A novel graph-based similarity measure for 2D chemical structures

Si Quang Le[1] **Tu Bao Ho**[1] **T.T Hang Phan**[2]
quang@jaist.ac.jp bao@jaist.ac.jp s0244@st.ube-k.ac.jp

[1] Japan Advanced Institute of Science and Technology, 923-1292, Ishikawa, Japan
[2] Ube National College of Technology, 755-8555, Yamaguchi, Japan

Abstract

In this paper, we propose a graph-based method to measure the similarity between chemical compounds described by 2D form. Our main idea is to measure the similarity between two compounds based on edges, nodes, and connectivity of their common subgraphs. We applied the proposed similarity measure in combination with a clustering method to more than eleven thousand compounds in the chemical compound database KEGG/LIGAND and discovered that compound clusters with highly similar structure compounds that share common names, take part in the same pathways, and have the same requirement of enzymes in reactions. Furthermore, we discovered the surprising sameness between pathway modules identified by clusters of similar structure compounds and that identified by genomic contexts, namely, operon structures of enzyme genes.

Keywords: chemical structures, graph similarity, clustering algorithms, genomic information.

1 Introduction

Determining the degree of similarity between chemical compounds (molecules) plays an important role in chemistry and, increasingly, biology, e.g. protein-ligand docking, database searching, the prediction of biological activity, reaction site modelling, and the interpretation of molecular spectra. Among molecule description methods, 2D description where each compound is presented as a graph of nodes (atoms) and edges (bonds) can be adequate for most purposes in practice [1, 2]. Thus, most real-life applications focus on this description.

The first approach to measure the similarity between molecules via the 2D structure description is fingerprint-based comparison. In this approach, a molecule is considered as a bit-string, each bit indicates the presence or absence of an atom or a predefined molecular substructure known as key descriptor or finger [3]. The similarity between two molecules is then determined by comparing their corresponding bit-strings [1, 2]. Also, the combination of numerical vector methods and fingerprint methods has been used as mathematical extension of bit-comparison methods [4, 5, 6, 7]. Although these methods are simple and easy in practical use, they contain some drawbacks in key descriptor selections - the heart of these methods [8, 9].

In the second approach, the similarity between compounds is determined by comparing directly their corresponding graphs. Current graph-based methods [10, 11, 12, 13, 14] measure the similarity between two graphs either by the maximum common subgraphs (MCS) [10, 11, 12] or the maximum common edge subgraphs (MCES) [13, 14]. The main drawback of these methods is that they measure the similarity between two graphs only by calculating the size (either number of nodes or edges) of their MCS or MCES regardless of its structure. However, in practice, both nodes and edges play the same important roles in compound structures, and meaningful substructures are those that are connected. In addition, since there are few types of atoms and bonds, MCS or MCES found are often large. This would mislead the measuring of similarity between compounds as large MCS or MCES is not guarantee of meaningful substructures. For instance, a large subgraph whose atoms are separated is of little meaning while a smaller connected subgraph may be much meaningful.

In this paper, we introduce an innovative graph-based similarity measure for compounds in 2D description. Our key idea to overcome the drawbacks of the above mentioned graph-based methods is to measure the similarity between two compounds presented by two graphs based on nodes, edges, and the connectivity of their common subgraphs. To this end, we weigh each common connected subgraph by its relations with the two whole graphs, which depend mainly on its nodes and edges. Then we define the weight of a set of non-overlap connected common subgraphs (NOCCS) based on the subgraphs' weights, and normalized by the sizes of the two graphs. Subsequently, we define the similarity between two graphs as the weight of the set of NOCCS whose weight is maximum.

2 Similarity measure

In the following part, we introduce our proposal similarity measure for two chemical compounds based on the 2D description. First, we recall some definitions of graph theory [15] used in this paper. Then, we describe our similarity measure and its properties in details.

2.1 Basic notions of graph theory

Definition 1 *A graph is a 4-tuple* $G =< V, E, \mu, \nu >$ *where* V *is a set of finite vertices,* $E \subseteq V \times V$ *is the set of edges,* $\mu : V \to L_V$ *is a function assigning labels to the vertices and* $\nu : E \to L_E$ *is a function assigning labels to the edges.*

For convenience, we denote a graph as a node set and an edge set, $G =< V, E >$.

