Finding Key Members in Compound Libraries by Analyzing Networks of Molecules Assembled by Structural Similarity

Zsolt Lepp,*,† Chunfei Huang,‡ and Takashi Okada§

Chemical Biology Department, Riken ASI, College of Computer Science and Technology, Jilin University, and Kwansei Gakuin University

Received March 24, 2009

Characterization of chemical libraries is an essential task in everyday chemoinformatics practice. This study describes some potential uses of network visualization and analysis methods to identify distinguished members of compound libraries. Molecules were ordered into networks by their structural similarity defined by molecular fingerprints. Various properties of such networks were examined. It was shown, that the correlation methods used to calculate the similarity between two structures radically determined the topology of networks. From the same set of molecules, the Russel−Rao and the Baroni−Urbani methods created sparser and denser networks, respectively, than using the Tanimoto method. Central nodes, corresponding to central compounds in the libraries, were determined for some example data sets. It was shown by the case of adenosine A1, A2, and dual antagonists that the methods used to identify central nodes could be divided into two groups: (1) centrality methods, exemplified by the centroid centrality, which could pick up structures that were the most similar to the largest number of other molecules and (2) group, exemplified by betweenness centrality, that could identify molecules that had intermediate structures between some homogeneous subsets of the library. The latter method gave significantly higher ranks to dual adenosine antagonists, hinting the suitability of this measure to identify molecules with multiple activities. Some practical applications of the method for clustering of and sample selection from chemical libraries are presented. In the frame of the study, a Jchem plug-in has been developed to the Cytoscape network visualization software, which makes the visual observation of molecular networks more convenient. The plug-in is included in the Supporting Information of the article for free usage.

1. INTRODUCTION

In recent years, the creation and analysis of networks has come into focus in various scientific domains, such as the social sciences and biosciences.[1,2] This has driven the rapid development of novel theories and software tools. In order to gather meaningful information from networks, the principles of graph theory have been utilized. Graph theory has also been used extensively for chemoinformatics research,[3] but the main focus has been describing atoms and bonds in molecules.

Networks made of independent molecules play an increasingly important role in the integration of chemical and biological sciences through the construction of metabolic networks.[4−10] However, there are few examples of their use in chemoinformatics research. The development and mindshare of network analysis methods offer an easy path to carry out more studies on the applicability of these methods for various chemoinformatics tasks.

The subjects of network analyses are ensembles in which members are connected by their relationships to each other. Typically, the members of compound libraries are not selected randomly, but rather, they are chosen for specific reasons (for example, synthesis methods or activity). These libraries may be natural choices for the creation and analysis of networks.

The most obvious approach is to organize molecules into networks based on structural similarity. It is expected that the graph would contain certain patterns that highlight some structural aspects of the data sets; for example, molecules with a given physiological activity are grouped together. Traditionally, the approach of minimum spanning trees has been applied to a limited extent to cluster chemical libraries and to select quantitative structure−activity relationship (QSAR) subseries from a larger data set.[11−17]

Recently, studies have been conducted to relate protein similarity to the structural similarity of the ligands of those proteins.[18] Although the resulting networks contained only proteins, the ligand molecules were represented by their average structural similarity; thus, they were incorporated indirectly into the graphs. This article was followed-up by further analyses to further study the distribution of the active molecules among target classes.[19] In the current study, we present additional potential applications of applying network analysis methods to networks based on molecular structural similarity in order to answer questions that are asked frequently in chemoinformatics. More emphasis is put toward the analyses of the topologies of molecular similarity networks.

When analyzing a chemical library, it is important to find molecules that hold central positions in the chemical space

* Corresponding author. Tel.: +81-48-467-4839. Fax.:+81-48-462-1353. E-mail: zsolt.lepp@riken.jp.
† Riken ASI.
‡ Jilin University.
§ Kwansei Gakuin University.

10.1021/ci9001102 CCC: $40.75
Published on Web 10/30/2009

occupied by the library. One way to perform this task is to perform clustering. The cluster centers (and/or nearby points) correspond to the central structures.

The identification of centrally located nodes and the analysis of their roles in the graph are also critical to the study of a network. This analogy leads to the idea that the central nodes in molecular similarity networks could be the molecules in the central positions of the chemical space. There are various methods to find network centralities, and thus, it would be a convenient and effective way to analyze compound libraries.

A method to identify key compounds[20] resembles finding central nodes in a graph by determining the centrality values of the nodes.

One of the most recent and thorough applications of molecular networks for drug research was reported by Santana et. al to predict monoamine oxidase (MAO) inhibitor activity.[21] The authors created a network of molecules and showed that the activity class(es) of molecules can be well-predicted by their connectivity to hubs in the network with known activity.

Here, we present some additional potential methods to identify key molecules in chemical libraries using network analysis methods. Different types of molecular similarity based networks are introduced. Various approaches are shown to assemble these networks. These approaches can be used to select representative subsets from libraries by effectively taking the distribution of structural features into consideretion. The effects of the variation of certain parameters, such as similarity threshold and correlation methods, on network topology is also disscussed.

2. METHODS

2.1. Terms Used to Create and Analyze the Molecular Networks. As molecular networks are used infrequently for chemoinformatics research, the following section explains the methods used in the current work.

2.1.1. Minimum Spanning Trees. A spanning tree of a graph is a subgraph, which is a tree and connects all of the nodes (also called vertices) together. A single graph can have many different spanning trees. A minimum spanning tree or minimum weight spanning tree is then a spanning tree with a weight less than or equal to the weight of every other spanning tree.

2.1.2. Threshold Network. The threshold network is a graph in which every edge (or arch) has a weight greater than a predefined threshold value. In molecular structure similarity networks, the weight has been defined by either molecular similarity or dissimilarity values.

2.1.3. Network Layouts. Although the nodes in a graph can be drawn anywhere, it is generally useful (or required) to present them in a specifically ordered structure. For example, it should be ensured that nodes do not overlap and stay at least a certain distance from one another, or that each node appears in a specific position relative to the others. The process to determine such favorable positions is called the layout application. There are a number of methods to calculate layouts, based on different ideas. The layouts can be trees, hierarchical structures, based on physical analogy (force-directed layouts), or combinations thereof.

For different problems, some layouts may present the information in more meaningful ways. We have found that for the networks of molecules presented in this study, the force-directed layouts, particularly the one designated organic, were the most suitable. For visualizing clusters in minimum spanning trees, the orthogonal layout could also be very meaningful. This does not mean that other methods are not usable, and personal preference might also influence the choice of layout.

2.1.3.1. Spring Layout. The spring layout is a force-directed layout algorithm designed to simulate a system of particles, each with some mass. The vertices simulate mass points repelling each other, and the edges simulate springs with attractive forces. The algorithm moves through a number of iterations trying to minimize the energy of this physical system. This means that a certain number of iterations are required to bring the system close to equilibrium, but further iterations will accomplish very small changes and simply waste CPU time.

2.1.3.2. Organic Layout. This force-directed layout is an implementation of a simulated annealing layout, which describes the following criteria as favorable in a graph layout: (1) distributes nodes evenly; (2) makes edge lengths uniform; (3) minimizes cross-crossings, and (4) keeps the nodes from coming too close to the edges. These criteria are translated into energy cost functions in the layout. Nodes or edges breaking these criteria create a larger cost function; the total cost they contribute is related to the extent that they break it. The idea of the algorithm is to minimize the total system energy. Factors are assigned to each of the criteria describing the importance of that criterion. The organic layout corresponds well to the clustering of the compound libraries, as exemplified by adenosine antagonists A1 and A2 (see Results and Discussion).

2.1.3.3. Force-Directed Layout. This layout positions graph elements based on a physics simulation of interacting forces; by default, nodes repel each other, edges act as springs, and drag forces (similar to air resistance) are applied. This algorithm can be run for multiple iterations for a run-once layout computation or repeatedly run in an animated fashion for a dynamic and interactive layout.

The running time of this layout algorithm is the greater of $O(N \log N)$ and $O(E)$, where N is the number of nodes and E is the number of edges.

2.1.4. Network Centrality Measures. Various centrality methods exist to measure the central role of a node in a graph. In this report, two of these were used and are described below. A more extensive list of methods can be found, for example, on the CentiScaPe home page.[22] To calculate the following centrality values, it is necessary to first calculate the shortest paths among all pairs of nodes in the network.

2.1.4.1. Betweenness Centrality. The betweenness centrality of a node is the portion of all of the shortest paths in a network that pass through the given node. To calculate this value, all of the shortest paths are determined between each possible pair. The higher the value, the more important a node is in linking together a number of other nodes.

2.1.4.2. Centroid. This is the most complex centrality index. A centroid can be defined in many ways, such as one based on connectivity or other centrality measures. The definition to determine the centroid given by CentiScaPe is as follows:

A centroid weights the distance between two given nodes by all of the other shortest paths in the graph. It is computed by focusing the calculation on pairs of nodes (v,w) and systematically counting the nodes that are closer (in term of shortest path) to v or to w. The calculation proceeds by comparing the node distance from other nodes with the distance of all other nodes from the others, such that a high centroid value indicates that a node v is much closer to other nodes. Thus, the centroid value provides a centrality index always weighted with the values of all other nodes in the graph. In other terms, a node v with the highest centroid value is the node with the highest number of neighbors separated by the shortest path to v. The centroid value suggests that a specific node has a central position within a graph region characterized by a high density of interacting nodes.

2.2. Data Sets. The three data sets used in the present study included 400 CRF antagonists, 479 adenosine antagonists, and 5000 randomly selected molecules. All molecule sets were selected from the 2004/02 version of MDDR.[23] Molecules were standardized by removing counterions, dearomatizing rings, and neutralizing ionic groups using JChem[24] software.

2.3. Similarity Measures. The molecules were characterized by molecular fingerprints. Chemistry development kit (CDK) fingerprints were calculated using the RCDK[25] module of R project, which integrates the open-source CDK library with the R environment. Default parameters were used (size, 1024 bit; search dept, 6). The Baroni−Urbani/Buser correlation coefficients to measure molecular similarities were determined by the Fingerprint module of R.[26] The hcluster[27] python library was used for the other three correlation methods. The default choice was the hcluster library due to its significantly faster speed; however, it lacks the Baroni−Urbani method (in the 0.1.6 version). Additionally, for the clustering comparisons, the ScreenMD module of Jchem was used to calculate similarity matrices.

The following four coefficients were used.

