Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 231780, 8 pages
doi:10.1155/2009/231780

Research Article

A Similarity Search Using Molecular Topological Graphs

Yoshifumi Fukunishi[1] and Haruki Nakamura[1, 2]

[1] *Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan*
[2] *Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan*

Correspondence should be addressed to Yoshifumi Fukunishi, y-fukunishi@aist.go.jp

Received 26 February 2009; Revised 26 July 2009; Accepted 19 September 2009

Recommended by George Karypis

A molecular similarity measure has been developed using molecular topological graphs and atomic partial charges. Two kinds of topological graphs were used. One is the ordinary adjacency matrix and the other is a matrix which represents the minimum path length between two atoms of the molecule. The ordinary adjacency matrix is suitable to compare the local structures of molecules such as functional groups, and the other matrix is suitable to compare the global structures of molecules. The combination of these two matrices gave a similarity measure. This method was applied to *in silico* drug screening, and the results showed that it was effective as a similarity measure.

1. Introduction

The molecular similarity measure is an important tool in chemometrics and chemoinformatics. The main applications include ligand-based *in silico* (virtual) drug screening, ADME-Tox (adsorption, distribution, metabolism, excretion, and toxicity) property prediction, physical molecular property prediction (1-octanol-water partition coefficient, solubility), and measurement of the diversity of chemical compounds in a library.

The molecular similarity measure generally assigns a one-dimensional (1D) and/or two-dimensional (2D) descriptor—that is, molecular fingerprints based on substructure, molecular mass, number of rotatable bonds, number of hydrogen donors/acceptors of the compound, and so forth—to compounds so that the similarities of the compounds can be evaluated [1–5]. Many methods have been proposed for the similarity search of chemical compounds, such as the comparison of overlapping substructures in the form of Daylight fingerprints (Daylight Chemical Information Systems Inc., Aliso Viejo, CA, USA), the chemically advanced template search (CATS) descriptor method developed by Pickett [6], and the Burden-CAS-University of Texas (BCUT) descriptor method [7]. One of the most widely used methods is to compare the existence of fragment structures; this is the technique employed by the MACSS key, which was developed by Molecular Design Limited (MDL, Santa Clara, CA, USA). Each element of the feature vector of the molecule represents the existence of a particular fragment structure in the molecule (dictionary based fingerprinting). A rather large example of a dictionary used for this fingerprinting technique is the software program DRAGON developed by Talete SRL (geographical information), which consists of more than 3200 molecular descriptors. The affinity fingerprint approach is a new type of similarity search method based on a multiprotein/multicompound affinity matrix [8–21]. In this method, each element of the feature vector of the molecule represents the binding affinity of the molecule with a particular protein. Usually, the binding affinity is measured by calculation using a protein-compound docking program.

There are various applications for molecular similarity, and thus many types of similarity measures are needed. Most of the conventional molecular descriptors aim scaffold hopping (lead hopping) to find a compound with a different scaffold from the known active compound. However, in some cases, we want to find similar compounds with similar scaffolds. For example, in lead generation, we want to find a series of similar compounds with the same or similar scaffold instead of performing an actual synthesis. Substructure searches have been used for this purpose [22, 23]. However, a comparison of the indices of molecular topologies is much faster than a substructure search. A topological index, which is any of several numerical parameters of a molecular graph, is also widely used [5]. The Wiener index, Hosoya index, and Randic's molecular connectivity index, are graph invariants and conventional topological indices. These topological indices show correlation to the physical or chemical properties of molecules, although these indices do not recognize atom types and they can be quite difficult to calculate.

In the current study, we proposed a new similarity measure for identifying topologically similar compounds based on their molecular topologies and evaluated this method by applying it to a ligand-based drug screening test.

2. Methods

2.1. Similarity and Distance between Compounds. First, the all-atom model compound structures are converted to united atom models, in which all hydrogen atoms are omitted and the atomic charge of the hydrogen atom is added to the atomic charge of the connected heavy atom.

In this method, the adjacency matrix E and the distance matrix D are used, and Figure 1 shows an example of these two matrices for a simple graph. The topology of the compound can be represented by an edge-adjacency matrix E [4, 5]:

$$\begin{aligned} e_{ab} &= 1, \quad \text{when the } a\text{th and the } b\text{th atoms are connected} \\ e_{ab} &= 0, \quad \text{otherwise,} \end{aligned} \tag{1}$$

where e_{ab} is the a-b element of matrix E. The value of e_{ab} could be the bond order between the ath and the bth atoms (the value of e_{ab} could be 1.5 for an aromatic bond). Just as in the BCUT method, the diagonal part (e_{aa}) is replaced by the converted atomic charge q_c:

$$q_c = \frac{1}{1 + \exp(-c \cdot q_A)}, \tag{2}$$

where q_A is an atomic partial charge and c is a coefficient. In this study, c was set to 1.0. The q_c value is > 0 for any q_A value.

