Ligand-based Virtual screening using Fuzzy Correlation Coefficient

Ali Ahmed[1,2] Ammar Abdo[1] Naomie Salim[1]

[1]Faculty of Computer Science and Information Systems, Universiti Technologi Malaysia, 81310, Skudai Malaysia
[2]Faculty of Engineering, University of Karary, 12304, Khartoum, Sudan

ABSTRACT

Selection and identification of a subset of compounds from libraries or databases, which are likely to possess a desired biological activity is the main target of ligand-based virtual screening approaches. The main challenge of such approaches is achieving of high recall of active molecules. In this paper we presented fuzzy correlation coefficients (FCC), which is used as a similarity coefficient. The new approach is based on mutually dependent between molecular features, while most common approaches (Tanimoto, Bayesian and other coefficients) based on mutually independent between features. Our experiments have shown that the new coefficient increases the recall of active molecules in high diversity database compared with other correlation coefficients and Tanimoto.

Keywords: Correlation coefficients, fingerprint features, similarity search, similarity coefficients, virtual screening.

1. INTRODUCTION

Virtual screening (VS) refers to the use of a computer-based method to process compounds from a library or database of compounds in order to identify and select ones that are likely to possess a desired biological activity, such as the ability to inhibit the action of a particular therapeutic target. Selection of molecules with a virtual screening algorithm should yield a higher proportion of active compounds, as assessed by experiment, relative to a random selection of the same number of molecules [1].

Currently, VS becomes widely used in computer-based search for novel lead molecules. Typically, there are two approaches to the general problem: virtual screening by docking, when the 3D structure of the biological target (protein or enzyme) involved in the disease is available, and similarity-based virtual screening, where no information on the protein is necessary, instead, structural information of one or more known (bind to protein) molecules are used as structural query. The screening procedure retrieves molecules from the database according to the molecular similarity principle which states that structurally similar molecules exhibit similar biological activities.

Similarity searches are now a standard tool for drug discovery. The idea behind such searches is that, given a compound with an interesting biological activity is compared to other compounds. The basic idea of similarity-based Virtual Screening is a very simple and it was first enunciated explicitly by Johnson and Maggiora [2]; in which Similar Property Principle states that molecules that are structurally similar are likely to have similar properties.

The main goal of any system for similarity based screening is to quantify the degree of similarity or resemblance between reference structure (target query or queries) and each of the structures in database that is being screened for both real and virtual screening. A similarity measure requires three components: the molecules' representation that is used to characterize them when are being compared, the weighting scheme that priorities the importance of various components of these representations and the coefficient that is used to calculate the degree of similarity or relatedness between two structural representations.

This paper suggests a new ligand-based VS approach for similarity search. The new approach is based on the relationship between the target's molecules features and all molecules' features of in database.

In the next section, fuzzy text retrieval method is explained. In section 3, we overviews some related works belong to this area. Materials and methods are discussed in section 4, including our proposed method FCC and all our experiments. In section 5, results were presented including evaluation of the new method based on measurement of recall of active molecules. Finally, our conclusions are presented in section 6.

2. FUZZY TEXT RETRIEVAL METHOD

In information retrieval, keyword connection matrix is composed of number of keywords and their relationship [3].The value of that relationship represents the similarity between that two keywords. Relationship values range between 0 and 1, the value 0 means there is no relationship between those two words and value 1 means there is strong relationship between them.

Term-term correlation between two terms k_i and k_l is shown in eq (1) :

$$W_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}} \qquad (1)$$

Where n_i is the total number of documents containing term k_i , n_l is the total number of documents containing term k_l and $n_{i,l}$ is the total number of documents containing term k_i and k_l.

Let d_i denote the set of keywords indexed to the i^{th} document d, R_{ij} represent the relationship between i^{th} document and j^{th} keyword in the user query q as shown in eq(2)

$$R_{i,l} = \bigoplus_{k_i \in d_j} W_{i,l} \qquad (2)$$

Where $W_{i,l}$ is the relationship value between i^{th} and l^{th} keyword in the keyword connection matrix.

$\oplus$ denote the algebraic sum defined by $\oplus_i \; X_i = 1 - \prod_i (1 - X_i)$. Then eq (2) becomes

$$R_{il} = 1 - \prod_{k_l \in d_i} (1 - W_{il})$$

then the similarity between query q and document d is calculated as follow

$$sim(q, d_j) = \prod_{k_i \in d_j} R_{k_i, d_j} \qquad (3)$$

Finally, fuzzy model ranks the documents relative to the user query. The fuzzy model uses a term-term correlation matrix to compute the similarity between a document *dj* and fuzzy set index terms. The new approach described here uses the same concept of fuzzy text retrieval approach after modified its formula. The new formula of our approach is discussed in the section (4).

