Large-Scale Evaluation of Molecular Descriptors by Means of Clustering

Matthias Dehmer[1]*, Frank Emmert-Streib[2], Shailesh Tripathi[1]

1 UMIT, Division for Bioinformatics and Translational Research, Eduard Wallnoefer Zentrum 1, Hall in Tyrol, Austria, 2 Queen's University Belfast, Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Belfast, United Kingdom

Abstract

Molecular descriptors have been explored extensively. From these studies, it is known that a large number of descriptors are strongly correlated and capture similar characteristics of molecules. In this paper, we evaluate 919 Dragon-descriptors of 6 different categories by means of clustering. Also, we analyze these different categories of descriptors also find a subset of descriptors which are least correlated among each other and, hence, characterize molecular graphs distinctively.

Citation: Dehmer M, Emmert-Streib F, Tripathi S (2013) Large-Scale Evaluation of Molecular Descriptors by Means of Clustering. PLoS ONE 8(12): e83956. doi:10.1371/journal.pone.0083956

Editor: Danilo Roccatano, Jacobs University Bremen, Germany

Received June 6, 2013; **Accepted** November 9, 2013; **Published** December 31, 2013

Funding: Matthias Dehmer and Shailesh Tripathi thank the Austrian Science Funds for supporting this work (project P22029-N13). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors declare that co-author Frank Emmert-Streib is a PLOS ONE Editorial Board member. This does not alter their adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: matthias.dehmer@umit.at

Introduction

Molecular descriptors map molecular structures to the reals by taking physical, chemical or structural information into account [1]. A large number of descriptors have been developed to describe different properties of molecular graphs. Therefore, these descriptors can be classified into different categories depending what kind of information is used (e.g., physical, chemical or structural information) to define such a measure. The commercial software package Dragon [2] (version 6.0.26) contains 4885 molecular descriptors which are classified into 29 categories.

The problem of analyzing molecular descriptors by applying clustering techniques has been already explored [3–6]. These are usually based on using principal component analysis (PCA) and correlation-based methods for the identification of different descriptors. For example, Todeschini et al. [6] and Basak et al. [3] evaluated descriptors on a rather small collection of molecular graphs using PCA and ranked them based on the intercorrelation. In order to find similarities between molecular descriptors, Basak et al. [4,5] used a PCA-based clustering technique on both a hydrocarbon dataset and mixed chemical compounds. Taraviras et al. [7] performed a cluster analysis with 240 descriptors by using different clustering algorithms. The weak point of the just sketched approaches is that the corresponding study has not been performed on a large scale (large data sets) and with distinct descriptors belonging to several categories. Also, the optimal number of different descriptors (dimension) has not been validated statistically. In this paper, we overcome these problems.

A thorough evaluation of the vast amount of developed descriptors [1] is required to identify categories of descriptors which capture structural information differently. In our analysis we evaluate 6 categories (see next section) of structural descriptors by means of clustering. The main contribution of this paper is to explore the *dimension* of the descriptor space, i.e., how many different descriptors exist among all which have been introduced so far. Here, we put the emphasis on 919 structural descriptors from Dragon. In particular, we find that only a very few descriptors are different. In this context that means they are least correlated and, therefore, capture structural information differently.

Methods and Results

Molecular Descriptors

To perform our study, we used six categories of descriptors implemented in Dragon (version 6.0.26) which are defined as follows:

1. **Connectivity indices** [1]: These indices are calculated from the vertex-degree of a molecular graph. The Randić index [8] is a prominent example thereof.
2. **Edge adjacency indices** [1]: These indices are based on the edge adjacency matrix of a graph. The resulting descriptor-value is the sum of all edge entries of the adjacency matrix of a graph. Balaban et al. [9] developed several indices by using graph-theoretical matrices.
3. **Topological indices** [1]: These structural graph measures which take various structural features into account, e.g., distances and eigenvalues. The term *topological index* has been firstly coined by Hosoya [10]. The first and the second Zagreb indices [11] are prominent examples thereof.
4. **Walk path counts** [1]: These indices are defined by counting paths or walks of a graph. Here, the term *walk* refers to random walks which is based on using a probability measure. We point out that such indices have been listed by Todeschini and Consonni [1].
5. **Information indices** [1]: These measures are based on using Shannon's entropy. To assign a probability value to a graph, Dragon uses so-called partition-based methods [12] by using

several graph invariants such as vertices, edges, vertex degrees and distances have been used [12]. The so-called topological information content [13] and the Bonchev-Trinajstić index [14] are prominent examples of partition-based information indices. So-called partition-independent information-theoretic measures for graphs have been developed by Dehmer [12].

6. **2D Matrix-based** [1]: These descriptors are calculated based on the elements of so-called graph-theoretical matrices [15] by using several algebraic operations. The Balaban-like indices inferred from the adjacency matrix [2,9] are important examples of this category.

We want to emphasize that the term 'Topological indices' is here misleading and ambiguous. For example, typical information indices are based on structural features of a graph by using Shannon's entropy. So, they represent topological indices too. The same holds for all other groups which have been defined by using structural features of molecular structures and, therefore, they are topological indices as well, see [1,9,16–19].

Data

In order to evaluate the above mentioned 6 categories of descriptors, we use 3 data sets namely:

1. MS_{2265} contains (non-isomorphic) molecular structures (only skeletons, i.e., without vertex- and edge labels) inferred from the NIST spectral database [20].
2. C_{15} contains exhaustively generated (non-isomorphic) tree structures with 15 vertices each [20].
3. N_8 contains exhaustively generated (non-isomorphic) graphs with 8 vertices each [21].

To perform our analysis, we calculate the descriptor values for these three datasets. We removed those descriptors which give constant and erroneous values by using the three data sets. The erroneous values are produced by those descriptors for which we have not been able to calculate a descriptor value of a network without additional physical or chemical information. Finally, we the above mentioned six categories contain 24, 301, 57, 28, 40, 469 descriptors.

