BAYESIAN INFERENCE NETWORK FOR MOLECULAR SIMILARITY SEARCHING USING 2D FINGERPRINTS AND MULTIPLE REFERENCE STRUCTURES

Ammar Abdo[1], Naomie Salim[2]

Faculty of Computer Science and Information Systems
University Teknologi Malaysia
81300 Skudai, Johor

[1]ammar_utm@yahoo, [2]naomie@utm.my

Abstract: 2D fingerprint based similarity searching using a single bioactive reference is the most popular and effective virtual screening tool. In our last paper, we have introduced a novel method for similarity searching using Bayesian inference network (BIN). In this study, we have compared BIN with other similarity searching methods when multiple bioactive reference molecules are available. Three different 2D fingerprints were used in combination with data fusion and nearest neighbor approaches as search tools and also as descriptors for BIN. Our empirical results show that the BIN consistently outperformed all conventional approaches such as data fusion and nearest neighbor, regardless of the fingerprints that were tested.

Keywords: Virtual Screening, Drug Discovery, Bayesian Network, Inference Network, Multiple Reference, Similarity Searching.

1. INTRODUCTION

Nowadays, virtual screening tools are fast growing and widely used to enhance the cost effectiveness of drug discovery. Similarity searching is one of the most widely used virtual screening approaches. The basic idea underlying similarity searching approach is the similar property principle, where structurally similar molecules tend to have similar properties.[1] Over the years, many types of similarity measure have been introduced in the literature,[2, 3] but by far the similarity measure based on the number of substructural fragments common to a pair of molecules and a simple association coefficient are the most common.[4, 5] Currently, the most prominent coefficient being used is the Tanimoto coefficient (TC). In similarity searching, a query involves the specification of an entire structure of a molecule. This specification is in the form of one or more structural descriptors and this is compared

with the corresponding set of descriptors for each structure in the database [2]. A measure of similarity is then calculated between the target structure and every database structure. Similarity measures quantify the relatedness of two molecules with a large number (or one) if their molecular descriptions are closely related and with a small number (large negative or zero) when their molecular descriptions are unrelated. The results of the similarity measure will be used to sort the database structures into the order of decreasing similarity with the target. This type of order then means that biological testing can be focused on just those few molecules that come on the top of the list.

Most of the studies that have been reported in the literature have considered on the similarity searching based on 2D fingerprints that use only a single bioactive reference molecule. However, studies of similarity searching when not one but several bioactive reference structures are available shown that is noticeably superior to that obtained from the use of a single bioactive reference structure and the search performance usually improves. Many approaches have been introduced to utilize multiple bioactive references, for instance, Shemetulskis et al.[6] developed modal fingerprint containing the common bits found in the molecular fingerprints of the set of bioactive reference molecules. In order to set a bit in the modal fingerprint, this bit must be common in the set of bioactive reference molecules. The degree to which a bit is considered common is determined from a user defined threshold value. This value ranges from 50% to 100%. In another study, Sheridan [7] proposed centroid approach where descriptor representations of set of reference molecules may be approximated as the descriptor average of its individual molecules. He suggested that, using the centroid approximation for similarity methods and quantitative structure-activity relationship (QSAR) methods can give useful results. There are two important advantages to a modal fingerprint and centriod representation of the descriptors in a multiple bioactive reference molecules. First, the generated representation is computationally inexpensive. Assume that the reference set contains n_a active molecules and that there are n molecules in the whole database, then the modal fingerprint and centriod approaches requires n similarity calculations instead of nn_a. Second, modal fingerprint and centriod can be conducted on existing software that is designed to run with a single molecule as query. Xue et al.[8, 9] have introduced a scaling technique that emphasizes consensus bit settings in keyed fingerprints conserved in a specific set of compounds having similar biological activity. Scaling procedure incorporates generating fingerprint profile for each activity class. These profiles are then used to create a consensus pattern, which consists of all bits that are always set "on" in a specific activity class, and then scaling factors that are weighted according to the bit frequencies in the fingerprint profiles applied to the bits present in the consensus pattern, so that bits that are in common and in the consensus patter give higher great contribution to the overall similarity calculation than nonconsensus bits. It was demonstrated that profile scaling consistently and often

significantly increased the similarity search performance of fingerprints. However, Hert et al.[10] have demonstrated that fingerprint scaling does not materially affect the performance of similarity searching approaches when a multiple reference molecules are available.

