2D-Based Virtual Screening using Multiple Bioactive Reference Structures

Jérôme Hert*, Peter Willett*, David J. Wilton*, Pierre Acklin†, Kamal Azzaoui†, Edgar Jacoby†, Ansgar Schuffenhauer†.

The University of Sheffield

NOVARTIS

INTRODUCTION

Virtual Screening (VS) methods can be classified according to the amount of chemical and biological data that they require. When several active compounds are available, the normal screening approach is pharmacophore mapping followed by 3D database search. However, the technique is not always applicable given the heterogeneity that characterises typical HTS hits. This study hence describes several approaches to the combination of the structural information that can be gleaned from multiple *reference structures* and then evaluates their effectiveness, as well as the effectiveness of several descriptors, by means of simulated VS experiments.

DATABASE SEARCHES

The experiments were carried on the *MDL Drug Data Report* (MDDR) database in which 102 535 compounds were available to us.[1] These molecules were searched using the 11 activity classes detailed in table 1.

For each activity class, 10 active compounds were randomly selected for use as the training set. Each search was repeated 10 times using 10 different training sets.

A note was made of the percentage of the active molecules (i.e., those in the same class as those in the training set) that occurred in the top 1% and the top 5% of the ranking resulting from that search. The results reported are the means and standard deviations for these recall values, averaged over each set of 10 searches.

Table 1. MDDR activity classes used in the study.

Activity name	Nb of compounds	Similarity Mean	Similarity SD
5HT3 antagonists	752	0.351	0.116
5HT1A agonists	827	0.343	0.104
5HT reuptake inhibitors	359	0.345	0.122
D2 antagonists	395	0.345	0.103
Renin inhibitors	1130	0.573	0.106
Angiotensin II AT1 antagonists	943	0.403	0.101
Thrombin inhibitors	803	0.419	0.127
Substance P antagonists	1246	0.399	0.106
HIV protease inhibitors	750	0.446	0.122
Cyclooxygenase inhibitors	636	0.268	0.093
Protein kinase C inhibitors	453	0.323	0.142

REFERENCES

1. The MDL Drug Data Report database is available from MDL Information Systems at www.mdli.com.
2. Shemetulskis, N.E. et al., *J. Chem. Inf. Comp. Sci.*, **1996**, *36*, 862–871.
3. Cramer, R.D. at al., *J. Med. Chem.*, **1974**, *17*, 533-535
4. Harper, G. et al., *J. Chem. Inf. Comp. Sci.*, **2001**, *41*, 1295–1300.
5. Barnard Chemical Information Ltd.
6. Daylight Chemical Information Systems Inc.
7. Tripos Inc.
8. Willett, P., *J. Chem. Inf. Comp. Sci.*, **1979**, 19, 159–162.
9. Scitegic Inc.
10. Schuffenhauer A. et al., *J. Chem. Inf. Comp. Sci.*, **2003**, 43, 391–405.
11. Schneider G. et al., *Angew. Chem. Int. Ed.*, **1999**, *38*, 2894–2896.

The first part of this work has been published in the Journal of Chemical Information and Computer Science, **2004**, ASAP.

COMPARISON OF VIRTUAL SCREENING METHODS

Single Fingerprint Methods. Following earlier work on Stigmata,[2] two methods involving the creation of a single, combined, fingerprint are described. The first method, referred to as *Modal*, generates a *modal fingerprint* from the training set (figure 1) which is then used as a query. The compounds of the database are ranked using the Tanimoto coefficient. During initial experiments the best results were obtained from a threshold percentage of 40% which was hence adopted as the default value. A second, *weighted*, approach computes a weighted vector (figure 1). The continuous version of the Tanimoto coefficient is used to rank the compounds.

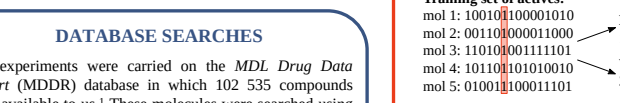

Figure 1. Illustration of the generation of the modal and the weighted fingerprints.