Russel−Rao

$$a/n$$

Jackard/Tanimoto

$$a/(a + b + c)$$

Baroni−Urbani/Buser

$$\mathrm{sqrt}(ad + a)/(\mathrm{sqrt}(ad) + a + b + c)$$

Yule

$$(ad - bc)/(ad + bc)$$

Where a is the number of bits common to both compounds, b and c are the number of bits unique to the query or database compound, respectively, d is the number of bits found in neither compound, and n is the fingerprint size $(n)a + b + c + d$.

2.4. Network Analyses. For building, visualizing, and analyzing networks, the Cytoscape,[28] Sagemath,[29] and the NetworkX[30] python library open-source programs were used. To view the molecular structures of network nodes, the Marvin module of Jchem through our own Cytoscape plug-in was used.

2.4.1. Minimum Spanning Trees. Minimum spanning trees were calculated by the Kruskal method as implemented in Sagemath (which uses the integrated BGL library for this task).

2.4.2. Threshold Networks. A python script was used to create the threshold networks from the similarity matrix and to determine the sizes of the components as a function of threshold. The NetworkX python APIs were used by the script.

2.4.3. Network Centrality. Centralities were determined using the CentiScaPe plug-in of Cytoscape. This can calculate nine different centrality measures, of which two (centroid, betweenness) were used for detailed analyses in the current study.

2.4.4. Network Visualization. Cytoscape was used to visualize the network graphs. Because there was no convenient way to view the molecular structures of molecules in the networks, we created a plug-in for Cytoscape in Java, which linked the molecular viewer MarvinView of the JChem software into Cytoscape.

A screenshot is shown in Figure 1. Using the plug-in, the molecular structure of the selected network node(s) can be seen on the control panel (as shown in Figure 1) or in a pop-up window. The mechanism of the plug-in is such that the structure of a compound, in smiles format, is stored as a network node attribute. The plug-in directs the smiles string to MarvinView, which creates the geometry and displays the structure. All of the structure visualization and manipulation features of JChem can be used (such as logP calculation). The plug-in and detailed installation instructions may be found in the Supporting Information. Various other scripts were used to automate the creation of networks. These will be made available on the Internet in the near future. Until then, researchers who are interested are welcome to contact the authors.

2.5. Clustering. To perform network clustering, the MCODE (Molecular Complex Detection) plugin of Cytoscape was used. It is a novel graph theoretic clustering algorithm, that was developed to detect densely connected regions in large protein−protein interaction networks that may represent molecular complexes. The method is based on vertex weighting by local neighborhood density and outward traversal from a locally dense seed protein to isolate the dense regions according to given parameters. The detailed description of the algorithm can be found in the publication by Bader et al.[31]

3. RESULTS AND DISCUSSION

3.1. Minimum Spanning Trees of CRF Inhibitors. The minimum spanning tree method is a convenient way to simplify a network. A very useful feature of the method is the lack of randomness; there is only one minimum tree, and it depends only on the data set. Furthermore, in a spanning tree, there is only one route between any pair of nodes. Below, we present an interesting and not-yet-discussed feature of calculating minimum spanning trees for molecular data sets.

This feature is based on the fact that in order to define the weights of an edge in a graph, either the (1) dissimilarity or (2) similarity value between two connected compounds can be used. Not only are the structures of the spanning trees very different, but the two kinds of graphs can be useful for different purposes. In case 1, the tree is useful for clustering or selecting QSAR series, but case 2 could be used for diversity analysis, or to find the most dissimilar compounds.

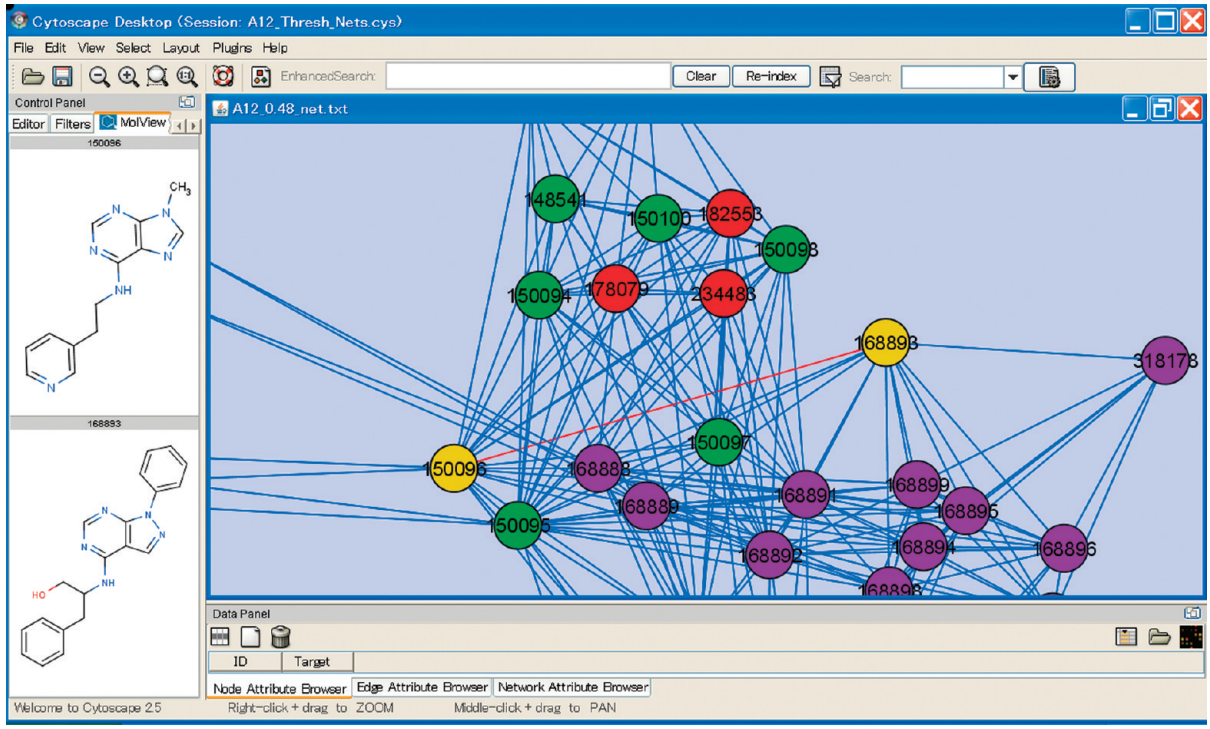


Figure 1. Integration of the MarvinView molecule visualizer into Cytoscape. The molecular structures of the two selected nodes (yellow) are drawn by the MarvinView plugin (on the left).

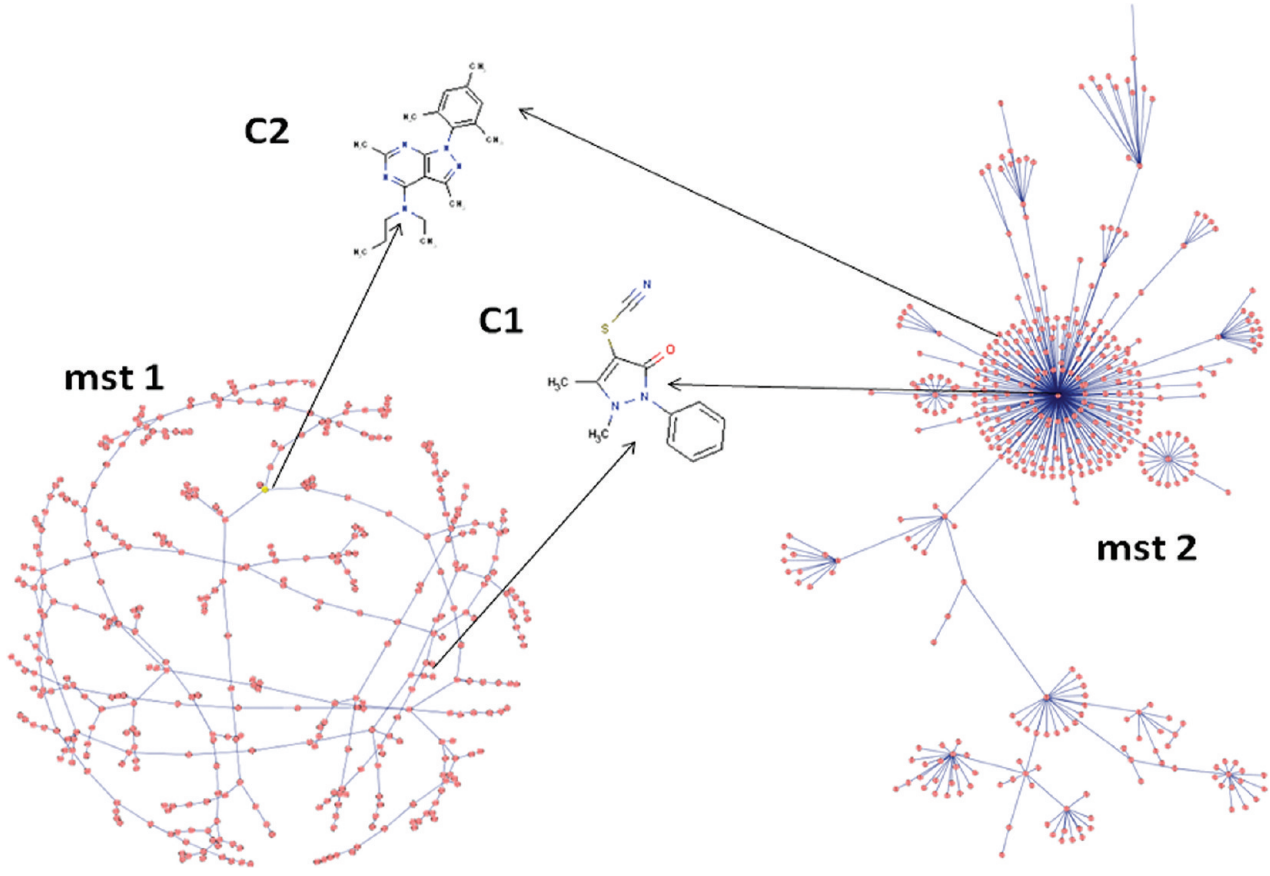


Figure 2. Minimum spanning trees of CRF inhibitors.

We have illustrated this point using a data set of CRF antagonist antidepressants. This data set was selected because it is a relatively homogeneous set, with the chemical structures relatively different from other drug molecules.

