New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching

Jérôme Hert, Peter Willett,* and David J. Wilton

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

Pierre Acklin, Kamal Azzaoui, Edgar Jacoby, and Ansgar Schuffenhauer

Novartis Institutes for BioMedical Research, Discovery Technologies, CH-4002 Basel, Switzerland

Received August 26, 2005

Similarity searching using a single bioactive reference structure is a well-established technique for accessing chemical structure databases. This paper describes two extensions of the basic approach. First, we discuss the use of group fusion to combine the results of similarity searches when multiple reference structures are available. We demonstrate that this technique is notably more effective than conventional similarity searching in scaffold-hopping searches for structurally diverse sets of active molecules; conversely, the technique will do little to improve the search performance if the actives are structurally homogeneous. Second, we make the assumption that the nearest neighbors resulting from a similarity search, using a single bioactive reference structure, are also active and use this assumption to implement approximate forms of group fusion, substructural analysis, and binary kernel discrimination. This approach, called turbo similarity searching, is notably more effective than conventional similarity searching.

INTRODUCTION

Similarity searching is perhaps the simplest tool available for ligand-based virtual screening of chemical databases, requiring just a single known bioactive molecule, the *reference* or *target* structure, as the starting-point for a database search.[1] The traditional approach to similarity searching, first described by Carhart et al.[2] and Willett et al.,[3] involves comparing the reference structure with each of the database structures in turn, computing a quantitative measure of structural similarity in each case, and then returning the most similar molecules, the *nearest neighbors* (NNs), as the output from the search. The similar property principle states that structurally similar molecules are likely to have similar properties, and the nearest neighbors of a bioactive reference structure are hence also expected to exhibit this activity.[4] There is considerable evidence to suggest that this is, indeed, often the case in practice,[5−10] albeit with exceptions to this general rule.[11]

Many types of similarity measure have been discussed in the literature,[1,12−14] but by far the most common involves the use of a simple association coefficient, normally the Tanimoto coefficient, with a 2D fragment bit-string representation of molecular structure. We have recently described two approaches that can enhance the effectiveness of virtual screening based on such similarity measures. First, we described the use of *group fusion* (vide infra) when not one but several bioactive reference structures are available.[15−17] Second, we have recently shown how nearest-neighbor information and group fusion can be used even when only the conventional, single reference structure is available, an approach that we refer to as *turbo similarity searching* (TSS)[18] and that involves the assumption that the nearest neighbors of a bioactive reference structure are also active. In this paper, we report further developments of these two approaches.

In the next section, we discuss the suitability of group fusion for *scaffold-hopping* applications.[19] Scaffold-hopping (other names that have been used include *leapfrogging*, *lead hopping*, and *scaffold searching*) involves finding chemical structures that exhibit the same biological activity but that have significant topological differences. The identification of such nonclassical bioisosteres[20] has occasioned much interest in recent years, as it provides ways of enhancing the pharmacological properties of known leads and of entering unpatented portions of chemical space.[21−26] We then report a development of turbo similarity searching that again utilizes information about the nearest neighbors of a reference structure but that here enables the use of machine-learning techniques for virtual screening without the need for an explicit training set of known active and known inactive molecules; we again focus on the use of these methods for scaffold-hopping applications. In both cases, we have evaluated the effectiveness of the approaches using simulated virtual screening searches on the MDL Drug Data Report (MDDR) database.[27] After removal of duplicates and molecules that could not be processed using local software, the version of the file used here contained a total of 102 514 molecules. Searches were then carried out for molecules belonging to specific activity classes (the selection of which is discussed further below) and the success of a search determined by the extent to which it was possible to retrieve molecules exhibiting the required activity. Fuller details of all of the experiments reported here are presented by Hert.[28]

* To whom all correspondence should be addressed. Tel.: +44-114-2222633. Fax: +44-114-2780300. E-mail: p.willett@sheffield.ac.uk.

10.1021/ci050348j CCC: $33.50
Published on Web 01/04/2006

USE OF MULTIPLE REFERENCE STRUCTURES FOR SCAFFOLD HOPPING

Data fusion (or *consensus scoring*) involves combining the results of different similarity searches of a chemical database.[29,30] The normal approach to data fusion, called *similarity fusion*, involves carrying out searches with a single reference structure but using multiple similarity measures. For example, one might combine searches for a specific reference structure that had been carried out with several different types of fingerprints or of similarity coefficients.[17] Group fusion, conversely, involves carrying out searches with a single similarity measure but using multiple reference structures. Specifically, assume that some database structure *i* yields similarity scores (Tanimoto coefficient values in our experiments) of s_1, s_2, ..., s_n with *n* different reference structures; then, we have shown that effective searches are obtained by ranking the database molecules on the basis of the maximum of these scores, that is, max{s_1, s_2...s_i...s_{n-1}, s_n}; we refer to this as the MAX fusion rule.[15]

Our initial experiments (as detailed in ref 15) involved searches for 11 MDDR activity classes that had been chosen from the database such that the mode of action is known, the activity is of current pharmaceutical interest, and there is a substantial number of MDDR molecules categorized as exhibiting that activity. The data sets chosen were quite disparate in nature, some of them being structurally homogeneous (e.g., rennin and HIV-1 protease inhibitors) while others were structurally diverse (e.g., cyclooxygenase and protein kinase C inhibitors); the diversity was estimated by the mean pairwise similarity (hereafter, MPS) across each set of active molecules. Hardly surprisingly, it was found that the search effectiveness increased broadly in line with this mean self-similarity; however, acceptable results were obtained even with the more heterogeneous activity classes.[15] Subsequent experiments using data sets from both MDDR and the Dictionary of Natural Products[31] suggested that, while the absolute level of performance might not be very high for such data sets, the increase in performance, relative to conventional similarity searching based on just a single reference structure, was greatest for the most diverse activity classes.[17] These experiments suggest that group fusion may be of greatest value in those cases where conventional similarity searching (and conventional similarity fusion) is least effective, that is, that it might prove to be a useful tool for scaffold hopping. We have hence carried out further experiments to test the generality of this observation; to this end, we have identified MDDR activity classes that, while satisfying the criteria mentioned above, additionally are as structurally diverse as possible.

