Research Journal of Applied Sciences, Engineering and Technology 9(5): 391-395, 2015
DOI:10.19026/rjaset.9.1418
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: September 22, 2014 Accepted: October 24, 2014 Published: February 15, 2015

Research Article

Retrieval Performance using Different Type of Similarity Coefficient for Virtual Screening

[1]Shereena Arif, [1]Noor Zeemah Shamsheh Khan, [2]Nurul Malim and [1]Suhaila Zainudin
[1]Centre of Artificial Intelligence, Faculty of Information Sciences and Technology,
Universiti Kebangsaan Malaysia, 43650 UKM Bangi, Malaysia
[2]School of Computer Science, Universiti Sains Malaysia, 11800 Penang, Malaysia

Abstract: Development of a new drug needs chemical databases as references to find lead compounds. This study aims to determine the best similarity coefficient to be used for virtual screening task using chemical databases. We calculated the structural resemblance between each pair of chemical structures in their own activity class to get the Mean Pairwise Similarity (MPS) value to see the nature of heterogeneity for each natural product and synthetic chemical databases. The process involves the 2D descriptor of type ECFC4 fingerprint to represent each structure and Tanimoto coefficient to calculate the similarity score between each pair of chemical structures in the same activity class. MPS for an activity class was obtained by taking the average of all similarity scores within that class. Next, three types of similarity coefficients have been used to calculate the similarity score between a query structure and each of the database structure. The results indicate that Tanimoto coefficient shows better performance compared to Russell Rao and Forbes in retrieval task using chemical database. This implies that Tanimoto coefficient is recommended to carry out virtual screening in drug development. More work should be carried out to determine the best combination of similarity coefficient and fingerprint type to get optimal retrieval performance.

Keywords: Chemoinformatics, mean pairwise similarity, retrieval, similarity search, virtual screening

INTRODUCTION

Chemoinformatics can be described as the application of computer and information retrieval technique to solve a problem in the field of chemistry (Prakash and Gareja, 2010). Virtual screening is a technique that is used in drug discovery to search libraries of compounds using computer programs. There are many methods in virtual screening for example similarity searching, 3D pharmacophore matching and ligand docking. The focus of this paper is similarity searching, which can be defined as a measure to compute the degree of similarity between active reference structure and the chemical structures in the database of 2D structures as an effective way of searching large chemical databases (Willett, 2011). Structures in the database that have a high ranking value based on the reference structure can be considered as having similar biological activity with the reference structure (Johnson and Maggiora, 1990). The focus of virtual screening task is to separate compounds that have low similarity values, which will eventually save time, energy and cost for the chemists to investigate compounds in drug discovery process.

Similarity coefficient is used for calculating the degree of resemblance of active reference chemical structure with the chemical structures in the database (Willett, 2003). There are three important components that is used in similarity searching, molcular descriptor to represent a chemical compound; similarity coefficient to measure the resemblance between a pair of chemical structures and a weighting scheme to differentiate importance of each fragment occurrence in a compound. However, No Free Lunch Theorem (Wolpert and Macready, 1997) suggests that an algorithm would not satisfy all condition of a problem. Thus, this study is to determine the best similarity coefficient to be used with ECFC fingerprint in carrying out virtual screening task.

Similarity measures:

Similarity coefficients: There are many types of similarity coefficients, but only three similarity coefficients that are used to calculate similarity search here which are Tanimoto, Russell-Rao and Forbes. Descriptors that represent a molecular structure can be in continuous and dichotomous (i.e., binary) form. Holliday *et al*. (2002) found that these three coefficients are grouped differently in a clustering work they carried out. Similar results were found when different database and fingerprint types were used (Salim *et al*., 2003). The list below shows the similarity coefficients for Tanimoto, Russell-Rao and Forbes in continuous form which is applicable to non-binary data representation.

Corresponding Author: Shereena Arif, Centre of Artificial Intelligence, Faculty of Information Sciences and Technology, Universiti Kebangsaan Malaysia, 43650 UKM Bangi, Malaysia, Tel.: +6038921 6086

The Tanimoto, Russell-Rao and Forbes coefficients is given by S^1, S^2 and S^3, respectively:

$$S^1 = \frac{\sum (x_{ru} . x_{rv})}{\sum (x_{ru})^2 + \sum (x_{rv})^2 - \sum (x_{ru} . x_{rv})} \quad (1)$$

$$S^2 = \frac{\sum (x_{ru} . x_{rv})}{n} \quad (2)$$

$$S^3 = \frac{n \sum (x_{ru} . x_{rv})}{n \sum (x_{ru})^2 . \sum (x_{rv})^2} \quad (3)$$

For the similarity coefficients (1), (2) and (3), x refers to the representation of the chemical structure for u and v where u is the representation for query structure and v is the representation for database structure and n refers to the number of bits of the representation.