Nearest neighbor methods can use a set of reference molecules instead of a single reference molecule as target for similarity searching.[11]. This can be done in several ways. Assume that, the set of reference molecules contains n active molecules, the representation of vector of molecule i in the reference set R be r_i and the representation vector of the molecule in a data set be x. Similarity function $S(x,r_i)$ is used to calculate the similarity between molecule x and query r_i. According to the centroid method introduced by Sheridan,[7] the similarity between the molecule x and the center of the reference set can be shown as:

$$S(x,R) = S(x,c), \qquad c = \frac{1}{n}\sum_{i=1}^{n} r_i$$

Another way to calculate the similarity between the molecule x and the target is by calculating the similarity between molecule x and each member in the reference set and then takes the average similarity over them according to

$$S(x,R) = \frac{1}{n}\sum_{i=1}^{n} S(x,r_i)$$

One may takes the average for only k highest values representing the similarities $S(x,r_i)$ between molecule x and reference set R. According to the value of k, these methods will be referred to as k nearest neighbor (k-NN) methods. 1-NN is a particular k-NN method in which the similarity between molecule x and R reference set is defined by choosing the highest similarity value. Recently, data fusion methods are described to enhance the effectiveness of the similarity searching.[10, 12-14] Data fusion involves combining the results of different similarity searches of a chemical database. Binary kernel discrimination (BKD) and support vector machine (SVM) are machine learning techniques that can be used when a multiple reference of molecules are available, in which a reference set will be used rather than a full training set.[10, 15, 16] Several studies have been conducted to compare these different search techniques. Hert et al.[10] have investigated many approaches such as single fingerprints, substructural analysis, data fusion and BKD. The single fingerprint approaches involves creating a single combined fingerprint from the fingerprints of the individual reference structures. Experiments on MDDR database demonstrate that data fusion based on similarity scores and BKD method are the best with BKD being slightly more effective, but notably the less efficient, of the two. Wilton et al.[15] have studied BKD, similarity searching, substructural analysis and SVM methods for virtual screening. Experiments on pesticide data set show that BKD outperforms the similarity searching and substructural analysis methods and inferior to SVM approach. Geppert et al.[16] have

investigated 1-NN, 5-NN, centroid and SVM methods for similarity searching using multiple reference molecules. Experiments show that the SVM method outperforms the 1-NN, 5-NN and centriod approaches.

Currently, Abdo and Salim have introduced a novel technology for similarity searching based on Bayesian inference network [17]. Bayesian inference networks were originally developed for text documents retrieval and have become popular in the information retrieval field.[18-21] however, their empirical results demonstrate that BIN outperforms current conventional similarity searching methods.

In this paper, we discuss Bayesian inference network with multiple reference structures. In this view, similarity searching is an inference or evidential reasoning process in which we estimate the probability that a reference, expressed as multiple structures, is met given compound as evidence.

2. METHODS

We have investigated three different ways of carrying out such search when multiple bioactive reference molecules are available. The detailed implementation of these approaches is discussed in below.

2.1 Bayesian Inference Network Method

The BIN was first introduced to molecular similarity searching using a single bioactive reference structure by Abdo and Salim [17]. In their network model, a set of bioactive reference structures can be utilized. BIN model for similarity searching using multiple references, shown in Figure 1, consists of two component networks: a compound network and a query network. The compound network represents the compound collection. The compound network is built once for a given collection and its structure does not change during query processing. The query network consists of a multiple nodes, which represents the reference structures. The reference set is expressing the target structure (activity-need node). A query network is built for each target and modified in some cases (refined or added) in an attempt to better characterize the target structure. The compound and query networks are connected though links between their feature nodes (more details see ref 17).