Clustering Techniques

Clustering is an unsupervised learning technique which aims to find different groups or clusters of objects in data [22]. The groups are described as a collection of objects which are closer to each other than the rest of the objects [22]. An example thereof is hierarchical clustering as groups of the objects are arranged in a hierarchical order by a so-called dendogram. The objects which are clustered in one group have a higher degree of similarity than the objects which are clustered in different groups. Thus a resulting clustering solution allows to determine clusters where each cluster shows distinct property of the data. The similarity or dissimilarity between two objects is usually determined by using a Similarity/distance function which measures the similarity/distance between data points of different objects. Examples are the Euclidean distance, the Manhattan distance or the correlation-based distance. A dissimilarity can be described as follows:

Several algorithms have been developed for cluster analysis [22]. These algorithms can be divided into several categories namely partition-based clustering, hierarchical clustering, density-based clustering, grid-based clustering and fuzzy clustering [22,23]. Thus k-means, soft k-means Clustering, k-medoids Clustering [22] are some examples representing non-hierarchical clustering methods. Hierarchical clustering itself can be divided into two categories called agglomerative and divisive clustering [22]. As known, several concrete methods thereof have been developed such as single linkage, complete linkage and average linkage, see [22].

In order to evaluate the descriptors, we perform hierarchical clustering (average linkage) by using the mentioned Dragon descriptors and the Spearman rank correlation as a distance measure. Here, we denote the correlation matrix between descriptors as Σ. Then, the distance between a pair of descriptors is defined by.

$$d_{ij} = 1 - \|\Sigma_{ij}\|. \tag{1}$$

In order to choose a clustering method we use the cophenetic correlation measure [24]. A high correlation coefficient shows that the distance between the data points is well preserved by the created dendogram of the hierarchical clustering solution. In our analysis, the cophenetic correlation coefficient is highest for the average clustering solution for all three data-set compared to other clustering algorithms. We calculate the cophenetic correlation for seven hierarchical clustering algorithms which are the Ward, Single, Complete, Average, Mcquitty, Median and the Centroid-method. The cophentic correlation coefficients for the average clustering solutions for three data-sets are 0.84, 0.89 and 0.93.

Cluster Validity

Cluster validity [23,25] is used to evaluate the quality of clustering solution (by using a certain clustering algorithm), e.g., the optimum number of clusters in the data, or whether the resulting cluster solution fits the data. Known clustering validation techniques are divided into three categories namely internal, external and relative validity criteria. External validation criteria evaluate clustering solutions with a predefined clustering structure. Using internal validation criteria relates to find the optimal number of clusters which is based on the intrinsic knowledge of data. Relative validation criteria are used to compare two different clustering solutions [23].

In order to perform analyses, we use external and internal clustering validation criteria. For the external validation, we compared the clustering solution with a predefined group of clusters which serve as reference clusters. The external clustering validity of a clustering solution with respect to the given reference cluster is estimated by using the information-theoretic quantity NMI_{max} (normalized mutual information) [26,27] defined by

$$NMI_{max} = \frac{I(U,V)}{max\{H(U),H(V)\}}. \tag{2}$$

where

$$H(U) = -\Sigma_{i=1}^{R} \frac{a_i}{N} \log\left(\frac{a_i}{N}\right), \tag{3}$$

$$H(V) = -\Sigma_{i=1}^{C} \frac{b_i}{N} \log\left(\frac{b_i}{N}\right), \tag{4}$$

$$I(U,V) = \Sigma_{i=1}^{R} \Sigma_{j=1}^{C} \frac{n_{ij}}{N} \log\left(\frac{n_{ij}/N}{a_i b_j/N^2}\right). \tag{5}$$

Table 1. A contingency table which defines the overlap between two cluster solutions, U and V.

U ↓ \ V →	V_1	V_2	.	.	.	V_C	**Sums**
U_1	n_{11}	n_{12}	.	.	.	n_{1C}	a_1
U_2	n_{21}	n_{22}	.	.	.	n_{2C}	a_2
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
U_R	n_{R1}	n_{R2}	.	.	.	n_{RC}	a_R
Sums	b_1	b_2	.	.	.	b_C	N

doi:10.1371/journal.pone.0083956.t001

Hereby, we assume that we have two clustering solutions U and V which have R and C clusters. The overlap between these two clusters is shown in the contingency Table 1. We calculated NMI_{max} for all three data-sets with different number of clusters.

The Optimal Number of Clusters

The optimal number of clusters (internal cluster validity) are determined by consensus clustering [27,28] which has been here performed as follows. Assume we evaluate N descriptors on a dataset containing n molecular graphs. Thus we get n descriptor values for each descriptor. First, we resample the data of sample-size, $p<n$, $B=100$ times for N descriptors to generate B clustering solutions $U_k=\{U_k^1,U_k^1\ldots U_k^B\}$, for k clusters, where $k=2,3,\ldots,200$. After that we calculate the consensus indices for each

