# Discovery of Chemical Compound Groups with Common Structures by a Network Analysis Approach (Affinity Prediction Method)

Shigeru Saito,[†,§] Takatsugu Hirokawa,[†] and Katsuhisa Horimoto*[,†,‡]

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan, Chem & Bio Informatics Department, INFOCOM Corporation, Sumitomo Fudosan Harajuku Building, 2-34-17 Jingumae, Shibuya-ku, Tokyo, 150-0001, Japan, and and Institute of Systems Biology, Shanghai University, 99 Shangda Road, Shanghai 200444, China

Received July 9, 2010

We developed a method in which the relationship between chemical compounds, characterized by the secondary dimensional descriptors by a standard method, is first determined by network inference, and then the inferred network is divided into the compound groups by network clustering. We applied this method to 279 active inhibitors of factor Xa found by the first screening. A large network of 266 active compounds connected with 408 edges emerged and was divided into 10 clusters. Surprisingly, the chemical structures that were common within the clusters, but diverse between them, could be extracted. The activity differences between the clusters provide rational clues for the systematic synthesis of derivatives in the lead optimization process, instead of empirical and intuitive inspections. Thus, our method for automatically grouping the chemical compounds by a network approach is useful to improve the efficiency of the drug discovery process.

## 1. INTRODUCTION

Novel computational approaches and methodologies are increasing the efficiency of drug discovery, which involves numerous processes.[1] Indeed, various computational approaches in virtual screening are utilized to predict the activity of hypothetical compounds, based on the quantitative structure−activity relationship (QSAR).[2−5] In particular, the selection of compounds from a library or database of compounds is widely used to identify those that are likely to possess a given activity, when a single bioactive reference structure is available.[6−8] In this approach, fingerprint-based similarity searching is performed to identify the database molecules that are most similar to a user-defined reference structure.[9] Furthermore, the support vector machine is utilized to predict the activity of newly synthesized compounds with high accuracy.[10] The principal component analysis (PCA) also presents the relationship between the compounds, to allow a visual investigation of their activities in the principal component space. In particular, it generates a concept for the distribution of chemical compounds, named the chemical space, where different chemical compounds are reasonably distributed, depending on their corresponding origins.[11]

In spite of the popularity of computational approaches, empirical and intuitive approaches are still employed in drug discovery processes.[1] One reason for retaining the empirical and intuitive approaches is that after the first screening, the active compounds are usually compared in terms of the relationship between the chemical structure and its activity, before the next step of synthesizing the derivatives for selecting the ultimate lead. Unfortunately, this step partially depends on the empirical selection of the candidates for the chemical synthesis of the drug target, with reference to the chemical structures of the active and inactive compounds obtained by the first screening. Indeed, the structural information on the active compounds after the first screening is not fully utilized for selecting candidates of seeds for the derivative synthesis. Thus, the extraction of useful information about chemical structure and activity, in an automatic and visual manner, is desirable to systematically and efficiently synthesize derivatives for drug discovery.

We now propose an automatic method to visually group chemical compounds based on their structures, by using two types of network analysis methods. One is the network inference method. In the present study, we use the path consistency algorithm,[12] one of the graphical models from the family of probability models simplified by the conditional independences inherent in the graph,[13] which can visually infer the relationships between variables in a network form. Another is the network clustering method. This is a method to extract one property, named the "community structure", which indicates that the vertices in networks are often clustered into tightly knit groups, with a high density of within-group edges and a lower density of between-group edges.[14] This method is useful to automatically group the variables into some clusters from the connected network structure. Here we utilized the two network analysis methods to assess the relationship between chemical compounds. The utility of the present method is demonstrated by a set of chemical compounds after the first screening. The merits and pitfalls of the present method are also discussed, in terms of the previous computational methods.

* Corresponding author. E-mail: k.horimoto@aist.go.jp.. Telephone: +81 3 3599 8711.

[†] National Institute of Advanced Industrial Science and Technology (AIST).

[§] INFOCOM Corporation.

[‡] Shanghai University.

10.1021/ci100262s
Published on Web 12/09/2010

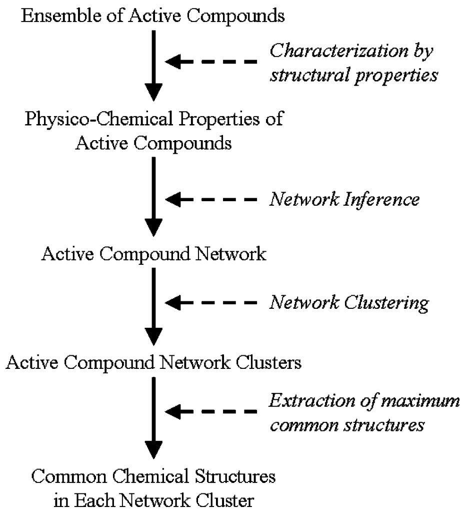

Ensemble of Active Compounds
Characterization by structural properties
Physico-Chemical Properties of Active Compounds
Network Inference
Active Compound Network
Network Clustering
Active Compound Network Clusters
Extraction of maximum common structures
Common Chemical Structures in Each Network Cluster

