Chemical Similarity Based on Map Edit Distance

Xin Li, Xiaoqing Lyu, Zhi Tang
Peking University
Beijing, China
{l_x, lvxiaoqing, tangzhi}@pku.edu.cn

Hao Zhang
Beijing Institute of Technology
Beijing, China
gcrth@outlook.com

***Abstract*—Quantifying the similarity of compounds is a particularly important problem, as it is able to give rise to many new possibilities for understanding the relationship between molecules. Here we develop a novel algorithm for computing the Molecular similarity. Experimental results on a broad range of real world graphs and tasks show that the proposed similarity measure is very effective.**

***Index Terms*—Chemical Similarity; Graph Edit Distance; Drug Discovery; Molecular Graph**

I. Introduction

Providing a measure of similarity between chemical compounds is a central task in cheminformatics [1] [2]. Based on the assumption that similar molecules are more likely to have similar biological or physicochemical properties than dissimilar ones [3], similarity measures are employed by diverse applications in drug discovery, particularly in ligand-based virtual screening and medicinal chemistry, but also in toxicology, chemogenomics, and pharmacology.

Graphs offer a natural way of representing chemical structures. In this paper, we introduce a novel measure of similarity between molecular graphs, which is an approximation of the minimum number of edit operation costs required to transform one molecule into another and it is more robust compared to other tolerant measures of graphs such as maximum common subgraph [4] measures based on relaxation labeling [5], measures based on spectral methods [6] and measures based on graph kernel [7].

Intuitively, by proper definition of the operation costs, we can integrate error tolerance into different tasks. As in bioinformatics, the ability to parameterize structural similarity by using weighted edit operations [8] allows our methods to outperform existing measures such as Tanimoto coefficients of binary fingerprints.

II. Methods

A. Molecular graph

In this paper, a molecule is represented by an undirected graph $G = (V, E, l)$ without self-loops, where V is the finite set of vertices associated with the atoms, E is the set of edges which correspond to bonds between atoms.

A graph can be transformed into another graph by elementary edit operations including insertion, deletion and relabeling of a/an vertex/edge. Each elementary edit operation comes with a non-negative **operation cost**. The costs of operations on vertices and edges are defined as $c_V : \Sigma_V \times \Sigma_V \to \mathbb{R}_{\geq 0}$ and $c_E : \Sigma_E \times \Sigma_E \to \mathbb{R}_{\geq 0}$. Σ_V and Σ_E both contain a special label ϵ. A vertex or an edge deletion can be considered as a relabeling from $\sigma \in \Sigma_V \cup \Sigma_E \setminus \{\epsilon\}$ to ϵ. Symmetrically, a vertex or an edge insertion can be considered as a relabeling from ϵ to $\sigma \in \Sigma_V \cup \Sigma_E \setminus \{\epsilon\}$.

Let G and H be two graphs, which satisfy $|V^G| = n$ and $|V^H| = m$. Suppose that $V_1 = V^G \cup \{\epsilon_1, ..., \epsilon_m\}$, $V_2 = V^H \cup \{\epsilon_1, ..., \epsilon_n\}$. Each matching pairs between V_1 and V_2 corresponds either with a vertex substitution (v, u), deletion (v, ϵ), insertion (ϵ, u) or empty substitution [9], where $v \in V^G$ and $u \in V^H$. When the matching pairs are determined, one way to transform G to H is also determined, hence the cost is determined. It is obvious that an optimal match associated with a minimum cost exists. But the number of mappings from V_1 to V_2 is exponential in the number of vertices. Our idea is to regard a transforming cost that is as low as possible as a dissimilarity (similarity) measure and a iterative method is proposed.

B. Algorithm

The top-level framework to compute the similarity between two molecular graphs is outlined in Algorithm 1 as described below. In the beginning, a map is randomly initialized in line 1 and its associated cost is calculated in line 2.