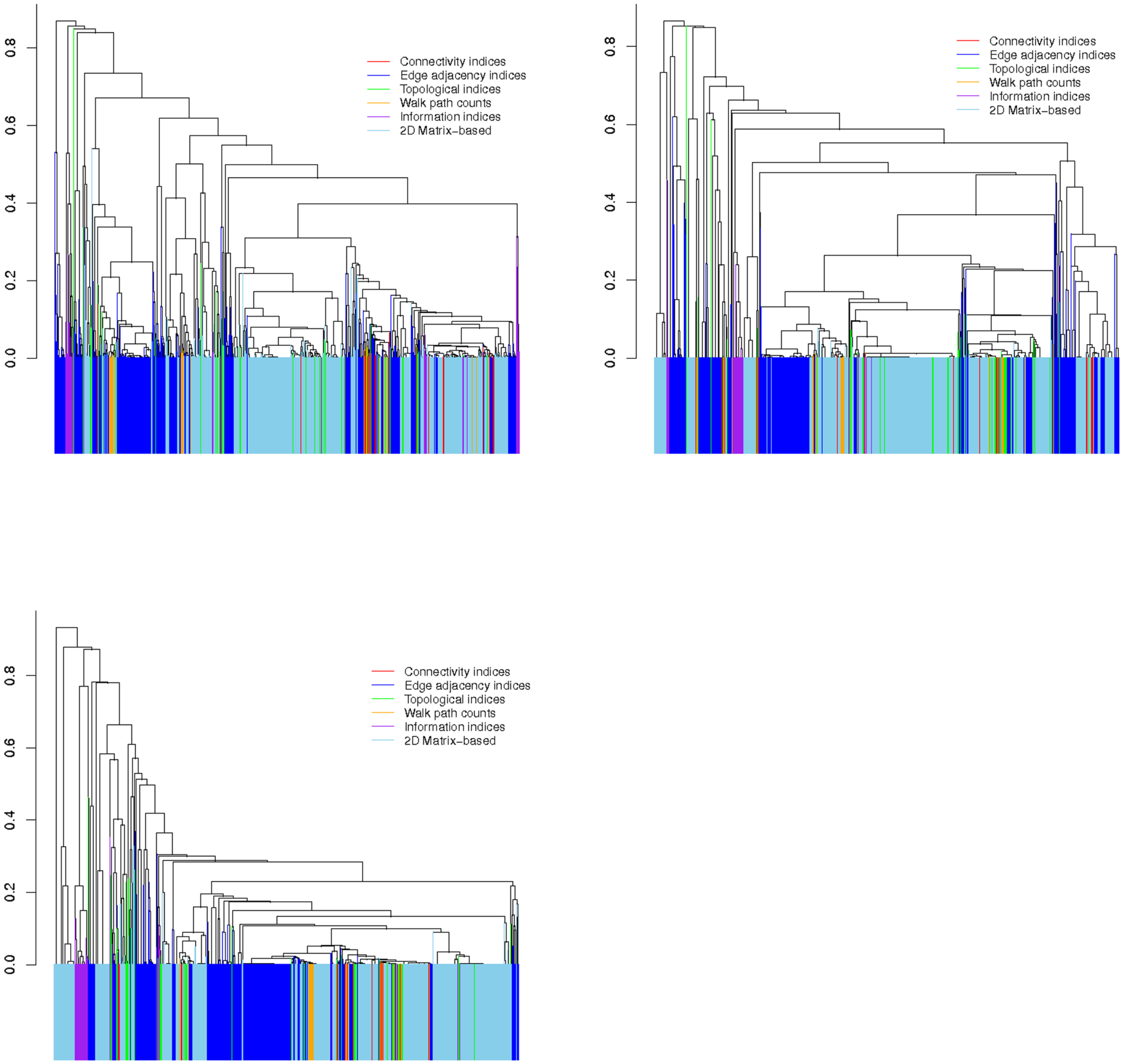

Figure 1. Hierarchical clustering using the average algorithm, MS_{2265} **(left),** C_{15} **(middle),** N_8 (right). The total number of descriptors equals 919. They belong to 6 different categories which are as follows: connectivity indices (24), edge adjacency indices (301), topological indices (57), walk path counts (28), information indices (40) and 2D Matrix-based (469).
doi:10.1371/journal.pone.0083956.g001

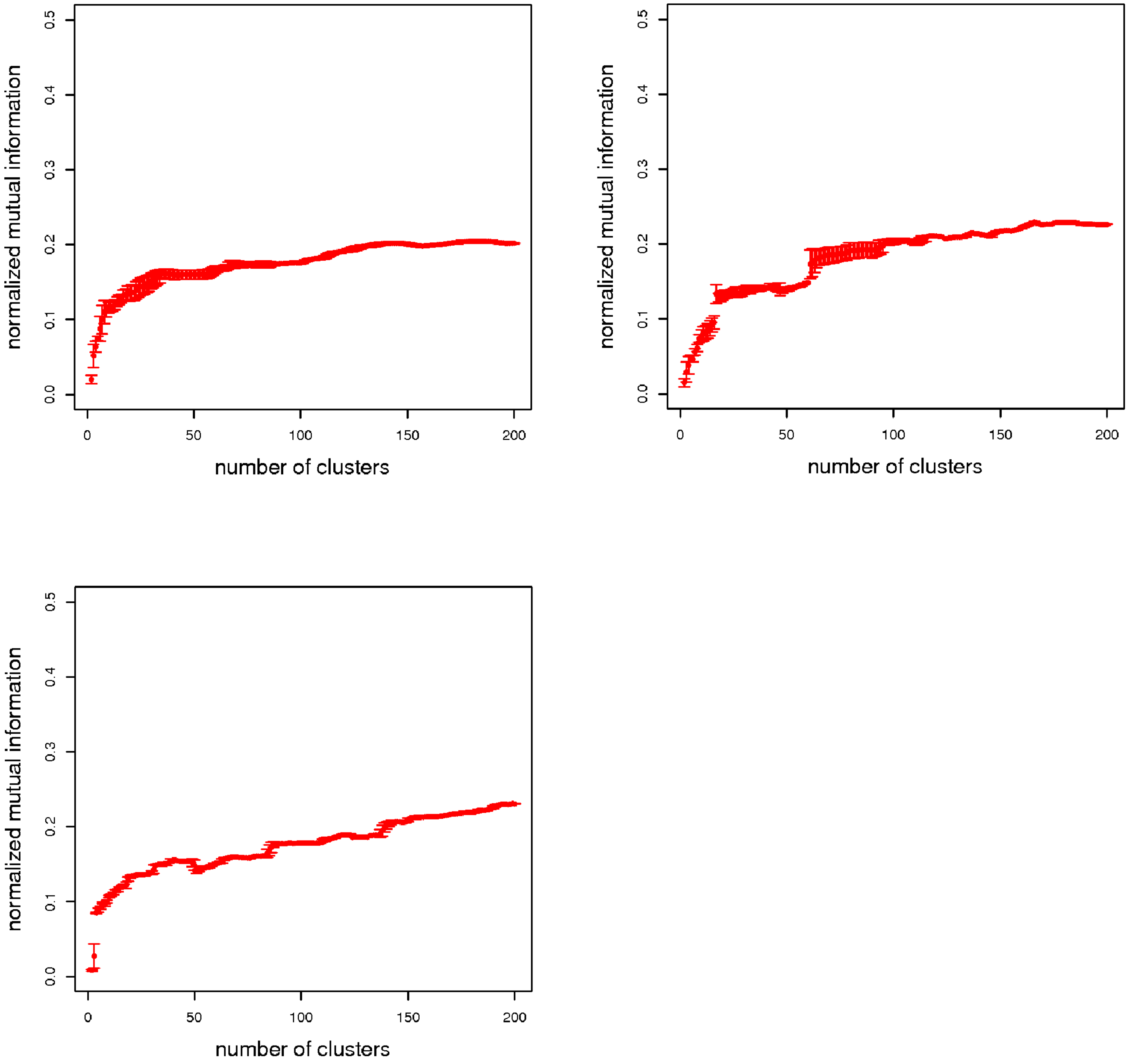

Figure 2. The normalized mutual information, NMI_{max}, between reference clusters, RC, and the number of clusters, K, obtained by hierarchical clustering for three data-sets MS_{2265} (left), C_{15} (right) and N_8 (bottom). NMI_{max} for each K has been generated by sampling the data sets D_B, where $B=1,\ldots,100$ (data set MS_{2265}). The total number of descriptors equals 919. They belong to 6 different categories which are as follows: connectivity indices (24), edge adjacency indices (301), topological indices (57), walk path counts (28), information indices (40) and 2D Matrix-based (469).
doi:10.1371/journal.pone.0083956.g002