One could use trained medicinal chemists to comment on the diversity of a set of molecules, but human judgment is known to be highly variable;[32,33] instead, we have again used the MPS, when averaged over all of the pairs of molecules within a specific MDDR activity class, to quantify the diversity. This is easy to compute and has been shown to give results that mirror those obtained with algorithmic scaffold definitions.[18] The MPS requires a similarity measure to be defined, that is, a coefficient (the Tanimoto coefficient in this case) and a structural representation, so that the individual intermolecular structural similarities can be computed. In our initial study of group fusion,[15] we used just a single descriptor (Unity 2D fingerprints[34]) to assess the diversities of the activity classes that were used. To ensure the generality of our results and to remove any unintended bias, we have here used multiple descriptors to choose the most diverse activity classes. Specifically, we have computed the MPS using the Tanimoto coefficient and the following types of 2D fingerprints: one structural-key fingerprint [the 1052-bit Barnard Chemical Information (BCI) fingerprints; three hashed fingerprints, 2048-bit Daylight, 988-bit Unity, and 2048-bit Avalon fingerprints]; four circular substructure fingerprints (ECFP_2, ECFP_4, FCFP_2, and FCFP_4 from Scitegic Inc.); and two pharmacophore representations (Similog and CATS). These fingerprints and their implementation are described by Hert et al.[16]

The local version of the MDDR database contains 707 different activity classes. Of these, we considered only the 309 activity classes that were covered in a ligand ontology linking the activity classes to a hierarchical classification of protein targets.[35] A total of 48 of these 309 classes had less than two members in our file, and the MPS was then computed for each of the remaining 261 activity classes, using each of the 10 different types of fingerprints listed above. The final MPS for each class was then taken as the mean when averaged over the values for the individual fingerprint types, and the 10 most diverse (lowest average MPS) classes were selected, subject to them containing at least 50 compounds. These activity classes are listed in Table 1a; only one of them, the cyclooxygenase inhibitors, had figured in our previous experiments. For comparison with the MPS values in this table, the MPS for 10 000 compounds selected at random from MDDR was 0.200, demonstrating the highly disparate natures of the data sets listed in Table 1a. In addition, 10 activity classes were similarly selected that had medium average MPS values and a further 10 that had the highest average MPS values (i.e., very homogeneous sets of compounds), as detailed in parts b and c of Table 1, respectively.

Simulated virtual screening searches were carried out on the MDDR database using the procedures described in our previous papers.[15−18] A set of 10 molecules was picked at random from an activity class. A similarity search of the database was carried out for each selected active, and the 10 resulting sets of similarity scores were combined using the MAX fusion rule to give the output from the search. The procedure was then repeated using nine further sets of reference structures, and in each search, a note was made of the *recall*, that is, the percentage of the active molecules (i.e., those in the same class as those in the reference set) that occurred in the top 5% of the ranking resulting from the group fusion search; the mean recall was then calculated when averaged over these 10 group fusion searches. The procedure was repeated over each of the 30 chosen activity classes (10 high diversity, 10 medium diversity, and 10 low diversity as detailed in Table 1) and the mean recall calculated when averaged over the 10 activity classes of each type. The whole procedure was repeated for each of the 10 different types of fingerprint. The results of these runs are summarized in Table 2. This table also contains the results for conventional similarity searches, that is, when just a single reference structure is used (with the results in the left-hand portion of the table being the mean values when averaged over similarity searches based on all of the active molecules

Table 1. (a) 10 MDDR Activity Classes Selected as Having the Lowest Average MPS Scores, and Hence as Being the Most Diverse for Virtual Screening Experiments; (b) 10 MDDR Activity Classes Selected as Having Medium Average MPS Scores, and Hence as Being of Medium Diversity for Virtual Screening Experiments; and (c) 10 MDDR Activity Classes Selected as Having the Highest Average MPS Scores, and Hence as Being the Least Diverse for Virtual Screening Experiments

(a) Lowest Average MPS Scores

activity class	MDDR activity key	number of compounds	average MPS
muscarinic (M1) agonists	09249	848	0.206
NMDA receptor antagonists	12455	1311	0.199
nitric oxide synthase inhibitors	12464	377	0.189
dopamine β-hydroxylase inhibitors	31281	95	0.229
aldose reductase inhibitors	43210	882	0.232
reverse transcriptase inhibitors	71522	519	0.218
aromatase inhibitors	75721	513	0.229
cyclooxygenase inhibitors	78331	636	0.220
phospholipase A2 inhibitors	78348	704	0.224
lipoxygenase inhibitors	78351	2555	0.224

(b) Medium Average MPS Scores

activity class	MDDR activity key	number of compounds	average MPS
CRF antagonists	06215	254	0.315
5 HT2B antagonists	06249	90	0.308
5 HT2C antagonists	06250	174	0.308
dopamine (D1) antagonists	07702	167	0.315
carbonic anhydrase inhibitors	16200	255	0.310
thrombin inhibitors	37110	803	0.321
CCK A antagonists	42712	161	0.323
oxytocin antagonists	44210	209	0.310
protease inhibitors	78330	574	0.313
phosphodiesterase V inhibitors	78437	164	0.320

