Consensus Methods for Combining Multiple Clusterings of Chemical Structures

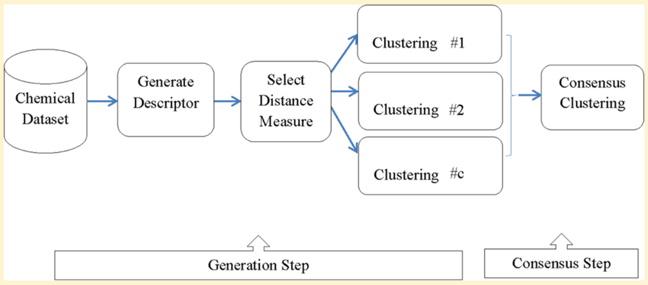
Faisal Saeed,*[†,‡] Naomie Salim,[†] and Ammar Abdo[§,∥]

[†]Faculty of Computing, Universiti Teknologi Malaysia, Malaysia

[‡]Information Technology Department, Sanhan Community College, Sana'a, Yemen

[§]Computer Science Department, Hodeidah University, Hodeidah, Yemen

[∥]LIFL UMR CNRS 8022 Université Lille 1 and INRIA Lille Nord Europe, 59655 Villeneuve d'Ascq cedex, France



ABSTRACT: The goal of consensus clustering methods is to find a consensus partition that optimally summarizes an ensemble and improves the quality of clustering compared with single clustering algorithms. In this paper, an enhanced voting-based consensus method was introduced and compared with other consensus clustering methods, including co-association-based, graph-based, and voting-based consensus methods. The MDDR and MUV data sets were used for the experiments and were represented by three 2D fingerprints: ALOGP, ECFP_4, and ECFC_4. The results were evaluated based on the ability of the clustering method to separate active from inactive molecules in each cluster using four criteria: F-measure, Quality Partition Index (QPI), Rand Index (RI), and Fowlkes−Mallows Index (FMI). The experiments suggest that the consensus methods can deliver significant improvements for the effectiveness of chemical structures clustering.

1. INTRODUCTION

The basic concept of consensus clustering is to cluster a set of objects by trying to find a clustering that agrees as much as possible with a number of individual clusterings. This approach has been used in various applications, including clustering of categorical data, improving clustering robustness, and detecting outliers.[1]

Consensus clustering involves two main steps: (i) ensemble generation and (ii) consensus function. Many mechanisms can be applied in the first step and include (i) different data representations, (ii) different individual clustering methods, (iii) different parameters initialization for clustering methods, and (iv) data resampling. The second step contains two main categories of approaches: object−co-occurrence-based and median-partition-based. In the first approach, the goal is to determine which cluster label must be associated with each object in the final consensus partition. To this end, the method analyses the number of times an object belongs to one cluster (for a voting process) or the number of times two objects belong together to the same cluster (to establish a similarity matrix). The consensus is obtained through a voting process among the objects such that each object should vote for the cluster to which it will belong in the consensus partition.[2] Co-association-based, graph-based, and voting-based consensus methods are examples of this approach. In the second consensus approach, the consensus partition is obtained by the solution of an optimization problem, which involves finding the median partition with respect to the cluster ensemble. The median partition is defined as the partition that maximizes the similarity with all partitions in the cluster ensemble.[2]

Consensus clustering methods have been successfully applied in many areas such as machine learning, applied statistics, pattern recognition, and bioinformatics.[3−8] Topchy et al.[4] and Fred and Jain[5] summarized the main advantages of using clustering, such as improving the robustness, consistency, novelty, and stability of individual clusterings. Moreover, consensus clustering is useful for distributed clustering. Over time, data sets sizes have grown rapidly because of the advances in technology and automated business processes. One way to cluster the distributed data sets is to use consensus clustering to cluster them locally in each site and subsequently combining the resulting clustering for all sites to obtain the consensus partition. Johnson and Kargupta[9] introduced a feasible approach for combining distributed agglomerative clusterings that generates local cluster models and subsequently combines them to generate the global cluster model of the data. Additionally, combining multiple clusterings provides a framework for knowledge reuse, which can be used to exploit the power of existing knowledge that is implicated in these clusterings.[3] Topchy et al.[10] proved this idea and showed that weak clustering algorithms can produce high-quality clusterings with the use of a proper consensus clustering method because different algorithms make different types of mistakes that can cancel each other out in the final aggregation.[1] Furthermore,

Received: September 18, 2012
Published: April 15, 2013

combining multiple clusterings provides a partition with lower sensitivity to noise and outliers.[4,5]

Many individual clusterings have been used in the literature of chemoinformatics.[11−17] The assessment of large compound-screening sets using various clustering methods and similarity metrics to explore the diversity and uniqueness of compounds was reported in ref 17. Brown and Martin[11] considered Ward's method as the most effective clustering method in cluster-based compound selection. In addition, Rivera-Borroto et al.[18] reported that the performance of clusterings depends on various factors, such as the molecular representation, mathematical method, algorithmical technique, and statistical distribution of the data. In their work, a comparison was performed among combinatorial methods to find the model that best fit the problem at hand, and the results were compared with the most classical algorithm, which is Ward's algorithm. However, it is known that no clustering method is capable of correctly finding the best clustering results for all data sets and applications,[2] and it is unlikely that any individual method will yield the best classification under all circumstances even if attention is restricted to a single type of application.[19] Therefore, the concept of combining different clustering results (known as consensus clustering) is considered as an alternative approach for improving the quality of the individual clustering algorithms.[2]

