https://doi.org/10.17576/apjitm-2018-0701-08

http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 7 No. 1, June 2018: 91 - 98
e-ISSN: 2289-2192

COMPARISON OF SIMILARITY METHOD TO IMPROVE RETRIEVAL PERFORMANCE FOR CHEMICAL DATA

SUHAILA ZAINUDIN
NEVY RAHMI NURJANA

ABSTRACT

Drug discovery is the process through which new drugs are discovered. One of the most common techniques in drug discovery is similarity searching based on virtual screening that involves comparing the similarity between molecule structures in chemical database using established similarity methods. The objective of this study is to identify the similarity of the structure in chemical dataset using Mean Pairwise Similarity (MPS) calculation and to determine the best coefficient to be used in similarity searching which involves of molecular descriptor ECFP2 fingerprint and three types of similarity coefficient which are Tanimoto, Soergel and Euclidean. From the results, it was deduced that Tanimoto and Soergel coefficients has a better performance than Euclidean coefficient. For future work, different combinations of fingerprints such as Daylight, BCI, Unity MDL and similarity coefficient can be studied further.

Keywords: mean pairwise similarity; virtual screening; similarity searching; retrieval; chemoinformatics

INTRODUCTION

Drug discovery is the process through which new medicine are discovered. The process involves lengthy procedures of developing the drug. Lab tests and clinical tests are carried out to ensure the safety and effectiveness of the drug. One of the earliest domain to support drug discovery and design is Chemoinformatics (Gasteiger 2016). Chemoinformatics methods were developed for use in all major pharmaceutical companies (Gasteiger 2016). Using chemoinformatics as the basis, computer methods for learning from massive chemical data were proposed.

Drug discovery are found in medicine, biotechnology and pharmacology fields where the new candidate medications are indicated. The traditional drug discovery process includes step by step process from lead discovery (duration: 3 years), preclinical development (duration: 1 year), clinical development (duration: 4 years) and Food and Drug Administration (FDA) filing (duration: 1.5 years) (Hughes et al. 2011). As can be seen from the time taken by each step, these traditional methods can be labour intensive and time-consuming (Al Qaraghuli et al. 2017). However, the new development of computational technology can simplify and speed up the drug discovery process.

Numerous factors have made drug discovery more of a challenging task. Drug discovery is a lengthy and costly process. There are significant expenses incurred in the process which includes purchase of the main materials used in drug making. There are insufficient qualified diagnostic and also biomarkers in the process to help in the detection and treatment of diseases in the industry. Scientists have resolved to the use of chimpanzees for disease exposure as they are believed to have the same genes as those of the humans.

In the past, drug researchers made their discoveries through identification of the active ingredient from their traditional remedies. Current modern drug discovery involves methods in chemoinformatics like similarity searching, virtual screening among others. These methods have helped drug discovery in a substantial way in that they optimize the discovery process with speed and accuracy.

Chemoinformatics is a computer and information-based technique that has been widely used in drug discoveries in pharmaceutical companies. This technique uses basic application of science from different fields of science such as chemistry, and computer and information science (Gasteiger 2016). Areas of computer and information that has been studied in chemical space include topology, data mining, data retrieving and chemical graph theory (Alexandre and Baskin, 2011).

Virtual screening is a computational method applied in drug discovery. It involves searching for small molecules in large libraries of compounds with the aim of identifying structures which have high chances of binding with the drug target. A lot of studies have been done that has improved the accuracy of Virtual Screening (VS) and therefore it has become a crucial part of the process of drug discovery. Virtual screening is done in two broad ways; one is ligand-based, and the other one is structure-based.

Ligand-based virtual screening (LBVS) is the technique uses the information which is present and known in the identified active ligands for both lead identification and optimization. It does not use the structure of the target enzyme or protein receptor. These techniques are chosen when 3D structures of the target protein do not exist, for instance, in G-protein-coupled receptor targets. Even if the protein structure for the target is unknown, it is possible to identify a set of ligands which are active against the target. Therefore, in such cases, ligand-based techniques are used. Basically, it involves finding new ligands by examining and analyzing similarities between known active ligands and the candidate ligands. Besides ligand-based virtual screening, another approach is structure-based virtual (Sonalkar and Jain, 2016).