In the BIN, an individual similarity search will be conducted for each active reference structure and then the resulting similarity scores are combined using weighted-sum link matrix, with weights expressing the importance of each score. In the inference net model we need to encode the dependency of activity-need node (target node) to the reference nodes. To

encode this probability, we use weighted-sum canonical link matrix form [17, 21]. By using weighted-sum canonical link matrix form, we assigned a weight to each of the n parents of the activity-need node, reflecting their influence on the activity-need node, as shows in Figure 1. The parents with larger weights have more influence on our belief *bel(A)*. The belief in the activity-need node is then determined by the parents that are involved and evaluated as

$$bel(A \mid q_{1..n}) = \sum_{i=1}^{n} w_{ij} p_{ij}, \quad w_{ij} = \frac{c_{ij}}{ql_i}$$

where c_{ij} is the number of common features between i^{th} reference and j^{th} structure, ql_i is the size of the i^{th} reference, w_{ij} is the assigned weight to i^{th} reference and j^{th} structure, and p_{ij} is the estimated probability that the i^{th} reference is met the j^{th} structure. The equation above is identical to SUM strategy, MAX strategy if applied to only reference node with highest value, and NN strategy if applied to k reference nodes with highest values and then averaged.

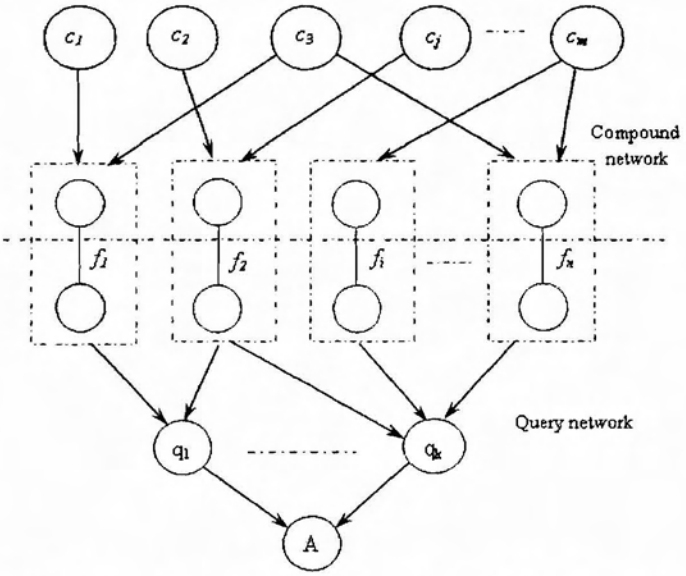

Figure 1. Similarity inference network model using multiple reference structures.

2.2 Data Fusion and Nearest Neighbor Methods

The BIN method is compared to two popular techniques for similarity searching using multiple bioactive reference structures, the DF [10, 12] and NN [11] approach in combination with Tanimoto similarity [4, 5]. Our application of DF involves fusion the similarity scores yield from similarity searches of a chemical database against each member of the reference set. Specifically, we have used the MAX and SUM fusion rules, for the maximum of the similarity scores and the sum of the similarity scores respectively. Assuming that, the set of

reference structures contains n active structures, and x is a structure in a database yield similarity scores of s_1, $s_2 \ldots s_n$ with the n different reference structures, a fusion rule can be applied on these scores according to

$$Maximum = \{s_1, s_2 \ldots s_{n-1}, s_n\}$$

$$Sum = \sum_{i=1}^{n} s_i$$

Similarly, the k nearest neighbor method (k-NN) applied on the k ($1 \leq k \leq n$) highest values similarity scores. The average of these k selected values represents the finial similarity score for x according to

$$k - NN = \frac{1}{k} \sum_{i=1}^{k} s_i, \quad 1 \leq k \leq n$$

Thus, the data fusion and nearest neighbor approaches conducts an individual similarity search for each reference molecule and then "fuses" the resulting similarity scores.