Consensus scoring has been successfully used for protein−ligand interactions[20] and for ligand-based virtual screening[21−23] by combining the results of two or more screening methods. In certain cases, the fused search may even perform better than the best individual screening method when averaged over large numbers of searches.[23] Therefore, the success of using consensus scoring for chemoinformatics has motivated researchers to apply consensus methods for combining multiple clustering of chemical data sets. Chu et al.[19] used selected consensus clustering methods on sets of chemical compounds represented by 2D fingerprints (ECFP_4) and concluded that consensus clustering can outperform Ward's method, which is the current standard clustering method for chemoinformatics applications. However, based on the implemented methods, this observation is not the case if the clustering is restricted to a single consensus method. In addition, Saeed et al.[24] examined the use of the graph-based consensus methods for combining multiple clusterings of the MDDR data set and concluded that they can improve the effectiveness of individual clusterings and provide robust and stable clustering. Moreover, Saeed et al.[25] used a cumulative voting-based aggregation algorithm (CVAA) for combining multiple clusterings of chemical structures and found that it could significantly improve the quality of clustering. In addition, an adaptive voting-based consensus method was used by Saeed et al.[26] to obtain a final consensus partition with a greater quantity of mutual information associated with its clusters. However, an enhanced voting-based consensus method was introduced in this paper and compared with other consensus clustering methods (four co-association-based, two graph-based, and two voting-based methods) and to the Ward's method to study the effectiveness of consensus methods for combining multiple clusterings of chemical structures.

2. METHODS

2.1. Co-Association-Based Consensus Methods.

The co-association consensus algorithms are developed based on the transformation of the set of clusterings into a similarity matrix. The generation of similarity matrix is described by Che et al.[19] as follows.

Given a data set containing N molecules and $NCLASS$ different clustering techniques applied on this data set, the consensus similarity matrix (CSM) is an $N \times N$ matrix, the JK-th element of which contains the number of times molecules J and K are found in the same cluster. The generation of the matrix is described in the simple pseudocode below in which all the elements of CSM have been initialized to zero and the counter I loops through each of the $NCLASS$ in turn.

```
FOR $I$:= 1 TO $NCLASS$
    FOR $J$:= 1 TO $N$-1
        FOR $K$:= $J$+1 TO $N$
            IF J and K are in the same cluster THEN $CSM[J,K]$:=$CSM[J,K]$ +1.
```

Then, each resulting element CSM $[J,K]$ can be converted to an interval between zero and unity by dividing by $NCLASS$.

After calculating the similarity matrix, four hierarchical agglomerative approaches are applied to cluster the matrix and obtain the final consensus partition, i.e., the single linkage, complete linkage, average linkage, and weighted average distance methods.

2.2. Graph-Based Consensus Methods.

Two graph-based consensus clustering algorithms proposed by Strehl and Ghosh[3] are used to obtain the consensus partition from the generated ensembles. The algorithms are developed based on transformation of the set of clusterings into a hyper-graph representation. The first is the cluster-based similarity partitioning algorithm (CSPA) in which a clustering signifies a relationship between objects in the same cluster and can thus be used to establish a measure of pairwise similarity. The consensus similarity matrix (CSM) is generated as described in the above section, and the matrix is subsequently used to recluster the objects using any reasonable similarity-based clustering algorithm. In this work, we view the similarity matrix as a graph (vertex = object, edge weight = similarity) and cluster it using the graph partitioning algorithm METIS.[27] The second algorithm is the hyper-graph partitioning algorithm (HGPA) in which the cluster ensemble problem is formulated as a partitioning of the hyper-graph by cutting a minimal number of hyper-edges. The HGPA partitions the hyper-graph directly by removing the lower number of hyper-edges. All hyper-edges have the same weight and are searched by cutting the minimum possible number of hyper-edges that partition the hyper-graph into k components of approximately the same dimension. Essentially, the cluster ensemble problem is posed as a partitioning problem of a suitably defined hyper-graph in which the hyper-edges represent clusters.[3] The hyper-graph partitioning package HMETIS[28] was used for the implementation of this method.

2.3. Voting-Based Consensus Methods.

The cumulative voting-based aggregation algorithm consists of two steps. The first step is used to obtain the optimal relabeling for all partitions, which is known as the voting problem. Next, the voting-based aggregation algorithm is used to obtain the consensus partition. The cumulative voting-based aggregation algorithm, which was described by Ayed and Kamel,[6,7] was used by Saeed et al.[25] to combine multiple clusterings of chemical structures.

Let χ denote a set of n data objects, and let a partition of χ into k clusters be represented by an $n \times k$ matrix U with a row for each object (j), and a column for each cluster (q), such that

$\sum_{q=1}^{k} u_{jq} = 1$, for $\forall$ j. Let $u = \{U^i\}_{i=1}^{b}$ denote an ensemble of partitions, where b is the number of partitions. The voting-based aggregation problem is concerned with searching for an optimal relabeling for each partition V^i with respect to representative partition U^0 (with k^0 clusters) and for a central aggregated partition denoted as $\bar{U}$ that summarizes the ensemble partitions. The matrix of coefficients W^i, which is a $k^i \times k^0$ matrix of w_{lq}^i coefficients, is used to obtain the optimal relabeling for ensemble partitions.

In this paper, the fixed-reference approach is applied whereby an initial reference partition is used as a common representative partition for all of the ensemble partitions and remains unchanged throughout the aggregation process. Instead of the selecting random partition as an initial reference, the partition generated by Ward's method is suggested to be the reference partition U^0. The cumulative voting-based aggregation algorithm is described as follows.

```
Cumulative Voting-based Aggregation Algorithm (CVAA)

select a partition $U^i \in \mathcal{U}$ generated by Ward's method and assign to $U^0$
for $i$=1 to $b$ do
    $W^i = (U^{iT}U^i)^{-1}U^{iT}U^0$
    $V^i = U^iW^i$
    $U^0 = \frac{i-1}{i}U^0 + \frac{1}{i}V^i$
end for
$\bar{U} = U^0$.
```

Although the CVAA can outperform Ward's method and give better results than other consensus clusterings, as shown by Saeed et al.,[25] different results are obtained every time the CVAA is applied because the final partition depends on the arrangement of partitions during the relabeling process. Therefore, the adaptive cumulative voting-based aggregation algorithm (A-CVAA)[7] was used to overcome this limitation.[26]

The initially selected A-CVAA reference partition determines the entropy $H(C^i)$ associated with the aggregated clusters, the initial value of the mutual information $I(C^i;X)$, and the upper bound on the amount of information that random cluster C contains for data objects X. This result motivates the use of a selection criterion for the initial reference partition based on the mutual information $I(Ci;X)$, which is equal to $H(C^i)$ for hard partitions (all individual clusterings used in this paper are hard clusterings).

