Small-World Phenomena in Chemical Library Networks: Application to Fragment-Based Drug Discovery

Naoki Tanaka, Kazuki Ohno, Tatsuya Niimi, Ayako Moritomo, Kenichi Mori, and Masaya Orita*

Chemistry Research Laboratories, Drug Discovery Research, Astellas Pharma Inc., 21 Miyukigaoka, Tsukuba-shi, Ibaraki 305-8585, Japan

Received April 3, 2009

A wide variety of networks in various fields have been characterized as small-world networks. In scale-free networks, a representative class of small-world networks, numbers of contacts (degree distributions) of nodes follow power laws. Although several examples of power-law distributions have been found in the field of chemoinformatics, the network structures of chemical libraries have not been analyzed. Here, we show that small-world phenomena are observed not only in existing chemical libraries but also in virtual libraries generated from structurally diverse fragments when represented as networks. On the basis of this observation, we propose that an efficient compound-prioritization method of fragment-based drug discovery (FBDD) would be to select those fragments as a starting point such that the linked compounds become hubs in the library and therefore allow identification of many similar compounds when all-to-all fragment linkings are performed. Moreover, our analyses indicated that the variety of linkers had a marked influence on the network structure and thus on the diversity of the compounds synthesized by linking fragment hits.

1. INTRODUCTION

With the advent of the Internet and other huge networks, many researchers have closely analyzed various network structures. One of the most significant outcomes has been the discovery of scale-free networks with degrees of connectivity of the nodes distributed according to power laws.[1] Here, scale-free networks are a representative class of small-world networks, which have small diameters and high clustering coefficients.[2] Scale-free networks occur in a wide variety of fields, such as computer science, biology, physics, and social science.[3] The term "scale-free" is derived from the fact that the same feature can be observed regardless of scale. A classic example of scale-freeness is the Pareto principle,[4] also known as the 80−20 rule, which states that roughly 80% of the effects come from 20% of the causes.

Most nodes in scale-free networks have few edges, representing interactions between components of the network. However, small numbers of nodes that have a large number of edges also exist. These nodes are called hubs because they are linked to many other nodes. Because of this unbalanced degree distribution, scale-free networks are said to be robust against random attacks, although they can be easily split if the attacker zeroes in on the hubs. Many real-world networks with scale-freeness benefit from this robustness. Networks related to life sciences are typical examples: several metabolic networks have been shown to be scale-free, for instance, and thus to possess robustness.[5] Thanks to the robustness of the network, substances critical to survival can be produced by alternative paths if any of the nonhub metabolic intermediates cannot be produced for any reason.

Several distributions of chemoinformatics data have also been found to follow power laws.[6] The results are insightful and have led to the development of many useful applications, such as lossless compression of chemical fingerprints. To date, however, despite the significant interest of the pharmaceutical industry in the network structures of chemical libraries, studies have not focused on the network structures of chemical libraries. One reason for this is the huge size of the chemical space. It is thought that as many as 10^{60} compounds are synthesizable.[7] The enormity of this number serves to highlight the desperate need for efficient compound exploration methods in synthesis and assay. The systematic network analysis of the structures of compound libraries represented as networks is expected to provide new insights in this field. Given that an understanding of the mechanisms underlying such properties will facilitate the exploration of compounds, this is particularly true for networks having the characteristic properties of scale-freeness and small-worldness. Moreover, network analyses can also be useful in library design because they provide structural insights into the compound diversity of the libraries.

With a view to these factors, we have systematically analyzed the network structures of chemical libraries. Our goals included defining whether chemical library networks show characteristic behaviors, such as scale-freeness and small-worldness, and making use of such properties for drug discovery. In our network representation of chemical libraries, compounds are considered as nodes and are linked by edges if two compounds are similar in terms of fingerprint-based Tanimoto similarity.[8] We selected the ZINC database[9] as our main target library because of its relatively large compound set and drug-likeness.

Given the recent attention to fragment-based drug discovery (FBDD), we have also examined fragment libraries. FBDD emerged in the past decade and has proven to be a novel drug-discovery paradigm,[10−12] as demonstrated by its success in the development of more than 10 clinical candi-

* Corresponding author phone: +81-29-863-6768; fax: +81-29-856-2558; e-mail: masaya.orita@jp.astellas.com.

10.1021/ci900123v
Published on Web 12/04/2009

dates.[13] Also, FBDD is now established as the principal alternative to traditional methods used for lead discovery, such as high-throughput screening (HTS) and virtual screening. Our group has previously developed two original indices that can be utilized in FBDD. Further, we have found that these two indices are based on the "Golden Ratio".[14] The further acceleration of FBDD is dependent on the development and availability of efficient compound-prioritization methods, particularly in fragment-linking strategy.

In this Article, our analyses included not only real fragment libraries from vendors but also structurally diverse fragment libraries constructed by extracting relatively small compounds from the ZINC database. To determine whether there were any network structural differences between the generated virtual library and original diverse fragments, we examined the network structure of the virtual library generated by linking fragments in the extracted diverse fragment libraries. We also analyzed the impact of the variety of linkers on the network structure.

These network analyses of chemical libraries indicated that these libraries possessed a small-world property when represented as networks. From this, we then proposed an efficient compound-prioritization method for fragment linking. Moreover, we found that the network structures of chemical libraries generated by fragment linking were significantly influenced by the variety of linkers, indicating the importance of this factor to molecular diversity when fragment linking was performed.

Here, we first explain our analysis procedures and data sets in detail and show the results of the network analyses of various chemical libraries. We then discuss the small-world behavior of chemical library networks and its application in compound prioritization, and finally we state our conclusions.

2. METHODS

2.1. Network Representation of Chemical Libraries. To represent chemical libraries as networks, we considered compounds as nodes. Edges exist between nodes with similarities above the threshold. This network representation of chemical libraries is defined as the graph

$$G(V, E)$$

where

$$V(G) = \{v | v \in L\}$$
$$E(G) = \{(v_1, v_2) | \mathrm{sim}(v_1, v_2) \geq T\}$$

Here, L denotes a chemical library, $\mathrm{sim}(v_1,v_2)$ denotes a similarity between molecules v_1 and v_2, and T denotes a similarity threshold. In this Article, we used fingerprint-based Tanimoto similarity.

2.2. Network Analyses of Chemical Libraries. To examine the scale-freeness of chemical library networks, we analyzed the degree distributions of the networks, because we can conclude that chemical library networks are scale-free if their degrees of nodes distribute according to power laws. Power-law distributions are written as

$$p(x) = kx^{-\alpha}$$

where k is the normalization constant, and $\alpha > 1$ is the exponent. When the degree distribution of a scale-free network is plotted on a log−log scale, the plot shows a straight line. Thus, one of the easiest ways to determine whether a certain network shows scale-free behavior is to plot the degree distribution on a log−log scale: if the plot is linear, then we can assume the network is scale-free. An alternative means of examining scale-freeness is the use of a rank/frequency plot, in which the frequency of each value is counted, the data are ranked by descending order, and the rank of each value is plotted versus its frequency. If a network is scale-free, this rank/frequency plot on a log−log scale occurs as a straight line.

