Southeast-Asian J. of Sciences Vol. 2, No. 2 (2013) pp. 160-164

MOLECULAR SIMILARITY SEARCHING USING MEANING-FACTOR REDUCED GRAPH

Somnuek Worawiset[*] and **Tawun Remsungnen**[†]

[*] *Department of Mathematics, Faculty of Science*
Khon Kaen University, Khon Kaen Thailand
e-mail: wsomnu@kku.ac.th

[†] *Faculty of Applied Science and Engineering*
Nong Khai Campus, Khon Kaen Univ., Nong Khai, Thailand
e-mail: rtawun@kku.ac.th

Abstract

Molecular similarity searching from molecular SMILES format using score base on reduced graphs with meaning factor has been proposed. The effectiveness of the method at is compared with searching using Tanimoto score. The searches have been carried out to find the potential drugs or molecules that form interactions by docking with Cov-Proteinase receptor which is SARS virus protein. Since the new proposed method yield almost higher scores than those obtained from Tanimoto method, the structurally diverse sets of active molecules are more retrieved.

1 Introduction

The graph theory is a subject of mathematics that has found a variety of applications such as the optimization of communication and transport networks, the design of electrical circuits, the synchronization of interacting oscillators with different topologies, the analysis of social networks, etc [1]. The ligand-based virtual screening is searching method for novel drug lead compounds by docking them to the 3D structure of the target protein and obtain estimated binding affinities. In order to avoid huge computation times for all possible ligand-

Key words: molecular similarity searching, reduced graph, molecular docking, Cov-Protease, SARS virus, SMILES.

traget calculations, ranking or classifying molecules in a database according to their similarity to known active and inactive molecules are useful.

In this work, the refinement of Tanimoto' coefficient score (TCS), so called RTCS is proposed. It is based on reduced graphs which are derived from the SMILES strings with the proposed realiable complete fragments of rings and functional groups instead of the conventional breakable fragments. The comparison between RTCS and original TCS are obtained by searching which carried out to find the potential drugs or molecules that form interactions by docking with Cov-Proteinase receptor in SARS virus protein.

2 Models and calculations

2.1 Molecular structure, SMILES string and reduced molecular graph.

The undirected labeled graphs can be used to represent topology of chemical compounds, where edge is labeled by bond properties like bond order or bond length, while node is labeled by atom properties, like atom type, partial charge, functional group, ring, etc. Some graph mining methods then can be used to deal with molecular structures base on a question as how the similarity of two or more molecular graphs can be precise and high-throughput obtained [2]. Thus, fundamental considerations of molecular-similarity concepts are likely to be as important as the design of novel computational approaches [3, 4].

The simplified molecular input line entry system (SMILES) is a chemical notation language specifically designed based on mathematical graph theory for computer use [5, 6]. This language can transform 3D or 2D structures into 1D as a string, so all the kernels that have been developed for sequences can be also applied to the SMILES strings too [7]. Figure 1, shows the transformation of molecular structure to labeled graph and reduced labeled graph, respectively. Then the fragments of molecular structure are represented by labeled nodes of the reduced graph and can transform to substring of molecular SMILES which can be counted in refined Tanimoto score.

2.2 Molecular similarity scores.

There are some proposed molecular similarity scores [8], however, the Tanimoto coefficient score (TCS) is commonly used and is used as improvement or refinement base score. This score is calculated based on the presence of common substructure (subgraph) between two interested molecules as follows:

$$TCS = \frac{||A \cap B||}{||A|| + ||B|| - ||A \cap B||}, \tag{1}$$

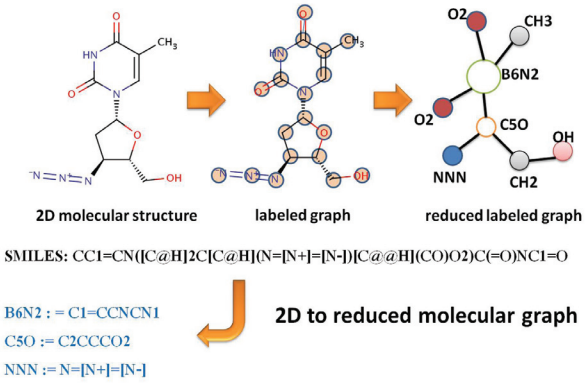

Figure 1: Molecular structure and its SMILES string and reduced graph.

where $||A \cap B||$ is the number of common sub-fragments of both structures, $||A||$ and $||B||$ are number of sub-fragments founded in structure A and B, respectively. There are some variety of tanimoto scores which depend on how fragment is defined. In this study the tanimoto score obtained by OpenBabel package [9]. The difference between proposed refined tanimoto score (RTCS) and the original TCS is how to define molecular fragment. In RTS, nodes of labeled molecular graphs are only labeled by complete defined rings or functional groups. These labled can be obtained from SMILES string using substring matching. Figure 2 and Figure 3, show the difference between obtained RTCS and original TCS scores.

