ARTICLES

Development of a Spectral Clustering Method for the Analysis of Molecular Data Sets

Mark L. Brewer*

Computational Chemistry Group, Evotec (UK) Limited, 111 Milton Park,
Abingdon, Oxfordshire OX14 4RZ, United Kingdom

Received December 21, 2006

A spectral clustering method is presented and applied to two-dimensional molecular structures, where it has been found particularly useful in the analysis of screening data. The method provides a means to quantify (1) the degree of intermolecular similarity within a cluster and (2) the contribution that the features of a molecule make to a cluster. In an application of the spectral clustering method to an example data set of 125 COX-2 inhibitors, these two criteria were used to place the molecules into clusters of chemically related two-dimensional structures.

INTRODUCTION

There is considerable interest in the development and utilization of clustering methods for the analysis of sets of molecules.[1−16] Clustering methods can be applied to a set of molecules to arrange similar molecules into the same cluster and dissimilar molecules into different clusters. The clusters of molecules that result from this exercise can be used to provide a condensed representation of the original set which can be employed in a variety of ways. For example, in the analysis of biological screening data, structure−activity relationships can be derived by clustering the active molecules.

Applications of clustering methods to chemical data sets have been reviewed by Downs and Barnard[1] and also previously by Willett.[2] Early work, largely by Willett and co-workers, identified the Jarvis−Patrick clustering method for the study of chemical data sets,[2,3] while later work by Brown and Martin highlighted Ward's method for the separation of biologically active and inactive molecules.[4−6] More recently, the fuzzy k-means clustering method has been used for simulated property prediction and the identification of multicluster membership and outlier compounds.[7] In addition, cluster-based approaches have been developed for the analysis of high throughput screening results[8,9] and large compound collections.[10−14] Finally, Volkmann and co-workers have used clustering methods in conjunction with novel molecular descriptors such as biological activity spectra and drug-induced side effects in an effort to forecast activity in protein networks[15] and likely clinical effect profiles,[16] respectively.

At a practical level the procedure to cluster a molecular data set begins with the assignment of descriptors to each molecule in the set. There are a vast number of molecular descriptors available ranging from straightforward one-dimensional properties (e.g., molecular weight) through to more complex properties of the three-dimensional nuclear coordinates (e.g., the molecular electrostatic potential).[17] Next, a criterion is established to quantify the similarity between any pair of molecules in the data set based upon the associated pair of molecular descriptors. There are also a large number of ways to compute molecular similarity, the precise form of which will depend on the underlying choice of molecular descriptor.[18,19] Finally a clustering algorithm is selected to process the intermolecular similarities and to arrange the molecules into clusters. There are clearly a number of choices to be made when it comes to the selection of molecular descriptors, the similarity criterion, and clustering method, and the optimal combination will depend on the set of molecules under consideration as well as the objective of the analysis.[1−19]

Our interest in this area stemmed from a desire to develop an automated procedure for the classification of two-dimensional molecular structures to assess the number and representation of different molecular scaffolds within chemical data sets. The concept of a molecular scaffold is important in drug discovery, when it can sometimes be necessary to consider molecules that are based on alternative scaffolds to avoid undesirable physical properties and/or circumvent intellect property restrictions that may be associated with a particular chemical series, a process referred to as scaffold hopping.[20] In this article we report a spectral clustering method that has been useful in this regard, especially in the analysis of screening data, but which at the same time exhibits additional attractive features that may make it a generically useful technique for the analysis of sets of molecules. In particular, the method provides a means to quantify the degree of intermolecular similarity within a cluster and also the contribution that a molecule makes to a cluster.

To demonstrate the practical utility of the spectral clustering method it has been applied to a set of cyclooxygenase-2 (COX-2) inhibitors, chosen to be representative of the confirmed hits that might be obtained from a (highly

* Corresponding author e-mail: mark.brewer@evotec.com.

10.1021/ci600565r CCC: $37.00
Published on Web 07/18/2007

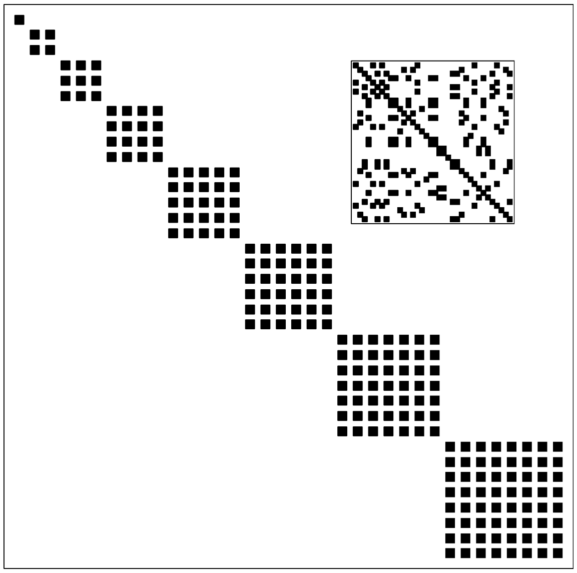

Figure 1. Schematic representation of the similarity matrix for an ideal data set of 36 molecules which can be partitioned into 8 clusters with populations 1, 2, . . ., 8. The black quares in the diagram represent pairs of molecules that are similar. The main panel displays the form of the matrix obtained with an order of molecules which reveals a block diagonal structure, where each block is associated with a particular cluster. The inset panel displays the same matrix but based upon a random order of molecules.