3. EXPERIMENTAL DETAILS

3.1 Fingerprint Designs

In order to make the evaluation of an approach independent of the characteristics of the specific fingerprint design, we included three different 2D fingerprints in our experiments: extended-connectivity fingerprint (ECFC), atom environment fingerprint (EEFC) and hashed atom environment fingerprint (EHFC) from SciTegic [22]. Extended-connectivity fingerprint generate higher-order features, each feature represents the presence of a structural (not substructural) unit. It was chosen because it was found to perform better when used in similarity searching to retrieve compounds with similar biological activity [10]. Atom environment fingerprint, generates higher-order features using a method developed by Bender et al.[23]. Hashed atom environment fingerprint, use a hashing algorithm to create an integer fingerprints representation of the atom environment fingerprints.

Maximum distance parameter is used with all fingerprint types above. For extended-connectivity fingerprints, it is the maximum diameter of the feature generated. For atom environment types, it is the maximum bond distance. To make the computational task manageable, we employed a maximum diameter size of four for all types in this study, and the fingerprints are folded to a fixed length of 1024 bits. Similar to other fingerprints, these fingerprints encodes whether a feature is present in a molecule or not. Furthermore, these

fingerprints show how many times this feature is present in the molecule. All the fingerprints types above were generated by Pipeline Pilot software,[24] from SciTegic.

Table 1. MDDR compound activity classes used in the study.

code	activity class	number of compounds	diversity	
			mean	SD
5H3	5HT3 antagonists	213	0.8537	0.008
5HA	5HT1A agonists	116	0.8496	0.007
D2A	D2 antagonists	143	0.8526	0.005
Ren	Renin inhibitors	993	0.7188	0.002
Ang	Angiontensin II AT1 antagonists	1367	0.7762	0.002
Thr	Thrombin inhibitors	885	0.8283	0.002
SPA	Substance P antagonists	264	0.8284	0.006
HIV	HIV-1 protease inhibitors	715	0.8048	0.004
Cyc	Cyclooxygenase inhibitors	162	0.8717	0.006
Kin	Tyrosin protein kinase inhibitors	453	0.8699	0.006
PAF	PAF antagonists	716	0.8669	0.004
HMG	HMG-CoA reductase inhibitors	777	0.8230	0.002

3.2 Database Preparation

For evaluation of the various approaches above, simulated virtual screening searches have been conducted on the MDL Drug Data Report [25] (MDDR) database. After removal of duplicates and molecules could not be processed using Pipeline Pilot software, a total of 40751 compounds were available for searching and forming our test database, including 6804 compounds belonging to 12 different activity classes. The activity classes and number of compounds per class are reported in Table 1.

Pipeline Pilot software[24] used to conduct a rigorous search on each of the chosen set of the bioactives, by matching each compound with every other in its activity class, calculating diversity using the ECFP_6 fingerprints and Tanimoto coefficient, and computing the mean and standard deviation for these intraset diversities. The resulting diversity scores are listed in Table 1, where it will be seen that the renin inhibitors are the most homogenous and the cyclooxygenase are the most heterogeneous. For each of the 12 activity class, 10 different sets of 10 active compounds were randomly selected as reference sets. Hence, each searching method was repeated 10 times using 10 different reference sets that for each type of fingerprint (ECFC_4, EEFC_4, and EHFC_4). For each combination of a fingerprint and activity class, the BIN, DF and NN methods were applied and the percentage of the recall active structures that monitored at the top 5% of the ranking list were generated. The results

presented in this study are the mean and standard deviations for these recall values, averaged over each set of the 10 searches. DF and NN methods were applied in combination with non-binary Tanimoto coefficient to compute the similarity scores due to the fingerprint combine integer values.

Table 2. Comparison of the Average Percentage of Actives Compounds Recalled over the Top 5% of the Ranked Test Set by Combining the Scores of Different Single Similarity Searches using BIN and DF Approaches with ECFC Fingerprints.