Definition 2 *Graph* G *is called a connected graph if and only if there is at least one path between any vertex pair, where a path is a list of vertices such that there is an edge between two adjacent vertices.*

Definition 3 *Given graph* $G =< V, E, \mu, \nu >$*, subgraph* $G_i =< V_i, E_i, \mu_i, \nu_i >$ *of* G *is a graph where* $V_i \subseteq V$*,* $E_i = E \cap (V_i \times V_i)$*, and* μ_i *and* ν_i *are the restrictions of* μ *and* ν *to* V_i *and* E_i*.*

$$\mu_i(v) = \begin{cases} \mu(v) & \text{if } v \in V_i \\ \text{undefined} & \text{otherwise} \end{cases} \qquad \nu_i(v) = \begin{cases} \nu(v) & \text{if } v \in E_i \\ \text{undefined} & \text{otherwise} \end{cases}$$

Definition 4 G_i *is called a common subgraph of* G *and* G' *when* G_i *is a subgraph of both* G *and* G'*.*

Definition 5 *A set of subgraphs of* G*,* $\mathbf{\Gamma} = \{G_i =< V_i, E_i >: i = 1 \ldots\}$*, is called a set of NOCCS of* G*, denoted by* $\pi(\mathbf{\Gamma}, G)$*, when the subgraphs are connected subgraphs of* G *and their node sets are not overlapped.*

$$\pi(\mathbf{\Gamma}, G) \Leftrightarrow G_i s \text{ are connected subgraphs}, V_i \cap V_j = \emptyset \ \ \forall\, i, j$$

2.2 Similarity measure method and properties

Denote $G =< V, E, \mu, \nu >$ and $G' =< V', E', \mu', \nu' >$ the graphs presenting two compared chemical structures. The similarity score between G and G' is defined based on the weights of sets of NOCCS of G and G'. The weight of each set is built on the weights of its member subgraphs, which depend on the nodes and edges of the subgraphs.

Now, consider a subgraph $G_i =< V_i, E_i >$ of graph $G =< V, E >$. For each node v of V_i, we define its weight with respect to G_i and G, denoted $\tau(v, G_i, G)$, as the ratio between the number of edges from v in G_i and that in G:

$$\tau(v, G_i, G) = \frac{|\{v' : (v, v') \in E_i)\}|}{|\{v' : (v, v') \in E)\}|} \tag{1}$$

The weight $\tau(v, G_i, G)$ is clearly proportional to the number of common edges between G_i and G at node v. In other words, the more complete the structure of G_i at node v with respect to G, the

greater $\tau(v, G_i, G)$. It is obvious that $0 \leq \tau(v, G_i, G) \leq 1$. The equality of the right side happens when and only when all connected nodes of v in V belong to V_i.

The weight of subgraph G_i with respect to graph G is then defined as the sum of weights of nodes in V_i, denoted by $\rho(G_i, G)$,

$$\rho(G_i, G) = \sum_{v \in V_i} \tau(v, G_i, G) \tag{2}$$

Theorem 1 *For any set of non-overlap subgraphs* $\mathbf{\Gamma} = \{G_j =< V_j, E_j >: j = 1\}$ *of* G_i*, it holds true that*

$$\sum_{G_j} \rho(G_j, G) \leq \rho(G_i, G)$$

The equality occurs when and only when $\mathbf{\Gamma} = \{G_i =< V_i, E_i >\}$

The proof is given in Appendix.

Theorem 1 means that connected subgraphs are considered to be more important than unconnected (separated) subgraphs. In fact, the weight of a subgraph is greater than total weights of any set of its non-overlap subgraphs.

