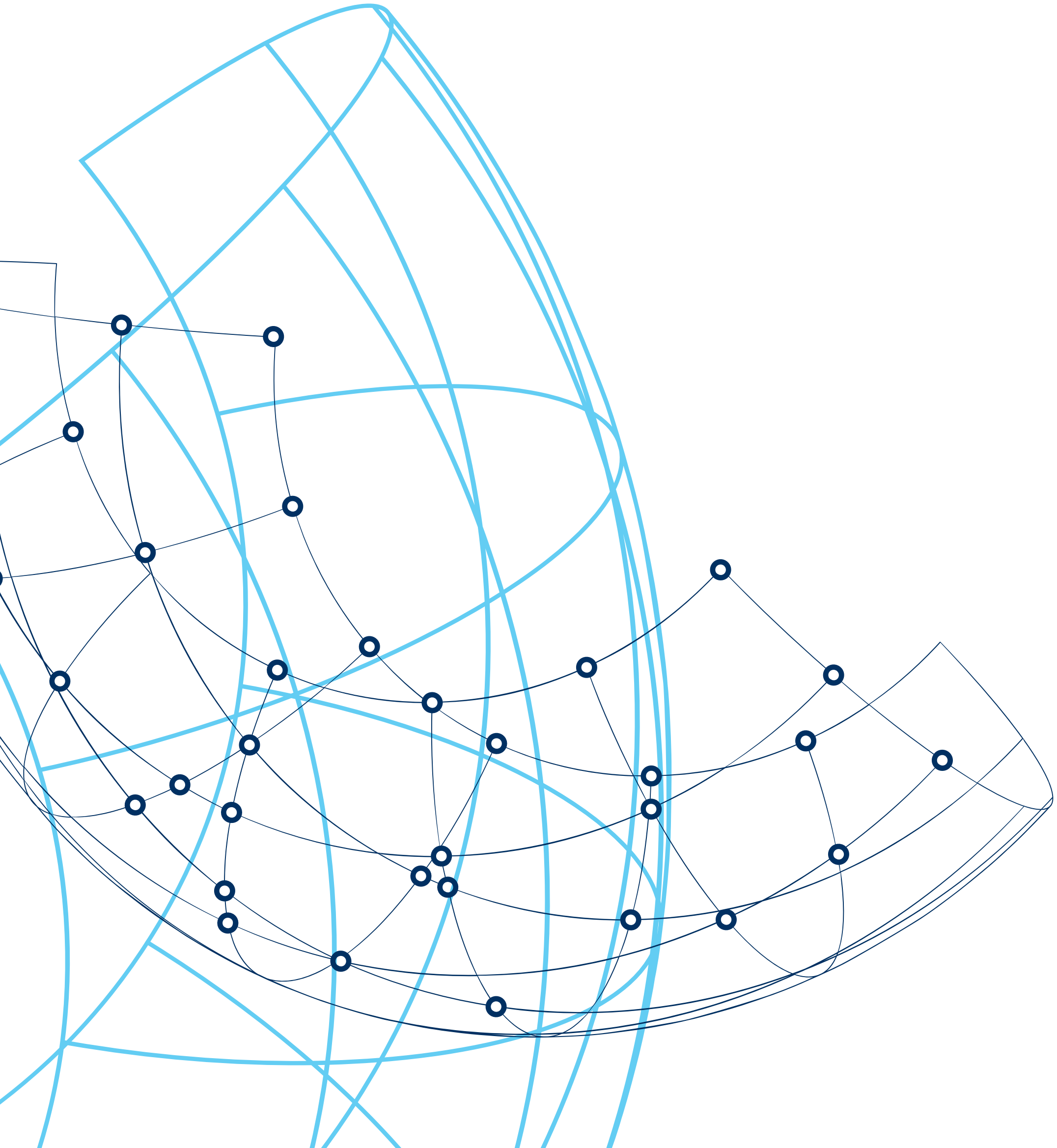
1QBit

Quantum Approaches to Graph Similarity

Measuring similarity among graph objects using quantum adiabatic optimization



White Paper

Quantum Approaches to Graph Similarity

Measuring similarity among graph objects using quantum adiabatic optimization

Maritza Hernandez, Arman Zaribafiyan, Maliheh Aramon,
and Mohammad Naghibi

Abstract

In this paper, we address the problem of measuring the similarity among graph objects. The problem is formulated as a quadratic unconstrained binary optimization that can be solved by a quantum annealer. The goal is to find the maximum co-k-plex of a graph called conflict graph which is induced from the graphs being compared. Our formulation not only quantifies the similarity between graph objects, but it also provides their common subgraph structures. The developed similarity measure is used in the context of molecular similarity to predict the muteganicity of small molecules. Our results show that the developed measure yields a higher prediction accuracy compared to the existing state-of-the-art classical approaches.