Shown in Figure 2 is the minimum spanning tree of the dissimilarity matrix of the CRF data set (**mst 1**). The spring layout was used, as we found that it reflected well the shape of a tree for this data set (the organic layout gave a similar

result, but the spring layout was a bit more expressive). The edge weights are the dissimilarities between the molecule pairs. This means that if two molecules are similar, then the dissimilarity value is low, and the two nodes are located close to each other in the tree. The visualization power of the method is easily perceptible; by changing the layout of the network to orthogonal (for the graph, see the Cytoscape document in the Supporting Information), we might view the library from an alternative perspective.

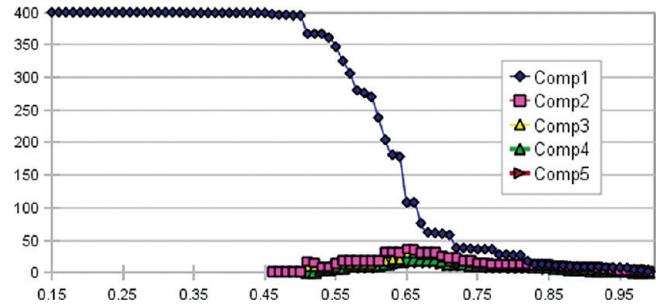
In **mst 1**, the similar molecules are located in close proximity. However, sometimes the opposite approach might be useful, e.g., when looking for the most dissimilar lead molecules or performing diversity analyses. In this case, we can use the same method, but replace dissimilarity values with similarity (1 − dissimilarity). The tree generated by such an approach is shown in Figure 2 **mst 2**. This is the inverse of **mst 1**. In this tree, each molecule is connected to the most dissimilar ones. The organic layout was used to draw the graph of **mst 2**, as it provided the most aesthetic, balanced view.

As the distances between the nodes in **mst 2** represent similarities, the lower the similarity between two given molecules, the closer the nodes. For example, compound **C1** is the most dissimilar to the largest number of compounds, thus it has a central position. The central nodes of **mst 2** can be considered to be the centers of dissimilarity clusters. This graph gives an interesting overview of the data set, as it shows the opposite of the common view of clustering. The central molecules such as **C1** are small branches in **mst 1**, and we could not deduce the central role they play in **mst 2**.

This method might provide an interesting way to select a diverse set of molecules. The most unique structures are located in the center. The molecules in close proximity to those are the most dissimilar molecules to those central points and might be structurally quite diverse (although not necessarily; the structures might be very similar, as well). Nevertheless, the method gives different results than other clustering methods and may provide some additional information. For example, it can be useful for understanding new lead generation processes. Following the path between any given node on **mst 2** orders the molecules by the largest structural variation. This is because the neighbors of each molecule are the most dissimilar ones; if one follows a path in this spanning tree, the structural variations will be significant in each step. Additionally, it might be useful to analyze the structural evolution of the lead molecules for a given target. On the other hand, **C2** in **mst1** is located at the cross of three branches, thus it is probably a central molecule in the library. However, in **mst 2** it has no central role, as **C2** is not a unique structure like **C1**. It can be a problem, that two molecules, which have a path from the same molecule may be similar even if two molecules are dissimilar to the molecule, respectively. Yet, the method can be very useful to extract series of most dissimilar molecules, by following the most connected molecules and selecting a number of neighboring nodes.

Another potential use of **mst 2** is to cluster molecules based on dissimilarity. The topologies of **mst 1** and **mst 2** are very different. The are few central nodes on the former tree, which makes it unsuitable to be used for clustering. On the other hand, the second tree consists of clusters of various sizes with the most dissimilar compounds in the central positions.

Chart 1. Sizes of the Five Largest Components in the Threshold Networks of CRF Inhibitors As a Function of Threshold[a]



[a] *x*-axis: similarity threshold. *y*-axis: size of network components.

The flexibility of the method should be emphasized; although the same method was used, **mst 1** and **mst 2** supply solutions to different chemoinformatics tasks. The former can be used to select homogeneous series, and the latter is useful for dissimilarity-based clustering.

The minimum spanning tree method also has some drawbacks. Probably the most serious disadvantage for chemoinformatics is the loss of important information. Traditionally, the method has been used successfully for tasks such as designing a telephone cable network among locations. For this application, laying the minimum amount of cables is desirable. In contrast, in the case of a molecular similarity network, all of the edges might represent important information: the similarity between any given pair of compounds. By extracting a tree subgraph, much information may be lost. Therefore, it might also be useful to study the full network. However, it would not be convenient to include all of the possible edges as the network would be too dense, nor is it very useful to directly connect very different molecules. The obvious solution is to create a threshold network in which only the edges that have a weight greater than a given threshold are drawn.

3.2. Threshold Networks of the CRF Inhibitors. Next, some threshold networks assembled from the CRF antagonists are introduced.

The topology of a threshold network depends heavily on the value of the threshold. The most basic question is whether, using a given threshold, all of the nodes are connected to form one large network or it is broken down into smaller components. Some nodes may not be connected at all. This must be verified before analyzing the network.

The five curves illustrated in Chart 1 show the sizes of the five largest network components in the CRF antagonist network, as a function of threshold. The *x* axis is the threshold, and the *y* axis is the number of connected nodes. For threshold values of up to 0.33, the network has only one component and all 400 nodes are connected. Beyond that threshold, one molecule leaves the network, followed later by others, but at a threshold of 0.50, 395 nodes remain connected. However, after this threshold, the network quickly disintegrates. Instead of breaking down into a few components of comparable sizes, the nodes leave the network one by one or in small groups. The size of the second largest component is never significant when compared to that of the largest one (Chart 1). This is very convenient for our purposes, because the methods to determine network cen-

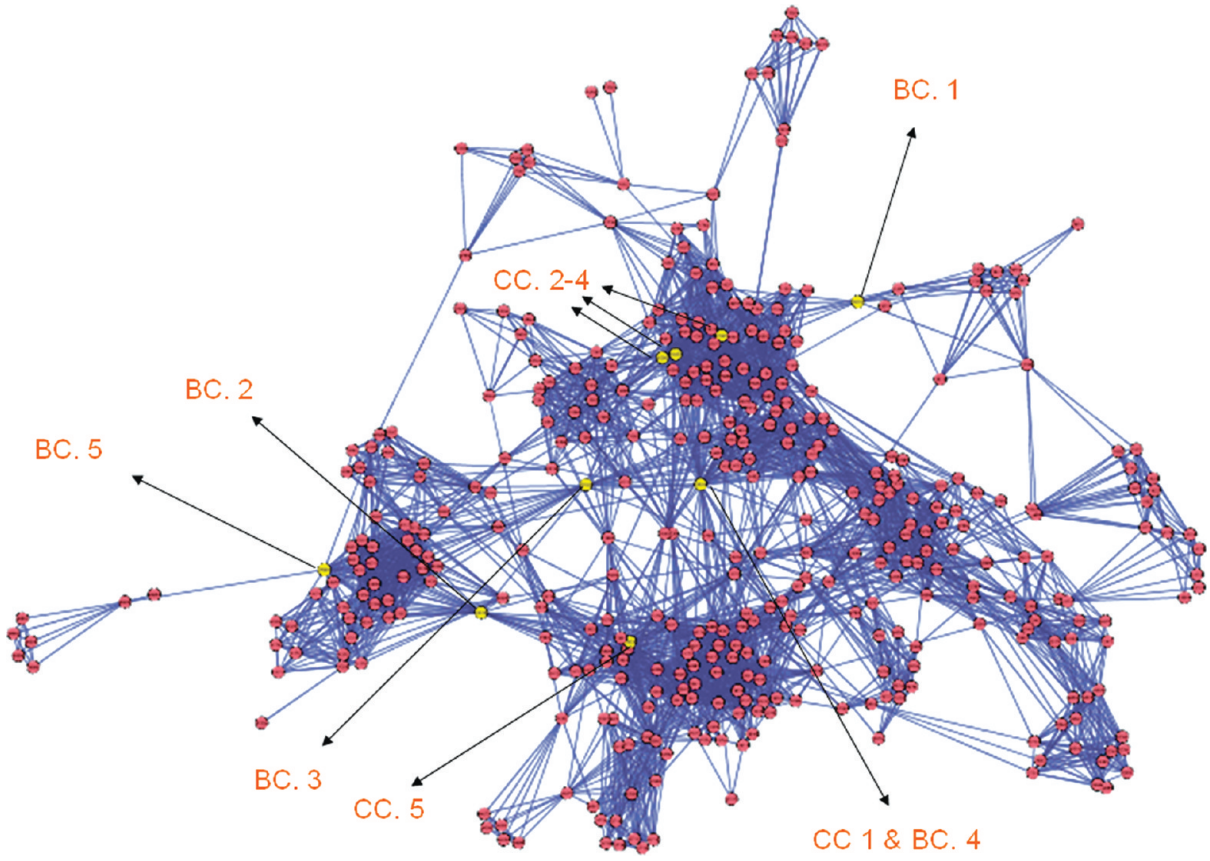


Figure 3. Threshold networks of CRF inhibitors. The threshold value to decide which edges to draw has been a Tanimo similarity of 0.49. The nodes with the top 5 centroid and betweenness centralities are named CC 1−5 and BC 1−5, respectively. These nodes are depicted by arrows, and they are yellow.

tralities might not work reliably for multicomponent networks. At least in the case of this data set, we could include nearly all (395 of 400) of the compounds into one network. It is also best to include as many data points as possible, because by increasing the threshold, the diversity of the molecules in the network decreases, and obviously, the most dissimilar structures leave first.

Thus, the best threshold to create a network for analysis would be 0.50, just before the sharp disintegration begins. Using this approach, we could also reduce the number of edges to the necessary minimum, reducing the complexity and redundancy.

3.2.1. Graphical Representation of the CRF Antagonist Threshold Network. The 0.50 threshold network consists of 395 nodes and 2865 edges (Figure 3). To present the network graphically, the organic layout was applied. This layout seems to be very informative for the visual observation of molecular networks because the strongly interconnected vertices (especially those interconnected by edges of low dissimilarity values) aggregate together. It is known that the organic layout corresponds to the highly interconnected clusters in a network. For chemical libraries, this is important because the researchers cannot only overview the structure of a library immediately, but it also might be useful for clustering in a novel way. For example, the Tanimoto dissimilarity values between two molecules might be replaced by the spatial distance between those in the graph. These distance values can then be used to calculate clusters.