**Figure 1.** Workflow of the present method. The present method is schematically described in four steps.

## 2. MATERIALS AND METHODS

**2.1. Overview of the Present Method.** An overview of our method is schematically described in Figure 1. First, the chemical compounds selected by the first screening, in terms of drug activity, are characterized by their secondary structure properties, by a standard procedure. Second, the relationships between the compounds are investigated by a network inference method, the path consistency algorithm.[12] Third, the inferred network structures are divided into groups by a network clustering method, the Newman algorithm.[14] Fourth, the maximum common structures of the compounds are extracted in each cluster by a standard method. Thus, the characteristic features of the chemical structures hidden in the active chemical compounds are revealed visually and automatically by network analysis methods. The details of each step are described below.

**2.2. Data Set.** The data set contains a wide series of inhibitors of factor Xa extracted from the literature, all sharing a benzamidine moiety.[15] The considered data set contains 279 very active compounds ($K_\mathrm{i}$ lower than 10 nM) among a total of 435 chemical compounds, also including 156 low-activity compounds ($K_\mathrm{i}$ higher than 1 $\mu$M).

**2.3. Descriptors.** The calculated 2D descriptors were derived from the commercially available software, MOE, by Chemical Computing Group Inc. (http://www.chemcomp.com/). As a preprocessing step for the following analyses, the values of each descriptor were standardized by their averages and standard deviations. In this step, the number of descriptors was reduced, by leaving only the continuous values of the descriptors. Finally, 158 descriptors were used.

**2.4. Network Inference by Path Consistency (PC) Algorithm with Modifications.** The path consistency (PC) algorithm is a network inference method based on the graphical model.[12] The original PC algorithm is composed of two parts: the undirected graph inference by the partial correlation coefficient and the following directed graph generated by using the orientation rule. The present method partially exploits the first part of the PC algorithm, because the aim of the present application of the network inference method is to scrutinize the relationships between the chemical compounds, without the causality.

The algorithm for the first part is simple. The relationship between two variables is tested from the lower partial correlation coefficient to the higher one. For example, the relationship between the two variables is first tested by the zero-th partial correlation coefficient. If the null hypothesis is accepted, i.e., no association between the two variables, then no further test is performed for the higher order of the partial correlation coefficient. If it is rejected, then the relationship between the two variables is tested by the first partial correlation coefficient. In general, the $(m - 2)$-th order of the partial correlation coefficient is calculated between two variables, given $(m - 2)$ variables, i.e., $r_{ij,\mathrm{rest}}$, between $X_i$ and $X_j$, given the 'rest' of the variables, $\{X_k\}$ for $k = 1, 2, ..., m$, and $k \neq i, j$, and after calculating the $(m - 2)$-th order of the partial correlation coefficient, the algorithm naturally stops. However, the algorithm does not usually request the $(m - 2)$-th order of the partial correlation coefficient for the natural stop. This is because no adjacent variables will be found after excluding the variables, even in the calculation of the lower order of the partial correlation coefficient. We provide the pseudocode of the algorithm in Figure 2.

In the sample data, the zero-th order (i.e., the condition where subset $S$ is empty) of the partial correlation coefficient is calculated by Pearson's correlation coefficient, $r_{ij|S = \phi}$, expressed by

$$r_{ij|S=\varphi} = \frac{\mathrm{cov}(X_i, X_j)}{\sqrt{\mathrm{var}(X_i)\ \mathrm{var}(X_j)}}$$

where $\mathrm{cov}(X_i, X_j)$ and $\mathrm{var}(X_i)$ are the covariance between $X_i$ and $X_j$ and the variance of $X_i$. The higher order of the partial correlation coefficients, $r_{ij|S}$, expressed by

$$r_{ij|S} = \frac{-r^{ij}}{\sqrt{r^{ii} \cdot r^{jj}}}$$

where $ij|S$ means $S=\{1, 2, ..., p\}\backslash\{i,j\}$, and $r^{ij}$ is the $i$-$j$ element of the inverse correlation coefficient matrix.[13] Note that the dimensions of the correlation coefficient matrix are related to the orders of the partial correlation coefficients. The $m$-th order partial correlation coefficient is calculated from the $(m + 2)$ dimension of the correlation coefficient matrix. The partial correlation coefficient is statistically tested by using the $Z$-statistic.[16] First, $z$-transforms of the partial correlation coefficients are calculated, by the following equation:

