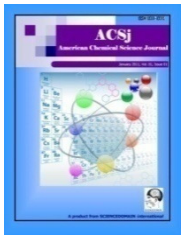

American Chemical Science Journal
4(6): 787-797, 2014

SCIENCEDOMAIN *international*
www.sciencedomain.org

Molecular Similarity Searching Method Based on Adaptive IR Technique

Mohammed Salem Binwahlan[1*] and Naomie Salim[2]

[1]Faculty of Applied Science, Hadhramout University, Yemen.
[2]Faculty of Computer Science and Information Systems, University Technology Malaysia, 81310 Skudai, Johor, Malaysia.

Authors' contributions

This work was carried out in collaboration between all authors. Author NS prepared the data set only and remaining work was done by author MSB. All authors read and approved the final manuscript.

Original Research Article

Received 17th March 2014
Accepted 21st April 2014
Published 31st May 2014

ABSTRACT

An age-old question remains as an open research problem in the field of chemoinformatics, which is how much could the proposed approach enhance the effectiveness of lead-discovery programmes? Answering that question is a target of any new virtual screening approach. The current research tries to contribute in this direction by improving the performance of molecular similarity searching process. In this paper, Okapi similarity measure, which is effective and widely used in text retrieval, is adapted to perform the role of molecular similarity measure in 2D fingerprints. The adapted similarity measure calculates the molecular similarity between a reference structure and a database structure. The experimental results showed that the proposed method performs well compared to Tanimoto coefficient.

Keywords: Molecular similarity searching; structure searching; similarity coefficients; virtual screening.

**Corresponding author: Email: moham2007med@yahoo.com;*

1. INTRODUCTION

Chemoinformatics is a discipline that exploits computational techniques to deal with chemical problems and provides suitable solutions for them [1]. Virtual screening techniques are the backbone of most chemical problems. In a virtual screening process, very large libraries of compounds are automatically evaluated based on an implementation of a computational technique proposed for this purpose. Virtual screening covers a category of such computational techniques, which enables chemists to control the size of a massive virtual library [2]. The role of the computational techniques is to score each molecule in the database of molecules based on its reaction to a specified biological target [3]. The virtual screening (VS) is widely employed to boost the cost-effectiveness of molecular database consultation during biological testing, where the database molecules are ranked in decreasing order based on their scores. This leads to consider just those few molecules that have the highest a priori scores of activity [4,5]. Thus far, its application and its usefulness have been realized widely by pharmaceutical industry.

Molecular similarity searching, which is one of the most broadly used virtual screening approaches, is a specific class of similarity search problems in which a given molecular query is compared against a collection of molecules, in order to retrieve those that most closely similar to the molecular query. Molecular similarity searching approaches work based on the fact stated by the Similar Property Principle [6], which is, molecules that share similar structures can be described by similar properties. Based on this fact, the properties of any molecule, unseen before, can be inferred by the properties of the structurally similar molecules to it. There are different mechanisms of chemical database searching based on matching of the molecular structures. Structure searching imposes an exact-match between a reference structure and a database structure. Substructure searching requires a partial-match of a user-defined query with a database structure to retrieve all those molecules that contain a user-defined query substructure [1,5,7].

The age-old question remains as an open research problem in the field of chemoinformatics, which is how much could the proposed approach enhance the effectiveness of lead-discovery programmes? Answering that question is a target of any new virtual screening approach. The Tanimoto coefficient [1] dominates on the peak of performance hierarchy of the existing molecular similarity searching techniques and provides the best performance. The current research tries to contribute in this direction by improving the performance of molecular similarity searching process. In this paper, Okapi similarity measure [8], which is effective and widely used in text retrieval, is adapted to perform the role of molecular similarity measure in 2D fingerprints, the adapted similarity measure calculates the similarity between a reference structure and a database structure.