3. RELATED WORKS

Chemical information systems can offer three principal types of searching facility. Early systems provided two types of retrieval mechanisms: structure searching and substructure searching. These mechanisms were later complemented by another access mechanism that is the similarity searching.

Structure search involves searching a molecule database for a specified query molecule. A user will supply the complete structure of the molecule and the database is searched for a compound that matches perfectly with the target structure. This type of search is used to get some data about a particular compound, for example, its associated biological test results. Another use for this type of search is during the registration process of a new molecule, a procedure by which new compounds and accompanying data are added to a structure file and associated data files, respectively. A structure file in which a single and unique record of each compound is maintained is known as a registry file [4]. This is because before a molecule can be registered, the structure files needs to be searched in order to make sure that the compound is novel and has never been identified before.

Substructure searching involves the retrieval of molecules from the database that contain a user-defined query substructure. Substructure search is especially useful for finding structures containing a specified functional group, thus allowing the properties common to that group to be observed. Another use of substructure search is in the implementation of pharmacophoric pattern searching, where compounds containing a specific 3D substructure that has been identified in a molecular modeling study, are sought.

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 molecule in the database [5]. 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. There are many measures available to quantify the degree of similarity between a pair of molecules. The computational requirements of these measures vary depending on the level of detail used to represent the molecules that are being compared. Measures designed for highly complex representations require a lot of processing, thus limiting the number of database structures that can be compared in a given amount of time, such as the use of maximal common substructure [6]. Maximal common substructure (MCS) is the largest set of atoms or bonds from the target structure that can be superimposed exactly onto another structure, and is identified by using a maximal common subgraph isomorphism algorithm. Due to its NP-complete computational requirement [7], MCS algorithms have not been widely used for similarity searching to date.

A common application of similarity searching is in the rational design of new drugs and pesticides where the nearest neighbors for an initial lead compound are sought in order to find better compounds. Similarity searching is also used for property prediction purposes [8], where the properties of an unknown compound are estimated from those of its nearest neighbors. Related to the similar property principle is the concept of neighborhood behavior [9], which states that compounds within the same neighborhood or similarity region have the same activity. Unknown biological or physicochemical properties of a molecule can be predicted from the properties of molecules that lie within the same neighborhood region. In lead finding, selection of compounds whose neighborhood regions overlap one another should be avoided. In lead optimisation, if a particular compound is found to be active, compounds that lie in the same neighborhood region can be tested to find one with the most optimum activity.

Several methods have been used to further optimise the measures of similarity between molecules. These methods include weighting, standardisation and data fusion. A weighting scheme is used to differentiate between different features in a molecule, based on how important they are in determining the similarity of that molecule with another molecule. Certain molecular features can be emphasised by associating higher weights with them when calculating similarity. Different types of statistical information can be extracted from computerised representations of molecules to form the basis for a fragment weighting schemes. Many weighting schemes used in chemical information systems are derived from the general information retrieval literature, like the term-frequency and inverse document frequency. For example, higher weights can be given to attributes that occur frequently in a molecule, attributes that occur in small molecules and also attributes that occur less frequently in a data set.

Standardisation involves re-scaling of the variables in a multivariate analysis to ensure that all of them are measured on the same scale and that some of them do not dominate the overall similarities. Bath et al. [10] have evaluated some different of standardisation approaches for fragment based similarity measures and found that their use did not give significant improvement in performance when compared with non-standardised fragment occurrence data.

Different types of similarity measure focus on different molecular characteristics. Data fusion is the process of combining inputs from several similarity measures with information from other similarity measures, information processing blocks, databases or knowledge bases, into one representational format. The use of data fusion to combine several similarity measures can give an overall estimate of similarity based on several characteristics [11].

The process of data fusion involves computing several types of similarity measures, and combining the results using one of several fusion rules. The combined scores output by the fusion rule are then used to re-order the compounds to give the final ranked output. Holliday et al. [12] found that data fusion results in an increase in search effectiveness. In some cases where the use of fusion rule results in the assignment of the same score to two or more items, a further sort key is specified for the tied compounds. An example would be to sort the canonicalised connection tables of the tied compounds alphabetically. Weights can also be allocated to individual rankings based on some statistical observations of the coefficients' historical performances.

