Finding a Maximum Common Subgraph from Molecular Structural Formulas through the Maximum Clique Approach Combined with the Ising Model

Yasuharu Okamoto*

ABSTRACT: We examined the maximum common subgraph (MCS) of four neuraminidase inhibitors that were antiviral medication for treating and preventing type A and B influenza viruses. The MCS was obtained by finding a maximum clique of an association graph constructed from the two input chemical structural formulas. Maximum clique problem was reformulated to Ising Hamiltonian to allow for applying various techniques for optimization. We observed that the combined label for a vertex composed of elemental species and chemical bonds significantly worked well for decreasing the number of vertices in the association graph, which in turn helped to reduce the computational cost.

1. INTRODUCTION

A chemical structural formula is often regarded as a kind of graph where atoms and chemical bonds correspond to vertices (V) and edges (E) of the graph $G = (V, E)$, respectively. It is for this reason that approaches based on graph theory have been often used to deal with problems related with the structural formula. For example, finding a maximum common subgraph (MCS) from multiple structural formulas is an essential work for drug discovery because the molecules that have similar partial structures are expected to have similar drug efficacy. In graph theory, two undirected graphs, $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, are isomorphic if there is a bijection between their vertex sets that preserves adjacency. An MCS stands for a graph that has the largest number of vertices of any graph isomorphic to the induced subgraph of $G_1$ and $G_2$. The MCS has an extremely high computational complexity where several well-known NP-complete problems reduced to it. In the worst case, the number of MCSs increases exponentially as the number of vertices in the graph increases. In addition to applying a structural formula, MCS is a familiar and challenging problem in graph theory because it offers a plethora of applications such as bioinformatics and pattern recognition.

There are two types of approaches to solve the MCS problem: one is clique-based approaches, and the other is nonclique-based backtracking approaches. The former approaches are the most widely used approaches in the literature, where the MCS problem results in finding a maximum clique in the association graph. An association graph is a particular form of a product graph generated from two input graphs $G_1$ and $G_2$. The definition of the association graph will be given in the Methods section. In graph theory, the maximum size of the adjacency matrix of the association graph is $|V_1||V_2|$. The value becomes large unless the size of each graph of $G_1$ and $G_2$ is sufficiently small. To reduce the computational requirements in solving the MCS problem, it appears to be indispensable to exploit attribute information to reduce the size of the adjacency matrix of the association graph. In the molecular structural formula, each vertex, namely, an atom is labeled by a chemical element. In addition, each edge corresponds to a chemical bond, which might also be used as another attribute label.

Although a maximum clique in an arbitrary graph can be found using the branch-and-bound algorithm combined with approximate coloring to obtain an upper bound on the size of a maximum clique, it also can be found through the Ising model. It is noteworthy that quantum and classical annealing techniques based on the Ising model have attracted great attention over several years as a promising tool for computationally intensive optimization problems. Although it remains to be clarified whether such techniques are more efficient and accurate than conventional metaheuristics, such as simulated annealing and genetic algorithm, or not, the Ising model approach appears to be promising because we can expect further advancement in both software and hardware associated with the annealing-related techniques.

In this study, we examined the MCS of four neuraminidase inhibitors. A neuraminidase inhibitor obstructs the action of viral neuraminidases of the influenza virus by preventing its...
neuraminidase inhibitor. The molecule is a prodrug of an

2. RESULTS AND DISCUSSION

2.1. Simple Example of the Association Graph. We first explain a simple case of an association graph generated from two input graphs \( G_1 = (V_1, E_1) \) and \( G_2 = (V_2, E_2) \) without attribute information (top panel in Figure 1). According to the definition of the association graph (see Methods section), two vertices \((A, \alpha)\) and \((B, \beta)\) in the association graph are adjacent because \((A, B) \in E_1\) and \((\alpha, \beta) \in E_2\). Similarly, two vertices \((A, \alpha)\) and \((D, \delta)\) in the association graph are adjacent because \((A, D) \notin E_1\) and \((\alpha, \delta) \notin E_2\). On the other hand, two vertices \((C, \gamma)\) and \((D, \delta)\) in the association graph are not adjacent because \((C, D) \in E_1\) but \((\gamma, \delta) \notin E_2\). There are four maximum cliques in the association graph: \{\((A, \alpha), (B, \beta), (C, \delta)\)\}, \{\((A, \alpha), (B, \beta), (D, \delta)\)\}, \{\((A, \delta), (B, \beta), (C, \alpha)\)\}, and \{\((A, \delta), (B, \beta), (D, \alpha)\)\}. Subsequently, we consider the case where the vertices have some kind of attribute information, as shown in the bottom panel in Figure 1, where the vertices were distinguished by the colors: yellow and blue. There is one maximum clique \{\((A, \alpha), (B, \beta), (D, \delta)\)\} in the association graph because vertices with different colors do not match. It is noteworthy that the number of vertices in the association graph is decreased from 16 to 8 by considering the attribution information in this example.