$$z_{ij} = \frac{1}{2}\ln\left(\frac{1 + |r_{ij|S}|}{1 - |r_{ij|S}|}\right)$$

Then, the $z$-statistic is obtained from the following equation:

$$Z = \frac{z_{ij}}{\sqrt{1/n - 3 - p}}$$

where $n$ is the number of samples and $p = |S|$ is the conditioning order of the partial correlation coefficient. The $z$-statistic follows the standard normal distribution, $N(0,1)$, and the significance probability can be set according to this distribution; i.e., we reject the null hypothesis $H_0$:$r_{ij|s} = 0$, if $Z > Z_{\alpha/2}$with significance level $\alpha$. If $H_0$ is not rejected, then

```
Let $Adj(G,X_i) \setminus \{X_j\}$ be the set of nodes (variables) adjacent to $X_i$, except for $X_j$, in the undirected graph $G$.

Let $p$ be the degree of conditioning.

1: $G \leftarrow$ complete undirected graph
2: $p = 0$
3: repeat
4:     for all $X_i$ such that $|Adj(G,X_i)| - 1 \geq p$ do
5:         for all $X_j \in Adj(G,X_i)$ do
6:             for all subset $S \subseteq Adj(G,X_i) \setminus \{X_j\}$ such that $|S| = p$ do
7:                 if $X_i$ _||_ $X_j$ | $S$ then
8:                     delete edge between $X_i$ and $X_j$ in $G$
9:                 end if
10:            end for
11:        end for
12:    end for
13:    $p = p + 1$
14: until $|Adj(G,X_i)| - 1 \leq p, \forall X$
15: return $G$

Where "$X_i$_||_$X_j$ | $S$" means $X_i$ and $X_j$ are conditionally independent on $S$; i.e., there is no edge between $X_i$ and $X_j$.
```

**Figure 2.** Pseudocode of the modified path consistency algorithm. A pseudocode of the modified PC algorithm is described. In line seven, statistical hypothesis testing for the partial correlation between $X_i$ and $X_j$ conditioning on $S$ is used to determine whether $X_i$ and $X_j$ are conditionally independent (for details, see text). If the partial correlation cannot be calculated, due to the multicollinearity, then we consider that $X_i$ and $X_j$ are always conditionally dependent on any other variables.

we consider $r_{ij|s} = 0$, and we judge the i-th and j-th nodes as being conditionally independent of $S$.

The key point in the present network inference is the two modifications of the original PC algorithm, for application to the chemical compounds. The first modification is the correction of the algorithm in the calculation of the partial correlation coefficient. Since many compounds frequently show very similar descriptor values, the difficulty increases in the numerical calculation of the partial correlation coefficients, due to the multicolinearity between the variables. The original PC algorithm accidentally stops if only one partial correlation between a pair of variables violates the numerical calculation, against the high similarity of the descriptors. To avoid the accidental stops by the highly associated compound pairs, the original PC algorithm is modified as follows: If the calculation of any order of the partial correlation coefficient between the variables is violated, then the corresponding pair of variables is regarded as being dependent. The second modification is the correction of the output by the algorithm. The network inference outputs the edges with positive and negative correlations. The edge with a positive correlation in the network can be interpreted as a relationship with direct similarity between the properties of the chemical compound structures, while the edge with a negative correlation indicates a relationship with dissimilarity in a linear fashion. Thus, the edges with the positive correlation are adopted, and those with the negative correlation are excluded from the inferred network.

**2.5. Grouping of Chemical Compounds by Network Clustering.** In networks, the vertices are often clustered into tightly knit groups, with a high density of within-group edges and a lower density of between-group edges. This property is called a "community structure", and the computer algorithms for identifying the community structure are based on the iterative removal of edges with high "betweenness" scores, which identify such structures with some sensitivity. Here, we applied one of these algorithms to group the chemical compounds in the inferred network.[14]

This method is based on the modularity that is measured by a parameter, the $Q$-value. The $Q$-value is defined as follows:

$$Q = \sum (e_{ii} - a_i^2)$$

where $e_{ii}$ means the fraction of edges in cluster $i$ with respect to all edges in the network, and $a_i$ means the fraction of the number of edges that end in cluster $i$. First, this method considers each node as a cluster. In each subsequent step, two clusters are combined to maximize the increment of the $Q$-value, $\Delta Q$. $\Delta Q$ is calculated as follows:

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j)$$

where $e_{ij}$ means the half of the fraction of edges between clusters $i$ and $j$ with respect to all edges in the network. In addition to the above definition, $e_{ij}$ is commutative, $e_{ij} = e_{ji}$, in the undirected graph. The complexity of the calculation is on the order $O(N)$, where $N$ is the number of nodes in the network, and we combine two clusters at most $(N - 1)$ times; therefore, in sparse networks, the clustering is complete after $O(N^2)$ times.

