J Comput Aided Mol Des (2014) 28:675–684
DOI 10.1007/s10822-014-9750-2

Weighted voting-based consensus clustering for chemical structure databases

Faisal Saeed · Ali Ahmed · Mohd Shahir Shamsir · Naomie Salim

Received: 27 January 2014 / Accepted: 7 May 2014 / Published online: 15 May 2014

Abstract The cluster-based compound selection is used in the lead identification process of drug discovery and design. Many clustering methods have been used for chemical databases, but there is no clustering method that can obtain the best results under all circumstances. However, little attention has been focused on the use of combination methods for chemical structure clustering, which is known as consensus clustering. Recently, consensus clustering has been used in many areas including bioinformatics, machine learning and information theory. This process can improve the robustness, stability, consistency and novelty of clustering. For chemical databases, different consensus clustering methods have been used including the co-association matrix-based, graph-based, hypergraph-based and voting-based methods. In this paper, a weighted cumulative voting-based aggregation algorithm (W-CVAA) was developed. The MDL Drug Data Report (MDDR) benchmark chemical dataset was used in the experiments and represented by the AlogP and ECPF_4 descriptors. The results from the clustering methods were evaluated by the ability of the clustering to separate biologically active molecules in each cluster from inactive ones using different criteria, and the effectiveness of the consensus clustering was compared to that of Ward's method, which is the current standard clustering method in chemoinformatics. This study indicated that weighted voting-based consensus clustering can overcome the limitations of the existing voting-based methods and improve the effectiveness of combining multiple clusterings of chemical structures.

Keywords Chemical dataset · Compound selection · Consensus clustering · Cumulative voting · Weighting schemes

F. Saeed (✉) · N. Salim
Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
e-mail: alsamet.faisal@gmail.com

F. Saeed
Information Technology Department, Sanhan Community College, Sana'a, Yemen

A. Ahmed
Faculty of Engineering, Karary University, Khartoum 12304, Sudan

M. S. Shamsir
Faculty of Faculty of Bioscience and Bioengineering, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Introduction

In simple terms, chemoinformatics is defined as the use of computer and information technology to solve a variety of problems in the field of chemistry. Different names have been used to describe it, such as cheminformatics and chemical informatics. The term chemoinformatics was initially defined by Brown [1, 2] as the collection, representation and organisation of chemical data to create chemical information, which is applied to generate chemical knowledge. The main purpose of using chemoinformatics is to make efficient decisions for the drug discovery and design process, especially in the lead identification and optimisation process.

Drug discovery is a very complex and multi-disciplinary task, which consists of many stages that requires a long time because there are a huge number of compounds that should be screened for biological activity testing. One of these stages is known as lead identification, which is slow

and poses time and cost constraints. Therefore, there is a need for the rational and effective selection of only a subset of compounds from a combinatorial chemical library such that the maximum amount of information can be obtained by testing the minimum number of chemical compounds.

The process of selecting a subset of compounds is known as compound selection. The four main approaches for compound selection include the cluster-based, optimisation-based, partition-based and dissimilarity-based approaches. The cluster-based compound selection, which is known as clustering, has been commonly used. The basic idea behind compound selection is the similar property principle [3], which states that structurally similar molecules will exhibit similar physicochemical and biological properties. Therefore, by using a clustering technique that clusters structurally similar compounds together, the application of this principle means that the selection and testing of representatives from each cluster should be sufficient to understand the structure–activity relationships of the whole set without testing all of the compounds in the dataset [4].

Clustering is a method for subdividing a number of objects into smaller clusters (groups) where each cluster exhibits a high degree of intra-cluster similarity and inter-cluster dissimilarity [5]. Different individual clustering techniques have been used for chemical structure clustering [6–12], and many studies [4, 9–11] have reported consistent performance of Ward's method [13]. Brown and Martin [4] reported that for compound selection, Ward's method provides the most effective clustering, and it is now the method of choice for clustering databases [10]. However, little attention has been focused on the use of consensus clustering for chemical structure clustering.

The two primary steps of consensus clustering are ensemble generation and consensus function. In the first step, many partitions are generated using different methods including the use of different object representations and different individual clustering methods [14]. In the consensus function, there are two main approaches including the object co-occurrence based and median partition based approaches. The co-association matrix-based, graph and hypergraph-based, and voting-based consensus clusterings are categorised in the first approach.

For clustering of chemical datasets, Chu et al. [15] reported that it is unlikely that any individual method will yield the best results under all circumstances, and they reported that the consensus clustering methods outperformed Ward's method. However, when the clustering is restricted to a single consensus method, no significant improvements were obtained because when the consensus clusterings were evaluated using only one single criterion, it was possible to find a consistent and effective consensus clustering method. However, when using different evaluation criteria, it is difficult to find one consensus clustering method that is best. In our previous work, Saeed et al. [16] used the graph and hypergraph-based consensus clustering methods for clustering of chemical datasets and concluded that these methods provided robust and stable clustering. In addition, Saeed et al. [17] combined multiple clusterings of chemical structures using the cumulative voting-based aggregation algorithm (CVAA) and reported that it improved the effectiveness of clustering. In addition, the adaptive and enhanced voting-based consensus methods were developed by Saeed et al. [18, 19] to obtain the final consensus partition. Unlike these consensus methods that have been previously used for chemical databases, the proposed weighted cumulative voting-based aggregation algorithm (W-CVAA) assigned different weights to the individual clustering methods (that are used to generate the ensemble) based on the entropy of each partition.