Even when the degree distributions appear to follow power laws, it is also probable that their networks belong to other classes of small-world networks. Thus, we also tried exponential distributions

$$p(x) = ke^{-\alpha x}$$

for curve fitting. If an exponential distribution better characterizes a network than power laws, then the network is of the broad-scale or single-scale network, both of which are other classes of small-world networks.[2]

We analyzed the degree distributions of the networks representing our data set libraries by the following procedures. First, we calculated all-to-all similarities among molecules in the libraries using Accelrys Pipeline Pilot.[15] We then calculated the number of similar compounds above the similarity threshold T, and subtracted 1 from this value to obtain the degree, because the all-to-all similarity calculation includes the similarity to itself. Finally, histograms and rank/frequency plots were drawn to check the scale-freeness. Here, given the benefits of comparing the degree distributions among different libraries, we used % degree and % frequency as the two axes of our histogram. These were calculated by dividing the degrees and frequencies by the number of molecules in the library. Histograms and rank/frequency plots were visualized with R.[16]

For quantitative analyses, the parameter α of power-law distributions was calculated using the scripts provided by Clauset et al.[17] This calculation was conducted using a maximum likelihood estimator (MLE) of the power-law distribution.[17,18] Previous studies indicate that the α of real-world scale-free networks tends to lie between 1.5 and 3.

In addition, curve fittings to both power laws and exponential distributions were performed using R. We adopted the Akaike information criterion (AIC)[19] as a measure of the goodness of fit.

To intuitively overview the network structures, they were also visualized using the igraph[20] package for R. The Fruchterman−Reingold algorithm,[21] a type of force-directed layout algorithm, was used for the graph layout. The idea of a force-directed layout algorithm is to consider a force between any two nodes, and to minimize the energy of the system by moving the nodes and changing the forces between them. This algorithm guarantees that topologically near nodes are placed in the same vicinity, and that far nodes are placed far from each other. GraphML,[22] an XML[23]-based text format for graph representation, was adapted as an input format for the graph. We wrote a Perl script to create GraphML format text files by converting the output files of the Accelrys Pipeline Pilot containing similarities among molecules in the libraries.

2.3. Data Sets. *a. Existing Chemical Libraries.* The ZINC database was used for our network analyses on the basis of

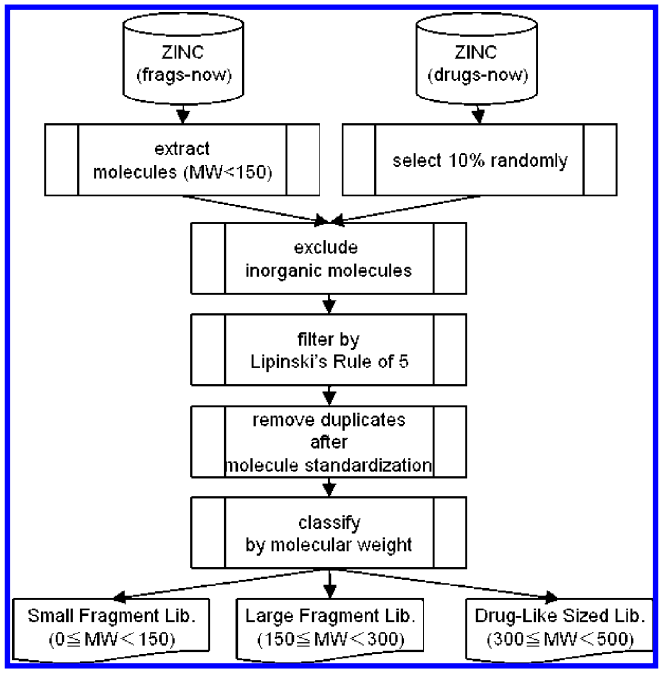

Figure 1. Extraction of input data sets from ZINC. Three libraries were produced on the basis of molecular weight after the filtering of drug-unlike molecules.

its status as a standard chemical database for virtual screening and provision of a relatively large compound set. Because of limited computational resources, however, we used two subsets of ZINC, frags-now and drugs-now, for the input of our analyses on the basis of our interest in FBDD and because our main focus is the application of network analyses to drug discovery. Because the drugs-now subset still includes a large number of molecules, we randomly selected 10% of them. From these inputs, we excluded drug-unlike molecules as follows: we first excluded molecules having inorganic atoms, filtered molecules by Lipinski's Rule of Five,[24] standardized stereochemical status and charges, and then removed duplicates, resulting in a set of 490 326 molecules. These 490 326 molecules were finally divided into the three library categories on the basis of their molecular weights: first, a ZINC small-fragment library of 30 797 molecules with a molecular weight less than 150; second, a ZINC large-fragment library of 106 513 molecules with a molecular weight between 150 and 300; and third, a ZINC drug-like-sized library of 353 016 molecules with a molecular weight between 300 and 500. Figure 1 illustrates this process, which was performed using the Accelrys Pipeline Pilot.

In addition to ZINC, we also analyzed fragment libraries from Enamine[25] and Life Chemicals,[26] both of which were released in 2007. We selected these because they are representative fragment libraries often used in FBDD, and because their library sizes are relatively large as compared to those of fragment libraries from other vendors. The numbers of fragments derived from these libraries was 37 436 and 8810, respectively.

b. Virtual Libraries Generated from Diverse Fragments. To obtain helpful insights for the FBDD, particularly in characterizing a fragment-linking strategy for the fragment hits-to-leads phase after fragment screening, we conducted additional in silico analyses with virtual libraries generated by the following procedures. First, after manually eliminating reactive molecules, we extracted 100 structurally diverse amine and carboxyl fragments each from our ZINC small-fragment library with a molecular weight less than 150 (see Tables S1 and S2 in the Supporting Information). The extracted fragments were dissimilar to each other, as ensured using a similarity threshold of 0.8. We extracted the two sets of 100 fragments on the assumption that a relatively large number of fragment hits is obtained in fragment screenings. Second, the extracted amine and carboxyl fragments were linked to generate virtual libraries. Here, to study the impact of linkers on the properties of network structures, we made several libraries that differed in the constituent linkers. Each step is explained in detail below.

Each molecule in our ZINC fragment library of 30 797 molecules with a molecular weight less than 150 was ionized at pH 7.4, and its charge was standardized using the *Standardize Molecule* component of the Accelrys Pipeline Pilot. From them, we extracted amine fragments whose SMILES[27] representations contained "C[NH3+]", and carboxyl fragments having a SMILES substring "[C](=O)[O-]". Fragments were regarded as amines only if they had one amine part and no carboxyl part. Similarly, carboxyl fragments were defined as those with only one carboxyl part and no amine parts. This extraction was performed by the *OESubSearch()* function of the OpenEye OEChem[28] python interface, as explained on the Website of Dalke Scientific Software, LLC.[29] To obtain structurally diverse fragment sets, we applied the *Diverse Molecules* component of the Accelrys Pipeline Pilot with MDLPublicKeys[30] to both the amine and the carboxyl fragments. Through this step, we obtained 100 amine and 100 carboxyl fragments.

We then linked these fragments in silico to generate chemical libraries. We made the following two types (Figure 2): direct linking of amine and carboxyl fragments without any linkers (Type 1), and linking with some linkers (Type 2). This process is detailed below.