**2.6. Maximum Common Structures of Clusters.** The maximum common structure within the constituent compounds belonging one cluster was obtained by using ChemAxon JKlustor libMCS.[17]

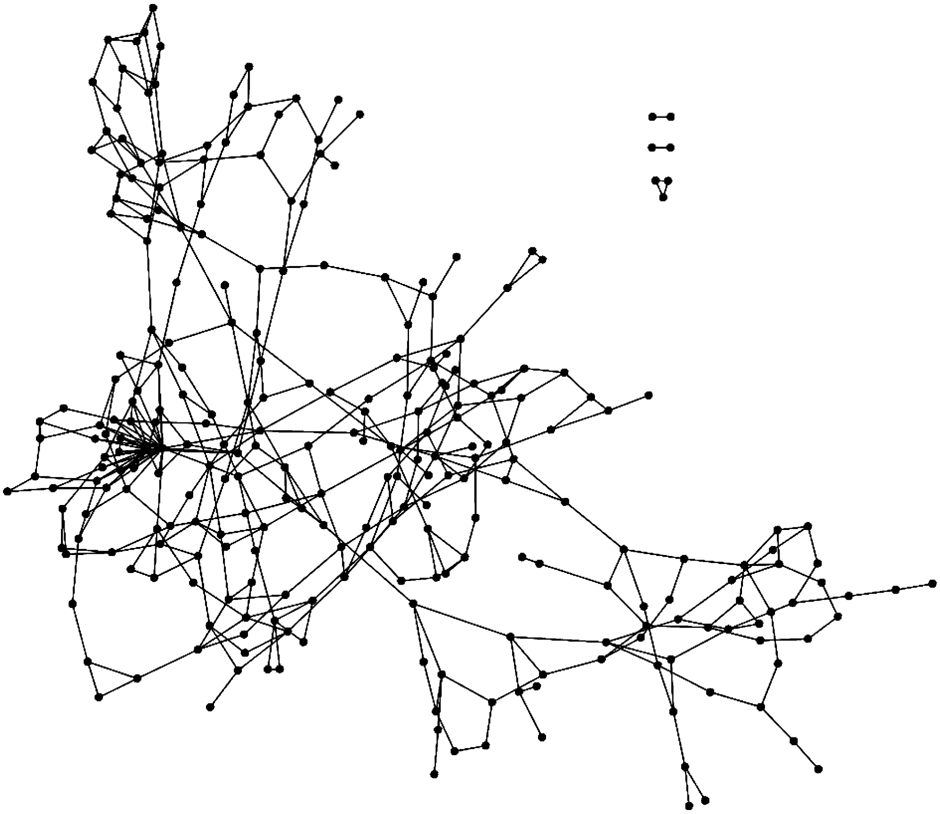


**Figure 3.** Chemical compound network inferred by path consistency algorithm. A large network of 266 compounds, inferred by the path consistency algorithm with 5% significance probability,[12] is described. The compounds and the established edges between compounds are denoted by open circles and straight lines, respectively.

## 3. RESULTS AND DISCUSSION

**3.1. Chemical Compound Network.** The relationships between the 279 active compounds were inferred by the PC algorithm. By the network inference, a large network containing 266 of the 279 active compounds emerged, as shown in Figure 3. Only seven compounds remained apart from the large network, and among them, five edges of the seven compounds were established. The emergence of a large network seems natural, because all of the compounds analyzed in this study share similar physicochemical properties, in terms of drug activity.

The large network contained 408 edges between compounds, and the average connectivity ([number of edges]/[$n(n - 1)/2$], where $n$ is the number of nodes) was about 0.0116. As shown in Figure 3, the inferred network was relatively sparse, in terms of edge connectivity. Although several hubs were observed in the network, it seems difficult to identify clear relationships between the compounds by visual inspection.

**3.2. Chemical Compound Network Clusters.** To scrutinize the compound relationships, we applied a network clustering method to rationally rearrange the connectivity in the inferred network of Figure 3. In Figure 4, 10 clusters naturally emerged from the entire connectivity in the inferred large network. Thus, the large, complicated network was transformed into distinctive clusters, with the number of compounds in each cluster ranging from 14 to 41. The emergence of the clusters indicates that some distinctive compound groups with similar structural properties exist in the network. The following step involves the investigation of the constituent compounds of each cluster that emerged by two network analyses, in terms of chemical structure and activity.

**3.3. Common Structures of Chemical Compound Network Clusters.** We surveyed the structural relationship between the constituent chemical compounds that belong to each cluster in the active network. Interestingly, the structures of the constituent compounds were common within each cluster, and they were diverse between the clusters.