4. MATERIALS AND METHODS

This study has compared the retrieval results obtained using three different similarity-based screening systems. The first system was based on FCC. The second screening system was based on the tanimoto (TAN) coefficient which has been used for ligand-based virtual screening for many years and has been considered as a reference standard. The third screening system was based on correlation coefficients [13,14,15]. In the following paragraphs we give a detail description of FCC approach.

4.1 FCC-based Similarity Searching

The FCC give the similarity measure between the unknown chemical molecule and all the molecules stored in the chemical following equations:

$$sim(q,m) = \frac{q_i}{ml_i} \sum_{i=1,\, j=i+1}^{n} (1 - CF_{ij}) \qquad (4)$$

Where CF_{ij} is the correlation factor between i^th feature of molecule query and j^th feature of data base molecule as follow:

$$CF_{ij} = \frac{Q_{ij}}{Q_i + Q_j - Q_{ij}} \qquad (5)$$

Where Q_i is the total number of molecules containing feature i , Q_j is the total number of molecules containing feature j and Q_{ij} is the total number of molecules containing both feature i and j,

$\frac{q}{ml_i}$ is the ratio of active feature number that molecule q was participated with molecule ml_i, which q is the number of common bits set in both query and molecule i and ml_i is the number of bits set in molecule i

4.2 Tanimoto-based Similarity Searching

The second similarity search system used the binary form of the Tanimoto coefficient, which is applicable to binary data. The similarity score $S_{X,Y}$, in eq(6), computes the similarity between two molecular fingerprints, X and Y, of length n, in which a is the number of bits set in both X and Y, b is the number of bits set exclusively in X, c is the number of bits set exclusively in Y.

$$S_{X,Y} = \frac{a}{a+b+c} \qquad (6)$$

4.3 Correlation Coefficients-based Similarity Searching

Similarity coefficients are used to obtain a numeric quantification of the degree of similarity between a pair of structures. There are four main types of similarity coefficients: distance coefficients, association coefficients, correlation coefficients and probabilistic coefficients [13,14,15]. Correlation coefficients are generally used to measure the degree of correlation between sets of values representing the molecules, like the proportionality and independence between pairs of real-valued molecular descriptors. Correlation coefficients used in this comparison are: Pearson, Yule, McCon-naughey, Stiles and Dennis as listed in Table 1.

Table 1. Standard Correlation Coefficients

Coefficient	Formula
Pearson	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Yule	$\frac{ad-bc}{(a+b)(a+c)}$
McCon-naughey	$\frac{a^2-bc}{(a+b)(a+c)}$
Stiles	$\log_{10} \frac{n\left(\lvert ad-bc\rvert - \frac{n}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}$
Dennis	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$

Each coefficient computes the similarity between two molecular fingerprints, X and Y, of length n, in which a is the number of bits set in both X and Y, b is the number of bits set exclusively in X, c is the number of bits set exclusively in Y and d in the number of bits set in neither X or Y , so $n=a+b+c+d$.

4.4 Simulated Virtual Screening Experiments

All molecules in the database were converted to Pipeline Piolt's ECFC_4 (Extended Connectivity) fingerprint and flooded to a size 1024 [16]. For screening experiments, three data sets (DS1-DS3) were chosen [17] from MDDR data base.

The data set DS1 contains 11 activity classes, with some of the classes involving actives that are structurally homogeneous and with other involving actives that are structurally heterogeneous.

The DS2 data set contains 10 homogeneous activity classes and DS3 10 heterogeneous activity classes. The three data sets are listed in Table 2-4, each row of a table contains an activity class and the number of molecules belongs to that class. An inactive molecule in any search using one of MDDR data sets is one that has not been allocated the appropriate database activity descriptor.

For each data set DS1, DS2 and DS3, the screening experiments were performed with 20 references structures selected randomly from each activity class and the similarity measure obtains activity score for all of its compounds. Then we sort these activity scores in a descending order and the recall of the active compounds provides a measure of the performance of our similarity method. By recall of active compound, we mean the percentage of the desired activity class compounds that are retrieved in the top 1% and 5% of the resultant sorted activity scores.

5. RESULTS AND DISCUSSION

The results for searches of DS1, DS2 and DS3 are shown in Table 5, 6 and 7 respectively using cutoffs of both 1% and 5%. The left part of each table contains the activity number; the right part represents the result for each class using the seven coefficients; FCC, Tanimoto and five correlation coefficients. Each row in a table corresponds to one activity class and the last bottom row is the mean (Avg) when averaged overall of the classes for a data set, the similarity method with the best recall rate in each row is strongly shaded and the recall value is boldfaced.

