Journal of Molecular Graphics and Modelling 28 (2009) 187–195

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Clustering files of chemical structures using the Székely–Rizzo generalization of Ward's method

Thibault Varin [a], Ronan Bureau [a], Christoph Mueller [b], Peter Willett [b,*]

[a] Centre d'Etudes et de Recherche sur le Médicament de Normandie, UPRES EA4258, INC3M FR CNRS 3038, Université de Caen, Boulevard Becquerel, 14032 Caen Cedex, France
[b] Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, 2111 Portobello Street, Sheffield S1 4DP, United Kingdom

ARTICLE INFO

Article history:
Received 6 March 2009
Received in revised form 22 June 2009
Accepted 27 June 2009
Available online 4 July 2009

Keywords:
Clustering method
Distance coefficient
Energy clustering
Fingerprint
Fragment substructure
Joint between-within distance
Minimum variance clustering method
Soergel coefficient
Székely–Rizzo clustering method
Ward's clustering method

ABSTRACT

Ward's method is extensively used for clustering chemical structures represented by 2D fingerprints. This paper compares Ward clusterings of 14 datasets (containing between 278 and 4332 molecules) with those obtained using the Székely–Rizzo clustering method, a generalization of Ward's method. The clusters resulting from these two methods were evaluated by the extent to which the various classifications were able to group active molecules together, using a novel criterion of clustering effectiveness. Analysis of a total of 1400 classifications (Ward and Székely–Rizzo clustering methods, 14 different datasets, 5 different fingerprints and 10 different distance coefficients) demonstrated the general superiority of the Székely–Rizzo method. The distance coefficient first described by Soergel performed extremely well in these experiments, and this was also the case when it was used in simulated virtual screening experiments.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Cluster analysis [1–3] is a multivariate technique that has been extensively used in chemoinformatics to partition a set of molecules into clusters that exhibit a high degree of both intra-cluster similarity and inter-cluster dissimilarity [4–6]. The molecules in such analyses are normally represented by 2D fingerprints, an approach that was first suggested over three decades ago by Adamson and Bush [7] and that has since been extensively used for applications such as the analysis of substructure-search outputs [8], the selection of molecules for biological screening [9], property-prediction [10], and molecular diversity analysis [11].

There are many different ways in which clustering can be carried out. This has encouraged comparative studies to determine the effectiveness of the various clustering methods that are available when they are applied to the clustering of chemical structures. Early property-prediction studies of over 30 hierarchic and non-hierarchic clustering methods [4] highlighted the consistent performance of the hierarchical agglomerative method first described by Ward [12], a finding that was confirmed in subsequent studies by Brown and Martin [10,13] and by Downs et al. [14]. With the advent of efficient implementations based on the reciprocal nearest neighbours algorithm [15], Ward's method is now arguably the method of choice for clustering databases containing up to *ca.* 500K structures (larger files may be clustered using an hierarchic divisive procedure based on the well-known *k*-means relocation clustering method [16,17]).

Székely and Rizzo have recently described a new hierarchical agglomerative clustering method [18]. This method generalizes Ward's method by defining a cluster distance and objective function in terms of a power in the interval (0,2] of the Euclidean distance between cluster centres, with Ward's method being obtained as the limiting case when the power is 2. Varin et al. have applied the Székely–Rizzo method to the clustering of chemical structures as part of a detailed study of ligands for the 5-hydroxytryptamine subtype-4 (5-HT_4) receptor, and found that this method led to better grouping of the actives than did conventional hierarchical agglomerative methods, including Ward's method [19]. In this paper, we report an extended evaluation of the Székely–Rizzo method, using a range of different distance metrics and using not just this 5-HT_4 data but also other 5-HT datasets from the Université de Caen and 10 public PubChem datasets.

* Corresponding author.
E-mail address: p.willett@sheffield.ac.uk (P. Willett).

1093-3263/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2009.06.006

2. The Székely–Rizzo clustering method

Hierarchical agglomerative methods have been used in a vast range of application domains [1,2]. They generate a classification in a bottom-up manner, by a series of agglomerations in which small clusters, initially containing individual molecules, are fused together to form progressively larger clusters, with the most similar pair of clusters being fused at each stage of the classification. The various hierarchic agglomerative methods differ only in the criterion that is used to select the most similar pair of clusters at each stage; indeed, they can all be implemented using a common algorithm first described in detail by Lance and Williams [20] (although more efficient algorithms are available for individual clustering methods [21]).

The fusion criterion for many hierarchic agglomerative methods focuses on between-cluster distances; for example, the single linkage method (or the complete linkage method) fuses that pair of clusters for which the distance between two existing clusters is the minimum (or the maximum) of all the distances between molecules in one cluster and molecules in the other cluster. Ward's method differs from other hierarchic agglomerative methods in its use of a fusion criterion that results in clusters that are both homogenous (in the sense that the criterion minimizes the within-cluster distances) and heterogeneous (in the sense that the criterion maximizes the between-cluster distances). This is achieved by minimizing the increase in the total within-cluster sum of squared errors when two clusters are fused, an increase that is proportional to the squared Euclidean distance between the mean centres of the two clusters that are being fused.

Székely and Rizzo have recently described "a joint between-within *e*-distance between clusters" that encompasses both intra-cluster homogeneity and inter-cluster heterogeneity, and use this distance as the criterion for a generalized clustering method that includes Ward's method as a limiting case [18]. Specifically, Székely and Rizzo define the *between-within distance*, or *e*-distance $e(A,B)$, between two clusters A and B, containing n_a and n_b objects, respectively, as

$$e^{\alpha}(A,B) = \frac{n_a n_b}{n_a + n_b}\left(\frac{2}{n_a n_b}\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}||a_i - b_j||^{\alpha} - \frac{1}{n_a^2}\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}||a_i - a_j||^{\alpha} - \frac{1}{n_b^2}\sum_{i=1}^{n_b}\sum_{j=1}^{n_b}||b_i - b_j||^{\alpha}\right),$$

where the exponent α is in the range $0 < \alpha \leq 2$ and where $||X - Y||$ is the Euclidean distance between two objects X and Y. Székely and Rizzo focus on two special cases: $\alpha = 2$ and $\alpha = 1$. They show that the first of these cases, $e^2(A,B)$ is proportional to the weighted squared distance between the cluster means for A and B, and that this will accordingly yield a Ward hierarchy. They note that the second of these cases, $e^1(A,B)$, has several desirable theoretical properties, one of which (that of statistical consistency as defined by Kaufman and Rousseuw [22]) is not exhibited by $e^2(A,B)$ (i.e., Ward's method). They suggest that clusters based on $e^1(A,B)$ will be superior to those based on $e^2(A,B)$ for separating clusters with the same, or close, cluster means but different distributions of points around those means, and use dermatological, gene expression and simulated multivariate normal datasets to demonstrate that $e^1(A,B)$ can indeed out-perform $e^2(A,B)$ in some circumstances [18]. The Székely–Rizzo clustering method is available as a contributed package, called Energy, in the *R* statistical system (available from http://www.r-project.org/), and this was used for all the experiments reported below.