Now we introduce the local search step (lines 3-11). The basic strategy is to rematch the current pairs (line 4), then to calculate the associated cost (line 5) and compare it with the current optimal value (lines 6-10). Loop will terminate after finding a local optimum. In other words, we can't get a better solution by rematching any two pairs, which means the current cost converges to a local optimum.

In Algorithm 1, each cost is associated with a mapping. The set of those mappings is denoted as Ψ. Formally, the solution returned by Algorithm 1 is defined in Definition 1.

Definition 1: **Map edit distance (MED).** The map edit distance is defined as

$$MED(G, H) = \min_{s \in \Psi} c_s, \tag{1}$$

C. Comparison between graph edit distance (GED) and MED

A sequence of edit operations that transforms a graph G into another graph H is called an edit path between G and H. The edit cost of an edit path is the sum of all the costs of edit operations in the edit path. The minimum edit cost is called graph edit distance (GED). GED is a widely used similarity

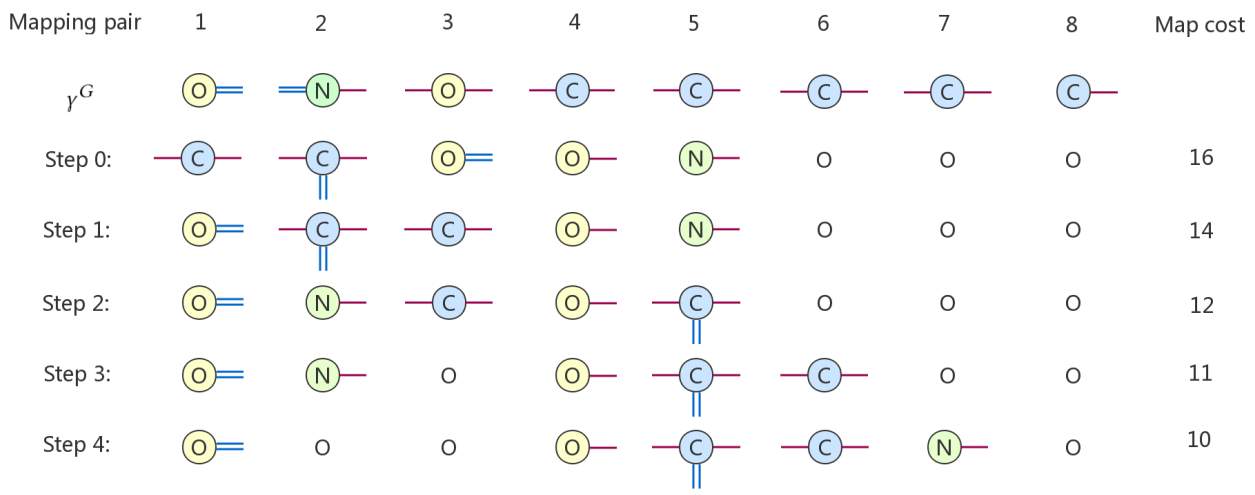

Fig. 1. An example of the proposed approach.

measure for graphs. MED is a local optimum and it is a upper bound of GED. In some cases, MED equals to GED. In terms of measuring the similarity between molecular graphs, MED has the following advantages over GED:

- The exact computation of GED is NP-hard [10]. The dramatic increasing computational time make it impossible to use for large graphs. Since MED don't require enumerations of all the mappings, the solution can be returned in much less time and MED can be used for large graphs.
- GED focuses overly on global topological structure similarity. However, some local molecular structures are more important since they are known to be responsible for activity in chemistry. Comparing to GED, MED focuses more on local structure, which will help to capture more local similarity.

Algorithm 1 MED (G, H)

Require: G, H
Ensure: suboptimal graph edit distance

```
initialize mapping with G and H;
compute initialized cost c*;
while c* not suboptimal do
    rematch the pairs (v1, u1) and (v2, u2) to (v1, u2) and (v2, u1);
    compute the new cost c_s' of the new mapping;
    if c_s' < c* then
        c* = c_s';
    else
        recover the pairs (v1, u2) and (v2, u1) to (v1, u1) and (v2, u2);
    end if
end while
return c*;
```

III. Experiments

The purpose of the experiments presented in this section is to empirically verify the feasibility of the proposed similarity measure. For the application for cheminformatics, we consider two tasks: chemical pattern recognition and similarity search in Sections III-A and III-B.