	BIN				DF			
	MAX		SUM		MAX		SUM	
activity class	mean	SD	mean	SD	mean	SD	mean	SD
5H3	68.92	5.89	65.17	3.71	41.48	6.63	42.07	13.20
5HA	75.19	5.72	69.62	10.05	58.30	8.73	42.74	11.39
D2A	67.97	5.64	64.74	5.63	50.60	8.36	42.71	6.03
Ren	94.18	1.31	93.76	1.03	89.74	3.01	89.20	3.03
Ang	86.91	5.97	83.02	9.02	73.32	3.91	76.07	3.27
Thr	50.53	12.45	43.65	12.26	26.73	7.09	32.74	6.23
SPA	63.86	13.98	46.85	23.03	70.12	8.19	54.96	9.97
HIV	61.71	6.31	54.85	9.51	61.77	2.61	57.59	5.34
Cyc	48.88	14.41	24.61	14.15	55.52	10.00	34.41	6.39
Kin	43.88	13.62	34.58	17.38	33.52	9.24	30.99	10.64
PAF	45.64	14.77	32.88	17.70	32.76	6.66	19.93	4.50
HMG	78.71	14.94	66.13	23.75	54.06	7.54	50.17	7.97
Average	65.53	9.58	56.66	12.27	53.99	6.83	47.80	7.33

4. RESULTS AND DISCUSSION

Our experiments were carried out in two different ways. First, the BIN method conducts an individual similarity search for each active reference structure, and then "fuses" the resulting similarity values. Second, the data fusion and nearest neighbor methods conducts an individual similarity search for each active reference structure, and then "fuses" the resulting similarity values. With the data fusion method we choices MAX and SUM fusion rules. With the nearest neighbor method, 3-NN and 5-NN were choosing.

Inspection of the results reported in the Table 2, suggest that fusion by MAX rule is better than SUM rule (in both BIN and DF), with difference in the performance being greatest for the more heterogeneous activity classes (cyclooxygenase inhibitors). The results in Table 2, show that, fuses scores resulting from BIN approach produced the overall highest average recall rate (in percentage) rather than fuses scores resulting from Tanimoto similarity using

DF technique, with greatest performance for the MAX fusion rule. The superiority of the MAX is in line with other previous studies [10, 11].

Inspection of Table 3, shows that 3-NN achieved the overall highest average recall rate (in percentage) than 5-NN (in both BIN and NN), with greatest performance for fuses scores resulting from BIN than Tanimoto similarity method using NN technique, with difference in the performance being greatest for the more heterogeneous activity classes. The superiority of fuses scores resulting from BIN than that from Tanimoto similarity method is ascribed to the fact that, an individual similarity search for each active reference structure by BIN generates ranked lists riches with active molecules than that generated by Tanimoto similarity method, and then fuses these lists being in high performance.

Table 3. Comparison of the Average Percentage of Actives Compounds Recalled over the Top 5% of the Ranked Test Set by Combining the Scores of Different Single Similarity Searches using the BIN and NN Approaches with ECFC Fingerprints.

activity class	BIN				NN			
	3-NN		5-NN		3-NN		5-NN	
	mean	SD	mean	SD	mean	SD	mean	SD
5H3	68.82	4.92	67.69	4.46	37.19	12.88	39.71	15.41
5HA	76.60	9.69	74.62	10.32	47.17	11.67	44.53	13.12
D2A	67.89	5.23	66.16	5.65	48.72	6.87	46.39	6.80
Ren	94.33	1.27	94.20	1.16	90.54	2.84	91.03	2.50
Ang	87.25	6.22	86.30	6.77	75.61	5.05	76.34	4.17
Thr	49.00	12.60	47.31	12.16	29.76	11.44	30.69	10.04
SPA	60.47	15.97	55.00	19.16	68.35	9.52	64.72	11.70
HIV	60.26	7.76	58.10	8.16	60.27	3.92	59.58	5.06
Cyc	42.11	15.98	34.87	15.68	53.22	9.62	41.45	11.50
Kin	41.08	16.21	38.29	17.63	33.41	13.12	33.47	12.61
PAF	40.74	17.46	37.42	18.08	27.21	4.86	23.64	5.96
HMG	76.05	19.00	73.19	20.66	47.59	9.33	45.17	8.83
Average	63.72	11.03	61.10	11.66	51.59	8.43	49.73	8.98

Inspection the results reported in Table 4, provide the basis for a detailed comparison with BIN method. This comparison revealed one major trend, regardless of the fingerprints and activity classes that were tested. The BIN approach produced consistently higher recall rates than conventional similarity searching approaches (data fusion and nearest neighbor). BIN (in combination with MAX and 3-NN) obtained highest recall rates than DF and NN approaches, with 21% and 24% performance improvement in overall average recall rate.