Denote $\mathbf{\Gamma} = \{G_i =< V_i, E_i >: i = 1 \ldots\}$ a set of NOCCS of G and G'. Having introduced how to determine the weight of a subgraph with respect to its super graph, we define the weight $\mathbf{\Gamma}$, denoted by $\delta(\mathbf{\Gamma})$, as the sum of products of weights of subgraphs G_is with respect to G and G' divided by the product of weights of G and G' with respect to themselves.

$$\delta(\mathbf{\Gamma}) = \frac{\sum_{G_i}(\rho(G_i, G)\ \rho(G_i, G'))}{\rho(G, G)\ \rho(G', G')}$$

Since $\rho(G, G) = |V|$ and $\rho(G', G') = |V'|$, $\delta(\mathbf{\Gamma})$ can be rewritten as

$$\delta(\mathbf{\Gamma}) = \frac{\sum(\rho(G_i, G)\ \rho(G_i, G'))}{|V||V'|}$$

It implies that the weight of a set of NOCCS of G and G' is defined based on its subgraphs and normalized by the size of G and G'.

Now we are ready to introduce the similarity measure for two graphs representing two compounds.

Definition 6 *The similarity between two graphs* G *and* G'*, denoted* $\psi(G, G')$*, is defined as the maximum weight of all possible sets of NOCCS of* G *and* G'*,*

$$\psi(G, G') = \max_{\mathbf{\Gamma}} \{\delta(\mathbf{\Gamma}) : \pi(\mathbf{\Gamma}, G)\ \textit{and}\ \pi(\mathbf{\Gamma}, G')\}$$

Theorem 2 *Let* $\mathbf{\Gamma} = \{G_1, \ldots, G_k\}$ *and* $\mathbf{\Gamma}' = \{G'_1, \ldots, G'_{k'}\}$ *be two sets of NOCCS of* G *and* G'*. If* $G_i \in \Gamma$ *is a subgraph of* $G'_j \in \Gamma'$ *for* $i = 1 \ldots k$*, then* $\delta(\mathbf{\Gamma}) \leq \delta(\mathbf{\Gamma}')$*.*

The proof is given in Appendix.

Theorem 2 means that when connected subgraph G_is are larger, the weight of $\mathbf{\Gamma}$ is greater.

Now we present the properties of the proposed similarity measure.

Propostion 1 *For any pair of graphs* (G, G')*, the following properties hold true:*

1. $0 \leq \psi(G, G') \leq 1$
2. $\psi(G, G') = \psi(G', G)$
3. $\psi(G, G') = 1$ *if and only if* G *and* G' *are isomorphic graphs.*
4. $\psi(G, G') = 0$ *if and only if* G *and* G' *have no common connected subgraphs of the size larger than 1.*

The proof is given in Appendix.

Procedure SimilarityComputing
In: G, G'
Out: $Sim(G, G')$
Begin
$\mathbf{\Gamma} = \emptyset$
repeat
cG = the largest connected common of (G, G')
$\mathbf{\Gamma} = \mathbf{\Gamma} + cG$
$G = G - cG, G' = G' - cG$
until $cG = \emptyset$
return $\delta(\mathbf{\Gamma})$
End

Figure 1: Algorithm for determining the similarity between two compounds

3 Experimental evaluation

To evaluate the usefulness of the proposed measure, we applied a clustering method to more than eleven thousands compounds obtained from the KEGG database [16] using the similarity calculated by our method. Then we performed several experiments to analyze the clusters of compounds: The first ones are to analyze relations between clusters of compounds of the whole database with other chemical information such as pathways, enzymes, etc. The second experiments are to analyze relations between pathway modules identified by clusters of similar structure compounds and that identified by genomic contexts, namely, operon structures of enzyme genes.

3.1 Methodology

Similarity determining Since finding the exact similarity between two compounds is an NP-hard problem, we compute heuristically the similarity between two compounds. According to Proposition 1, Theorem 1, and Theorem 2, larger connected common subgraphs have greater weights. Thus, we designed an algorithm to find the similarity between two graphs by finding sequently the largest common connected subgraph of the two graphs. The largest common connected subgraph is determined by a back tracking algorithm. Then the similarity between two compounds is estimated by the weight of the set of found subgraphs (see Fig. 1). This algorithm is parallelized to determine similarity of 11,149 compounds on 16 x 2GHz node PC-Clusters to speed up the computation.

Clustering method Clustering methods can be divided into two main approaches: partitioning and hierarchical. Since partitioning methods [17, 18] are not suitable for non-continuous data, we chose a hierarchical-based clustering method to cluster compounds. Among hierarchical-based clustering methods, we chose the method with the average complete linkage condition [19].