Székely and Rizzo suggest that it is possible to obtain additional clustering methods by replacing the Euclidean distance in the formula above with other distance coefficients, but note that the resulting methods may not possess the same theoretical properties as those obtained using the Euclidean distance. This suggestion was developed by Varin et al. in their study of ligands for the 5-HT_4 receptor: they used not just Euclidean distance but also six other similarity and distance coefficients [19]. They also used a range of clustering methods (single linkage, complete linkage, group average, Ward and energy methods) and structural descriptors (ChemAxon JChem fingerprints, Unity fingerprints and ChemAxon 2D pharmacophore fingerprints). They found that the best results were obtained using the energy method (i.e., Székely–Rizzo $e^1(A,B)$ clustering) with the Canberra distance [23], rather than Euclidean distance, as the fusion criterion. In this paper, we report a more detailed comparison of the effectiveness of the energy and Ward methods, using: five different types of fingerprint (all of which take account of the frequency of occurrence of the encoded fragments substructures, rather than just their presence or absence as with conventional binary fingerprints); 14 different datasets; and 10 different distance coefficients. These are detailed in the following sections.

3. Experimental details

3.1. Fingerprints

Our experiments have used five different types of fingerprint, these being selected from three broad classes, as summarized in Table 1. The first class contains circular substructure fingerprints (specifically those popularized in the Pipeline Pilot system and available from Accelrys Software Inc. at http://www.accelrys.com). The elements of the ECFC_4 fingerprints contain the counts of fragment occurrences, in which the atoms are encoded by their atomic types and in which the circular substructures are of diameter four bonds. In the related FCFC_4 fingerprints, the atoms are encoded by their functional types. The fragments in each molecule were encoded in a vector containing 1024 integer elements. The second class contains 2D pharmacophore fingerprints (specifically those available from ChemAxon at http://www.chemaxon.com). Two of these fingerprints were used here: pharmacophore fingerprints (referred to as PFP), in which the elements contain the counts of fragments occurrences, and the fragments consist of pairs of atoms encoded by one of six pharmacophore types together with their through-bond separations; and the related fuzzy pharmacophore fingerprints (referred to as FPFP), in which the counts are smoothed to take account of the number of rotatable bonds separating each pair of atoms. The fragments for each molecule were encoded in a vector containing 210 integer (for PFP) or real (for FPFP) elements. The third class

Table 1
Fingerprints used in the clustering experiments.

Code	Name	Source	Atom abstraction	Fragment type	Elements
D1	ECFC_4	Pipeline Pilot	Elemental type	Circular substructures	1024 integers
D2	FCFC_4	Pipeline Pilot	Functional class	Circular substructures	1024 integers
D3	PFP	Chemaxon	Functional class	Atom pairs	210 integers
D4	FPFP	Chemaxon	Functional class	Atom pairs	210 reals
D5	Holograms	Tripos	SYBYL atom type	Atom chains	997 integers

Table 2
Datasets used in the clustering experiments.

Code	Target	Molecules	Active molecules
E1	5-HT_{1E}	3080	442
E2	5-HT_{1A}	2048	365
E3	Thyroid stimulating hormone receptor	2303	343
E4	Protein kinase D	727	109
E5	5-HT_{1E}	4322	634
E6	Acetylcholine muscarinic M1 receptor	701	136
E7	ras and ras-related GTPase	1489	225
E8	5-HT_{1A}	2908	413
E9	Prostaglandin EP2 receptor	1313	139
E10	Hydroxyprostaglandin dehydrogenase	945	91
E11	5-HT_{1A}	278	69
E12	5-HT_4	995	170
E13	5-HT_6	1020	39
E14	5-HT_7	992	166

contains strings of 4–7 atoms encoded in Unity holograms (available from Tripos International Inc. at http://www.tripos.com) using the default parameters; the holograms involve a superimposed coding procedure in which a hashing procedure is used to associate each specific string with multiple elements in the molecule's hologram. The fragments for each molecule were encoded in a vector containing 997 integer elements.

3.2. Datasets

We have used 14 datasets in our experiments, as detailed in Table 2. Four of these were created by CERMN (Centre d'Etudes de Recherche sur le Médicament de Normandie, see http://www.cermn.unicaen.fr/chimiotheque.html) as part of an ongoing project to develop ligands for a range of 5-HT receptors [19]. In addition, we have used 10 datasets downloaded from the PubChem database at http://pubchem.ncbi.nlm.nih.gov/, with the bioassays selected to cover not just 5-HT receptors but also a range of other types of biological target. For each dataset, we filtered all molecules not definitely active or not definitely inactive, duplicate molecules, molecules containing non-organic elements, and molecules with a molecular weight of less than 150. The remaining molecules were then processed to remove salts and to standardize charges and stereochemistry. These procedures yielded datasets containing between 278 and 4332 molecules, with between 3.8% (for E13) and 24.8% (for E11) recorded as being active.

3.3. Distance coefficients

Many different types of similarity coefficient have been used in chemoinformatics, most commonly association coefficients such as the Tanimoto coefficient [24]. As noted above, the Székely–Rizzo method requires the use of a distance, and we have hence used the 10 distance coefficients listed below, taken from the extensive review of metric coefficients presented by Gower and Legendre [25]. Assume that a molecule X_i is represented by a p-element vector, with the element X_{ik} containing the frequency of occurrence of the k-th fragment in X_i (and similarly for another molecule X_j). Then the coefficients studied here are as follows:

$$\mathrm{M1} = \sqrt{\frac{1}{p}\sum_{k=1}^{p}(X_{ik} - X_{jk})^2}$$

$$\mathrm{M2} = \sqrt{\frac{1}{p}\sum_{k=1}^{p}\frac{(X_{ik} - X_{jk})^2}{R_k^2}}$$

$$\mathrm{M3} = \sqrt{\frac{1}{p}\sum_{k=1}^{p}\frac{(X_{ik} - X_{jk})^2}{\sigma_k^2}}$$

$$\mathrm{M4} = \frac{1}{p}\sum_{k=1}^{p}|X_{ik} - X_{jk}|$$

$$\mathrm{M5} = \frac{1}{p}\sum_{k=1}^{p}\frac{|X_{ik} - X_{jk}|}{R_k}$$

$$\mathrm{M6} = \frac{1}{p}\sum_{k=1}^{p}\frac{|X_{ik} - X_{jk}|}{\sigma_k}$$

$$\mathrm{M7} = \sqrt{\frac{1}{p}\sum_{k=1}^{p}\frac{(X_{ik} - X_{jk})^2}{(X_{ik} + X_{jk})^2}}$$