successful) biological screen against COX-2, a target which is associated with inflammatory disorders.[21]

SPECTRAL CLUSTERING

Consider a set of N molecules, for which an $N \times N$ similarity matrix $\mathbf{S}$ can be defined which is composed of matrix elements S_{ij} that are equal to the similarity between molecule i and molecule j.

Now, by way of example, consider an ideal set of molecules that can be divided into k distinct clusters of population p_i ($i = 1, . . ., k$). For simplicity, the similarity matrix for this ideal set $\mathbf{S}^I$ is in binary form so that molecules which belong to the same cluster are similar with $S_{ij}^I = 1$, and molecules that belong to different clusters are not similar with $S_{ij}^I = 0$. For most orders of the molecules in the ideal set, the form of $\mathbf{S}^I$ will correspond to a permuted block diagonal matrix with k distinct blocks that contain matrix elements of unit value. In principle, the ideal set could be clustered simply by rearranging the order of the molecules so that a block diagonal form of $\mathbf{S}^I$ is revealed; this approach is illustrated schematically in Figure 1.

One route to identify an order of the molecules that results in a block diagonal form of ideal similarity matrix $\mathbf{S}^I$ lies in an eigenvalue and eigenvector decomposition of the matrix. By definition, $\mathbf{S}^I$ will by have k nonzero eigenvalues (i.e., one for each block), the values of which will be equal to the cluster populations p_i ($i = 1, . . ., k$). Furthermore, each nonzero eigenvalue will have an associated eigenvector that has p_i nonzero elements (of magnitude $p_i^{-1/2}$), and the indices of these nonzero elements will correspond to those molecules that belong to the ith cluster. Hence, with the eigenvectors of the nonzero eigenvalues of $\mathbf{S}^I$ at hand, it would be possible to identify an order of the set with which to rearrange the matrix to block diagonal form and reveal the k distinct clusters in the ideal set.

Clearly, the ideal situation outlined above will not in general be applicable to realistic data sets, where elements of $\mathbf{S}$ will not be restricted to values of 0 or 1, and it may not be possible to arrange the similarity matrix into a purely block diagonal form. However, it does suggest that probing the eigenvalues and eigenvectors of $\mathbf{S}$ may be useful for the identification of molecular clusters. In fact, spectral (i.e., eigenvalue-based) clustering methods are routinely applied in the analysis of digital images, where eigenvalues and eigenvectors derived from affinity matrices (cf. similarity matrices in this work) are used for clustering.[22−24] The key step here is to recognize the similarity matrix $\mathbf{S}$ as the weighted adjacency matrix of a graph of N nodes (or molecules) with each pair of nodes connected by edges that are weighted by the similarity between the two nodal molecules. The theory of graph spectra can then be employed to show that the eigenvectors of $\mathbf{S}$ constitute a set of weights which can be used to cluster similar nodes.[24] We now continue to describe our implementation of a spectral clustering method noting that a more formal justification of these approaches has been presented elsewhere in the literature.[22−24]

The starting point for the spectral clustering approach is to diagonalize a similarity matrix $\mathbf{S}$ so that

$$\mathbf{S} = \mathbf{c}\lambda\mathbf{c}^T \qquad (1)$$

where λ is the diagonal matrix of eigenvalues, and $\mathbf{c}$ is the matrix of eigenvectors.[25] Each eigenvalue and its corresponding eigenvector is associated with a cluster: the magnitude of the eigenvalues λ_i ($i = 1, . . ., N$) quantify the cohesiveness of cluster i, meaning the degree to which the molecules in the cluster are similar to one another, and the magnitude of the eigenvector elements c_{ij} ($j = 1, . . ., N$) quantify the degree to which the features of the molecule j contribute to cluster i.[24,26] $\mathbf{S}$ is a real symmetric matrix, and the eigenvectors are normalized so that

$$1 = \sum_{j=1}^{N} c_{ij}^2 \qquad (2)$$

The coefficients in the summation in eq 2 are analogous to membership functions introduced in fuzzy clustering techniques.[7] The extent to which eq 2 is satisfied when summed over individual contributions from a subset of molecules determines the completeness of cluster i.

