0 | 3/0 | 4/1 | – | – | – | – | – | – | – | – | All methods include a nitroimidazole antitumor agent |
| Nidazoles | 4 | 0.610 | 1/0 | 1/0 | 1/0 | 1/1 | 4 | 4 | 4 | 3 | 1 | 80 | 80 | 75 | – |
| Conazoles | 7 | 0.735 | 1/4 | 1/0 | 2/0 | 3/0 | 3 | 7 | 5 | 3 | 100 | 100 | 100 | 100 | – |
| β-Lactams | 50 | 0.842 | 5/12 | 6/3 | 3/3 | 4/4 | 16 | 22 | 45 | 38* | 100 | 100 | 90 | 76 | *One compound from misc class included (loracarbef) |
| Cephalosporin | 22 | 0.874 | 3/4 | 1/0 | 1/0 | 3/0 | 12 | 22 | 45 | 17 | 100 | 100 | 49 | 45 | Smaller clusters are also pure in all methods |
| Penicillin | 23 | 0.895 | 2/3 | 3/0 | 1/0 | 2/1 | 16 | 17 | 45 | 20 | 100 | 100 | 49 | 53 | – |
| Miscellaneous | 5 | 0.453 | 0/5 | 2/3 | 2/3 | 2/3 | 1 | 1 | 2 | 1 | – | 13 | 50 | 33 | Not grouped together or with other β-lactam clusters |
| ACE inhibitors | 17 | 0.797 | 2/5 | 2/1 | 4/2 | 3/0 | 8 | 12 | 8 | 13 | 1 | 86 | 26 | 50 | Largest clusters contain morphine analogs; small clusters are <50% pure for CAST methods |
| Opiates | 37 | 0.856 | 4/5 | 9/5 | 5/2 | 4/2 | 11/10/8 | 10/8 | 23 | 24 | 2/100/100 | 100/100 | 72 | 57 | |

Table 6
List of compounds

| | | |
|----------------------------------|---------------------------|----------------------|
| <i>Angiotensin-2 antagonists</i> | Cefazolin | <i>Antifungals</i> |
| Candesartan | Cefdinir | Benznidazole |
| Irbesartan | Cefixime | Metronidazole |
| Losartan | Cefmetazole | Misonidazole |
| Valsartan | Cefoperazone | Tinidazole |
| Proprietary compound | Cefotaxime | Econazole |
| <i>Dihydropyridines</i> | Cefoxitin | Fluconazole |
| Amlodipine | Cefpodoxime | Itraconazole |
| Felodipine | Cefprozil | Ketoconazole |
| Isradipine | Ceftriaxone | Miconazole |
| Lacidipine | Cefuroxime | Sertaconazole |
| Nicardipine | Cefuroxime .axetil | Voriconazole |
| Nifedipine | Cephacetrile | |
| Niguldipine | Cephalexin | <i>Opiates</i> |
| Nilvadipine | Cephaloglycin | Acetylnormethadol |
| Nimodipine | Cephalothin | Alfentanil |
| Nisoldipine | Cephapirin | Buprenorphine |
| Nitrendipine | Cephadrine | Butorphanol |
| Oxodipine | Amdinocillin | Butylmorphine |
| Proprietary compound | Amoxicillin | Codeine |
| <i>ACE inhibitors</i> | Ampicillin | Dextromethorphan |
| Benazepril | Azidocillin | Dezocine |
| Candoxatril | Carbenicillin | Dihydrocodeine |
| Captopril | Carbenicillin.indanyl | Ethylmorphine |
| Cilazapril | Carbenicillin.phenyl | Etorphine |
| Enalapril | Cloxacillin | Fentanyl |
| Enalaprilat | Cyclacillin | Heroin |
| Fosinopril | Dicloxacillin | Hydrocodone |
| Indolapril | Flucloxacillin | Hydromorphone |
| Lisinopril | Hetacillin | Ketobemidone |
| Moexipril | Methicillin | Levallorphan |
| Moexiprilat | Nafcillin | Meperidine |
| Perindopril | Oxacillin | Meptazinol |
| Quinapril | Penicillin.G | Methadone |
| Quinaprilat | Penicillin.V | Methadyl.acetate |
| Ramipril | Piperacillin | Morphine |
| Trandolapril | Piridicillin | Nalbuphine |
| Zofenoprilat | Pivampicillin | Nalmefene |
| | Sulbenicillin | Nalorphine |
| | Ticarcillin | Naloxone |
| | Proprietary compound | Naltrexone |
| <i>β-Lactams</i> | | Oxycodone |
| Aztreonam | <i>Tetracyclines</i> | Pentazocine |
| Clavulanic acid | Chlortetracycline | Pholcodine |
| Imipenem | Demethylchlortetracycline | Prolididine |
| Loracarbef | Doxycycline | Propiram |
| Moxalactam | Methacycline | Propoxyphene |
| Cefaclor | Minocycline | Sufentanil |
| Cefadroxil | Oxytetracycline | Tilidine |
| Cefamandole | Rolitetracycline | Proprietary compound |
| Cefatrizine | Tetracycline | |