$$\mathrm{M8} = \frac{1}{p}\sum_{k=1}^{p}\frac{|X_{ik} - X_{jk}|}{|X_{ik}| + |X_{jk}|}$$

$$\mathrm{M9} = \sum_{k=1}^{p}\frac{|X_{ik} - X_{jk}|}{Max(X_{ik}, X_{jk})}$$

$$\mathrm{M10} = \frac{1}{p}\sum_{k=1}^{p}\left(1 - \frac{Min(X_{ik}, X_{jk})}{Max(X_{ik}, X_{jk})}\right)$$

where R_k and σ_k are the range and the standard deviation, respectively, for the k-th variable. Of these, M1 and M4 are examples of Minkowski metrics and M2, M3, M5 and M6 are special cases of them. A potential problem with many of these coefficients is that both X_{ik} and X_{jk} are frequently zero: this corresponds to a substructural fragment that is absent from both of the molecules that are being compared and results in the denominator in the expression for the coefficient having a value of zero. In each such case, we have ignored the k-th fragment and reduced the value of p by one, as recommended by Gower and Legendre [25].

3.4. Evaluation of clusterings

A long-standing problem in cluster analysis is that of evaluating the effectiveness of the classification produced by a specific clustering procedure. One procedure that has been used in the chemical context is to determine the extent to which a procedure is able to cluster together the active molecules in a dataset, whilst simultaneously separating them from the inactives. This procedure was first applied on a large scale in the much-cited papers of Brown and Martin [10,13] and we have used a development of their procedure in the work reported here. Brown and Martin defined an *active cluster* as a non-singleton cluster that contained at least one active molecule, and the *active cluster subset* as the set of molecules, both active and inactive, in the active clusters; they then evaluated their clustering experiments using an index, P_a, describing the fraction of the active cluster subset that were active molecules.

A limitation of P_a is that it is severely affected by large clusters containing just a single active, a common occurrence even when the bulk of the actives are tightly clustered together. Varin et al. hence developed an alternative performance measure in which an active cluster is now defined as a non-singleton cluster for which the percentage of active molecules is greater than the percentage in the dataset as a whole [19]. Let p be the number of actives in active clusters, q the number of inactives in active clusters, r the number of actives in inactive clusters (i.e., clusters that are not active clusters) and s the number of singleton actives. Then the quality partition index, *QPI*, is defined to be

$$QPI = \frac{p}{p + q + r + s}.$$

This expression will have its upper-bound value of unity when p is the total number of actives and when q, r and s are all zero (i.e., when the actives are clustered tightly together on their own) and its lower-bound value of zero when none of the actives are in active clusters. These would seem to be appropriate characteristics for a measure of clustering effectiveness. However, assume that, e.g., $p = 12$ for a dataset; then, other things being equal, the value of *QPI* will be the same irrespective of whether there is a single active cluster containing all 12 actives or whether there are 3 active clusters each containing 4 actives, despite the fact that one would, arguably prefer the former situation (i.e., a single, large cluster rather than multiple smaller clusters). A modified form, of QPI_w, was hence employed in which the QPI_w value at level l in the hierarchy was weighted by:

$$QPI_w = QPI \times \frac{nc + 1 - l}{nc}$$

where nc is the number of compounds. The $-l$ term in the penalty function means that we focus attention towards the top of the hierarchy, so as to summarize the dataset in a small number of clusters. Similar results to those presented below were obtained when QPI_w was defined with other weights based on

$$\left(\frac{nc + 1 - l}{nc}\right)^{1/2}, \left(\frac{nc + 1 - l}{nc}\right)^{2}, \quad \text{and} \quad \log\left(\frac{nc + 1 - l}{nc}\right)$$

The QPI_w value can be computed at each level of a cluster hierarchy and the result displayed as in Fig. 1. The operation of the QPI_w approach is illustrated in Fig. 1, which shows the variation of QPI_w, *QPI* and P_a with the level of the hierarchy. This figure is based on clustering the 995 molecules comprising dataset E12, and using the D1 descriptor (i.e., ECFC_4 fingerprints), the M9 distance coefficient and the energy (i.e., Székely–Rizzo $e^1(A,B)$) clustering method. The hierarchy level (on the X-axis) has been plotted as a natural logarithm to focus attention on the maxima in the *QPI* and QPI_w curves at low numbers of clusters. It will be seen that P_a (shown in green) increases with cluster level, whereas both *QPI* (shown in blue) and QPI_w (shown in red) reach a maximum value, with the latter achieving its maximum at a much lower level in the hierarchy. The index was developed for the analysis of hierarchic clusterings, but is also applicable to non-hierarchic clusterings, such as the partitions produced by a single-pass or a k-means method, by letting l denote the number of partitions in the classification.

The maximum QPI_w value corresponds to that level in the hierarchy for which the corresponding partition results in the best possible separation of the active and inactive molecules in the dataset. It would be possible to use this maximum value as a performance criterion to compare different classifications and to give an upper-bound to clustering performance, in the same way that the *optimal cluster* approach has been used to set an upper-bound for the evaluation of hierarchic document clustering methods in information retrieval (where the aim is to cluster together textual documents relevant to the same query, rather than to cluster together molecules with the same bioactivity as here) [26,27]. However, while the maximum value highlights the best possible partition, it may not be representative of the effectiveness of the hierarchy as a whole; our chosen performance criterion is hence based on the QPI_w values at each level. Specifically, we define the *Quality Hierarchy Index*, *QHI*, by:

$$QHI = \sqrt{\sum_{l=1}^{n-1} QPI_w^2}.$$

where n is the number of compounds in the dataset, and hence $n - 1$ the number of levels in the hierarchy.

We have used the *QHI* values as the performance criterion to compare the very large numbers of different clusterings that can be obtained by combining the various parameters listed previously. Specifically, we generated a total of 100 different classifications (resulting from the combination of two different clustering methods, 10 different distance coefficients and five different structure representations) for each of the 14 datasets listed in Table 2. The *QHI* value was recorded for each of the 1400 resulting classifications, and then this data (or subsets thereof) was analyzed to determine the effect of the various factors (clustering method, distance coefficient and structure representation) on clustering performance.