Methods

The cumulative voting-based aggregation algorithm (CVAA) was described by Ayed and Kamel [20, 21] and used by Saeed et al. [17] for clustering of chemical structures. In this study, the CVAA is enhanced by assigning different weights to the individual clustering methods that are used to generate the ensemble.

Let χ denote a set of N data objects (molecules in this context). The $N \times k$ matrix U represents the partitioning of χ into k clusters such that $\sum_{q=1}^{k} u_{jq} = 1$, for $\forall\, j$ (each row j represents one object and each column q represents one cluster in the partition). Let $u = \{U^i\}_{i=1}^{b}$ denote an ensemble where b is the size of the ensemble (b = 6 partitions in this paper). The first step of this algorithm is to search for optimal relabeling of each partition V^i (with k^i clusters) with respect to a representative partition U^0 (with k^0 clusters) and then to obtain a consensus partition denoted $\bar{U}$. The matrix of coefficients W^i, which is a $k^i \times k^0$ matrix of w_{lq}^i coefficients, is used for the partition relabeling process such that:

$$w_{lq}^i = \frac{n_{lq}^i}{n_l^i} \tag{1}$$

where n_{lq}^i is the number of objects that are assigned to both clusters c_l^i and c_q^0 and n_l^i is the number of objects that are assigned to cluster $c_l^i = |c_l^i|$.

Let $H(C)$ denote the Shannon entropy that measures the mutual information associated with partition C and is defined as [22]:

$$H(c) = -\sum_{c \in C} p(c) \log p(c) \quad (2)$$

and

$$P(c) = \frac{n}{N} \quad (3)$$

where n refers to the total number of molecules in a certain cluster and N refers to the total number of molecules in the dataset.

In this algorithm, different weights are assigned to the individual clustering methods that are used to generate the ensemble based on the mutual information associated with each method, which is measured by the entropy, as shown in Eq. 4. For each partition, each compound will be assigned to any cluster with a weight that is associated with that partition. Therefore, the weights of assigning compounds to clusters will not be the same using different individual clusterings.

$$\mathrm{T}^i = \frac{H(\mathrm{C}^i)}{\sum_{i=1}^{b} H(\mathrm{C}^i)}. \quad (4)$$

The W-CVAA is described as follows:

Weighted Cumulative Voting-based Aggregation Algorithm

1: Re-order the ensemble $\mathcal{U}$, such that:

U^i partitions are sorted as:

Weighted Average Distance partition> Ward partition>

Average Linkage partition > Single - Linkage partition > Complete Linkage partition > K-means partition.

2: Assign U^1 to U^0

3: for i =2 to b do

4: $\mathbf{W}^i = (\mathbf{U}^{i\,\mathrm{T}}\,\mathbf{U}^i)^{-1}\,\mathbf{U}^{i\,\mathrm{T}}\,\mathbf{U}^0$

5: $\mathbf{V}^i = \mathbf{U}^i\,\mathbf{W}^i$

6: $\mathbf{U}^0 = \frac{i-1}{i}\,\mathbf{U}^0 + \frac{\mathrm{T}^i}{i}\,\mathbf{V}^i$

7: end for

8: $\bar{\mathbf{U}} = \mathbf{U}^0$.

Experimental design

Consensus clustering experiments

A subset of the MDL Drug Data Report (MDDR) database [23] has been used for the consensus clustering experiments [15–19]. This subset included 8,294 homogeneous and heterogeneous active molecules (eleven activity classes). Table 1 shows a description of this dataset where each row contains an activity index, activity class, the number of active molecules, and the mean pairwise Tanimoto similarity for all pairs of molecules in the class. The dataset is represented by two descriptors that were generated by Scitegic's Pipeline Pilot [24]. These were 120-bit ALOGP and 1,024-bit ECFP_4 fingerprints (more details about these fingerprints in [25–27]).

Table 1 MDDR activity classes for the MDDR data set

Activity index	Activity class	Active molecules	Pairwise similarity Mean
31420	Renin inhibitors	1,130	0.290
71523	HIV protease inhibitors	750	0.198
37110	Thrombin inhibitors	803	0.180
31432	Angiotensin II AT1 antagonists	943	0.229
42731	Substance P antagonists	1,246	0.149
06233	5HT3 antagonist	752	0.140
06245	5HT reuptake inhibitors	359	0.122
07701	D2 antagonists	395	0.138
06235	5HT1A agonists	827	0.133
78374	Protein kinase C inhibitors	453	0.120
78331	Cyclooxygenase inhibitors	636	0.108

Each ensemble includes six partitions that were generated using six individual clustering methods. These methods included the hierarchical agglomerative methods (single, complete, average, weighted average distance, and Ward) and K-means clustering. In addition, each method clusters the dataset six times to generate six ensembles where each has a different number of clusters using thresholds of 500, 600, 700, 800, 900 and 1,000. This ensemble generation process is carried out for AlogP and ECFP_4.