The first type, direct linking, is relatively simple and was again performed in accordance with the explanation on the Dalke Scientific Software, LLC Website. Initially, attachment points were inserted into the SMILES representations of the fragments by transforming the molecules using the *OEUniMolecularRxn()* function of the OpenEye OEChem python interface with SMIRKS[31] "[C:1][NH3+]≫[C:1][NH2]*" and "[C:1](=[O:2])[O-]≫[C:1](=[O:2])*". Following this, "*" symbols were replaced with "%90". We then obtained a virtual library by simply concatenating the SMILES of the amine fragments with those of the carboxyl fragments, both of which have an attachment point represented by the symbol "%90", using the dot disconnect character.

The library of the second type (linking with some linkers) was generated by a slight modification of the methods used for the first type. The "**" and "%91" symbols were used for the amine fragments, and "*" and "%90" were used for the carboxyl fragments. We selected 15 linkers manually from our ZINC small-fragment library with molecular weight less than 150, with consideration given to their diversity (see Table S3 in the Supporting Information). Employing a methodology similar to that for the SMILES representations of amine and carboxyl fragments, we generated the SMILES of these linkers. Here, "*" and "%90" symbols were used for amino groups, and "**" and "%91" were used for carboxyl groups. After obtaining the SMILES of the amine and carboxyl fragments and the amino-acid chains, we then

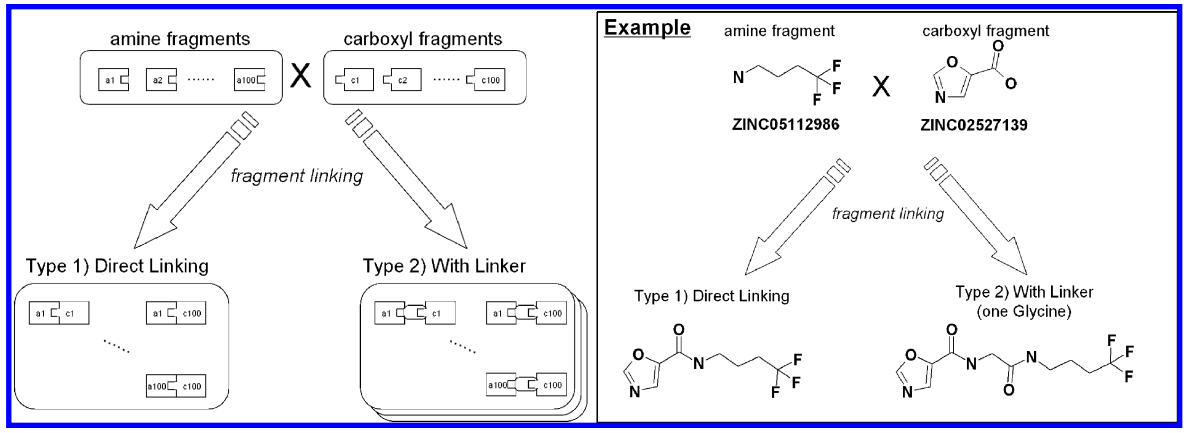

Figure 2. Two types of fragment linking: direct linking of amine and carboxyl fragments without linkers (Type 1), and linking with some linkers (Type 2). An example is shown at the right.

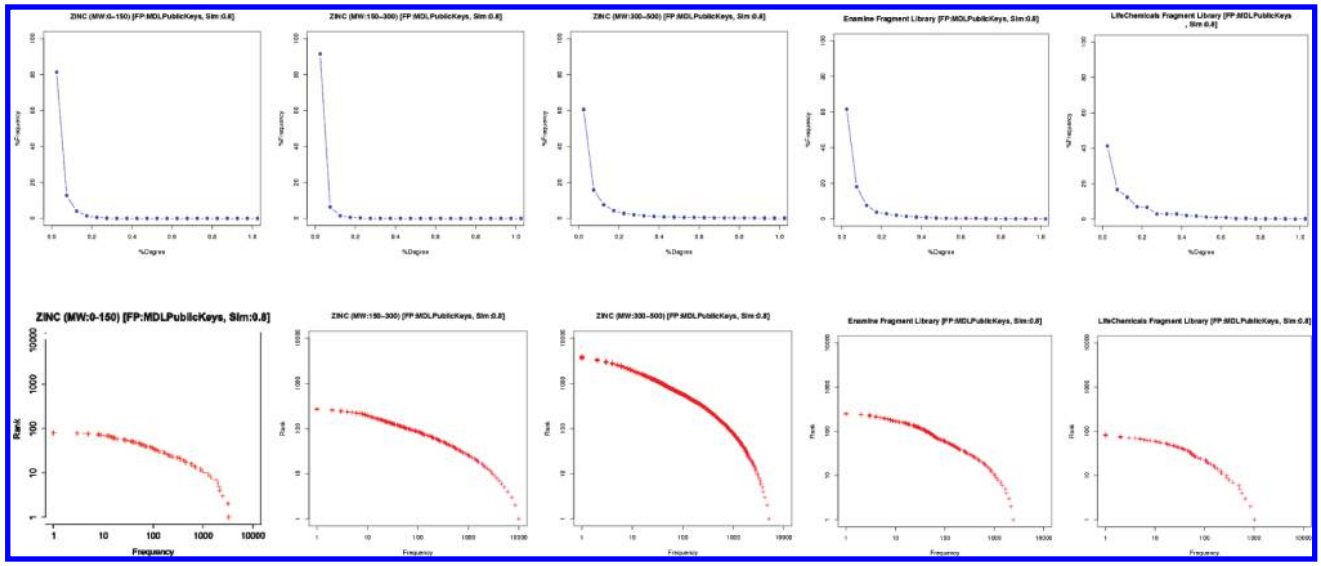

Figure 3. Histograms (top) and rank/frequency plots (bottom) of molecules in the ZINC database with a molecular weight less than 150, between 150 and 300, and between 300 and 500 (first three on the left), the Enamine fragment library (second to the right), and the Life Chemicals fragment library (right). MDLPublicKeys is used, and similarity threshold is set to 0.8.

generated virtual libraries by concatenating the SMILES of these three parts using the dot disconnect character in the same way as for the first type. A virtual library was prepared by linking each amine and carboxyl fragment with one glycine linker using the above procedure. A virtual library with one randomly selected linker was also generated for comparison.

3. RESULTS

3.1. Small-Worldness of Chemical Library Networks. We examined the degree distributions of three libraries derived from the ZINC database (small fragment library, large fragment library, and drug-like-sized library), and fragment libraries from Enamine and Life Chemicals. The ZINC database was selected because it is a standard chemical database for virtual screening and provides an exhaustive range of molecules. Here, we excluded drug-unlike molecules from our analyses by using filters such as Lipinski's Rule of Five. We divided molecules by the molecular weights 300 and 150 because, in FBDD, fragments are usually considered to have a molecular weight less than 300 or even less than 250, and because a molecular weight less than 150 is considered to be appropriate to the conduct of our analyses of virtual libraries in which such small parts are linked. Fragment libraries from Enamine and Life Chemicals were selected because they are representative fragment libraries often used in FBDD.