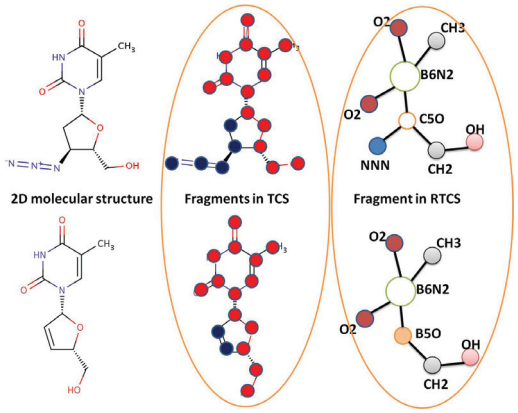

Figure 2: Reduced graphs and molecular fragments in TCS and RTCS scores.

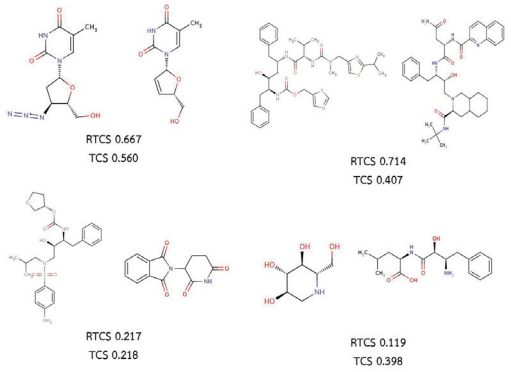

Figure 3: Comparison between TCS and RTCS scores of some drug pairs in database.

3 Results and discussions

The 37 molecules of drugs/inhibitors/ligands for SARS-Cov Protease obtained from DrugBank database [10] are used as testing data. Figure 4 shows the comparison between RTCS abd original TCS scores for all possible 666 drug pairs. The number of pairs that have obtained RTCS scores higher than its corresponding TCS scores is 454 pairs which more than $\frac{2}{3}$ of the number of taotal pairs. According to this computational experiment, it can say that if the RTCS is used instead of original TCS for similarity criteria i.e. > 0.60, the number of similarity structures can be expected increase (60 instead of 12). However, some few structures which should be found with TCS are possible missed and the increase of hit compounds also mean the increase of time for further calculations. In further works the estimation binding affinities for hit compounds will obtained by molecular docking as well as the searching addition compounds from other open access databases.

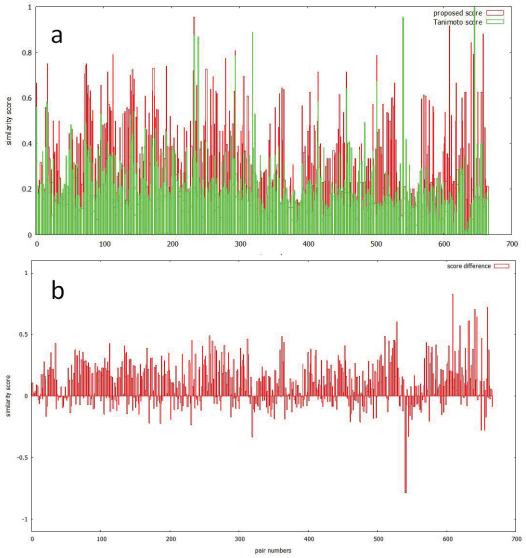

Figure 4: a) Comparision of RTCS (red) abd TCS (green) and b) the differences between two methods.

References

[1] J. M. Amig, J. Glvez, V. M. Villar, *A review on molecular topology: applying graph theory to drug discovery and design* Naturwissenschaften, 96 (2009), 749-761

[2] J. W. Raymond, P. Willett, *Maximum common subgraph isomorphism algorithms for the matching of chemical structures* J. Comp.-Aided Mol. Design 16 (2002), 521-533.

[3] H. Eckert, J. Bajorath, *Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches* Drug Discovery Today 12 (2007), 225-233.

[4] H. C. Ehrlich, M. Rarey, *Systematic benchmark of substructure search in molecular graphs - From Ullmann to VF2*, J. Cheminfo. 4 (2012), [doi:10.1186/1758-2946-4-13].

[5] D. Weininger, A. Weininger, J. L. Weininger, *SMILES. 2. Algorithm for Generation of Unique SMILES Notation* J. Chem. ZnJ Comput. Sci. 29 (1989), 97-101.

[6] N. M. O'Boyle, *Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI* J. Cheminfo. 4 (2012), [doi:10.1186/1758-2946-4-22].

[7] A. G. Maldonado, J.P. Doucet, M. Petitjean, B. T. Fan, *Molecular similarity and diversity in chemoinformatics: From theory to applications*, Mol.r Diver. 10 (2006), 39-79.

[8] P. Willett, *Chemical Similarity Searching*, J. Chem. Inf. Comput. Sci., 38 (1998), 983-996.

[9] N. M. O'Boyle, M. Banck, C. A. James et al., *Open Babel: An open chemical toolbox*, J. Cheminfo.3 (2011), 33.

[10] C. Knox, V. Law, T. Jewison et al., *DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.*, Nucleic Acids Res. 39 (2011),D1035-1041. (Database issue)