Structure-based virtual screening (SBVS) are methods of virtual screening that involves docking of candidate ligands into a protein target and then afterward applying a scoring function which will help in generating the probability of the ligand binding to the target protein with high affinity. These methods are very significant in drug discovery processes. They help in optimization of the discovery process. Structure-based discovery helps in understanding the molecular design of a disease by the use employing the knowledge of the 3D structure of the target. Structure-based computational approaches together with the 3D structure information of the compound target help in evaluating the molecular interactions between the ligand and the protein. Basically, in virtual screening, large libraries of huge numbers of drug-like compounds that are readily available (commercially) are computationally screened against targets of known structure. Numerous attempts have been made to develop computational algorithms to predict the binding affinity of a ligand to a given receptor, which would allow potential compounds to be screened in silico, reducing costs and saving time (Lee et al. 2016).

This work focuses on similarity searching. A similarity searching is done by matching or overlapping elements for purposes of qualitative or quantitative characterization. Characterization using similarity searching is a matter of trial and error. Queries are used in object specification, and when multiple searches are undertaken using a single query, it results in a hyperlinked screen that gives highly reliable information. These similarity searches retrieve information of objects similar to the query, and the data is sorted in order of decreasing similarity. The similarity scores illustrate the effectiveness of similarity searching (Wang and Bajorath, 2010).

Similarity searching has turned out to be the simplest and cost effective way for analyzing information among various chemical databases to identify the relationship between active structures of target references in the database. Through this approach, it is now easier to make a follow up when tracing the original active aspect basing on the level of resemblance between the structures. Due to its simplicity and effectiveness, most of chemoinformatics software systems are exploiting similarity searching using a sole target structure approach. In order to perform multiple search or to analyze target structure that are not structurally related, the similarity searching is performed through chemical database like MDL Drug Data Report (MDDR) (Finn and Morris, 2012).

SIMILARITY MEASURES

SIMILARITY COEFFICIENT

Similarity coefficient is used to determine the similarity between the query and the target in a form of fingerprint (Syuib et al., 2013). In chemoinformatics fields, there are many similarity coefficients that can be used to investigate similarity searching in virtual screening. There are two types of coefficient which can be calculated; either using distance coefficients or similarity coefficient. In this works, the focus is on 3 similarity coefficients which are Tanimoto coefficient, Soergel coefficient and Euclidean coefficient.

STRUCTURAL REPRESENTATIONS

Structural representation in chemoinformatics is describing the structural features of chemical structures. The representation known as “fingerprints” which are mathematically presented strings of binary bits. They are set in such a way that they produce a bit pattern of a specific molecule. In this work, the focus is on Extended Connectivity Fingerprints or as known as ECFP with the length of 2 bounds (ECFP2) to calculate the mean of recall and ECFP fingerprint with the length of 4 bounds (ECFP4) to calculate the Mean Pairwise Similarity (MPS). ECFPs are the new class of topological fingerprints used in molecule characterization. Topological fingerprints were mainly developed to assist in similarity searching as well as in substructure and today ECFPs are mostly used in activity modeling. ECFPs are the type of binary fingerprints and can be tailored to develop different types of fingerprints which can be optimized for the various applications. Seal et.al (2015) used ECFP6 to optimize drug target interactions.

METHODS

The datasets used in this experiment is MDDR datasets. MDDR is one of the database which commercially available and in this case the database used is purchased by Universiti Kebangsaan Malaysia. From this database, 15 random classes were chosen as the datasets for further investigation. The number of active molecules in the class are between 293 to 1355 molecules with total of active molecules of 9.941 molecules.

MEAN PAIRWISE SIMILARITY

This part involves selecting 15 activity classes from MDDR database as the datasets in this experiment. The first task is to calculate the Mean Pairwise Similarity (MPS) for every class in this datasets. Mean pairwise similarity is the similarity of the molecules in each activity class (Saeed et al., 2012). From the calculation of MPS, we can see whether each of activity class has similar molecules to each other (homogeny) or has dissimilar molecules to each other (heterogenic). In this task, MPS is calculated using Tanimoto coefficient and ECFP4 for the fingerprint representation.