Preliminary studies of example similarity matrices derived from binary fingerprints based upon a two-dimensional molecular structure revealed that each molecule contributes to a given cluster by a similar amount irrespective of apparent differences in the structures. To obtain clusters with increased discrimination between structurally different molecules, it was necessary to use the eigenvalues and eigenvectors of a modified form of the similarity matrix, which is referred to as a filtered similarity matrix, $\mathbf{S}^F$. The filtered similarity matrix $\mathbf{S}^F$ can be regarded as a more sparse representation of the original similarity matrix with a reduced level of background similarity; it has greater emphasis on more similar pairs of molecules and a reduced emphasis on less similar pairs of molecules.

Improved results were obtained with a discretized representation of $\mathbf{S}$ where a threshold (or exclusion) similarity

S_{cut} was applied so that $S_{ij}^F = 1$ if S_{ij} is greater than S_{cut} and $S_{ij}^F = 0$ if S_{ij} is less than S_{cut}. Although simple to implement, this discretized approach does result in a loss of information from the original matrix **S**, since in this scheme a pair of molecules is either similar or not similar, depending on the choice of S_{cut}, and intermediate levels of similarity are not represented. As an alternative to the discrete filter, a continuous filtering function

$$S_{ij}^F = e^{-\alpha(S_{ij}-1)^2} \tag{3}$$

can be applied to the elements of **S** in order to reduce the level of background similarity but at the same time to preserve information contained in the original similarity matrix.

In practice, the introduction of a single tunable parameter α in eq 3 provides a convenient way to adjust the eigenvalues and eigenvectors that are obtained from $\mathbf{S}^F$, which in turn modulates cluster cohesiveness and the cluster contributions from each molecule. Further insight into the effect of the continuous filtering function is given by considering a similarity matrix for a molecular data set with diagonal elements (i.e., the self-similarities) $S_{ii} = 1$ and off-diagonal elements in the range $0 \leq S_{ij} < 1$. In the limit $\alpha \rightarrow \infty$ the filtered similarity matrix will be a unit matrix with N nonzero eigenvalues equal to 1, and each eigenvector will have a single nonzero element so that each molecule contributes to a single cluster. Conversely, in the limit as $\alpha \rightarrow 0$ the matrix elements in the filtered similarity matrix will all be equal to 1, the matrix will have a single nonzero eigenvalue equal to N, all the corresponding eigenvector elements will be equal, and each molecule will make the same contribution to a single cluster. Finite positive values of α will produce eigenvector and eigenvalue combinations that lie between these two limiting cases.

APPLICATION

COX-2 inhibitors were downloaded from the cheminformatics.org Web site,[27] and 125 unique SMILES strings were used to calculate standard Unity 992 bit fingerprints.[28] Proprietary computer programs were used to compute a similarity matrix of intermolecular Tanimoto coefficients, filtered similarity matrices using eq 3, and, finally, the eigenvalues and eigenvectors of these matrices so that cluster assignment could be performed. The normalized cluster contributions (i.e., the squared eigenvector elements that appear in eq 2) were read into a MOE database[29] containing the molecular structures of the COX-2 inhibitors, and clusters were revealed by sorting the database entries according to decreasing cluster contributions. Results are presented in two parts: First we present example clusters obtained with a single value of α in eq 3, and second the effect of this parameter is demonstrated using a single cluster for reference.

Clusters with a Single Value of the Parameter α. Example molecular clusters obtained with $\alpha = 25$ in eq 3 are shown in Figures 2−6. The two-dimensional structures associated with the five most cohesive clusters (i.e., the five largest eigenvalues λ_i ($i = 1, . . ., 5$)) are presented along with their corresponding normalized cluster contributions. The eigenvalues for each of these clusters are provided in the figure captions, and chemical structures in each of the figures have been arranged by decreasing cluster contribution.

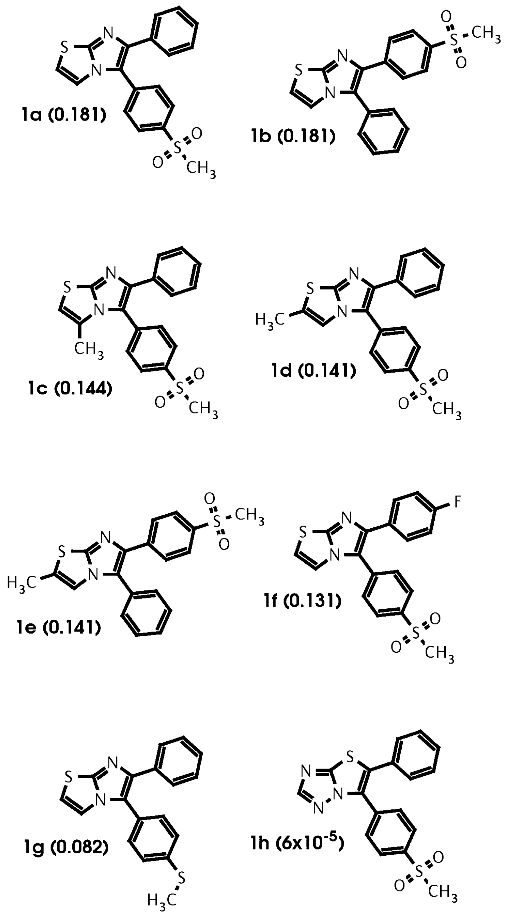

Figure 2. Example structures from cluster 1, $\lambda_1 = 5.586$.