(c) Highest Average MPS Scores

activity class	MDDR activity key	number of compounds	average MPS
adenosine (A1) agonists	07707	88	0.524
adenosine (A2) agonists	07708	71	0.536
renin inhibitors	31420	1130	0.459
CCK agonists	42710	79	0.452
monocyclic β-lactams	64100	76	0.549
cephalosporins	64200	1312	0.501
carbacephems	64220	73	0.487
carbapenems	64500	896	0.457
tribactams	64530	74	0.548
vitamin D analogues	75755	279	0.574

Table 2. Mean Recalls at 5% Averaged over the 10 Fused Searches for Each Class and over the Classes Selected as Having High, Medium, or Low MPS Scores (See Table 1)[a]

	similarity search			group fusion			BKD		
fingerprint type	high	medium	low	high	medium	low	high	medium	low
BCI	92.3	41.2	20.8	91.3	73.8	47.4	91.5	74.9	52.2
Daylight	92.8	36.1	18.3	91.1	70.4	42.6	88.0	62.1	45.6
Unity	92.6	38.4	16.6	90.9	70.3	40.3	89.6	69.0	51.0
Avalon	92.9	40.0	20.1	91.2	71.0	45.1	86.3	66.2	49.5
ECFP_2	93.3	43.6	22.0	91.8	76.8	49.5	91.9	77.2	53.5
ECFP_4	92.2	42.9	21.0	91.9	77.2	50.1	91.8	77.8	51.0
ECFP_6	92.9	41.8	19.9	91.9	76.4	48.2	91.7	77.2	47.8
FCFP_2	83.5	37.0	18.3	89.4	67.7	41.2	90.2	71.3	48.2
FCFP_4	91.8	39.2	20.0	91.6	74.5	48.5	91.6	75.6	49.4
FCFP_6	92.1	38.7	20.2	91.6	75.0	49.3	91.6	75.6	47.5
Similog	78.7	35.3	18.5	87.6	64.7	32.6	75.1	46.4	38.8
CATS	71.0	29.5	16.4	82.1	48.3	20.4	88.0	55.3	35.5

[a] The fingerprint types are discussed in detail in ref 16.

in each of the chosen classes), and for searches using binary kernel discrimination (BKD; vide infra).

Two conclusions may be drawn from the results in Table 2. First, it is clear that the fingerprints based on circular substructures are the most effective of those tested here in the group fusion experiments, with the best performance being obtained across the various types of activity class with the ECFP_2 and ECFP_4 fingerprints from Scitegic's Pipeline Pilot system.[36] This result is in line with our previous study,[16] which considered just 11 activity classes spanning

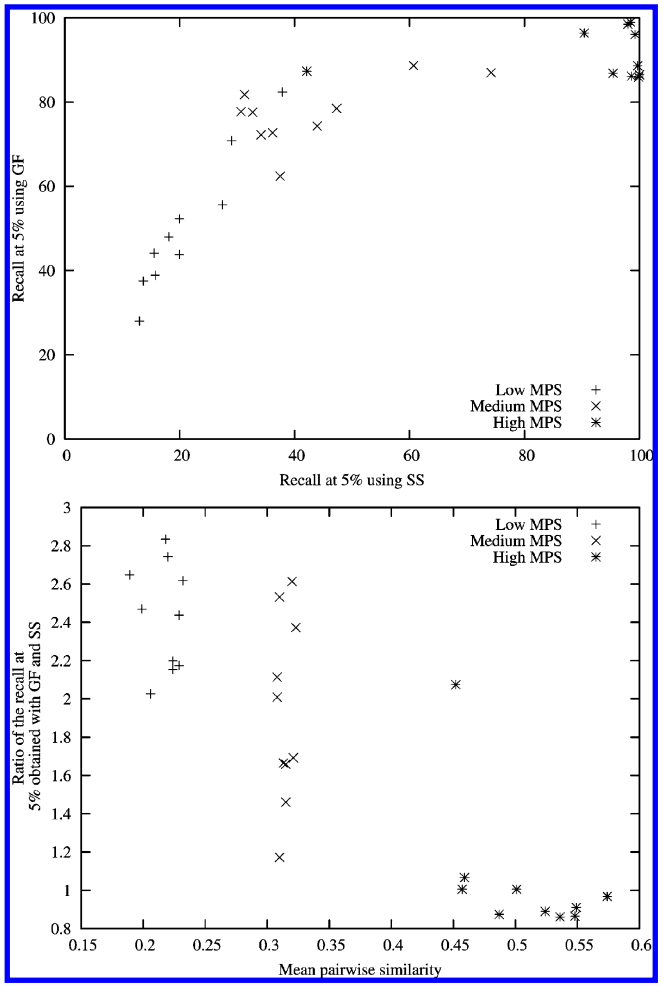

Figure 1. Comparison of the average recall at 5% obtained with group fusion (GF) and similarity searching (SS) using the ECFP_4 descriptors and 30 MDDR activity classes chosen as being of low, medium, and high diversity. The upper part of the figure shows the recall at 5% obtained with GF versus recall at 5% obtained with SS, while the lower part shows the diversity (as measured by the mean pairwise similarity) versus the ratio of the recalls at 5% obtained with GF and with SS.

a range of levels of diversity. The second conclusion that can be drawn, and the one that is of importance in the context of scaffold hopping, is the effectiveness of group fusion when used with the high-diversity (low MPS score) data sets. As would be expected, the absolute recall values for these searches are much lower than those for the medium-MPS and high-MPS searches; however, the recall values relative to those for conventional similarity searching are much higher, more than doubling the numbers of retrieved actives for most of the fingerprint types. It may be argued that merely doubling the recall is rather disappointing given that 10 times as many structures are being used as input to the search algorithm. However, there is likely to be at least some structural redundancy in any set of reference structures; moreover, in many cases, even conventional similarity searching with a single reference structure can retrieve a significant fraction of the total actives, and there is hence only a limited number of additional actives available for retrieval. It is hence most unlikely that one can achieve a linear increase in recall commensurate with the number of reference structures.