3.2.2. Centrality Values in the CRF Antagonist Threshold Network. The CentiScaPe plug-in for Cytoscape integrates nine different centrality measures. We selected two measures, the centroid value and betweenness centrality, for use in our analysis. At first, all nine of the centrality values for each molecule were calculated by CentiScaPe. The correlations between these values were calculated, and the correlation matrix is presented in Table S1 in the Supporting Information. It is clear that two groups can be distinguished by the two chosen measures: (1) Betweenness centrality and stress highly correlate only with each other, and there are significantly lower or no correlations between these and any other measures. (2) The other seven available measures, on the other hand, show the same high in-group correlation. Because of this, we focused on the one−one method from each group: the betweenness centrality from the first and the centroid value from the other group.

To understand the meaning of high and low centrality values, the organic layout is again very helpful. In Figure 3, the five−five nodes that ranked the highest by betweenness centrality and centroid values are designated BC and CC, respectively. The number is the ranking by the given method; for example, BC 1 refers to the node ranked highest by betweenness centrality.

Looking at the position of nodes, which are ranked at the top by the betweenness centrality method, might give a hint of a possible explanation. Most of these nodes are at the intersection of some large clusters (as seen very clearly for BC 1−3). This means that all of the shortest paths from each molecule in one cluster to each molecule in the other cluster go through these nodes and, thus, are ranked high by the betweenness centrality method, which uses the shortest paths to describe centrality. These molecules can be considered bridge structures between two (or more) sets of homogeneous

compounds. These bridge structures can be useful in selecting representative subsets from a library. Also, these molecules may serve as new lead compounds, because of the lack of other similar molecules between the large homogeneous sets.

On the other hand, the centroid measure gives higher rankings to the most interconnected nodes. These nodes can be viewed as central molecules of homogeneous subsets. Obviously, the molecules of a larger subset have better ranking by centroid centrality than those of a smaller subset. This means that when the task is the selection of a subset of a compound library, the centroid value is probably used best in combination with a clustering method.

3.2.3. Dependence of Centralities on the Threshold. The choice of threshold value, as we have seen, fundamentally determines the topology of a threshold network. The question is how much it influences the centrality values for each node. For the present case, two networks of CRF antagonists were compared. One was that discussed above (threshold = 0.50, termed **CRF50** in the following), and the other was **CRF49**, in which the threshold value was 0.49. The two were very similar: 396 vs 395 nodes and 3467 vs 2865 edges for **CRF50** and **CRF49**, respectively. Both networks can be found in the Cytoscape.cys file in the Supporting Information (pdf format is also included). The centralities for these two networks were calculated, and the findings are summarized below.

The calculated values for the 395 common nodes of the two networks are included in the same Cytoscape file as the node attributes (Supporting Information). Also in the Supporting Information (Table S1) is the correlation matrix between the centrality values for the nodes in **CRF49** and **CRF50**, due to its rather large size (18 × 18). The centroid values of the nodes in **CRF49** correlate relatively well with the centroid values of the same nodes in **CRF50**, with a coefficient of 0.92. However, the betweenness centrality values change significantly when the threshold is altered. The reason for this phenomenon is that the betweenness similarity, as we have seen, is the strongest for intermediate nodes between clusters. The consequence, however, is that if many or all connections between two clusters disappear as the threshold is increased, these nodes suffer the largest decrease in centrality values. An analogy for increasing the threshold would be the stretching of an irregular web. The biggest stress will be on the thinnest parts, and breakage will occur there first. After that, the stress quickly drops at these points and grows in other parts of the web. The betweenness centrality is analogous to this stress in a web (as is the other highly correlated and aptly named centrality measure, stress centrality).

The ranking of a node is just as important, if not more so, than its centrality value. The correlations of rankings among the centralities of the two networks are in Table S2 (Supporting Information). In the case of the betweenness centrality, the correlation of node rankings is about the same as the correlation of the betweenness centrality values. However, this is very different in the case of centroid values. As we have seen, there was a 0.92 correlation for these centroid values. However, the node rankings change more. The correlation coefficient between these rankings in **CRF49** and **CRF50** is only 0.52. There are many nodes with very similar centroid values in the network, and a small change in its topology can significantly alter the rankings of the nodes, even if the values change slightly. This may have an important consequence: if we use the centroid values to select molecules from clusters, it is better to choose a group of molecules instead of just the first few that are highly ranked.

Table 1. Change in the Centroid-Ranking of Nodes When the Threshold Is Increased to 0.50 from 0.49 in the Threshold Network of CRF Antagonists[a]

	CRF50_rank		
CRF49_rank	high	middle	low
high	110	24	0
middle	24	91	15
low	0	15	116

[a] The ranks were categorized into three tiers. The high, middle, low values mean the number of nodes belonging to the first, second, and third tiers, respectively.

Table 2. Change in the Betweenness Centrality Ranking of Nodes When the Threshold Is Increased to 0.50 from 0.49 in the Threshold Network of CRF Antagonists[a]

	CRF50_rank		
CRF49_rank	high	middle	low
high	87	42	5
middle	23	51	56
low	24	37	70

[a] The ranks were categorized into three tiers. The high, middle, low values mean the number of nodes belonging to the first, second, and third tiers, respectively.

To further study the tendency of change for centroid rankings, the nodes were divided approximately into thirds: upper-, middle-, and bottom-ranked nodes. Table 1 shows the migration of nodes among the three categories. Almost all (110 of 134) of the nodes that were in the upper (top third) category in **CRF49** remain in the same category in **CRF50**. It is also true for the bottom third. The biggest change happens in the middle, but still 91 of the 130 nodes do not move to another group. We can conclude that even if the individual rankings of the centroid values of many nodes change (slightly) with the topology of the network, we can clearly separate most of the central nodes from the peripheral ones.

The same tendency can be observed for the betweenness centrality values (Table 2), although here, the migration is slightly more definite. There are nodes with a drastic decrease in their centrality value. This is caused by the separations of large clusters: when these are connected through a few nodes, the nodes have high betweenness values, which are lost after the break.

The data set in the previous example included ligands for a single target protein. It is also interesting to examine the method for more diverse data sets. Two further examples are presented in the following chapters. The first is a slightly more diverse library composed of antagonists of adenosine A1 and A2 receptors. The second example consists of a random data set of 5000 molecules from the MDDR.

3.3. Networks of a Multicategory Data Set: Adenosine A1 and A2 Antagonists. The MDDR database that we used contains 479 adenosine antagonists, divided into three categories: 192 and 224 A1 and A2 antagonists, respectively, and 63 dual antagonists. The first question addressed how the members of the three classes were distributed in the

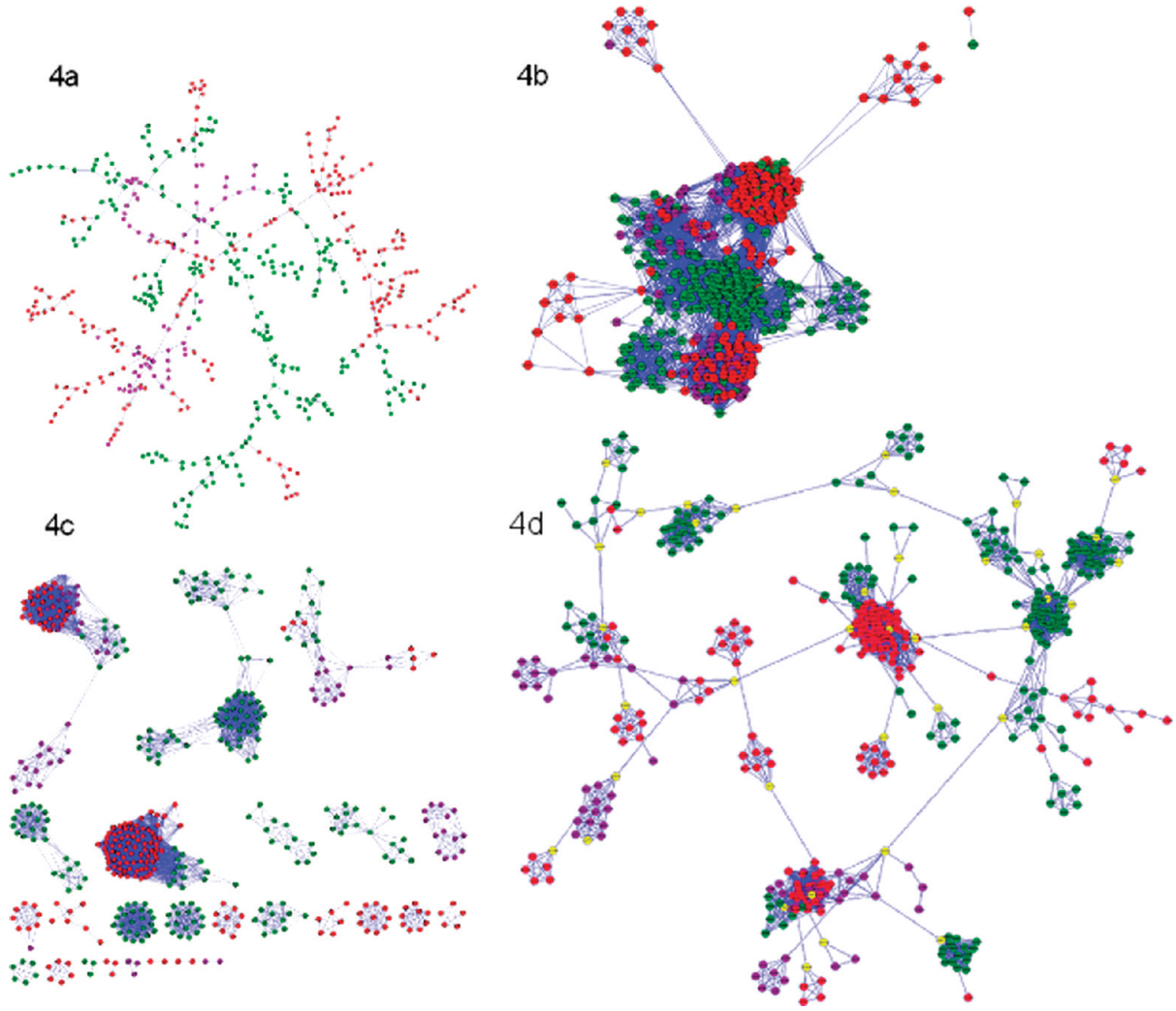


Figure 4. Molecular similarity networks of adenosine A1 and A2 inhibitors. (a) Minimum spanning tree. (b and c) Threshold networks with thresholds values of Tanimoto similarity of 0.36 and 0.48, respectively. (d) Hybrid network of part a and a threshold network of 0.58 Tanimoto similarity. Meaning of colors: (red) A1; (green) A2; (purple) dual antagonists; (yellow) cluster centroids.

structure similarity network(s). To answer this question, both the minimum spanning tree and threshold networks with varying similarity thresholds were created.