Keywords: quantum annealing, graph similarity, molecule similarity

1 Graph Similarity

Graphs have shown promising potentials in representing real-world data objects including social networks, webpages, DNA, and molecules due to their high abstractional power [1]. To obtain valuable insights from graph objects, one of the main challenges is to develop analytical tools and algorithms that can infer new correlations among graphs by measuring their similarity and dissimilarity. An intuitive way to measure the similarity between two graphs is to find their maximum common subgraph (MCS). The MCS of two arbitrary graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is the largest subgraph of G_1 that is isomorphic to a subgraph of G_2. Two graphs are called isomorphic if there is a bijection between their vertex sets such that a mapping between the adjacent pairs of their vertices exists.

Since the MCS problem is in general NP-hard, various classical heuristic algorithms are developed in the literature to approximately measure the similarity between graphs. However, due to recent advances in the ability to utilize quantum mechanical effects in computation, we model the problem of finding the exact MCS between arbitrary graphs as a novel quadratic unconstrained binary optimization (QUBO). The proposed formulation is amenable to a commercial *adiabatic quantum computation* (AQC) device. The machine developed by *D-Wave Systems* is an example of such a device. Before describing our formulation, we first introduce some concepts and definitions.

1.1 Conflict Graph

To formalize the MCS problem, a third graph $G_c = (V_c, E_c)$, called *conflict graph*, is generated. The main purpose of having such a graph is to find the maximal possible one-to-one correspondence $f : V_1 \rightarrow V_2$ between the sets V_1 and V_2 of vertices in graphs G_1 and G_2, respectively. The nodes of the conflict graph are the ordered pairs of vertices (v_i, v_j), where $v_i \in V_1$ and $v_j \in V_2$. There is an edge between two arbitrary nodes (v_i, v_j) and (v_k, v_l) in the conflict graph if and only if having $f(v_i) = v_j$ and $f(v_k) = v_l$ breaks the isomorphism between the two vertex sets.

The usual criteria for adding edges in the conflict graph are the existence of a repeated vertex in the corresponding ordered pairs or different connectivity structure of the mapped vertices. Since vertices and/or edges of graphs of objects in the real-world are associated with different types of information whose levels of importance vary, these criteria can be generalized to find a label-preserving correspondence between graphs. In this case the MCS problem becomes a *labelled maximum weighted common subgraph* (LMWCS) problem. Application-specific criteria for adding or removing edges can also be introduced for labelled graphs.

1.2 Maximum co-k-plex

Since the edges of the conflict graph break the isomorphism, they identify the pairs that cannot be present in the common subgraph simultaneously. Therefore, it is straightforward to see that the MCS of graphs G_1 and G_2 is equivalent to the problem of finding the largest set of vertices in the conflict graph such that there is no edge between all selected pairs, forming the largest conflict-free mapping. This problem is another NP-hard optimization problem know as the *maximum independent set* (MIS) problem [2].

The MIS is an overly restrictive measure because it seeks to find an exact correspondence between the two initial graphs. However, since the real-world's graph objects usually contain noisy data, the definition of similarity can be relaxed by allowing the existence of a few conflict edges among the selected vertices. In this paper, we use a relaxation called the maximum co-k-plex problem, where the goal is to find the largest set of vertices in the conflict graph such that each vertex has at most $k-1$ edges connecting it to the other vertices [2, 3]. The developed formulation of the maximum co-k-plex at 1QBit is

$$\max \left(\sum_{v_i \in V_c} w_{v_i} x_{v_i} - \Big(\sum_{(v_1,\ldots,v_{k+1})} a_{v_1,\ldots,v_{k+1}} \mathcal{A}_{v_1,\ldots,v_{k+1}} \prod_{i=1}^{k+1} x_{v_i} \Big) \right) \tag{1}$$

where x_{v_i} is a binary variable equal to 1 if the vertex v_i is included in the MIS or 0 otherwise, w_{v_i} is the weight of vertex v_i and $a_{v_1,\ldots,v_{k+1}} > \min\{w_{v_1}, \ldots, w_{v_{k+1}}\}$. Furthermore, $\mathcal{A}_{v_1,\ldots,v_{k+1}}$ is a binary parameter equals 1 if the vertex set $(v_1, \ldots, v_{k+1})$ forms a specific subgraph called a star, and 0 otherwise.

