

Graph Theoretic Analysis of Chemical Compounds in Biological Pathways

Atsuko Yamaguchi ¹

atsuko@kuicr.kyoto-u.ac.jp

Yasushi Okuno ²

okuno@pharm.kyoto-u.ac.jp

Hiroshi Mamitsuka ¹

mami@kuicr.kyoto-u.ac.jp

¹ Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji 611-0011, Japan

² Graduate School of Pharmaceutical Sciences, Kyoto University Sakyo-ku, Kyoto 606-8501, Japan

Keywords: chemical compounds, molecular graph, graph theory, tree-width

1 Introduction

Graphs are very suitable models for representing the constitutions of molecules [4]. However, in the field of chemoinformatics, methods using graph algorithms have not entered the mainstream because graph problems comparing two graphs are often intractable. For example, the problem of finding the maximum common subgraph of two graphs is known to be NP-hard [1], even for two graphs of bounded degree [3].

In [5], we analyzed the characteristics of the structures of chemical compounds in biological pathways, in order to search for chemical structures in a biological database efficiently using graph representations. We investigated the complexity of graphs using tree-width, which is one of measures for the complexity of graphs. Our experimental results have shown that out of 10,000 chemical compounds in KEGG LIGAND [2], approximately 500 and only one compounds have tree-widths of three and four respectively, and the others have that of less than three.

In this paper, we analyze chemical compounds of tree-width three which are relatively complex in biological pathway. First, we classify the chemical compounds of tree-width three focusing on the most complex parts of these compounds. Furthermore, we evaluate this classification from a variety of biological viewpoints.

2 Method and Results

The tree-width is a complexity measure of graphs that takes an integer in the range of 1 to $N - 1$ for a graph with N nodes and increases with increasing complexity of the graph. The definition of tree-width is as follows. The *tree-decomposition* of a graph G is a pair (T, X) , where T is a tree and $X : V(T) \rightarrow 2^{V(G)}$ that satisfies the following three conditions: (1) $\cup_{t \in V(T)} X(t) = V(G)$, (2) for every edge $(u, v) \in V(G)$, there exists a vertex $t \in V(T)$ such that $u, v \in X(t)$, (3) for any three vertices $r, s, t \in V(T)$, if s is on the path from r to t , $X(r) \cap X(t) \subseteq X(s)$. The *width* of a tree-decomposition (T, X) is $\max_{t \in V(T)} |X(t) - 1|$. The *tree-width* of a graph G is the minimum width of all tree-decompositions of G .

The *local tree-width with range r* of a graph G is defined as $\max\{\text{The tree-width of } G_r(v) \mid v \in V(G)\}$, where $G_r(v)$ is the induced subgraph by the vertex set $\{u \in V(G) \mid \text{There is a path of length at most } r \text{ between } u \text{ and } v\}$.

For each compound of tree-width three in the LIGAND database, we checked the smallest range r with local tree-width three. By definition of local tree-width, the size of the range indicates the size of

the most complex part of a compound. From our experiment, we found that the range of more than 70% of the compounds of tree-width three is three. In addition, we found that the compounds with the same ranges tended to include similar substructures in their most complex parts. For example, all chemical compounds with the range twelve had a same core.

However, the compounds with the range three have various types of structures. For these compounds, we attempt to classify them using the most complex part of a chemical compound. We first call the most complex part of a chemical compound a *core* of the compound which can be formally defined as a subgraph that consists of minimal subgraphs with tree-width k of the molecular graph. We compute all cores of molecular graphs of the range three.

Second, we make a directed graph G such that $V(G)$ is a set of cores found in the previous step and $E(G) = \{(u, v) \mid u \text{ is a subgraph of } v\}$. Since the binary relation between a graph and its subgraph is a partial order, the graph G is a directed acyclic graph. Therefore, we can make a list of cores which are sources of G . We call a core which is a source of G a *base*.

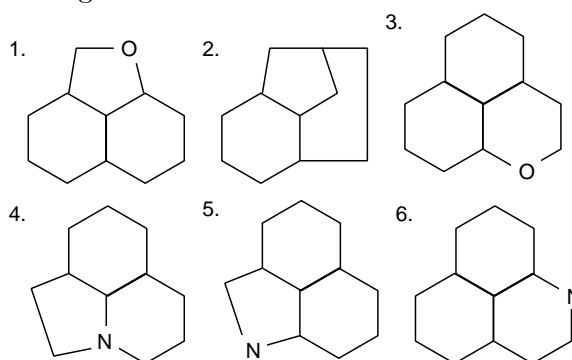


Figure 1: Examples of bases.

3 Discussions

We checked the distribution of chemical compounds, whose locations on pathway maps are already known. For all cases, chemical compounds in a base are found in a series of consecutive chemical reactions. This result indicates that our measure classified chemical compounds from a viewpoint of metabolic pathways, despite that it is purely derived from graph theory. Our possible future work is to progress our detailed analysis further from a variety of viewpoints including statistical tests.

References

- [1] Garey, M.R. and Johnson, D.S., *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, 1987.
- [2] Goto, S., Okuno, Y., Hattori, M., Nishioka, T. and Kanehisa, M., LIGAND: Database of chemical compounds and reactions in biological pathways, *Nucleic Acids Res.*, 30:402–404, 2002.
- [3] Kann, V., On the approximability of the maximum common subgraph problem, *Proc. of 9th Ann. Symp. on Theoretical Aspects of Comput. Sci.*, 377–388, 1998.
- [4] Trinajstic, N., *Chemical Graph Theory*, CRC Press, 1992.
- [5] Yamaguchi, A., Aoki, K. F. and Mamitsuka, H., Graph complexity of chemical compounds in biological pathways, *Genome Informatics*, 14:376–377, 2003.