Data Fusion Methods. In data fusion, inputs from different sensors are combined. We fuse the ranks and the scores obtained with the Tanimoto coefficient from the different molecules in the training set. Combination using the SUM and the MAX fusion rules are investigated (figure 2). The fusion of the scores using the MAX rule provide the most effective combination in initial experiments.

$$\text{SUM}: \sum_{i=1}^{N} s_i, \quad \text{MAX}: \text{Maximum}\ \{s_1, s_2, ..s_i..s_{n-1}, s_n\}$$

Figure 2. Fusion rules used for the data fusion method, where s_i is the score obtained with the ith member of the training set.

Substructural Analysis Methods. In substructural analysis (SSA), a weight is assigned to each bit.[3] The weights of the bits set in a given molecule are summed to produce the score. As we do not have information about inactive compounds, we make the assumption that the overall characteristics of the inactive training set are mirrored by those of the entire database and choose a weighting scheme that does not make explicit use of the inactives (figure 3).

Binary Kernel Discrimination (BKD) uses a kernel function to weight the different contributions of a training set of active and inactive compounds (figure 4).[4] The score is given by the probability that a compound is active (figure 4). We suggest that the characteristics of the inactives should be approximated by the characteristics of the entire database and generate inactive training sets of 100 compounds by randomly selecting compounds from the entire database.

$$R_1(i) = \log\left(\frac{A_j}{T_j}\right)$$

Figure 3. Weighting scheme used for the Substructural Analysis method where A_j is the number of actives and T_j the total number of compounds with bit j set.

$$K_\lambda(i,j) = \lambda^{N-d_{i,j}}(1-\lambda)^{d_{i,j}}$$

$$L_A(j) = \frac{\sum_{i \in actives} K_\lambda(i,j)}{\sum_{i \in inactives} K_\lambda(i,j)}$$

Figure 4. K_λ is the kernel function used in the BKD method, given two compounds, i and j represented by fingerprints containing N bits and differing in $d_{i,j}$ of those bits. λ is a smoothing parameter that needs to be optimised. $L_A(j)$ is an estimate of compounds j to be active.

Results. Data fusion and BKD are clearly the methods of choice, consistently outperforming the other approaches. The performance of the methods tends to increase with the self-similarity of the activity classes; however a good virtual screening is obtained even for diverse sets of compounds (table 2 and figure 5).

By comparing the average and maximum results obtained in a single similarity search, we can demonstrate that using 10 active reference structures and combining them using the BKD or the data fusion methods gives comparable results to the best possible conventional similarity search.

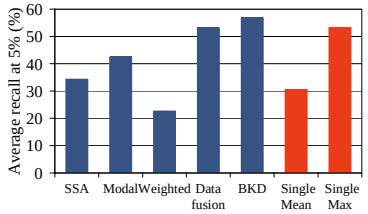

Figure 5. Comparison of the average recall at 5% over all activity classes obtained for the various methods investigated using Unity fingerprints. In addition, the average results for the average and the maximum recalls of single similarity searching are showed.

Table 2. Comparison of the average percentage of active compounds retrieved by the various methods over the top 5% of the ranked test set using Unity fingerprints

Activity name	SSA Mean	SSA SD	Modal Mean	Modal SD	Weighted Mean	Weighted SD	Data fusion Mean	Data fusion SD	BKD Mean	BKD SD
5HT3 antagonists	29.27	5.07	30.31	5.21	2.99	2.14	49.03	5.43	52.32	8.27
5HT1A agonists	30.13	2.21	21.85	2.29	1.05	0.79	37.15	4.06	38.19	7.03
5HT reuptake inhibitors	33.12	4.72	39.63	3.67	10.14	6.19	49.68	5.45	45.82	7.93
D2 antagonists	27.51	3.06	27.12	5.35	4.83	3.47	37.40	4.92	38.65	7.38
Renin inhibitors	52.94	6.67	88.77	3.15	73.84	4.92	88.62	1.90	93.34	1.35
Angiotensin II AT1 antagonists	43.40	6.66	73.63	5.64	51.55	3.58	80.44	6.08	84.47	6.59
Thrombin inhibitors	35.64	7.69	49.43	7.31	22.50	4.24	58.58	8.98	63.06	7.66
Substance P antagonists	36.52	6.60	36.80	5.08	14.51	3.12	47.14	5.16	58.39	8.27
HIV protease inhibitors	34.05	6.22	53.53	4.45	40.88	8.45	61.62	7.85	68.45	8.31
Cyclooxygenase inhibitors	19.20	3.48	10.96	3.07	5.30	2.53	26.52	7.15	33.15	4.68
Protein kinase C inhibitors	35.58	6.69	35.60	10.33	21.67	5.25	48.01	8.99	49.37	10.84