For an arbitrary value of k, Formulation (1) becomes a binary polynomial of degree $k+1$. Thus, the maximization problem becomes a higher-order binary optimization (HOBO). Since the available commercial AQC device handles QUBO problems, the degree of our developed polynomial is reduced to two by introducing new auxiliary variables. We use 1QBit implementation of HOBO2QUBO.

1.3 Similarity Measure

The solution to the formulation presented in the previous section is a binary vector with size $|V_c|$. The non-zeros in the solution vector identify the pair of vertices present in the maximum weighted independent set of the conflict graph. To measure the similarity between graphs, we use the metric

$$\mathcal{S}(G_1, G_2) = \delta \max \left\{ \frac{|V_c^1|}{|V_1|}, \frac{|V_c^2|}{|V_2|} \right\} + (1-\delta) \min \left\{ \frac{|V_c^1|}{|V_1|}, \frac{|V_c^2|}{|V_2|} \right\}, \quad \delta \in [0,1], \tag{2}$$

where $|V_c^1|$ and $|V_c^2|$, respectively, denote the number of distinct vertices of G_1 and G_2 in the maximum weighted independent set of the conflict graph. This metric quantifies the contribution of each graph to the MIS. Our metric is the convex combination of two existing similarity measures---Bunk and Shearer, and Asymmetric [4]--- providing the user with more flexibility.

2 Molecular Similarity

The study of small molecules and their properties plays an important role in many aspects of chemical and pharmaceutical research. For example, to develop an effective drug, it is necessary to test whether the drug possesses certain properties like absorption, distribution, metabolism, excretion, and toxicity. Experimental tests to identify different properties in molecules are often expensive and laborious. For this reason, it is important to develop alternative methods of analysis and classification of small molecules. In order to predict properties or functionality in molecules, various molecular similarity measures have been proposed. Determining similarity between molecules is practical due to the *similar property principle* [5], which states that structurally similar molecules are expected to display similar properties. That is to say, properties of chemical molecules can be predicted, to some extent, by establishing and comparing their molecular structure.

The molecular similarity measures can be categorized into two classes: graph-based and fingerprint-based. We discuss each of these measures below.

2.1 Graph-based Molecular Similarity

Graph-based measures are based on the intuitive graph representation of the molecular atom-bond structure where nodes represent single atoms or aromatic rings, and edges represent the chemical bonds between the atoms. Each node and edge label can carry specific properties of the atom and bond, respectively; moreover, the graph itself can have a label carrying some overall properties or characteristics of the respective molecule. At 1QBit, we have developed a tool which finds the reduced graph representations of molecules, builds their associated conflict graph, and quantifies their similarity by solving the QUBO problem formulation discussed in Formulation (1).

Figure 1 illustrates the steps of reducing molecules to graphs. Some examples of conflict rules we define for our specific molecular similarity algorithm include different atomic number, the difference between relative distances of the atoms in two molecules, the type of bonds for each atom, and whether a node is an aromatic ring or an individual atom.

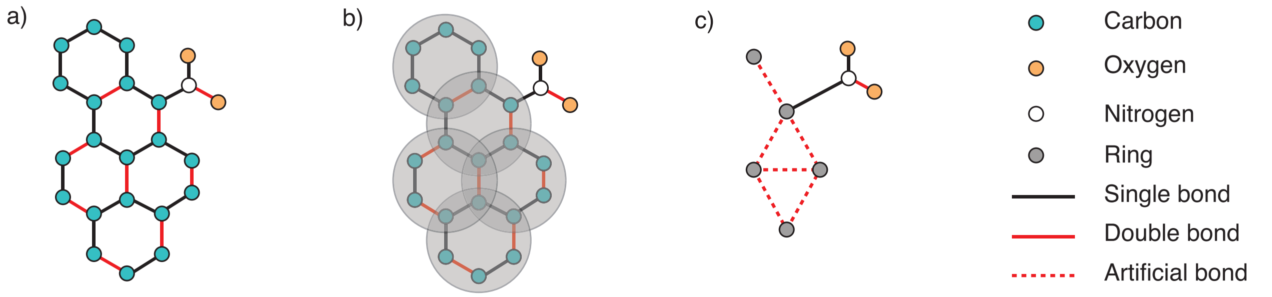


Figure 1: Steps for modelling molecules as graphs: the first step is to a) build a basic structural representation of a molecule, then b) identify ring structures, which subsequently are added as a vertex to the reduced graph as shown in c).