cluster, k, which is defined as follows:

$$CI(U_k) = \frac{\Sigma_{i<j} AM(U_k^i, U_k^j)}{B(B-1)/2}. \qquad (6)$$

As to the measure AM, we use the adjusted rand index ARI [29] defined by.

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_i}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{N}{2}}. \qquad (7)$$

The number of clusters k for which CI attains its maximum is chosen as the optimal number of clusters, namely.

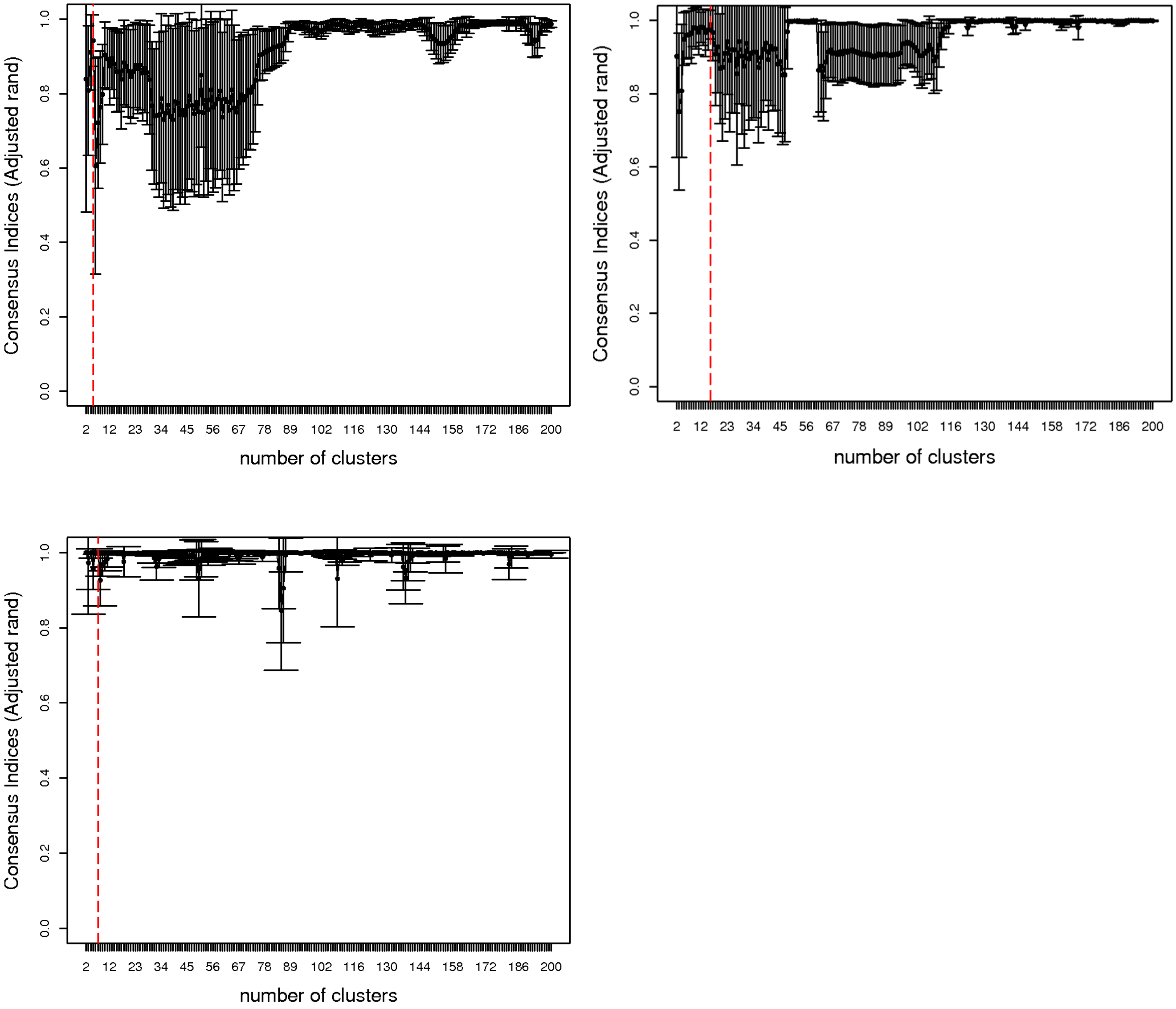

Figure 3. Consensus indices using the *adjusted rand index* for estimating the number of clusters in the data. These plots have been generated by sampling the data sets B, where $B = 1, \dots, 100$ for the three data sets, MS_{2265} (left), C_{15} (right), N_8 (bottom). The dotted red line shows the optimal number of clusters.
doi:10.1371/journal.pone.0083956.g003

$$k_{optimal} = arg\,max_{k=2,\dots,k_{max}} CI(U_k). \tag{8}$$

Determining a Highly Correlated Subset of Descriptors

Let D be a set of descriptors and $|D|$ is its cardinality. Let S be a subset of D. The selected $|D| = 919$ descriptors can be reduced to a set of descriptors, $S \subseteq D$. The remaining $|D| - |S|$ descriptors will have a significant correlation with at least one of the descriptor in the set S and the descriptors in S are not significantly correlated. If two descriptors are showing a significant correlation with each other, then we conclude that they capture structural information similarly. In order to predict the significance of the correlation between two descriptors, we perform the following approach:

Let M be a dataset of N descriptors and n samples. First, we generate bootstrap datasets, M_k, $k \dots B = 500$ possessing sample size $p = 200$, where $p < n$. Then, for each dataset, M_k, we perform a correlation test [30,31] between each pair of descriptors and obtained a *p value* p_{ij} for each pair. Thus, we test $N(N-1)/2$ hypotheses for all pairs. In order to control the false positives in the multiple hypothesis testing problem, we use the *bonferroni correction* method for multiple testing correction (MTC) [32] and obtained adjusted *p-values*. For each pair these adjusted *p-values* are denoted by q_{ij}. In order to decide whether the correlation between a pair is significant, we choose $\alpha = 0.00001$. After applying the correlation test and MTC, we obtain a binary matrix I_{M_k} which is defined follows:

$$I_{M_k}(i,j) = \begin{cases} 1 & \text{if } q_{ij} < = \alpha \\ 0 & \text{if } q_{ij} > \alpha \end{cases} \tag{9}$$

Finally we calculate a summary-statistic, T(i,j), for each pair of descriptors by averaging the values, i.e.,

Table 2. The optimal number of clusters for the three data-sets obtained by using consensus indices (CI).