The a-b matrix element of the pseudodistance matrix D represents the minimum path length between the ath and bth atoms:

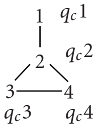
$$D_{ab} = \ln(1 + d_{ab}), \tag{3}$$

where d_{ab} is a shortest path length between the ath and the bth atoms. We also tried $D_{ab} = d_{ab}$ (shortest-path topological distance matrix [4, 5]) and $D_{ab} = {d_{ab}}^{1/2}$ and found that $D_{ab} = \ln(1+d_{ab})$ gave the best result among these definitions.

Let ε_i and N be the ith eigenvalue of the matrix E or D and the number of atoms of the united-atom model of the compound. We define the eigenvalue histogram $g(\varepsilon)$ as follows:

$$g(\varepsilon) = \sum_{i=1}^{N} \exp\left(-c(\varepsilon_i - \varepsilon)^2\right). \tag{4}$$

Here ε and c are the energy and the arbitral coefficient.



$$E = \begin{pmatrix} q_c1 & 1 & 0 & 0 \\ 1 & q_c2 & 1 & 1 \\ 0 & 1 & q_c3 & 1 \\ 0 & 1 & 1 & q_c4 \end{pmatrix} \quad D = \begin{pmatrix} 0 & \ln 2 & \ln 3 & \ln 3 \\ \ln 2 & 0 & \ln 2 & \ln 2 \\ \ln 3 & \ln 2 & 0 & \ln 2 \\ \ln 3 & \ln 2 & \ln 2 & 0 \end{pmatrix}$$

Figure 1: Example of the matrices E and D.

The distance $S(A, B)$ between molecules A and B is defined based on the eigenvalue histogram of molecule $A(g_A(\varepsilon))$ and that of molecule $B(g_B(\varepsilon))$ as follows:

$$S(A, B) = \int_{-\infty}^{\infty} |g_A(\varepsilon) - g_B(\varepsilon)| \, d\varepsilon. \tag{5}$$

In the current *in silico* drug screening, candidate hit compounds are selected using the following method. Let $S^E(A, B)$ and $S^D(A, B)$ be the distance between A and B based on the adjacency matrix E and that based on the distance matrix D. These two distances give the consensus distance $S'(A, B)$ with the weight parameter λ:

$$S'(A, B) = \lambda S^E(A, B) + (1 - \lambda) S^D(A, B). \tag{6}$$

Compounds that are close to the known active compounds are selected as the candidate hit compounds. For this purpose, the distance to the known active compounds is introduced. The distance from the kth compound to the average position of the active compounds (Dist_k) is defined as

$$\mathrm{Dist}_k = \sqrt{\sum_{i=1}^{M} S'(A_k, C_i)^2 / M}, \tag{7}$$

where A_k, C_i, and M are the kth compound, ith active compounds, and the total number of the active compounds. When the number of active compounds is one, $\mathrm{Dist}_k = S'$. We call Dist_k the molecular-graph (MG) distance and we call this screening procedure the molecular-graph distance (MGD) method. The eigenvalues (ε_i) of S^E and S^D of each compound of the compound library are stored in a database file *a priori*. For a query compound, the eigenvalues of S^E and S^D must be calculated, which costs less than 1 second. The database search is conducted only to perform the calculations in (4)–(7) and thus is quite fast.

3. Preparation of Materials

For the drug screening test, our target proteins were the macrophage migration inhibitory factor (MIF), cyclooxygenase-2 (COX-2), human immunodeficiency virus protease-1 (HIV), thermolysin (THR), glutathione S-transferase (GST), the histamine H1 receptor, the adrenaline beta receptor, the serotonin receptor, and the dopamine D2 receptor. For validation of the present method, we used the same set of compounds as used in our previous study [20]. Namely, the compound set consisted of 12 inhibitors of MIF, 28 inhibitors of THR, 15 inhibitors of COX-2,

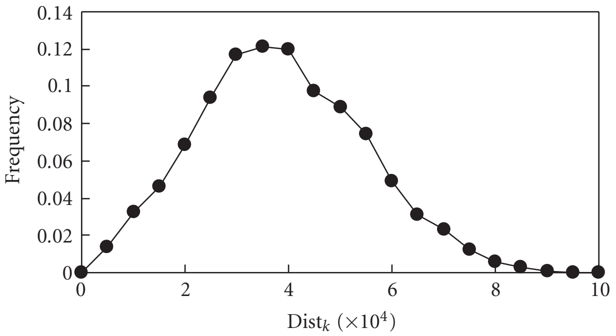


Figure 2: Distribution of the values of Dist$_k$. The template is diphenhydramine. The Dist$_k$ values are multiplied by 10000 times and the frequency is normalized.