For DS1 and DS2 searches Tanimoto and correlation coefficients have best overall performance at the 1% cutoff than FCC, but FCC has best result than Yule and McCon-naughey. And the same state happened at the 5% cutoff, that FCC has best performance than Yule and McCon-naughey.

Table 2. MDDR Activity Classes for DS1 Data Set

Activity Index	Activity class	Active molecules
31420	renin inhibitors	1130
71523	HIV protease inhibitors	750
37110	thrombin inhibitors	803
31432	angiotensin II AT1 antagonists	943
42731	substance P antagonists	1246
06233	substance P antagonists	752
06245	5HT reuptake inhibitors	359
07701	D2 antagonists	395
06235	5HT1A agonists	827
78374	protein kinase C inhibitors	453

78331	cyclooxygenase inhibitors	636

Table 3. MDDR Activity Classes for DS2 Data Set

Activity Index	Activity class	Active molecules
07707	Adenosine (A1) agonists	207
07708	Adenosine (A2) agonists	156
31420	Renin inhibitors 1	1300
42710	CCK agonists	111
64100	Monocyclic _lactams	1346
64200	Cephalosporins	113
64220	Carbacephems	1051
64500	Carbapenems	126
64350	Tribactams	388
75755	Vitamin D analogous	455
07707	Adenosine (A1) agonists	207

Table 4. MDDR Activity Classes for DS3 Data Set

Activity Index	Activity class	Active molecules
09249	Muscarinic (M1) agonists	900
12455	NMDA receptor antagonists	1400
12464	Nitric oxide synthase inhibitors	505
31281	Dopamine _hydroxylase inhibitors	106
43210	Aldose reductase inhibitors	957
71522	Reverse transcriptase inhibitors	700
75721	Aromatase inhibitors	636
78331	Cyclooxygenase inhibitors	636
78348	Phospholipase A2 inhibitors	617
78351	Lipoxygenase inhibitors	2111
09249	Muscarinic (M1) agonists	900

For DS3, FCC has best overall performance at both 1% and 5% cutoff compared with Tanimoto and all correlation coefficients. The DS3 searches are of particular interest since they include the most heterogonous activity classes.

Additional comparison was also done against previous result using Bayesian Inference Network (BIN) and Bayesian belief network (BNN) [18] and our method has shown best overall performance on DS3 at both 1% and 5% cutoff as shown in Table 8 It will be seen that the FCC provide a high level of performance compared with Tanimoto for the structurally diverse, DS3. These experiments suggest that FCC coefficient, is well suited for similarity search when we used high diverse, DS3.

Table 5. The recall is calculated using the top 1% and top 5% of the DS1 data sets when ranked using the FCC, Tanimoto and correlation coefficients

Activity Index	1%							5%						
	FCC	Tan	Pearson	Yule	MacCon	Stiles	Dennis	FCC	Tan	Pearson	Yule	MacCon	Stiles	Dennis
31420	55.04	69.55	67.81	58.42	65.99	68.08	66.61	79.66	86.28	85.42	81.48	78.47	85.5	85
71523	26.36	27.53	27.09	24.07	24.15	27.17	26.96	53.96	54.3	52.26	45.73	39.33	52.48	51.7
37110	22.5	23.18	22.69	19.11	18.63	22.86	22.62	36.31	45.54	44.2	36.82	30.69	44.33	43.49
31432	38.85	39.95	39.11	34.67	33.21	39.18	38.8	69.33	78.85	76.23	66.77	54.69	76.46	75.54
42731	15.45	16.95	16.59	13.34	12.01	16.64	16.21	22.59	25.93	25.23	21.63	17.05	25.38	24.92
06233	12.49	14.43	14.3	14.14	12.62	14.31	14.39	23.98	26.1	26.02	26.74	23.44	25.87	26.13
06245	5.17	6.23	5.87	5.31	5.28	6.01	5.81	12.51	14.05	13.91	12.99	11.28	14.11	13.94
07701	8.78	10.28	9.85	8.98	8.68	9.92	9.82	20.84	25.56	25.13	22.16	16.98	25.38	25.08
06235	9.02	10.86	10.71	10.13	9.38	10.74	10.67	21.02	23.81	23.96	23.79	20.59	23.93	24.15
78374	11.7	12.65	12.81	12.5	11.81	12.77	12.88	18.34	19.31	19.38	19.71	18.41	19.38	19.42
78331	6.76	6.74	6.96	7.32	7.01	6.93	6.98	13.94	12.74	12.94	14.44	15.56	12.93	13.06
Avg	19.28	21.67	21.25	18.91	18.98	21.33	21.07	33.86	37.5	36.79	33.84	29.68	36.89	36.58

Table 6. The recall is calculated using the top 1% and top 5% of the DS2 data sets when ranked using the FCC, Tanimoto and correlation coefficients.