Data-set	CI	# of clusters ($\|P\|$)	# Descriptors in each cluster
MS_{2265}	0.942	5	$\|c_1\|=863$, $\|c_2\|=22$, $\|c_3\|=18$, $\|c_4\|=1$, $\|c_5\|=15$
c_{15}	0.9878	16	$\|c_1\|=764$, $\|c_2\|=32$, $\|c_3\|=12$, $\|c_4\|=26$, $\|c_5\|=2$, $\|c_6\|=10$, $\|c_7\|=9$, $\|c_8\|=6$, $\|c_9\|=6$, $\|c_{10}\|=1$, $\|c_{11}\|=1$, $\|c_{12}\|=1$, $\|c_{13}\|=2$, $\|c_{14}\|=6$, $\|c_{15}\|=24$, $\|c_{16}\|=17$
N_8	1.00	7	$\|c_1\|=834$, $\|c_2\|=3$, $\|c_3\|=12$, $\|c_4\|=26$, $\|c_5\|=27$, $\|c_6\|=14$, $\|c_7\|=3$

The optimal numbers of clusters (for three data-sets) for a clustering solution P is represented by the set $P=\{c_1,c_2,\dots c_{|p|}\}$, where $|P|$ is the optimal number of clusters in the data.

doi:10.1371/journal.pone.0083956.t002

$$T(i,j)=\frac{\sum_{k=1}^{B=500} I_{M_k}(i,j)}{B}. \qquad (10)$$

In order to decide whether the correlation between two descriptors is strong, we choose a cut-off threshold $\alpha_{sum}=0.99$. If for the summary-statistic between two descriptors holds the inequality $T(i,j)>=\alpha_{sum}$, then we define two descriptors to be strongly correlated with each other. The descriptors in the set S

Table 3. The descriptors in predicted clusters (rows) overlapping with different categories of descriptors.

MS_{2265}						
Number of cluster	**connectivity indices**	**edge adjacency indices**	**topological indices**	**walk path counts**	**information indices**	**2D Matrix-based**
1	24	261	56	28	25	469
2	0	22	0	0	0	0
3	0	18	0	0	0	0
4	0	0	1	0	0	0
5	0	0	0	0	15	0
C_{15}						
1	17	214	51	22	34	426
2	4	21	3	2	0	2
3	3	6	1	2	0	0
4	0	26	0	0	0	0
5	0	2	0	0	0	0
6	0	10	0	0	0	0
7	0	9	0	0	0	0
8	0	6	0	0	0	0
9	0	6	0	0	0	0
10	0	1	0	0	0	0
11	0	0	1	0	0	0
12	0	0	1	0	0	0
13	0	0	0	2	0	0
14	0	0	0	0	6	0
15	0	0	0	0	0	24
16	0	0	0	0	0	17
N_8						
1	24	287	56	28	14	425
2	0	2	1	0	0	0
3	0	12	0	0	0	0
4	0	0	0	0	26	0
5	0	0	0	0	0	27
6	0	0	0	0	0	14
7	0	0	0	0	0	3