2.2. MCS of Zanamivir and Laninamivir Octanoate. Zanamivir \((C_{12}H_{22}N_4O_7)\) is the first neuraminidase inhibitor commercially developed. It is a medication used to treat influenza caused by type A and B influenza viruses, but it does not have efficacy for the type C virus which does not have neuraminidase. Laninamivir octanoate \((C_{16}H_{30}N_2O_4)\) is also a neuraminidase inhibitor. The molecule is a prodrug of an octanoic acid ester and undergoes hydrolysis to form an active metabolite of laninamivir \((C_{17}H_{22}N_2O_4)\) by the body. In MCS problem concerning a chemical structural formula, hydrogen atoms are often ignored to emphasize the structural similarities of the basic skeleton. It is noteworthy that the basic skeleton of zanamivir obtained by ignoring hydrogen atoms is contained in that of laninamivir octanoate, as shown in Figure 2.

Although identifying the MCS between zanamivir and laninamivir octanoate was rather a trivial task for this reason, computation through the Ising model correctly assigned the entire basic skeleton of zanamivir as the MCS of the two molecules. In this case, one connected graph was obtained as the MCS by finding a maximum clique of the association graph generated by the two input graphs. Note that, however, the present approach provides the MCS that has the maximum number of constituent vertices, which does not necessarily mean that the obtained MCS is connected. The number of vertices in the association graph was reduced from 759 to 324 and to 118 by considering the label of elemental species only and the combined label of both elemental species and bond information, respectively. It is noteworthy that the value of 78.6% of the off diagonal elements in the reduced adjacency matrix obtained using the combined labels becomes 1. This indicates that most vertices are connected by edges. Although the structural similarity between zanamivir and laninamivir is notable, zanamivir and laninamivir have different dose regimens. The former requires twice a day for five days as the duration of drug exposure. By contrast, the latter needs just one inhalation administration. Octanoic acid in laninamivir octanoate is an 8-carbon straight-chain fatty acid that is almost insoluble in water. Therefore, when it is inhaled, it is not taken up into the blood but passes through the lipid bilayer of mucosal cells and is slowly broken down in the cytoplasm, enabling it to act for a long time. That makes the difference in the dose regimen between laninamivir and zanamivir.

2.3. MCS of Zanamivir and Oseltamivir. Subsequently, we examined the MCS of zanamivir and oseltamivir \((C_{16}H_{28}N_2O_4)\). Oseltamivir is also a neuraminidase inhibitor, and it is used internally unlike inhalation drug such as zanamivir and laninamivir. We can observe the structural similarity between the two molecules even from plain visual comparison of Figure 3: both molecules contain a six-membered ring and have the same groups of atoms such as \(-COO\) and \(-NHCOCH_3\). Figure 3 also shows an extracted MCS through a maximum clique of the association graph generated by the two input graphs. Unlike the connected graph in the above-mentioned case, the MCS consists of two disconnected graphs designated by red and sky blue parts. Disconnected graphs are usually obtained by the present approach. In addition, as shown in the figure, we observed that there were four other cases with respect to the possible matching of vertices in the small subgraph comprising four atoms OCCC: \((a,c), (a,d), (b,c), (b,d)\).