The Shannon entropy associated with cluster C is defined over the cluster labels of the partition $\bar{U}$; H(C) measures the average amount of information associated with C and is defined as a function of its distribution p(c)[29] as follows

$$H(C) = -\sum_{c \in C} p(c) \log p(c) \tag{1}$$

The mutual information I(C;X) between C and X measures the amount of information that the random variable C contains about X, and vice versa. It is defined as

$$I(C; X) = H(C) - H(C|X) \tag{2}$$

It is noted that for a hard partition U^i, $I(C^i;X) = H(C^i)$ because the value of C^i is completely determined by the value of X (i.e., $H(C^i|X) = 0$).

The algorithm selects the initial reference partition, which is the one with highest entropy $H(C^i)$, and the ensemble partitions are sorted in descending order of their entropies. The adaptive cumulative voting based aggregation algorithm is described below.

In addition to the above two voting-based algorithms and in order to enhance the performance of the voting-based consensus algorithms, we used all possible permutations of arrangements of partitions $b!$ (in this experiment, $b = 6$) when using ECFP_4 for the MDDR data set. We found that the best results were obtained by the partitions arrangement: weighted average distance > Ward > average linkage > single-linkage > complete linkage > K-means. This enhancement may have occurred because the initial partition and its successors have high ability to separate the active from inactive molecules. We observed that when partitions are sorted according to their ability to separate the actives from inactives, the effectiveness of the final consensus partition is better. Because we cannot measure the ability to separate active from inactive molecules for all chemical data sets because no information is available for all active molecules, we use this arrangement for the enhanced voting-based consensus algorithm E-CVAA. The adaptive and enhanced cumulative voting-based aggregation algorithms are described as follows.

```
Adaptive and Enhanced Cumulative Voting-based Aggregation Algorithms

Re-order $\mathcal{U}$, s.t.
    For (A-CVAA), $U^i$ partitions are sorted in decreasing order of $H(C^i)$.
    For (E-CVAA), $U^i$ partitions are sorted as: weighted average distance> Ward >
    average linkage>  single-linkage> complete linkage> K-means.
Assign $U^1$ to $U^0$
for $i$=2 to $b$ do
    $W^i = (U^{iT}U^i)^{-1}U^{iT}U^0$
    $V^i = U^iW^i$
    $U^0 = \frac{i-1}{i}U^0 + \frac{1}{i}V^i$
End for
$\bar{U} = U^0$.
```

3. EXPERIMENTAL SECTION

3.1. Molecular Fingerprints. For the experiments, three molecular fingerprints were developed using Scitegic's Pipeline Pilot software,[30] the 120-ALOGP, which includes the octanol–water partitioning coefficient based on Ghose and Crippen's method,[31,32] and the two Scitegic extended-connectivity fingerprints, ECFP_4 and ECFC_4, with a length of 1024.[33,34] The extended-connectivity fingerprints are circular topological fingerprints designed for molecular characterization, similarity searching, and structure–activity modeling and are used in many applications such as clustering and virtual screening. The first character E in the fingerprint name denotes the atom abstraction method used to assign the initial atom code, which is derived from the number of connections to an atom, element type, charge, and atomic mass.[35]

3.2. Chemical Data Sets. Experiments were conducted using the most popular chemoinformatics databases: MDL Drug Data Report database.[36] This database consists of 102,516 molecules and has been used for many virtual screening experiments.[37−39] According to Hert et al.,[22] the subset data set was chosen from MDDR database, which involves structurally homogeneous molecules (e.g., rennin and HIV-1 protease inhibitors) and structurally heterogeneous molecules (e.g., cyclooxygenase and protein kinase C inhibitors). The diversity was estimated by the mean pairwise Tanimoto similarity across each set of active molecules (activity class), and the calculations of the similarity were conducted using the Pipeline Pilot software. The MDDR data set contains 11 activity classes (8294 molecules), and the details of this data set are listed in Table 1.

Table 1. MDDR Data Set Activity Classes

			Tanimoto pairwise similarity
activity index	activity class	active molecules	mean
31420	renin inhibitors	1130	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	1246	0.149
06233	substance P antagonists	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 row in the table contains an activity class, the number of molecules belonging to the class and the diversity of the class. The molecules were represented using the ALOGP and ECFP_4 fingerprints.

Moreover, to study the effectiveness of consensus clusterings on different chemical data sets, the experiments involved the maximum unbiased validation (MUV) data set reported recently by Rohrer and Baumann.[40] This data set contains 17 activity classes, and each class includes 30 active molecules (510 molecules were included in this data set). Details of this data set are given in Table 2. The molecules were represented using the ALOGP and ECFC_4 fingerprints.

3.3. Ensemble Generation and Combination. Every consensus clustering method consists of two steps: (i) ensemble generation and (ii) consensus functions. In this paper, the partitions were generated by a single run of multiple individual clustering algorithms (single-linkage, complete linkage, average linkage, weighted average distance, Ward's and K-means methods). For the MDDR data set, thresholds of 500, 600, 700, 800, 900, and 1000 were used to generate partitions with different sizes (number of clusters). For the MUV data set, a threshold of 17 was applied to obtain the required number of clusters. The same process was carried out for each 2D fingerprint in each data set. The Jaccard (Tanimoto) distance measure was used with each clustering technique because it was considered the method of choice for ensemble generations,[25] and the Euclidean distance was used for Ward's method.