3.2 Clustering results analysis

3.2.1 Analysis on clustering results of the whole database

In this subsection, we analyze the clustering result of the whole database with the threshold similarity degree of 0.5. We found 2629 clusters and 1261 of them contain a single compound.

We found that compounds in the same clusters are strongly alike in structures. For example, for the five largest clusters, the common structure of each (see Fig. 2) is little different form its original compounds. Also, compounds in the same cluster share common names that indicate common

Table 1: Common formula, names, etc. of the five largest clusters

No.	Size	Common formula	common name	description of member	KEGG pathways map numbers						
					C	L	AA	BX	second	AtR	P&NP
1	188	C22O17N6P3S	CoA	Coenzyme A	640, 650	62, 71, 120	280	632			
2	115		rna	Ribonucleotid			251, 252, 260, 450			970	
3	98	C19O	one	cyclopenta[a] phenanthrene		140, 150					
4	82	C9O12P2	dp-, ose	pyran, diphosphate, methyl cyclopenta	51, 500, 520, 530				521		522
5	61	C6O	benz	containing benzene ring			380	362, 632, 623	622		

C: Carbon hydrate Metabolism; L: Lipid Metabolism; AA: Amino Acid Metabolism; BX: Biodegradation of Xenobiotics; Second: Biosynthesis of Secondary Metabolites; AtR: Genetic Information Processing (Translation);P&NP: Biosynthesis of Polyketides and Nonribosomal Peptides

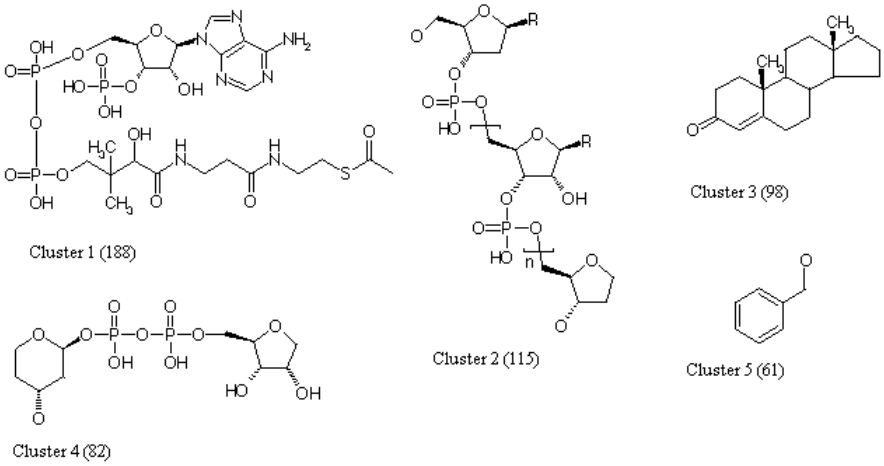

Figure 2: Common structures of the five largest clusters

properties of the compounds. As an example, compounds in Cluster 1 have the common name CoA (Coenzyme A), therefore possess the properties of Coenzyme A such as being required to metabolize fat, carbohydrate and protein and convert them into energy at the cellular level, or being the initiation of the body's energy cycle.

With a deeper analysis of the relation between compound clusters and pathways, we found that compound clusters highly associate with specific pathways in the KEGG database, e.g. Cluster 1 has 28 compounds taking part in *Fatty acid biosynthesis (path 2)*(map00062), Cluster 2 has 40 compounds joining *Aminoacyl-tRNA biosynthesis*(map00970). In addition, it can be seen from Table 1 that each cluster tends to associate with certain classes of pathways, e.g. compounds in Cluster 3 strongly associate with *Lipid Metabolism*(map00140, map00150), or compounds in Cluster 2 are assigned mainly to *Amino Acid Metabolism* and *Aminoacyl-tRNA biosynthesis* of genetic information processing.

Moreover, compounds in the same clusters are found to share the same groups of enzymes working on specific radicals in compounds, accordingly catalyzing the reactions they join. For example, compounds in cluster 2 use enzymes of EC 6.1.1 (Ligases Forming Aminoacyl-tRNA and Related Compounds) which mainly catalyze reactions that *rna* compounds take part in. Other introduced groups of enzymes also works on radicals that each cluster's common structure carries (Table 2).