This increase in relative performance as the sought structures become more heterogeneous is illustrated in Figure 1, which shows the recall at 5% obtained with group fusion and similarity searching using the ECFP_4 descriptors and the 30 MDDR activity classes listed in Table 1. The upper part of the figure compares the recalls resulting from the use of the two methods, and it will be seen that many of the group fusion results lie well-above the 45° diagonal that would be obtained if the two approaches gave comparable recalls; moreover, the distance above the line is greatest for the most diverse and medium diverse data sets. This difference is further demonstrated in the lower part of the figure, which plots the ratio of the two recalls against the MPS, with the largest performance enhancements resulting from the most diverse data sets (i.e., lowest MPS values).

A comparable increase in relative performance is obtained if an approximate form of BKD[37] is used with the multiple reference structures, as described in detail by Hert et al.[15] In brief, BKD involves computing a kernel function that is based on the Hamming distance between a pair of fingerprints and on a smoothing parameter, λ, the value of which is optimized using the available training-set information. The fingerprint representing a database molecule, j, is matched against the fingerprints for each of the active and inactive molecules in the training-set and its score then computed as the ratio of the sum of the kernel scores for the training-set actives to the corresponding sum for the inactives. In the similarity context, the multiple reference structures comprise the actives in the training-set actives and the training-set inactives (or, rather, assumed inactives) are obtained by random selection from the database that is to be searched, subject to the qualification that none of the resulting pairs of molecules had a similarity coefficient greater than 0.80 using Unity fingerprints and the Tanimoto coefficient. Figure 2 illustrates the variation in relative performance with the diversity of the sought actives; the trends here are entirely comparable to those observed with group fusion.

USE OF MACHINE-LEARNING METHODS WITH A SINGLE REFERENCE STRUCTURE

We have recently introduced what we refer to as turbo similarity searching. In much the same way that the turbocharger in an automobile increases the power of an engine by using the engine's exhaust gases, TSS increases the power of a similarity searching engine by using the reference structure's nearest neighbors. TSS is based on the general applicability of the similar property principle and on our studies of group fusion as reported above and previously: if the similar property principle is applicable to a particular biological system, then the NNs of a bioactive reference structure are also likely to possess that activity, and we would hence expect to find further active structures if we were to use them, in turn, as reference structures. The resulting rankings of the database that is being searched can then be combined using a group fusion search to give a fused output that would be expected to have a greater percentage of high-ranked actives than would a conventional similarity search using just a single reference structure. This expectation is based on the *assumption* that the NNs are also active. In practice, of course, only some of them will be active, but the similar property principle would suggest that we could obtain improvements in performance even if this is not

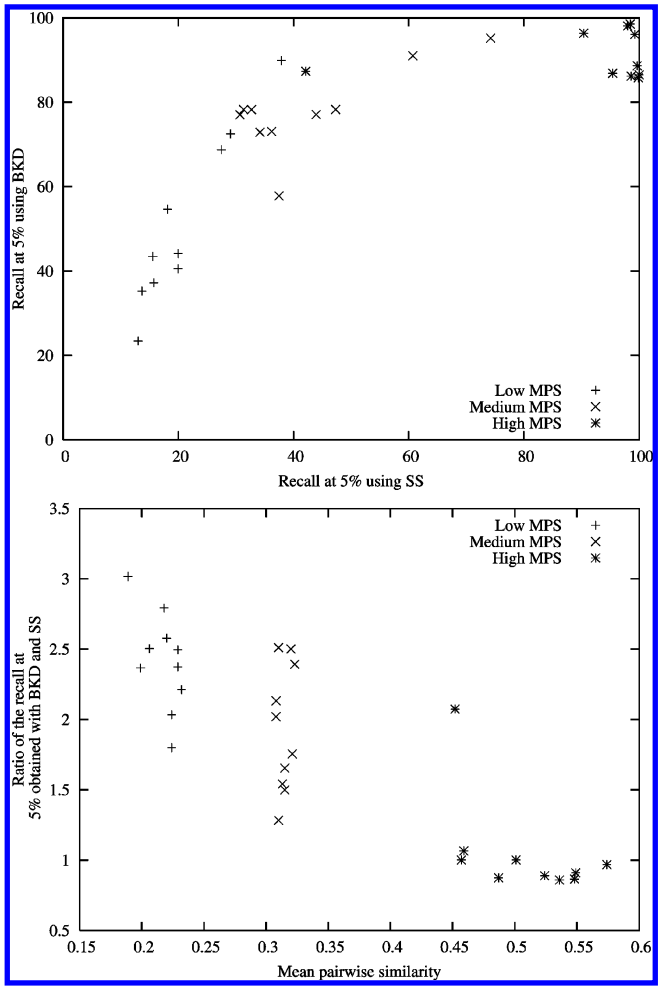

Figure 2. Comparison of the average recall at 5% obtained with binary kernel discrimination (BKD) and similarity searching (SS) using the ECFP_4 descriptors and 30 MDDR activity classes chosen as being of low, medium, and high diversity. The upper part of the figure shows the recall at 5% obtained with BKD versus recall at 5% obtained with SS, while the lower part shows the diversity (as measured by the mean pairwise similarity) versus the ratio of the recalls at 5% obtained with BKD and with SS.