Representations: Representation describes the structural features of the chemical structures. These representations are fragment bit strings also known as "fingerprints". In this study we only focus on continuous representation which is Extended Connectivity Count vector or ECFC with the length of four bonds (ECFC4), containing 1024 bit-string. This continuous fingerprint is the non-binary representation of fragment bit string and is a 2D fingerprint. ECFC fingerprint are based on counts of how many times each fragment present in the chemical structure rather than binary strings which only encodes the presence and absence of a fragment (Todeschini and Consonni, 2009).

METHODOLOGY

The datasets used in this investigation were Taiwan Traditional Chinese Medicine (TCM) and MDL Drug Data Report (MDDR) database. TCM is one of the natural products database that is freely available at http://tcm.cmu.edu.tw (Chen, 2011) with 12,289 compounds that focuses on plant-based traditional remedies data repositories. In another hand, MDDR represents a synthetic chemical database with 211,061 compounds. MDDR is a commercial database subscribed from Accelrys Inc (available from http://www.accelrys.com) (Sheridan and Joseph, 2004).

Mean pairwise similarity: This task involves 17 activity classes from TCM database and 15 activity classes that has been chosen from MDDR database. First, we calculated the Mean Pairwise Similarity (MPS) for all the activity classes. Mean pairwise similarity is the similarity of chemical structures in each activity class (Saeed *et al*., 2012). MPS is conducted using Tanimoto coefficient as it is the most popular coefficient used in computing chemical similarity. While ECFC4 is chosen for the representation of the chemical structures as recent work found that it shows the best retrieval performance among many (Franco *et al*., 2014; Bender *et al*., 2009; Medina-Franco *et al*., 2009).

The TCM and MDDR datasets were filtered to remove duplicates of chemical structures in each activity class. Then all the active molecules in each activity class were converted to ECFC4 fingerprints using Pipeline Pilot software (available from http://www.accelrys.com) that gives 1024-element fingerprints (Warr, 2012). MPS is calculated using the Tanimoto coefficient, which will compare the reference structure with all the structures in the activity class thus giving the similarity value between structures in the activity class. The formula used for calculating the MPS is given below:

$$\text{MPS} = \frac{\text{Similarity value}}{\text{\# of actives in the activity class}}$$

Similarity search: The next task is to compute the similarity search. In this task we will use ECFC4 fingerprint with Tanimoto, Russell-Rao and Forbes as the coefficients to calculate the similarity between two chemical structures. First, TCM database are filtered to remove duplicates of the chemical structures. Then the database are converted into ECFC4 (1024 bit) fingerprint using the Pipeline Pilot to represent the chemical structures. Ten reference structures were randomly selected from each activity class. Each reference structure similarity value is calculated against the whole datasets to get the similarity value and only the top 1% of the highest ranked result was chosen for further investigation.

Next, the results that were obtained are then investigated to see how many of them belong to the same activity class which is known as true positives. True positive is the number of successful retrieved chemical structures (Wolpert and Macready, 1997). The next task is to calculate the Mean of Recall (MR) using the frequency of true positives obtained. The equation below shows the formula to calculate mean of recall:

$$\text{MR} = \frac{\sum \text{Number of true positive}}{\sum \text{Number of actives in activity class}}$$

RESULTS AND DISCUSSION

Table 1 indicates the number of active molecules in each activity in TCM. From here, it is clear that AM and PE activity class ID has the highest and lowest value of MPS, respectively. We can see the activity class that has the high value of MPS is from the class that has the lower number of active molecules in its

Table 1: Mean pairwise similarity for TCM activity classes

Activity class ID	Activity class	Number of actives	Mean pairwise similarity
AM	Anti-malaria medicinal	30	0.385
AS	Astringent medicinal	262	0.217
CM	Traditional Chinese medicine	8893	0.216
DM	Digestant medicinal	144	0.184
EM	Emetic medicinal	8	0.352
ER	Exterior-releasing medicinal	894	0.207
HC	Heat-clearing medicinal	1597	0.250
IW	Interior-warming medicinal	453	0.187
LP	Liver-pacifying and wind-extinguishing medicinal	100	0.238
NC	No category	4984	0.217
PE	Parasites elimination, dampness reduction and itchiness relief	81	0.156
PM	Purgative medicinal	284	0.280
TR	Tonifying and replenishing medicinal	1458	0.238
WD	Wind-dampness dispelling medicinal	517	0.240
WE	Worm-expelling medicinal	92	0.230
DR1	Dampness-resolving medicinal 1	560	0.230
DR2	Dampness-resolving medicinal 2	169	0.251