4. Results and discussion

4.1. Analysis of results

The *QHI* values (rounded to two decimal places) for the classifications are listed in Table 3. The values for all 1400 combinations of descriptor (D1–D5), distance coefficient (M1–M10), clustering method (Ward or Energy) and experimental dataset (E1–E14) are listed in the Supplementary Material; in the published paper, Table 3 contains just 10% of this Material, specifically the 10 top-ranked combinations of descriptor, distance coefficient and clustering method. The final column in each row of the table gives the overall mean rank for that particular combination of descriptor, metric and clustering method when analyzed using Kendall's coefficient of concordance, W. This is a non-parametric statistic that is used to evaluate the consistency of k different sets of ranked judgments of the same set of N different objects [28]. The basic form of the W statistic is:

$$\frac{12 \sum_{i=1}^{N} (R_i - R)}{N^3 - N}$$

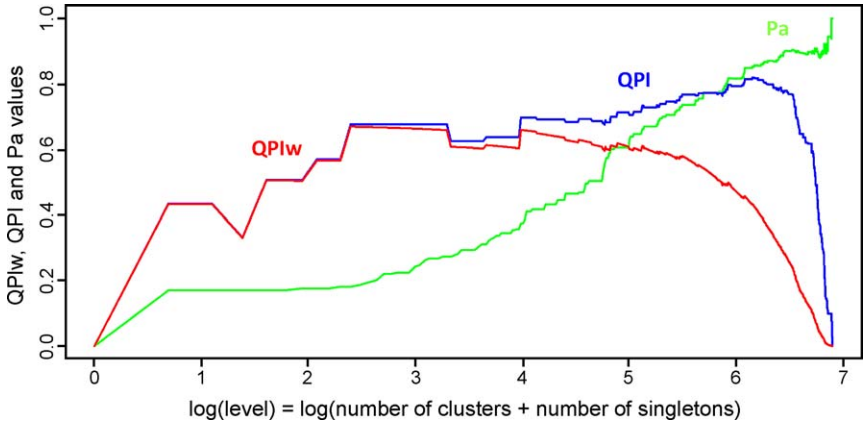

Fig. 1. QPI_w (red), *QPI* (blue) and P_a (green) curves for E12, with the parameter combination D1/M9/Energy.

Table 3
QHI values (rounded to two decimal places) for combinations of descriptor (D1–D5), distance coefficient (M1–M10), clustering method (Ward or Energy) and experimental dataset (E1–E14). The final column in each row gives the overall mean rank for that particular combination of descriptor, metric and clustering method in the Kendall *W* analysis. The reader should note that this table contains the E1–E14 data for just the 10 top-ranked combinations of descriptor, distance coefficient and clustering method: the full table, containing all 100 such combinations is provided in Supplementary Material.

Combination			Dataset													Mean Rank
			E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	
D1	M9	Energy	11.97	9.88	9.88	5.42	16.79	8.43	9.98	14.21	6.06	5.02	6.51	13.26	7.31	10.08	17.50
D2	M7	Energy	11.98	9.94	9.89	5.32	16.84	8.81	10.05	14.04	6.22	5.21	6.33	12.92	7.16	9.48	19.36
D2	M10	Energy	11.71	9.80	9.86	5.28	17.19	8.82	10.00	14.28	6.24	5.49	6.49	12.53	7.11	9.54	19.93
D2	M9	Energy	11.77	9.66	10.02	5.11	16.56	8.92	9.43	14.64	6.09	4.83	6.37	13.28	7.53	9.94	20.07
D2	M8	Energy	11.83	9.96	9.88	5.37	16.99	8.78	9.92	14.19	6.24	5.50	6.47	12.47	7.05	9.37	20.14
D1	M7	Energy	11.84	9.91	9.84	5.33	16.79	8.48	10.11	14.11	6.25	5.07	6.51	13.04	6.44	9.75	21.43
D1	M8	Energy	11.76	9.76	9.76	5.33	17.01	8.30	10.09	14.17	6.20	4.87	6.74	12.86	6.76	9.61	22.14
D2	M7	Ward	11.83	9.88	9.89	5.23	16.97	9.01	10.00	14.04	6.18	5.22	6.37	12.55	6.98	9.25	22.36
D1	M7	Ward	11.70	9.92	9.93	5.29	16.89	8.38	10.24	14.04	6.16	5.09	6.69	12.85	6.40	9.77	22.43
D1	M10	Energy	11.85	9.75	9.93	5.20	17.01	8.32	9.92	14.01	6.17	5.00	6.78	12.82	6.76	9.60	22.50

where R_i is the mean of the ranks assigned to the i-th object and R is the grand mean of the ranks assigned to all N objects (there is a more complex form, used here, that takes account of ties in the rankings).

To illustrate its use in the present context, we consider the analysis of Table 3, i.e., the problem of comparing the 100 different classifications (each representing one particular combination of clustering method, distance coefficient and structure representation) that can be generated for each of the 14 datasets. We consider each of the datasets as a judge ranking the different classifications in order of decreasing effectiveness (as measured by the *QHI* value), i.e., $k = 14$ and $N = 100$. The first step is hence to convert the data in Table 3 (*QHI* values) to ranks, so that each column contains 100 integers in the range 1–100 (although not all of these values may be present if, as is normally the case, there are tied rank positions). The ranks are used to compute the Kendall statistic, for which the observed value is 0.54. The statistical significance of this W value can be tested using the χ^2 distribution since for $N > 7$,

$$\chi^2 = k(N-1)W$$

with $N - 1$ degrees of freedom (alternatively, Siegel and Castellan provide a table of critical values for W when $N <= 7$ [28]). The value is significant ($p < 0.001$). Given that a significant level of agreement has been achieved, Siegel and Castellan suggest that the best overall ranking of the N objects can be obtained using their mean ranks averaged over the k judges, i.e., the R_i values in the expression for W. This analysis shows that the best single combination is that used in Fig. 1, i.e., D1/M9/Energy (the ECFC_4 descriptor, the M9 distance coefficient, and the energy clustering method), a fact that we shall discuss in more detail below. The mean rank for this combination is 17.5, as listed in the right-hand column of Table 3, the rows of which have been arranged in order of increasing mean rank, i.e., decreasing effectiveness of clustering.

An exactly comparable procedure can be used for alternative analyses (e.g., enabling one to compare the different descriptors using each combination of dataset, distance metric and clustering method) or to use subsets of the data by holding one factor constant (e.g., enabling one to compare the distance metrics when used for energy clustering of the 5-HT datasets).

4.2. Comparison of descriptors

When we compared the five different descriptors, the computed value of W was 0.25, which is statistically significant ($p < 0.01$) and which identified the D1 descriptor (ECFC_4) as the best, with a mean rank of 2.14 when averaged over the 280 classifications involving it. Similar statistically significant ($p < 0.01$) conclusions were obtained when the two different clustering methods were considered separately and when the 10 different coefficients were considered separately.

