Journal of Systemics, Cybernetics and Informatics (2023) 21(2), 83-86
https://doi.org/10.54808/JSCI.21.02.83

Integrating Quantum Computing into De Novo Metabolite Identification

Li-An TSAI
School of Literature, Media, and Communication, Georgia Institute of Technology
Atlanta, GA 30332, USA

Estelle NUCKELS
Department of Natural Sciences, Middle Georgia State University
Macon, GA 31206, USA

Yingfeng WANG*
Department of Computer Science and Engineering, University of Tennessee at Chattanooga
Chattanooga, TN 37403, USA
*Corresponding Author. Email address: yingfeng-wang@utc.edu

ABSTRACT [1]

Tandem mass spectrometry (MS/MS) is a widely used technology for identifying metabolites. De novo metabolite identification is an identification strategy that does not refer to any spectral or metabolite database. However, this strategy is time-consuming and cannot meet the need for high-throughput metabolite identification. Böcker et al. converted the de novo identification problem into the maximum colorful subtree (MCS) problem. Unfortunately, the MCS problem is NP-hard, which indicates there are no existing efficient exact algorithms. To address this issue, we propose to apply quantum computing to accelerate metabolite identification. Quantum computing performs computations on quantum computers. The recent progress in this area has brought the hope of making some computationally intractable areas trackable, although there are still no general approaches to converting regular computer algorithms into quantum algorithms. Specifically, there is no efficient quantum algorithm for the MCS problem. The MCS problem can be considered as the combination of many maximum spanning tree problems that can be converted into minimum spanning tree problems. This work applies a quantum algorithm designed for the minimum spanning problem to speed up de novo metabolite identification. The possible strategy for further improving the performance is also briefly discussed.

Keywords: Metabolite Identification, Quantum Computing, Algorithm Design, Tandem Mass Spectrometry, Undergraduate Research.

[1] Peer editing and final proofreading for this article by Dr. Yu Liang of the University of Tennessee at Chattanooga.

1. INTRODUCTION

Metabolites are small molecules found in biological samples whose weights are generally less than 1500 Da [1]. Detecting, identifying, and quantifying metabolites is critical in studying metabolic activities. Tandem mass spectrometry (MS/MS) is a popular technology for measuring metabolites [2]. Scientists have developed several computational strategies for identifying metabolites from MS/MS data. Spectral library search is considered the most reliable strategy. This strategy searches a spectral library consisting of annotated MS/MS spectra to match the experimental MS/MS spectrum to a known spectrum. Though widely used, database searches only perform if the structure candidates of metabolites are available. If neither spectrum nor structure candidate is available, users must apply the de novo metabolite identification strategy to identify unknown metabolites. Since de novo metabolite identification can only use the experimental MS/MS spectrum, this strategy remains challenging.

Researchers have developed several de novo identification approaches [3–5]. However, the application of this strategy may be limited by the high time complexity issue [6]. For example, Böcker et al. proposed a method of molecular formula identification, which can be used as the first step of de novo metabolite identification [7]. The method converts the formula identification problem into the NP-hard maximum colorful subtree (MCS) problem. Recently, quantum computing has been applied to bioinformatics for reducing running time and improving performance [8]. These advancements encourage the integration of quantum computing into de novo metabolite identification.

This study proposes to accelerate molecular formula identification by using quantum computing. Böcker et al. suggested that the MCS problem can be considered as a combination of many maximum spanning tree problems [7]. A tree in graph theory is a special connected simple graph. There is exactly one path between each pair of vertices of a tree. Figure 1a shows an example of a tree. For a given connected simple graph, if we can generate a tree by covering all vertices, this tree is a spanning tree. Figure 1b shows an example of a spanning tree. A connected simple graph may have multiple spanning trees. If each edge is assigned a weight, the spanning trees with the largest weight summations are called maximum spanning trees. Figure 2a shows a connected simple graph where each edge of this graph is assigned a weight. Figure 2b shows an example of a maximum spanning tree on the connected simple graph whose weight summation is 17. Similarly, the spanning trees with the smallest weight summations are called minimum spanning trees. Figure 2c shows an example of a minimum spanning tree on the connected simple graph where the weight summation is 10. Aghaei et al. developed a quantum algorithm for finding the minimum spanning tree of the graph [9]. This undergraduate research project simply updates this algorithm for finding maximum spanning tree and applies it to speeding up the existing de novo molecular identification method. An undergraduate student can learn how to search existing methods and apply them to solve a real-world research problem.

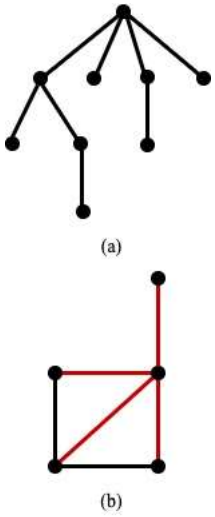

Figure 1. Examples of a tree and a spanning tree in graph theory. (a) An example of a tree with nine vertices and eight edges. (b) The connected simple graph contains a spanning tree consisting of all red edges and all vertices of this graph.