A. Chemical Pattern Recognition

In this paper, we utilize k-Nearest Neighbors (kNN) classifier to show the effectiveness of the proposed MED for graph classification tasks in the area of cheminformatics since it can be directly used without any additional training.

Data. We conducted experiments for pattern recognition tasks on four data sets. The MUTAG data set consists of 188 aromatic and heteroaromatic chemical compounds divided into two classes according to their mutagenic effect on a bacterium. The Enzymes is a data set of protein tertiary structures consisting of 600 enzymes from the BRENDA enzyme database. The NCI1 and NCI109 data set represent two balanced subsets of data sets of chemical compounds screened for activity against non-small cell lung cancer and ovarian cancer cell lines respectively.

Results. To investigate the effectiveness of tackling the graph classification tasks in the area of cheminformatics, We compared the developed classifier with other two types of methods. One type of these methods is to make a kNN classifer by utilizing other graph edit distances including BR [11] and HY [12]. Our methods were also compared against SVM classifiers using different kinds of graph kernels, including WL [13], WL-Edge [7], SP [14], Graphlet [7], pRW [15] and MLG [16]. As shown in the TABLE I, the developed classifier achieves higher accuracy on the MUTAG and Enzymes but gets average results on NCI1 and NCI109. One possible reason is that the proposed similarity measure is good at capturing

TABLE I
CLASSIFICATION RESULTS (MEAN ACCURACY ± STANDARD DEVIATION)

kernel / dataset	MUTAG	ENZYMES	NCI1	NCI109
MED-kNN	**87.37(±5.89)**	**58.42(±4.84)**	72.52(±1.35)	72.43(±1.33)
BR-kNN	82.24(±5.04)	50.17(±4.07)	70.94(±1.64)	70.92(±1.98)
HY-kNN	60.66(±14.47)	TIMED OUT	TIMED OUT	TIMED OUT
WL	84.50(±2.16)	53.75(±1.37)	**84.76(±0.32)**	85.12(±0.29)
WL-Edge	82.94(±2.33)	52.00(±0.72)	84.65(±0.25)	**85.32(±0.34)**
SP	85.50(±2.50)	42.31(±1.37)	73.61(±0.36)	73.23(±0.26)
Graphlet	82.44(±1.29)	30.95(±0.73)	62.40(±0.27)	62.35(±0.28)
p–RW	80.33(±1.35)	28.17(±0.76)	TIMED OUT	TIMED OUT
MLG	84.21(±2.61)	57.92(±5.39)	80.83(±1.29)	81.30(±0.80)

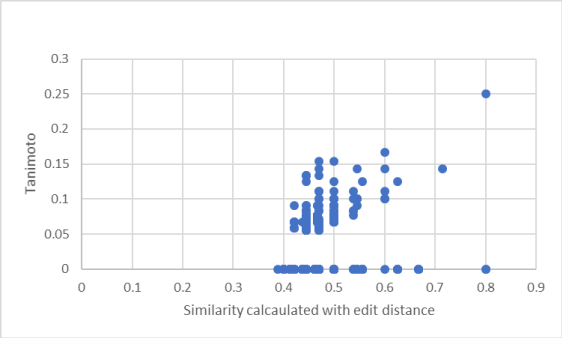

Fig. 2. Tanimoto vs MED similarity

global similarity of graphs, which is useful for the whole-compound view of similarity characteristic of the perspective of chemoinformatics [2]. To be specific, computing MED requires global consideration since it is implicitly associated with one way to transform one graph into another where the transfomring cost is quite low.