4.3. Comparison of distance coefficients

A similar situation pertains when we consider the 10 different distance coefficients. Here, $W = 0.54$; this value is highly significant ($p < 0.001$) and we can hence rank the coefficients, as shown in Table 4, which lists the mean ranks when averaged over the 140 different classifications involving each coefficient. The best results were obtained with M9, followed by M10, M7 and M8. It is noticeable that all the Minkowski-related metrics perform relatively poorly here (except M2 with an overall mean rank of 4.92), despite their frequent use for clustering applications. This result is particularly surprising since Euclidean distance (coefficient M1) is the basis for the generalized Székely–Rizzo method. The same, highly significant correlations ($p < 0.001$) were also obtained when the two different clustering methods were considered separately. There was some variation when the five different descriptors were considered separately, as shown in Table 4, although all the W values were again highly significant ($p < 0.001$): M9 was the best coefficient with D1 (ECFC_4), D4 (FPFP) and D5 (holograms) but was second-best to M7 for D2 (FCFC_4) and to M10 for D3 (PFP).

In view of the very high level of performance achieved by the M9 coefficient, which uses the Max{X_{ik},X_{jk}} normalisation first described by Soergel [29], we have carried out additional experiments to assess its suitability for virtual screening applications. This study builds on earlier work by Fechner and Schneider who noted the excellent screening performance of the Soergel distance in experiments using a small file from the COBRA database [30]: our experiments are on a much larger scale and are

Table 4
Mean rank for distance metrics using all descriptors and using each of the five individual descriptors. The best performing (lowest mean rank) metric in each case is shaded.

Coefficient	All	D1	D2	D3	D4	D5
M1	6.19	7.04	6.86	7.89	3.68	5.46
M2	4.92	4.96	4.46	5.32	4.75	5.11
M3	8.46	8.79	8.86	8.93	8.25	7.46
M4	5.10	5.43	5.39	5.36	3.36	5.96
M5	6.40	6.07	5.96	6.68	6.54	6.75
M6	9.67	9.82	9.93	9.79	10.00	8.82
M7	4.11	3.00	3.11	3.57	7.07	3.79
M8	4.19	3.46	3.68	3.11	5.96	4.71
M9	2.54	2.61	3.21	2.61	1.82	2.46
M10	3.43	3.82	3.54	1.75	3.57	4.46

discussed below in the section *Additional comparison of distance coefficients*.

4.4. Comparison of clustering methods

The Kendall W analysis is not appropriate for comparing just two objects, i.e., the energy and Ward clustering methods in the present context. Instead, we have used the large-sample version of the Wilcoxon signed ranks test: this is used to analyze paired observations in which one notes the magnitude and the direction of the difference between the two observations that are being compared [28]. The difference between each pair of observations is noted and these differences ranked. Let T be the sum of the ranks for the positive differences; then, for large N, T is normally distributed with a mean of

$$\frac{N(N+1)}{4}$$

and a variance of

$$\frac{N(N+1)(2N+1)}{24},$$

and the significance of T can hence be determined using the Z statistical test.

The two clustering methods gave different results for every single one of the 700 combinations of dataset, distance metric and descriptor. The mean ranks for the two methods are listed in Table 5 where it will be seen that the energy method was notably superior to Ward's method. The computed value for Z in the Wilcoxon test is 19.50: this value is highly significant ($p < 0.001$) and this was again the case when the five descriptors were considered separately (mean ranks also shown in Table 5) and when the 10 distance metrics were considered separately (data not shown).

4.5. Comparison of all combinations

We have discussed previously the analysis of the complete data (i.e., 100 combinations of clustering method, descriptor and distance coefficient for each of the 14 datasets), and noted that a highly significant value of 0.54 was obtained for W. The best single combination was D1 (the ECFC_4 descriptor), M9 (the Soergel coefficient) and the energy method, with a mean rank position of 17.50 when averaged across the complete set of 1400 classifications.

We can make some further general observations when the 100 combinations are sorted into decreasing order of mean rank, as shown in the Supplementary Material. Thus, if we consider the top-20 combinations in the sorted list (from the top row down to and including the combination D1/M4/Energy), then noteworthy appearances (or non-appearances) include: the descriptors D1 and D2 appear 10 and 9 times, with no other descriptor appearing more than once; distance coefficient M9 appears 5 times, M7 and M8 both appear 4 times, with no other metric appearing more than twice; energy appears 12 times and Ward appears 8 times, with the former out-performing the latter for all cases where the combination of descriptor and metric was the same. Conversely, if we consider the last-20 combinations (from the bottom row up to and including the combination D1/M3/Ward in the Supplementary Material), then: the descriptors D4, D3, D2 and D1 appear 8, 5, 4 and 3 times, respectively, with no appearances of D5; the coefficients M6 and M3 appear 8 and 7 times, respectively, with no other coefficient appearing more than once; Ward appears 13 times and energy appears 7 times. Based on such observations, a high-performing combination might be expected to involve D1 or D2, M7, M8 or M9 and Energy, and the highest ranked combination is indeed found to be D1/M9/Energy.

Table 5
Mean rank for clustering methods using all descriptors and using each of the five individual descriptors. The energy method always performs better (lower mean rank) than the Ward method.

Method	All	D1	D2	D3	D4	D5
Ward	1.86	1.76	1.82	1.92	1.92	1.89
Energy	1.14	1.24	1.18	1.08	1.08	1.11

4.6. Characteristics of clusters

Thus far, we have considered the clusters only by means of the QPI_w values associated with the optimal classifications for each combination of parameters. Here, we look briefly at the composition of these clusters for each of the datasets. We have chosen to illustrate the clusters with three combinations: D1/M9/Energy (the best overall combination), D1/M9/Ward (differing from the best combination only in the clustering method) and D1/M1/Energy (differing from the best combination only in the distance coefficient).

Table 6 details the make-up of the optimal classifications for these three combinations for each of the datasets, listing the level of the best partition, the number of active clusters (as defined previously), the mean size of the active clusters, and the values of p, q, r, s and QPI_w. It will be seen that the Energy and Ward methods normally result in similar optimal partitions in terms of the level in the hierarchy and the number of active clusters. The one obvious exception to this general behaviour is with dataset-E6 where Energy identified just a single active cluster in the optimal partition, as against 23 for the Ward classification: this is explained by Fig. 2, which shows an extended plateau for Energy with the very highest value at the extreme left-hand end. The comparison between D1/M9/Energy and D1/M1/Energy reveals a greater difference in behaviour. For 10 of the datasets – the exceptions are E6 and E11-13 – the best partition for M1 is at a lower level (i.e., a smaller number of clusters) than for M9 with a lesser number of active clusters. For eleven of the datasets, the M9 p values are greater than the M1 p values: the exceptions are E13 (where the two are equal) and E6, E11 and E12 (where the M1 value is greater).

4.7. Analysis of dataset types

A referee noted that eight of the 14 datasets considered in this study are associated with 5-HT receptors, and wondered whether similarities between these sets of ligands might have affected the results. Two further sets of experiments were hence carried out to investigate this possibility: clustering the datasets on the basis of their constituent molecules; and repeating the analyses described above but omitting all but the largest of the eight 5-HT datasets, i.e., E5.