Cluster 1 is mostly based upon 5,6-diphenylimidazo[2,1-*b*]thiazole derivatives, as shown in Figure 2. Regioisomers 1a and 1b make the largest contributions to cluster 1, both equal to 0.181. 1c is similar to 1a except for a 3-methyl group on the thiazole core, and 1c has a slightly lower contribution of 0.144. Another pair of single point change regioisomers, 1d and 1e, with 2-methyl substituents on the thiazole core have equal contributions of 0.141. 1f has a contribution equal to 0.131 and is similar to 1a except for a *para* fluorine atom on the 6-phenyl ring. 1g retains the general form of 1a but with a reduced *para* methylthioether on the 5-phenyl ring instead of the methyl sulfones in 1a−f; this has a correspondingly lower cluster 1 contribution of 0.082. The next contribution to cluster is considerably lower at 6×10^{-5} from 1h, as reflected in a change of core from imidazothiazole to thiazolotriazole.

Cluster 2 is dominated by 2-pyridyl-3-phenyl-5-chloropyridines, as shown in Figure 3. 2a−d are single point change regioisomers, where a methyl group is placed at the 4 unique carbons on the 2-pyridyl ring, and these structures have an average contribution of 0.21. 2e retains alkyl functionality on the 2-pyridyl ring in the form of a cyclopropyl ring and has a lower contribution of 0.093. In contrast to 2a−e, 2f has a primary sulfonamide at the *para* position of the 3-phenyl group instead of a methyl sulfone and has a lower contribution of 0.059. 2g displays a 2,3,5,6 tetra-substitution pattern (i.e., it has a different parent ring system) and is the next member of the cluster with a low contribution of 9 ×

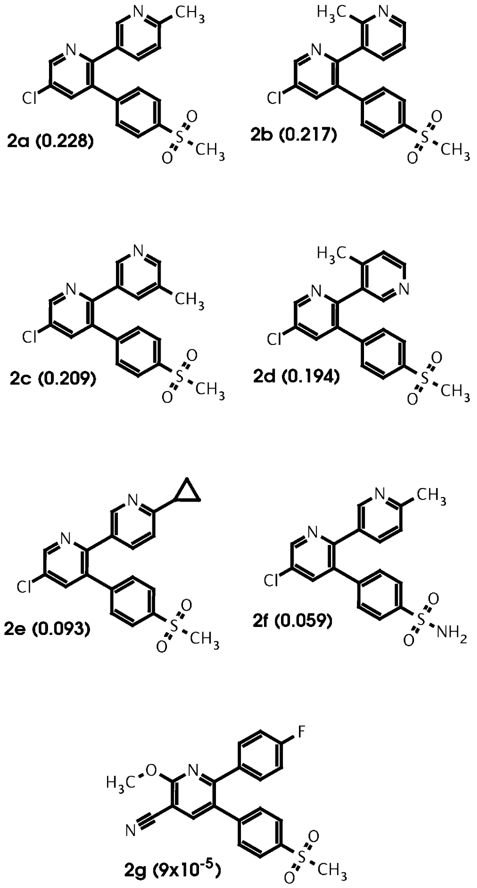

Figure 3. Example structures from cluster 2, $\lambda_2 = 4.332$.

10^{-5}, as reflected in the greater variability of substituents about the central pyridyl in comparison to other cluster members.

Cluster 3 is composed of 1,5-diarylpyrazoles as shown in Figure 4. The most dominant molecules, 3a−e, have a trifluoro methyl group at the 3-position of the pyrazole and various *para* substituents on the 1,5 pendant aryl groups. 3f exhibits a change from trifluoro methyl to chlorine at the pyrazole 3 position accompanied by a drop in contribution to 0.007. 3g and 3h are the next largest contributors to the cluster but with relatively small contributions of 0.002 and 6×10^{-4}, respectively.

Cluster 4 contains molecules with a *cis*-stilbene feature and is dominated by the diaryl cyclobutenone derivatives, 4a−c, as shown in Figure 5. 4a and 4b are geminal dimethyl regioisomers with equal contributions of 0.307. 4c is similar to 4b, except for a cyclopentyl spiro center on the 4-membered ring, and it has a slightly lower contribution of 0.297. Remaining members of cluster 4 are 5-membered ring derivatives; 4d−f are cyclopentenone analogues with averaged contributions of 0.02, and the last two members of the cluster, 4g and 4h, have lower contributions of 0.005 and 0.004, respectively, and show further structural variation.