In short, compounds in the same clusters not only share common structures and names but also strongly associate with specific pathways, mainly metabolic pathways, and share common groups of enzymes catalyzing their reactions.

Table 2: Compound clusters with their main enzyme requirements in related reactions

Cluster ID	EC number	Fre.	Functions
Cluster 1	EC 1.1	31	Acting on the CH-OH group of donors
	EC 1.2	35	Acting on the aldehyde or oxo group of donors
	EC 1.3	68	Acting on the CH-CH group of donors
	EC 2.3	210	Acyltransferases
Cluster 2	EC 6.1.1	23	Ligases Forming Aminoacyl-tRNA and Related Compounds
Cluster 3	EC 1.1	21	Acting on the CH-OH group of donors
	EC 1.14	18	Acting on paired donors, with incorporation or reduction of molecular oxygen
Cluster 4	EC 1.1	18	Acting on the CH-OH group of donors
	EC 2.4	203	Glycosyltransferases
Cluster 5	EC 1.14	27	Acting on paired donors, with incorporation or reduction of molecular oxygen

3.2.2 Analysis on clustering results of pathway oriented databases

Clustering analysis of the whole database shows a tendency of similar structures to be assigned to specific pathways. Thus, the clustering of compounds along the pathway maps provided by KEGG is an important step in order to learn more about the metabolic pathways and predict possible operon structures [12].

In this part, we analyze the result of clustering compounds and the correlation between compound clusters and enzyme clusters within metabolic pathways. Due to space limit, we present the analysis result on one pathway(pathway map00860) as an example. The analysis of other pathways can be downloaded at www.jaist.ac.jp/q̃uang/chemical/PathwayAnalysis/

Clustering of compounds on pathways: The result of clustering similar compounds on the pathway maps shows that there is a clear tendency of highly similar structure compounds to take up adjacent positions in reaction steps of the maps. As a result, the pathway maps can be divided into several parts depends on the chemical compounds achieved in each cluster. For example, the clustering compounds on the pathway map00860 (*Porphyrin and chlorophyll metabolism*-Fig. 3) identifies 5 noticeable compound clusters that are noted on the map as areas enclosed by thinner line, named C1 to C5 accordingly.

Correlation of enzyme clusters and compound clusters on the pathways: To find out about the relation between chemical information and genomic information, it is necessary to discover the correlation of compound clusters and enzyme clusters on the metabolic pathways. The enzyme clusters are derived from the ortholog table [20, 21] which contain the information about orthologous sets of enzyme genes. Analysis of the correlation between compound clusters and enzyme clusters helps to predict possible operon-like structures in selected genomes [12].

The most surprising discovery we achieved when examining the pathway oriented clustering is that in many cases, compound clusters and enzyme clusters almost completely overlap each other on the pathway maps. For instance, in Fig. 3, the area of C3 overlap most of that of E1. The intersection of compound clusters and enzyme clusters helps to point out the operon-like structures, e.g., in the intersection of enzyme clusters with C4, the possible operon-like structure(such as in *Pseudomonas aeruginosa*) consists of E2.5.1.17 E6.3.5.10 E6.3.1.10 E2.7.1.156 E2.7.7.62 E2.7.8.26, and another operon-like structure (such as in *Mycobacterium tuberculosis H37Rv*) consisting of E2.1.1.130 E1.14.13.83 E2.1.1.131 (CbiG) E2.1.1.133 is found within C3. Other possible operon-like structures are shown in Table 3.