doi:10.1371/journal.pone.0083956.t003

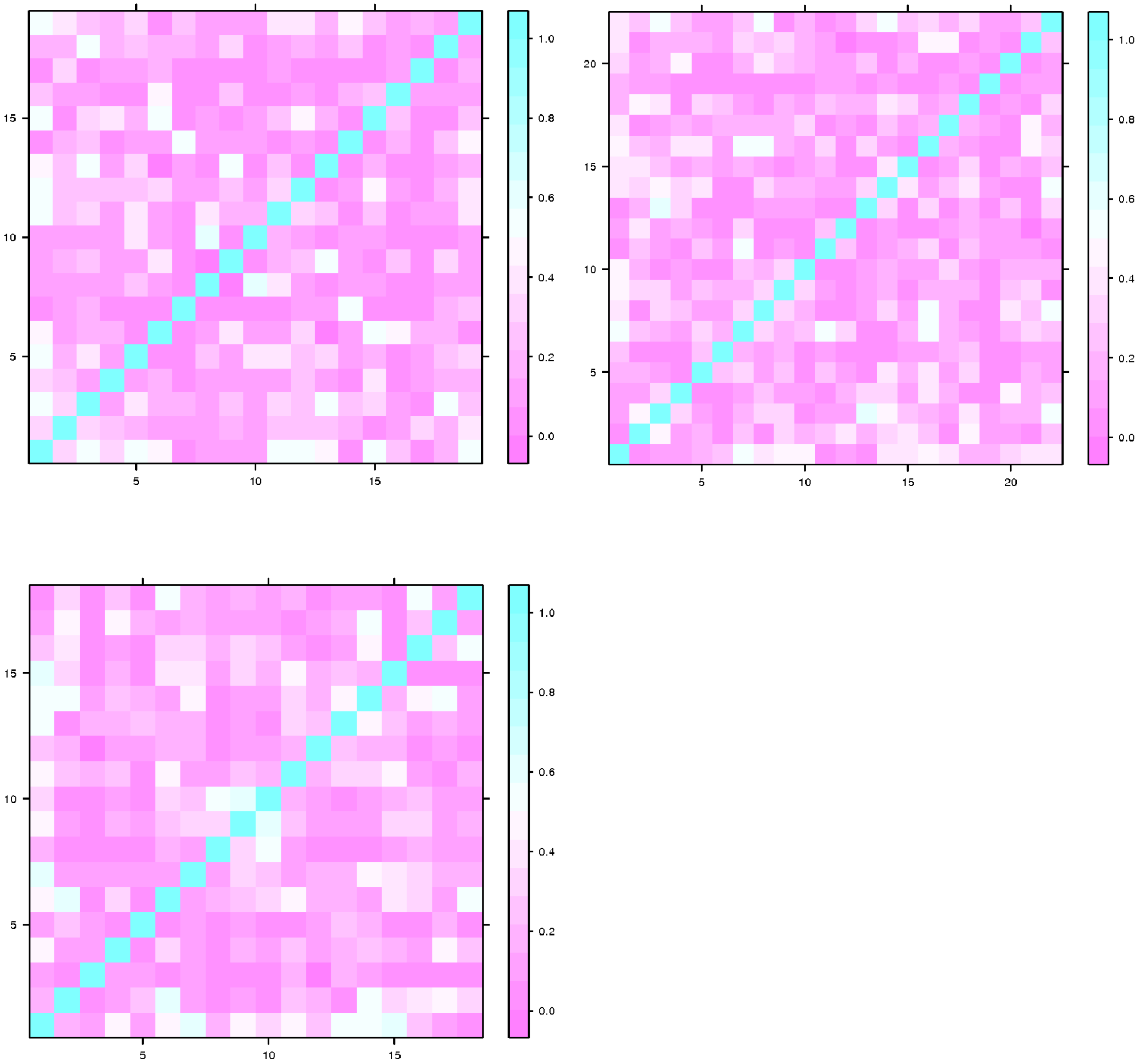

Figure 4. Levelplot of the correlation between the subset S for the three data sets, MS_{2265} (left), C_{15} (right), N_8 (bottom).
doi:10.1371/journal.pone.0083956.g004

have been chosen as follows. Suppose a descriptor D_i has a maximum number of summary-statistics greater or equal α_{sum} (i.e., $\#(T(i,j)>=\alpha_{sum})$, where $j=1\ldots i-1,i+1,\ldots|D|$), then the descriptor D_i is ranked first, and D_i is included in the subset S. Then we remove the descriptor D_i and the other descriptors with which D_i has summary-statistic $\geq\alpha_{sum}$. Then, we apply the same procedure to the remaining descriptors until we find any descriptor having maximum number of summary-statistics with remaining descriptors $\geq\alpha_{sum}$. Note that some of the descriptors do not have any summary-statistic greater than $\geq\alpha_{sum}$ with any of the other descriptors. These descriptors are described as lowly correlated descriptors and such descriptors are also included in the subset S.

This procedure reduces $|D|$ descriptors to $|S|$ descriptors. That means starting with a set of D descriptors, we hypothesize that the set S identify structural properties of a graph class distinctly. The remaining $|D|-|S|$ descriptors are showing stronger similarity (correlation) with at least one of the descriptor of set S.

Interpretation of the Results

The clustering of descriptors for three datasets is shown by Figure 1. In this figure, the six categories of descriptors are shown in different colors. The figure indicates that the descriptors of each categories have not been clustered correctly regarding their respective groups. For the external validity of the resulting clustering solution, we estimated NMI_{max} (normalized mutual information) [26] between reference cluster, $RC=\{c_1,c_2,c_3,c_4,c_5,c_6\}$ (the descriptors of six categories, $|RC|=6$, and $\{|c_1|=24,|c_2|=301,|c_3|=57,|c_4|=28,|c_5|=40,|c_6|=469\}$ are considered as the groups of the reference cluster) and the number of clusters of the clustering solution by cutting at different heights.

Table 4. Given the subset S; then, the remaining $|D|-|S|$ descriptors have at least one pair for which the summary statistic $T(i,j)$ is greater than $\alpha_{sum}=0.99$ with $|S|$ descriptors.

Data-set	Names of the descriptors
MS_{2265}	SM3_L, H_Dt, AVS_B.v., SM02_EA.dm., Eig11_AEA.bo., SpMAD_AEA.ed., CIC2, Eig13_AEA.bo., AVS_B.s., SM06_AEA.dm., Eig14_AEA.dm., MAXDP, J_Dz.v., BIC4, SpDiam_AEA.dm., SpMAD_X, PJI2, SpPosA_B.m., IDDE
C_{15}	SM2_B.s., PW4, Chi1_EA.ri., SM02_EA.dm., VE1_A, IC2, CENT, SM13_AEA.bo., Eig03_EA.bo., SM03_AEA.dm., VE3_Dz.p., piPC05, Eig04_AEA.bo., SpDiam_AEA.dm., piPC06, Eig02_AEA.dm., IVDE, MAXDP, PJI2, Eig05_AEA.dm., Chi0_EA.dm., Eig07_AEA.ed.
N_8	QW_L, TIE, VE3_B.i., BIC1, VE3_Dz.i., Eig10_AEA.dm., SpPosLog_B.m., SM03_AEA.dm., Eig11_AEA.ri., SM04_AEA.dm., CSI, VE1_Dt, Eig08_EA.ed., SpMaxA_AEA.bo., Yindex, Ram, IVDE, Chi1_EA.dm

doi:10.1371/journal.pone.0083956.t004

The estimated normalized mutual information is calculated by sampling the data $B=200$ times. Results for the three data-sets (average NMI) are shown in Figure 2. The average normalized mutual information plot between the reference cluster and the clusters created by performing average hierarchical clustering shows that they are quite dissimilar, that is the predicted clusters and the reference cluster are not similar at all. Also, the descriptors of different categories are strongly correlated with each other.

Next, we predict the optimal number of clusters, $P=\{c_1,c_2,\dots c_{|P|}\}$ by using consensus indices measure for different number of clusters generated by a clustering solution. The plots for the consensus indices for the three data sets are shown in Figure 3. The consensus indices are calculated for $k=2,\dots,200$ clusters. CI for different number of clusters for the three data-sets does not show an absolute maximum. Therefore we selected the first local maxima which gives the optimal number of clusters. The optimal number of clusters are shown with a dotted red line in the Figure 3. The consensus indices (CI) for the optimal number of clusters ($|P|$) and the total number of descriptors ($|c_i|$, where $i=1,\dots,|P|$) in each cluster for the three data-sets, MS_{2265}, C_{15} and N_8 are shown in Table 2. The optimal number of clusters are very little for all three data-sets and for all data-sets. The first cluster is the largest one which contains more than 80% of 919 descriptors. The cardinalities of the remaining clusters are smaller as they contain much less descriptors. The largest cluster for all three datasets contains descriptors from all six categories which means that most of the descriptors from different categories have a strong correlation among the descriptors and, therefore, they measure structural information similarly.