Therefore, to evaluate the effectiveness of E-CVAA, this method was used to combine the ensembles generated using the MDDR and MUV data sets for different fingerprints, and the results were compared with those from other consensus clusterings (four co-association based, two graph-based, and two voting-based method) and compared with individual clustering (Ward's method).

Table 2. MUV Data Set Activity Classes

			Tanimoto pairwise similarity
activity index	activity class	active molecules	mean
466	S1P1 rec. (agonists)	30	0.285
548	PKA (inhibitors)	30	0.293
600	SF1 (inhibitors)	30	0.288
644	rho-kinase2 (inhibitors)	30	0.267
652	HIV RT-RNase (inhibitors)	30	0.258
689	Eph rec. A4 (inhibitors)	30	0.267
692	SF1 (agonists)	30	0.247
712	HSP 90 (inhibitors)	30	0.260
713	ER-a-Coact. bind. (inhibitors)	30	0.261
733	ER-b-Coact. bind. (inhibitors)	30	0.266
737	ER-a-Coact. bind. (potentiators)	30	0.297
810	FAK (inhibitors)	30	0.277
832	cathepsin G (inhibitors)	30	0.319
846	FXIa (inhibitors)	30	0.281
852	FXIIa (inhibitors)	30	0.295
858	D1 rec. (allosteric modulators)	30	0.247
859	M1 rec. (allosteric inhibitors)	30	0.275

3.4. Performance Evaluation. The MDDR data set results were evaluated based on the effectiveness of the methods in separating the active from inactive molecules using two measures: F-measure[41] and Quality Partition Index (QPI).[42] As defined by,[19] if the cluster contains n compounds, a of these are active, and there is a total of A compounds with the chosen activity. The precision, P, and recall, R, for that cluster can be expressed by

$$P = \frac{a}{n} \tag{3}$$

$$R = \frac{a}{A} \tag{4}$$

$$F = \frac{2PR}{P + R} \tag{5}$$

This calculation is carried out on each cluster, and the F-measure is the maximum value across all clusters.

According to ref 16, 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 data set as a whole. Let p be the number of actives in the active clusters, q be the number of inactives in the active clusters, r be the number of actives in the inactive clusters (i.e., clusters that are not active clusters), and s be the number of singleton actives. The high value occurs when the actives are clustered tightly together and separated from the inactive molecules. The QPI is defined as

$$\mathrm{QPI} = \frac{p}{p + q + r + s} \tag{6}$$

For the MUV data set, which was clustered into 17 classes (the number of different activity classes), the Rand Index (RI)[43] and Fowlkes–Mallows Index (FMI)[44] were used to compute the difference or (dis)agreement between the observed distribution

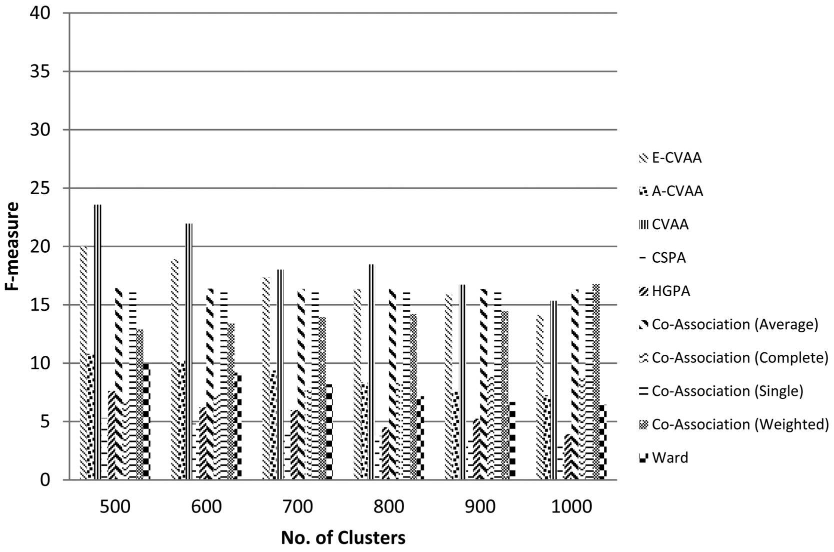


Figure 1. Effectiveness of clustering of the MDDR data set using the F-measure: ALOGP Fingerprint.

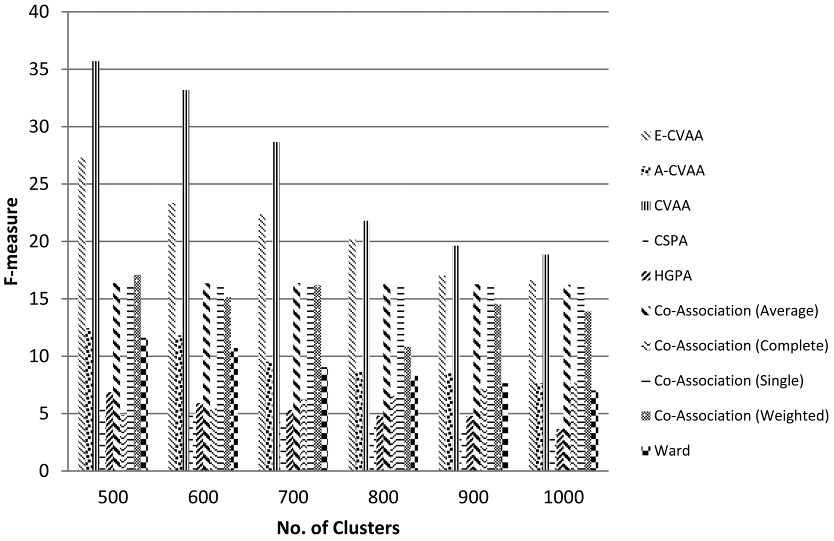


Figure 2. Effectiveness of clustering of the MDDR data set using the F-measure: ECFP_4 Fingerprint.