Table 4. Comparison of the Average Percentage of Actives Compounds Recalled by the Various Methods over the Top 5% of the Ranked Test Set Using ECFC Fingerprints.

activity class	MAX				3-NN			
	BIN		DF		BIN		NN	
	mean	SD	mean	SD	mean	SD	mean	SD
5H3	68.92	5.89	41.48	6.63	68.82	4.92	37.19	12.88
5HA	75.19	5.72	58.30	8.73	76.60	9.69	47.17	11.67
D2A	67.97	5.64	50.60	8.36	67.89	5.23	48.72	6.87
Ren	94.18	1.31	89.74	3.01	94.33	1.27	90.54	2.84
Ang	86.91	5.97	73.32	3.91	87.25	6.22	75.61	5.05
Thr	50.53	12.45	26.73	7.09	49.00	12.60	29.76	11.44
SPA	63.86	13.98	70.12	8.19	60.47	15.97	68.35	9.52
HIV	61.71	6.31	61.77	2.61	60.26	7.76	60.27	3.92
Cyc	48.88	14.41	55.52	10.00	42.11	15.98	53.22	9.62
Kin	43.88	13.62	33.52	9.24	41.08	16.21	33.41	13.12
PAF	45.64	14.77	32.76	6.66	40.74	17.46	27.21	4.86
HMG	78.71	14.94	54.06	7.54	76.05	19.00	47.59	9.33
Average	65.53	9.58	53.99	6.83	63.72	11.03	51.59	8.43

Table 5. Comparison of the Average Percentage of Actives Compounds Retrieved over the Top 5% of the Ranked Test Set with Single Similarity Searches Using ECFC Fingerprints.

activity class	BIN		Tanimoto	
	mean	SD	mean	SD
5H3	34.43	2.94	28.69	4.23
5HA	35.98	4.64	27.90	3.59
D2A	28.13	1.68	25.11	2.88
Ren	84.27	6.01	74.32	9.64
Ang	58.50	10.69	59.48	4.68
Thr	30.86	9.14	18.97	3.35
SPA	28.59	14.58	34.98	4.62
HIV	37.05	10.24	40.64	3.03
Cyc	12.84	5.88	23.01	2.02
Kin	20.66	10.88	21.44	5.69
PAF	14.55	7.08	14.63	1.48
HMG	35.99	15.56	30.13	3.16
Average	35.15	8.28	33.28	4.03

Inspection the results reported in Table 5, revealed the benefit that can be achieved using multiple reference structures, rather than single reference structure as in BIN and in conventional similarity searching. Such search carrying out by conducts an individual similarity search for each reference structure in each of the 10 different sets (100 individual similarity searches for each activity classes) and then averaged over each set of the 10 searches. The values under mean column in Table 5 shows the expected recall rate using a single reference structure are clearly much lower than results reported in Table 4 for the BIN, DF and NN approaches, with 86%, 81%, 62% and 55% performance improvement in overall average recall rate when multiple reference structures used rather than just one.

5. CONCLUSION

Similarity searching using Bayesian inference network and 2D fingerprint provide an effective and an efficient technique for virtual screening when a single reference structure is available [17]. In this work we have investigated the similarity searching based on Bayesian inference network and conventional similarity searching approaches when multiple reference structures are available. The BIN was found to outperform the data fusion, and nearest neighbor approaches. The observed improvements in recall rates by BIN were because of the understanding of the contents fingerprints of the entire database (including reference structures), and uses the information gained from that understanding to calculate the similarity degree between structures.

REFERENCES

[1] Johnson, M. A. and Maggiora, G. M., "Concepts and Application of Molecular Similarity," John Wiley & Sons, New York 1990.

[2] Willett, P., Barnard, J. M., and Downs, G. M., "Chemical Similarity Searching," J. Chem. Inf. Comput. Sci., vol. 38, pp. 983-996, 1998.

[3] Bender, A. and Glen, R. C., "Molecular similarity: a key technique in molecular informatics," Org. Biomol. Chem., vol. 2, pp. 3204 - 3218, 2004.