The histograms and rank/frequency plots of molecules in the ZINC database with a molecular weight less than 150, between 150 and 300, between 300 and 500, as well as in the Enamine Fragment Library and Life Chemicals Fragment Library are shown in Figure 3. These used the MDLPublicKeys, and similarity threshold was set to 0.8. All degree distributions in Figure 3 appear to show power-law behavior with highly right-skewed plots, indicating the presence of several hubs in the networks. Further, although slight deviations are seen at large and small frequencies, linear trends exist in all rank/frequency plots.

To examine the robustness of our network analysis results, we repeated the same procedures using different similarity thresholds and fingerprints in the Life Chemicals fragment library, which has a relatively small number of molecules. First, Figure 4 shows the histograms and rank/frequency plots of the Life Chemicals fragment library when MDLPublic-Keys were used and similarity threshold was changed from 0.7 to 0.9 at 0.05 intervals. These plots indicate that the trend

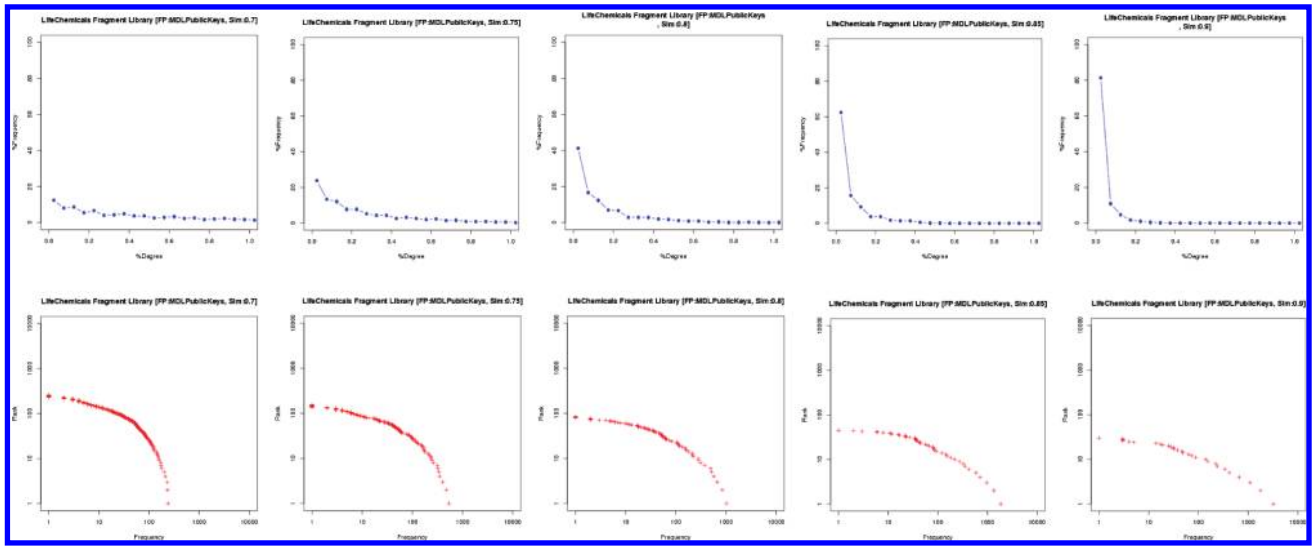

Figure 4. Histograms (top) and rank/frequency plots (bottom) of the Life Chemicals Fragment Library. MDLPublicKeys is used, and the similarity threshold is changed from 0.7 (left) to 0.9 (right) at 0.05 intervals.

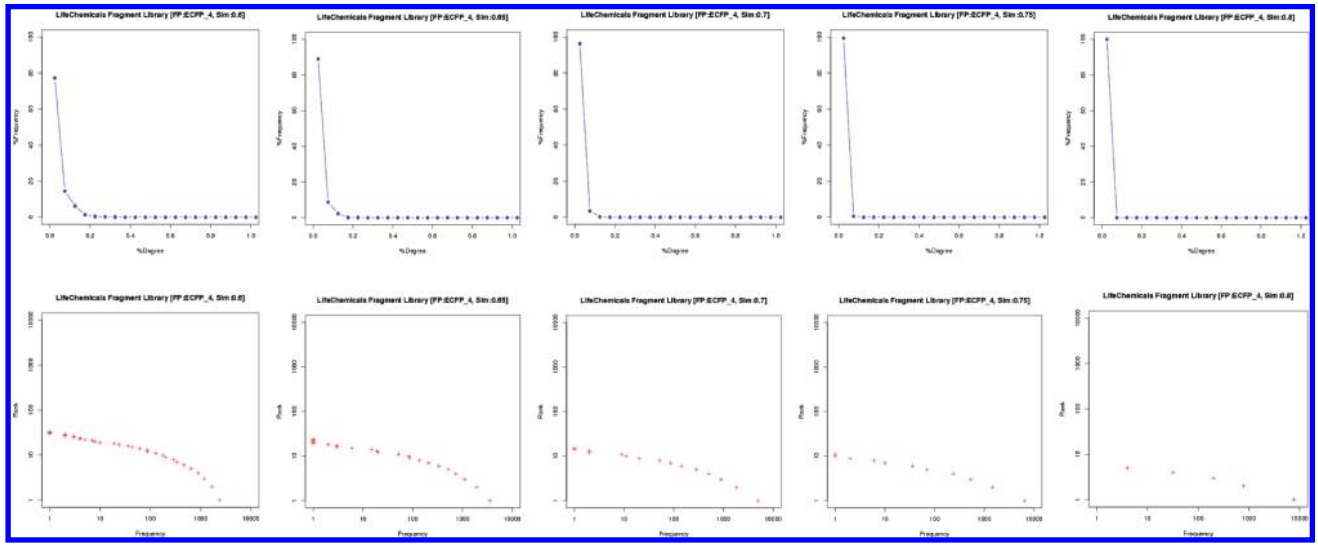

Figure 5. Histograms (top) and rank/frequency plots (bottom) of the Life Chemicals fragment library. ECFP_4 is used, and similarity threshold is changed from 0.6 (left) to 0.8 (right) at 0.05 intervals.

is maintained with a reasonable similarity threshold range for normal use, albeit that the behavior of the chemical library network is gradually lost as the threshold becomes too high or low. Considering that the number of similar compounds decreases with high similarity thresholds and increases with low thresholds, this result appears reasonable.

Next, Figure 5 shows plots developed using ECFP_4 as calculated by the Accelrys Pipeline Pilot and a change in similarity threshold from 0.6 to 0.8 at 0.05 intervals. Although the range of similarity threshold is different from that of the plots with MDLPublicKeys, the network using ECFP_4 shows power-law behavior. The existence of a difference in the right threshold range among different fingerprints is also reasonable, given that all fingerprints have their own characteristics and that the number of similar compounds differs among fingerprints when the same similarity threshold is used. We therefore conclude that the chemical library network retains its power-law behavior if the right similarity threshold is used regardless of fingerprint.

The calculated values of α are shown in Table 1, which includes the α value of the Type 1 virtual library by direct linking, as mentioned below. All α values in Table 1 lie in the typical range between 1.5 and 3.

Also, α values of the Life Chemicals fragment library with various fingerprints and similarity thresholds are shown in Table 2. Here, several α values in the table are 3.5, because the specification of the script we used allows the calculation of α values less than or equal to 3.5.