The common structures of the member compounds in the clusters of the active network are shown in Figure 5A. It is readily apparent that common structures were found for all of the clusters, and high densities of the constituent compound structures were present in all of the clusters. Indeed, on average, ca. 63.7% of the compounds shared common structures: the highest and lowest share rates were 100.0% in cluster 6 and 35.5% in cluster 3. In addition, the average density of heavy atoms over all constituent compounds in each cluster was high: 9 of the 10 clusters showed more than 50% of the average density, and the exceptional cases were found in cluster 8. Furthermore, the common structures of each cluster were distinctive between them, as seen in Figure 5A. To estimate the differences between the common structures, the Tanimoto coefficients were calculated between them, as follows:

$$T_{ij} = \frac{\sum_k (X_{ik}X_{jk})}{\sum_k X_{ik}^2 + \sum_k X_{jk}^2 - \sum_k (X_{ik}X_{jk})}$$

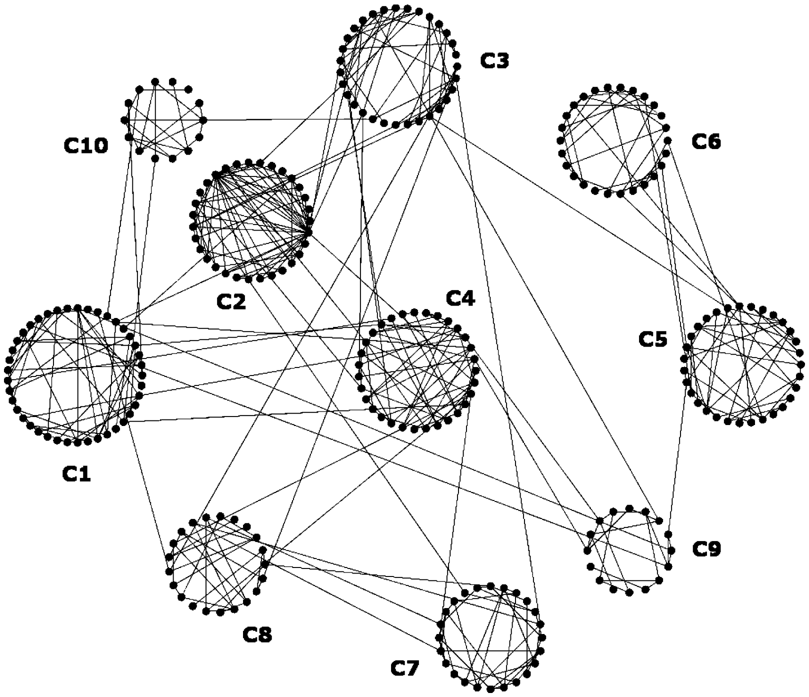


**Figure 4.** Network clusters of a large network of active chemical compounds. The network clusters estimated for the large network of active compounds in Figure 3 are depicted. Ten clusters emerged, and they are numbered in the order of the numbers of constituent compounds within the clusters.