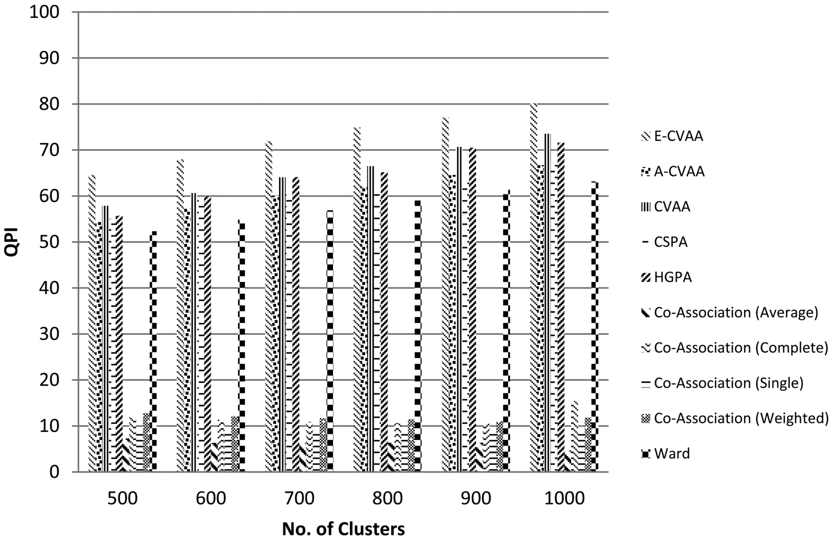


Figure 3. Effectiveness of clustering of the MDDR data set using the QPI: ALOGP Fingerprint.

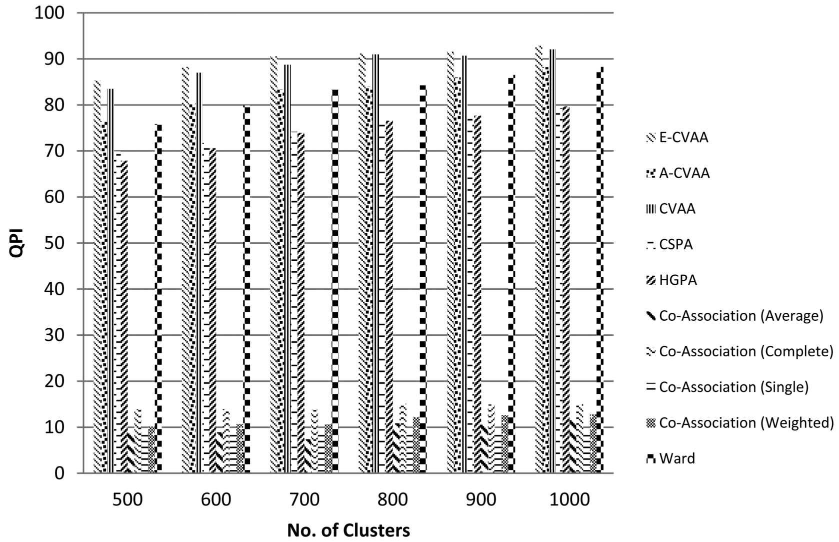


Figure 4. Effectiveness of clustering of the MDDR data set using the QPI: ECFP_4 Fingerprint.

Table 3. T-test Statistical Significance Testing Using F-measure Paired Differences for MDDR Data Set: ALOGP Fingerprint

	paired differences					
			95% confidence interval of the difference			
	mean	std. deviation	std. error mean	lower	upper	sig. (2-tailed)
Pair 1: E-CVAA−Ward	9.16	0.82	0.34	8.30	10.02	0.000001
Pair 2: A-CVAA−Ward	1.00	0.15	0.06	0.84	1.16	0.000017
Pair 3: CVAA−Ward	11.08	1.84	0.75	9.15	13.00	0.000026
Pair 4: CSPA−Ward	−3.82	0.60	0.25	−4.45	−3.18	0.000020
Pair 5: HGPA−Ward	−2.35	0.52	0.21	−2.89	−1.80	0.000109
Pair 6: Co-Association (Average)−Ward	8.44	1.38	0.56	6.99	9.87	0.000024
Pair 7: co-association (complete)−Ward	0.08	2.23	0.91	−2.26	2.41	0.934731
Pair 8: co-association (single)−Ward	8.37	1.36	0.55	6.94	9.80	0.000023
Pair 9: co-association (weighted)−Ward	6.34	2.64	1.08	3.57	9.11	0.002028

of molecules using clustering techniques and the ideal distribution of molecules in the data set. The ideal distribution is one in which the data set was partitioned into 17 clusters, and each cluster only contains molecules from a single activity (30 molecules).

If X is the partition generated by the clustering method (consensus or individual), and Y is the ideal partition, then a is the number of pairs that are in the same cluster in X and also in the same cluster in Y; b is the number of pairs that are in different clusters in X and different clusters in Y; c is the number of pairs that are in the same cluster in X and in different clusters in Y; and d is the number of pairs that are in different clusters in X and in the same cluster in Y.

The Rand Index (RI) and Fowlkes−Mallows Index (FMI) are computed as

$$\mathrm{RI} = \frac{a + b}{a + b + c + d} \tag{7}$$

$$\mathrm{FMI} = \frac{a}{\sqrt{(a + c)(a + d)}} \tag{8}$$

4. RESULTS AND DISCUSSION

In the experiments, nine consensus clustering methods were applied to combine multiple clusterings of chemical structures. Every ensemble contained six partitions for each data set (MDDR and MUV) that were represented by the ALOGP and the Scitegic extended-connectivity fingerprints (ECFP_4 and ECFC_4).