[4] Salim, N., Holliday, J., and Willett, P., "Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion," J. Chem. Inf. Comput. Sci., vol. 43, pp. 435-442, 2003.

[5] Willett, P., "Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures," J. Med. Chem., vol. 48, pp. 4183-4199, 2005.

[6] Shemetulskis, N. E., Weininger, D., Blankley, C. J., Yang, J. J., and Humblet, C., "Stigmata: An Algorithm To Determine Structural Commonalities in Diverse Datasets," J. Chem. Inf. Comput. Sci., vol. 36, pp. 862-871, 1996.

[7] Sheridan, R. P., "The Centroid Approximation for Mixtures: Calculating Similarity and Deriving Structure-Activity Relationships," J. Chem. Inf. Comput. Sci., vol. 40, pp. 1456-1469, 2000.

[8] Xue, L., Stahura, F. L., Godden, J. W., and Bajorath, J., "Fingerprint Scaling Increases the Probability of Identifying Molecules with Similar Activity in Virtual Screening Calculations," J. Chem. Inf. Comput. Sci., vol. 41, pp. 746-753, 2001.

[9] Xue, L., Godden, J. W., Stahura, F. L., and Bajorath, J., "Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys," J. Chem. Inf. Comput. Sci., vol. 43, pp. 1218-1225, 2003.

[10] Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A., "Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures," J. Chem. Inf. Comput. Sci., vol. 44, pp. 1177-1185, 2004.

[11] Schuffenhauer, A., Floersheim, P., Acklin, P., and Jacoby, E., "Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins," J. Chem. Inf. Comput. Sci., vol. 43, pp. 391-405, 2003.

[12] Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A., "Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures," Org. Biomol. Cheem., vol. 2, pp. 3256-3266, 2004.

[13] Whittle, M., Gillet, V. J., Willett, P., Alex, A., and Losel, J., "Enhancing the Effectiveness of Virtual Screening by Fusing Nearest-Neighbour Lists: A Comparsion of Similarity Cofficients," J. Chem. Inf. Comput. Sci., vol. 44, pp. 1840-1848, 2005.

[14] Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and 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., vol. 46, pp. 462-470, 2006.

[15] Wilton, D. J., Harrison, R. F., Willett, P., Delaney, J., Lawson, K., and Mullier, G., "Virtual Screening Using Binary Kernel Discrimination: Analysis of Pesticide Data," J. Chem. Inf. Model., vol. 46, pp. 471-477, 2006.

[16] Geppert, H., Horváth, T., Gärtner, T., Wrobel, S., and Bajorath, J., "Support-Vector-Machine-Based Ranking Significantly Improves the Effectiveness of Similarity Searching Using 2D Fingerprints and Multiple Reference Compounds," J. Chem. Inf. Model., vol. 48, pp. 742-746, 2008.

[17] Abdo, A. and Salim, N., "Similarity-Based Virtual Screening with a Bayesian Inference Network," ChemMedChem, vol. 9999, p. NA, 2008.

[18] Howard, R. T. and Croft, W. B., "Evaluation of an inference network-based retrieval model," ACM Trans. Inf. Syst., vol. 9, pp. 187-222, 1991.

[19] Berthier, A. N. R. and Richard, M., "A belief network model for IR," in Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval Zurich, Switzerland: ACM, 1996.

[20] Luis, M. d. C., Juan, M. F.-L., and Juan, F. H., "The BNR model: foundations and performance of a Bayesian network- based retrieval model," Int. J. Approx. Reasoning, vol. 34, pp. 265-285, 2003.

[21] Howard, R. T., "Inference networks for document retrieval," University of Massachusetts - USA, PhD. Thesis (1991)

[22] SciTegic Accelrys Inc. *http://www.SciTegic.com*

[23] Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S., "Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier," J. Chem. Inf. Comput. Sci., vol. 44, pp. 170-178, 2004.

[24] Pipeline Pilot Basic Chemistry Component collection v6.1 Student Edition

[25] The MDL Drug Data Report Database is available from MDL Information Systems Inc. *http://www.mdli.com*