The MDDR datasets were filtered to remove the duplicates and null data from each activity class. Then all the active molecules in each activity class were converted to ECFP4 fingerprint using Pipeline Pilots software (available from http://www.accelerys.com). Mean Pairwise Similarity would be calculated using Tanimoto coefficient which will compare the similarity of each molecule in each activity classes. The formula in Equation (1) is used for calculating Mean Pairwise Similarity in this datasets is

$$\text{Mean Pairwise Similarity} = \frac{\text{Similarity Value}}{\text{Number of active molecules in activity class}}. \qquad (1)$$

SIMILARITY SEARCH

The next part would be to compute the similarity search. In this task, the ECFP2 Fingerprint and Tanimoto, Soergel and Euclidean Coefficient to calculate the similarity search between two chemical structures using Mean of Recall formula in order to compare the similarity coefficients and the other task would be using Precision formula to compare the fingerprints which will be using Tanimoto as the coefficient and ECFP4, ECFP6 and FCFP6 for the fingerprint comparison. First, we filtered the MDDR datasets to remove duplicates or null data from each activity classes in this datasets. Then the datasets are converted to ECFP2 (1024bit) fingerprint using Pipeline Pilot software. Ten reference structures were chosen based on the most representatives ID/query from each class. The most representative ID are the 10 most similar molecules in each activity classes. In order to find the most representative ID/molecules in each class, the calculation using Tanimoto coefficient and ECFP2 fingerprint were involved. Each query of 10 the most representative ID will then be used to calculate the similarity value in each class in MDDR datasets. Only top 1% high ranked value will then be analysed for further investigation.

After obtaining the top 1% high ranked value, this result will be analysed to see how many of these values belong to the same activity class (true positive). After determining the true positive number, the mean of recall and precision will be calculated where the equation of the mean of recall and precision are as below (Equations (2) and (3)): -

$$\text{Mean of Recall} = \frac{\text{Number of True Positive}}{\text{Number of active molecules in activity class}}. \qquad (2)$$

$$\text{Precision} = \frac{\text{Number of True Positive}}{\text{Number of molecule top 1\%}}. \qquad (3)$$

RESULT AND DISCUSSION

The results are shown in Table 1, Table 2 and Table 3. Table I is a compilation of Mean Pairwise Similarity.

TABLE 1. Mean Pairwise Similarity

Activity Class ID	Activity class Name	Number of active molecules	Mean Pairwise Similarity
01252	Agent for Neurogenic Pain	634	0.1273
33451	Agent for Restenosis	695	0.1330
35100	Agent for Urinary Incontinence	913	0.1421
41270	Agent for Erectile Dysfunction	532	0.1602
42102	Growth Hormone Release Promoting Agent	398	0.2332
42731	Substance P Antagonist	366	0.2015
43200	Symptomatic Antidiabetic	980	0.1338
55210	Agent for Inflammatory Bowel Disease	293	0.1341
59500	Antiacne	444	0.1388
62210	Agent for Autoimmune Diseases	747	0.1304
64200	Cephalosporin	1355	0.3861
73000	Anthelmintic	541	0.1972
75400	Antineoplastic Antibiotic	921	0.1783
78329	Dipeptidyl Aminopeptidase IV Inhibitor	490	0.1716
80000	Diagnostic Agent	632	0.1193

Based on the Main Pairwise Similarity calculation in Table 1 it is clear that class ID for 64200 has the highest MPS value on these datasets which also has the highest number of active molecules among other activity classes and class ID for 80000 has the lowest MPS value on these datasets. From the MPS result shows that class ID for 64200 has the molecules which most similar to each other and class ID for 80000 have the molecules which are dissimilar to each other.

TABLE 2. Mean of Recall for MDDR Datasets

Activity Class ID	Activity class name	Similarity Coefficient		
		Tanimoto	Soergel	Euclidean
01252	Agent for Neurogenic Pain	0.029	0.029	0.037
33451	Agent for Restenosis	0.059	0.059	0.056
35100	Agent for Urinary Incontinence	0.039	0.039	0.040
41270	Agent for Erectile Dysfunction	0.199	0.199	0.138
42102	Growth Hormone Release Promoting Agent	0.217	0.217	0.253
42731	Substance P Antagonist	0.191	0.191	0.200
43200	Symptomatic Antidiabetic	0.073	0.073	0.067
55210	Agent for Inflammatory Bowel Disease	0.094	0.094	0.095
59500	Antiacne	0.125	0.125	0.118
62210	Agent for Autoimmune Diseases	0.049	0.049	0.035
64200	Cephalosporin	0.067	0.067	0.067
73000	Anthelmintic	0.166	0.166	0.166
75400	Antineoplastic Antibiotic	0.098	0.098	0.098
78329	Dipeptidyl Aminopeptidase IV Inhibitor	0.183	0.183	0.181
80000	Diagnostic Agent	0.097	0.097	0.104