The number of vertices in the association graph was reduced from 506 to 228 and to 86 by considering the label of elemental species only and the combined label of both elemental species and bond information, respectively. The value of 75.8% of the off diagonal elements in the reduced adjacency matrix obtained by the combined labels becomes 1. It is noteworthy that although the carbon atom designated by a green arrow in Figure 3 should
be included in the MCS at first glance, the atom does not belong to the MCS because of the constraint that originates from the definition of the isomorphism of the association graph. The O atom designated by the blue arrow is adjacent to the carbon atom in the left graph. The atom, however, is not adjacent to the carbon atom in the right graph. Consequently, two vertices \((C_{\text{left}} C_{\text{right}})\) and \((O_{\text{left}} O_{\text{right}})\) are not adjacent in the association graph, which indicates that the carbon atom cannot be included in the clique because all vertices in the clique must be adjacent. This result illustrates that the present approach does not necessarily provide the largest connected MCS.

2.4. MCS of Zanamivir and Peramivir. Because peramivir \((C_{15}H_{28}N_{4}O_{4})\) has three inhibition sites for neuraminidase, it shows a high breeding depression effect toward influenza A and B viruses, and it is effective in a single dose through intravenous drip injection.19 As shown in Figure 4, we observe that the structural similarity between the two molecules is not so much high as the above-mentioned cases at first glance: the size of the ring is different between the two molecules; nonetheless, they have in common that the same groups of atoms are attached to each ring. The MCS consists of 17 vertices, which indicates that the size of it is larger than that in Section 2.3. Figure 4 shows an example of the computed MCS through a maximum clique of the association graph generated by the two input graphs. We observed that the MCS consist of four subgraphs distinguished by colors: red, blue, green, and purple. The smallest subgraph contains just one O atom. Note that there are many MCSs that have the same size. The number of vertices in the association graph was reduced from 529 to 224 and to 84 by considering the label of elemental species only and the combined label of both elemental species and bond information, respectively. The value of 77.5% of the off diagonal elements in the reduced adjacency matrix obtained by using the combined label becomes 1. It is noteworthy that although the carbon atom designated by arrow in Figure 4 should be included in the MCS at first glance, the atom is not included in the MCS because of the definition of isomorphism of the association graph, as explained in Section 2.3. We used 0.9 for the value of hyperparameter C to preserve the constraint conditions in this case.

Because combinatorial optimization based on the Ising model is a heuristic approach, the optimality of the results is not guaranteed; however, it is obvious in the case of laninamivir octanoate and zanamivir because all heavier atoms in zanamivir are included in the obtained MCS. For the cases of oseltamivir and zanamivir as well as peramivir and zanamivir, we confirmed that the number of vertices of the obtained MCSs agrees with the number of vertices of the maximum clique of the association graphs computed using an API (find_cliques) of NetworkX20,21 a tool for complex network research. In addition, MCSs of other three pairs of molecules (laninamivir octanoate, oseltamivir), (laninamivir octanoate, peramivir), and (oseltamivir, peramivir) were also examined to make the study more exhaustive, and the results were shown as Figure S1 in the Supporting Information.

3. CONCLUSIONS

We examined the MCSs among zanamivir, laninamivir octanoate, oseltamivir, and peramivir. The MCS was determined by finding the maximum clique of the association graph that were generated from two input chemical structural formulas combined with the Ising model. The combined label determined from elemental species and chemical bond significantly worked well for decreasing the number of vertices in the association graph, which in turn, reduced the computational cost.

The following point shall be remarked: the present approach finds the MCS that is not necessary a connected graph but rather a graph with the maximum number of vertices. This might cause somewhat of a problem because most tools for the MCS in cheminformatics such as the RDkit21 try to find the maximum connected graph at least in the default setting. Note that the connected part of the graph in the MCS determined by the present approach does not necessarily correspond to the maximum connected graph of the association graph. Thus, it would be desirable to apply the present approach to the problems that place a premium in deciding the MCS comprising the disconnected graphs.

It is also noteworthy that the MCS among three or more molecules can be found by the straightforward extension of the isomorphic mapping for generating the association graph. However, the approach results in significant increase of the computational cost: the number of vertices in the extended association graph of m molecules becomes as many as \(\prod_{i=1}^{m} |V_i|\) in the worst case. Consequently, it would be more practical to extract the MCS of many molecules on a one-by-one basis instead of considering all molecules at a time by extending the association graph.