Activity Index	1%							5%						
	FCC	Tan	Pearson	Yule	MacCon	Stiles	Dennis	FCC	Tan	Pearson	Yule	MacCon	Stiles	Dennis
07707	70.04	71.99	71.65	69.56	66.6	71.75	71.55	73.54	74.81	74.42	73.83	71.89	74.47	74.32
07708	83.16	97.16	96.39	90.65	88.32	96.65	95.81	96.45	99.95	99.94	99.42	96	99.94	99.94
31420	50.81	73.48	72.31	63.57	68.72	72.45	71.79	82.52	94.08	93.24	89.14	85.04	93.35	93.05
42710	77.82	80.55	81.64	80.64	78.27	81.45	81.36	90.36	91.36	91.09	90.55	88.73	91.27	91.09
64100	82.47	90.05	89.91	88.74	90.38	89.91	89.62	96.64	99.5	99.38	98.03	94.51	99.41	99.31
64200	71.43	68.13	68.48	65.45	60.89	68.21	68.84	90.89	98.21	97.95	94.55	87.14	97.95	97.77
64220	60.68	67.92	67.7	64.2	62.68	67.67	67.54	89.06	90.11	90.06	89	83.28	90.06	90.02
64500	51.52	73.44	70.96	63.28	57.52	71.12	69.92	64.32	89.44	86.48	77.04	68.08	86.88	85.28
64350	65.48	82.12	80.8	71.55	61.21	80.88	80.26	81.29	89.84	86.1	82.64	72.66	86.2	85.79
75755	97.86	97.8	97.84	97.8	96.85	97.84	97.84	98.24	98.26	98.26	98.26	98.04	98.26	98.27
Avg	71.12	80.26	79.77	75.54	73.14	79.79	79.45	86.33	92.56	91.69	89.25	84.54	91.78	91.48

Table 7. The recall is calculated using the top 1% and top 5% of the DS3 data sets when ranked using the FCC, Tanimoto and correlation coefficients.

Activity Index	1%							5%						
	FCC	Tan	Pearson	Yule	MacCon	Stiles	Dennis	FCC	Tan	Pearson	Yule	MacCon	Stiles	Dennis
09249	18.72	16.07	16.36	17.96	17.35	16.3	16.47	32.56	27.96	28.42	31.52	33.98	27.99	28.64
12455	10.24	8.55	8.84	9.51	9.78	8.77	8.9	15.39	12.33	12.87	15.08	18.18	12.63	13.05
12464	12.12	10	10.06	10.67	12.22	10.06	10.12	25.34	18.93	19.05	22.04	27.46	18.55	19.33
31281	30.38	23.33	23.43	27.24	32.19	23.14	23.62	59.85	35.71	36.19	42.57	58	34.76	36.48
43210	8.73	9.16	9.34	9.02	7.9	9.33	9.39	17.96	16.75	17.32	18.23	17.37	17.17	17.45
71522	5.55	5.41	5.46	5.38	4.88	5.46	5.51	9.84	9.99	10.19	10.03	9.5	10.13	10.19
75721	26.77	26.19	26.27	26.55	22.28	26.19	26.46	37.97	35.8	36.41	37.24	34.27	36.22	36.46
78331	10.16	9.81	9.86	9.95	9.5	9.8	9.81	19.65	16.83	17.15	18.6	19.73	16.98	17.28
78348	9.8	9.33	9.66	10.15	9.69	9.63	9.79	21.95	20.63	21.32	21.68	20.02	21.25	21.28
78351	18.06	14.97	15.49	16.53	15.23	15.44	15.65	19.81	16.98	17.65	19.21	18.65	17.59	17.82
Avg	15.03	13.28	13.48	14.3	14.1	13.41	13.57	25.98	21.19	21.66	23.67	25.72	21.33	21.8

Table 8. The recall is calculated using the top 1% and top-5% of the DS3 databases when ranked using the FCC, BIN and BNN.