20 inhibitors of HIV, 12 inhibitors of GST, 10 antagonists of the histamine H1 receptor [24], 12 agonists and 13 antagonists of the adrenaline beta receptor [25], 8 agonists and 9 antagonists of the serotonin receptor [26], and 6 agonists and 15 antagonists of the dopamine D2 receptor [27] as the active compounds, along with 11050 potentially negative compounds from the random compound library of the Coelacanth Chemical Corporation (East Windsor, NJ, USA). Typically, only one hit compound could be found out of 10^4 randomly selected compounds; we therefore expected that there would be no more than a few, if any, hit compounds among these 11212 compounds. The 160 active compounds are listed in the Supplemental Materials available online at doi:10.1155/2009/231780.

The size distribution of compounds was as follows: percentage of compounds with 0 ~ 19 atoms, 0.1%; with 20 ~ 29 atoms, 1.2%; with 30 ~ 39 atoms, 1.6%; with 40 ~ 49 atoms, 9.3%; with 50 ~ 59 atoms, 22.5%; with 60 ~ 69 atoms, 37.9%; with 70 ~ 79 atoms, 20.5%; and with more than 80 atoms, 7.0%. The average number of heavy atoms was 30.9.

The atomic charge of each ligand was determined by the Gasteiger method [28, 29]. To calculate the Gasteiger charge, an all-atom model is necessary for each compound. The three-dimensional (3D) coordinates of the 11050 random compounds above were generated to add the hydrogen atoms by the Concord program (Tripos, St. Louis, MO, USA) from 2D Sybyl SD files provided by the Coelacanth Chemical Corporation. The 3D coordinates of the active compounds (inhibitors, substrates, agonists, and antagonists) were generated by Chem3D (Cambridge Software, Cambridge, MA, USA).

4. Results

To evaluate the efficiency of this method, the leave-one-out cross-validation test was applied; namely, the active compounds of each target protein were selected one by one as the known active compounds for this software and the other unknown active compounds were discovered by the software. The test dataset consists of these active compounds and the other approximately 10^4 potential inactive compounds (decoy set). The total number of compounds was 11212. A total of 160 (= total 160 active compounds) database enrichment curves were calculated for these 9 target proteins and 11212 compounds and the results were averaged.