uniformly the case. We have recently carried out experiments using the initial set of 11 MDDR activity classes (vide supra) and shown that such improvements can be obtained in practice. Specifically, we have used the search procedure shown in Figure 3 and found that using searches with $k = 100$ can enhance the recall of virtual screening by about 5% at minimal computational cost. Thus, even though the assumption underlying TSS is not entirely correct (and even though the incomplete annotation that characterizes the MDDR may further confuse the picture), it is sufficiently correct to yield enhancements in performance in practice, and this has suggested the further studies reported here. In like vein, Klon et al. have considered high-scoring molecules in a docking study to be active (irrespective of whether this is the case) and then used this information for substructural analysis (SSA).[38]

We note in passing at this point that there is a further way in which NN information could be used. Rather than using the procedure detailed in Figure 3, one could adopt a sequential procedure: the first search is based on the original reference structure, the second search is based on the top-ranked molecule from the first search, the third search is based on the top-ranked molecule from the second search, and so forth. We have not, thus far, evaluated this approach since we believe that the search may move too rapidly away from the known active starting point for the assumption of NN activity to hold to any great extent. It may be, however, that some form of this approach could be combined with our current search procedure (as detailed in Figure 3), and we hope to investigate this in the future.

Machine-learning methods for virtual screening require the availability of a *training set*, that is, sets of both known active and known inactive molecules that can be used to develop a tool that can be applied to molecules of unknown activity (the *test set*) and predict their (in)activities with a fair degree of confidence. The best-established approach of this sort is substructural analysis,[39,40] but the past few years have also seen increasing interest in approaches based on kernel discrimination and support vector machines.[37,41,42] If we wish to use such methods when the only activity information available is that represented by a single reference structure, then means must be found to identify inactives and further actives that can be used as training data. Given the success of our TSS experiments, it seems possible that the NNs of the known reference structure could comprise the actives in the training set, with inactives being obtained by noting that the characteristics of inactives are approximated with a high degree of accuracy by the characteristics of the entire database that is to be searched. Hence, the training-set inactives are obtained by selecting molecules at random from the database that is to be searched (subject, in our experiments, to none of these presumed inactives having a similarity coefficient greater than 0.40 using ECFP_4 fingerprints and the Tanimoto coefficient with the reference structure and its NNs that comprise the training-set actives). Our new TSS procedure is hence as shown in Figure 4.

We have tested two machine-learning procedures: SSA and BKD. Full details of these have been presented elsewhere,[37,39−41] and we hence provide only a summary description here. SSA assigns a weight to each bit (or substructure) in a fingerprint that describes that bit's differential occurrence in the active and inactive molecules constituting a training set for which biological testing has already taken place. The resulting weights are then used to rank the test database, with molecules at the top of this ranking being candidates for testing. Many different substructural analysis weights have been described in the literature,[40] but they are all based on some or all of the following pieces of information about the training set: A_j and I_j are the numbers of active and inactive compounds with bit j set; T_j is the total number of compounds with bit j set (so $T_j = A_j + I_j$); and N_A and N_I are the total number of active and inactive molecules, respectively, with N_T being the total number of molecules (so $N_T = N_A + N_I$). Our experiments have used the R1, R2, R3, R4, and AVID weights that are discussed in the Appendix to this paper. Of these, the best results were obtained using the R4 weight,[40] and we shall focus on this here. The weight has the form

$$\mathrm{R4}(j) = \log\left[\frac{A_j/(N_A - A_j)}{I_j/(N_I - I_j)}\right]$$

with the frequencies in the weight for each fragment j being derived as described above (i.e., from a training set contain-

Input the reference structure *R*
Compute the similarity of *R* with every molecule in the database *D*
Sort *D* in decreasing order of the calculated similarity values to give a sorted database *SD*(0)
Identify the *k* NNs of *R* from the top of the list *SD*(0)
For each such nearest-neighbour, *NN*(*i*)
    Compute the similarity of *NN*(*i*) with every molecule in *D*
    Sort *D* in decreasing order of the calculated similarity values to give a sorted database *SD*(*i*)
Fuse the sorted lists *SD*(0)-*SD*(*k*) to give the final output from the turbo similarity search

Figure 3. Turbo similarity searching using group fusion.

Input the reference structure *R*
Compute the similarity of *R* with every molecule in the database *D*
Sort *D* in decreasing order of the calculated similarity values
Assume that the *k* NNs at the top of the ranking are active
Select *k*+1 molecules at random from *D*, subject to the constraint that none of them has similarity ≥ 0.40 (Tanimoto coeficient and Scitegic ECFP_4 fingerprints) with the reference structure or with any of the top-*k* NNs
Assume that these *k*+1 molecules are inactive
Use *R*, the *k* NNs and the *k*+1 randomly selected molecules as the training-set for a machine learning procedure

Figure 4. Generation of a training set for machine learning using turbo similarity searching.