In the second step of the consensus clustering, the ensembles were combined using nine consensus clustering methods including the proposed method. These methods include four co-association matrix based (average, single, complete, weighted) [19], two graph- and hypergraph-based [16] and three voting-based (CVAA, A-CVAA and W-CVAA) [17, 18] methods.

Performance evaluation

The performance evaluation of the chemical structure clustering was performed by measuring the ability of the clustering to separate active molecules from inactive ones using the Quality Partition Index (QPI) measure [28], F-measure [29] and Active Cluster Subset (ACS) measure [4].

For the QPI measure, let p be the number of active molecules in the active clusters (an active cluster is defined as a non-singleton cluster for which the percentage of active molecules in the cluster is greater than the percentage of active molecules in the dataset as a whole), q be the number of inactive molecules in the active clusters, r be the number of active molecules in the inactive clusters (i.e.,

clusters that are not active clusters) and s be the number of singleton active molecules. The QPI is defined as [10]:

$$QPI = \frac{p}{p+q+r+s} \quad (5)$$

The QPI calculation was performed for the eleven activity classes, and the results were averaged. When the value of the QPI measure is higher, the performance of a clustering method is better.

For the F-measure, if the number of compounds in a cluster is n and a of these compounds are active, the total number of compounds with the chosen activity is A. Then, the precision (P), recall (R) and F-measure value for that cluster are [15]:

$$P = \frac{a}{n} \quad (6)$$

$$R = \frac{a}{A} \quad (7)$$

$$F = \frac{2PR}{P+R} \quad (8)$$

The calculation of P, R, and F was performed for each cluster, and the maximum F value across all of the clusters was considered to be the F-measure. Then, this process was carried out for the eleven activity classes, and the results were averaged. When the value of the F-measure is higher, the performance of a clustering method is better.

The third evaluation criterion is the active cluster subset (ACS) measure that was proposed by Brown and Martin [4] and used for evaluating the clustering methods based on the ability of clustering to separate active and inactive molecules into different sets of clusters. Therefore, the active cluster subset includes compounds from all clusters in which the percentage of active molecules is greater than the percentage of active molecules in the dataset as a whole. The difference in the proportion of active molecules in an active cluster subset (P_a) to the proportion of active compounds in the entire dataset (P_0) can provide a measure of the separation of active molecules from inactive ones. The proportion of active compounds in an active cluster subset (P_a) was calculated as:

$$P_a = \frac{No.\ of\ Active\ Compounds\ in\ the\ Active\ Cluster\ Subset}{Total\ No.\ of\ Compounds\ in\ the\ Active\ Cluster\ Subset} \quad (9)$$

The proportion of active compounds in the whole dataset (P_0) was calculated as:

$$P_0 = \frac{No.\ of\ Active\ Compounds\ in\ the\ Dataset}{Total\ No.\ of\ Compounds\ in\ the\ Dataset}. \quad (10)$$

For the overall performance, this process was carried out for the eleven activity classes, and the results were averaged. When the value of the ACS is higher, the performance of a clustering method is better.

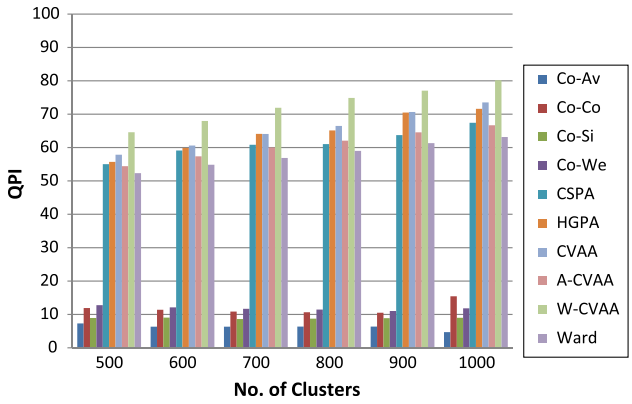

Fig. 1 Effectiveness of the weighted voting-based consensus clustering for the ALOGP descriptor using the QPI measure

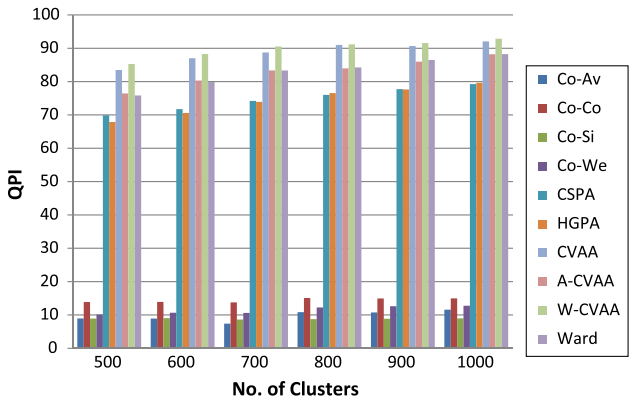

Fig. 2 Effectiveness of the weighted voting-based consensus clustering for the ECFP_4 fingerprint using the QPI measure

Finally, the W-CVAA consensus clustering results were compared to the those of Ward's method and other consensus clustering methods, such as the co-association based methods, CSPA, HGPA, CVAA and A-CVAA.