or were “extraneous” compounds “interesting” in any sense? A summary of the observations on these series is given in Table 5 for the compounds listed in Table 6. For each class or major subclass examined, the number of members and the ANN similarity (Tanimoto using Daylight fingerprints) as an indication of the diversity of the collection are recorded. This is followed, for each method, by the number of clusters and singletons, the size of the largest single cluster and

the purity of that cluster. An ideal result would be a single pure cluster with no singletons for each grouping of interest.

7. Results

The first three classes have large and common ring templates. They show a decreasing level of internal similarity as measured by their mean nearest neighbor similarities. The tetracyclines, with their unique and characteristic template are efficiently grouped and discriminated by all methods. The smaller and somewhat more diverse dihydropyridines are also effectively grouped by all methods except for one analog which is clearly a substantial structural variant from the rest and is a singleton in all methods. The RASCAL method also fails to include one additional analog which is grouped appropriately by all the other methods. This compound, nilvadipine, differs from all the others by having one of the ring methyl groups replaced by a cyanide group. The common substructure method might have been expected to be the most sensitive to this minor structural change. The angiotensin-2 antagonists show a very low internal similarity by fingerprint methods: clustering based on fingerprints might then be expected to find these too far apart to group, in spite of the presence of the biphenyl tetrazole as a large common substructure. Indeed, the Ward’s method fails to group these compounds at all, while the Jarvis–Patrick and CAST/Daylight methods do find a subset of three compounds to group. These compounds, however, are grouped with other non-class compounds as shown by the low purity of the clusters. The commonality that is keyed upon appears to be a smaller fragment related to the benzyl imidazole moiety rather than the “pharmacophoric” biphenyl tetrazole. This is deduced by examining the non-class compounds retrieved. The RASCAL method does what is expected and groups all five appropriately.

The next two groups represent therapeutic groupings which each contain two clearly recognizable structural subclasses. In the case of the antifungals, the two classes have very little in common structurally. The conazoles, all of which are characterized by an elaborated phenethyl imidazole or triazole moiety are grouped cleanly by all methods. However, the Jarvis–Patrick method is the only one to get all of them into one cluster. Ward’s method fails to include four which appear as singletons, while the two CAST methods break this group into two or three clusters, respectively, with RASCAL generating the most partitions. The four compounds of the nitroimidazole subclass are grouped into a single cluster by all methods except RASCAL which misses one as a singleton. In this group, there is one extraneous compound included by all methods. Upon examination, this turned out to be an antitumor candidate which also contained the nitroimidazole moiety and which was otherwise quite similar to the compounds in this subset. The Jarvis–Patrick and CAST/Daylight methods also put

these same five compounds cleanly into their own cluster, while Ward's method failed to discriminate them from a large number of other small compounds. The extra compound included is correct from a chemical viewpoint, if not from a biological one.