It should be noted that, for this kind of database, the possibility for missing information cannot be overlooked. It is very likely that some (or many) of the molecules were not tested against both protein targets. It is an interesting question whether, by looking at the topology of the network, we could point to some nodes that because of their locations in the network could be suspected to be active for the other protein as well.

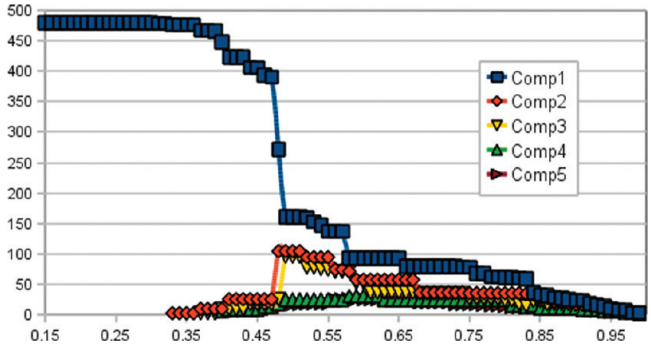
The analyses were performed in a manner similar to that for the CRF inhibitors. CDK fingerprints and the Tanimoto similarity matrix of the compounds were calculated. The networks were generated from the similarity matrix. The networks in Cytoscape format can be found in the Supporting Information.

3.3.1. Minimum Spanning Tree of Adenosine Antagonists. The minimum spanning tree (**mst 3**) is shown in Figure 4a. The three activity categories are colored differently. Traditionally, the minimum spanning tree method has been used for clustering. For this data set, it works quite well. Most of the data points are located in homogeneous regions. This network view is also very useful to identify suspicious molecules by the points found inside a large group of a different color or at the intersection of two groups. These could be dual antagonists, or unexpected "cliffs" in the chemical space of the antagonists (if we can exclude the technical errors in the database or in the algorithm of fingerprint calculation).

3.3.2. Threshold Networks of Adenosine Antagonists. As the mst cannot be used to find central molecules, various threshold networks were also created. At first, it was necessary to select a threshold value. In order to do this, the sizes of the network components as a function of the threshold value were determined in the same way as that for the CRF data set (Chart 2).

Although similar, the tendency of disintegration is not the same as that of the CRF data set. At a threshold value of 0.48, a large group (104 nodes) consisting of mainly A2 antagonists separates from the main component (Figure 4c). This behavior points out that this data set consists of a number of homogeneous subsets that are relatively different from each other. The clusters in these threshold networks contain mostly compounds with the same activity. Thus, by clustering these networks, we could obtain some very important information about the structure of the molecule library. It is also potentially useful for virtual screening.

Chart 2. Sizes of the Five Largest Components in the Threshold Networks of Adenosine Inhibitors As a Function of Threshold[a]



[a] *x*-axis: similarity threshold. *y*-axis: size of network components.

Table 3. Average Rankings of Adenosine A1, A2, and Dual (A12) Antagonists in the **A36** Network, by Three Centrality Methods

	average centrality ranking		
target	betweenness	closeness	centroid
A1	253	226	236
A12	212	313	295
A2	235	229	226

Some important observations can also be made by calculating the centrality values for the network nodes. When calculating centralities, in order to avoid potential false results, a network with only one component is necessary. By looking at Chart 2, a threshold of 0.36 seems to be a good compromise between the need to include as many nodes as possible and avoiding unnecessary complexity of the network. Using this threshold, 477 nodes form one network component (at 0.37, only 467 nodes belong to the largest component; Figure 4b). Thus, the network with a 0.36 Tanimoto similarity threshold (**A36**) was used to calculate centralities.

The correlations among centralities are similar to those observed for the CRF data set (correlation values in the cys file, Supporting Information). Because of this, the same two types of centralities were analyzed as before: the centroid and the betweenness similarity.

The molecules can be ranked by these similarity values, as seen in the previous example (CRF). The average rankings for the compounds grouped by the target classes are summarized in Table 3. The most interesting result is that the dual antagonists (A12) have the largest average betweenness rankings. This makes sense, as the compounds that are active against both proteins are expected to somehow have an average molecular structure of some of the two types of single antagonists. Thus, in a structural similarity network, these compounds are usually located between some groups of others, resulting in a relatively large betweenness centrality value. Hence, calculating betweenness centrality values can be useful to identify compounds with multiple activities.

The A2 antagonists are ranked slightly higher on average than the A1s by betweenness values, but slightly lower by the centroid similarity.

The centroid centrality is the average distance of a node to all other nodes. A node with many close neighbors can also have a high centroid value, even if it has many distant neighbors.

Table 4. Some of the Conclusions from Reference 32[a]

coefficient	hierarch clust	nonhierarch clust	cmpnd sel
Russel−Rao	−	−	−
Tanimoto	+	++	+
Baroni−Urbani/Buser	+	++	++
Yule	++	−	+

[a] The table shows the wellness of four correlation methods for three tasks: (−) not suitable; (+) good; (++) best.

In Figure 3, the A1 nodes (red) are concentrated mostly in two large and compact clusters, with many strong (low dissimilarity) interconnections among them. This explains the high average centroid values. In contrast, the A2 ligands are more diverse; the green nodes are distributed more widely in the network graph, but still have many connections and neighbors. The dual antagonists show the same tendency as the A2 group, but even more pronounced, faring better by the betweenness than by the centroid measures.

3.4. Effect of Correlation Coefficients on the Topology of Molecular Similarity Networks. There are many different methods to calculate the similarity between two molecules. The most common method is to calculate molecular fingerprints and use a correlation method to quantify the difference between them. The most popular correlation method is the Tanimoto coefficient. In a recent article, the suitability of 13 different correlation methods and 4 types of fingerprints for 3 different tasks was compared[31] The tasks were hierarchical and nonhierarchical clustering and dissimilarity-based compound selection. The authors concluded that the type of fingerprint had little influence on the results, but the choice of correlation coefficient had radical effects. Some of the coefficients were simply not suitable for certain tasks, and others were inefficient.

The molecular similarity network-based approach also depends heavily on how the similarity values are calculated. Because of this, we have examined how the selection of correlation methods influences the topology of the network, and how this corresponds to the observations reported previously. We selected four representative correlation methods, which are described in the Methods section. The conclusions from the previous paper regarding the success of the four methods for the three tasks are summarized in Table 4. The Russel−Rao method proved to be useless for these chemoinformatics tasks. The Yule method was the best for hierarchic clustering, but interestingly, it scored very poorly for the nonhierarchic clustering. Both Baroni−Urbani/Buser and Tanimoto coefficients were solid and nearly equal performers, with the Baroni−Urbani being a bit better for compound selection, due to its lower bias to the size of structure.

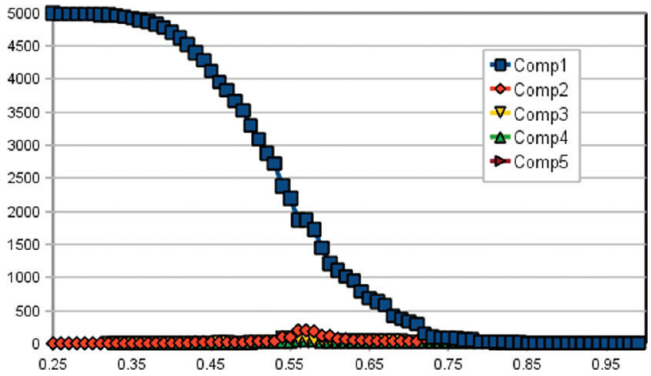
It should be noted, that using different types of descriptors, these coefficients might score differently.[33,34] In the current study however, we applied similar fingerprints as the introduced work; thus, we think their findings are relevant.

The compounds for the current comparison were selected from the MDDR, similar to the aforementioned article. However, only 5000 random molecules were used, in contrast to the 20 000 used previously. The reason for this was to decrease the computational requirements for network analyses. Some of the procedures discussed later are rather large consumers of CPU time and/or memory. A state-of-the-art

Table 5. Minimum and Maximum Similarity Values in the Similarity Matrices of the Random 5000 MDDR Compounds, Calculated by the Four Correlation Methods

	similarity value	
correlation method	min	max
Russel−Rao	0.003	0.33
Tanimoto	0.012	1
Baroni−Urbani/Buser	0.125	1
Yule	0.66	2

Chart 3. Sizes of the Five Largest Components in the Threshold Networks of the Randomly Selected 5000 Molecules of the MDDR, As a Function of Threshold



(single core) CPU with 1−2 GB of RAM could easily handle the data set of 5000, with none of the calculations presented taking more than a couple of hours.

Fingerprints were generated and similarity matrices calculated using the four correlation methods. Note that in order to compare the different methods, the similarity values had to be rescaled to between 0 and 1 using eq 1.

$$S_{sc} = T_{low} + (T_{up} - T_{low})(S - S_{min})/(S_{max} - S_{min}) \quad (1)$$

where S_{sc} is the scaled value of the original similarity (S), T_{low} and T_{up} are the theoretical minimum and maximum values for the given method (usually 0 and 1, except for the Yule method, in which they are 0 and 2), and S_{min} and S_{max} are the minimum and maximum values in the given similarity matrix. These values are presented in Table 5. The scaling influenced neither the distribution of similarity values, nor the topologies of networks but made it possible to draw all of the results in the same chart.

Threshold networks were created from each matrix, with the threshold values being varied between two marginal threshold values (e.g., S_{Sc} is between 0.25 and 0.99 with a step of 0.01).

Displayed in Chart 3 are the sizes of the five largest network components obtained using Tanimoto similarity. The most interesting observation is that even in the case of a random data set of 5000 molecules, there is only one significant component and most of the nodes are connected to it, and when S_{Sc} reaches about 0.35−0.4, this sole network component begins to disintegrate. Similar to the case of 400 CRF antagonists, the nodes leave one by one or in small groups. At 0.57, there is a small bump in the graph, when a somewhat larger group leaves the main component, but its size is still minor compared to the largest one. The same tendency was observed in the cases of the other correlation coefficients; thus, only the largest components had to be analyzed, making the comparison much easier (Figure 5).