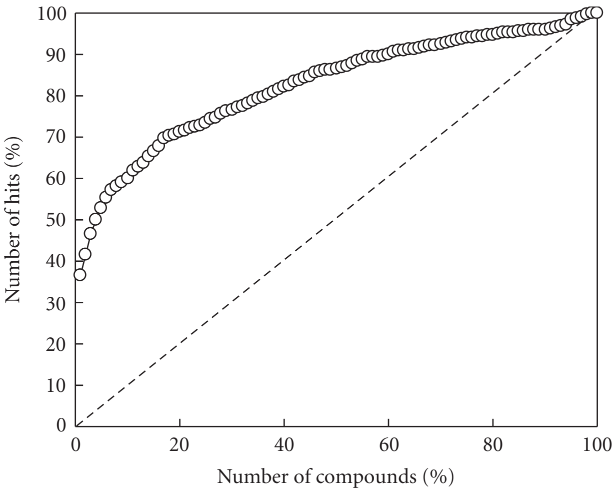


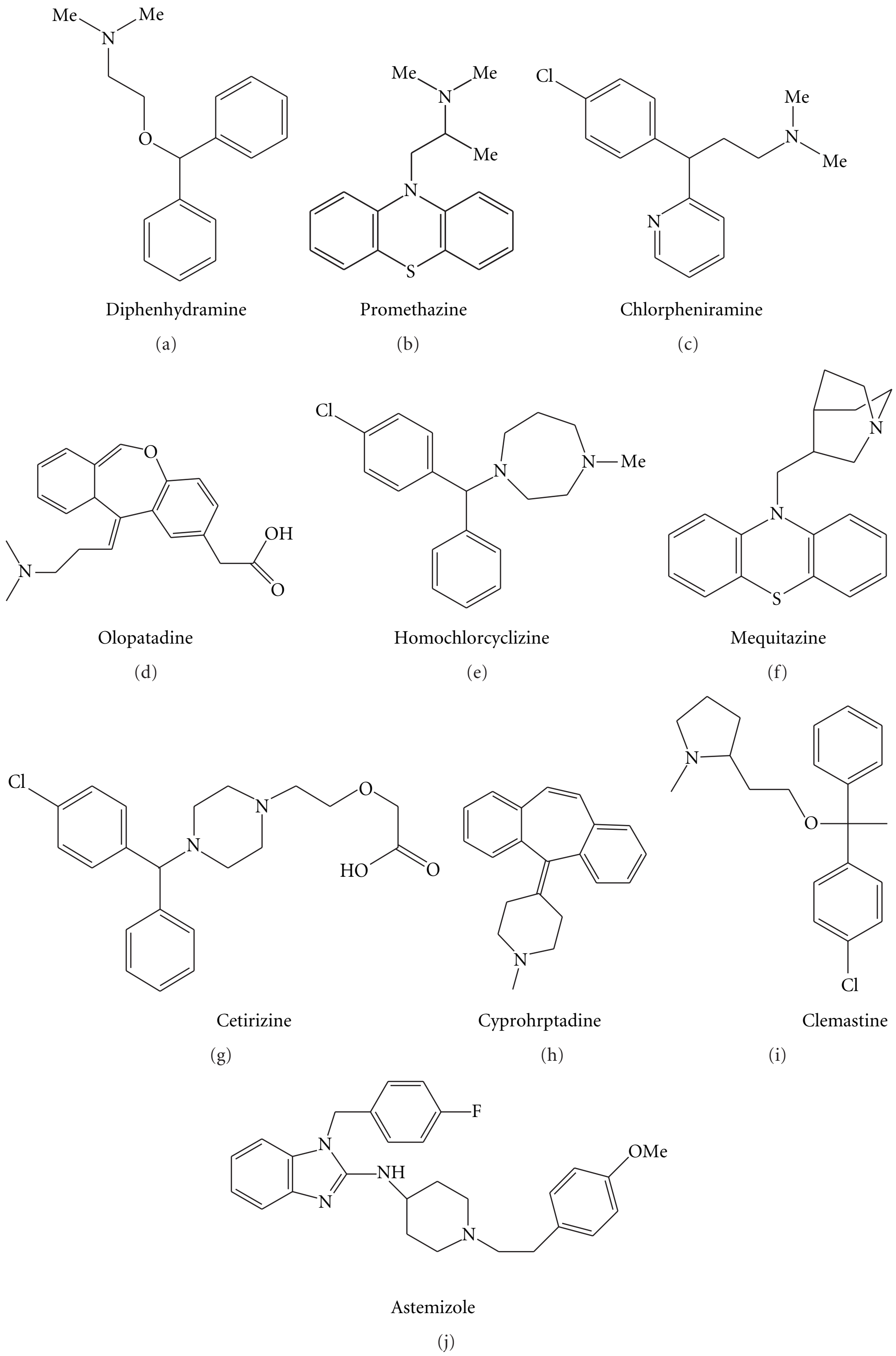
Figure 3: Average database enrichment curves of 160 active compounds with the current similarity measure. Open circles represent the average database enrichments with the distance measure defined by (6) (MGD) with $\lambda = 0.25$. The dashed line represents the random screening result.

The surface area (q or "area under curve": AUC) under the total database enrichment curve (f) is a measure of the database enrichment:

$$q = \int_0^{100} f(x)dx, \tag{8}$$

where x and $f(x)$ are the percentages of compounds that are selected from the total compound library and the database enrichment curve, respectively. A higher q value corresponds to better database enrichment, and the q value is always greater than zero and less than 100. For the random screening, $q = 50$.

First, the λ dependence of the hit ratio was examined in the MGD method. The average q values and the hit ratio of the 160 trials with various λ values are summarized in Table 1. The coefficients c for matrices E and D were optimized for every λ to maximize the hit ratio. When $\lambda = 0$ or $\lambda = 1$, the hit ratio and the q values were lower than those in the other cases. This result showed that the combination of matrices D and E is more effective than the single usage of either D or E. The optimized coefficients were used in the following study. The average eigenvalues of D and E were $-3.21 * 10^{-6}$ and 0.505 for the decoy set, respectively. The histograms $g(\varepsilon)$ of (4) were close to single Gaussian distributions. We show the distributions of $g(\varepsilon)$ of H1 antagonist diphenhydramine and COX-2 inhibitor indomethacin in the supplementary data. For diphenhydramine, the average $g(\varepsilon)$ values of S^E and S^D were $1.46*10^{-3}$ and $1.009*10^{-3}$, respectively. The deviations of $g(\varepsilon)$ values of S^E and S^D were 75.75 and 9.712, respectively. For indomethacin, the average $g(\varepsilon)$ values of S^E and S^D were



Diphenhydramine

(a)

Promethazine

(b)

Chlorpheniramine

(c)

Olopatadine

(d)

Homochlorcyclizine

(e)

Mequitazine

(f)

Cetirizine

(g)