2.2 Fingerprint-based Molecular Similarity

The second class of measures uses a vector-based representation called fingerprint, a conventional concept in chemical informatics and related fields. Fingerprints are binary vectors representing specific substructures in the molecule. Each bit can be either 1 or 0, indicating whether or not the molecule contains an associated substructure with some probability. Several measures to quantify similarity between fingerprint-based representations of molecules have been developed such as Tanimoto, Cosine, Dice, and Euclidean distance [4]. We use Euclidean distance in this paper.

Although fingerprints are easy to use and their pairwise comparisons are computationally efficient, they have certain drawbacks. They cannot be used to assess for certain whether a particular pattern is present or not in a molecular graph [4] and do not usually consider underlying information about the molecular topology.

3 Illustrative Example

Table 1 shows the conflict graphs and the similarity measures of two molecules, ``1,8-Dinitrobenzo[e]pyrene" (Mol 0) and ``6-Nitro-7,8,9,10-tetrahydrobenzo[pqr]tetraphene" (Mol 1), for different parameter values using the tools developed at

1QBit. The details on the optional parameters and d_t can be found in our research paper for the interested readers.

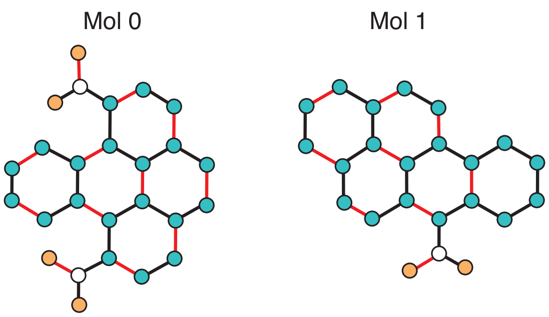


Figure 2: The molecular representations of molecules 0 and 1.

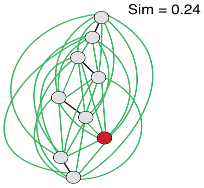
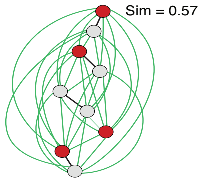
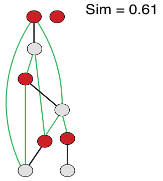
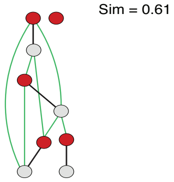
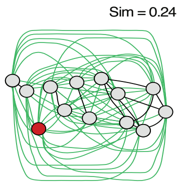
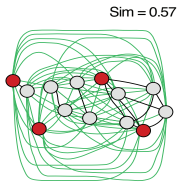
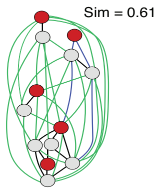
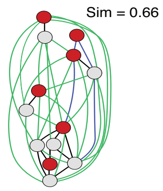
Optional Parameters	$d_t = 0$		$d_t = 1.5$	
	$k = 1$	$k = 4$	$k = 1$	$k = 4$
FC = 1, DN = 1				
RH = 1, FC = 1				

Table 1: The conflict graphs and similarity measures of Mol 0 and Mol 1 for different parameter values.

4 Benchmarking and Computational Results

In order to validate the efficiency of our proposed graph similarity method, we benchmark this method and the fingerprint-based method to classify a specific property in a group of molecules. Put simply, a more appropriate similarity measure leads to a more accurate classification. In particular, we are interested in classifying the mutagenic property of small molecules. Detecting mutagenicity can be critical as an early alert system for potential carcinogenecity in a molecule. A molecule is either mutagenic or non-mutagenic so we can model this as a binary feature and train a binary classifier over a large data set of molecules. We use κ-nearest neighbours (κ-NN) statistical method as the classifier to assign a ``mutagen'' or ``non-mutagen'' label to a molecule. The trained classifier can then be used as an *in silico* approach to predict this property in molecules whose mutagenicity is not known.

We compare the performance of binary classifiers using different similarity measures on two external test sets. The first test set is a balanced set provided by Xu et al. [6], and the second test set is built using the sets given by Hansen et al. [7] and Xue et al. [6]. We use three performance metrics: accuracy (the normalized number of correct predictions), sensitivity (true positive rate), and specificity (true negative rate).