The curves shown in Figure 5a are very different for each of the correlation methods. In the case of Russel−Rao, the network disintegrates very early, the molecules are very dissimilar to each other, and the distribution curve of similarity values is shifted very much to the left. This demonstrates its inadequacy for chemoinformatics tasks. On the other hand, using the Yule method, the compounds appear to be very similar (the distribution of similarities shifted to the right) and there is only one network component, which practically does not disintegrate. The network has a very homogeneous topology; most of the nodes are connected to many of the others. This homogeneity is probably favorable for hierarchic clustering. However, it makes the formation of islands, i.e., subgraphs consisting of strongly interconnected nodes that are weakly connected to each other, difficult. Hence, it is not well-suited for nonhierarchic clustering. The Tanimoto- and Baroni−Urbani-based networks lie between the two extremes. The curve of the latter is closer to that of the Yule graph, whereas that of the former is closer to the Russel−Rao graph. Both offer a compromise between a dense network and a sparse one.

The previous statistics alone could not describe the topology of the networks. For this reason, additional analyses were conducted. One such statistic was the average number of edges per node.

The curves for the four cases are shown in Figure 5b. These curves correspond well to the previous ones. For example, the disintegration of the network starts at S_{Sc} values of about 0.4 and 0.7 in the cases of Tanimoto and Baroni−Urbani correlations, respectively. The edge/node graphs exhibit similar tendencies, as the curves become less steep above the aforementioned threshold values. The same observation is made when we consider the average number of triangles of which an edge belongs (Figure 5c). This curve is even steeper than the previous ones, highlighting how complex a molecular similarity network can become. This is especially true when we use a correlation method such as Baroni−Urbani or Yule. The latter is probably not suitable for this kind of task, given the cost of computation that such a large scale of interconnection would require. The advantage of using the Tanimoto correlation is that all (or almost all) compounds can be included in the network, without the risk of it becoming highly complex.

A library might contain a large number of highly similar compounds. These would form highly interconnected regions in a threshold network. To reduce the complexity, these regions could be identified and the excess nodes removed. One possible solution could be to find cliques in the network. A clique is a group of nodes in which each member is connected to all of the others. Figure 5d illustrates a four-membered clique (top right). The numbers of all such groups, as a function of the similarity threshold, are also shown. These are very steep curves; for example, the Baroni−Urbani curve has a cubic trend. The huge number of cliques probably makes the analysis impractical (at least for large data sets), unless the selected threshold is between 0.77−0.80. However, in these cases, about one-third to one-half of the molecules are already eliminated (Figure 5a). The problem is that the most dissimilar structures are most likely to leave first, thus the diversity of the network decreases, which is not usually

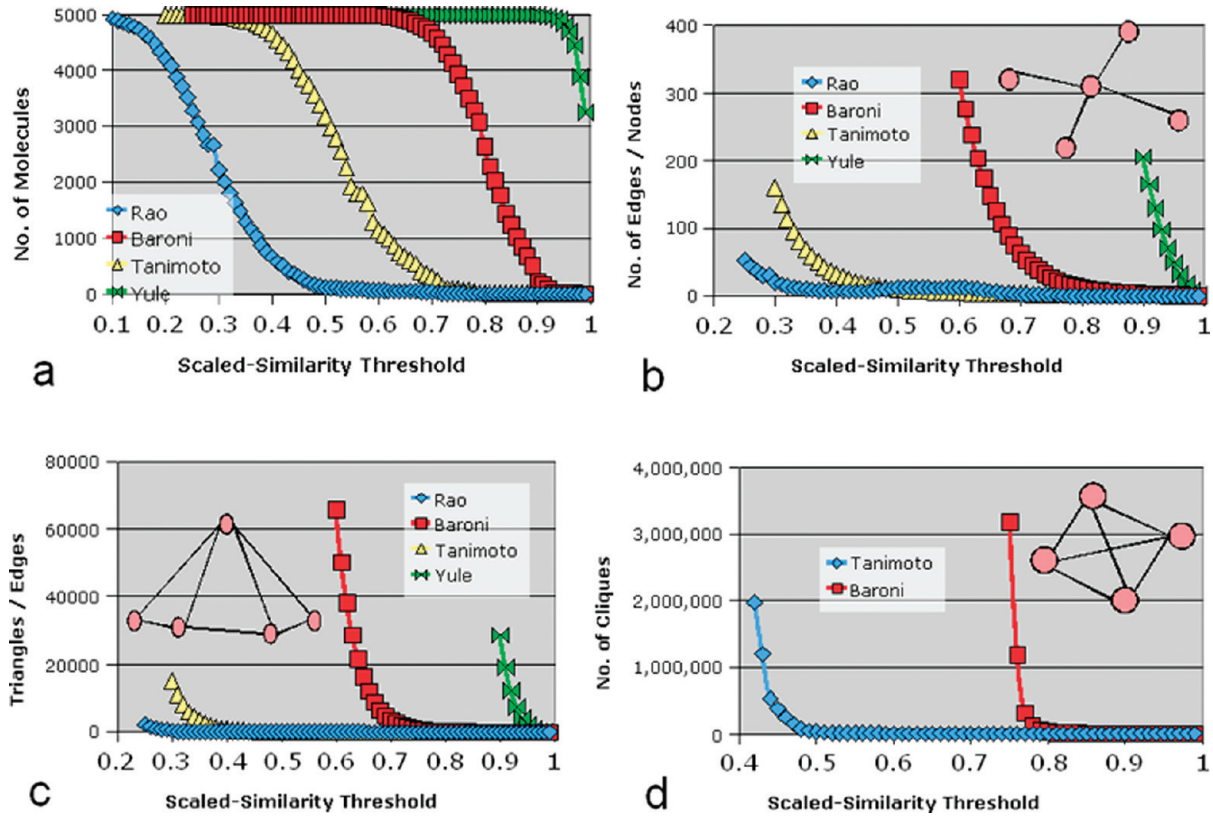


Figure 5. Alteration of network topology parameters, as the function of threshold, in the largest one−one components in the threshold networks of the random 5000 MDDR molecules. The colors indicate the correlation method used for calculating similarity (see the legends). The measured parameters were the following: (a) number of nodes; (b) ratio of edges and nodes; (c) number of triangles; (d) number of full cliques.

what we would like to achieve. However, by identifying cliques and removing some members of each clique (such as removing the ones with low centrality values), we can reduce the most homogeneous parts of a library. For the analyses of cliques, the Tanimoto similarity seems to be optimal.

3.5. Potential Application Fields of Complex Network Analysis Methods. A great advantage of the similarity network-based method is flexibility, as it can be applied to various purposes. But this fact also makes it more difficult to exhaustively compare the method to all the other techniques. In this chapter, a few examples are discussed to show that promising results can be obtained for important chemoinformatics tasks by the use of the approach.

3.5.1. Theoretical Applications. The applicability of the concept for a large variety of tasks has not been thoroughly studied. Thus, we can only list some promising application domains.

1. Visualization. Probably the most evident advantage of the use of networks or trees to analyze molecular libraries is their rich visualization capabilities. Various examples were already shown in the previous chapters to illustrate this fact.

2. Clustering and Virtual Screening. Among the most important tasks in chemoinformatics, it is of great importance to figure out whether and how the approach the can be effectively used for these tasks. Two successful applications are going to be discussed, later in this chapter.

3. Library Analyses. It is one of the main topics of this article. We showed previously that central and transitionary elements of data sets can be conveniently detected by centrality methods. Analyzing the diversity of a library also seems to be a well-suited area for the approach, but this has neither been thoroughly analyzed yet, as far as we know. One way to perform diversity analysis is to calculate the connectedness of nodes. This is certainly an area worth further investigations.

4. Selection of Molceule Series. As it was mentioned, the method of minimum spanning tree has been already applied to the selection of QSAR series. Another frequent and important task is choosing a representative subset from a library. This is also an area where a similarity network could be effectively used. All the tools exist (and are routinely used in various scientific domains) to perform the job: clustering the network, detecting subnetworks and cliques, identifying central and peripheral nodes, etc. It is also not to be despised that a meaningful graphical representation of a compound library can greatly help in hand-picking entities.

5. Integrating Knowledge. Human knowledge is frequently organized into trees (e.g., mind-maps) or networks. The construction, interpretation, and use of metabolic networks are fundamental parts of systems biology. On the other hand, the similarity networks that are the subject of the article are generated in an automatic way. How to integrate these different approaches is a very complicated and diverse problem. It largely depends on the purpose of the research. This decides the kind of information that may be added to enhance an automatically assembled similarity network. A generally useful example may be the addition of molecular fragments, such as natural compoundlike fragments (SCONP[35]), in order to link the nodes in an alternative way, besides fingerprint similarity. This is also a very important and promising area for further research.

3.5.2. Network Effects to Amplify Structural Differences. Layouts are applied to networks in order to create a graphical representation that helps humans to easily understand the structure of a network. Some of the layout methods result in the network vertices aggregating into clusters, as a joint-effect of neighboring and strongly interconnected nodes. One such network layout technique is the force-directed layout in Cytoscape.

The concept of force-directed layout is similar to that of a molecular mechanics force field. The nodes of a network behave like charged particles, which repulse each other. The edges, on the other hand, are analogous to springs with an optional force constant (weight of an edge) that attract the nodes. The forces in the system are minimized to obtain a, hopefully, aesthetically pleasing graph. This also means that vertices that are heavily interconnected with strong bonds aggregate together and unconnected groups of nodes are pushed further away. Thus, the differences among subgroups are amplified.

The question is if this phenomenon can be applied to enhance the discrimination power of similarity methods.

The dissimilarity values between molecules correspond to the lengths of edges in the similarity network. When a layout is applied, these length values change. The question is whether desirable results can be obtained by replacing the original dissimilarity values with the altered lengths. Such a desirable effect would be if molecules with various structure-dependent properties (like pharmacological effect) could be better separated.

To investigate the usability of concept, the previously examined threshold networks of adenosine antagonists were used (Figure 4b).