These results indicate that chemical library networks appear to have a power-law property, albeit that slight deviations are present. Their small world phenomena can also be confirmed intuitively by visualization of the network structures. Figure 6a illustrates correspondence between the histogram of degree distribution of a chemical library network and the visualization of the network. As we adopted the Fruchterman−Reingold algorithm for the layout of this graph, nodes with low degrees tend to be located in the external side of the visualization, while those with high degrees tend to be clustered in the internal side. As an example, Figure 6b shows the network structure of the Life Chemicals fragment library with MDLPublicKeys and a similarity threshold of 0.8. This visualization indicates that

Table 1. α and Other Descriptive Values of Input Libraries[a]

library (without singleton)	no. of molecules	MW range	MW average	α
ZINC small-fragment library	30 797	0−150	129.63	2.31
ZINC large-fragment library	106 513	150−300	257.09	2.56
ZINC drug-like-sized library	353 016	300−500	385.63	2.16
Enamine fragment library	37 436	60−300	253.559	2.19
Life Chemicals fragment library	8810	150−300	246.689	2.25
virtual library (Type 1)	10 000	150−280	246.46	2.19

[a] MDLPublicKeys is used, and similarity threshold is set to 0.8.

Table 2. α Values of the Life Chemicals Fragment Library with Various Fingerprints and Similarity Thresholds

library (without singleton)	fingerprint	similarity threshold	α
Life Chemicals fragment library	MDLPublicKeys	0.70	1.66
		0.75	1.92
		0.80	2.25
		0.85	2.28
		0.90	3.05
	ECFP_4	0.60	3.17
		0.65	3.5
		0.70	3.5
		0.75	3.32
		0.80	3.5

Table 3. AIC Values of the Input Libraries[a]

library	MW range	AIC (power-law)	AIC (exponential)
ZINC small-fragment library	0−150	1408.016	1209.356
ZINC large-fragment library	150−300	7010.481	6145.185
ZINC drug-like-sized library	300−500	84 986.27	76 956.63
Enamine fragment library	60−300	3725.104	3016.674
Life Chemicals fragment library	150−300	1069.444	912.91

[a] MDLPublicKeys is used, and similarity threshold is set to 0.8.

the network of the Life Chemicals fragment library has many nodes with small degrees and several nodes with large degrees, called hubs. Figure S1 in the Supporting Information shows visualizations of the network structures of other libraries with MDLPublicKeys and a similarity threshold of 0.8, while Figure S2 shows visualizations of the Life Chemicals fragment library with various thresholds.

To better understand network properties, we next tried curve fitting with exponential distributions as well as power laws. This approach is based on the likelihood that, even when the degree distributions look to follow power laws, networks belong to classes of small-world networks other than scale-free networks. Table 3 shows the values of Akaike information criterion (AIC), a measure of the goodness of fit, of our data sets for comparison. Because the preferred model is the one with the lowest AIC value, the chemical library networks appear to fit better to exponential distributions than power laws, in turn indicating that the networks belong to broad-scale or single-scale networks rather than to scale-free networks. However, the differences in AIC values are not particularly significant. Moreover, scale-free, broad-scale, and single-scale networks are just three classes of small-world networks, and network properties, such as the existence of hubs inferred from our previous results, do not change even if the networks are more properly categorized as broad-scale or single-scale networks than scale-free networks.

3.2. Network Structure of Virtual Libraries Generated from Diverse Fragments. In addition to the existing chemical libraries, we also analyzed the degree distributions of the Type 1 and Type 2 virtual libraries. First, the degree distribution of the virtual library Type 1 (direct linking of amine and carboxyl fragments without linkers) was examined with MDLPublicKeys and a similarity threshold 0.8. The histogram and rank/frequency plot of the virtual library Type 1 is shown in the upper left of Figure 7. The calculated value of α is 2.41, as shown in Table 1. From these plots, we assume that the virtual library Type 1 also has power-law behavior, even though it was derived from structurally diverse fragments.

Similar analyses were done for the virtual library Type 2. The histogram and rank/frequency plot of the virtual library Type 2 with one glycine linker are shown in the upper middle of Figure 7. We cannot assume from these plots that the virtual library with the same linker displayed power-law behavior. Interestingly, only one linker had a major impact on the network structures. While the Type 1 library retained its power-law behavior, the Type 2 library with one linker lost it. For comparison, the histogram of the virtual library with one randomly selected linker from 15 diverse, manually extracted linkers in the upper right side of Figure 7 shows

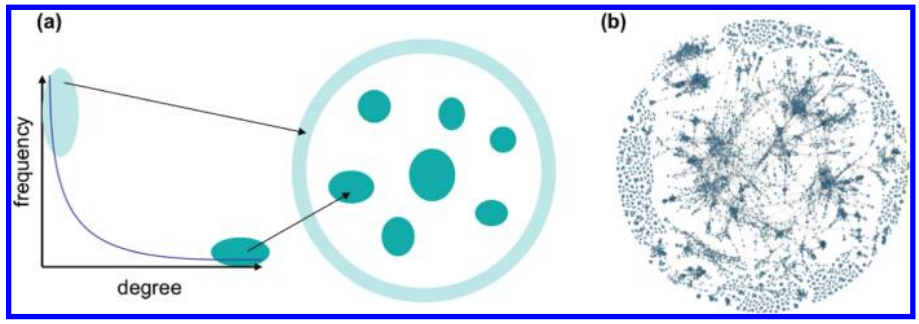

Figure 6. Histogram of the degree distribution of a chemical library network and visualization of the network. Part (a) illustrates correspondence between the histogram and visualization. Part (b) shows the network structure of the Life Chemicals fragment library as an example. MDLPublicKeys is used, and similarity threshold is set to 0.8. Singletons (i.e., nodes without any edges) are omitted in this visualization.

