Graph Complexity of Chemical Compounds in Biological Pathways

Atsuko Yamaguchi atsuko@kuicr.kyoto-u.ac.jp
Kiyoko F. Aoki kiyoko@kuicr.kyoto-u.ac.jp
Hiroshi Mamitsuka mami@kuicr.kyoto-u.ac.jp
Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho,
Uji 611-0011, Japan

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

1 Introduction

Graph theory for chemical compounds is often studied for the fact that labeled graphs are suited to express the connectivity of chemical compounds [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]. Therefore, we focus on chemical compounds in biological pathways and analyze the characteristics of the chemical compounds to reduce the problem space to allow for tractable comparisons of graphs.

In analyzing these characteristics, we can focus on consistent measures for estimating the similarity of chemical compounds. Such measures are essential for gaining an understanding of the structural aspects of chemical compounds. They also assist in useful tasks such as querying chemical databases. As an example of the utility of similarity measures of chemical compounds, we were able to develop a polynomial-time algorithm for the maximum common subgraph problem [5]. This was possible due to the identification of similarity measures based on structural characteristics, as we explain below.

2 Method and Results

As a first step towards capturing the properties of chemical compounds, we examined the *tree-width*, a measure indicating the complexity of a given graph, of chemical compounds found in molecular biology. Since experimental results show that compounds with tree-width 3 or 4 have some particular structural characteristics, we further analyzed the *local tree-width*.

2.1 Tree-Width

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 $X : V(T) \to 2^{V(G)}$ 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.

We obtained 9712 chemical compounds from the LIGAND database [2] and examined the tree-width of each compound. In all but one of the 9712 cases, the tree-width was between 1 and 3, and the tree-width of the one exception was 4 (Table 1). In fact, the high complexity of the structure of this one exception, as illustrated in Figure 1, may allow us to ignore it. Thus, we are able to claim that the structures of chemical compounds in biological pathways are generally simple in terms of tree-width.

Tree-width	# compounds
1	1881 (19.4%)
2	7336 (75.7%)
3	477 (4.9%)
4	1 (0.01%)
≥ 5	0

Table 1: Tree-width analysis of the LIGAND database.

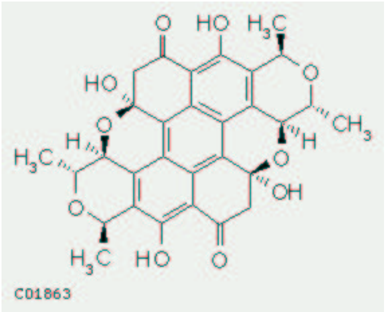

Figure 1: Xanthoaphin, the chemical compound found of tree-width 4.

2.2 Local Tree-Width

Next, we analyzed these chemical compounds whose tree-width was 3. The local tree-width with range r of a graph G is defined as the maximum tree-width of a subgraph $G_r(v)$ of G such that the vertex set of $G_r(v)$ is the set of vertices from some vertex v.

For each compound of the LIGAND database, we checked the smallest range r with local tree-width 3. 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 was 3. Furthermore, it is worth noting that there is a peak in the number of compounds with range 9, as well as small peaks for ranges 6 and 12. In addition, we found that the compounds with the same ranges tended to include similar substructures in their most complex parts.

Table 2: The number of compounds corresponding to the smallest ranges for local tree-width 3.

Range	2	3	4	5	6	7	8	9	10	11	12	13–16	17	≥ 18
# compounds	9	348	2	9	15	4	12	27	1	0	12	0	1	0

3 Conclusion

Based on our results, we claim that the tree-width of compounds is not only a useful complexity measure, but it is also an appropriate factor to consider in characterizing chemical compounds.

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. and Mamitsuka, H., Finding the maximum common subgraph of a partial k-tree and a graph with a polynomially bounded number of spanning trees, *Proc. 14th Ann. Inter. Symp. Algo. Comput., ISSAC 2003*, in press.