As a next step, we examine the so-called overlap between the optimal number of clusters shown in Table 2 and the six categories of descriptors. That means we have to determine how many different descriptors are distributed over different groups (belonging to the optimal number of clusters). This distribution over different clusters could give some information namely which category might capture structural information of the graphs more uniquely than others. The results are shown in Table 3 and we are going to interpret these results as follows. The intersection of the descriptors between the optimal clusters and the categories of descriptors show that the edge adjacency indices are grouped into different cluster for all three data-sets in comparison to the remaining categories. The 2D Matrix-based descriptors are grouped into different clusters by using C_{15} and N_8. The information indices are grouped into two different clusters by using all three data-sets. The measures from the category walk path counts and topological indices are grouped into different clusters by using C_{15} only. This shows that these descriptors behave differently on trees. The overlap indicates that the group of edge adjacency indices contains more descriptors which capture structural information of the graphs differently compared to other categories.

Next, we find a subset of descriptors $S \subseteq D$, $|D|=919$. The main idea is to find a smaller set of descriptors which are little correlated with each and, hence, those graph measures captures structural information uniquely. If they would be strongly correlated, they would capture similar structural information of the graphs. Importantly, the remaining descriptors have much stronger correlation with them. The procedure to obtain a subset of descriptors $S \subseteq D$ is described in the section 'Methods and Results'. We obtained $|S|=\{19,22,18\}$ for MS_{2265},C_{15},N_8 datasets shown in Table 4. The levelplot of correlation for the subset of descriptors of three data-sets are shown in Figure 4. For all three data-sets, we can clearly see that the descriptors of these subsets are not strongly correlated. These subset of descriptors for all three data-set might detect structural features of the molecular graphs uniquely.

Moreover we now examine for all data-sets which descriptors from S (shown in Table 4) belong to which group out of the six categories of descriptors. The results are summarized in Table 5. For each data-set, we start with a different number of descriptors for the different categories. The subset S does not contain any descriptor from the *connectivity indices* for all three data-sets, however, only two descriptors from *walk path counts* are contained in S by using C_{15}. Two, four and three descriptors from the category *topological indices* are contained in S for all three data-sets. Three, two and three descriptors from the category *information indices* are in S for three data-sets. Seven, three and three descriptors from the category *2D Matrix-based* are in S for three data-sets. Seven, eleven and seven descriptors from the category *edge adjacency indices* are in S for MS_{2265}, C_{15}, N_8. These are the maximal numbers of descriptors compared to other categories of descriptors. The large occurrence of the descriptors from the category *edge adjacency indices* shows again that these descriptors quantify structural information more uniquely than others.

Table 5. The number of descriptors of S which belong to six different categories by using three data sets.

Descriptor category	MS_{2265}	C_{15}	N_8
Connectivity indices	0	0	0
Edge adjacency indices	7	11	7
Topological indices	2	4	3
Walk path counts	0	2	0
Information indices	3	2	3
2D Matrix-based	7	3	5

doi:10.1371/journal.pone.0083956.t005

Table 6. The overlap between S and the predicted clusters (rows).

MS_{2265}	
Number of cluster	**Descriptors of S**
1	SpMAD_AEA.ed., SpDiam_AEA.dm., Eig13_AEA.bo., Eig14_AEA.dm., MAXDP, IDDE, SM3_L, SpMAD_X, H_Dt, J_Dz.v., SpPosA_B.m., AVS_B.v., AVS_B.s.
2	SM02_EA.dm.
3	SM06_AEA.dm., Eig11_AEA.bo.
4	PJI2
5	CIC2, BIC4
C_{15}	
1	SpDiam_AEA.dm., Eig03_EA.bo., Eig07_AEA.ed., Eig02_AEA.dm., PW4, IC2, SM2_B.s.
2	Chi1_EA.ri.
3	CENT, piPC05
4	SM02_EA.dm.
5	Chi0_EA.dm.
6	Eig04_AEA.bo.
7	SM13_AEA.bo.
8	SM03_AEA.dm.
9	_
10	Eig05_AEA.dm.
11	PJI2
12	MAXDP
13	piPC06
14	IVDE
15	VE1_A
16	VE3_Dz.p.
N_8	
1	Eig08_EA.ed., Eig10_AEA.dm., Eig11_AEA.ri., CSI, TIE, Yindex, QW_L, SpMaxA_AEA.bo., IVDE, SpPosLog_B.m.
2	Chi1_EA.dm., Ram
3	SM03_AEA.dm., SM04_AEA.dm.
4	BIC1
5	VE3_B.i.
6	VE3_Dz.i.
7	VE1_Dt

doi:10.1371/journal.pone.0083956.t006

Also, we examine the overlap between the descriptors from S and the descriptors in the found clusters; the intersections between them are shown in Table 6. Interestingly, at least one descriptor (for all data-sets) overlap with the descriptors of each cluster, except for the ninth cluster by using C_{15}. The overlap with the found clusters show that the measures contained in S (for three data-sets) have the potential to quantify unique structural features of molecular graphs.

Summary and Conclusions

In this paper, we have evaluated 919 Dragon descriptors to investigate to what extent these measures quantify structural information of molecular graphs uniquely. From our analysis, it is clear that measures which are strongly correlated are not useful as they capture structural information similarly. From this, the question of determining the usefulness or quality of topological indices arises.

We found by calculating the information-theoretic quantity NMI that the used six categories of descriptors are strongly correlated with other categories of descriptors. This indicates that despite being categorized into different groups, these descriptors are providing similar information. From this, one can conclude that many of them they have been introduced in an unconsidered manner. Again, the question how useful such indices are seems to be quite important and deserves further attention.