Table 2: Mean pairwise similarity for MDDR activity classes

Activity class ID	Activity class	Number of actives	Mean pairwise similarity
5HT1AG	5 HT1A agonist	250	0.417
5HTR	5 HT reuptake	625	0.403
AP	Antineoplastic	921	0.388
AD	Autoimmune disease	747	0.385
CCB	Calcium channel blocker	257	0.391
CD	Cardiovascular disorders	417	0.356
CCK	CCK antagonist	208	0.549
5HT3	5 HT3 antagonist	536	0.371
5HT1AN	5 HT1A antagonist	277	0.514
IB	Inflammatory bowel disease	293	0.345
RH	Rhinitis	870	0.418
SH	Sedative/hypnotic	600	0.357
SD	Sleep disorders	1307	0.424
SPA	Substance P antagonist	366	0.513
UI	Urinary incontinence	913	0.387

class which consist of 30 active molecules. This shows that the chemical structures in AM activity class ID are the most similar to each other than other activity classes while the chemical structures in PE activity class ID are the most dissimilar to each other.

Table 2 shows the MPS values for MDDR activity class. The activity class ID that has the highest value of MPS is CCK (i.e., 0.549) while IB has the lowest (i.e., 0.345). Based on this, the activity class that has lower number of active molecules has the highest value of MPS for TCM and MDDR activity classes which are 30 and 208 actives molecules, respectively. Further analysis shows that MDDR activity classes has a higher value of MPS compared to TCM activity classes.

As TCM shows more heterogeneity which represents a more challenging dataset, we further the work in determining the best similarity coefficients using the natural product database. Table 3 shows the mean of recall for 14 activity class for TCM. Here we can see that PE activity class ID has the highest mean of recall of 0.043 when using Tanimoto as the similarity coefficient. TR activity class ID has the lowest mean of recall using this coefficient with the value of 0.008.

There is a relationship between the MPS value and mean of recall based on results outlined in Table 3. PE activity class ID has the lowest MPS but highest mean of recall using Tanimoto coefficient. This is also true when using Forbes coefficient, where PE activity class ID gives a high mean of recall of 0.035 while HC activity class ID which has lower MPS gives the poorest retrieval performance with mean of recall of 0.005. However, in the case when using Russell-Rao similarity coefficient, it shows LP and DR2 activity class ID which represents high level of homogeneity (i.e., high MPS) has the highest mean of recall (i.e., value 0.030) and lowest (i.e., value 0.00) for the mean of recall. This indicates that Russell-Rao alone should not be considered for chemical similarity task as it is unsuitable to both homogenous and heterogenous datasets.

Recently, there exist more interest in producing molecular descriptors based on physicochemical properties and Structure-Activity Relationship (SAR) in a molecule based on statistical techniques (Hancock *et al*., 2005; Andersson *et al*., 2000; Mridha *et al*., 2014) and machine learning approaches (Kovačević *et al*., 2014; Nantasenamat *et al*., 2014). These works found that these molecular descriptors able to give comprehensive coverage in solving chemical problems.

Table 3: Mean of recall for TCM activity classes

		Similarity coefficient		
Activity class ID	Activity class	Tanimoto	Russell-rao	Forbes
AS	Astringent medicinal	0.029	0.021	0.007
CM	Traditional Chinese medicine	0.010	0.009	0.011
DR1	Dampness-resolving medicinal 1	0.017	0.007	0.007
DR2	Dampness-resolving medicinal 2	0.027	0.000	0.018
DM	Digestant medicinal	0.023	0.001	0.010
EM	Exterior-releasing medicinal	0.026	0.001	0.012
HC	Heat-clearing medicinal	0.011	0.006	0.005
IW	Interior-warming medicinal	0.033	0.009	0.016
LP	Liver-pacifying and wind-extinguishing medicinal	0.030	0.030	0.027
PE	Parasites elimination, dampness reduction and itchiness relief medicinal	0.043	0.005	0.035
PM	Purgative medicinal	0.012	0.007	0.006
TR	Tonifying and replenishing medicinal	0.008	0.013	0.006
WD	Wind-dampness dispelling medicinal	0.016	0.015	0.008
WE	Worm-expelling medicinal	0.037	0.001	0.008

Thus, future work can be done to compare the performance of these QSAR descriptors and fingerprint-based descriptors to determine the best descriptors used with Tanimoto coefficient.