The β -lactam antibiotics are slightly more complicated in that, in addition to the well recognized subclasses of the cephalosporins and penicillins, which themselves have a high degree of internal similarity, there are five additional β -lactams more distantly related structurally. The internal similarity, except for the miscellaneous class, is much higher than for the antifungal subclasses. Here, there is a clear difference between the CAST methods and the Ward or Jarvis–Patrick methods: the former tend to group the two subclasses together, with CAST/Daylight being more efficient (one clean cluster as against three or two for RASCAL on cephalosporins or penicillins, respectively). Ward gives a high number of singletons for both subclasses as well as multiple clusters; Jarvis–Patrick gives one clean cluster for all cephalosporins and three for the penicillin analogs. Neither of these latter methods mixes penicillins and cephalosporins at the default settings selected. The miscellaneous class compounds are either singletons or members of larger, undifferentiated clusters in all methods. Surprisingly, RASCAL generated more subclusters than expected. This is basically “subsetting” of the sort more commonly observed in Ward's method (as seen here also). However, not all of these smaller clusters contain the same compounds across methods.

Peptide-like drugs represent a challenge for automated series organization methods. We examined the behavior of these methods with the set of 17 angiotensin converting enzyme (ACE) inhibitors present in the collection. Interestingly, Jarvis–Patrick gave the best result. It grouped 12 of the 17 into one cluster with only two other non-class compounds. A subset of four, also segregated by all of the other methods, was grouped into a second class which, however, was not very pure (22%). Captopril was a singleton. The small class of four was cleanly found by Ward, but not by CAST/Daylight (38%). RASCAL groups a different subset of 13 together, but at low purity (50%). The non-class compounds in each case are primarily non-peptide drugs, not compounds from other peptide classes. The operational commonality keyed on appears to be related to the region of the ACE inhibitors including the phenylalanine-derived moiety.

As a final example, the structurally complex class of opiate drugs was examined. This collection does not include any of the peptide opiates, and the one kappa compound was excluded from the comparison leaving 37 compounds. Not unexpectedly, all methods split this collection into several clusters. Both Ward and Jarvis–Patrick isolated clusters of 10 and 8 compounds that cleanly contained natural product analogs related to morphine. In addition, a smaller clean cluster of four compounds with related polycyclic structures was found. The two CAST methods gave larger single clusters (23 for CAST/Daylight and 24 for RASCAL), which grouped the compounds in all three of these clusters together,

but at the expense of including non-class compounds (purities of 72% for CAST/Daylight and 57% for RASCAL). Examination of the incorrect compounds surprisingly showed that several estrogenic compounds had been included. This led to the conclusion that the substructure keyed upon by the CAST methods isolated the phenol and alicyclic carbon ring systems but did not include the tertiary piperidine substructure characteristic of the opiate analgesics. The remaining compounds in this class are the simplified piperidine analogs of morphine. Small subsets of these are isolated by the fingerprint methods, but most fall into undifferentiated clusters in all methods.

8. Discussion

Several conclusions can be drawn from the earlier observations. When series are characterized by relatively large or unique ring templates, all methods tend to group their members effectively. There is a suggestion that the RASCAL method may do a better job if the internal fingerprint similarity of the collection to be grouped drops too low (angiotensin antagonists). The β -lactams and opiates illustrated an interesting difference between Ward and Jarvis–Patrick on the one hand and the two CAST methods on the other. The former methods achieved a finer but cleaner grouping of related structures of high complexity at the cost of generating a larger number of clusters, whereas the two CAST methods found regions of commonality that could consolidate these subgroups, but at the expense of purity or diminished coverage. This suggests that further examination of the appropriate option settings for the latter methods may be necessary to tune these for particular types of structures. We already know that this is the case for the better-studied Ward and Jarvis–Patrick methods, especially when applied in a single pass to datasets with high structural diversity. In particular, the Ward/Kelley method we have used typically chooses a level where there is one large cluster (30–50% of the dataset) in the first pass on datasets with the sort of diversity represented here. This accounts for the cases in Table 6 where there are very low purities for Ward's method and where recursive clustering of this large cluster is necessary to generate additional groupings for more structurally similar classes. Collections of small compounds with relatively simple structural commonality are still not easily differentiated by any of the methods. The Jarvis–Patrick method performed quite well across the board in this study, and the newly proposed CAST/Daylight method also did quite well.