First, the force-directed layout of Cytoscape was applied to the 36% threshold network. The edges were not weighted in order to avoid any influence of the original fingerprint-dissimilarities, other than defining the connectivity of nodes. The lengths of all of the edges, after the layout was applied, were determined by a Cytoscape plug-in. There were no correlation (correlation value was 0.36) among the original dissimilarity values and these lengths; thus, the application of layout had significantly altered the relative similarities of molecules to each other.

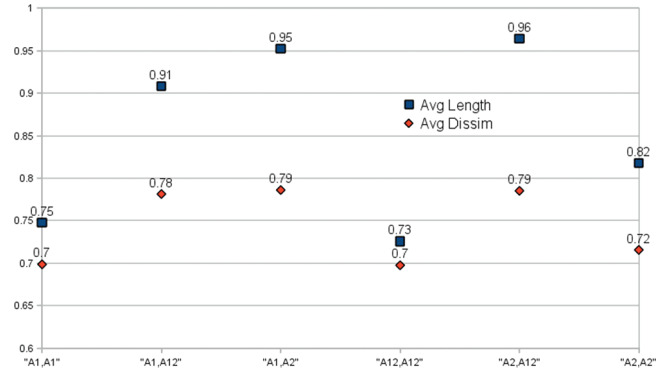
The original dissimilarity matrix was modified by replacing the original values by the normalized edge lengths. The normalization was done by scaling between 0.0 and 1.0. The dissimilarity values between all of those molecule pairs that were not connected in the network were uniformly 1.0.

The averages of the pairwise similarities of molecules, grouped by all the target combinations, were calculated for both the original and modified similarity matrices.

The results are in Chart 4. The average pairwise dissimilarities from the original dissimilarity matrix are shown by the rhombus marks. The corresponding values for the modified matrix are squares.

The most significant consequence of the modification is the increase of gaps between average pairwise intra- and intergroup dissimilarities. For example the average dissimilarities between all of the A1−A1 and A1−A2 pairs are 0.70 and 0.79 in the original matrix, respectively. On the other hand, the same values for the modified matrix are 0.75 and 0.95, respectively. As the fundamental aim of many tasks, such as virtual screening or clustering, is to discriminate compounds belonging to the same group from outliers, the increase of gap between in-group and out-group (dis)-similarities is a desirable effect.

Chart 4. Average Pairwise Similarities between Molecule Pairs of Each Target−Target Combination[a]



[a] The symbols are as follows: (♦) in the original dissimilarity matrix and (■) in the modified adjacency matrix of the network (average edge length) in Figure 4d.

The average dissimilarity values in the modified matrix are higher for each target pair. The reason is that the dissimilarity of nonbonded compound pairs has been set to 1.0, and this has increased the averages, too. Because there are few edges between compounds belonging to different targets, the average dissimilarities for such pairs have grown considerably.

This means that the amplification of distances between target groups has been caused by the aggregative effect of the application of a threshold value and the force-directed layout. It also means that when a larger threshold is applied, for example 46% instead of 36%, the averages shift more toward 1.0 and the differences between groups decrease.

We can conclude that, using the aforementioned network-based technique, it is possible to increase the contrast among structure subgroups in a library, which can be a desirable effect for a number of purposes.

Another advantage of using the similarity network and a force-based layout method for the visualization of chemical library is its relatively low sensitivity to outliers. The topology of the network is mainly decided by the most dense areas. The outliers may be "pushed" outside, but they hardly influence the position of populous areas. This is in contrast with e.g. the self-organizing map (SOM) approaches, where the data points are placed to a fixed matrix, and the inclusion of distant outliers squeezes the other points into a smaller area. Thus, using SOM techniques, the outliers should be carefully removed. In the network approach, these outliers could be included through the use of minimum spanning tree (e.g., as it is shown in the next example), so they can be studied together with the major population.

Another differentiating feature of the use of a force-based layout compared to other dimension reduction methods (e.g., principal components or multidimensional scaling) is that it alters the relative position of data points (similarly to SOM). The basic topic of this chapter has been to show that it can be advantageous if we can include the joint (network) effects of data points into our analyses.

3.5.3. Clustering and Library Selection. Clustering chemical libraries is a very common and fundamental task. A large variety of methods have been used to perform it. Empirical

experiments suggest that the hierarchical methods generally produce "better" clusters in terms of their ability to bring compounds with similar properties into the same cluster.[36] Among the hierarchal methods, Ward's method has become one of the most widely used.

Clustering is also an important topic in network analysis studies. But, the analysis of complex networks is a recently emerging field, and it is less well-studied, yet. For example, there is no exact definition of what a cluster in a network is. Thus, clustering methods have been developed by the requirements of the specific domains. One such domain is proteomics, where the detection of densely connected regions in large protein−protein interaction networks is of great importance, because those regions may represent molecular complexes.

Clustering is a very commonly used tool for selecting representative subsets from chemical libraries. Thus, in the followings, we are presenting an integrated approach for library analyses and subset selection, using those complex network analyses concepts, which were introduced previously. The steps of the approach are described below.

1. Creation of a hybrid of a minimum spanning tree and a threshold network, to combine the advantages of both of them. The biggest problem with threshold networks is how to select a threshold, which gives a good balance between complexity and connectedness. In a minimum spanning tree, all the nodes are connected, but the number of edges is too low (loss of information) and centralities cannot be calculated, either. But simply adding those edges, of which similarity values are above a defined similarity threshold, to a spanning tree, we could get a better balanced complex network.

2. Calculation of the centrality values of all nodes.

3. Determination of clusters by the MCODE method.

4. Selection of a subset from each clusters, based on their centrality values.

5. Optional steps, like recognizing cliques in large clusters and selecting entities from those cliques.

As an example, a hybrid network of the adenosine inhibitors, is shown in Figure 4d.

The similarity threshold for edges is 58%. The reason for the relatively high threshold is to obtain better separation of substructures (see also Chart 2). Because all of the nodes are already connected by the mst, the disintegration of network into smaller pieces cannot happen.

Looking at the graph in Figure 4d, it is clear that the distribution of structures is not uniform. There are some highly interconnected regions. These regions can be clustered and ranked by the Cytoscape plugin, MCODE (see the methods chapter for description). This method was developed to identify highly interconnected nodes in proteomics networks, because it was assumed that these nodes correspond to protein complexes. In our case, we can expect that the strongly linked vertices are structurally homogeneous and probably belong to the same group regarding pharmacological activity. The graphical view (colors) in the figure confirms this assumption, but it is necessary to prove it numerically.

Thus, the quality of clustering by using this network and MCODE was compared to what can be obtained by the Ward method. To carry out the reference clustering, the ward module of Jchem was used. The number of clusters cannot be directly defined by the MCODE method, as it is a function of chosen cutoff parameters of the method. Because of this, the default parameters of the plugin were used, which resulted in 37 clusters. One more cluster was added to contain the unclustered (singleton) nodes, which were 66 altogether. For the best comparison, the number of clusters in Ward clustering was also 38. The Kelley method[37] could also be used to predict the optimal number of clusters in Ward clustering. However, the prediction gave 17 as the optimal number, which was significantly lower.

The compositions of each cluster obtained by both methods are summarized in Chart 5. There are also some screenshots of MCODE clusters in the Supporting Information.

While the sizes of MCODE clusters are uneven and correspond to those of the highly interconnected areas in the network, the sizes of Ward clusters are more evenly distributed. There are no singletons in the case of the Ward method.

The two approaches gave comparable results, regarding their ability to group compounds with the same activity together. There are 8 mixed clusters with 120 compounds among the Ward clusters. The corresponding values for the MCODE method are 5 and 133; however, about one-half of these nodes (66) are singletons. Also, when the optimal number of clusters (17), as predicted by the Kelley method was used, 257 compounds were in 8 mixed clusters, which was a significantly worse performance for separation. All the dual antagonists were placed into mixed clusters. The reason for this result is that the Kelley method can be used to obtain clusters with the most uniform number of data points, as possible. But if the data set is not homogeneous as in this case, this is not the most optimal solution.

Thus, apart from the problem of singletons, the network-based clustering approach seems to perform well. But, because the two approaches are rather complementary, we can expect that the combined use of the two techniques could significantly improve the efficiency of clustering (for example by reclustering the nonclustered nodes with the Ward method).

The combination of the aforementioned techniques—the creation of a mixed minimum spanning tree—threshold network, determination of clusters, and calculation of centralities can be an effective way to select a subset from a library. The nodes with the highest centroid values in each MCODE clusters are painted yellow in Figure 4d.

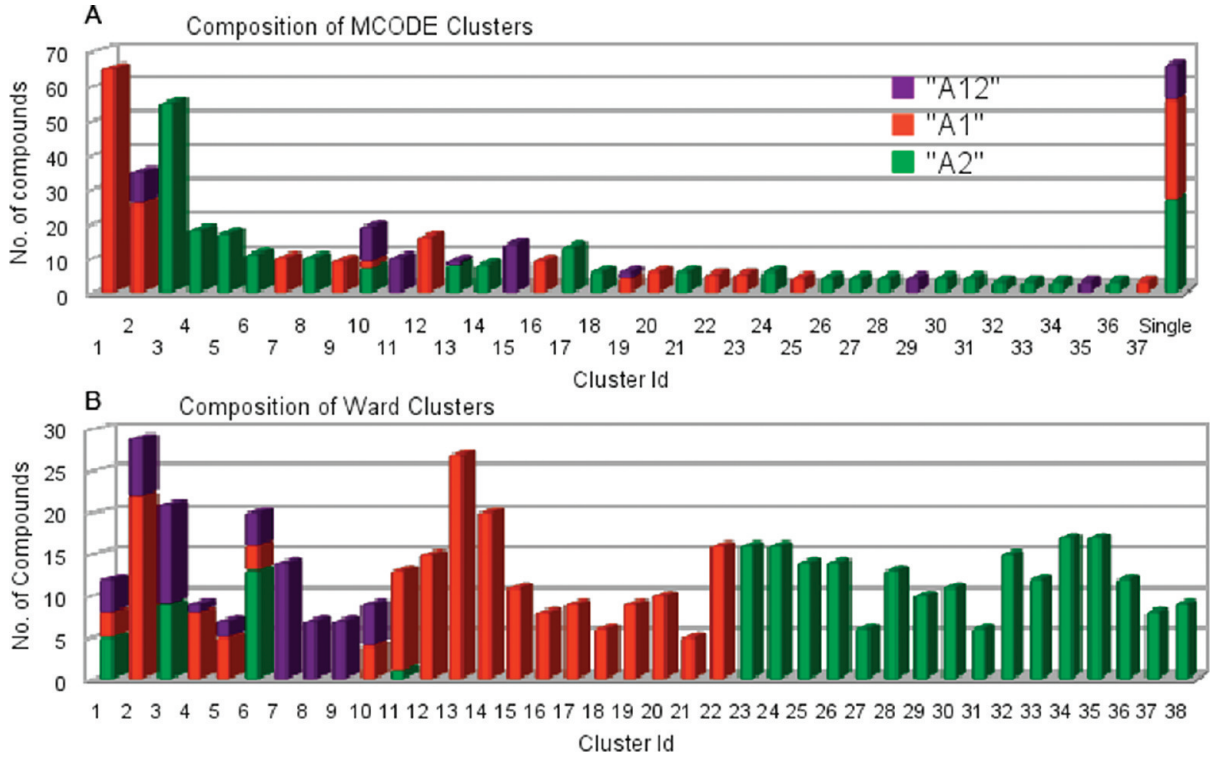
The 38 (8% of all) compounds are already good starting points for selection. Because the centralities are calculated from the full network, these nodes not necessarily identical to cluster centers, which are local centers. In the current example, only in 29 clusters out of 38 are these nodes identical.