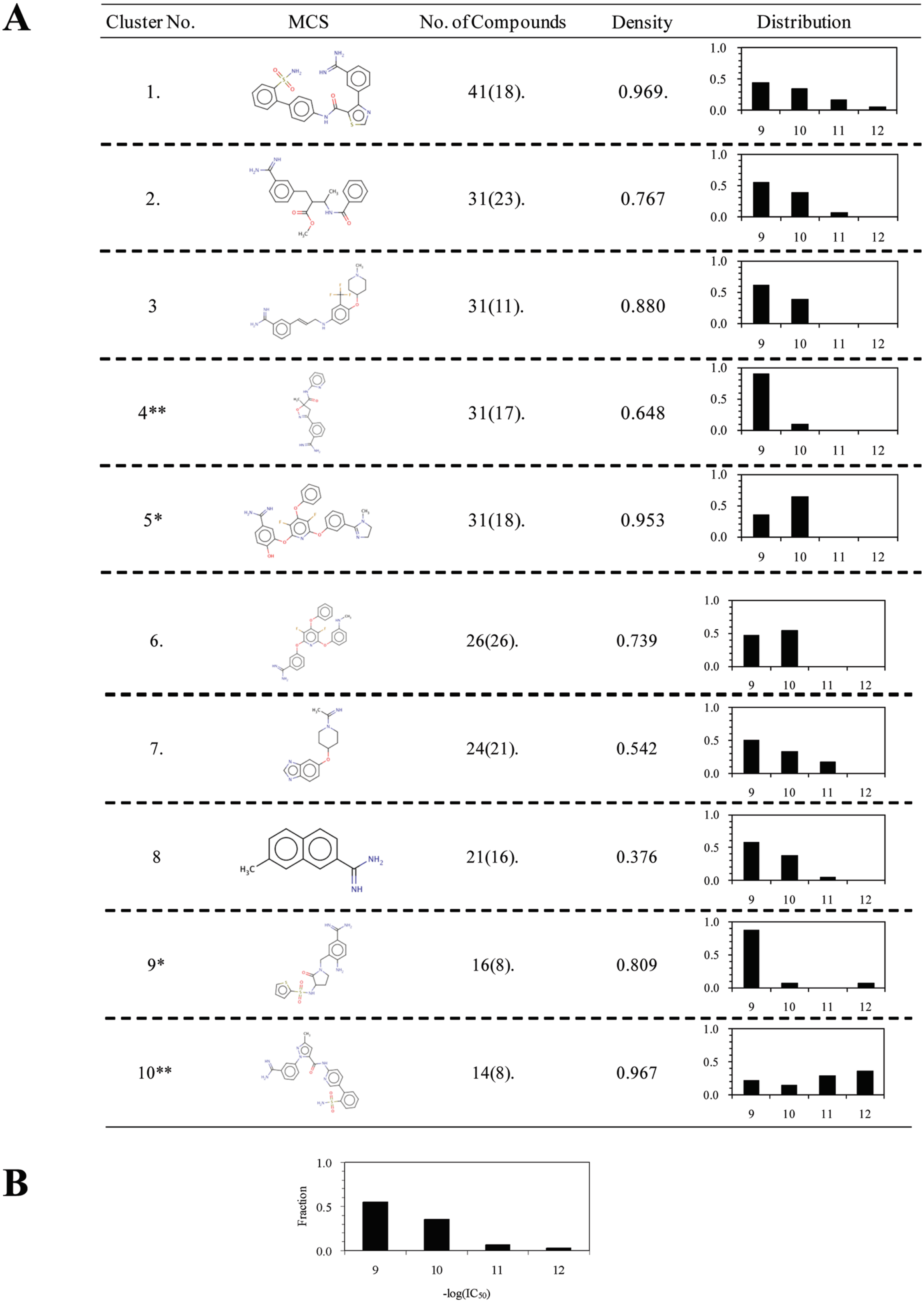
As shown in Table 1, the Tanimoto coefficients for all pairs of common structures were much less than 0.85, a value that is generally considered to reflect similarity to each other. All of the coefficients were less than 0.4, except for only 0.688 between the common structures of clusters 5 and 6.

The structures common within clusters and diverse between clusters were further investigated in terms of the activity distribution, expressed by the $-\log(IC_{50})$ histogram of the constituent compounds. For reference, the histogram of all compounds was also drawn in Figure 5B, and in the histogram, the compounds with a $-\log(IC_{50})$ value of less than 9 (10 nM), which is generally regarded as the lead compound, were frequently included (29.3% of compounds). Subsequently, the compounds with $IC_{50}$ values less than 10 nM were frequently observed in the histograms of each cluster in Figure 5A. Interestingly, some exceptions were also observed. A statistical difference between the total $IC_{50}$ distribution in Figure 5B and the distributions in Figure 5A was found in several clusters. In the distributions of clusters 5 and 10, the frequency of observing an $IC_{50}$ value than 10nM was relatively high, in comparison with the total distribution. This indicates that the common structures in clusters 5 and 10 may show a robust $IC_{50}$ for any chemical modification. Thus, the common structure may be a candidate for lead optimization. In contrast, the frequencies of $IC_{50}$ values less than 10 nM in clusters 4 and 9 were much lower than that in the total $IC_{50}$ distribution. This indicates the possibility that many compounds with an $IC_{50}$ activity of less than 10 nM can be synthesized from the common structures of the two clusters. Thus, the correspondence between the common structures and the $IC_{50}$ distributions of each cluster provides some clues for the synthesis of new compounds in the lead optimization process.

**3.4. Related Methods.** For comparison with the performance of the present method, the PCA was performed for the same data. Figure 6 shows the projection of the cluster members of the active network in Figure 4 into the principal component space. As easily seen in the figure, the cluster members with each common structure are scattered in the space. Indeed, the constituent compounds in each cluster were projected into some duplicated spaces, while the compounds of clusters 5 and 6 were relatively separated from the other clusters in the projected space. As indicated in the preceding subsection, the Tanimoto coefficient between the common structures of clusters 5 and 6 was exceptionally large, and this similarity reflects the common configuration of the constituent compounds in the two clusters in the principal component space. In contrast, the Tanimoto coefficients between the common structures of the other clusters were small, and therefore the compounds were not clearly discriminated in the space. Thus, the PCA may be a low-resolution method to clearly detect the groups with common chemical structures in the data, followed by the first screening.

The fingerprint approach is well-known as another method to detect common structures in an ensemble of compounds.[9] Actually, we used this approach to identify the common structure of the constituent compounds in the respective clusters. As a trial, we applied the fingerprint approach to all of the compounds but were unable to find the common structure (data not shown). We expected this failure from the fact that the common structures of each cluster show much less similarity, as depicted in Figure 5A. In contrast to the PCA, therefore, the fingerprint method may be a high-resolution technique to detect the distinctive groups with common chemical structures.