Activity Index	1%			5%		
	FCC	BIN	BBN	FCC	BIN	BBN
09249	18.72	17.89	21.57	32.56	28.94	34.10
12455	10.24	6.28	7.47	15.39	12.17	15.47
12464	12.12	8.73	11.60	25.34	16.25	18.10
31281	30.38	26.26	31.37	59.85	34.95	43.26
43210	8.73	10.61	12.89	17.96	19.31	23.70
71522	5.35	3.29	3.54	9.84	7.04	7.57
75721	26.77	22.75	23.73	37.97	28.52	30.68
78331	10.16	5.10	5.88	19.65	10.08	12.68
78348	9.8	3.60	4.84	21.45	11.25	14.35
78351	18.06	4.12	4.84	19.81	9.58	12.59
Avg	15.03	10.86	12.77	25.98	17.81	21.25

6. CONCLUSION

This paper has further investigated the development and use of the FCC for ligand-based virtual screening. Experiments with MDDR database showed that this approach allows effective screening searches to be carried out especially on DS3.

7. REFERENCES

[1] Johnson, M. A. and Maggiora, G. M., Concepts and Application of Molecular Similarity:

John Wiley & Sons, New York 1990.

[2] Y. Ogawa, T. Morita and K. Kobayashi, Fuzzy document retrieval system and its learning method based on the keyword connection, in: Proc. Int. Workshop on Fuzzy System Applications (1988) 143-14.

[3] Ash, J., Chubb, P., Ward, S., Welford, S. and Willett, P. (1985). Communication, Storage and Retrieval of Chemical Information. Ellis Horwood Limited, Chichester.

[4] Willett, P., Barnard, J.M. and Downs, G.M. (1998). Chemical similarity searching. Journal of Chemical Information and Computer Sciences. 38:983-996.

[5] Willett, P. (1999a). Matching of chemical and biological structures using subgraph and maximal common subgraph isomorphism algorithms. In: Truhlar, D.G., Howe, W.J., Hopfinger, A.J., Blaney, J.D. and Dammkoehler, R. (Eds.) Rational Drug Design. Springer Verlag, New York. p. 11-38.

[6] Garey, M.R. and Johnson, D.S. (1977). Computers and Intractability: A Guide to the Theory of NP-Completeness. WH Freeman, San Francisco, CA.

[7] Carhart, R.E., Smith D.H. and Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definitions and applications. Journal of Chemical Information and Computer Science. 25:64-73.

[8] Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. and Weinberger, L.E. (1996). Neighbourhood behaviour: a useful concept for validation of molecular diversity descriptors. Journal of Medicinal Chemistry. 39 : 3060-3069.

[9] Bath, P.A., Morris, C.A. and Willett, P. (1993). Effect of standardisation of fragment-based measures of structural similarity. Journal of Chemometrics. 7:543.

[10] Salim, N., Holliday, J., and Willet, P., “Combination of Fingerprint-Based Similarity Coefficient Using Data Fusion”, Journal of Chemical Information and Computational Science, 2003, 43 pp. 435-442.

[11] Holliday, J.D., Hu, C-Y. and Willett, P. (2002). Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. Combinatorial Chemistry & High Throughput Screening. 5:155-166.

[12] Sneath PHA and Sokal RR (1973) Numeric taxonomy: the principles and practice of numerical classification. W.H. Freeman, San Francisco, 573 pp.

[13] Willett, P. (1987). Similarity and Clustering in Chemical Information Systems; Research Studies Press: Letchworth,

[14] Ellis, D.; Furner-Hines, J.; Willett, P. Perspect. (1994) Measuring the degree of similarity between objects in text retrieval systems.Inf. Manag. 3,128-149.

[15] Pipeline Pilot; Accelrys Software Inc.: San Diego, CA, 2008.

[16] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, New methods for ligand-based virtual screening: use of data-fusion and machine learning techniques to enhance the effectiveness of similarity searching, J. Chem. Inf. Model. 46 (2006) 462–470.

[17] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, New methods for ligand-based virtual screening: use of data-fusion and machine learning techniques to enhance the effectiveness of similarity searching, J. Chem. Inf. Model. 46 (2006) 462–470.

[18] Ammar Abdo, Beining Chen, Christop Muller, Naomie Salim, and Peter Willett, Ligand-Based Virtual Screening Using Bayesian Networks , J.Chem.Inf.Model., 50,1012-1020, 2010.