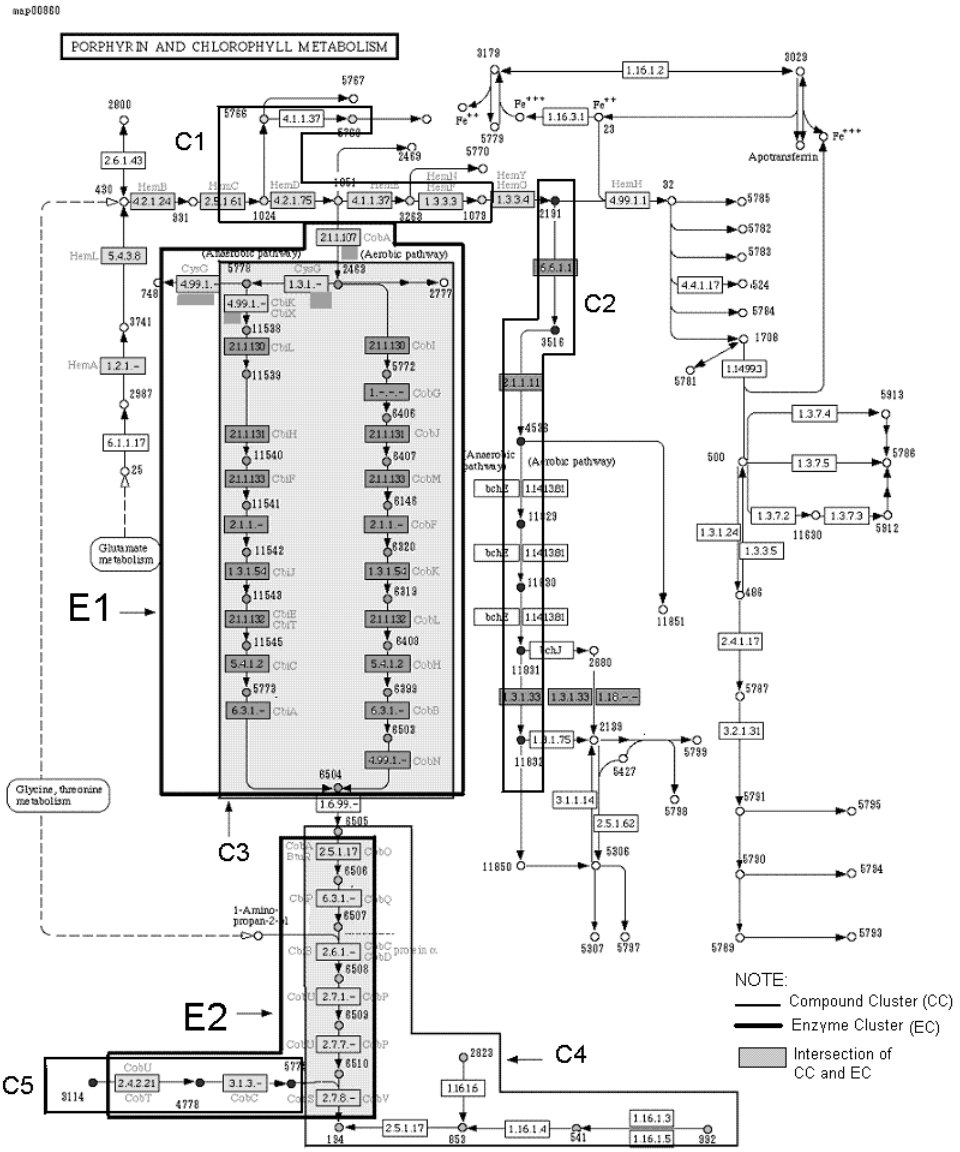

Figure 3: Example of compound/enzyme clusters in pathway oriented

In brief, the clustering of compounds on pathway maps reveals the tendency of similar compounds to take up adjacent steps of reactions on the pathways. Moreover, it shows that a set of enzyme genes encoded in an operon often corresponds to a set of enzymes catalyzing successive reaction steps (where compounds in the clusters are nodes) in a specific metabolic pathway. This encourages the new way of discovering knowledge on genome by analyzing structural similarity of chemical compounds.

4 Conclusion

In this paper we present an innovative similarity measure for 2D chemical structures. In our approach, we measure the similarity between two compounds (two graphs) by using not only atoms (nodes) and chemical bonds (edges) but also the connectivity of common substructures (common subgraphs).