Cyprohrptadine

(h)

Clemastine

(i)

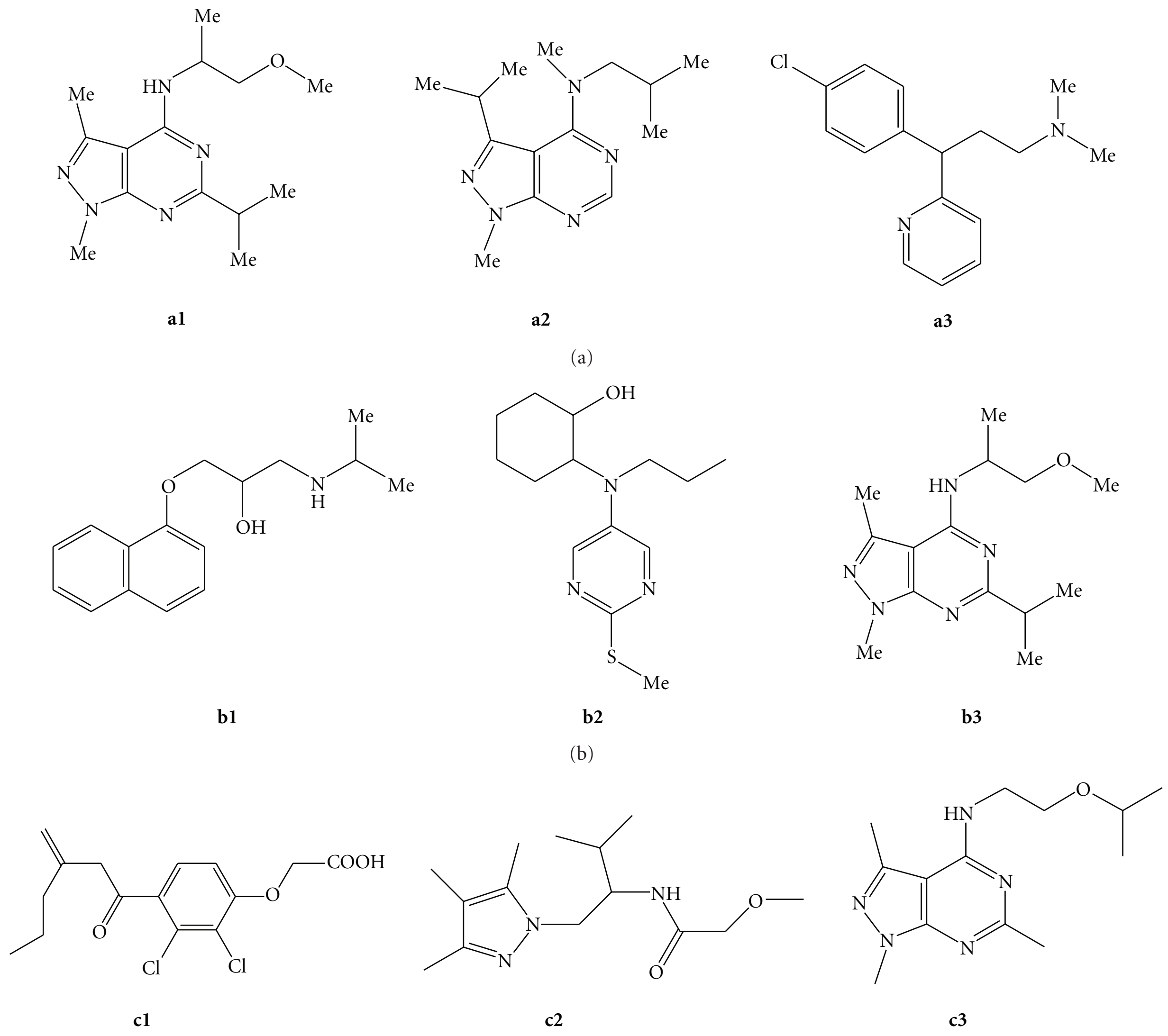
Astemizole

(j)

FIGURE 4: Histamine H1 receptor antagonists.

$1.46 * 10^{-3}$ and $1.009 * 10^{-3}$, respectively. The deviations of $g(\varepsilon)$ values of S^E and S^D were 75.88 and 9.739, respectively.

Second, the score distribution was examined by the MGD method. The average values and the standard deviations for the λ values are summarized in Table 1. Figure 2 shows a score distribution with $\lambda = 0.25$ using diphenhydramine as the template (see Figure 4). The template corresponds to $\text{Dist}_k = 0$. The frequency was normalized; the surface area under the curve was set to 1. The distribution is similar to Gaussian distribution.