Table 2 shows the three performance metrics of the κ-NN ($\kappa = 3$) classifier on both test sets, where the QUBO-based maximum weighted co-3-plex, the QUBO-based maximum weighted co-1-plex, and the fingerprint-based approaches are used. In Table 2, set 1 and set 2 refer to the first test set and the second test set, respectively. As shown, both QUBO-based maximum weighted co-k-plex methods have better performance than the fingerprint-based method in two metrics---accuracy, and specificity---in both test sets. Although the fingerprint-based method has a higher sensitivity in both sets, its low specificity value implies that the majority of non-mutagenic molecules are blindly labelled as mutagens, increasing the sensitivity. It is worth mentioning that our graph-based maximum weighted co-k-plex

		Method		
Test Set	Performance Metric	QUBO-based Maximum Weighted Co-3-plex	QUBO-based Maximum Weighted Co-1-plex	MACCS Fingerprint
Set 1	Accuracy	**0.81**	0.80	0.76
	Sensitivity	0.84	0.85	0.95
	Specificity	0.78	0.75	0.58
Set 2	Accuracy	**0.80**	0.79	0.77
	Sensitivity	0.83	0.85	0.95
	Specificity	0.77	0.74	0.60

Table 2: Accuracy, sensitivity, and specificity of the 3-NN classifier, where the QUBO-based maximum weighted co-3-plex, the QUBO-based maximum weighted co-1-plex, and the MACCS fingerprint-based approaches are used.

relaxation similarity measures not only yield a higher classification quality, but also provide complete information on the common molecular substructures while the fingerprint-based measure reports only one value as the similarity measure without providing any insights.

Table 2 also shows the superiority of the maximum weighted co-3-plex relaxation method over the maximum weighted co-1-plex relaxation. The higher accuracy of the maximum weighted co-3-plex relaxation method provides evidence that accounting for the noisy data by relaxing the definition of similarity results in a more accurate prediction of mutagenicity.

5 Conclusion

Object comparison, and specifically graph similarity, plays a key role in machine learning and artificial intelligence. Finding common substructures in graph objects is in general an NP-hard problem. Thus, researchers use approximation algorithms to determine similarity. At 1QBit we have developed a method of measuring similarity between graph objects based on the NP-hard common subgraph problem. The NP-hard problem is transformed into a quadratic unconstrained binary optimization problem that can be solved by a quantum annealer. In contrast to conventional methods, our method is capable of identifying the common graph substructures between two or more graphs. The ability to accept user-defined similarity rules makes our method very flexible and allows for its adoption in different contexts. In order to validate our approach we use our method in a biochemical scenario classifying the toxicity properties of a library of molecules based on their similarity to labelled molecules. Our benchmarking results show that our general-purpose similarity determination method results in a higher prediction accuracy than the best-known classical solution tailored for this problem.

References

[1] M. A. Eshera and K-S Fu. An image understanding system using attributed symbolic representation and inexact graph-matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:604--618, 1986.

[2] B. Balasundaram and F. Mahdavi Pajouh. Graph theoretic clique relaxations and applications. In Panos M. Pardalos, Ding-Zhu Du, and Ronald L. Graham, editors, *Handbook of Combinatorial Optimization*, pages 1559--1598. Springer New York, 2013.

[3] B. Balasundaram, S. Butenko, and I. V. Hicks. Clique relaxations in social network analysis: The maximum k-plex problem. *Operations Research*, 59(1):133--142, 2011.

[4] D. Baum. *A Point-Based Algorithm for Multiple 3D Surface Alignment of Drug-Sized Molecules*. PhD thesis, Free University of Berlin, 2007.

[5] Mark A. Johnson and Gerald M. Maggiora. *Concepts and applications of molecular similarity*. Wiley, New York, 1990.

[6] C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P. W. Lee, and Y. Tang. In silico prediction of chemical Ames mutagenicity. *Journal of Chemical Information and Modeling*, 52(11):2840--2847, 2012.

[7] Katja Hansen, Sebastian Mika, Timon Schroeter, Andreas Sutter, Antonius ter Laak, Thomas Steger-Hartmann, Nikolaus Heinrich, and Klaus-Robert Müller. Benchmark data set for in silico prediction of Ames mutagenicity. *Journal of Chemical Information and Modeling*, 49(9):2077--2081, 2009.

White Paper Summary - http://1qbit.com/research/1QBit_Graph_Similarity_white_paper.pdf
Academic Paper - http://arxiv.org/abs/1601.06693
Demonstration Software - http://demo.1qbit.com/

Visit 1QBit.com for more information.