Results and discussion

The voting-based (CVAA, A-CVAA and W-CVAA), co-association matrix-based (average, complete, single and weighted) and graph-based (CSPA and HGPA) consensus clustering methods were independently used to combine the multiple partitions in each ensemble.

Figures 1 and 2 show the effectiveness of the proposed method using the QPI measure compared to the other consensus clustering and Ward's methods. By visual inspection of Figs. 1 and 2, W-CVAA exhibited the best performance and outperformed the other consensus clusterings and Ward's method over all numbers of clusters for

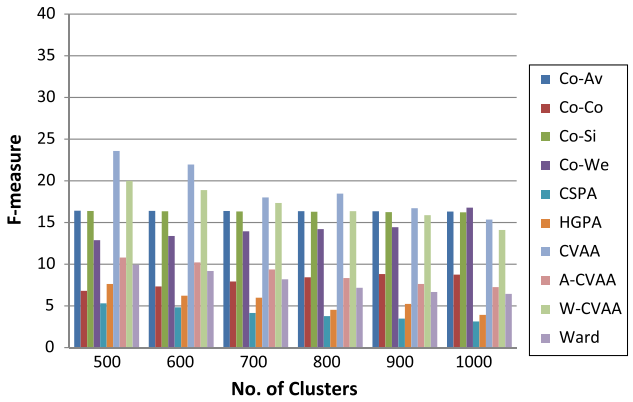

Fig. 3 Effectiveness of the weighted voting-based consensus clustering for the ALOGP descriptor using the F- measure

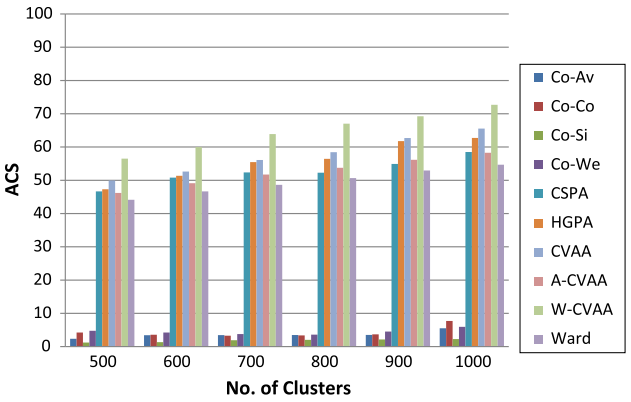

Fig. 5 Effectiveness of the weighted voting-based consensus clustering for the ALOGP descriptor using the ACS measure

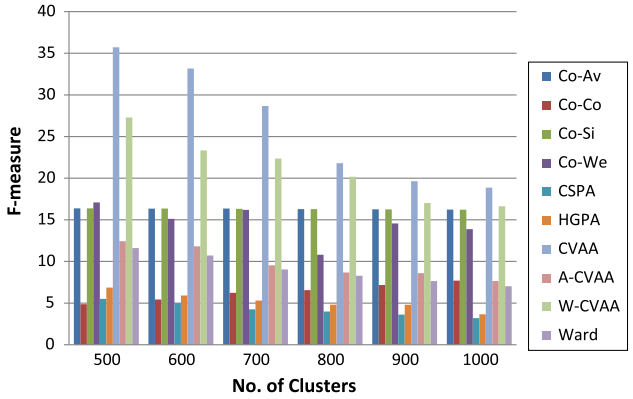

Fig. 4 Effectiveness of the weighted voting-based consensus clustering for the ECFP_4 fingerprint using the F- measure

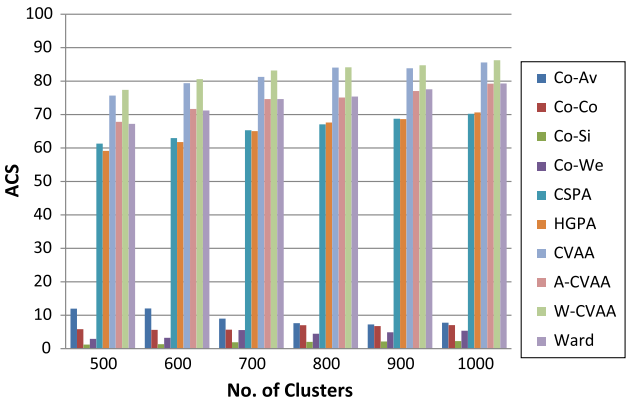

Fig. 6 Effectiveness of the weighted voting-based consensus clustering for the ECFP_4 fingerprint using the ACS measure

the ALOGP and ECFP_4 descriptors. The worst performance was observed for the co-association based matrix methods because they generate many active singletons where a few large clusters are found with a mix of active and inactive molecules leading to lower QPI values.