The graph-based RASCAL method generally did not perform as well in partitioning this large and structurally diverse 1325-member dataset in the desired manner as the fingerprint-based methods. Examination of the compound groupings suggests that the substructures keyed upon by this method are more akin to queries than to “series templates” in many cases, resulting in decreased purity in the groupings. A better strategy for partitioning such datasets might

be to use a fingerprint method for the initial partitioning and then apply RASCAL to the clusters to extract a more appropriate MCES for that grouping. Finally, it must be pointed out that none of these methods can be expected in general to group compounds efficiently based on pharmacophoric patterns, even if the latter have a large topological content.

9. Conclusions

Most approaches to the clustering of 2D chemical databases structures have been based on similarity measures calculated using fingerprint representations of chemical structure. In this paper, we have discussed clusterings that are based on similarity measures calculated using graph-based representations. Specifically, we have applied our recent algorithm for the identification of maximum common edge subgraphs to the calculation of inter-molecular similarities based on the graph similarity coefficient of Wallis; these similarities have then been used for the implementation of several different clustering methods, with comparable experiments being carried out using a conventional, fingerprint- and Tanimoto-based similarity measure.

Two groups of experiments were carried out. The first involved an extensive series of simulations that were designed to identify the most appropriate parameter values for the various clustering methods that were studied (CAST, Jarvis–Patrick, Raymond–Willett, Yin–Chen and Ward), and the extent to which these values were dataset-dependent. These simulations suggested that the CAST and Yin–Chen methods appear to be the most robust of those tested. The second involved consideration of the bioactivity of several sets of compounds, focusing principally on the ability of the various approaches to highlight meaningful chemical series in datasets comparable to those resulting from HTS analyses. No obvious advantage appeared to result from the use of the more sophisticated, graph-based similarity measures when compared to conventional, fingerprint-based measures. We hence draw three principal conclusions from the study. First, while the results obtained from the use of graph-based similarities are different from fingerprint-based similarities, there is no evidence to suggest that one approach is consistently better than the other: each approach has its strengths and weaknesses, and it may be that an investigation should employ both approaches to obtain a fuller view of the structural relationships present within a dataset. Second, the CAST and Yin–Chen methods warrant further investigation as a potential alternative to the Ward and Jarvis–Patrick methods for the clustering of chemical structure databases as they have proven effective in the evaluations carried out here. Third, this study supports an earlier observation that there are significant differences in the behavior of a clustering algorithm with respect to the relative diversity of a data set [20]. It is suggested that further work is needed to assess this effect and that it may be possible to develop a formulaic correlation between optimal values for a clustering algorithm's

adjustable parameter(s) and some measure of the relative diversity of a data set.

Acknowledgements

We thank the following: Pfizer (Ann Arbor) for funding; Alain Calvet, Christine Humblet, Sherry Marcy, and David Wild for their support. We would also like to thank David Wild for the use of his program for performing Ward's clustering and Rafael Marti for making publicly available his code for the scatter search optimization algorithm (at <http://www.uv.es/~rmarti/sscode.html>). The Krebs Institute for Biomolecular Research is a designated centre of the Biotechnology and Biological Sciences Research Council.