If more samples are needed then, for example, dissimilarity based selection starting from these entities could be a good strategy.

4. CONCLUSION

The identification of certain elements of chemical libraries, such as central molecules in the chemical space or transitional structures, is an important task in chemoinformatics. Here, we have shown that organizing the molecules into networks

Chart 5. Composition of Ward (a) and MCODE (b) Clusters[a]



[a] *x*-axis: cluster id. *y*-axis: number of compounds.

defined by structural similarity can be a convenient method to analyze libraries.

Applying meaningful graph layouts of the network is a very powerful way to visualize the relationships among compounds and the hierarchical structure of a library. A Cytoscape plug-in was developed that integrates the MarvinView molecular graphics program into this network visualizer and analysis system, offering a convenient way to view the chemical structure of a network node.

Various chemoinformatics tasks can be performed through the analyses of the molecular networks. These include clustering, selection of QSAR subseries, identification of central compounds in a library, and potentially other tasks, such as virtual screening.

The weight of edges connecting any pair of nodes in the network can be defined by both similarity and dissimilarity. Minimum spanning trees have long been used for clustering molecular libraries. In previous studies, the edge weight was the Tanimoto dissimilarity between the two compounds. We have shown that the similarity coefficient can also be used to define the weight. In this case, the graph is the inverse, with the most dissimilar molecules becoming the most connected, which can help to identify outliers in a library.

Threshold networks, which generate graphs containing only edges with a weight larger than a predefined threshold, have not yet been applied widely for chemoinformatics studies. Here, we have shown that threshold networks, in which the weights are Tanimoto similarities of molecular fingerprints, are very suitable for cluster analysis (e.g., by applying a force-directed layout). Central molecules in libraries can be identified by determining the central nodes in the networks. It was found that two different centrality methods, centroid and betweenness centrality, give very different results. The former is useful to find cluster centers and highly connected molecules. The latter centrality can be used to identify the average structures between two or more homogeneous subsets. In the case in which the subsets are ligands of different proteins, the average structures have a good chance of having multiple activities, as exemplified by the data sets of adenosine A1 and A2 antagonists.

The correlation methods used to define the similarity between two fingerprints were found to have a significant impact on the topology of structure similarity networks. A comparison of different correlation methods revealed that the complexity of the networks, created from 5000 random druglike molecules, increased sharply in the order of Russel−Rao, Tanimoto, Baroni−Urbani/Buser, and Yule correlations. The Tanimoto dissimilarity had a good balance between connectedness and complexity of the networks, which may be in line with the success of the method for various chemoinformatics tasks.

It was shown, through the example of adenosine antagonists, that network-based clustering can be effectively used for clustering compound libraries with a performance comparable or better than Ward clustering (depends on the number of clusters). Furthermore, the application of a force-based network layout to a similarity network can amplify the differences among groups of homogeneous structures in the library.

Among other areas, the rapid development in the field of network analysis has made it possible that all of the tools used in this report (except the data set) exist as open-source software. A significant advantage of using network analysis methods for chemoinformatics projects is that researchers can benefit immediately from developments in this growing field.

Supporting Information Available: Molecular networks studied in Cytoscape format. Development snapshot of the Jchem plug-in for Cytoscape. Installation information provided in readme.txt.

This material is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Newman, M.; Barabasi, A. L.; Watts, D. J. *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*; Princeton University Press: Princeton, NJ, 2006.

(2) Bornholdt, S.; Schuster, H. G. *Handbook of Graphs and Networks: From the Genome to the Internet*; Wiley-VCH: Weinheim, 2003.

(3) Ivanciuc, O.; Downs G. M.; Bangov, I. P.; Weininger, D. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Chapters II. 4 and 5.

(4) Schuster, S.; Fell, D. A.; Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **2000**, *18*, 326–332.

(5) Palsson, B. O. *Systems Biology: Properties of Reconstructed Networks*; Cambridge University Press: New York, NY, 2006.

(6) Hopkins, A. L. Network pharmacology. *Nat. Biotechnol.* **2007**, *25*, 1110–1111.

(7) Schrattenholz, A.; Soskić, V. What does systems biology mean for drug development. *Curr. Med. Chem.* **2008**, *15*, 1520–8.

(8) Long, T. A.; Brady, S. M.; Benfey, P. N. Systems approaches to identifying gene regulatory networks in plants. *Annu. Rev. Cell. Dev. Biol.* **2008**, 81–103.

(9) Babu, M. M. Computational approaches to study transcriptional regulation. *Biochem. Soc. Trans.* **2008**, *36* (Pt 4), 758–65.

(10) González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics* **2008**, *8*, 750–78.

(11) Ritter, G. L.; Isenhour, T. L. Minimal spanning tree clustering of gas chromatographic liquid phases. *Comput. Chem.* **1977**, *1*, 145–154.

(12) Miyashita, Y.; Takahashi, Y.; Yotsui, Y.; Abe, H.; Sasaki, S. Application of pattern recognition to structure-activity problems Use of minimal spanning tree. *Anal. Chem. Acta.* **1981**, *133*, 614–624.

(13) Kenndler, E.; Reich, G. Characterization and selection of electrolyte systems for isotachophoresis of anions by cluster analysis. *Anal. Chem.* **1988**, *60*, 120–124.

(14) Mount, J.; Ruppert, J.; Welsh, W.; Jain, A. N. IcePick: A flexible surface based system for molecular diversity. *J. Med. Chem.* **1999**, *42*, 60–66.

(15) Dore, J. C.; Gilbert, J.; Bignon, E.; Crastes de Paulet, A.; Ojasoo, T.; Pons, M.; Raynaud, J. P.; Miquel, J. F. Multivariate analysis by the minimum spanning tree method of the structural determinants of diphenylethylenes and triphenylacrylonitriles implicated in estrogen receptor binding, protein kinase C activity, and MCF7 cell proliferation. *J. Med. Chem.* **1992**, *35*, 573–583.

(16) Schlich, P.; Guichard, E. Selection and classification of volatile compounds of apricot using the RV coefficient. *J. Agric. Food Chem.* **1989**, *37*, 142–150.

(17) Shen, Q.; Jiang, J. H.; Jiao, C. X.; Huan, S. Y.; Shen, G. l.; Yu, R. Q. Optimized Partition of Minimum Spanning Tree for Piecewise Modeling by Particle Swarm Algorithm. QSAR Studies of Antagonism of Angiotensin II Antagonists. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 2027–2031.

(18) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.

(19) Hert, J.; Keiser, M. J.; Irwin, J. J.; Oprea, T. I.; Shoichet, B. K. Quantifying the Relationships among Drug Classes. *J. Chem. Inf. Model.* **2008**, *48* (4), 755–765.

(20) Hattori, K.; Wakabayashi, H.; Tamaki, K. Predicting Key Example Compounds in Competitors' Patent Applications Using Structural Information Alone. *J. Chem. Inf. Model* **2008**, *48* (1), 135–142.

(21) Santana, L.; Gonzlez-Daz, H.; Quezada, E.; Uriarte, E.; Yez, M.; Via, D.; Orallo, F. Quantitative Structure-Activity Relationship and Complex Network Approach to Monoamine Oxidase A and B Inhibitors. *J. Med. Chem.* **2008**, *51*, 6740–6751.

(22) *Centiscape*, version 1.0; http://profs.sci.univr.it/~scardoni/centiscape/centiscapepage.php (accessed Sep 8, 2009).

(23) *MDL Drug Data Report*; Elsevier MDL: Hayward, CA, 2004.

(24) *JChem 4.0*; ChemAxon Ltd: Budapest, Hungary.

(25) Guha, R. Chemical Informatics Functionality in R. *J. Stat. Soft.* **2007**, *18*, 5.

(26) Guha, R.; Schurer, S. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 367–384.

(27) Eads, D. hcluster: Hierarchical Clustering for SciPy. http://scipy-cluster.googlecode.com/ (accessed Jun 8, 2009).

(28) *Cytoscape*, version 2.6.3; http://www.cytoscape.org (accessed Jun 8, 2009).

(29) *Sage: Open Source Mathematics Software*, version 3.2; http://www.sagemath.org (accessed Nov 21, 2008).

(30) *Networkx*, version 0.36; http://networkx.lanl.gov/ (accessed Jun 12, 2008).

(31) Bader, G. D.; Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf.* **2003**, *4* (1), 2.

(32) Haranczyk, M.; Holliday, J. Comparison of Similarity Coefficients for Clustering and Compound Selection. *J. Chem. Inf. Model.* **2008**, *48* (3), 498–508.

(33) Willett, P. Similarity-based approaches to virtual screening. *Biochem. Soc. Trans.* **2003**, *31* (3), 603–606.

(34) Marín, R. M.; Aguirre, N. F.; Daza, E. E. Graph Theoretical Similarity Approach To Compare Molecular Electrostatic Potentials. *J. Chem. Inf. Model.* **2008**, *48*, 109–118.

(35) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (48), 17272–17277.

(36) Efficiency and Effectiveness of Clustering Methods. In *An Introduction to Chemoinformatics*, revised ed.; Leach, A. R., Gillet, V. J., Eds.; Springer: Dordrecht, The Netherlands, October 12, 2007; pp 127−128.

(37) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.* **1996**, *9*, 1063–1065.

CI9001102