By using all three data sets, the most suitable descriptor subset S contains those measures which have the largest number of significant correlations with the remaining descriptors but they are not significantly correlated with each other. S forms a reduced set of descriptors (the original sets contains 919 descriptors) and their sizes are feasible approximations of the effective dimension of the descriptor space by using all three datasets. For each individual data set, we found the size of S to be 19 (MS_{2265} dataset), 18 (N_8 dataset) and 22 (C_{15} dataset). Because most of the descriptors we have used are redundant, i.e., they are highly correlated, the estimation of the effective dimension is an intriguing problem. In our context, the dimension is the number of different descriptors among all. By performing our analysis, we obtained a lower bound on the dimension of descriptors space regarding the different classes. Note that these descriptors (the ones in S) depend on the used data set. By inspecting these subsets, we see that the majority thereof are from the category of the edge-adjacency indices. This implies that the edge-adjacency based descriptors can capture more structural diversity when quantifying structural properties of molecular graphs. As another result of this paper, we see that it would not be appropriate to select descriptors more or less randomly for QSAR problems. Neither the random selection nor using all available descriptors would be appropriate as demonstrated in our paper. To tackle this problem, we suggested a statistical analysis evidenced by using clustering. Again, we note that our method applied to six categories of descriptors reduces the descriptor space for three datasets. In this paper we have presented a statistical approach by using correlation test to select a smaller subset of descriptors which captures information similarly. By employing bootstrapping and a probabilistic measure for the selection process, we have identified the most informative set of descriptors. As seen, a set of descriptors can cover a dataset best, but studying this important issue in depth might be future work.

Author Contributions

Analyzed the data: ST FES MD. Wrote the paper: MD FES ST.

References

1. Todeschini R, Viviana C (2010) Molecular Descriptors for Chemoinformatics (2nd edition). Wiley- VCH, Weinheim, Germany.
2. Mauri A, Consonni V, Pavan M, Todeschini R (2006) Dragon software: An easy approach to molecular descriptor calculations. MATCH Communications in Mathematical and in Computer Chemistry 56: 237–248.
3. Basak SC, Magnuson VR, Niemi GJ, Regal RR (1988) Determining structural similarity of chem- icals using graph-theoretic indices. Discrete Applied Mathematics 19: 17–44.
4. Basak SC, Balaban AT, Grunwald GD, Gute BD (2000) Topological indices: their nature and mutual relatedness. Journal of Chemical Information and Computer Sciences 40: 891–8.
5. Basak SC, Gute BD, Balaban AT (2004) Interrelationship of major topological indices evidenced by clustering. Croatica Chemica Acta 77: 331–344.
6. Todeschini R, Cazar R, Collina E (1992) The chemical meaning of topological indices. Chemometrics and Intelligent Laboratory Systems 15: 51–59.
7. Taraviras SL, Ivanciuc O, Cabrol-Bass D (2000) Identification of groupings of graph theoretical molecular descriptors using a hybrid cluster analysis approach. Journal of Chemical Information and Computer Sciences 40: 1128–1146.
8. Randić M (1975) On characterization of molecular branching. Journal of the Americam Chemical Society 97: 6609–6615.
9. Devillers J, Balaban AT (1999) Topological indices and related descriptors in QSAR and QSPR. Amsterdam: Gordon & Breach, The Netherlands.
10. Hosoya H (1971) Topological index. a newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. Bulletin of the Chemical Society of Japan 44: 2332–2339.
11. Khalifeh M, Yousefi-Azari H, Ashrafi A (2009) The first and second zagreb indices of some graph operations. Discrete Applied Mathematics 157: 804–811.
12. Dehmer M, Mowshowitz A (2011) A history of graph entropy measures. Information Sciences 1: 57–78.
13. Mowshowitz A (1968) Entropy and the complexity of the graphs I: An index of the relative complexity of a graph. The Bulletin of Mathematical Biophysics 30: 175–204.
14. Bonchev D, Trinajstić N (1977) Information theory, distance matrix, and molecular branching. The Journal of Chemical Physics 67: 4517–4533.
15. Janežić D, Miležević A, Nikolić S, Trinajstić N (2007) Graph-Theoretical Matrices in Chemistry. Mathematical Chemistry Monographs. University of Kragujevac and Faculty of Science Kragujevac.
16. Kier L, Hall L (1976) Molecular connectivity in chemistry and drug research. Medicinal chemistry. Academic Press. Available: http://books.google.co.in/books?id = gQfwAAAAMAAJ.
17. Kier L, Hall L (1986) Molecular connectivity in structure-activity analysis. Chemometrics series. Research Studies Press. Available: http://books.google.co.in/books?id = qpA0AAAAMAAJ.
18. Kier L, Hall L (1999) Molecular Structure Description: The Electrotopological State. Academic Press. Available: http://books.google.co.in/books?id = QixqQgAACAAJ.
19. Karelson M (2000) Molecular descriptors in QSAR/QSPR. New York: Wiley-Interscience.
20. Dehmer M, Varmuza K, Borgert S, Emmert-Streib F (2009) On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures. Journal of Chemical Information and Modeling 49: 1655–1663.
21. Dehmer M, Grabner M, Varmuza K (2012) Information indices with high discriminative power for graphs. PLoS ONE 7: e31214.
22. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning, volume 1. Springer.
23. Halkidi M, Batistakis Y, Vazirgiannis M (2001) On Clustering Validation Techniques. Journal of Intelligent Information Systems 17: 107–145.
24. Sokal RR, Rohlf JF (1962) The comparison of dendrograms by objective methods. Taxon 11: 33–40.
25. Halkidi M, Batistakis Y, Vazirgiannis M (2002) Clustering validity checking methods. ACM SIGMOD Record 31: 19.
26. Kvalseth, Tarald O (1987) Entropy and correlation: Some comments. Ieee Transactions On Systems Man And Cybernetics 17: 517–519.
27. Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clustering comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research 11: 2837–2854.
28. Zhongmou L, Hui X, Xuedong G, Junjie W (2010) Understanding of internal clustering validation measures. 2010 IEEE International Conference on Data Mining 0: 911–916.
29. Hubert L, Arabie P (1985) Comparing partitions. Journal of Classification 2: 193–218.
30. Best DJ, Roberts DE (1975) Algorithm AS 89: The upper tail probabilities of Spearman's rho. Journal of the Royal Statistical Society Series C (Applied Statistics) 24: 377–379.
31. Hollander M, Wolfe D (1999) Nonparametric statistical methods (Wiley Series in Probability and Statistics). John Wiley & Sons.
32. Dudoit S, Van Der Laan MJ (2007) Multiple Testing Procedures and Applications to Genomics. Springer.