Given two datasets A and B, the similarity between A and B was computed as the sum of the pair-wise inter-molecular similarities, where one molecule of each pair was in A and the other in B. These inter-molecular similarities were computed using either Unity or ECFP_4 fingerprints with the cosine coefficient, and the resulting matrix of inter-dataset similarities clustered using the complete link hierarchic agglomerative clustering method. Both types of fingerprint yielded a classification with three well-marked clusters: one containing E10 on its own, one containing E6 and E11–E14, and the third containing the eight remaining datasets (E1–E5 and E7–E9). Thus, rather than clustering together in a single group, the eight 5-HT datasets have split so that there are four of them in each of the two non-singleton clusters, i.e., they do not form a single group that is structurally distinct from the other datasets.

Table 6
Comparison of the best partition (i.e., that with the maximum value for QPI_w) for each dataset using the combinations D1/M9/Ward, D1/M9/Energy and D1/M1/Energy. The table also lists the p, q, r and s values, and the number and the mean size of the active clusters in the best partition.

Dataset	Combination	Best partition							
		Level	Active clusters	Mean size	p	q	r	s	QPI_w
E1	D1/M9/Ward	1072	296	3.3	433	548	3	6	0.29
	D1/M9/Energy	979	290	3.5	434	572	6	2	0.29
	D1/M1/Energy	812	252	4.2	409	638	29	4	0.28
E2	D1/M9/Ward	387	151	5.1	306	466	59	0	0.30
	D1/M9/Energy	424	162	4.7	308	451	57	0	0.30
	D1/M1/Energy	401	146	4.9	288	428	75	2	0.29
E3	D1/M9/Ward	708	214	3.8	333	480	7	3	0.28
	D1/M9/Energy	750	226	3.6	337	468	4	2	0.28
	D1/M1/Energy	710	206	3.7	306	456	17	20	0.27
E4	D1/M9/Ward	242	76	3.2	104	140	3	2	0.28
	D1/M9/Energy	229	71	3.4	103	135	5	1	0.29
	D1/M1/Energy	157	52	5.3	94	182	13	2	0.25
E5	D1/M9/Ward	808	239	5.2	565	672	69	0	0.35
	D1/M9/Energy	845	248	4.9	571	648	63	0	0.36
	D1/M1/Energy	631	212	6.5	549	830	85	0	0.32
E6	D1/M9/Ward	74	22	9.9	119	98	17	0	0.46
	D1/M9/Energy	4	1	168.0	100	68	36	0	0.49
	D1/M1/Energy	2	1	258.0	90	168	46	0	0.30
E7	D1/M9/Ward	352	106	4.2	210	231	14	1	0.35
	D1/M9/Energy	332	103	4.2	206	226	18	1	0.36
	D1/M1/Energy	285	92	5.1	197	272	26	2	0.32
E8	D1/M9/Ward	968	216	3.4	404	322	0	9	0.37
	D1/M9/Energy	922	218	3.5	408	355	0	5	0.36
	D1/M1/Energy	773	201	4.0	377	420	20	16	0.33
E9	D1/M9/Ward	528	106	3.3	134	211	0	5	0.23
	D1/M9/Energy	494	107	3.2	135	208	0	4	0.24
	D1/M1/Energy	434	102	3.9	134	268	1	4	0.22
E10	D1/M9/Ward	427	85	2.3	90	103	0	1	0.25
	D1/M9/Energy	427	85	2.3	90	104	0	1	0.25
	D1/M1/Energy	402	49	3.5	56	114	12	23	0.16
E11	D1/M9/Ward	22	6	14.0	63	22	6	0	0.64
	D1/M9/Energy	20	6	14.0	62	21	7	0	0.64
	D1/M1/Energy	29	9	9.9	64	25	5	0	0.61
E12	D1/M9/Ward	12	2	86.0	138	34	32	0	0.67
	D1/M9/Energy	12	2	86.0	138	34	32	0	0.67
	D1/M1/Energy	108	17	12.0	158	46	12	0	0.65
E13	D1/M9/Ward	232	19	4.6	39	48	0	0	0.35
	D1/M9/Energy	256	18	4.4	39	40	0	0	0.37
	D1/M1/Energy	272	22	3.6	39	41	0	0	0.36
E14	D1/M9/Ward	130	32	8.4	154	114	12	0	0.48
	D1/M9/Energy	133	35	7.7	154	114	12	0	0.48
	D1/M1/Energy	78	21	11.0	140	86	26	0	0.51

When the experiments are repeated with seven datasets (i.e., the six non-5-HT datasets and E5) the important conclusions from above remain largely unchanged. Thus, the energy clustering method is significantly superior to Ward's method: the mean ranks for energy and Ward's from Table 5 are 1.86 and 1.14, respectively, whereas with just the seven datasets, the corresponding figures are 1.85 and 1.15. The best-performing distance coefficient is again M9 with very little change in the significant rankings in the two cases: when all datasets are used the ordering is M9 > M10 > M7 > M8 > M2 > M4 > M1 > M5 > M3 > M6, and when just the seven datasets are used the ordering is M9 > M10 > M8 > M7 > M2 > M4 > M5 > M1 > M3 > M6. The best-performing descriptor is now D5: when all datasets are used the ordering is D1 > D2 > D5 > D3 > D4, and when just the seven datasets are used the ordering D5 > D2 > D1 > D3 > D4. That said, the differences between D1, D2 and D5 in these later experiments are very small, and an analysis using just these three descriptors, i.e., discarding the D3 and D4 data, reveals no significant difference between them. Given that D5 now ranks above D1, it is hardly surprising that D5 figures more prominently than previously when all the combinations are considered. Consider the top-10 combinations: when all datasets are used, both D1 and D2 appear 5 times; when just the seven datasets are used, D5 appears 7 times and D2 3 times. Consider the next-10 combinations: the same result is obtained whether all or just seven of the datasets are used, with D1 appearing 5 times, D2 4 times and D5 once.

We hence conclude that the choice of datasets has had only a slight effect on the overall results, and none on our conclusions regarding the effectiveness of the energy method and of coefficient M9.

4.8. Additional comparison of distance coefficients

As its title makes clear, the principal focus of this paper has been the comparison of the Ward and energy clustering methods.