4. COMPUTATIONAL METHODS

4.1. Association Graph. Given two undirected graphs \(G_1 = (V_1, E_1)\) and \(G_2 = (V_2, E_2)\), the association graph \(G = (V, E)\) is an
undirected graph defined on the vertex set \( V = V_1 \times V_2 \) with two vertices \((u_1, v_1)\) and \((u_2, v_2)\) being adjacent whenever

\[ u_1 \neq u_2 \text{ and } v_1 \neq v_2, \text{ and either } ((u_1, u_2) \in E_1 \text{ and } (v_1, v_2) \in E_2) \]

or \(((u_1, u_2) \notin E_1 \text{ and } (v_1, v_2) \notin E_2)\).

4.2. Labeling of Atoms. There are two types of labels with respect to structural formula: elemental species and chemical bonds corresponding to vertices and edges, respectively. It is noteworthy that the label for the edge can be combined into the label for the vertex if we define the vertex label as elemental species having specific bond information. This significantly reduces the number of possible matching of vertices in the construction of the association graph. We defined the combined label as follows: tens place and ones place digits correspond to elemental species and bond information labels, respectively. In fact, the combined label of atom \( X \) \((L[X])\) was assigned by \( L[X] = (\text{atomic number of } X) \times 10 + (\text{bond order of } X)\). For example, in the case of \( \text{C} \equiv \text{O}, L[\text{C}] = 6 \times 10 + 3 \) and \( L[\text{O}] = 8 \times 10 + 3 \). To further reduce the actual number of vertices and to focus on a basic skeleton of a molecule formed by elements and bonds corresponding to vertices and edges, respectively. It is noteworthy that the label for the edge can be combined into the vertex label as elemental species and bond information labels, respectively. In other words, we did not distinguish double bonds from single bonds in the ring.

Let us consider a hydrocarbon \((\text{C}_n\text{H}_{2n+2} \to 2m)\) consisting of \(n\) carbon atoms with \(m\) \(\text{C} \equiv \text{C}\) double bonds. (However, we ignored \(= = =\) to simplify the analysis.) It is noteworthy that the molecule has \( (n - 2m) > 2 < 2m > 2 \). If we construct an association graph from two \(\text{C}_n\text{H}_{2n+2} \to 2m\) molecules, the number of vertices in the graph will be \(n^2\) for the vertex label that considers only elements, but \((n - 2m)^2 + 4m^2\) for the vertex label consolidates elements and bonds. The ratio of the latter to the former is significantly smaller than \(1\) over a wide range of \(n\) for a certain magnitude of \(m\), as shown in Figure 5. Because quadratic unconstrained binary optimization (QUBO) is on the scale of the square of the number of vertices and because the increase of the number of vertices requires more optimization iterations, the computational complexity can be considered to be substantially proportional to the third power of the number of vertices or more, the reduction of the number of vertices is very important for improving the efficiency of the computation.

4.3. Solving Maximum Clique Problem through Ising Model. A clique is an induced subgraph of \(G\) that forms a complete graph: every two distinct vertices in the clique are adjacent. A maximum clique of a graph \(G\) is a clique such that there is no clique with more vertices. The NP-complete decision problem of whether or not a clique of the size \(K\) exists can be written as an Ising-like model using a binary bit variable \(x_i\) \((0\ or\ 1)\). The relevant Hamiltonian \(\{H\}\) for the maximum clique problem is the sum of three partial Hamiltonian operators

\[
H = H_A + H_B + H_C,
\]

where

\[
H_A = A \left(1 - \sum_{n=2}^{N} y_n \right)^2 + A \left(\sum_{n=2}^{N} ny_n - \sum_{n} x_n \right)^2,
\]

\[
H_B = B \left[\frac{1}{2} \sum_{n=2}^{N} ny_n + 1 + \sum_{n=2}^{N} ny_n - \sum_{n} x_n x_n \right]
\]

and

\[
H_C = -C \sum_{n} x_n
\]