(a)

(b)

(c)

Figure 5: The three top-ranked compounds by the MGD method for the template: diphenhydramine. (a) The result with $\lambda = 0.25$. (b) The result with $\lambda = 0$. (c) The result with $\lambda = 1$.

Figure 3 shows the average database enrichment results of the 160 trial screening tests by the leave-one-out cross-validation test. The MGD method worked well and showed good enrichment. In this calculation, λ was set to 0.25. Within the first 1%, 5%, and 10% of the database, 36.5%, 52.8%, and 60.0% of the active compounds were found by the similarity measure defined by (7), respectively. The average q value by the MGD method using (7) was 82.53. This result is worse than the result by the machine-learning docking score index method reported previously. Namely, about 70% of the active compounds were found within the first 1% of the database and the average q value was 98.5.

Ten histamine H1 receptor antagonists were included in the compound library (see Figure 4). Figure 5 shows the known active compound (template) and the best ranked molecules, when $\lambda = 0$, 0.25 and 1. The Dist_k values and z-scores (= (score – average score)/standard deviation) of the three top-ranked compounds are summarized in Table 1. These z-scores show that the score distribution is slightly different from the Gaussian distribution. In the Gaussian distribution, the number of compounds with a z-score > 3 is 0.1% of the database (10 compounds in this case). The z-scores of the top-ranked compounds were only 2, in this case. The selected molecules look similar to the template compound, diphenhydramine. For $\lambda = 0.25$, the other H1 receptor antagonist, chlorpheniramine, was found as the third compound. Both diphenhydramine and chlorpheniramine have a diphenyl group. The other compounds (**a1** and **a2**) are not very similar to the template. For $\lambda = 0$, compound **b2**, which is not an H1 antagonist, is similar to the template. Compound **b3** is identical to compound a1. For $\lambda = 1$, the three best-ranked compounds are not particularly similar to the query. Compounds **a1**, **a2**, **b3**, and **c3** are similar to each other.

Table 1: The average q values and the hit ratio with optimized coefficients for various λ values. The coefficient1 is the c parameter in (4) for the E matrix and the coefficient2 is the c parameter in (4) for the D matrix. The hit ratio is the % of the active compounds found within the first 1% of selected compounds of the database. The average and σ are the average value and the standard deviation of the scores. $\ast$: the top-ranked compounds when diphenhydramine is the template.

λ	0	0.25	0.5	0.75	1
Hit ratio	28.86%	36.48%	35.76%	35.69%	26.93%
q value	80.14	82.53	80.37	82.09	72.72
Coefficient1	0.002	0.01	0.02	0.01	0
Coeffcient2	0	0.00005	0.00005	0.00001	0.0001
Average	$0.100\ast10^{-4}$	$3.61\ast10^{-4}$	$7.13\ast10^{-4}$	$10.64\ast10^{-4}$	$14.163\ast10^{-4}$
σ	$7.50\ast10^{-6}$	$1.62\ast10^{-4}$	$3.24\ast10^{-4}$	$4.87\ast10^{-4}$	$6.50\ast10^{-4}$
1st compound*					
z-score	1.321	2.218	2.189	2.179	2.174
Score	0.00008295	0.001723	0.001944	0.002164	0.002383
2nd compound*					
z-score	1.315	2.202	2.179	2.171	2.167
Score	0.0001266	0.004425	0.005333	0.006242	0.00715
3rd compound*					
z-score	1.314	2.192	2.169	2.161	2.157
Score	0.0001338	0.005944	0.008402	0.01085	0.01331

5. Discussion

Figure 3 shows that 36.5% of the active compounds were found within the first 1% of the compounds of the whole library. Our previous study showed that 12.4%, 43.4%, and 67.5% of the active compounds were found within the first 1% of the database by the docking score index (DSI), factor-selection DSI (FS-DSI), and machine-learning DSI (ML-DSI) methods, respectively, when 180 proteins were used to calculate the affinity fingerprint [19]. The three-dimensional (3D) shape and charge distribution of a compound govern the protein-compound binding energy. The DSI, FS-DSI, and ML-DSI methods utilize the 3D shape and charge distribution of the compound through the affinity fingerprint. On the other hand, the 2D structure of the compound does not govern the protein-compound binding energy. Thus, the current similarity measure was not better than the previously developed screening methods for *in silico* drug screening, when it was used as a single measure to describe the molecular similarity. However, the MGD method did have an advantage in terms of its computational speed. The MGD method can search 10 000 000 compounds within 1 hour on a Xeon 3 GHz computer, which is 1000 times faster than the MSM-DSI method.