Table 2 shows the mean of recall for MDDR datasets using Tanimoto, Soergel and Euclidean Similarity Coefficient. Based on this result, class ID for 42102 has the highest mean of recall of 0.253 when using Euclidean Similarity. However, the mean of recall for the same class using Tanimoto and Soergel similarity resulting not much difference with the Euclidean Similarity which is 0.217. This table shows that the mean of recall for Tanimoto and Soergel similarity has the same value in every classes and more than half of the classes shown that Tanimoto and Soergel similarity coefficient result with higher mean of the recall.

TABLE 3. Precision for MDDR Datasets

Activity Class ID	Activity class name	Similarity Coefficient		
		ECFP4	ECFP6	FCFP6
01252	Agent for Neurogenic Pain	0.214	0.214	0.229
33451	Agent for Restenosis	0.497	0.497	0.469
35100	Agent for Urinary Incontinence	0.434	0.434	0.464
41270	Agent for Erectile Dysfunction	0.864	0.864	0.874
42102	Growth Hormone Release Promoting Agent	0.902	0.902	0.800
42731	Substance P Antagonist	0.852	0.852	0.881
43200	Symptomatic Antidiabetic	0.753	0.753	0.719
55210	Agent for Inflammatory Bowel Disease	0.315	0.315	0.335
59500	Antiacne	0.649	0.649	0.616
62210	Agent for Autoimmune Diseases	0.353	0.353	0.367
64200	Cephalosporin	1.0	1.0	1.0
73000	Anthelmintic	1.0	1.0	1.0
75400	Antineoplastic Antibiotic	1.0	1.0	1.0
78329	Dipeptidyl Aminopeptidase IV Inhibitor	0.995	0.995	0.989
80000	Diagnostic Agent	0.744	0.744	0.837

Table 3 showed the precision for MDDR datasets using fingerprints ECFP4, ECFP6 and FCFP6. Based on this results, class ID 64200, 73000 and 75400 has reached the highest result of precision of 1.0 in 3 fingerprints meaning that these activity classes are able to retrieve the same molecule from the original class using 3 different fingerprints. The classes of 35100, 41270, 42731, 55210, 62210 and 8000 scored the highest result from FCFP6 fingerprint. Whereas, the scores for ECFP4 and ECFP6 are lower. In general, for this MDDR datasets, FCFP6 has the highest value of precision for this experiment.

Figure 1 shows the highest scores achieved by each similarity methods for different fingerprints. From Figure 1, each similarity method has different performance for each fingerprint. Tanimoto dan Soergel has high frequencies for ECFP4, ECFP6, FCFP4 and FCFP6. However, both has low frequencies for ECFP2 and FCFP2. Past research (Arif et al. (2015), Fatimah Zawani (2014), Syuib et al. (2013), Todeschini et al. (2012)) also shows that Tanimoto coefficient produced acceptable results. On the other hand, Euclidean scored high frequencies for ECFP2 and FCFP2 and scored low frequencies for ECFP6 and FCFP6.

