MapLE: Matching Molecular Analogues Promptly with Low Computational Resources by Multi-Metrics Evaluation (Student Abstract)

Xiaojian Chen[1, 2], Chuyue Liao[1], Yanhui Gu[1], Yafei Li[3], Jinlan Wang[4], Yi Chen[1], Masaru Kitsuregawa[5]

[1]School of Computer and Electronic Information Science, Nanjing Normal University
[2]Department of Biomedical Engineering, Johns Hopkins University
[3]School of Chemistry and Materials Science, Nanjing Normal University
[4]School of Physics, Southeast University
[5]Institute of Industrial Science, The University of Tokyo

Abstract

Matching molecular analogues is a computational chemistry and bioinformatics research issue which is used to identify molecules that are structurally or functionally similar to a target molecule. Recent studies on matching analogous molecules have predominantly concentrated on enhancing effectiveness, often sidelining computational efficiency, particularly in contexts of low computational resources. This oversight poses challenges in many real applications (e.g., drug discovery, catalyst generation and so forth). To tackle this issue, we propose a general strategy named ***MapLE***, aiming to promptly match analogous molecules with low computational resources by multi-metrics evaluation. Experimental evaluation conducted on a public biomolecular dataset validates the excellent and efficient performance of the proposed strategy.

Introduction

Matching analogues involves the identification of molecules that resemble original compounds based on their chemical, pharmacological, and structural characteristics. When applied to drug discovery, this procedure is of paramount importance, as it could considerably hasten the investigation of potential therapeutic agents. Indeed, there are numerous untapped opportunities for drug discovery originating from natural products (DeCorte 2016).

Most recent studies on matching analogues have primarily focused on improving similarity screening accuracy rather than on time efficiency optimization (Chen et al. 2023). Nonetheless, current methodologies tend to slow down when employed with fully enumerated chemical libraries, which may contain billions of compounds.They often become impractical due to the substantial computational resources required (Sadybekov and Katritch 2023). Warr et al. introduced Arthor, utilizing the RoundTable algorithm, which could search for patterns over a billion molecules in a few seconds (Warr et al. 2022). However, search time scales with database size, and the vast growth of chemical space may pose challenges.

To address these challenges, we introduce a general framework that incorporates several efficient strategies. Specifically, we perform multiple feature extraction on processed molecular objects and a progressive evaluation strategy to match the analogs promptly. By integrating data from these multiple features, our framework results in competitive evaluation performance and a better understanding of the impact of various information sources. In addition, we streamline the molecule accesses in the process of molecule matching by a progressive prompt evaluation, ultimately reducing the execution time and computational resources.

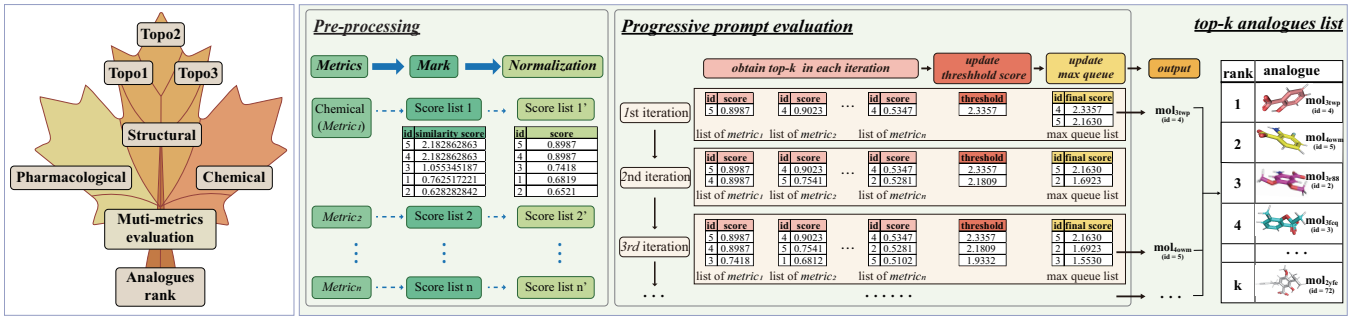
Proposed Strategy

As shown in Figure 1, we introduce ***MapLE***—a general framework for **M**atching molecular **a**nalogues **p**romptly with **L**ow computational resources by multi-metrics **E**valuation—which integrates multiple similarity metrics and introduces a progressive prompt evaluation technique tailored to speed up the screening process.

Multi-metrics Fusion. Our approach intuitively considers multiple attributes of a molecule and synthesize them into a cohesive evaluation. As depicted in Figure 1 (a), our general framework is hierarchical, incorporating various features of the molecule. Specifically, we rank the similarity among molecules within a set by considering pharmacological, structural, and chemical features in a multi-metrics evaluation. Moreover, within the category of structural features, we further employ sub-metrics (e.g., $topo1$, $topo2$, and $topo3$) to analyze the molecule from multiple topological perspectives (Kim et al. 2023).