However, the current similarity measure was still effective for *in silico* drug screening. Our active compounds were chosen based on literature. As shown in Figure 4, some compounds were very similar to each other by human-eye inspection. The main reason for this similarity was likely that these compounds were generated from a progenitor compound by lead optimization. Diphenhydramine, chlorpheniramine, homochlorcyclizine, cetirizine, and clemastine have a diphenyl-like group. In promethazine, olopatadine, mequitazine, and cyprohrptadine, the conformations of two phenyl groups are fixed. Most of these antagonists are structurally similar, which should be the reason why the current similarity measure was effective for the *in silico* drug screening. In other words, this method is not suitable for scaffold hopping (lead hopping) [30]. For scaffold hopping, other methods have been developed [30, 31].

If all matrix elements of the off-diagonal part are zero, the eigenvalues are equal to the values of the diagonal part. The off-diagonal part shifts the eigenvalues from the values of the diagonal part. In (1), the bond information close to the ith atom can give major perturbation on the ith diagonal value (atomic charge of the ith atom). Thus the matrix E represents the short-range information of the molecular topology. On the other hand, in (3), the i-j matrix element of the matrix D becomes large when the ith atom is far from the jth atom on the molecular topology. Thus the matrix D can represent the long-range information of the molecular topology.

The information of the matrix E and that of D are independent of each other. For each compound in the test database of 10^4 compounds, the values of S^E and S^D in (6) were calculated. The correlation coefficient of the values of S^E and S^D was only 0.08, indicating that there was no correlation between these values. Thus, the compounds in the compound library are widely distributed in the (S^E, S^D) two-dimensional (2D) space, and the MGD method selects the compounds around the query compound from the compound library in the 2D space.

6. Conclusion

We developed a similarity measure for chemical compounds that is based on the molecular topology, the atomic charge, and the minimum path length between atoms. The histograms of eigenvalues of these matrices were smoothed to generate a continuous curve. The difference and overlap between these two histograms define the distance between the two compounds.

This similarity measure was applied to ligand-based *in silico* drug screening. In this calculation, compounds whose molecular topology structures are similar to the given active compounds were selected by using this similarity measure. This measure actually worked to find unknown active compounds from a random compound library.

Acknowledgments

This work was supported by the New Energy and Industrial Technology Development Organization of Japan (NEDO), the Ministry of Economy, Trade, and Industry (METI) of Japan, and the Japan Biological Informatics Consortium (JBiC).

References

[1] H. van de Waterbeemd, B. Testa, and G. Folkers, Eds., *Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry*, Wiley-VCH, Weinheim, Germany, 1997.

[2] A. R. Leach, *Molecular Modelling: Principles and Applications*, Pearson Education, Edinburgh, UK, 2nd edition, 2001.

[3] W. G. Richards and D. D. Robinson, "Molecular similarity in Rational Drug Design," in *Rational Drug Design*, D. G. Truhlar, W. J. Howe, A. J. Hopfinger, J. Blaney, and R. A. Dammkoehler, Eds., pp. 39–49, Springer, New York, NY, USA, 1999.

[4] A. Varnek and A. Tropsha, Eds., *Chemoinformatics Approaches to Virtual Screening*, Royal Society of Chemistry, Cambridge, UK, 2008.

[5] J. Gasteiger and T. Engel, Eds., *Chemoinformatics: A Textbook*, Wiley-VCH, Weinheim, Germany, 2003.

[6] S. Pickett, "Pharmacophore fingerprints in Protein-ligand Interactions," in *Protein-Ligand Interactions: From Molecular Recognition to Drug Design*, H.-J. Böhm, G. Schneider, R. Mannhold, H. Kubinyi, and G. Folkers, Eds., Methods and Principles in Medicinal Chemistry, pp. 88–91, Wiley-VCH, Weinheim, Germany, 2003.

[7] R. S. Pearlman and K. M. Smith, "Metric validation and the receptor-relevant subspace concept," *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 1, pp. 28–35, 1999.

[8] L. M. Kauvar, D. L. Higgins, H. O. Villar, et al., "Predicting ligand binding to proteins by affinity fingerprinting," *Chemistry & Biology*, vol. 2, no. 2, pp. 107–118, 1995.

[9] H. Briem and I. D. Kuntz, "Molecular similarity based on DOCK-generated fingerprints," *Journal of Medicinal Chemistry*, vol. 39, no. 17, pp. 3401–3408, 1996.

[10] U. F. Lessel and H. Briem, "Flexsim-X: a method for the detection of molecules with similar biological activity," *Journal of Chemical Information and Modeling*, vol. 40, no. 2, pp. 246–253, 2000.

[11] H. Briem and U. F. Lessel, "In vitro and in silico affinity fingerprints: finding similarities beyond structural classes," *Perspectives in Drug Discovery and Design*, vol. 20, pp. 231–244, 2000.