B. Chemical Structure Similarity Search

A similarity function which can be constructed with the MED to obtain a numerical value that quantifies the similarity between molecular structures is defined as

$$sim(G,H) = \frac{1}{MED(G,H)/(|V^G| + |V^H|) + 1}. \quad (2)$$

Fingerprint similarity searching is one of the most popular molecular searching method. In cheminformatics, the Tanimoto coefficient (Tc) is one of the most popular similarity function. We compared our similarity measure using 2 with ECFP4 using Tc. As shown in Figure 2, overall, the proposed similarity measure behaves similarly to Tanimoto while the presented similarity may be more robust since it can overcome the limitation that a lot of zeros are produced by the Tanimoto which may result in the lack of searching results in some situations.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grant 61876003. It is a research achievement of Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

REFERENCES

[1] A. Bender and R. C. Glen, "Molecular similarity: a key technique in molecular informatics," *Org. Biomol. Chem.*, vol. 2, pp. 3204–3218, 2004. [Online]. Available: http://dx.doi.org/10.1039/B409813G

[2] G. Maggiora, M. Vogt, D. Stumpfe, and J. Bajorath, "Molecular similarity in medicinal chemistry," *Journal of Medicinal Chemistry*, vol. 57, no. 8, pp. 3186–3204, 2014, pMID: 24151987. [Online]. Available: https://doi.org/10.1021/jm401411z

[3] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies, "How similar are similarity searching methods? a principal component analysis of molecular descriptor space," *Journal of Chemical Information and Modeling*, vol. 49, no. 1, pp. 108–119, 2009, pMID: 19123924. [Online]. Available: https://doi.org/10.1021/ci800249s

[4] G. Levi, "A note on the derivation of maximal common subgraphs of two directed or undirected graphs," *Calcolo*, vol. 9, no. 4, p. 341, 1973.

[5] R. Myers, R. Wison, and E. R. Hancock, "Bayesian graph edit distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 6, pp. 628–635, 2000.

[6] A. Robles-Kelly and E. R. Hancock, "A riemannian approach to graph embedding," *Pattern Recognition*, vol. 40, no. 3, pp. 1042–1056, 2007.

[7] N. Shervashidze, S. Vishwanathan, T. Petri, K. Mehlhorn, and K. Borgwardt, "Efficient graphlet kernels for large graph comparison," in *Artificial Intelligence and Statistics*, 2009, pp. 488–495.

[8] G. Harper, G. S. Bravi, S. D. Pickett, J. Hussain, and D. V. S. Green, "The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 6, pp. 2145–2156, 2004, pMID: 15554685. [Online]. Available: https://doi.org/10.1021/ci049860f

[9] K. Riesen and H. Bunke, "Approximate graph edit distance computation by means of bipartite graph matching," *Image and Vision computing*, vol. 27, no. 7, pp. 950–959, 2009.

[10] Z. Zeng, A. K. Tung, J. Wang, J. Feng, and L. Zhou, "Comparing stars: On approximating graph edit distance," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 25–36, 2009.

[11] D. B. Blumenthal and J. Gamper, "Improved lower bounds for graph edit distance," *IEEE Transactions on Knowledge & Data Engineering*, vol. 30, no. 3, pp. 503–516, 2018.

[12] W. Zheng, L. Zou, X. Lian, D. Wang, and D. Zhao, "Efficient graph similarity search over large graph databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 964–978, 2015.

[13] N. Shervashidze, P. Schweitzer, E. J. v. Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels," *Journal of Machine Learning Research*, vol. 12, no. Sep, pp. 2539–2561, 2011.

[14] K. M. Borgwardt and H.-P. Kriegel, "Shortest-path kernels on graphs," in *Fifth IEEE international conference on data mining (ICDM'05)*. IEEE, 2005, pp. 8–pp.

[15] T. Gartner, "Exponential and geometric kernels for graphs," in *NIPS^* 02 Workshop on Unreal Data: Principles of Modeling Nonvectorial Data*, 2002.

[16] R. Kondor and H. Pan, "The multiscale laplacian graph kernel," in *Advances in Neural Information Processing Systems*, 2016, pp. 2990–2998.