Experiments with clustering for more than eleven thousand compounds in database KEGG/LIGAND discovered (revealed) clusters with highly similar structures compounds that share the same common

Table 3: Possible operon-like structure from KEGG Pathway map00860

Cluster area	Possible operon
Cluster 1	E4.1.1.37 E1.3.3.3
Cluster 2	E6.6.1.1 E2.1.1.11
Cluster 2	E1.3.1.33
Cluster 3	E2.1.1.130 E1.14.13.83 E2.1.1.131 (CbiG) E2.1.1.133 E2.1.1.152 E1.3.1.54 E2.1.1.132 (CbiD) E5.4.1.2 E6.3.5.9 E6.3.1.- E6.6.1.2
Cluster 3	E1.3.1.- E4.99.1.-
Cluster 4	E2.5.1.17 E6.3.5.10 E6.3.1.10 E2.7.1.156 E2.7.7.62 E2.7.8.26
Cluster 5	E3.1.3.73 E2.4.2.21

names, take part in the same pathways with the same requirement of enzymes in actions. Analysis on clustering results of pathway oriented databases showed that clusters of compounds and clusters of enzymes on the same pathway have a tight relation, and this encourages the new way of discovering knowledge on genome by analyzing structural similarity of chemical compounds.

Acknowledgments

This work was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Culture, Sports, Science, and Technology of Japan; and the COE project JCP KS1 from Japan Advanced Institute of Science and Technology. The first and third authors have been supported by the Japanese Government Scholarship (Monbukagakusho). We appreciate Le Sy Vinh at the Heinrich-Heine University of Duesseldorf, Germany for helpful comments on the manuscript.

References

[1] Brown R. D. and Martin Y. C. Use of structure - activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.*, 36:572–584, 1996.

[2] Brown R. D. and Martin Y. C. The information content of 2d and 3d structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.*, 37:1–9, 1997.

[3] Weiniger D. Introduct of encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[4] Flower D.R. On the properties of bit-string measures of chemical similarity. *J. Chem. Inf. Comput. Sci.*, 38:379–386, 1998.

[5] Gower J.C. Coeffecients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biometrics*, 27:857–871, 1971.

[6] Gower J.C and Legendre P. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification*, (3):5–48, 1986.

[7] Liebetrau A.M. *Measures of association.* Newbury Park, CA: Sage publications, 1983.

[8] Wegner J. K., Frhlich H., and Zell A. Feature selection for descriptor based classification models: Part i theory and ga-sec algorithm. *J. Chem. Inf. Comput. Sci.*, 44:921–930, 2004.

[9] Wegner J. K., Frhlich H., and Zell A. Feature selection for descriptor based classification models: Part ii human intestinal absorption (hia). *J. Chem. Inf. Comput. Sci.*, 44:921–930, 2004.

[10] Raymond J.W and Willet P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.*, 16:521–533, 2002.

[11] Hattori M., Okuno Y., Goto S., and Kanehisa M. heuristics for chemical compound matching. In Michael G., Minoru, K. Satoru M., and Toshihisa T., editors, *Genome Informatics*, pages 144–153, 2003.

[12] Hattori M, Okuno Y, Goto S, and Kanehisa M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, 125(39):11853–65, 2003.

[13] Raymond J.W, Gardiner E.J, and Willet P. Rascal: Calculation of graph similarity using maximum common edge subgraphs. *Comput. J.*, 45:631, 644 2002.

[14] Raymond J.W, Gardiner E.J, and Willet P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithms. *J. Chem. Inf. Comput. Sci.*, 42:305–316, 2002.

[15] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press and McGraw-Hill., the third edition, 2002.

[16] Kanehisa M., Goto S., Kawashima S., and Nakaya A. The kegg databases at genomenet. *Nucleic Acids Res.*, 30:42–46, 2002.

[17] MacQueen J. Some methods for classification and analysis of multivariate observation. In *Proceedings 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[18] Kaufmann L. and Rousseeuw P.J. Clustering by means of medoids. *Statistical Data Analysis based on the L1 Norm*, pages 405–416, 1987.

[19] Sokal R.R. and Michener C.D. Statistical method for evaluating systematic relationships. *University of Kansas science bulletin*, 38:1409–1438, 1958.

[20] Ogata H., Fujibuchi W., Goto S., and Kanehisa M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res*, 28:4021–8, 2000.

[21] Fujibuchi W., Ogata H., Matsuda H., and Kanehisa M. Automatic detection of conserved gene clusters in multiple genomes by graph comparison and p-quasi grouping. *Nucleic Acids Research*, 28:4029–36, 2000.