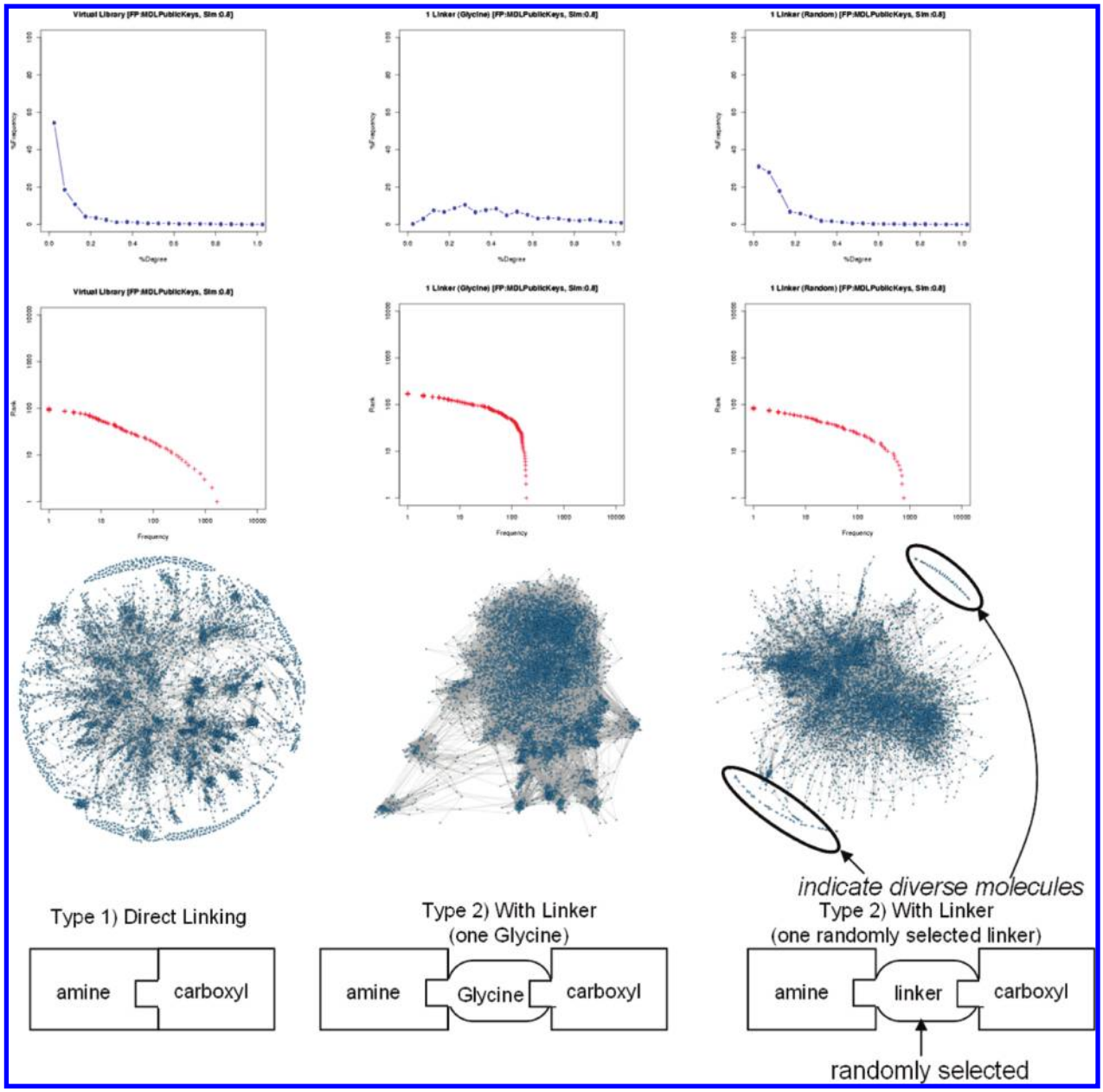

Figure 7. Histograms, rank/frequency plots, and network structures of the virtual library Type 1 (left), Type 2 with one glycine linker (middle), and Type 2 with one randomly selected linker (right).

that the distribution approaches power-law behavior when compared to that of the virtual library with a single glycine linker. Also, the number of molecules with small degrees is much larger than that of the Type 2 library with one glycine. From these observations, we conclude that the structural diversity of linkers greatly influences the diversity of the libraries generated by fragment linking.

To understand these differences in network structure more intuitively, we visualized the network structures as illustrated in the bottom of Figure 7, which shows from left the network structures of the virtual library Type 1, Type 2 with one glycine linker, and Type 2 with one randomly selected linker. Here, MDLPublicKeys is used and similarity threshold is set to 0.8, and singletons are omitted. The network of the virtual library Type 1 has many nodes with small degrees, and several nodes with large degrees, which illustrates its power lawness. On the contrary, the network of Type 2 with one glycine linker shows relatively dense connections, indicating the small diversity of the library. The network of Type 2 with one randomly selected linker also has many links, albeit with lower density than that of the Type 2 network with one glycine linker. Also, because we used the Fruchterman−Reingold algorithm for graph layout, nodes in the outer regions indicate the existence of relatively diverse molecules, which guarantees that topologically near nodes are placed in the same vicinity and that far nodes are placed far from each other. These illustrations highlight the marked influence a single linker has on network structure and indicate the need to consider the variety of linkers for molecular diversity when fragment linking is performed.

4. DISCUSSION

4.1. Small-World Behavior of Chemical Library Networks. Our analyses indicate that existing chemical libraries have small-world behaviors with power-law properties when

represented as networks. This result supports the argument by Benz et al.,[6] who reported that the sizes of molecular similarity clusters distribute according to power laws, because both analyses are based on the number of similar molecules in chemical libraries. However, while Benz et al. divided chemical libraries into clusters based on similarity, we adopted a different approach that takes equal account of the number of similar types of every molecule. Considering that clusters produced by clustering algorithms are heavily influenced by the algorithms and thresholds, we believe our approach better explains the underlying properties of chemical libraries.

Given our study results, we are interested in the mechanisms by which chemical library networks acquire small-world properties with power law. Other investigators have argued that creation of scale-free networks involves two mechanisms, growth and preferential attachment.[1] If a network has a growth mechanism, then the number of nodes and edges increases linearly in time. In contrast, preferential attachment is the assumption that the likelihood of receiving new edges increases with the node's degree.

Chemical library networks can be considered to have both of these mechanisms. With regard to the growth mechanism, the size of chemical libraries increases in time because new molecules are continually synthesized or purchased. The preferential attachment mechanism is also applicable because molecules similar to those already existing tend to be added to the libraries for several reasons. One such reason is the use of combinatorial chemistry technology, which allows many similar molecules with the same scaffolds to be added to the libraries simultaneously. In the case of in-house libraries in the pharmaceutical industry, many molecules similar to biologically active compounds such as patent compounds and HTS hits are synthesized or purchased to obtain more promising candidates or to reveal structure−activity relationships (SARs). These factors support a preferential attachment mechanism.

With our in silico experiments, we also saw the emergence of power-law behavior in a virtual library generated by direct linking of diverse fragments, even though each fragment had no fragments similar to it. As compounds are synthesized from small building blocks, this phenomenon might indicate that chemical libraries obtain their power-law behavior during the synthesis of constituent molecules from small parts.

4.2. Impact of the Variety of Linkers on Network Structures. The many different approaches to converting fragment hits to leads are categorized into the following four types: fragment evolution, fragment linking, fragment optimization, and fragment self-assembly.[32] In this Article, we analyzed the structures of chemical library networks generated by one of these approaches, fragment linking.

As we have shown, varieties of linkers profoundly influence the network structures of virtual libraries generated by fragment linking. Use of a single linker only destroys the power-law behavior relatively easily. The differences among the network structures of the three virtual libraries mentioned in section 3.2 can be attributed to the extent of the linker's effect on the similarities among molecules in the libraries. This is because similarities tend to be higher when molecules contain the same substructure as a linker.

Thus, to attain sufficient diversity of the library, fragment linking should be done using as many diverse linkers as possible. Because the distance between two active sites is small,[33] the best option is thought to be direct linking of fragments without any linkers.

4.3. Prioritization Method for Fragment Linking Strategy. With the recent strong interest in FBDD, efficient compound-prioritization methods are in great demand, particularly fragment-linking strategies for the fragment hits-to-leads phase after fragment screening. One simple approach to choosing two fragment hits to be linked is to select those having top activities as the first candidate for combination. Although this strategy appears reasonable, the combination of two top hits is often not the best, and other combinations of fragment hits with lower activities may show more potent activity. This in turn makes the prioritization of compound syntheses by fragment linking a relatively challenging task.