Here, \(N\) is the number of vertices of graph \(G\), and \(y_n\) is an ancillary binary bit to represent a clique whose size is \(n\). It is noteworthy that \(H_A\) becomes zero only when the number of selected vertices \(x_i\) is \(n\). Similarly, \(H_B\) becomes zero only when the selected vertices \(x_i\) form a complete graph. \(H_C\) means that it is energetically favorable to include as many vertices as possible into the clique, although the aggressive inclusion may be penalized by the terms \(H_A\) and \(H_B\). The constructed Ising Hamiltonian of the clique problem is a quadratic expression of the binary variables \((x_i, y_i)\), which is consistent with the QUBO form to which various optimization techniques are applicable. Here, we used Qbsolv (tabu search) that is an application programming interface supported in D-Wave Ocean (software development kit developed by D-Wave Systems). We set the hyperparameters \(A, B, C\) in the Hamiltonian to be \(A = (\Delta + 2) B\) and \(B = C\) as suggested by Lucas. In particular, we set \(\Delta = \min \{W_1, W_2\}\) and \(C = 1\) unless otherwise stated. In actual computation, we used the following binary approach to reduce the number of ancillary bits from \(N\) to \(M + 1\) as explained by Lucas:

\[
2^M \leq N < 2^{M+1}
\]

\[
M = \text{floor}(\log N)
\]

\[
\sum_{n=1}^{N} ny_n \rightarrow \sum_{n=0}^{M-1} 2^n y_n + (N + 1 - 2^M)y_M
\]

**ASSOCIATED CONTENT**

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.0c00987.
Pairs of structural formula of laninamivir octanoate and oseltamivir, laninamivir octanoate and peramivir, and oseltamivir and peramivir (PDF)

AUTHOR INFORMATION

Corresponding Author

Yasuharu Okamoto — The System Platform Research Laboratories, NEC Corporation, Tsukuba, Ibaraki 305-8568, Japan; orcid.org/0000-0001-6751-2447; Email: y-okamoto@df.jp.nec.com

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c00987

Author Contributions

Y.O. performed all calculations and the analysis of the results.

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

This paper is based on the results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

(1) Balaban, A. T. Application of Graph Theory in Chemistry. J. Chem. Inf. Comput. Sci. 1985, 25, 334–343.
(2) Schultz, H. P. Topological organic chemistry. I. Graph theory and topological indices of alkanes. J. Chem. Inf. Comput. Sci. 1989, 29, 227–228.
(3) Hansen, P. J.; Jurs, P. C. Chemical applications of graph theory. Part I. Fundamentals and topological indices. J. Chem. Educ. 1988, 65, 574–580.
(4) Cao, Y.; Jiang, T.; Girke, T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. Bioinformatics 2008, 24, i366–i374.
(5) Duesbury, E.; Holliday, J. D.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms. MATCH Commun. Math. Comput. Chem. 2017, 77, 213–232.
(6) Borgwardt, K. M.; Ong, C. S.; Schonauer, S.; Vishwanathan, S. V. N.; Smola, A. J.; Kriegel, H.-P. Protein function prediction via graph kernels. Bioinformatics 2005, 21, i47–i56.
(7) Han, L.; Wilson, R. C.; Hancock, E. R. A supergraph-based generative model. 20th International Conference on Pattern Recognition; IEEE Comp. Soc., Istanbul, 2010; pp 1566–1569.
(8) Park, Y.; Reeves, D.; Mulukutla, V.; Sundaravale, B. Fast Malware Classification by Automated Behavioral Graph Matching. Proceedings of the 6th Annual Workshop on Cyber Security and Information Intelligence Research, 2010, article no. 45.
(9) Abu-Khazam, F. N.; Samatova, N. F.; Rzik, M. A.; Langston, M. A. The maximum common subgraph problem: Faster solutions via vertex cover. ACS International Conference on Computer Systems and Applications; IEEE, 2007; pp 367–373.
(10) Pardalos, P. M.; Rodgers, G. P. A branch and bound algorithm for the maximum clique problem. Comput. Oper. Res. 1992, 19, 363–375.
(11) Lucas, A. Ising formulations of many NP problems. Front. Phys. 2014, 2, 5.
(12) Johnson, M. W.; et al. Quantum annealing with manufactured spins. Nature 2011, 473, 194–198.
(13) Goto, H.; Tatsumura, K.; Dixon, A. R. Combinatorial optimization by simulating adiabatic bifurcations in nonlinear Hamiltonian systems. Sci. Adv. 2019, 5, No. eaav2372.
(14) Okuyama, T.; Sonobe, T.; Kawarabayashi, K.; Yamaoka, M. Binary optimization by momentum annealing. Phys. Rev. E 2019, 100, 012111.