CONCLUSION

Previous works in this field has investigate the effect of similarity coefficients in synthetic chemical database and found that Tanimoto is the best coefficient to be used in virtual screening. This study extends the application and shows that it also perform better than Russell-Rao and Forbes when used with natural product database. In future work we will extend the research by using new molecular descriptors that are produced based on physicochemical properties to see the effect of different types of representations on the retrieval of TCM database.

ACKNOWLEDGMENT

This study is supported by the UKM Grant GUP-2013-010 and FRGS-2-2013-ICT02-UKM-02-2.

REFERENCES

Andersson, P.M., M. Sjöström, S. Wold and T. Lundstedt, 2000. Comparison between physicochemical and calculated molecular descriptors. J. Chemometr., 14(5-6): 629-642.

Bender, A., J.L. Jenkins, J. Scheiber, S.C.K. Sukuru, M. Glick and J.W. Davies, 2009. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. J. Chem. Inf. Model., 49(1): 108-119.

Chen, C.Y.C., 2011. TCM database@Taiwan: The world's largest traditional Chinese medicine database for drug screening *In silico*. PloS One, 6(1): e15939.

Franco, P., N. Porta, J.D. Holliday and P. Willett, 2014. The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation. J. Chem. Inf., 6(1): 1-10.

Hancock, T., R. Put, D. Coomans, Y. Vander Heyden and Y. Everingham, 2005. A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies. Chemometr. Intell. Lab., 76(2): 185-196.

Holliday, J.D., C.Y. Hu and P. Willett, 2002. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. Comb. Chem. High T. Scr., 5(2): 155-166.

Johnson, M.A. and G.M. Maggiora, 1990. Concepts and Application of Molecular Similarity. John Wiley and Sons, New York.

Kovačević, S.Z., S.O. Podunavac-Kuzmanović, L.R. Jevrić, E.A. Djurendić and J.J. Ajduković, 2014. Non-linear assessment of anticancer activity of 17-picolyl and 17-picolinylidene androstane derivatives-chemometric guidelines for further syntheses. Eur. J. Pharm. Sci., 62: 258-266.

Medina-Franco, J.L., K. Martínez-Mayorga, A. Bender, R.M. Marín, M.A. Giulianotti, C. Pinilla and R.A. Houghten, 2009. Characterization of activity landscapes using 2D and 3D similarity methods: Consensus activity cliffs. J. Chem. Inf. Model., 49(1): 477-491.

Mridha, P., P. Pal and K. Roy, 2014. Chemometric modelling of triphenylmethyl derivatives as potent anticancer agents. Mol. Simulat., 40(15): 1-18.

Nantasenamat, C., A. Worachartcheewan, P. Mandi, T. Monnor, C. Isarankura-Na-Ayudhya and V. Prachayasittikul, 2014. QSAR modeling of aromatase inhibition by flavonoids using machine learning approaches. Chem. Pap., 68(5): 697-713.

Prakash, N. and D.A. Gareja, 2010. Cheminformatics. J. Proteomics Bioinform., 3(1): 249-252.

Saeed, F., N. Salim and A. Abdo, 2012. Voting-based consensus clustering for combining multiple clusterings of chemical structures. J. Cheminformat., 4(1): 1-8.

Salim, N., J. Holliday and P. Willett, 2003. Combination of fingerprint-based similarity coefficients using data fusion. J. Chem. Inf. Comp. Sci., 43(2): 435-442.

Sheridan, R.P. and S. Joseph, 2004. Calculating similarities between biological activities in the MDL drug data report database. J. Chem. Inf. Comp. Sci., 44(2): 727-740.

Todeschini, R. and V. Consonni, 2009. Molecular Descriptors for Chemoinformatics. John Wiley and Sons, New York.

Warr, W.A., 2012. Scientific workflow systems: Pipeline pilot and KNIME. J. Comput. Aid. Mol. Des., 26(7): 1-4.

Willett, P., 2003. Similarity-based approaches to virtual screening. Biochem. Soc. T., 31(3): 603-606.

Willett, P., 2011. Similarity searching using 2D structural fingerprints. Method. Mol. Biol., 672(1): 133-158.

Wolpert, D.H. and W.G. Macready, 1997. No free lunch theorems for optimization. IEEE T. Evolut. Comput., 1(1): 67-82.