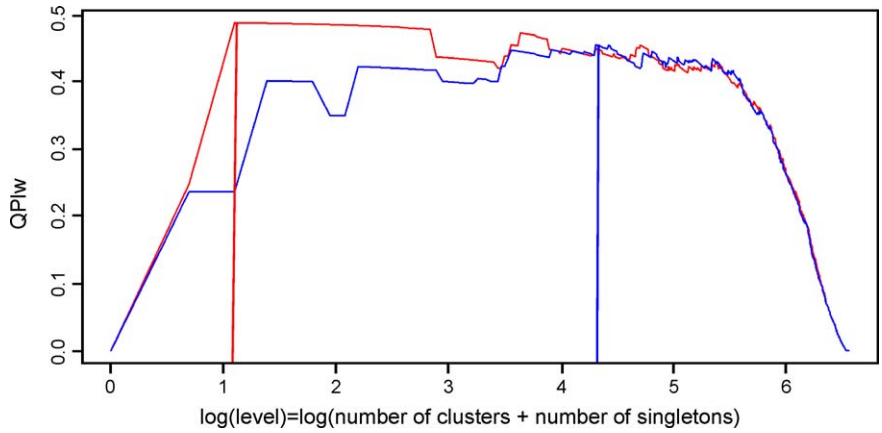

Fig. 2. QPI_w curves for E6 with D1, M9, and Energy (red) or Ward (blue) clustering. The vertical lines indicates the best partition for each method.

However, the consistently high level of effectiveness of the M9 distance coefficient encouraged us to investigate the effectiveness of this coefficient in a different application area, that of simulated virtual screening experiments. Specifically, we used the M1, M4 and M7–M10 coefficients (i.e., all but the standardised versions of M1 and M4, none of which performed particularly well in the experiments discussed thus far) for similarity-based virtual screening of the *MDL Drug Data Report* (MDDR) and *World of Molecular Bioactivity* (WOMBAT) databases. For comparison with the M1, M4 and M7–M10 coefficients, we also used the cosine (COS) and Tanimoto (TAN) coefficients [31]:

$$\mathrm{COS} = \frac{\sum_{k=1}^{p} X_{ik}X_{jk}}{\sqrt{\sum_{k=1}^{p} X_{ik}^2 \sum_{k=1}^{p} X_{jk}^2}}$$

and

$$\mathrm{TAN} = \frac{\sum_{k=1}^{p} X_{ik}X_{jk}}{\sum_{k=1}^{p} X_{ik}^2 + \sum_{k=1}^{p} X_{jk}^2 - \sum_{k=1}^{p} X_{ik}^2 X_{jk}^2}$$

We also used the binary version of the Tanimoto coefficient (TAN-B), i.e., a fingerprint denoting merely the presence or absence of each fragment rather than its frequency of occurrence. This coefficient was included since it has been used previously in very many studies of virtual screening: indeed, it is arguably the standard coefficient for this purpose [32]. Note that the Tanimoto and Soergel coefficients are identical if processing binary data; however, they are not identical if processing non-binary data.

The version of MDDR used here contained 102,514 molecules, and searches were carried out for eleven classes of active compounds described by Hert et al. [33]; searches were also carried out for a set of 10 activity classes chosen to be as structurally homogeneous as possible (MDDR-HOM) and another set of 10 activity classes chosen to be as structurally heterogeneous as possible (MDDR-HET) [34]. The version of WOMBAT used here contained 138,127 molecules, and searches were carried out for the 14 activity classes described by Gardiner et al. [35]. The molecules were represented by ECFC_6 fingerprints, analogous to the ECFC_4 fingerprints (i.e. D1) used in some of the clustering experiments but here encoding circular substructures of diameter six bonds. Twenty molecules were chosen from each activity class in turn and used as a reference structure for a similarity search, in which all the molecules in a database were ranked in decreasing similarity order and the top-1% of the molecules returned as the output of the search (the relative performance of the different coefficients – as discussed below – was unaffected when an alternative cut-off of 5% was used). Search effectiveness was measured by the *recall*, i.e., the percentage of the active molecules retrieved above the 1% cut-off. The mean recall was averaged over all of the reference molecules for each activity class, and then these mean values averaged over all of the activity classes for a dataset.

The results of the screening experiments, in terms of the overall mean recall for each coefficient for each of the four datasets (MDDR, MDDR-HET, MDDR-HOM and WOMBAT), are shown in Table 7. It will be seen that the coefficients M7–M9 provide a high level of performance across all of the datasets, and that M1 (and M4) performs well for the structurally diverse MDDR-HET. If the sets of recall values for the activity classes are analysed using the *W* test, then statistically significant correlations are obtained in all cases ($p < 0.001$ for all bar MDDR-HET, where $p < = 0.05$). It is hence appropriate to rank the coefficients, as shown in Table 8 where it will be seen that M9 gives the best overall screening performance. It is noteworthy that it is superior to TAN, COS and TAN-B, all of which have been used in previous studies of similarity searching. We note also that TAN-B is superior to TAN, despite the fact that several previous studies have suggested that weighted-searching is superior to binary searching [10,36,37];

Table 7
Percentage recall averaged over 20 searches for each activity class and over the eleven activity classes for MDDR, the 10 activity classes for MDDR-HOM and MDDR-HET, and the 14 activity classes for WOMBAT. The recall is calculated using the top-1% of the databases when ranked using the coefficient in the left-hand column.

Coefficient	All	MDDR	MDDR-HOM	MDDR-HET	WOMBAT
M1	28.74	16.85	62.27	15.60	23.52
M4	28.16	14.86	64.98	15.18	21.58
M7	37.20	22.84	84.84	12.54	32.06
M8	37.39	22.98	85.09	12.61	32.34
M9	37.16	23.06	84.65	13.05	31.53
M10	34.99	21.82	80.13	11.78	29.68
COS	29.99	18.39	71.08	8.10	25.38
TAN	31.78	19.27	73.11	11.86	26.32
TAN-B	36.66	22.39	84.47	12.07	31.29

Table 8
Kendall *W* analysis for percentage recall figures listed in Table 7. The figures listed for each coefficient is the rank of that coefficient when the coefficients are ranked in decreasing order of screening effectiveness for each dataset.

Coefficient	All	MDDR	MDDR-HOM	MDDR-HET	WOMBAT
M1	7.09	7.27	8.60	4.20	7.93
M4	7.68	8.50	8.10	5.30	8.43
M7	3.06	2.68	2.75	4.25	2.71
M8	2.77	2.82	2.45	3.90	2.14
M9	2.71	2.05	2.90	3.35	2.64
M10	4.62	4.73	4.30	5.20	4.36
COS	6.98	6.73	6.40	7.90	6.93
TAN	5.78	5.91	6.40	5.10	5.71
TAN-B	4.32	4.32	3.10	5.80	4.14
W	0.51	0.68	0.79	0.24	0.74

however, our results here are in accord with recent work by Bender et al. [38].

These experiments hence suggest that M9, the Soergel coefficient, is well suited for the processing of molecular fingerprints that encode fragments' frequencies of occurrence; it is already known (since it is then identical to TAN-B) to be well suited for processing their binary equivalents.