Cluster 5 is dominated by 2,4,5-triarylthiazole derivatives as shown in Figure 6. 5a−d make the largest contributions to the cluster at an average of 0.24. 5e contains the same thiazole core but has a 2-benzyl group and makes a lower contribution of 0.046. The final example member of the cluster, 5f, has a much lower contribution of 9×10^{-5} and is identical to 1h, the thiazolotriazole structure which also made a small contribution to cluster 1.

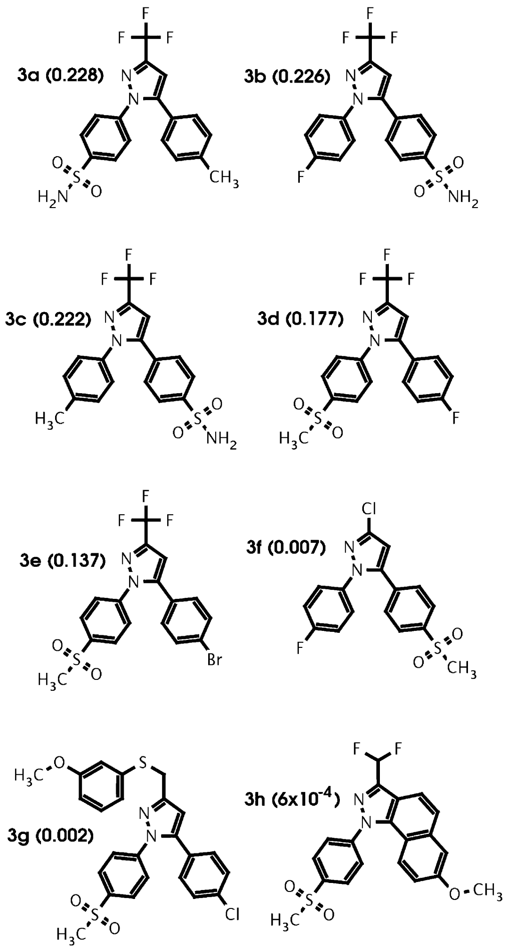

Figure 4. Example structures from cluster 3, $\lambda_3 = 3.853$.

The Effect of the Parameter α. The normalized cluster contributions of molecules 1a−h to the cluster associated with the largest eigenvalue of the filtered similarity matrix are shown in Table 1 for a series of different values of the parameter α in eq 3. At lower values of α the contributions of 1a−g, which all share an imidazothiazole core, tend to be more equivalent, and the contribution from 1h is small but still significant. As α is increased the contributions of 1a−g become less equivalent, and molecules 1a and 1b emerge with the dominant contributions. Indeed, for the smallest value of α (12.5) molecules 1a and 1b account for approximately 30% of the total cluster membership, whereas for the largest value of α (400) these two molecules alone account for roughly 70% of the total cluster membership. The contribution from 1g is reduced as α is increased; this molecule is based on an imidazothiazole core but does not possess a methyl sulfone group which is shared by 1a−f, and the lack of this feature becomes increasingly apparent at higher values of α.

DISCUSSION

The application of the spectral clustering method to the example set of COX-2 inhibitors, represented with Unity molecular fingerprints, has demonstrated several attractive features of the technique for the identification and classification of two-dimensional molecular structures. In particular, the eigenvalues and eigenvectors that are central to the method provide a natural order and means to inspect

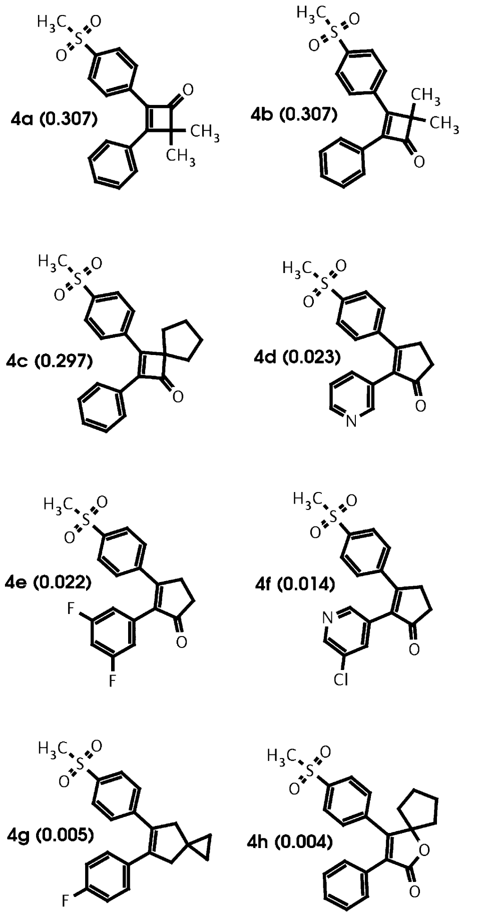

Figure 5. Example structures from cluster 4, $\lambda_4 = 3.019$.