2. MOLECULAR SIMILARITY SEARCHING

In modern chemical research, structural storages such as databases have become essential tools for storing the amount of information accumulated by chemists and facilitate the accessibility to the interested people of chemical data [9]. There are different ways of consulting such structural storages for specific information, structure searching and substructure searching. The later (substructure searching) is less difficulty than the former (structure searching) and more reasonable, but has some limitations such as the user, who consults the chemical database for specific query, must already get prior knowledge about the output structures returned from the database. The difficulty of this condition is when the

information about the particular feature(s), which led to select only one or two active structures for the target activity, is absent [5]. To rid of such difficulty and other limitations, chemical similarity searching is used as an alternative[10]. Similarity searching involves the description information of the whole structure in hand, unlike substructure searching that involves only a partial structure. The similarity degree between the target structure (query) and each structure in the database is measured by comparing the structural descriptors, which are in common between the target structure (query) and each structure in the database. The similarity degrees, resulting in the comparison process, are then reranked into an order of decreasing similarity with the target. The top *n* of the molecules in the ranked list will be the most likely relevant to the user' need, given an appropriate degree of intermolecular structural similarity.

Since the early works on similarity searches that appeared in the mid-1980s, based on the work carried out at Lederle Laboratories [11] and Pfizer [12], the choosing of a similarity measure is still purely trial and error process [13]. In the Lederle study, molecules were represented by their constituent atom pairs, where an atom pair is a substructural fragment comprising two non-hydrogen atoms together with a number of intervening bonds. The similarity search allowed users to request either some number of the top-ranked molecules or all those that had a similarity with the target structure greater than a minimal value. In the Pfizer system, together with a conventional substructural query, a user can submit a target molecule typical of the type of the structure that was required. The conventional screen search and atom-by-atom search were used to identify matches in the substructure searching, after which a similarity measure based on the screens common to the target and the matches was used to rank the substructure search output. The subsequent development of a faster, inverted-file-based, nearest neighbour search algorithm allowed the ranking of the entire database against the target structure in real time, without the need for the specification of the initial substructural query.

Later, similarity searching has undergone further investigation. An example is Hagadone's work on substructure similarity searching [14]. Substructure similarity searching is used to identify molecules containing a substructure similar to a target structure or substructure. Another extension of similarity search was described by Fisanick et al. [15] on facilities developed for Chemical Abstracts Service (CAS) Registry File. It focuses on different types of similarity relationships that can be identified between a structure in the query and a database structure.

Most recent works in this direction were done by Chen et al. [3], Abdo and Salim [16] and Abdo and Salim [17]. Chen et al. [3] adapted Bayesian inference network for molecular similarity searching problem, encouraged by the performance of such a network in texts ranking based on their relevance to a user-defined query. Bayesian inference network is a probability based technique. The researcher found that the Bayesian inference network could present a virtual screening performance better than the virtual screening performance of Tanimoto coefficient; particularly, when the vast majority of the molecules, contained in the data set being searched, has homogeneous molecular structures. Nevertheless, when the vast majority of those molecules has heterogeneous molecular structures, the virtual screening performance of Bayesian inference network is much less. Chen et al., [3] pointed out that a similar study to their study was done by Abdo and Salim [16]. The main difference between the two studies is only in terms of the size and types of data sets being used for evaluation purposes, where the former used a set of 102 K MDDR structures and eleven associated activity classes (MDDR-HOM, MDDR-HET and WOMBAT datasets), while the later used a small subset of the MDDR database, containing just 40 K structures.

As Bayesian inference network is used by both research groups (Chen et al.'s group and Abdo and Salim's group), Abdo and Salim [17] followed the same methodology explained in Chen et al.'s study [3] to run an experiment for testing the virtual screening performance of the Bayesian inference network using a number of weighting functions, which are mostly similar to those used by Chen et al. [3].

Recent work on similarity searching was proposed by Riniker and Landrum [18], the researchers focused on the visualization of both fingerprint similarities between two molecules and machine-learning (ML) model predictions. Random forest (RF) and naïve Bayes (NB) were, used as machine-learning (ML) methods, trained and employed to predict the probability to be active of new molecules.