5. Conclusions

Clustering sets of chemical structures represented by fragment substructures is a common and important application in chemoinformatics. One of the most extensively used clustering methods is the minimum variance method first described by Ward. Székely and Rizzo have recently described a class of clustering methods that includes Ward's method as a limiting case. They suggest that one specific member of this class of methods, called the energy method, may be superior to Ward's method in some cases, and in this paper we have described an extensive series of experiments that demonstrates that this is certainly the case when the two methods are used to generate chemical classifications. Specifically, our results show that the energy method outperforms Ward's method across a range of types of substructural descriptor, of dataset and of distance metric. Our results also show the consistently high level of performance of the Soergel distance coefficient. We hence conclude that this method and this distance coefficient merit consideration in future studies of the classification of chemical structure databases.

Acknowledgements

We thank the following: Conseil Regional de Basse Normandie for funding; Jean-Charles Delarue and John Holliday for helpful advice; Sylvain Rault, Sunset Molecular Discovery LLC and Symyx Technologies Inc. for provision of the ATB1, WOMBAT and MDDR data, respectively; and Accelrys Software Inc., Chemaxon, the Royal Society, Tripos International Inc. and the Wolfson Foundation for software and laboratory support.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2009.06.006.

References

[1] B.S. Everitt, S. Landau, M. Leese, Cluster Analysis, Edward Arnold, London, 2001.
[2] P.H.A. Sneath, R.R. Sokal, Numerical Taxonomy, W.H. Freeman, San Francisco, 1973.
[3] J.R. Kettenring, The practice of cluster analysis, J. Classif. 23 (2006) 3–30.
[4] P. Willett, Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, 1987.
[5] G.M. Downs, J.M. Barnard, Clustering methods and their uses in computational chemistry, Rev. Comput. Chem. 18 (2002) 1–40.
[6] J.W. Raymond, C.J. Blankley, P. Willett, Comparison of chemical clustering methods using graph-based and fingerprint-based similarity measures, J. Mol. Graph. Model. 21 (2003) 421–433.
[7] G.W. Adamson, J.A. Bush, A method for the automatic classification of chemical structures, Inf. Stor. Retriev. 9 (1973) 561–568.
[8] P. Willett, V. Winterman, D. Bawden, Implementation of non-hierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output, J. Chem. Inf. Comput. Sci. 26 (1986) 109–118.
[9] M.S. Lajiness, Dissimilarity-based compound selection techniques, Perspect. Drug Discov. Design 7/8 (1997) 65–84.
[10] 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.
[11] N.E. Shemetulskis, J.B. Dunbar, B.W. Dunbar, D.W. Moreland, C. Humblet, Enhancing the diversity of a corporate database using chemical database clustering and analysis, J. Comput.-Aid. Mol. Design 9 (1995) 407–416.
[12] J.H. Ward, Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc. 58 (1963) 236–244.
[13] 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.
[14] 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.
[15] F. Murtagh, Multidimensional Clustering Algorithms, Physica Verlag, Vienna, 1985.
[16] A. Boecker, S. Derksen, E. Schmidt, A. Teckentrup, G. Schneider, A hierarchical clustering approach for large compound libraries, J. Chem. Inf. Model. 45 (2005) 807–815.
[17] A. Schuffenhauer, N. Brown, P. Ertl, J.L. Jenkins, P. Selzer, J. Hamon, Clustering and rule-based classifications of chemical structures evaluated in the biological activity space, J. Chem. Inf. Model. 47 (2007) 325–336.
[18] G.J. Szekely, M.L. Rizzo, Hierarchical clustering via joint between-within distances: extending Ward's minimum variance method, J. Classif. 22 (2005) 151–183.
[19] T. Varin, N. Saettel, J. Villain, A. Lesnard, F. Dauphin, R. Bureau, S. Rault, 3D pharmacophore, hierarchical methods, and 5-HT_4 receptor binding data, J. Enzyme Inhibit. Med. Chem. 23 (2008) 593–603.
[20] G.N. Lance, W.T. Williams, A general theory of classificatory sorting strategies. I. Hierarchical systems, Comput. J. 9 (1967) 373–380.
[21] F. Murtagh, A survey of recent advances in hierarchical clustering algorithms, Comput. J. 26 (1983) 354–359.
[22] L. Kaufman, P.J. Rousseuw, Finding Groups in Data: an Introduction to Cluster Analysis, Wiley, New York, 1990.
[23] G.N. Lance, W.T. Williams, Mixed-data classificatory programs. I. Agglomerative systems, Aust. Comput. J. 1 (1967) 15–20.
[24] P. Willett, Similarity methods in chemoinformatics, Ann. Rev. Inform. Sci. Technol. 43 (2009) 3–71.
[25] J.C. Gower, P. Legendre, Metric and Euclidean properties of dissimilarity coefficients, J. Classif. 5 (1986) 5–48.
[26] C.J. van, Rijsbergen, Information Retrieval, Butterworth, London, 1979.
[27] P. Willett, Recent trends in hierarchic document clustering: a critical review, Inf. Proc. Manage. 24 (1988) 577–597.
[28] S. Siegel, N.J. Castellan, Nonparametric Statistics for the Behavioural Sciences, McGraw-Hill, New York, 1988.
[29] D. Soergel, Mathematical analysis of documentation systems, Inf. Stor. Retriev. 3 (1967) 129–173.
[30] U. Fechner, G. Schneider, Evaluation of distance metrics for ligand-based similarity searching, ChemBioChem 5 (2004) 538–540.
[31] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
[32] P. Willett, Similarity-based virtual screening using 2D fingerprints, Drug Discov. Today 11 (2006) 1046–1053.
[33] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures, J. Chem. Inf. Comput. Sci. 44 (2004) 1177–1185.
[34] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, New methods for ligand-based virtual screening: use of data-fusion and machine-learning techniques to enhance the effectiveness of similarity searching, J. Chem. Inf. Model. 46 (2006) 462–470.
[35] E.J. Gardiner, V.J. Gillet, M. Haranczyk, J. Hert, J.D. Holliday, N. Malim, Y. Patel, P. Willett, Turbo similarity searching: Effect of fingerprint and dataset on virtual-screening performance, Stat. Anal. Data Mining, in press, doi:10.1002/sam.
[36] X. Chen, C.H. Reynolds, Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients, J. Chem. Inf. Comput. Sci. 42 (2002) 1407–1414.
[37] U. Fechner, J. Paetz, G. Schneider, Comparison of three holographic fingerprint descriptors and their binary counterparts, QSAR Comb. Sci. 24 (2005) 961–967.
[38] A. Bender, J.L. Jenkins, J. Scheiber, S.C.K. Sukuru, M. Glick, J.W. Davies, How similar are similarity searching methods? A principal components analysis of molecular descriptor space, J. Chem. Inf. Model. 49 (2009) 108–119.