By visual inspection of the F-measure values in Figs. 3 and 4, W-CVAA exhibited the second best performance (after the CVAA method) and outperformed the other consensus clusterings and Ward's method over all numbers of clusters for both descriptors. Although W-CVAA exhibited inferior performance to that of CVAA using this evaluation criterion, W-CVAA was introduced to overcome the main limitations of CVAA, such as the order dependent problem and assigning similar weights to all partitions. In addition, the improvements that were observed with the co-association matrix-based methods are considered abnormal because they cluster the datasets into a few large clusters and many active singletons that lead to high F values.

Similar to the results obtained by the QPI measure, Figs. 5 and 6 indicated that when the ACS measure was used, W-CVAA exhibited the best performance and outperformed the other consensus clusterings and Ward's method over all numbers of clusters for both descriptors.

To provide a rigid assessment of the significance of the improvements obtained using W-CVAA, two statistical significance tests were employed including the *T* test and Kendall W test of concordance [30].

The *T* test procedure compares the means of two variables that represent the results of two clusterings (for each evaluation criterion such as the QPI measure) at different partition sizes (i.e., 500, 600, …, 1,000). A low significance value for the *T* test, which is typically < 0.05, indicates that a significant difference exists between two variables. Tables 2, 3 and 4 list the mean value, standard deviation, standard error mean and significance values for the pairs of the QPI, F and ACS values of (W-CVAA-Co-Average), (W-CVAA-Co-Complete), (W-CVAA-Co-Single), (W-CVAA-Co-Weighted), (W-CVAA-CSPA), (W-CVAA - HGPA), (W-CVAA-CVAA), (W-CVAA-A-CVAA) and (W-CVAA-Ward) compared in the Paired Samples *T* Test procedure.

Table 2 *T* test statistical significance testing using the QPI measure for the weighted voting-based consensus clustering of the MDDR dataset

	Paired differences					
	Mean	Std. deviation	Std. error mean	95 % Confidence interval of the difference		Sig. (2-tailed)
				Lower	Upper	
(a) ALOGP:						
Pair 1: W-CVAA-Co-Average	67.146	5.957	2.432	60.895	73.398	0.000
Pair 2: W-CVAA-Co-Complete	61.583	4.823	1.969	56.521	66.644	0.000
Pair 3: W-CVAA-Co-Single	64.510	5.293	2.160	58.955	70.064	0.000
Pair 4: W-CVAA-Co-Weighted	61.556	5.731	2.340	55.541	67.571	0.000
Pair 5: W-CVAA-CSPA	12.173	1.701	0.694	10.387	13.959	0.000
Pair 6: W-CVAA-HGPA	8.885	1.171	0.478	7.655	10.114	0.000
Pair 7: W-CVAA-CVAA	7.831	0.762	0.311	7.031	8.632	0.000
Pair 8: W-CVAA-A-CVAA	12.513	0.875	0.357	11.594	13.432	0.000
Pair 9: W-CVAA-Ward	15.441	1.310	0.535	14.066	16.817	0.000
(b) ECFP_4:						
Pair 10: W-CVAA-Co-Average	80.193	2.205	0.900	77.878	82.508	0.000
Pair 11: W-CVAA-Co-Complete	75.503	2.314	0.944	73.074	77.932	0.000
Pair 12: W-CVAA-Co-Single	81.068	2.788	1.138	78.141	83.994	0.000
Pair 13: W-CVAA-Co-Weighted	78.423	1.802	0.735	76.531	80.314	0.000
Pair 14: W-CVAA-CSPA	15.126	1.111	0.453	13.960	16.293	0.000
Pair 15: W-CVAA-HGPA	15.553	1.845	0.753	13.616	17.490	0.000
Pair 16: W-CVAA-CVAA	1.113	0.599	0.244	0.484	1.741	0.006
Pair 17: W-CVAA-A-CVAA	6.886	1.484	0.606	5.328	8.444	0.000
Pair 18: W-CVAA-Ward	6.925	1.818	0.742	5.016	8.833	0.000

The performance of the weighted voting-based consensus method was superior to the other consensus clusterings and Ward's methods. For example, W-CVAA significantly outperformed the standard clustering method and Ward's method using all of the evaluation criteria (QPI, F and ACS) for both descriptors (ALOGP and ECFP_4). In addition, W-CVAA significantly outperformed all of the other consensus clustering methods using the QPI and ACS measures for both descriptors (for the F-measure, W-CVAA significantly outperformed all of the other consensus clustering methods except the CVAA method, as discussed earlier).