Table 3. Mean Recalls at 5% for Similarity Searching (SS), Conventional Turbo Similarity Searching using Group Fusion (TSS), and Turbo Similarity Searching Using Substructural Analysis (TSS−SSA) and Using Binary Kernel Discrimination (TSS−BKD) of the 11 Activity Classes Used in Refs 15 and 16

activity class	SS	TSS-GF	TSS−SSA	TSS−BKD
5HT3 antagonists	31.7	44.0	39.9	36.3
5HT1A agonists	26.3	36.2	34.3	30.9
5HT reuptake inhibitors	21.6	24.1	25.0	25.0
D2 antagonists	25.1	30.3	30.4	30.6
renin inhibitors	90.4	94.7	94.4	94.4
angiotensin II AT1 antagonists	77.4	92.0	90.3	92.1
thrombin inhibitors	44.5	50.7	50.8	52.3
substance P antagonists	28.6	34.1	31.2	33.0
HIV Protease inhibitors	51.6	55.2	55.9	58.0
cyclooxygenase inhibitors	13.7	14.4	16.6	14.3
protein kinase C inhibitors	21.0	20.6	22.5	23.1
average over all classes	39.2	45.1	44.7	44.5

Table 4. Mean Recalls at 5% for Similarity Searching (SS), Conventional Turbo Similarity Searching Using Group Fusion (TSS−GF), and Turbo Similarity Searching Using Substructural Analysis (TSS−SSA) and Using Binary Kernel Discrimination (TSS−BKD) of the 10 Diverse Activity Classes Detailed in Table 1a

activity class	SS	TSS-GF	TSS−SSA	TSS−BKD
muscarinic (M1) agonists	27.4	31.0	46.6	42.2
NMDA receptor antagonists	15.7	17.1	20.7	18.7
nitric oxide synthase inhibitors	18.1	16.9	21.0	18.7
dopamine β-hydroxylase inhibitors	37.5	37.1	51.7	42.9
aldose reductase inhibitors	19.9	22.8	23.8	22.7
reverse transcriptase inhibitors	15.5	14.6	18.0	16.8
aromatase inhibitors	29.0	30.3	33.5	32.5
cyclooxygenase inhibitors	13.7	14.4	16.7	14.4
phospholipase A2 inhibitors	19.9	20.8	21.2	21.2
lipoxygenase inhibitors	13.0	15.3	15.2	13.5
average over all classes	21.0	22.0	26.8	24.4

ing the original reference structure and its NNs and molecules selected at random from the search file).

Searches were carried out using ECFP_4 fingerprints on the initial set of 11 activity classes, and the results are presented in Table 3. Here, SS represents conventional similarity searching and TSS−GF, TSS−SSA, and TSS−BKD represent turbo similarity searching based on the use of group fusion, substructural analysis, and binary kernel discrimination, respectively, for the reranking of the original similarity-search output. Taking the average over all of the actives in all of the activity classes, it will be seen from Table 3 that TSS−SSA and TSS−BKD outperform conventional similarity searching but that they are both marginally inferior to TSS−GF as originally described in ref 18. However, there are some cases where they are superior to TSS−GF, especially for the more diverse (low MPS) activity classes. Comparable searches were hence carried out using the 10 highly diverse activity classes listed in Table 1a, and the results here are shown in Table 4.

Inspection of the figures in this table shows very clearly that the machine-learning versions of TSS are to be preferred to the original TSS−GF when diverse sets of active structures

are to be searched for in a chemical database. The increase in performance here is really quite marked: the mean recalls at 5% for SS and TSS−SSA are 21.0 and 26.8, respectively, which represents an increase of over one-quarter in the number of active molecules identified in a conventional similarity search. It may appear surprising at first sight that TSS−SSA is superior to TSS−BKD, given that our previous experience of SSA and BKD had suggested that the latter was to be preferred.[15,41] However, we have recently conducted studies on the robustness of BKD and SSA when false positives and false negatives are present in a training set. This work will be presented elsewhere;[43] in brief, it has shown that the performance of BKD is more affected by the presence of noisy activity data than is SSA. In the present context, the assumption that the *k* NNs are all active inevitably means that there are many false positives in the training set (however, the assumption that database molecules picked at random are inactive may result in only a few false negatives). There are obvious differences with the work reported in ref 43, most obviously in the fact that the false positives here are all structurally related to the reference structure, but it does seem likely that the results in Table 3 can be explained by the quality of the activity data that is available to the two machine-learning procedures.

CONCLUSIONS

Similarity searching based on 2D fingerprints and the Tanimoto coefficient has become a well-established technique for virtual screening since it was first described some two decades ago. In this paper, we have described two ways of enhancing the retrieval effectiveness of this approach: one that can be used when just a single, bioactive reference structure is available and one that can be used when several reference structures are available. These enhancements have involved the use of both data-fusion and machine-learning techniques in combination with similarity searching.

The experiments reported here and previously suggest the following guidelines for similarity-based virtual screening of a 2D database. If a single reference structure is available, then search performance will be increased if the nearest neighbors resulting from the initial ranking of a database are assumed to be active, and this assumption used to rerank the database using group fusion or an approximate form of substructural analysis. The increase in effectiveness would appear to be maximized (when averaged over a range of types of search) by using group fusion to generate the second ranking (TSS−GF); when the sought actives are expected to be structurally diverse, that is, for scaffold-hopping applications, then the best results are obtained using the substructural-analysis approach (TSS−SSA). If multiple reference structures are available, then search performance will be considerably increased (when averaged over a range of types of search) by the use of group fusion or an approximate form of binary kernel discrimination. The increases are particularly marked when the sought actives are structurally diverse; if this is not the case, that is, if the actives are tightly clustered, then any improvement over conventional similarity searching is likely to be marginal.

The increases in search performance that we have described can be achieved with minimal increases in computational requirements: we hence believe that the techniques described here provide attractive ways of enhancing the cost-effectiveness of ligand-based virtual screening in chemical databases.

ACKNOWLEDGMENT

We thank the following: Novartis Institutes for BioMedical Research for funding J.H.; MDL Information Systems Inc. for the provision of the MDDR database; and Barnard Chemical Information Ltd., Daylight Chemical Information Systems Inc., the Royal Society, Scitegic Inc., Tripos Inc., and the Wolfson Foundation for software and laboratory support.