2. RELATED WORKS

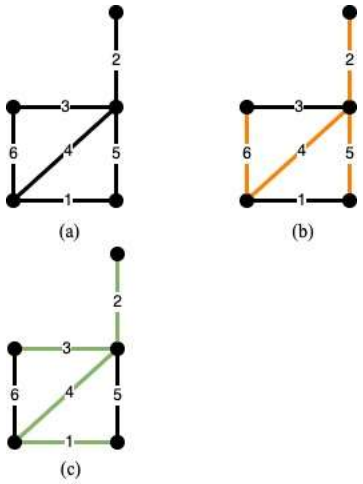

Figure 2. Examples of maximum and minimum spanning trees. The numbers on edges refer to the weights. (a) A connected simple graph. (b) A maximum spanning tree of the graph covers all yellow edges and all vertices with a weight summation of 17. (c) A minimum spanning tree of the graph covers all green edges and all vertices where the weight summation is 10.

De Novo Molecular Formula Identification

De novo molecular formula identification can be used as the first step of de novo metabolite identification. Böcker et al. converted the problem of identifying molecular formulas from MS/MS into the problem of finding a maximum colorful subtree [7] in detail [10]. An MS/MS instrument breaks a metabolite into small fragments. Böcker et al. built a directed acyclic graph for describing the relationships between a possible fragment and its molecular formula. Each vertex on the graph refers to an MS/MS spectrum consisting of some peaks that correspond to possible metabolite fragments. All vertices with the same MS/MS peak share the same color. If any two vertices of a graph (or subgraph) have different colors, this graph (or subgraph) is colorful. If this graph (or subgraph) is a tree (or subtree), this graph (or subgraph) is called a colorful tree (or subtree). If a formula, represented by vertex A, is the subset of the formula of vertex B, there is a directed edge from B to A. This edge suggests that the fragments with the formula represented by vertex A can be generated based on the fragments with the formula represented by vertex B. Böcker et al. developed an approach for scoring the match between a peak and its possible molecular formula. This score is the weight of the incoming edge of the vertex representing the formula. For example, the weight of edge BA is the score of the match between the peak and the formula related to vertex A. The goal is to find a colorful subtree with the maximum edge weight, in essence, the maximum colorful subtree problem. Böcker et al. proved

that this is an NP-hard problem, although there are some approximation algorithms. One brute-force exact solution is to compare the weights of all colorful subtrees. As for each combination of colorful vertices, this solution needs to find the maximum spanning tree, which includes all vertices and has the maximum weight. This paper attempts to accelerate this solution by using quantum computing.

The Quantum Algorithm of Finding Minimum Spanning Tree

Grover developed a quantum search algorithm for unconstructed search [11]. The time complexity of searching a number on an unsorted N-element array is $O(N)$ on a classical computer. Applying Grover's algorithm finishes the search with time complexity $O(\sqrt{N})$ on a quantum computer. Aghaei et al. applied Grover's quantum algorithm to find the minimum spanning tree on a quantum computer [9]. This quantum algorithm can also lower the time complexity from $O(N)$ on a classical computer to $O(\sqrt{N})$ on a quantum computer. Our study converts this algorithm to find the maximum spanning tree and applies it to accelerate the de novo metabolite molecular identification.

3. METHOD AND RESEARCH DESIGN

The goal of this undergraduate research project is to guide undergraduate students to develop a quantum computing algorithm for speeding up de novo metabolite identification. This project is divided into three steps: (1) identify research opportunities by literature review, (2) identify existing tools, and (3) develop the solution based on existing tools.

Research Opportunity Identification

In the first step, the existing framework of de novo molecular formula identification proposed by Böcker et al. [7] is identified. The outcome of this framework can be applied in de novo metabolite identification. Unfortunately, the exact algorithm of this framework is NP-hard. The high time complexity issue could be a concern of adopting this framework in real-world metabolite identification applications. Therefore, we decided to start with this framework and address the low-speed issue by using quantum computing.

Existing Tool Identification

It is challenging for undergraduate students to develop a new quantum algorithm for solving the maximum colorful subtree problem or the maximum spanning tree problem. To fit the background of undergraduate students, this project attempts to look for related works by literature review. Although quantum algorithms for these two problems were not found, Aghaei et al. developed the algorithm for finding a minimum spanning tree, which is close to the problem this project focuses on.

Solution Development Based on Existing Tools

After finishing the above two steps, the concept is to adapt the maximum spanning tree problem on metabolites into a minimum spanning tree problem. The conversion is completed by multiplying all weights of the graph of interest by (-1) and can be implemented with linear time complexity. Therefore, the approach developed by Aghaei et al. can be applied to the converted dataset to find the maximum edge weight and subsequently its corresponding molecular formula. Figure 3 outlines the idea of converting algorithms for solving the molecular formula identification problem.