3. SIMILARITY MEASURES

Different kinds of molecular representations can be used as chemical reference spaces, but the main observation which should be kept in mind that nature of relationships between molecules is not invariant to the selection of such spaces. Therefore, molecular similarity can be evaluated only based on a given molecular representation. Upon that, molecular similarity or dissimilarity is calculated by intermolecular distance in the selected reference space [19]. Conventional distance metrics such as Euclidean [20,21] or Divergence measure the distance between molecules in chemical space, whereas similarity coefficients (e.g., Tanimoto, Russell or Cosine coefficient) directly assess intermolecular similarity [22]. The adapted measure in this study belongs to similarity coefficient category.

4. TANIMOTO COEFFICIENT (TC)

Association coefficients are those similarity coefficients that return the similarity degree in the range [1], 0 mean complete dissimilarity and 1 self-similarity. A degree of molecular similarity is calculated using bit string matching in a case of the molecular representations being used in the binary fingerprint format. The association coefficient most widely used in chemical applications is theTanimoto coefficient (Tc) [1,22,23], which accumulates the number of bits common to two binary fingerprints with respect to the total number of bits that are set in each fingerprint. The Tc dominates on the peak of performance hierarchy of the existing molecular similarity searching techniques and provides the best performance. This is the reason of using Tc as benchmark method to compare the performance of the current study with. The Tc for two binary fingerprint representations A and B is calculated as follows.

$$Tc(A,B) = \frac{N_{AB}}{N_A + N_B - N_{AB}} \qquad (1)$$

Where N_{AB} is the number of bits set on in both fingerprints and N_A and N_B refer to the number of bits set on in A and B, respectively.

5. THE PROPOSED METHODS

The role of virtual screening techniques is to score each molecule in the database of molecules based on its reaction to a specified biological target. Consequently, the cost-effectiveness of molecular database consultation during biological testing is boosted, where

the database molecules are ranked in decreasing order based on their scores. This leads to consider just those few molecules that have the highest a priori scores of activity.

The current research tries to contribute in this direction by improving the performance of molecular similarity searching process. In this study, Okapi similarity measure Eq. 2 [8], which is effective and widely used in text retrieval, is adapted to be suitable for chemical data representation and to perform the role of molecular similarity measure in 2D fingerprints, the adapted similarity measure, as in Eq. 5, calculates the similarity between a reference structure and a database structure.

$$Okapi(d_1, d_2) = \sum_{t \in d_1 \cap d_2} \frac{3 + tf_{d_2}}{0.5 + 1.5 * \frac{len_{d_2}}{len_{avg}} + tf_{d_2}} * log \frac{N - df + 0.5}{df + 0.5} * tf_{d_1} \quad (2)$$

$$Okapi(q, m) = \sum_{f \in q \cap m} \frac{3 + ff_m}{0.5 + 1.5 . \frac{len_m}{len_{avg}} + ff_m} . ff_q . IMF(q) \quad (3)$$

$$IMF(q) = log \frac{N - mf + 0.5}{mf + 0.5} \quad (4)$$

$$Okapi(q, m) = \sum_{ff \in q \cap m} \frac{3 + ff_m}{0.5 + 1.5 . \frac{len_m}{len_{avg}} + ff_m} . ff_q . \sqrt{IMF(q)} \quad (5)$$

Where *f* is the fragment, *q* is the query, *m* is a molecule, ff_m is the fragment frequency in a molecule, len_m is a molecule length, len_{avg} is the average length of all molecules in the database, ff_q is the fragment frequency in the query, *IMF(q)* is the IMF (inverse molecule frequency) weight of a query fragment, *N* is the number of molecules in the whole database and *mf* is the number of the molecules containing the fragment.

In addition to the adaptation of Okapi similarity measure Eq. 2, a simple modification was added, which is a square root of the IMF (inverse molecule frequency) weight of a query fragment.