[12] A. Weber, A. Teckentrup, and H. Briem, "Flexsim-R: a virtual affinity fingerprint descriptor to calculate similarities of functional groups," *Journal of Computer-Aided Molecular Design*, vol. 16, no. 12, pp. 903–916, 2002.

[13] N. Hsu, D. Cai, K. Damodaran, et al., "Novel cyclooxygenase-1 inhibitors discovered using affinity fingerprints," *Journal of Medicinal Chemistry*, vol. 47, no. 20, pp. 4875–4880, 2004.

[14] G. P. A. Vigers and J. P. Rizzi, "Multiple active site corrections for docking and virtual screening," *Journal of Medicinal Chemistry*, vol. 47, no. 1, pp. 80–89, 2004.

[15] Y. Fukunishi, Y. Mikami, and H. Nakamura, "Similarities among receptor pockets and among compounds: analysis and application to in silico ligand screening," *Journal of Molecular Graphics and Modelling*, vol. 24, no. 1, pp. 34–45, 2005.

[16] Y. Fukunishi, Y. Mikami, S. Kubota, and H. Nakamura, "Multiple target screening method for robust and accurate in silico ligand screening," *Journal of Molecular Graphics and Modelling*, vol. 25, no. 1, pp. 61–70, 2006.

[17] Y. Fukunishi, Y. Mikami, K. Takedomi, M. Yamimouehi, H. Shima, and H. Nakamura, "Classification of chemical compounds by protein-compound docking for use in designing a focused library," *Journal of Medicinal Chemistry*, vol. 49, no. 2, pp. 523–533, 2006.

[18] Y. Fukunishi, S. Kubota, C. Kanai, and H. Nakamura, "A virtual active compound produced from the negative image of a ligand-binding pocket, and its application to in-silico drug screening," *Journal of Computer-Aided Molecular Design*, vol. 20, no. 4, pp. 237–248, 2006.

[19] Y. Fukunishi, S. Kubota, and H. Nakamura, "Finding ligands for G protein-coupled receptors based on the protein-compound affinity matrix," *Journal of Molecular Graphics and Modelling*, vol. 25, no. 5, pp. 633–643, 2007.

[20] Y. Fukunishi, S. Kubota, and H. Nakamura, "Noise reduction method for molecular interaction energy: application to in silico drug screening and in silico target protein screening," *Journal of Chemical Information and Modeling*, vol. 46, no. 5, pp. 2071–2084, 2006.

[21] Y. Fukunishi, S. Hojo, and H. Nakamura, "An efficient in silico screening method based on the protein-compound affinity matrix and its application to the design of a focused library for cytochrome P450 (CYP) ligands," *Journal of Chemical Information and Modeling*, vol. 46, no. 6, pp. 2610–2622, 2006.

[22] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa, "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways," *Journal of the American Chemical Society*, vol. 125, no. 39, pp. 11853–11865, 2003.

[23] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa, "Heuristics for chemical compound matching," *Genome Informatics*, vol. 14, pp. 144–153, 2003.

[24] T. Watanabe and Y. Fukui, "Chapter 7," in *Saiboumaku no Jyuyoutai*, I. Takayanagi, Ed., pp. 121–131, Nanzandou, Tokyo, Japan, 1988.

[25] K. Koike and T. Nagatomo, "Chapter 6," in *Saiboumaku no Jyuyoutai*, I. Takayanagi, Ed., pp. 103–118, Nanzandou, Tokyo, Japan, 1998.

[26] M. Sasa and K. Ishihara, "Chapter 8," in *Saiboumaku no Jyuyoutai*, I. Takayanagi, Ed., pp. 135–147, Nanzandou, Tokyo, Japan, 1998.

[27] Y. Nakata and A. Inoue, "Chapter 10," in *Saiboumaku no Jyuyoutai*, I. Takayanagi, Ed., pp. 169–182, Nanzandou, Tokyo, Japan, 1998.

[28] J. Gasteiger and M. Marsili, "Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges," *Tetrahedron*, vol. 36, no. 22, pp. 3219–3228, 1980.

[29] J. Gasteiger and M. Marsili, "A new model for calculating atomic charges in molecules," *Tetrahedron Letters*, vol. 19, no. 34, pp. 3181–3184, 1978.

[30] G. Schneider, W. Neidhart, T. Giller, and G. Schmid, ""Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening," *Angewandte Chemie International Edition*, vol. 38, no. 19, pp. 2894–2896, 1999.

[31] M. M. Ahlstrom, M. Ridderstrom, K. Luthman, and I. Zamora, "Virtual screening and scaffold hopping based on GRID molecular interaction fields," *Journal of Chemical Information and Modeling*, vol. 45, no. 5, pp. 1313–1323, 2005.