Note that there are two reasons why we use the Newman method, instead of the standard hierarchical clustering by using the partial correlation matrix as a distance measure. One reason is that the edges in the inferred network are established by considering the higher order of correlation between multiple variables, instead of the distance between



| Cluster No. | MCS | No. of Compounds | Density | Distribution |
|---|---|---|---|---|
| 1. | | 41(18). | 0.969. | |
| 2. | | 31(23). | 0.767 | |
| 3 | | 31(11). | 0.880 | |
| 4** | | 31(17). | 0.648 | |
| 5* | | 31(18). | 0.953 | |
| 6. | | 26(26). | 0.739 | |
| 7. | | 24(21). | 0.542 | |
| 8 | | 21(16). | 0.376 | |
| 9* | | 16(8). | 0.809 | |
| 10** | | 14(8). | 0.967 | |

**Figure 5.** Maximum common structures of the active compound network clusters, together with $IC_{50}$ histograms of constituent compounds. (A) The numbers of clusters in the first column are those described in Figure 4. The common structures of the 10 clusters in the second column were extracted by using ChemAxon JKlustor libMCS.[17] The total number of constituent compounds in each cluster is denoted in the third column, and the number of compounds sharing the corresponding common structures is also denoted in parentheses. In the fourth column, the average densities of heavy atoms in the common structures over the structures of all compounds are denoted. In the fifth column, the histograms of the $IC_{50}$ values of the constituent compounds are depicted: the vertical and horizontal axes are the frequency of the compounds and the $-\log(IC_{50})$ values, respectively. In addition, the differences between each histogram of $IC_{50}$ values for the respective clusters and that for the total active compounds (B) were tested by Fisher's exact test. The significance of the differences between the histograms is indicated at the cluster number in the first column: 5%, '**'; and 10%, '*'.

**Table 1.** Tanimoto Coefficients between Maximum Common Structures in Respective Active Compound Clusters

| cluster no. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | – | | | | | | | | | |
| 2 | 0.357 | – | | | | | | | | |
| 3 | 0.243 | 0.254 | – | | | | | | | |
| 4 | 0.261 | 0.304 | 0.244 | – | | | | | | |
| 5 | 0.179 | 0.197 | 0.199 | 0.202 | – | | | | | |
| 6 | 0.160 | 0.166 | 0.226 | 0.232 | 0.686 | – | | | | |
| 7 | 0.187 | 0.193 | 0.192 | 0.123 | 0.132 | 0.124 | – | | | |
| 8 | 0.137 | 0.224 | 0.176 | 0.152 | 0.126 | 0.150 | 0.073 | – | | |
| 9 | 0.254 | 0.271 | 0.234 | 0.229 | 0.213 | 0.198 | 0.151 | 0.157 | – | |
| 10 | 0.213 | 0.165 | 0.228 | 0.274 | 0.222 | 0.246 | 0.142 | 0.095 | 0.246 | – |

pairs of variables in the clustering, The clustering technique in the present study is therefore suitable for keeping the inferred relationships between variables. The other reason is that the Newman method can automatically determine the number of clusters in terms of the network structure. In contrast, the number of clusters is determined by setting a threshold, as in hierarchical clustering, or the cluster number is done before the clustering, as in a self-organization map (SOM).

In summary, the PCA provides a coarse-grinning relationship between compounds from the macroscopic resources, and the fingerprint approach provides a fine relationship between limited ensembles of compounds. With these situations in mind, our procedure provides a medium relationship between compounds, to enrich the selection of molecules with a desired activity. Thus, it bridges the gap between the two methods, by finding the groups of common structures in the step after the first screening, during the process of the lead optimization.

**3.5. Merits and Pitfalls of the Present Method.** One of the merits of the present method is that it simply detects the structural similarity relationships between active compounds. Indeed, only one parameter, the significance probability in the path consistency algorithm, is set in the network analyses. Thus, the present method is highly automatic and visual, to help reveal a rational synthesis route of chemical compounds for new drug discovery.