APPENDIX. RANKING DATASETS USING SUBSTRUCTURAL ANALYSIS AND NAÏVE BAYESIAN CLASSIFIERS

Recent studies have suggested the utility of naïve Bayesian classifiers for ligand-based virtual screening.[44,45] In fact, these classifiers are closely related to some of the fragment weighting schemes that have been suggested previously for substructural analysis.[39]

Studies by Ormerod et al.[40] and Cosgrove and Willett[46] have demonstrated the effectiveness of fragment weights that derive from work by Robertson and Sparck Jones on the weighting of keywords in text retrieval systems.[47] These authors carried out a detailed Bayesian analysis of the keyword weighting problem and reported four weighting functions, R1, R2, R3, and R4. It is straightforward to draw an analogy between a collection of documents, some of which are relevant and represented by keywords, and a library of compounds, some of which are active and represented by fragment substructures, thus enabling the keyword weights to be applied to the virtual screening problem.[40] The analysis of Robertson and Sparck Jones involves two kinds of assumption in the derivation of their weighting schemes: an *independence assumption* and an *ordering principle*. When put in the chemical context, one can assume that the distribution of fragments in active compounds is independent and their distribution in all compounds is independent (independence assumption I1) or that the distribution of fragments in active compounds is independent and their distribution in inactive compounds is independent (independence assumption I2). In like vein, one can assume that the probability of activity is based only on the presence of fragments in compounds (ordering principle O1) or that the probability of activity is based on both the presence of fragments in compounds and their absence from compounds (ordering principle O2).

The fragment weight R1 is based on I1 and O1, R2 on I2 and O1, R3 on I2 and O1, and R4 on I2 and O2. These weights are listed below in eqs 1−4. In these equations: T_j is the total number of compounds in the training set which contain fragment (bit) j, being made up of A_j and I_j, the number of active and inactive compounds, respectively, in the training set which contain fragment j; N_T is the total number of compounds in the training set, this being made up of N_A and N_I, the number of active and inactive compounds in the training set, respectively.

$$R1(j) = \log\left(\frac{A_j/N_A}{T_j/N_T}\right) \quad (1)$$

$$\mathrm{R2}(j) = \log\left(\frac{A_j/N_\mathrm{A}}{I_j/N_\mathrm{I}}\right) \quad (2)$$

$$\mathrm{R3}(j) = \log\left[\frac{A_j/(N_\mathrm{A} - A_j)}{T_j/(N_\mathrm{T} - T_j)}\right] \quad (3)$$

$$\mathrm{R4}(j) = \log\left[\frac{A_j/(N_\mathrm{A} - A_j)}{I_j/(N_\mathrm{I} - I_j)}\right] \quad (4)$$

Xia et al.[45] described the naïve Bayesian classifier implemented in the Pipeline Pilot software.[36] Consider a training set containing N_T compounds of which N_A are active as defined above; then, the baseline probability of a random compound being active is given by

$$P(A) = \frac{N_\mathrm{A}}{N_\mathrm{T}} \quad (5)$$

and an initial estimate of the probability of activity if a fragment j is present is given by

$$P(A|j) = \frac{A_j}{T_j} \quad (6)$$

This can be made into a relative estimate by dividing $P(A|j)$ by $P(A)$, so as to obtain

$$\frac{P(A|j)}{P(A)} = \frac{A_j/N_\mathrm{A}}{T_j/N_\mathrm{T}} \quad (7)$$

which is clearly very closely related to R1(j) above. In fact, eq 6 provides a poor estimate for $P(A|j)$, especially when T_j is small. A better estimate is obtained by using the Laplacian correction,[44] yielding after simplifications

$$P_\mathrm{final}(A|j) = A_j + 1/\left(T_j\frac{N_\mathrm{A}}{N_\mathrm{T}} + 1\right) \quad (8)$$

This weight is exactly the same as that for a substructural analysis weighting scheme described by Avidon et al.[48] However, in the Pipeline Pilot scheme, the logs of the fragment weights are summed to give the final probability of activity for a molecule; in the Avidon scheme, the fragment weights themselves are summed to give the final score.

In the naïve Bayesian classifier of Bender et al.,[44] the probability of activity given some molecular description F, $P(A|F)$, can be calculated from the product of the probabilities of occurrence of the individual fragments that are present in active molecules, $P(f_j|A)$, and similarly for the probability of inactivity, $P(I|F)$. The score for a test-set molecule is then calculated as the ratio of these two probabilities, that is,

$$\frac{P(A|F)}{P(I|F)} = \frac{P(A)\prod_j P(f_j|A)}{P(I)\prod_j P(f_j|I)} \quad (9)$$

$$\frac{P(A|F)}{P(I|F)} = \frac{P(A)}{P(I)}\prod_j \frac{P(f_j|A)}{P(f_j|I)} \quad (10)$$

Substituting for the various probabilities, we obtain

$$\frac{P(A|F)}{P(I|F)} = \frac{N_\mathrm{A}/N_\mathrm{T}}{N_\mathrm{I}/N_\mathrm{T}}\prod_j \frac{A_j/N_\mathrm{A}}{I_j/N_\mathrm{I}} \quad (11)$$

which is the MOLPRINT 2D weight.[44] Manipulation of eq 11 yields eq 12.

$$\log\left[\frac{P(A|F)}{P(I|F)}\right] = \log\left(\frac{N_\mathrm{A}/N_\mathrm{T}}{N_\mathrm{I}/N_\mathrm{T}}\right) + \sum_j \log\left(\frac{A_j/N_\mathrm{A}}{I_j/N_\mathrm{I}}\right) \quad (12)$$

The term

$$\log\left(\frac{N_\mathrm{A}/N_\mathrm{T}}{N_\mathrm{I}/N_\mathrm{T}}\right)$$

is a constant that will not affect the ranking of the test-set molecules, and we then obtain the weight

$$\log\left(\frac{A_j/N_\mathrm{A}}{I_j/N_\mathrm{I}}\right)$$

which is identical with eq 2, that is, the weight R2(j). It should be noted that Bender et al. additionally use a feature selection procedure as part of their virtual-screening method; here, we have focused only on the form of the weight that is associated with each of the substructural descriptors.