COMPARISON OF DESCRIPTORS

Sutructural keys. Structural keys are represented as a Boolean array, each entry of which represents the presence or absence of a specific 2D fragment (figure 6). We investigate the use of the BCI[5] fingerprints.

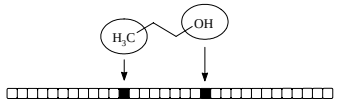

Figure 6. Simple illustration of bit-string encoding using the dictionary-based approach.

Hashed fingerprints. Hashed fingerprints indexe a set of patterns, e.g., (atom-bond-atom)$_n$. Each pattern is then hashed to produce one or a few bits over the bit-string (figure 7). We consider the Unity[7], Daylight[6] and Avalon fingerprints.

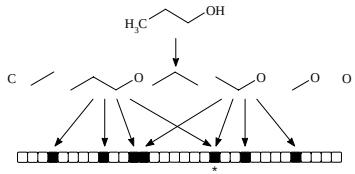

Figure 7. Simple Illustration of a hashing scheme. The asterisk denotes an element in the bit-string where a collision has resulted from the hashing procedure.

Circular Substructures. A circular substructure is a fragment descriptor where each atom is represented by a string of integers which are obtained by an adaptation of the Morgan algorithm.[8]

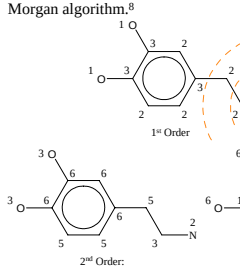

Figure 8. Illustration of the first, second and third order connectivity of dopamine.

The discrimination between the atoms is based upon the extended connectivity values where the n^{th} order connectivity is calculated by summing the $(n-1)^{th}$ order connectivities of all immediately adjacent atoms (figure 8). We include the Scitegic Extended Connectivity Fingerprints (ECFPs) and Functional Connectivity Fingerprints (FCFPs) in our study.[9]

Pharmacophore vectors. Two pharmacophore vectors are investigated: Similog keys[10] and CATS[11]. Similog keys are build according to a compact "DABE" atom-typing scheme based on four properties: hydrogen-bond donor, hydrogen-bond acceptor, electropositivity and bulkiness (figure 9). CATS is based on five atom-types: hydrogen-bond donor, hydrogen-bond acceptor, positive and negative and lipophilic. The vector is composed of counts of pairs of those generalised atoms for distances up to 10 bonds.

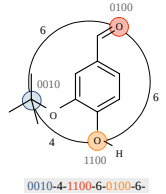

Figure 9. Example of a Similog key.

Results. The circular substructure fingerprints significantly outperform the other descriptors (figure 10). The best results are obtained when using ECFP_4, which suprisingly is more effective when used with the data fusion method.

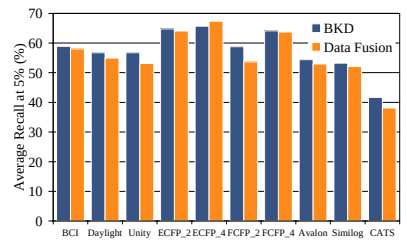

Figure 10. Comparison of the average recall at 5% over all activity classes obtained from the data fusion and the BKD methods.

CONCLUSION

We reviewed the use of virtual screening when multiple reference structures are available, evaluating five different methods and ten different descriptors. Experiments on the MDDR database demonstrate that data fusion based on similarity scores and an approximation of the BKD method are by far the best approaches and that circular substructures, especially ECFP_4, are the most effective fingerprints we have considered.

* Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.

† Novartis Institutes for Biomedical Research, Discovery Technologies, Compound Logistics and Properties Unit, Molecular and Library Informatics Program, CH-4002 Basel, Switzerland