6. EXPERIMENTAL SETUP

We run our experiment on the MDDR database [24], which was formerly explained and used by Hert et al. [25]. This database is in two dimensions. First dimension consists of 102516 compounds and second dimension consists of 1024-element fingerprints produced using the Pipeline Pilot software [26]. Each element in the second dimension of the fingerprint holding the occurrences of that a particular substructure in a molecule. For fair evaluation of the proposed method, three data sets of different content type were used in the conducted experiment. The first data set (DS1) is a mixture of structurally homogeneous (MDDR-HOM) and structurally heterogeneous (MDDR-HET) classified into 11 activity classes, the second data set (DS2) holds 10 homogeneous activity classes, and the third data set (DS3) holds 10

heterogeneous activity classes. However, these data sets are similar to those explained in Hert et al.' work [25] with a bit difference concluded in two different activity classes in the DS2 data set and the size of activity classes in DS2 and DS3.

Tables 1-3 show descriptions of these three data sets. Each row of a table holds the activity class, the number of molecules under the class, and the diversity of the class, which was computed as the mean pairwise Tanimoto similarity calculated across all pairs of molecules in the class using ECFP6. The pairwise similarity calculations for all data sets were conducted using Pipeline Pilot software [26]. The experiment was conducted with 10 reference structures selected randomly from each activity class. The recall, which is the percentage of the active molecules retrieved at both top1% and top-5% cut-off points in the ranking, was calculated as averaged over each set of active molecules.

Table 1. MDDR activity classes for DS1

Activity class	Active molecules	Pairwise similarity (mean)
Renin inhibitors	1130	0.290
HIV protease inhibitors	750	0.198
Thrombin inhibitors	803	0.180
Angiotensin II AT1 antagonists	943	0.229
Substance P antagonists	1246	0.149
5HT3 antagonists	752	0.140
5HT reuptake inhibitors	359	0.122
D2 antagonists	395	0.138
5HT1A agonists	827	0.133
Protein kinase C inhibitors	453	0.120
Cyclooxygenase inhibitors	636	0.108

Table 2. MDDR activity classes for DS2

Activity class	Active molecules	Pairwise similarity (mean)
Adenosine (A1) agonists	207	0.229
Adenosine (A2) agonists	156	0.305
Renin inhibitors	1130	0.290
Monocyclic β-lactam	111	0.361
Cephalosporins	1346	0.336
Carbacephems	113	0.322
Carbapenems	1051	0.269
Penicillin	126	0.260
Antibiotic, macrolide	388	0.305
Vitamin D analogous	455	0.386

Table 3. MDDR activity classes for DS3

Activity class	Active molecules	Pairwise similarity (mean)
Muscarinic (M1) agonists	900	0.111
NMDA receptor antagonists	1400	0.098
Nitric oxide synthase inhibitors	505	0.102
Dopamine-hydroxylase inhibitors	106	0.125
Aldose reductase inhibitors	957	0.119
Reverse transcriptase inhibitors	700	0.103
Aromatase inhibitors	636	0.110
Cyclooxygenase inhibitors	636	0.108
Phospholipase A2 inhibitors	617	0.123
Lipoxygenase inhibitors	2111	0.113

7. EXPERIMENTAL RESULTS AND DISCUSSION

Although similarity searching has undergone further investigation, the choosing of a similarity measure is still purely trial and error process. Therefore, the main objective of this experiment is to try and investigate the screening performance of adaptive similarity measure in chemical reference space. However, molecular similarity or dissimilarity is calculated by intermolecular distance in the selected reference space. The current study is dedicated to examine the ability of okapi similarity measure for boosting the cost-effectiveness of molecular database consultation during biological testing and to show how the okapi similarity measure could be a competitive virtual screening approach. To this end, the introduced method is evaluated using three data sets: DS1, DS2 and DS3 with different content types of a mixture of structurally homogeneous (MDDR-HOM) and structurally heterogeneous (MDDR-HET) activity classes, homogeneous activity classes, and heterogeneous activity classes, respectively. For fair comparison, Tanimoto coefficient (Tc) was tested based on the same data sets.