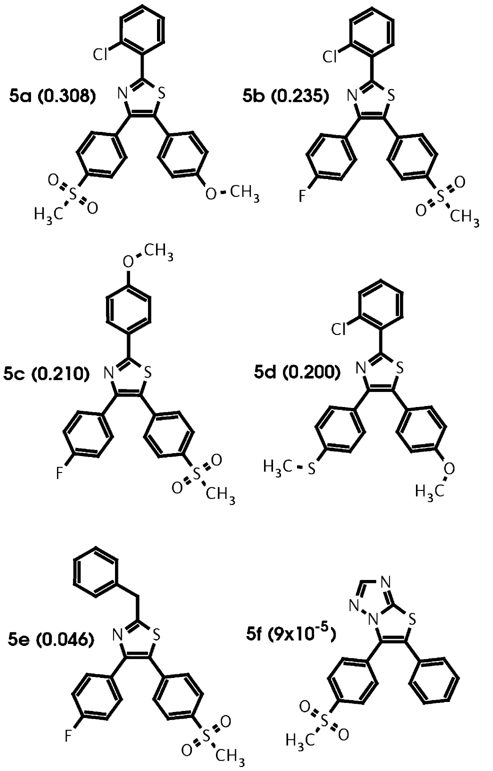

Figure 6. Example structures from cluster 5, $\lambda_5 = 2.836$.

Table 1. Normalized Cluster Contributions of Molecules 1a–h to the Cluster Associated with the Largest Eigenvalue (λ_1) of the Filtered Similarity Matrix for a Series of Different Values of the Width Parameter α

	α					
	12.5	25	50	100	200	400
λ_1	6.367	5.586	4.752	3.806	2.917	2.272
1a	0.149	0.181	0.208	0.243	0.286	0.348
1b	0.149	0.181	0.208	0.243	0.286	0.348
1c	0.132	0.144	0.142	0.133	0.115	0.075
1d	0.131	0.141	0.139	0.134	0.119	0.097
1e	0.131	0.141	0.139	0.134	0.119	0.097
1f	0.127	0.131	0.118	0.098	0.073	0.034
1g	0.097	0.082	0.046	0.016	0.002	4×10^{-5}
1h	0.004	6×10^{-5}	3×10^{-8}	0	0	0

molecular clusters. As stated earlier, the magnitude of the eigenvalues quantify cluster cohesiveness and the magnitude of the associated eigenvector elements quantify the contribution of each molecule to a cluster. So, by inspecting clusters in order of decreasing eigenvalue one can visualize clusters which display decreasing overall intermolecular similarity, and by arranging molecules within a cluster according to decreasing magnitude of the eigenvector elements one can view the molecules in terms of successively smaller cluster contributions.

Example structures from the 5 most cohesive clusters obtained from the COX-2 data set were presented according to the scheme outlined above for a single value of the width parameter, α = 25, in the filtering function described in eq 3. For this value of α, the most dominant molecular structures in each cluster were shown, as well as examples of less closely related structures which had smaller cluster contributions. One molecule, 1h/5f with a thiazolotriazole core, gave low contributions to clusters 1 and 5, which were dominated by molecules with imidazothiazole and thiazole cores, respectively. The effect of the parameter α was demonstrated using a single reference cluster. This showed that the parameter can be used to modulate the cluster contributions of the molecules, and as α is increased fewer molecules within the cluster made significant contributions.

The spectral clusters provide convenient arrangements of related molecules, and the quantitative measures of cluster cohesiveness and contribution appear to provide a useful means of analyzing and mining a data set. The ability of the method to combine different but chemically related structures both within and between different clusters is particularly attractive from a medicinal chemistry perspective, where such classification schemes taken in conjunction with activity data could prove useful in the generation of ideas for chemical synthesis. Based on the results presented here, a strategy for analyzing a data set emerges whereby one can inspect clusters in order of decreasing cohesiveness and arrange the molecules in each cluster by decreasing contribution. At some point during this inspection an alternative structure may be found (e.g., 1h in cluster 1) which could itself be useful as a starting point for further chemistry. In addition, other eigenvectors can be processed to identify other clusters where the alternative structure makes contributions (e.g., 5f in cluster 5) so that the structures of molecules in these new clusters can be considered as well.