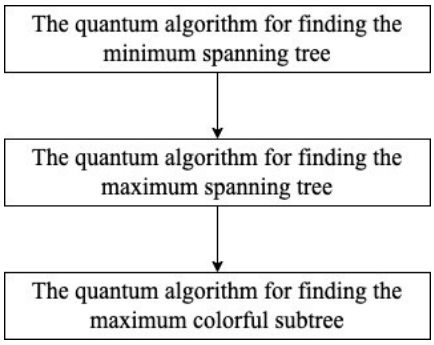

Figure 3. The idea of converting the algorithms for solving the molecular formula identification problem.

4. DISCUSSION

The study addresses the time complexity issue of de novo molecular identification by using quantum computing. Designing quantum algorithms requires expertise in both classical algorithm design and quantum computing characterizations. Our research design lowers the barrier of familiarity with quantum computing characterizations by using existing quantum algorithms. It allows undergraduate students to gain experience in quantum algorithm development and achieve some success. However, this design is limited from fully utilizing the power of quantum computing in this problem. It motivates us to develop a new path to develop quantum algorithms in this area in our future work.

A possible strategy for further improving the performance is to design a fixed-parameter quantum tractable (FPQT) algorithm. Böcker et al. proposed a fixed-parameter tractable (FPT) method for the MCS problem [7]. If we fix the parameters of this FPT algorithm at constant values, the MCS problem can be solved in polynomial time. Although it does not change its NP-hard nature, the

FPT framework provides efficient solutions at some specific scenarios. Under the quantum computing context, the FPT framework can be extended to the FPQT framework [12]. Therefore, a FPQT algorithm for the MCS problem may take the advantage of the FPT framework and quantum computing.

5. CONCLUSIONS AND FUTURE WORK

This paper presents our undergraduate research project that integrates quantum computing into the first step of de novo metabolite formula identification. This project converts an existing quantum algorithm for finding the minimum spanning tree into finding the maximum spanning tree, and then applies the converted algorithm in a brute force approach for finding the maximum colorful subtree. We will design more efficient quantum algorithms for the maximum colorful subtree problem in our future work.

6. ACKNOWLEDGMENT

We thank the support of NSF RUI #2053286 and REU #1852042 and #2149956, and the internal support from the Department of Computer Science and Engineering and the Office of the Vice Chancellor of Research at the University of Tennessee at Chattanooga.

7. REFERENCES

[1] W. B. Dunn, "Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes," Physical biology, Jan. 2008, doi: 10.1088/1478-3975/5/1/011001.

[2] E. M. Harrieder, F. Kretschmer, S. Böcker, and M. Witting, "Current state-of-the-art of separation methods used in LC-MS based metabolomics and lipidomics," Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences, vol. 1188, 2022, doi: 10.1016/j.jchromb.2021.123069.

[3] J. Guo, S. Shen, S. Xing, H. Yu, and T. Huan, "ISFrag: De novo recognition of in-source fragments for liquid chromatography-mass spectrometry data," Analytical Chemistry, vol. 93, pp. 10243–10250, 2021, doi: 10.1021/acs.analchem.1c01644.

[4] J. E. Peironcely et al., "Automated pipeline for de novo metabolite identification using mass-spectrometry-based metabolomics," Analytical chemistry, vol. 85, pp. 3576–83, Apr. 2013, doi: 10.1021/ac303218u.

[5] A. D. Shrivastava, N. Swainston, S. Samanta, I. Roberts, M. W. Muelas, and D. B. Kell, "Massgenie: A transformer-based deep learning method for identifying small molecules from their mass spectra," Biomolecules, vol. 11, 2021, doi: 10.3390/biom11121793.

[6] R. R. da Silva, P. C. Dorrestein, and R. A. Quinn, "Illuminating the dark matter in metabolomics," Proceedings of the National Academy of Sciences, vol. 112, pp. 12549–12550, 2015, doi: 10.1073/pnas.1516878112.

[7] S. Böcker and F. Rasche, "Towards de novo identification of metabolites by analyzing tandem mass spectra," Bioinformatics, vol. 24, pp. 49–55, 2008, doi: 10.1093/bioinformatics/btn270.

[8] A. K. Fedorov and M. S. Gelfand, "Towards practical applications in quantum computational biology," Nat Comput Sci, vol. 1, no. 2, pp. 114–119, 2021, doi: 10.1038/s43588-021-00024-z.

[9] I. Rauf, F. Rasche, F. Nicolas, and S. Böcker, "Finding maximum colorful subtrees in practice," J Comput Biol, vol. 20, no. 4, pp. 311–321, 2013, doi: 10.1089/cmb.2012.0083.

[10] L. K. Grover, "A fast quantum mechanical algorithm for database search," in Proceedings of the 28th Annual ACM Symposium on the Theory of Computing (STOC), 1996, pp. 212–219.

[11] M. R. S. Aghaei, Z. A. Zukarnain, A. Mamat, and H. Zainuddin, "A quantum algorithm for minimal spanning tree," in Proceedings of the International Symposium on Information Technology (ITSim), 2008, pp. 1–6. doi: 10.1109/itsim.2008.4632038.

[12] M. J. Bremner, Z. Ji, R.L. Mann, L. Mathieson, M.E.S. Morales, and A.T.E. Shaw (2022) Quantum parameterized complexity. Arxiv. https://doi.org/10.48550/arxiv.2203.08002