For the MDDR data set, Figures 1−4 show the effectiveness of the consensus clustering methods using the F and QPI measures compared with that of Ward's method. The mean of the F and QPI values were subsequently averaged over the 11 activity classes of the data set.

Visual inspection of the MDDR results for ALOGP (Figures 1 and 3) shows that the voting-based methods (CVAA and E-CVAA) performed quite well and outperformed all other methods (including Ward's method) using both criteria. In addition, the A-CVAA method outperformed Ward's method using F and QPI measures. Other consensus methods performed somewhat well. The performance of the graph-based methods outperformed Ward's method using the QPI measure, whereas the co-association based methods outperformed Ward's method for the F-measure. The best choice for the graph-based methods was the HGPA, whereas the average linkage method was the best choice among the co-association-based methods, which obtained similar results for different partition sizes.

With the clustering of the MDDR for ECFP_4, as shown in Figures 2 and 4, the performance of the consensus clustering methods is consistent with the ALOGP results. Both the CVAA and E-CVAA outperformed the other methods using the F and QPI measures, and the A_CVAA outperformed Ward's method

Table 4. T-test Statistical Significance Testing Using F-measure Paired Differences for MDDR Data Set: ECFP_4 Fingerprint

	paired differences					
				95% confidence interval of the difference		
	mean	std. deviation	std. error mean	lower	upper	sig. (2-tailed)
Pair 1: E-CVAA−Ward	12.08	2.38	0.97	9.59	14.58	0.000059
Pair 2: A-CVAA−Ward	0.73	0.28	0.11	0.44	1.03	0.001357
Pair 3: CVAA−Ward	17.25	5.49	2.24	11.49	23.01	0.000589
Pair 4: CSPA−Ward	−4.79	0.93	0.38	−5.77	−3.81	0.000056
Pair 5: HGPA−Ward	−3.83	0.79	0.32	−4.65	−3.00	0.000074
Pair 6: co-association (average)−Ward	7.26	1.74	0.71	5.44	9.08	0.000152
Pair 7: co-association (complete)−Ward	−2.73	2.84	1.16	−5.71	0.25	0.065363
Pair 8: co-association (single)−Ward	7.25	1.73	0.71	5.43	9.07	0.000151
Pair 9: co-association (weighted)−Ward	5.56	1.83	0.75	3.64	7.47	0.000684

Table 5. T-test Statistical Significance Testing Using QPI Measure Paired Differences for MDDR Data Set: ALOGP Fingerprint

	paired differences					
				95% confidence interval of the difference		
	mean	std. deviation	std. error mean	lower	upper	sig. (2-tailed)
Pair 1: E-CVAA−Ward	14.83	1.81	0.74	12.93	16.77	0.000006
Pair 2: A-CVAA−Ward	2.93	0.53	0.22	2.37	3.49	0.000040
Pair 3: CVAA−Ward	7.61	1.93	0.79	5.59	9.63	0.000200
Pair 4: CSPA−Ward	3.27	1.01	0.41	2.21	4.32	0.000507
Pair 5: HGPA−Ward	6.57	2.16	0.88	4.29	8.82	0.000687
Pair 6: co-association (average)−Ward	−51.70	4.76	1.94	−56.70	−46.71	0.000001
Pair 7: co-association (complete)−Ward	−46.14	3.72	1.52	−50.05	−42.24	0.000001
Pair 8: co-association (single)−Ward	−49.07	4.06	1.66	−53.33	−44.80	0.000001
Pair 9: co-association (weighted)−Ward	−46.12	4.52	1.85	−50.86	−41.37	0.000002

Table 6. T-test Statistical Significance Testing Using QPI Measure Paired Differences for MDDR Data Set: ECFP_4 Fingerprint

	paired differences					
				95% confidence interval of the difference		
	mean	std. deviation	std. error mean	lower	upper	sig. (2-tailed)
Pair 1: E-CVAA−Ward	6.94	1.86	0.76	4.98	8.89	0.0002662
Pair 2: A-CVAA−Ward	0.04	0.42	0.17	−0.41	0.48	0.8331521
Pair 3: CVAA−Ward	5.81	1.60	0.65	4.13	7.49	0.0003016
Pair 4: CSPA−Ward	−8.20	1.19	0.48	−9.45	−6.96	0.0000131
Pair 5: HGPA−Ward	−8.63	0.70	0.29	−9.36	−7.89	0.0000007
Pair 6: Co-Association (Average)−Ward	−73.27	3.77	1.54	−77.22	−69.31	0.0000001
Pair 7: co-association (complete)−Ward	−68.58	4.08	1.66	−72.86	−64.30	0.0000002
Pair 8: co-association (single)−Ward	−74.14	4.56	1.86	−78.93	−69.36	0.0000002
Pair 9: co-association (weighted)−Ward	−71.50	3.52	1.44	−75.19	−67.80	0.0000001

for the F-measure and performed similarly to Ward's method for the QPI measure. Again, the HGPA obtained the best results for the graph-based methods and outperformed Ward's method for QPI measure. The average linkage was the best choice for the co-association-based methods with the F-measure.