The second statistical test is the Kendall W test of concordance, which was used for ranking the clustering methods (based on the effectiveness of each method). Here, the values of each evaluation measure (e.g., QPI) for all numbers of clusters (i.e., 500, 600, … , 1,000) were considered as a judge ranking (raters) of the clustering methods (ranked objects), i.e., k = 6 and N = 10, in decreasing order. The outputs of the Kendall W test are the Kendall coefficient (W), Chi Square (χ^2) and the significance level (p value). Therefore, the p value is considered significant and can provide an overall ranking if $p < 0.05$ (and the critical value for Chi square χ^2 at $p = 0.05$ for nine degree of freedom is 16.92).

The results of the Kendall analysis are reported in Tables 5, 6 and 7. For example, the results in Table 5 indicated that the values of the Kendall coefficient (for the AlogP and ECFP_4 using the QPI measure) are significant ($p < 0.05$, $\chi^2 > 16.92$), and the performance of W-CVAA significantly outperformed all of the other methods. The overall rankings for ALOGP are as follows: W-CVAA > CVAA > HGPA > CSPA > A-CVAA > Ward > Co-Weighted > Co-Complete > Co-Single > Co_Average, and for ECFP_4 is: W-CVAA > CVAA > A-CVAA = Ward > CSPA > HGPA > Co-Complete > Co-Weighted > Co_Average > Co-Single.

For the F-measure, the results in Table 6 indicate that the value of the Kendall coefficient is significant ($p < 0.05$, $\chi^2 > 16.92$), and W-CVAA significantly outperformed Ward's method for both descriptors. The overall ranking for ALOGP is as follows: CVAA > W-CVAA > Co_Average > Co-Single > Co-Weighted > A-CVAA > Co-Complete > Ward > HGPA > CSPA, and for ECFP_4 is: CVAA > W-CVAA > Co_Average > Co-Single > Co-Weighted > A-CVAA > Ward > Co-Complete > HGPA > CSPA.

In addition, the results in Table 7 indicate that the values of the Kendall coefficient (for the AlogP and ECFP_4 using the ASC measures) are significant ($p < 0.05$,

Table 3 *T* test statistical significance testing using the F-measure for the weighted voting-based consensus clustering of the MDDR dataset

	Paired differences					
	Mean	Std. deviation	Std. error mean	95 % Confidence interval of the difference		Sig. (2-tailed)
				Lower	Upper	
(a) ALOGP:						
Pair 1: W-CVAA-Co-Average	2.390	3.085	1.259	−0.848	5.6282	0.116
Pair 2: W-CVAA-Co-Complete	10.748	3.915	1.598	6.638	14.8578	0.001
Pair 3: W-CVAA-Co-Single	2.456	3.062	1.250	−0.757	5.6707	0.107
Pair 4: W-CVAA-Co-Weighted	4.485	4.367	1.782	−0.098	9.0682	0.053
Pair 5: W-CVAA-CSPA	14.648	2.301	0.939	12.233	17.0635	0.000
Pair 6: W-CVAA-HGPA	13.171	1.983	0.809	11.089	15.2536	0.000
Pair 7: W-CVAA-CVAA	−0.255	0.381	0.155	−0.655	0.1454	0.163
Pair 8: W-CVAA-A-CVAA	9.825	1.757	0.717	7.980	11.6697	0.000
Pair 9: W-CVAA-Ward	10.826	1.765	0.720	8.974	12.6789	0.000
(b) ECFP_4:						
Pair 10: W-CVAA-Co-Average	4.400	3.952	1.613	0.252	8.547	0.041
Pair 11: W-CVAA-Co-Complete	14.385	5.043	2.059	9.091	19.678	0.001
Pair 12: W-CVAA-Co-Single	4.408	3.949	1.612	0.264	8.552	0.041
Pair 13: W-CVAA-Co-Weighted	6.100	3.319	1.355	2.616	9.583	0.006
Pair 14: W-CVAA-CSPA	16.450	3.182	1.299	13.110	19.789	0.000
Pair 15: W-CVAA-HGPA	15.483	3.003	1.226	12.331	18.635	0.000
Pair 16: W-CVAA-CVAA	−5.595	3.577	1.460	−9.349	−1.840	0.012
Pair 17: W-CVAA-A-CVAA	10.925	2.364	0.965	8.443	13.406	0.000
Pair 18: W-CVAA-Ward	11.658	2.336	0.954	9.206	14.110	0.000

Table 4 *T* test statistical significance testing using the ACS measure for the weighted voting-based consensus clustering of the MDDR dataset