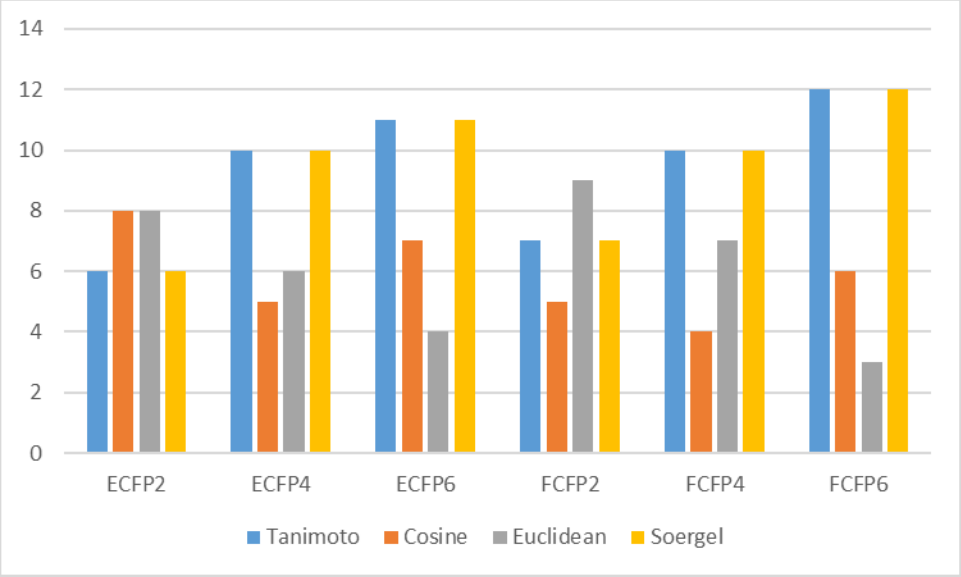

FIGURE 1. The Frequency of Scores for Each Similarity Methods

CONCLUSION

From this investigation, we clearly see that Tanimoto and Soergel has the same and higher value of the mean of recall. From the previous research in chemical similarity has also found that Tanimoto is the best coefficient among others to be used in similarity searching. The results reported above have shown that not only Tanimoto coefficient but also Soergel coefficient performs the same result in this MDDR datasets. In the future, the research can be extended by using many more of similarity coefficient with different types of molecular descriptors to this MDDR datasets. Consequently, this will lead to the discovery of new computational methods for prediction of drug target discovery.

REFERENCES

Al Qaraghuli, M. M., Alzahrani, A. R., Niwasabutra, K., Obeid, M. A. and Ferro, V. A. 2017. Where traditional drug discovery meets modern technology in the quest for new drugs. *Annals of Pharmacology and Pharmaceutics*, 2 (11): 1-5.

Arif, S., Khan, N. Z. S., Malim, N., and Zainudin, S. 2015. Retrieval performance using different type of similarity coefficient for virtual screening. *Research Journal of Applied Sciences, Engineering and Technology*, *9*(5): 391-395.

Fatimah Zawani bt Abdullah. 2014. Prestasi fingerprint dan pekali persamaan dalam informatik kimia menggunakan set data pubchem. Master Thesis. Universiti Kebangsaan Malaysia.

Finn, P., and Morris, G. 2012. Shape‐based similarity searching in chemical databases. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *3*(3), 226-241.

Gasteiger, J. 2016. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules*, 21(2): 151.

Hughes, J., Rees, S., Kalindjian, S., and Philpott, K. 2011. Principles of early drug discovery. British Journal of Pharmacology, 162(6): 1239–1249. http://doi.org/10.1111/j.1476-5381.2010.01127.x

Lee, A. A., Brenner, M. P., and Colwell, L. J. 2016. Predicting protein–ligand affinity with a random matrix framework. PNAS November 29, 2016. 113(48):13564-13569.

Saeed, F., Salim, N. and Abdo, A. 2012. Voting-based consensus clustering for combining multiple clusterings of chemical structures. *Journal of Cheminformatics*, 4(1): 1-8. doi:10.1186/1758-2946-4-37.

Seal, A., Ahn, Y.-Y., and Wild, D. J. 2015. Optimizing drug–target interaction prediction based on random walk on heterogeneous networks. Journal of Cheminformatics, 7, 40. http://doi.org/10.1186/s13321-015-0089-z.

Sonalkar, K and Jain, A. 2016. Virtual Screening through Substructure Matching. *International Journal of Science and Research.* 5(1): 663 – 666

Syuib, M. Arif, S. M. and Malim, N. 2013. Comparison of Similarity Coefficients for Chemical Database Retrieval. *AIMS '13 Proceeding of the 2013 1st International Conference on Artificial Intelligence, Modeling and Simulation*. pp. 129-133.

Wang, Y. and Bajorath, J. 2010. Advanced Fingerprint Methods for Similarity Searching: Balancing Molecular Complexity Effects. *Combinatorial Chemistry and High Throughput Screening*, *13*(3), 220-228.

Suhaila Zainudin
Nevy Rahmi Nurjana
Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia.

Received: 23 December 2017
Accepted: 28 February 2018
Published: 28 June 2018