Given this situation, we present an efficient compound-prioritization method for fragment linking based on the observation of small-world phenomena in virtual libraries generated by the direct linking of diverse fragments. In our strategy, compound prioritization is realized by selecting fragments as a starting point such that the linked compounds become hubs in the library, or nodes with the highest degrees. Here, we assume that molecules similar to nonhits are also nonhits. If this is the case, then when hubs turn out to be nonhits, many combinations of fragment hits can be eliminated as candidates, on the basis that many molecules linked to the hubs can also be considered as nonhits.

To examine the efficiency of hub-based selection, we compared the following three selection approaches: (1) selection of hub compounds, (2) selection of cluster centers, and (3) random selection of compounds. Here, previously selected compounds and their neighbors are excluded when the next one is chosen. Figure 8 illustrates the difference by visualizing the network of the Type 1 virtual library. The red nodes are 10 linked compounds and their adjacent compounds. Part (a) corresponds to hub selection, part (b) to cluster center selection, and part (c) to random selection, respectively. As shown in Figure 8a, many nodes are regarded as evaluated when only 10 molecules in the library are assayed, indicating the efficiency of our strategy. Although the result of the cluster center selection approach illustrated in Figure 8b is not particularly bad, few molecules are regarded as evaluated when 10 molecules are assayed, particularly in Figure 8c. This illustrates the inefficiency of other strategies. In approaches other than hub selection, it is highly likely that many trials would need to be repeated until promising compounds are found because their elimination of unlikely compounds is inefficient.

In addition to the network visualizations, Figure 9 shows the number of linked compounds and their neighbors when the number of selected compounds is changed from 1 to 100. Here, three types are prepared for the cluster center selection approach by changing the average number of molecules in each cluster with MDLPublicKeys 5, 10, and 20. Three types are also prepared for random selection. Because the number of linked compounds and their neighbors by hub selection is always significantly higher than those of the others, this figure also highlights the distinct advantage of our approach.

Although our proposal is relatively simple and counter-arguments might be presented, considering the current lack of any "Golden Rule" for compound prioritization on

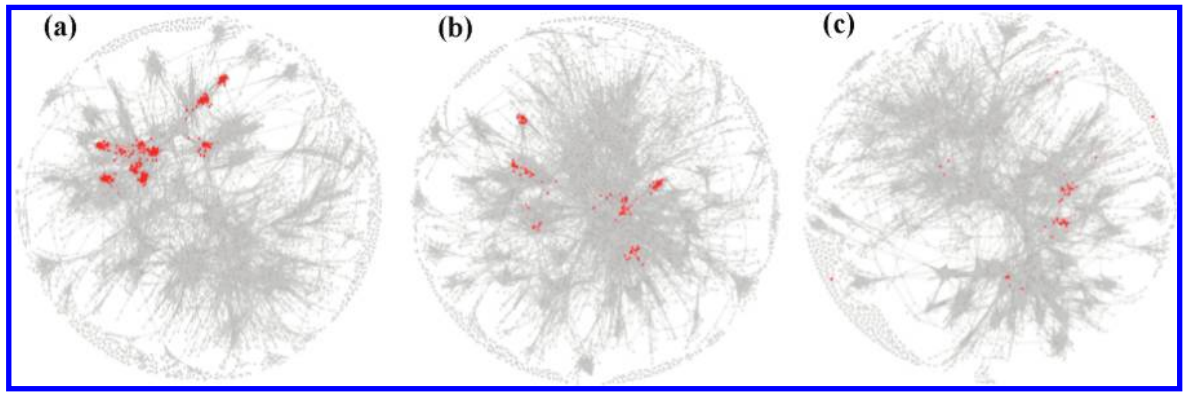

Figure 8. Network structures of three compound-prioritization methods. The red nodes are 10 linked compounds and their adjacent compounds. Part (a) corresponds to our approach, in which fragments are selected as a starting point such that the linked compounds become hubs, part (b) to cluster center selection, and part (c) to random selection.

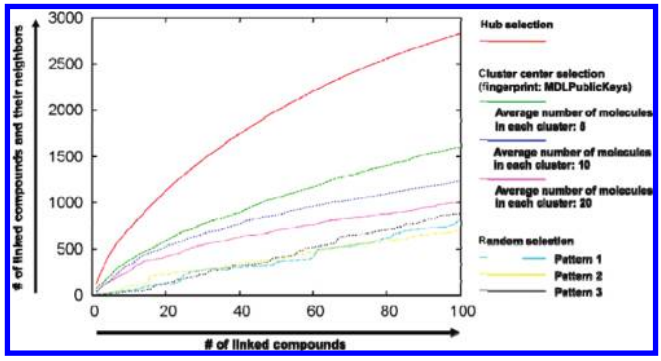

Figure 9. Graph of number of selected compounds versus number of selected compounds and their neighbors. Three types are prepared for the cluster center selection approach by changing the average number of molecules in each cluster with MDLPublicKeys to 5, 10, and 20. Three types are also prepared for random selection.

fragment linking, we believe that this simple approach can be used as a guiding principle for this purpose.

4.4. Other Probable Applications of Network Analyses of Chemical Libraries. Although our analyses of chemical libraries focused on the degree distribution of networks generated with the metrics of the fingerprint-based Tanimoto similarity, there are other possible methods for network analyses, including the use of other metrics by which edges are made, and to consider other indices for node importance in place of degree.

With regard to the metrics by which nodes are linked, other similarity metrics may be used, such as shape similarity. It might also be interesting to link nodes by edges if molecules are active in the same theme. The generation of networks of chemical libraries from these and various other points of view should allow network analyses to provide insightful information on chemical libraries.

Another possible method for network analysis is to adopt other indices for node importance in place of degrees. When the importance of nodes is considered, the centrality of networks is often discussed. There are three well-known policies for the centrality of networks: degree centrality, closeness centrality, and betweenness centrality.[34] The first, degree centrality, is the most widely adopted because of its simplicity, and we also adopted it here for our network analyses. In this policy, hubs, or nodes, with large degrees, are regarded as center points in a network. Alternatively, closeness centrality and betweenness centrality consider center points from different standpoints, as illustrated in Figure 10. Nodes with high closeness are those that have a short distance to other nodes. In the closeness centrality policy, such nodes are regarded as center points in networks, because nodes with high degrees are not always suitable for consideration as center points, as illustrated in the left side of Figure 10. In the application of this policy to chemical libraries, nodes with high closeness may be used as representative molecules for SAR analyses among similar types. As for betweenness centrality, nodes with high betweenness are those that occur on many of the shortest paths between other nodes. Such nodes are considered important because they act as bridges between two separated subnetworks, as illustrated in the right side of Figure 10. In network analyses of chemical libraries, nodes with high betweenness might provide key SAR information when there are two clusters of interest, and the nodes exist between the two, because they work as bridges between the two different scaffolds.

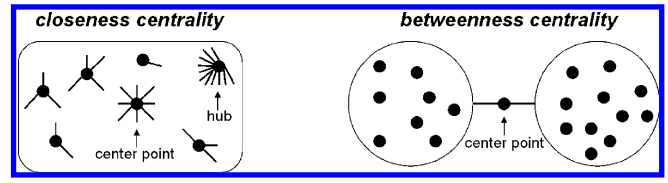

Figure 10. Schematic illustrations of closeness centrality and betweenness centrality.