The relative size of the molecular contributions to a given cluster are sensible from a chemical viewpoint with structures that display fewer of the dominant features within a cluster having smaller cluster contributions. The precise values of these contributions will clearly depend on several factors including the choice of molecular descriptor, similarity metric, and filtering function. Indeed, the results presented here depend specifically on the filtered Tanimoto similarities computed from Unity fingerprints; other choices of molecular descriptor, similarity metric, and filtering function will yield alternative eigenvalues and eigenvectors, hence alternative clusters. It is interesting to note that certain pairs of regioisomers have equal cluster contributions (e.g., 1a and 1b), whereas other pairs have different contributions (e.g., 1c and 1d or 1e); this is an artifact of the fingerprints employed since pairs of regioisomers with equivalent contributions have identical bit strings, while the pairs with nonequivalent contributions have different bit strings.

In this application of the spectral method we have explored overlapping clusters where all molecules effectively belong to all clusters, but we note here that the eigenvalues and eigenvectors can be used to generate nonoverlapping classifications. For example, clusters can be inspected according to decreasing eigenvalue, and molecules can be assigned to them based on the magnitude of their cluster contributions or the degree to which eq 2 is satisfied.[22] This would preclude the possibility of multicluster relationships such as the thiazolotriazole compound 1h/5f but may be useful in instances when a crisp classification and/or a specific number of clusters is required.

The spectral clustering method proceeds in a fundamentally different (global) way compared to other clustering methods that are commonly applied to molecular data sets. In the spectral approach all pair similarities are processed at once during the diagonalization of the similarity matrix, while in other approaches the elements of the matrix are more commonly explored in an iterative (local) fashion, for example, by combining molecules according to some closest distance criteria. In view of this difference we have conducted a preliminary comparison of the present spectral clustering results for the COX-2 data set with those of Ward's method,[6] which is a sequential agglomerative hierarchical nonoverlapping technique that results in the generation of a classification tree (or dendogram) which can be used to partition molecules into clusters.[1]

On application of Ward's method to the COX-2 data set, the 5 classes of dominant two-dimensional structures identified in spectral clusters 1−5 are found grouped together at the ends of individual branches of the classification tree, while the structures with lower spectral cluster contributions are found on separate branches. For example, at the 36th level of the hierarchy, structures 1a−g of cluster 1 occur on a single self-contained branch, but structure 1h is not combined with them until the 121st level, by which time there is a total of 43 molecules on the branch, including the structures 2a−g and 5a−f. Based on this preliminary comparison it appears there is consensus between the two approaches for the most dominant structures identified by the spectral method, but there is less agreement between them when it comes to the classification of the more weakly contributing structures. We emphasize here that we are not advocating one method over the other but merely demonstrating by example that the two approaches provide alternative, and quite possibly complementary, ways to analyze a data set.

FUTURE WORK

As mentioned in the Introduction clustering approaches have been developed for large data sets.[10−14] One potential drawback of the present spectral clustering implementation is the scaling of the method since diagonalization of a matrix formally requires $O(N^3)$ operations, which will limit its applicability to large data sets. [Although we have found it routinely possible with this implementation to study data sets containing up to 5000 compounds using a desktop PC.] However, we note that the Lanczos method which scales more favorably with N can be used to address this issue[23] and could permit the application of spectral clustering methods to even larger molecular data sets.

In this work two schemes were presented to compute filtered similarity matrices, one a discrete scheme based upon a threshold similarity and the other, a continuous scheme based upon a suitable functional form, but it is likely that other filtering schemes can also be applied successfully. The continuous function employed here introduces a single parameter which can be used to modulate the eigenvalues and eigenvectors of the filtered similarity matrix so as to achieve clustering at different resolutions. The choice of $\alpha = 25$ is based on our experience of applying the method in conjunction with Unity fingerprints and Tanimoto similarity coefficients, but systematic ways of choosing this parameter should be considered. Although Unity fingerprints and Tanimoto coefficients were chosen in this application, since all that is required to apply the spectral clustering technique is a similarity matrix, other types of descriptor and metric can be readily accommodated to enable spectral clustering of molecules according to different properties and similarity measures.

More generally, it would clearly be of interest to perform more detailed comparisons of spectral clustering methods with the other clustering algorithms that are firmly established for the study of molecular data sets. In addition, it will be revealing to apply spectral clustering techniques to additional data sets of different size and internal diversity than the COX-2 inhibitor data set used in this study.

CONCLUSION

A spectral clustering method has been presented and applied to the analysis of two-dimensional molecular structures. Given a suitable similarity matrix the method is generic and, through the provision of its eigenvalues and eigenvectors, provides a means to quantify the intermolecular similarity within a cluster and also the contribution that a molecule makes to a cluster. Upon application to an example data set of 125 COX-2 inhibitors, represented with Unity fingerprints, the method was shown to place the molecules into clusters of chemically related structures. Based on these results and other applications of the method we anticipate that spectral clustering approaches could enjoy widespread use in the analysis of molecular data sets but may be of particular use in the analysis of screening data, both for the prioritization of hit series and the generation of suggestions and ideas for chemical synthesis. To our knowledge this is one of the initial applications of a spectral clustering method to a molecular data set, and we conclude that it is a promising technology worthy of further investigation.