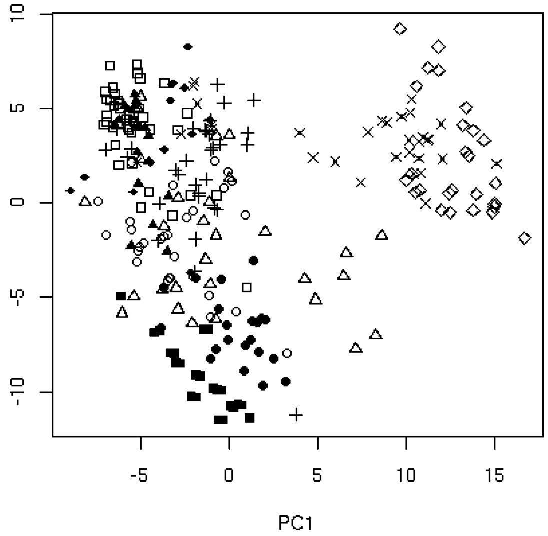


**Figure 6.** Distribution of members of network clusters in principle component space. The 266 active compounds in the network of Figure 3, which were characterized by the same number of descriptors (158 descriptors) as in the present analysis, were subjected to the PCA. The inertias of the first and second principal components (PC1 and PC2 in the figure) were 0.309 and 0.198, respectively. The constituent compounds of the 10 clusters in Figure 4 are indicated by the following symbols: cluster 1, □; 2, ○; 3, Δ; 4, +; 5, ×; 6, ◇; 7, ■; 8, ●; 9, ▲; and 10, ◆.

One of the key points of our method is the application of network inference, based on the graphical model, to the chemical compounds. Among the similar chemical structures, the present network inference detects the 'well-balanced' similarity, by using the partial correlation coefficient. In general, the graphical model distinguishes between real correlation and pseudocorrelation, based on the calculation of a partial correlation coefficient that realizes the concept of conditional independence.[13] The merit of this graphical model is that it only establishes the connection between the compounds with common structures and not between those lacking common structures. This discriminative ability is useful for classifying a large number of active compounds into various groups with different common structures in a rational manner.

In the present analysis, one large network was inferred, and 10 clusters emerged. The numbers of networks and clusters naturally depend on the user-defined descriptors and one parameter in the network inference. In the present analysis, the chemical compounds were characterized by as many secondary structure descriptors as possible. In general, the kinds of descriptors in the analysis may be changed, according to the analyzed data and the analysis aim. Fortunately, the quantification of chemical compounds by descriptors can be easily and quickly performed, due to recent advances in high-performance computing. Although the heuristic choice of descriptors is important to characterize the compound set, the descriptor optimization responsible for the compound set can be included as a preprocessing step in the present work. Furthermore, the size of the network and the following cluster numbers can be controlled by the user-defined significance probability in the network inference. For example, if one chooses a more significant probability than that of the present study, then a smaller network and fewer clusters will be obtained, in which more similar common structures will be found. In addition, the computational time for the present data in the two network analyses was about 5 s, using a personal computer (one CPU with a 2.4 GHz Pentium IV processor and 1GB of memory, under the Linux system). At any rate, the easy manipulation of the data, using only one user-defined parameter, may promote the use of the present method in applications to discriminate between various active compounds in drug discovery.

## 4. CONCLUSIONS

We have proposed a novel method to group active chemical compounds, by first screening with a combination of two network analysis methods. The scrutinization of active inhibitors of factor Xa by our method revealed reasonable grouping in terms of chemical structure and significant differences between each group in terms of activity. The present results illustrate the possibility that our method will bridge the gap between the compound activity test by the first screening and the following synthesis of lead derivatives.

## ACKNOWLEDGMENT

K.H. was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas "ystems Genomics" (grant

20016028) and by a Grant-in-Aid for Scientific Research (A) (grant 19201039), from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

## REFERENCES AND NOTES

(1) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861.

(2) Todeschini, R.; Consonni, V. The Handbook of Molecular Descriptors, in the Series of Methods and Principles in Medicinal Chemistry; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley-VCH: New York, 2000; Vol. 11.

(3) Katritzky, A. R.; Gordeeva, E. V. Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835–857.

(4) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027–1043.

(5) Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach: Amsterdam, The Netherlands, 1999.

(6) Bajorath, J. Integration of virtual and high throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.

(7) Bajorath, J. Virtual screening: methods, expectations, and reality. *Curr. Drug Discovery* **2002**, *2*, 24–28.

(8) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods. *Drug Discovery Today* **2002**, *7*, 903–911.

(9) Hert, J.; Willett, P.; Wilton, D. J. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(10) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.

(11) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824–828.

(12) Spirtes, P.; Glymour, C.; Scheines, R. *Causation, Prediction, and Search (Springer Lecture Notes in Statistics)*, 2nd ed., revised; MIT Press, Cambridge, MA, 2001.

(13) Whittaker, J. *Graphical Models in Applied Multivariate Statistics*; Wiley: New York, 1990.

(14) Newman, M. E. J. Fast algorithm for detecting community structure in networks. *Phys. Rev. E.* **2004**, *69*, 066133.

(15) Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: Filling the Gap between Standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005**, *48*, 2687–2694.

(16) Sokal, R. R.; Rohlf, F. J. *Biometry: The Principles and Practices of Statistics in Biological Research*, 3rd ed.; W. H. Freeman: 1994.

(17) *JChem JKlustor LibMCS*, version 5.3.8; ChemAxon: Budapest, Hungary, 2010.

CI100262S