As shown by these examples, network analyses of chemical libraries should provide a range of benefits to various aspects of the drug-discovery process, including those not mentioned in this Article. For future research, we plan to apply network analysis to fields such as library design.[35]

5. CONCLUSIONS

We have shown that chemical libraries have small-world behavior when represented as networks. Our analyses indicated that chemical libraries acquire such properties in the process of synthesizing constituent molecules from small parts. Making use of this phenomenon, we proposed an efficient compound-prioritization method for fragment linking. The variety of linkers was also shown to be relatively important for molecular diversity when fragment linking was performed. These results indicate the benefits of network analyses of chemical libraries. These analyses are also likely to benefit other fields of drug-discovery research, such as library design.

Abbreviations. AIC, Akaike information criterion; FBDD, fragment-based drug discovery; HTS, high-throughput screening; MLE, maximum likelihood estimator; SARs, structure−activity relationships.

ACKNOWLEDGMENT

We would like to thank Ms. Naoko Katayama, Dr. Makoto Oku, Dr. Hideyoshi Fuji, Dr. Takeshi Hondo, and Dr. Hitoshi Sakashita for their many insightful suggestions and helpful discussions, and Dr. Yuzo Matsumoto for carefully reviewing the manuscript.

Supporting Information Available: 100 amine fragments; 100 carboxyl fragments; 15 linkers; network structures of various chemical libraries with MDLPublicKeys and a similarity threshold of 0.8; and network structures of the Life Chemicals fragment library with various thresholds. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Barabasi, A.-L.; Albert, R. Emergence of Scaling in Random Networks. *Science* **1999**, *286*, 509–512.
(2) Amaral, L. A. N.; Scale, A.; Barthelemy, M.; Stanley, H. E. Classes of Small-World Networks. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 11149–11152.
(3) Baldi, P.; Frasconi, P.; Smyth, P. *Modeling the Internet and the Web*; John Wiley and Sons Ltd.: Chichester, England, 2003.
(4) Pareto, V. *Le Cours d'Economie Politique*; Macmillan: London, 1897.
(5) Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabasi, A.-L. The Large-Scale Organization of Metabolic Networks. *Nature* **2000**, *407*, 651–654.
(6) Benz, R. W.; Swamidass, S. J.; Baldi, P. Discovery of Power-Laws in Chemical Space. *J. Chem. Inf. Model.* **2008**, *48*, 1138–1151.
(7) Bohacek, R. S.; Mcmartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.
(8) Tanimoto, T. T. *An Elementary Mathematical Theory of Classification and Prediction; IBM Internal Report Nov. 1958*; IBM Corp.: New York, 1958.
(9) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(10) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-Based Drug Discovery. *J. Med. Chem.* **2004**, *47*, 3463–3482.
(11) Alex, A. A.; Flocco, M. M. Fragment-Based Drug Discovery: What Has It Achieved So Far. *Curr. Top. Med. Chem.* **2007**, *7*, 1544–1567.
(12) Hajduk, P. J.; Greer, J. A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned. *Nat. Rev. Drug Discovery* **2007**, *6*, 211–219.
(13) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent Developments in Fragment-Based Drug Discovery. *J. Med. Chem.* **2008**, *51*, 3661–3680.
(14) Orita, M.; Ohno, K.; Niimi, T. Two 'Golden Ratio' Indices in Fragment-Based Drug Discovery. *Drug Discovery Today* **2009**, *14*, 321–328.
(15) *Pipeline Pilot, version 7.5*; Accelrys Software Inc.: San Diego, CA, 2008.
(16) *R, version 2.9.1*; The R Project for Statistical Computing: Auckland, New Zealand, 2009.
(17) Clauset, A.; Shalizi, C. R.; Newman, M. E. J. Power-law distributions in empirical data. 2007, arXiv:physics/0706.1062. arXiv.org ePrint archive. http://arxiv.org/abs/0706.1062 (accessed Aug. 11, 2009).
(18) Newman, M. E. J. Power Laws, Pareto Distributions and Zipf's Law. *Contemp. Phys.* **2005**, *46*, 323–351.
(19) Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Auto. Control* **1974**, *19*, 716–723.
(20) *The igraph library, version 0.5.2*; The igraph project: Budapest, Hungary, 2009.
(21) Fruchterman, T. M. J.; Reingold, E. M. Graph Drawing by Force-Directed Placement. *Software: Practice and Experience* **1991**, *21*, 1129–1164.
(22) Brandes, U.; Eiglsperger, M.; Herman, I.; Himsolt, M.; Marshall, M. S. GraphML Progress Report: Structural Layer Proposal. *Proc. 9th Intl. Symp. Graph Drawing (GD '01)*; Springer-Verlag: Berlin, 2002; Vol. LNCS 2265, pp 501−512.
(23) Bray, T.; Paoli, J.; Sperberg-McQueen, C. M.; Maler, E.; Yergeau, F. *Extensible Markup Language (XML) 1.0*, 5th ed.; W3C Recommendation 26 November 2008. http://www.w3.org/TR/2008/REC-xml-20081126 (accessed Aug. 11, 2009).
(24) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
(25) Enamine Ltd. Libraries for Fragment-Based Drug Discovery. http://enamine.net/index.php?option=com_content&task=view&id=116 (accessed Aug. 11, 2009).
(26) Life Chemicals Inc. Fragments Library. http://www.lifechemicals.com/services/fragments (accessed Aug. 11, 2009).
(27) Daylight Chemical Information Systems, Inc. Daylight Theory: SMILES - A Simplified Chemical Language. http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed Aug. 11, 2009).
(28) *OEChem TK, version 1.7.0*; OpenEye Scientific Software: Santa Fe, NM, 2009.
(29) Dalke, A. At Dalke Scientific Software, LLC. Attachment points. http://www.dalkescientific.com/writings/diary/archive/2005/05/07/attachment_points.html (accessed Aug 11., 2009).
(30) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Model.* **2002**, *42*, 1273–1280.
(31) Daylight Chemical Information Systems, Inc. Daylight Theory: SMIRKS - A Reaction Transform Language. http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (accessed Aug. 11, 2009).
(32) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-Based Lead Discovery. *Nat. Rev. Drug Discovery* **2004**, *3*, 660–672.
(33) Cherry, M.; Mitchell, T. Introduction to Fragment-based Drug Discovery. In *Fragment-Based Drug Discovery, A Practical Approach*, 1st ed.; Edward, R. Z., Michael, J. S., Eds.; Wiley: Chichester, UK, 2008; pp 1−14.
(34) Koschützki, D.; Lehmann, K. A.; Peeters, L.; Richter, S.; Tenfelde-Podehl, D.; Zlotowski, O. Centrality Indices. In *Network Analysis: Methodological Foundations*; Brandes, U., Erlebach, T., Eds.; Springer-Verlag: Berlin, 2005; Vol. LNCS 3418, pp 16−61.
(35) Gregori-Puigjane, E.; Mestres, J. Coverage and Bias in Chemical Library Design. *Curr. Opin. Chem. Biol.* **2008**, *12*, 359–365.

CI900123V