As shown in Tables 3−6, statistical significance tests (*t* tests) were performed to verify the improvements achieved by the consensus clustering methods. The tables display a number of parameters, i.e., mean value, standard deviation, standard error, and significance values, for the pairs of F and QPI values of each consensus clustering method with Ward's method. The *t* test procedure compares the means of two variables that represent the results of two clusterings at different cluster sizes. A low significance value for the *t* test (typically less than 0.05) indicates that a significant difference exists that was satisfied between two variables. In Tables 3 and 4, it was noted that the significant improvements (significance field) associated with the F-measure for AlogP are (Table 3): E-CVAA−Ward (0.000001), A-CVAA−Ward (0.000017), CVAA−Ward (0.000026), co-association (average)−Ward (0.000024), co-association (single)−Ward (0.000023), and co-association (weighted)−Ward (0.002028). For ECFP_4 (Table 4): E-CVAA−Ward (0.000059), A-CVAA−Ward (0.001357), CVAA−Ward (0.000589), co-association (average)−Ward (0.000152), co-association (single)−Ward (0.000151), and co-association (weighted)−Ward (0.000684). Similarly, the significance field in terms of QPI values for AlogP (Table 5): E-CVAA−Ward (0.000006), A-CVAA−Ward (0.000040), and CVAA−Ward (0.000200), CSPA−Ward (0.000507), and HGPA−Ward (0.000687). For ECFP_4 (Table 6): ECVAA−Ward (0.0002662) and CVAA−Ward (0.0003016). It was noted that the E-CVAA and CVAA significantly outperformed Ward's method using both criteria for the ALOGP and

Table 7. Effectiveness of Clustering of MUV Data Set

		RI		FMI	
	clustering method	ALOGP	ECFC_4	ALOGP	ECFC_4
consensus	E-CVAA	0.68	0.59	**0.14**	**0.16**
	A-CVAA	0.89	**0.88**	0.09	0.09
	CVAA	0.46	0.13	**0.19**	**0.23**
	CSPA	**0.90**	**0.90**	0.09	0.10
	HGPA	0.89	**0.90**	0.08	0.09
	co-association (average)	0.11	0.11	**0.24**	**0.23**
	co-association (complete)	0.87	**0.88**	0.07	0.06
	co-association (single)	0.11	0.11	**0.23**	**0.23**
	co-association (weighted)	0.37	0.19	**0.19**	**0.22**
individual	mean of indv. methods (std)	0.70 (0.31)	0.63 (0.30)	0.13 (0.06)	0.15 (0.05)
	Ward	0.89	0.82	0.09	0.13

ECFP_4. In addition, the A-CVAA method significantly outperformed Ward's method using the F-measure for both fingerprints and using the QPI measure for ALOGP (it obtained a performance similar to that of Ward's method using the QPI for ECFP-4). Moreover, the co-association-based methods significantly outperformed Ward's method using the F-measure for both fingerprints (except for the co-association complete method, which obtained results with no significant difference from those of Ward's method), whereas the graph-based methods significantly outperformed Ward's method using the QPI measure for the ALOGP. It is concluded that the best significant improvements from both criteria and two fingerprints were obtained using the voting-based methods: E-CVAA, CVAA, and A-CVAA.

Inspection of the MUV data set results (Table 7) shows that the voting-based method (A-CVAA), graph-based methods (CSPA and HGPA), and co-association (complete) method outperformed Ward's method and the mean of the individual clustering methods for the RI measure. Although the RI measure is widely used, it contains limitations in that its value approaches its upper bound as the number of clusters increases. Therefore, the Fowlkes–Mallows Index (FMI) measure has been designed to address the limitations of the basic RI.[19] In Table 7, when using the FMI measure, the results show that the voting–based methods (CVAA and E-CVAA) and co-association-based methods (average, single, and weighted) outperformed Ward's method using the FMI measure for both fingerprints. Additionally, these methods outperformed the average (mean) performance of all individual clustering methods for the ALOGP and ECFC-4.

Because the best partition arrangement was suggested by the E-CVAA method using the ECFP_4 for the MDDR data set, we have evaluated the method using same data set with a different fingerprint (MDDR with ALOGP) and also a different data set with different fingerprints (MUV with ALOGP and ECFC_4) and concluded that it performs quite well and outperforms Ward's method for the F, QPI and FMI measures.

5. CONCLUSION

The experimental results show that voting-based consensus clustering methods significantly improve the effectiveness of chemical structure clusterings compared with those of individual clustering methods. The enhanced cumulative voting-based aggregation algorithm E-CVAA, which is introduced in this paper, was the method of choice among the consensus clustering methods. In addition, the co-association-based consensus methods significantly outperformed Ward's method using the F-measure for the ALOGP and ECFP_4, whereas the graph-based consensus methods significantly outperformed Ward's method using the QPI measure for the ALOGP. In addition, the computational complexity of voting-based methods is $O(n)$, 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 data objects. Significant improvements for the clustering effectiveness of the voting-based consensus methods compared with those of the individual clustering methods and the efficient performance of voting-based methods compared with that of other consensus methods may overcome the disadvantage of the high computational cost required for consensus clusterings. In future work, we will examine the use of soft consensus methods for chemical structure clustering.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: alsamet.faisal@gmail.com.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported by the Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at Universiti Teknologi Malaysia (UTM) under Research University Grant Category (VOT Q.J130000.7128.00H72).

■ REFERENCES

(1) Gionis, A.; Mannila, H.; Tsaparas, P. Clustering aggregation. *ACM Trans. Knowl. Discovery Data* **2007**, *1*, Article 1.

(2) Vega-Pons, S.; Ruiz-Schulcloper, J. A survey of clustering ensemble algorithms. *Int. J. Pattern Recogn.* **2011**, *25* (3), 337–372.

(3) Strehl, A.; Ghosh, J. Cluster ensembles: A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.

(4) Topchy,A.; Jain, A. K.; Punch, W. A Mixture Model of Clustering Ensembles. SIAM International Conference on Data Mining, Lake Buena Vista, FL, U.S.A., April 22–24, 2004; pp 379–390.

(5) Fred, A. L. N.; Jain, A. K. Combining multiple clustering using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 835–850.