Tables 5-6 present the screening performance of the presented method on DS1-DS3, respectively. The reported results in each row in those tables, which have been taken as both top1% and top-5% cut-off points in the ranking, were calculated as averaged over 10 reference structures for each activity class. Tables 5-6 also show the screening performance of Tanimoto coefficient (Tc) on the same data sets for the purposes of the comparison. The mean row shows the final average recall, which is calculated as an average of all recall results of all activity classes over the total number of the activity classes in the current data set. As shown in Tables 5-6, Okapi similarity measure noticeably effective and highly outperforms Tanimoto coefficient (Tc) across the 10 activity classes for the three datasets (DS1-DS3). The bold and highlighted cells in Tables 4-7 mean the value contained in any cell of those cells is the best.

Table 7, which consists of two main parts, gives a quick glance for all results shown in Tables 4-6. The first part is the highlighted cells part, which shows the number of the highlighted cells of Okapi similarity measure and Tanimoto coefficient (Tc) for both top1% and top-5% cut-off points in the ranking in each data set (Ds1-Ds3). The second part is the means part, which presents the means of Okapi similarity measure and Tanimoto coefficient (Tc) for both top1% and top-5% cut-off points in the ranking in each data set (Ds1-Ds3). The bottom raw in Table 7 contains the summations of the number of the highlighted cells and the summations of the means of both similarity measures (Okapi and Tanimoto) for both top1% and top-5% cut-off points in the ranking in all data sets (Ds1-Ds3).

The results drawn in Table 4 show that the proposed method obtained, a mean value of 2.1663 for recall of actives in the top-1% and a mean value of 2.9573 for recall of actives in the top-5%, higher than Tanimoto coefficient. For the number of active classes, it can be seen that the proposed method performs well with 7 active classes, while Tanimoto coefficient performs well with only 4 active classes (in the top-1%). Also, the proposed method performs well with 6 active classes, while Tanimoto coefficient performs well with only 5 active classes (in the top-5%). Therefore, it can be reported that the proposed method (using Okapi similarity measure) outperforms Tanimoto coefficient.

Table 4. Recall of actives in the top-1% and the top-5% of the ranked MDDR database (DS1) using the Okapi similarity measure and Tanimoto coefficient

Activity class	Tan		Okapi	
	1%	5%	1%	5%
Renin inhibitors	69.69	83.49	71.79	86.94
HIV protease inhibitors	25.94	48.92	27.13	53.72
Thrombin inhibitors	9.63	21.01	23.53	47.46
Angiotensin II AT1 antagonists	35.82	74.29	40.15	78.49
Substance P antagonists	17.77	29.68	18.73	26.93
5HT3 antagonists	13.87	27.68	13.4	23.69
5HT reuptake inhibitors	6.51	16.54	6.34	15.28
D2 antagonists	8.63	24.09	11.5	26.93
5HT1A agonists	9.71	20.06	10.91	24.13
Protein kinase C inhibitors	13.69	20.51	12.48	19.29
Cyclooxygenase inhibitors	7.17	16.2	6.3	12.14
Mean	19.8573	34.77	22.0236	37.7273
No. of highlighted cells	4	5	7	6

Table 5. Recall of actives in the top-1% and the top-5% of the ranked MDDR database (DS2) using the Okapi similarity measure and Tanimoto coefficient

Activity class	Tan		Okapi	
	1%	5%	1%	5%
Adenosine (A1) agonists	61.84	70.39	71.94	75.24
Adenosine (A2) agonists	47.03	56.58	97.23	100
Renin inhibitors	65.1	88.19	74.9	94.2
Monocyclic β-lactam	81.27	88.09	81	91.82
Cephalosporins	80.31	93.75	89.57	99.39
Carbacephems	53.84	77.68	70.27	98.75
Carbapenems	38.64	52.19	68.28	90.9
Penicillin	30.56	44.8	79.04	93.92
Antibiotic, macrolide	80.18	91.71	82.07	90.7
Vitamin D analogous	87.56	94.82	98.02	98.26
Mean	62.633	75.82	81.232	93.318
No. of highlighted cells	2	1	8	9

From the results shown in Table 5, it can be noticed that the proposed method achieved, a mean value of 18.599 for recall of actives in the top-1% and a mean value of 17.498for recall of actives in the top-5%, higher than Tanimoto coefficient. For the number of active classes, it can be realized that the proposed method does well with 8 active classes, while Tanimoto

coefficient does well with only 2 active classes (in the top-1%). Also, the proposed method does well with 9 active classes, while Tanimoto coefficient does well with only 1 active classes (in the top-5%). Therefore, it can be stated that the proposed method (using Okapi similarity measure) outperforms Tanimoto coefficient.