References

- [1] B.S. Everitt, Cluster Analysis, 3rd ed., Arnold, London, 1993.
- [2] F.J. Rohlf, Single-link clustering algorithms, in: Handbook of Stat., 1982, pp. 267–284.
- [3] R.A. Jarvis, E.A. Patrick, Clustering using a similarity measure based on shared near neighbors, IEEE Transac. Comput. 22 (1973) 1025–1034.
- [4] P.H.A. Sneath, R.R. Sokal, Numerical Taxonomy, Freeman, San Francisco, CA, 1973.
- [5] P.H.A. Sneath, Relations between chemical structure and biological activity in peptides, J. Theor. Biol. 12 (1966) 157–195.
- [6] P.J. Harrison, A method of cluster analysis and some applications, Appl. Stat. 17 (1968) 226–236.
- [7] G.W. Adamson, J.A. Bush, A method for the automatic classification of chemical structures, Inf. Storage Ret. 9 (1973) 561–568.
- [8] G.W. Adamson, D. Bawden, Comparison of hierarchical cluster analysis techniques for the automatic classification of chemical structures, J. Chem. Inf. Comput. Sci. 21 (1981) 204–209.
- [9] P. Willett, Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, 1987.
- [10] G.M. Downs, P. Willett, W. Fisanick, Similarity searching and clustering of chemical-structure databases using molecular property data, J. Chem. Inf. Comput. Sci. 34 (1994) 1094–1102.
- [11] R.D. Brown, Y.C. Martin, Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection, J. Chem. Inf. Comput. Sci. 36 (1996) 572–584.
- [12] R.D. Brown, Y.C. Martin, The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding, J. Chem. Inf. Comput. Sci. 37 (1997) 1–9.
- [13] J.H. Ward, Hierarchical grouping to optimize an objective function, J. Am. Statist. Ass. 58 (1963) 236–244.
- [14] F. Murtagh, A survey of recent advances in hierarchical clustering algorithms, Comput. J. 26 (1983) 354–359.
- [15] P. Willett, J. Barnard, G. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
- [16] J. Raymond, E. Gardiner, 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.
- [17] J. Raymond, E. Gardiner, P. Willett, RASCAL: calculation of graph similarity using maximum common edge subgraphs, Comput. J. 45 (2002) 631–644.
- [18] J. Raymond, P. Willett, Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases, J. Comput.-Aided Mol. Des. 16 (2002) 59–71.
- [19] R. Diestel, Graph Theory, Springer, New York, 2000.

- [20] D.J. Wild, 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.
- [21] R.M. Umesh, A technique for cluster formation, *Patt. Recog.* 21 (1988) 393–400.
- [22] A. Ben-Dor, R. Shamir, Z. Yakhini, Clustering gene expression patterns, *J. Comput. Biol.* 6 (1999) 281–297.
- [23] J. Raymond, P. Willett, A line graph algorithm for clustering chemical structures based on common substructural cores. *MATCH*, in press.
- [24] L.A. Kelley, S.P. Gardner, M.J. Sutcliffe, An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally-related subfamilies, *Protein Eng.* 9 (1996) 1063–1065.
- [25] P.Y. Yin, L.H. Chen, A new non-iterative approach for clustering, *Patt. Recog. Lett.* 15 (1994) 125–133.
- [26] G.W. Milligan, A Monte Carlo study of 30 internal criterion measures for cluster analysis, *Psychometrika* 46 (1980) 187–199.
- [27] D. Gusfield, Partition-distance: a problem and class of perfect graphs arising in clustering, *Inf. Proc. Lett.* 82 (2002) 159–164.
- [28] G. Carpaneto, S. Martello, P. Toth, Algorithms and codes for the assignment problem, *Ann. Oper. Res.* 13 (1988) 193–223.
- [29] R. Jonker, A. Volgenant, Linear assignment procedures, *Eur. J. Oper. Res.* 116 (1999) 233–240.
- [30] F. Glover, M. Laguna, R. Marti, Scatter search, in: A. Ghosh, S. Tsutsui (Eds.), *Theory and Applications of Evolutionary Computation: Recent Trends*, Springer, Berlin, 2001.
- [31] M. Laguna, R. Marti, Experimental testing of advanced scatter search designs for global optimization of multimodal functions, Technical Report, TR11-2000, University of Valencia, Valencia, 2000.