There are thus close relationships between some of the scoring schemes that have been suggested for ligand-based virtual screening.

REFERENCES AND NOTES

(1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(2) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(3) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nearest Neighbour Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.
(4) Johnson, M. A., Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley: New York, 1990.
(5) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(6) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighbourhood Behaviour: a Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(7) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules have Similar Biological Activities? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(8) Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1407−1414.
(9) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.
(10) He, L.; Jurs, P. C. Assessing the Reliability of a QSAR Model's Predictions. *J. Mol. Graphics Modell.* **2005**, *23*, 503−523.
(11) Kubinyi, H. Similarity and Dissimilarity: a Medicinal Chemist's View. *Perspect. Drug Discovery Des.* **1998**, *9*−*11*, 225−232.

(12) Sheridan, R. P.; Kearsley, S. K. Why Do We Need So Many Chemical Similarity Search Methods? *Drug Discovery Today* **2002**, *7*, 903−911.
(13) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity−a Review. *Quant. Struct.-Act. Relat. Comb. Sci*. **2003**, *22*, 1006−1026.
(14) Bender, A.; Glen, R. C. Molecular Similarity: a Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.
(15) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.
(16) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.
(17) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Losel, J. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest-Neighbour Lists: A Comparison of Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840−1848.
(18) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbour Information. *J. Med. Chem.* **2005**, *48*, 7049−7054.
(19) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discovery Today: Technol.* **2004**, *1*, 217−224.
(20) Patani, G. A.; LaVoie, E. J. Bioisosterism: a Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96*, 3147−3176.
(21) Schneider, G.; Neidhart, W.; Giller, T.; Schmidt, G. Scaffold-Hopping by Topological Pharmacophore Search. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.
(22) Stanton, D.; Morris, T.; Roychoudhury, S.; Parker, C. Application of Nearest Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.
(23) Bohl, M.; Dunbar, J. B.; Gifford, E. M.; Heritage, T.; Wild, D. J.; Willett, P.; Wilton, D. J. Scaffold Searching: Automated Identification of Similar Ring Systems for the Design of Combinatorial Libraries. *Quant. Struct.-Act. Relat.* **2002**, *21*, 590−597.
(24) Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; McLay, I. M.; Bradshaw, J. Drug Rings Database with Web Interface. A Tool for Identifying Alternative Rings in Lead Discovery Programmes. *J. Med. Chem.* **2003**, *46*, 3257−3274.
(25) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004**, *47*, 6144−6159.
(26) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. Lead Hopping. Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities. *J. Med. Chem.* **2004**, *47*, 6777−6791.
(27) The MDL Drug Data Report database is available from MDL Information Systems Inc. at http://www.mdli.com.
(28) Hert, J. Ph.D. Thesis, University of Sheffield, Sheffield, U. K., 2005.
(29) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures using Data Fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.
(30) Charifsen, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: a Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(31) The Dictionary of Natural Products database is available from Chapman & Hall/CRC Press at http://www.chemnetbase.com.
(32) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a Method for Evaluating Drug Likeness and Ease of Synthesis Using a Data Set in which Compounds are Assigned Scores Based on Chemists' Intuition. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1269−1275.
(33) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *J. Med. Chem*. **2004**, *47*, 4891−4896.
(34) The Unity software is available from Tripos Inc. at http://www.tripos.com
(35) Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J. J.; Lecchini, S.; Jacoby, E. An Ontology for Pharmaceutical Ligands and its Application for in silico Screening and Library Design. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 947−955.
(36) The Pipeline Pilot software is available from Scitegic Inc. at http://www.scitegic.com.
(37) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.
(38) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: a Simple and Efficient Method for Improving High-Throughput Docking Results. *J. Med. Chem*. **2004**, *47*, 2743−2749.
(39) Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural Analysis. A Novel Approach to the Problem of Drug Design. *J. Med. Chem.* **1974**, *17*, 533−535.
(40) Ormerod, A.; Willett, P.; Bawden, D. Comparison of Fragment Weighting Schemes for Substructural Analysis. *Quant. Struct.-Act. Relat.* **1989**, *8*, 115−129.
(41) Wilton, D. J.; Willett, P.; Lawson, K.; Mullier, G. Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 469−474.
(42) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Comput. Sci*. **2005**, *45*, 549−561.
(43) Chen, B.; Harrison, R. F.; Pasupa, K.; Wilton, D. J.; Willett, P.; Wood, D. J. Virtual Screening Using Binary Kernel Discrimination: Effect of Noisy Training Data and the Optimization of Performance. Submitted for publication.
(44) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 170−178.
(45) Xia, X. Y.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463−4470.
(46) Cosgrove, D. A.; Willett, P. SLASH: A Program for Analysing the Functional Groups in Molecules. *J. Mol. Graphics Modell.* **1998**, *16*, 19−32.
(47) Robertson, S. E.; Sparck Jones, K. Relevance Weighting of Search Terms. *J. Am. Soc. Inf. Sci.* **1976**, *27*, 129−146.
(48) Avidon, V. V.; Arolovich, V. S.; Kozlova, S. P.; Piruzyan, L. A. Statistical Study of Information File on Biologically Active Compounds. II. Choice of Decision Rule for Biological Activity Prediction. *Khim.-Farm. Zh.* **1978**, *12*, 88−93.

CI050348J