The results presented in Table 6 indicate that the proposed method got, a mean value of 18.599for recall of actives in the top-1% and a mean value of 17.498for recall of actives in the top-5%, higher than Tanimoto coefficient. For the number of active classes, it can be noticed that the proposed method executes well with 8 active classes, while Tanimoto coefficient executes well with only 2 active classes (in the top-1%). Also, the proposed method executes well with 9 active classes, while Tanimoto coefficient executes well with only 1 active classes (in the top-5%). Therefore, it can be said that the proposed method (using Okapi similarity measure) outperforms Tanimoto coefficient.

Table 6. Recall of actives in the top-1% and the top-5% of the ranked MDDR database (DS3) using the Okapi similarity measure and Tanimoto coefficient

Activity Class	Tan		Okapi	
	1%	5%	1%	5%
Muscarinic (M1) agonists	14.73	30.67	16.38	26.64
NMDA receptor antagonists	7.65	12.53	9.23	12.44
Nitric oxide synthase inhibitors	6.94	14.92	9.19	17.4
Dopamine -hydroxylase inhibitors	19.9	30.67	20.95	31.05
Aldose reductase inhibitors	7.69	16.89	8.83	15.23
Reverse transcriptase inhibitors	2.89	7.41	5.58	10.16
Aromatase inhibitors	25.92	32.36	25.24	35.84
Cyclooxygenase inhibitors	11.06	18.61	9.8	16.17
Phospholipase A2 inhibitors	10.6	27.32	8.62	21.4
Lipoxygenase inhibitors	9.84	13.01	13.86	16.18
Mean	11.722	20.439	12.768	20.251
No. of highlighted cells	2	5	7	5

Table 7. Overall results

Data Sets	No. of highlighted cells				Means			
	1%		5%		1%		5%	
	Tan	Okapi	Tan	Okapi	Tan	Okapi	Tan	Okapi
DS1	4	7	5	6	19.8573	22.0236	34.77	37.7273
DS2	2	8	1	9	62.633	81.232	75.82	93.318
DS3	2	7	5	5	11.722	12.768	20.439	20.251
Sum	8	22	11	20	94.2123	116.0236	131.029	151.2963

Table 7 presents the overall performance of the proposed method and Tanimoto coefficient with each data set. The proposed method performs well with 22 active classes of the three data sets, while Tanimoto coefficient performs well with only 8 active classes (in the top-1%). Also, the proposed method performs well with 20 active classes of the three data sets, while Tanimoto coefficient executes well with only 11 active classes (in the top-5%). For the sum of recall means of actives of the three data sets, the proposed method achieved a value of 21.8113 in the top-1% and a mean value of 20.2673 in the top-5%, higher than Tanimoto coefficient. Therefore, it can be concluded that the overall performance of the proposed

method (using Okapi similarity measure) better than the overall performance of Tanimoto coefficient.

The reported screening performance of the introduced method in Tables 5-7 enables us to draw two main conclusions. First, that Okapi similarity measure is a greatly promising technique for molecular similarity searching. Second, the proposed method performs well on structurally homogeneous (MDDR-HOM) data set (DS2). It is not clear why there is such a noticeable difference in the screening performances of Okapi similarity measure when different types of dataset are screened.