	Paired differences					
	Mean	Std. deviation	Std. error mean	95 % Confidence interval of the difference		Sig. (2-tailed)
				Lower	Upper	
a) ALOGP:						
Pair 1: W-CVAA-Co-Average	62.188	4.764	1.945	57.188	67.188	0.000
Pair 2: W-CVAA-Co-Complete	61.510	4.811	1.964	56.460	66.559	0.000
Pair 3: W-CVAA-Co-Single	63.986	5.147	2.101	58.584	69.388	0.000
Pair 4: W-CVAA-Co-Weighted	61.338	5.241	2.139	55.837	66.838	0.000
Pair 5: W-CVAA-CSPA	13.226	2.187	0.892	10.931	15.521	0.000
Pair 6: W-CVAA-HGPA	9.960	1.086	0.443	8.819	11.100	0.000
Pair 7: W-CVAA-CVAA	8.261	0.545	0.222	7.689	8.833	0.000
Pair 8: W-CVAA-A-CVAA	13.273	1.275	0.520	11.934	14.611	0.000
Pair 9: W-CVAA-Ward	16.191	1.733	0.707	14.372	18.011	0.000
(b) ECFP_4:						
Pair 10: W-CVAA-Co-Average	73.565	5.396	2.203	67.902	79.228	0.000
Pair 11: W-CVAA-Co-Complete	76.511	2.844	1.161	73.526	79.496	0.000
Pair 12: W-CVAA-Co-Single	81.006	2.914	1.189	77.948	84.065	0.000
Pair 13: W-CVAA-Co-Weighted	78.421	2.473	1.009	75.825	81.017	0.000
Pair 14: W-CVAA-CSPA	16.891	0.683	0.279	16.174	17.609	0.000
Pair 15: W-CVAA-HGPA	17.350	1.149	0.469	16.144	18.555	0.000
Pair 16: W-CVAA-CVAA	1.201	0.566	0.231	0.607	1.796	0.003
Pair 17: W-CVAA-A-CVAA	8.575	0.807	0.329	7.727	9.422	0.000
Pair 18: W-CVAA-Ward	8.615	1.112	0.454	7.447	9.782	0.000

Table 5 Rankings of clustering methods based on Kendall W test results using the MDDR dataset (the QPI measure)

FP	W	χ^2	*p*	Ranks									
AlogP	0.99	53.35	0.000	Technique	C-Average	Co-Complete	Co-Single	Co-Weighted	CSPA	HGPA	CVAA	A-CVAA	W-CVAA	Ward
				Mean Ranks	10.0	7.8	9.0	7.2	4.3	2.8	2.2	4.7	1.0	6.0
ECFP4	0.98	53.09	0.000	Technique	C-Average	Co-Complete	Co-Single	Co-Weighted	CSPA	HGPA	CVAA	A-CVAA	W-CVAA	Ward
				Mean Ranks	9.3	7.0	9.7	8.0	5.3	5.7	2.0	3.5	1.0	3.5

Table 6 Rankings of clustering methods based on Kendall W test results using the MDDR dataset (the F-measure)

FP	W	χ^2	*p*	Ranks									
AlogP	0.91	49.16	0.000	Technique	C-Average	Co-Complete	Co-Single	Co-Weighted	CSPA	HGPA	CVAA	A-CVAA	W-CVAA	Ward
				Mean Ranks	2.8	7.1	3.8	4.3	10.0	8.8	1.6	6.5	2.3	7.5
ECFP4	0.96	51.89	0.000	Technique	C-Average	Co-Complete	Co-Single	Co-Weighted	CSPA	HGPA	CVAA	A-CVAA	W-CVAA	Ward
				Mean Ranks	3.3	8.2	4.0	4.7	9.8	8.7	1.0	6.1	2.1	7.1

Table 7 Rankings of clustering methods based on Kendall W test results using the MDDR dataset (the ACS measure)

FP	W	χ^2	*p*	Ranks										
AlogP	0.99	53.16	0.000	Technique	C-Average	Co-Complete	Co-Single	Co-Weighted	CSPA	HGPA	CVAA	A-CVAA	W-CVAA	Ward
				Mean Ranks	8.8	8.2	10.0	7.1	4.3	3.0	2.0	4.6	1.0	6.0
ECFP4	0.99	53.38	0.000	Technique	C-Average	Co-Complete	Co-Single	Co-Weighted	CSPA	HGPA	CVAA	A-CVAA	W-CVAA	Ward
				Mean Ranks	7.0	8.0	10.0	9.0	5.3	5.7	2.0	3.5	1.0	3.5

$\chi^2 > 16.92$), and W-CVAA significantly outperformed all of the other methods. The overall ranking of the clustering methods using the ACS for ALOGP is as follows: W-CVAA > CVAA > HGPA > CSPA > A-CVAA > Ward > Co-Weighted > Co-Complete > Co_Average > Co-Single, and for ECFP_4 is: W-CVAA > CVAA > A-CVAA = Ward > CSPA > HGPA > Co_Average > Co-Complete > Co-Weighted > Co-Single.

Therefore, we can conclude that the weighted voting-based consensus method exhibited superior performance, and it significantly outperformed Ward's method using the QPI, F and ACS evaluation criteria for both descriptors. In these experiments, the leading group includes the voting based methods, which include W-CVAA, A-CVAA and CVAA. However, none of the co-association matrix based and graph based consensus clustering methods exhibited significant improvements over all of the evaluation criteria.

In addition, the computational complexity of W-CVAA is *O(N),* which is similar to other voting-based methods (A-CVAA and CVAA), which is more desirable than that of consensus clustering methods, such as the CSPA and co-association-based methods that have a complexity of $O(N^2)$ where *N* is the number of molecules in the chemical dataset.