To efficiently manage these features, we construct a set of inverted lists. It is important to note that each new molecule is broken down into several feature indices. When adding a new molecule to the database, we only need to update the lists containing these specific indices. Thereby, this approach minimizes irrelevant traversal queries, thus saving time, especially when dealing with large databases.

Progressive Prompt Evaluation. Inspired by the search for the top-k semantically similar sentences in the field of Natural Language Processing (NLP), we build a mapping between molecular structures and word sentences. Specifically, we consider the substructures or features of a molecule analogous to the words in sentences. We outline our strategy for measuring the similarity between query molecules Q and



(a) The overall framework (b) The procedure of our strategy

Figure 1: The general framework of ***MapLE***.

the top-k candidate similar molecule L, as follows:

$$\widetilde{Sim}_{metric}(L,Q) = \frac{\sum_{i=1}^{\delta}(f_L(i) \wedge f_Q(i))}{\sum_{i=1}^{\delta}(f_L(i) \vee f_Q(i))} \quad (1)$$

where $f(i)$ is the frequency of the i-th feature in the molecule, δ is the size of the whole molecular features, and we use the tanimoto coefficient to compute similarity under a single metric of L and Q as $Sim_{metric}(L,Q)$. To consolidate these different metrics of similarity, we present the similarity as:

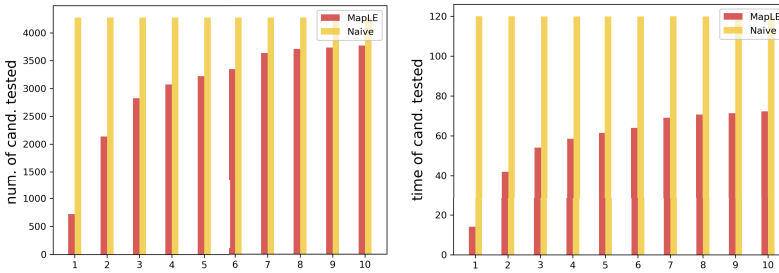
$$\widetilde{Sim}(L,Q) = \sum w_j \cdot \widetilde{Sim}_j(L,Q) \quad (2)$$

where $\widetilde{Sim}_j$ is measured under a specific metric similarity, and w_j denotes the corresponding weight of the metric.

We propose an efficient strategy to assemble the on-the-fly. We denote that $\widetilde{Sim}_j^{\,top-k}$ is the normalized similarity score of the top-k molecule in the metric j, and the t is the threshold where $t = \sum w_j \cdot \widetilde{Sim}_j^{\,top-k}$. We will progressively output the top-k result when a candidate molecule L meet the condition that $\widetilde{Sim}(L,Q) \geq t$, since it is at least for one metric that $\widetilde{Sim}_j(P,Q) \geq \widetilde{Sim}_j^{\,top-k}$. It means we only matching the top of molecule list rather than traversal the whole list, which will save a lot query time to get top-k analogues. More details shown in Figure 1 (b) make a example of our strategy.

Experiments

We conducted our experiments on the CASF-2016 dataset, a widely-used biomolecular dataset with thousands of high-quality molecular structures. As shown in Figure 2 (a), the baseline accesses all the ligand molecules in the dataset, with the query time remaining consistent regardless of the value of k. Notably, the top-1 values are returned almost instantly, and as k increases, a progressively greater number of results are obtained. In comparison, our proposed method showcases its efficiency, particularly with smaller values of k. Additionally, we monitored the number of accessed candidate values. Figure 2 (b) details the count of candidates accessed during the data collection phase.



(a) #Candidates accessed (b) Query time of different k-value

Figure 2: The evaluation result of the ***MapLE***.

Conclusion

In this paper, we introduce a general framework for matching molecular analogues based on multi-metrics evaluation. This framework incorporates several strategies that are both time-efficient and computational resource-saving. Our experimental evaluation, conducted on a real bio-molecular dataset, attests to the framework's effectiveness. In the future, we aim to validate the performance of various metrics and assess their impact on the results, thereby enhancing the interpretability of the framework.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No.92370127 and No.22033002).

References

Chen, Q.; Li, X.; Geng, K.; and Wang, M. 2023. Context-Aware Safe Medication Recommendations with Molecular Graph and DDI Graph Embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6): 7053–7060.

DeCorte, B. L. 2016. Underexplored Opportunities for Natural Products in Drug Discovery. *Journal of Medicinal Chemistry*, 59(20): 9295–9304.

Kim, S.; Lee, D.; Kang, S.; Lee, S.; and Yu, H. 2023. Learning Topology-Specific Experts for Molecular Property Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7): 8291–8299.

Sadybekov, A. V.; and Katritch, V. 2023. Computational Approaches Streamlining Drug Discovery. *Nature*, 616(7958): 673—685.

Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; and Rarey, M. 2022. Exploration of Ultralarge Compound Collections for Drug Discovery. *Journal of Chemical Information and Modeling*, 62(9): 2021–2034. PMID: 35421301.