ACKNOWLEDGMENT

It is my pleasure to thank Mark Whittaker, Osamu Ichihara, Michael Charlton, Paul Mortenson, and Stephen East for various helpful discussions and suggestions during the work described in this article and Sudeep Sarkar for his help with the interpretation of graph eigenvectors.

REFERENCES AND NOTES

(1) Downs, G. M; Barnard J. M. Clustering Methods and Their Uses in Computational Chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1−40.

(2) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.

(3) Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Near Neighbours. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.

(4) Brown, R. D.; Martin, Y. C. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(5) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(6) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.

(7) Holliday, J. D.; Rodgers S. L.; Willett, P.; Chen, M.; Madfouf, M.; Lawson, K.; Mullier, G. Clustering Files of Chemical Structures Using the Fuzzy *k*-Means Clustering Method. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 894−902.

(8) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Heirarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 3182−3193.

(9) Stahl, M.; Mauser, H.; Tsui, M.; Taylor, N. R. A Robust Clustering Method for Chemical Structures. *J. Med. Chem.* **2005**, *48*, 4358−4366.

(10) Stahl, M.; Mauser, H. Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. *J. Chem. Inf. Model.* **2005**, *45*, 542−548.

(11) Böcker, A.; Derksen, S.; Schmidt, E.; Teckentrup, A.; Schneider, G. A Hierarchical Clustering Approach for Large Compound Libraries. *J. Chem. Inf. Model.* **2005**, *45*, 807−815.

(12) Böcker, A.; Schneider, G.; Teckentrup, A. NIPALSTREE: A New Hierarchical Clustering Approach for Large Compound Libraries and Its Application to Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 2220−2229.

(13) Li, W. A Fast Clustering Algorithm for Analyzing Highly Similar Compounds of Very Large Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 1919−1923.

(14) Engels, M. F.; Gibbs, A. C.; Jaeger, E. P.; Verbinnen, D.; Lobanov, V. S.; Agrafiotis, D. K. A Cluster-Based Strategy for Assessing the Overlap between Large Chemical Libraries and Its Application to a Recent Acquisition. *J. Chem. Inf. Model.* **2006**, *46*, 2651−2660.

(15) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 261−266.

(16) Fliri, A. F.; Loging, W. T.; Thadeio, P. F.; Volkmann, R. A. Analysis of drug-induced effect patterns to link structure and side effects of medicines. *Nat. Chem. Biol.* **2005**, *1*, 389−397.

(17) Livingstone, D. J. The Characterisation of Chemical Structures using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(18) Good, A. C; Richards, W. G. Explicit Calculation of 3D Molecular Similarity. *Perspect. Drug Discovery Des.* **1998**, *9−11*, 321−338.

(19) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(20) Böhm, H.; Flohr, A.; Stahl, M.; Scaffold Hopping. *Drug Discovery Today: Technol.* **2004**, *1*, 217−224.

(21) de Leval, X.; Delarge, J.; Somers, F.; de Tullio, P.; Henrotin, Y.; Pirotte, B.; Dogné, J. Recent Advances in Inducible Cyclooxygenase (COX-2) Inhibition. *Curr. Med. Chem.* **2000**, *7*, 1041−1062.

(22) Forsyth, D. A.; Ponce, J. Segmentation by Clustering. In *Computer Vision A Modern Approach;* Prentice Hall: India, 2003; pp 301−328.

(23) Shi, J.; Malik, J. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Machine Intelligence* **2000**, *22*, 888−905.

(24) Sarkar, S.; Boyer, K. L. Quantitative Measures of Change Based on Feature Organization: Eigenvalues and Eigenvectors. *Comput. Vision Image Understanding* **1998**, *71*, 110−136.

(25) Press, W. H; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. Eigensystems. In *Numerical Recipes in Fortran 77;* Cambridge University Press: Cambridge, 2001; pp 449−489.

(26) Eigenvectors that yield discriminating clusters are positive, meaning all elements of the eigenvector share the same sign, but as Sarkar and Boyer have demonstrated this condition can be partially relaxed, and eigenvectors with dominant positive components also provide discriminating clusters. The condition may be completely relaxed if overlap is acceptable between the resultant clusters.

(27) cheminformatics.org. http://www.cheminformatics.org/ (accessed Dec 21, 2006).

(28) *SYBYL, version 7.2*; Tripos: St. Louis, MO, 2006.

(29) *MOE, version 2006.08*; Chemical Computing Group: Montreal, Quebec, 2006.

CI600565R