Conclusion

The experimental results indicated the superior performance of weighted voting-based consensus clustering (W-CVAA) compared with Ward's method and other consensus clusterings. In addition, W-CVAA overcomes the main limitations of the CVAA by using a pre-defined order for the ensemble partitions (to solve the ordering dependent problem), and by assigning different weights for the individual clusterings that generate the ensemble. The evaluation of W-CVAA on different descriptors and using different criteria suggests that the W-CVAA consensus method can deliver significant improvements for the effectiveness of chemical structure clustering.

Acknowledgments Faisal Saeed is a Researcher of Universiti Teknologi Malaysia under the Post-Doctoral Fellowship Scheme for the project "Consensus Clustering Methods for Chemical Structure Databases" and this work is supported by Research Management Centre (RMC) at Universiti Teknologi Malaysia under Research University Grant Category (VOT Q.J130000.2528.07H89).

References

1. Brown FK (1998) Chemoinformatics what is it and how does it impact drug discovery. Annu Rep Med Chem 33:375–384
2. Brown FK (2005) Chemoinformatics-a ten year update. Curr Opin Drug Discov Devel 8(3):298

3. Johnson MA, Maggiora GM (1990) Concepts and application of molecular similarity. Wiley, New York
4. Brown RD, Martin YC (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. J Chem Inf Comput Sci 36:572–584
5. Everitt BS, Landau S, Leese M (2001) Cluster analysis, 4th edn. Edward Arnold, London
6. Down GM, Barnard JM (2003) Clustering methods and their uses in computational Chemistry. Rev Comput Chem 18:1–40
7. Holliday JD, Rodgers SL, Willett P, Chen MY, Mahfouf M, Lawson K, Mullier G (2004) Clustering files of chemical structures using the fuzzy k-means clustering method. J Chem Inf Comput Sci 44(3):894–902
8. Downs GM, Willett P, Fisanick W (1994) Similarity searching and clustering of chemical-structure databases using molecular property data. J Chem Inf Comput Sci 34:1094–1102
9. Willett P (1987) Similarity and clustering in chemical information systems. Research Studies Press, Letchworth
10. Varin T, Bureau R, Mueller C, Willett P (2009) Clustering files of 549 chemical structures using the Székely – Rizzo generalization of Ward's 550 method. J Mol Graph Model 28(2):187–195
11. Brown RD, Martin YC (1997) The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding. J Chem Inf Comput Sci 37(1):1–9
12. Willett P (2000) Textual and chemical information processing: different domains but similar algorithms. Inf Res 5(2). http://informationr.net/ir/5-2/paper69.html
13. Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244
14. Vega-Pons S, Ruiz-Schulcloper J (2011) A survey of clustering ensemble algorithms. Int J Pattern Recogn 25(3):337–372
15. Chu C-W, Holliday J, Willett P (2012) Combining multiple classifications of chemical structures using consensus clustering. Bioorg Med Chem 20(18):5366–5371
16. Saeed F, Salim N, Abdo A, Hentabli H (2013) Graph-based consensus clustering for combining multiple clusterings of chemical structures. Mol Inf 32(2):165–178
17. Saeed F, Salim N, Abdo A (2012) Voting-based consensus clustering for combining multiple clusterings of chemical structures. J Cheminform 4:37
18. Saeed F, Salim N, Abdo A (2013) Information theory and voting based consensus clustering for combining multiple clusterings of chemical structures. Mol Inform 32(7):591–598
19. Saeed F, Salim N, Abdo A (2013) Consensus methods for combining multiple clusterings of chemical structures. J Chem Inf Model 53(5):1026–1034
20. Ayad HG, Kamel MS (2008) Cumulative voting consensus method for partitions with a variable number of clusters. IEEE Trans Pattern Anal Mach Intell 30(1):160–173
21. Ayad HG, Kamel MS (2010) On voting-based consensus of cluster ensembles. Pattern Recogn 43:1943–1953
22. Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York
23. Sci Tegic Accelrys Inc., The MDL Drug Data Report (MDDR) database (2014). http ://www.accelrys.com/. Accessed 1 Jan 2014
24. Pilot P (2008) Accelrys Software Inc., San Diego
25. Ghose AK, Crippen GM (1986) Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure–activity relationships 1. Partition coefficients as a measure of hydrophobicity. J Comput Chem 7:565–577
26. Ghose AK, Viswanadhan VN, Wendoloski JJ (1998) Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. J Phys Chem A 102:3762–3772
27. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754
28. Varin T, Saettel N, Villain J, Lesnard A, Dauphin F, Bureau R, Rault SJ (2008) 3D Pharmacophore, hierarchical methods, and 5-HT4 receptor binding data. Enzyme Inhib Med Chem 23:593–603
29. Van Rijsbergen CJ (1979) Information retrieval. London, Butterworth
30. Siegel S, Castellan NJ (1988) Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York