8. CONCLUSION

In this paper, we have examined the use of Okapi similarity measure for molecular similarity searching, which was the main objective of this experiment. Our experiment with three data sets of the MDDR database showed that the proposed method presents an effective tool for similarity-based virtual screening in terms of the calculation of the intermolecular similarity in a selected reference space. Specifically, our experiment has demonstrated a significant superiority of the proposed method for screening the activity class types when compared to a conventional screening system based on the Tanimoto coefficient. However, this study broke the basis saying that Tanimoto coefficient dominates on the peak of performance hierarchy of the existing molecular similarity searching techniques and provides the best performance. We look forward to improving and enhancing of the proposed method in different ways such as feature weighting or selection.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Leach AR, Gillet VJ. An introduction to chemoinformatics dordrecht: Kluwer; 2007.
2. Walters WP, Stah MT, Murcko MA. Virtual Screening-An Overview. Drug Discovery Today. 1998;3:160-178.
3. Chen B, Mueller C, Willett P. Evaluation of a Bayesian inference network for ligand-based Virtual Screening, Journal of Cheminformatics. 2009;1(1):5.
4. Bohm HJ, Schneider G, (Editors). Virtual Screening for Bioactive Molecules. Wiley-VCH; 2000.
5. Willett P. Digital libraries in chemistry: providing access to chemical structure information. In: Tabata, K., Ishii, H. and Sugimoto, S., (eds.) Proceedings of International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society 2004. International Symposium. Tsukuba, Japan: University of Tsukuba. 2004;33-39.
6. Johnson MA, Maggiora GM, (Editors). Concepts and Applications of Molecular Similarity. Wiley; 1990.
7. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular Similarity in Medicinal Chemistry. Journal of Medicinal Chemistry. doi: 10.1021/jm401411z. 2013.
8. Zhang B, Gon M, Fan A, Chen W, Fox Y EA, Calado P, Cristo M. Intelligent fusion of structural and citation-based evidence for text classification. ACM Thirteenth Conference on Information and Knowledge Management , CIKM'04, November 8–13, 2004, Washington D.C., USA.

9. Gasteiger J, Engel T. Chemoinformatics: A Textbook, Wiley-VCH Verlag; 2003.
10. Willett P, Barnard JM, Downs GM. Chemical Similarity Searching. Journal of Chemical Information and Computer Sciences. 1998;38:983-996.
11. Carhart RE, Smith DH, Venkataraghavan R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. J. Chem. Inf. Comput. Sci. 1985;25:64-73.
12. Willett P, Winterman V, Bawden D. Implementation of Nearest Neighbor Searching in an Online Chemical Structure Search System. J. Chem. Inf. Comput. Sci. 1986;26:36-41.
13. Monev V. Introduction to Similarity Searching in Chemistry. Measurement, 51 (3592). 2005;7-38.
14. Hagadone TR. Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. Journal of Chemical Information and Computer Science. 1992;32:515-521.
15. Fisanick W, Cross KP, Rusinko A. A similarity search on CAS Registry Substances. 1. Global molecular property and generic atom triangle geometric searching. Journal of Chemical Information and Computer Sciences. 1992;32:664-674.
16. Abdo A, Salim N. Similarity-based virtual screening with a Bayesian inference network. Chem Med Chem. 2009;4:210-18.
17. Abdo A, Salim N. New Fragment Weighting Scheme for the Bayesian Inference Network in Ligand-Based Virtual Screening. J. Chem. Inf. Model. 2011;51(1):25–32.
18. Riniker S, Landrum G. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. Journal of cheminformatics 5:43. doi: 10.1186/1758-2946-5-43. 2013.
19. Varnek A, Tropsha A, (Editors), Chemoinformatics Approaches to Virtual Screening. The Royal Society of Chemistry; 2008.
20. Carbó-Dorca RJ. Math Chem. 2012;50:734-740
21. Carbó-Dorca R, Besalú E. J Math Chem. DOI 10.1007/s10910-011-9960-y
22. Gillet VJ, Wild DJ, Willet P, Bradshaw J. Similarity and dissimilarity methods for processing chemical structure databases. The Computer Journal. 1980;41:547-558.
23. Godden JW, Xue L, Bajorath J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. J. Chem. Inf. Comput. Sci. 2000;40:163-166.
24. MDL Drug Data Report; Symyx Technologies: San Diego, CA. Available:http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp. Accessed November 19, 2010.
25. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. J. Chem. Inf. Model. 2006;46:462–470.
26. Pipeline Pilot; Accelrys Software Inc.: San Diego, CA; 2008.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid= 528&id=16&aid=4762