Appendix

Proof of Theorem 1

- $\sum_{G_j} \rho(G_j, G) \leq \rho(G_i, G)$

  Since G_js are subgraphs of G_i, $\tau(v, G_j, G) \leq \tau(v, G_i, G)$ when $v \in V_j$.

  Since V_js are not overlapped,

$$\sum_{G_j} \sum_{v \in V_j} \tau(v, G_j, G)) \leq \sum_{G_j} \sum_{v \in V_j} \tau(v, G_i, G)) \leq \sum_{v \in V_i} \tau(v, G_i, G) \tag{3}$$

$$\Rightarrow \sum_{G_j} \rho(G_j, G) \leq \rho(G_i, G) \ \square \tag{4}$$

- $\sum_{G_j} \rho(G_j, G) = \rho(G_i, G) \Leftrightarrow \mathbf{\Gamma} = \{G_i\}$
  - " $\Rightarrow$ " Since

$$\sum_{G_j} \sum_{v_j \in V_j} \tau(v_j, G_j, G) \leq \sum_{G_j} \sum_{v_j \in V_j} \tau(v_j, G_i, G) \leq \sum_{v_i \in V_i} \tau(v_i, G_i, G),$$

$$\bigcup_{V_j} = V_i, \text{ and } \tau(v_j, G_j, G) = \tau(v_j, G_i, G) \ \forall v_j \in V_i.$$

    Thus $G_j \equiv G_i$ or $\mathbf{\Gamma} = \{G_i\}$
  - " $\Leftarrow$ " is obvious.

Proof of Theorem 2

According to 1, it is clear that for $G_{i_k}(\mathrm{k} = 1..)$ in $\mathbf{\Gamma}$ being subgraphs of G'_j in $\mathbf{\Gamma}'$,

$$\sum_k \rho(G_{i_k}, G) \leq \rho(G'_j, G), \ \sum_k \rho(G_{i_k}, G') \leq \rho(G'_j, G')$$

On the other hand, we have

$$\sum_k \rho(G_{i_k}, G)\rho(G_{i_k}, G') \leq \sum_k \rho(G_{i_k}, G) \sum_k \rho(G_{i_k}, G') \leq \rho(G'_j, G)\rho(G'_j, G')$$

Consequently,

$$\delta(\mathbf{\Gamma}) = \frac{\sum_{G_i}(\rho(G_i, G)\ \rho(G_i, G'))}{\rho(G, G)\ \rho(G', G')} \leq \frac{\sum_{G'_j}(\rho(G'_j, G)\ \rho(G'_j, G'))}{\rho(G, G)\ \rho(G', G')} = \delta(\mathbf{\Gamma}') \ \square$$

Proof of Proposition 2

1. From the definition in Equation 1, $\rho(G_i, G) \leq |V_i|$.

   Thus, for any common subgraph set $\mathbf{\Gamma}$ of (G, G'),

$$\delta(\mathbf{\Gamma}) \leq \frac{\sum_i (|V_i|)^2}{|V|\ |V'|}$$

   Meanwhile, since G_is are disjoint common subgraphs of (G, G'),

$$\sum_i |V_i| \leq \min(|V|, |V'|).$$

   Hence,

$$\sum_i |V_i|^2 \leq \left(\sum_i |V_i|\right)^2 \leq \min(|V|, |V'|)^2 \leq |V||V'| \tag{5}$$

   This leads to $\psi(G, G') \leq 1$.

   The left part of Property 1 can be obviously seen.

2. It is apparent that $\delta(\mathbf{\Gamma})$ is the same no matter (G, G') or (G', G). Thus, Property 2 is true.

3. $\psi(G, G') = 1 \Leftrightarrow$ the equality in inequality (5) happens. This is equivalent to $|V| = |V'|, |\mathbf{\Gamma}| = 1$, and $\rho(G_1, G) = \rho(G_1, G') = |V_1| = |V|$, which means G and G' are isomorphic.

4. $\psi(G, G') = 0$ is equivalent to $\rho(G_i, G) = 0$ and $\rho(G_i, G') = 0 \ \forall G_i \in \mathbf{\Gamma}$. That means $|V_i| = 1$ for all G_i, or G and G' have no connected common subgraphs of the size larger than 1.