(6) Ayad, H. G.; Kamel, M. S. Cumulative voting consensus method for partitions with a variable number of clusters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30* (1), 160–173.

(7) Ayad, H. G.; Kamel, M. S. On voting-based consensus of cluster ensembles. *Pattern Recognit.* **2010**, *43*, 1943–1953.

(8) Monti, S.; Tamayo, P.; Mesirov, J.; Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **2003**, *52*, 91−118.

(9) Johnson, E.; Kargupta, H. Collective, Hierarchical Clustering from Distributed, Heterogeneous Data. In *Large-Scale Parallel KDD Systems*; Lecture Notes in Computer Science 1759; Springer-Verlag: Berlin, 1999; pp 221−244.

(10) Topchy, A. P.; Jain, A. K.; Punch, W. F. Clustering ensembles: Models of consensus and weak partitions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27* (12), 1866−1881.

(11) 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.

(12) Downs, G. M.; Willett, P.; Fisanick, W. Similarity searching and clustering of chemical−structure databases using molecular property data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.

(13) 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.

(14) Downs, G. M.; Barnard, J. M. Clustering methods and their uses in computational chemistry. *Rev. Comput. Chem.* **2002**, *18*, 1−40.

(15) Holliday, J. D.; Rodgers, S. L.; Willett, P. Clustering files of chemical structures using the fuzzy k-means clustering method. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 894−902.

(16) Varin, T.; Bureau, R.; Mueller, C.; Willett, P. Clustering files of chemical structures using the Székely−Rizzo generalization of Ward's method. *J. Mol. Graphics Modell.* **2009**, *28* (2), 187−195.

(17) Gillet, V. J. Diversity selection algorithms. *Wiley Interdisc. Rev.: Comput. Mol. Sci.* **2011**, *1*, 580−589.

(18) Rivera-Borroto, O. M.; Marrero-Ponce, Y.; García de la Vega, J. M.; Grau-Ábalo, R. C. Comparison of combinatorial clustering methods on pharmacological data sets represented by machine learning-selected real molecular descriptors. *J. Chem. Inf. Model,* **2011**, *51* (12), 3036−3049.

(19) Chu, C-W; Holliday, J.; Willett, P. Combining multiple classifications of chemical structures using consensus clustering. *Bioorg. Med. Chem.* **2012**, *20* (18), 5366−5371.

(20) Feher, M. Consensus scoring for protein−ligand interactions. *Drug Discovery Today* **2006**, *11*, 421−428.

(21) Salim, N.; Holliday, J. D.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−442.

(22) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New methods for ligand-based virtual screening: Use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **2006**, *46*, 462−470.

(23) Willett, P. Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR Comb. Sci.* **2006**, *25*, 1143−1152.

(24) Saeed, F.; Salim, N.; Abdo, A.; Hentabli, H. Graph-based consensus clustering for combining multiple clusterings of chemical structures. *Mol. Inf.* **2013**, *32*, 165−178.

(25) Saeed, F.; Salim, N.; Abdo. Voting-based consensus clustering for combining multiple clusterings of chemical structures. *J. Cheminf.* [Online] **2012**, *4*, Article 37. http://www.jcheminf.com/content/4/1/37 (accessed March 20, **2013**).

(26) Saeed, F.; Salim, N.; Abdo, A. Information theory and voting-based consensus clustering for combining multiple clusterings of chemical structures. *Mol. Inf.* **2013**.

(27) Karypis, G.; Kumar, V. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.* **1998**, *20*, 359−392.

(28) Karypis, G.; Aggarwal, R.; Kumar, V.; Shekhar, S. Multilevel Hypergraph Partitioning: Application in VLSI Domain. In *Proceedings of the 34th Anual Design Automation Conference*; ACM, 1997; pp 526−529.

(29) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley: New York, 1991.

(30) *Pipeline Pilot*; Accelrys Software Inc.: San Diego, 2008.

(31) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure−activity relationships 1. Partition coefficients as a measure of hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565−577.

(32) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A.* **1998**, *102*, 3762−3772.

(33) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(34) Hu, Y.; Lounkine, E.; Bajorath, J. Improving the search performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit-density-dependent similarity function. *ChemMedChem* **2009**, *4*, 540−548.

(35) Chen, L.; Li, Y.; Zaho, Q.; Peng, H.; Hou, T. ADME evaluation in drug discovery. 10. Predictions of Pglycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol. Pharm.* **2011**, *8*, 889−900.

(36) Accelrys, Inc.; MDL Drug Data Report (MDDR) database. http://www.accelrys.com/ (accessed November 1, 2012)

(37) Abdo, A.; Chen, B.; Mueller, C.; Salim, N.; Willett, P. Ligand-Based Virtual Screening Using Bayesian Networks. *J. Chem. Inf. Model.* **2010**, *50*, 1012−1020.

(38) Abdo, A.; Salim, N. New Fragment weighting scheme for the Bayesian inference network in ligand-based virtual screening. *J. Chem. Inf. Model.* **2011**, *51*, 25−32.

(39) Abdo, A.; Saeed, F.; Hentabli, H.; Ahmed, A.; Salim, N. Ligand expansion in ligand-based virtual screening using relevance feedback. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 279−287.

(40) Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* **2009**, *49*, 169−184.

(41) Van Rijsbergen, C. J. *Information Retrieval*, 2nd ed.; Butterworth: London, 1979.

(42) Varin, T.; Saettel, N.; Villain, J.; Lesnard, A.; Dauphin, F.; Bureau, R.; Rault, S. J. 3D Pharmacophore, hierarchical methods, and 5-HT4 receptor binding data. *Enzyme Inhib. Med. Chem.* **2008**, *23*, 593−603.

(43) Rand, W. M. J. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **1971**, *66*, 846−850.

(44) Fowlkes, E. B.; Mallows, C. L. A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* **1983**, *78*, 553−569.