JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Calculation of Substructural Analysis Weights Using a Genetic Algorithm

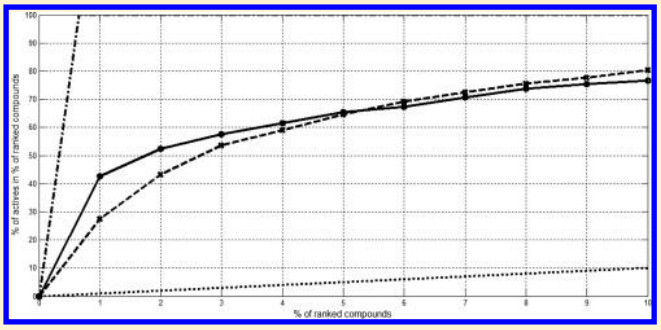
John D. Holliday,[†] Nor Sani,[†,‡] and Peter Willett*[,†]

[†]Information School, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, United Kingdom

[‡]Faculty of Information Science and Technology, National University of Malaysia, 43600 UKM Bangi, Malaysia

Supporting Information

ABSTRACT: This work describes a genetic algorithm for the calculation of substructural analysis for use in ligand-based virtual screening. The algorithm is simple in concept and effective in operation, with simulated virtual screening experiments using the MDDR and WOMBAT data sets showing it to be superior to substructural analysis weights based on a naive Bayesian classifier.



INTRODUCTION

Machine learning methods are widely used for ligand-based virtual screening. Such methods take as input a training set of molecules, each of which is known to be either active or inactive, and produce as output a set of rules that can then be used to predict whether previously unseen molecules in the test set should be classified as active or as inactive. Many different computational techniques have been used for this purpose, such as binary kernel discrimination, neural networks, random forests, recursive partitioning, and support vector machines *inter alia*, as reviewed by, e.g., Goldman,[1] Melville et al.,[2] Mitchell,[3] and Plewczynski et al.[4]

In this work we focus on substructural analysis, which was developed by Cramer and his co-workers in the 1970s[5,6] and which was thus one of the first machine learning approaches to be applied to the analysis of chemical data sets. In substructural analysis, it is assumed that each molecule in the training set and test set is characterized by a set of binary descriptors, most commonly in the form of a two-dimensional (2D) fingerprint in which each bit denotes the presence or absence of a substructural feature (often referred to as a fragment). Associated with each such bit is a weight that is a function of the numbers of active and inactive training-set molecules that have that bit switched on, i.e., that contain the corresponding fragment. This weight reflects the probability that a molecule containing that substructural feature will be active (or inactive); for example, the weight might be the fraction of the active training-set molecules containing that particular fragment. A test-set molecule is then scored by summing (or otherwise combining) the weights of those bits that are set in its fingerprint, the resulting score representing the overall probability that the molecule will be active. Substructural analysis was studied in considerable detail by Hodes in a National Institutes of Health project to develop novel anticancer agents,[7−10] but it is only quite recently that the approach has become widely used.[11−15] An operational example of the use of substructural analysis is the PASS (for prediction of activity spectra for substances) system developed by the Poroikov group.[11,16,17] Some of the weighting schemes that have been used in substructural analysis are closely related to those obtained using a naive Bayesian classifier[18] (hereafter NBC), a well-established approach to machine learning that has become popular in chemoinformatics with the availability of the Bayesian modeling routine in the Pipeline Pilot software system.[19−21]

The cited references provide several examples of the successful use of substructural analysis in drug discovery projects. In this brief communication, we describe an approach to the calculation of fragment weights using a genetic algorithm (hereafter GA).[22] Given a training set of molecules and associated bioactivity data, the GA (which is described in detail in the next section) computes fragment weights that can then be used as an alternative to those resulting from existing substructural analysis approaches. A GA provides a (non-deterministic) way of exploring combinatorial spaces, such as the set of all possible fragment weights in a weighting scheme in the present context, and may hence provide a way of identifying sets of weights that are different from, and possibly superior to, those identified by existing approaches to substructural analysis.

GENETIC ALGORITHM

Each chromosome in the GA is a vector, each element of which contains a real-valued number representing the weight of one of

Received: September 8, 2014
Published: January 23, 2015

DOI: 10.1021/ci500540s
J. Chem. Inf. Model. 2015, 55, 214−221

the fragments used to characterize the molecules in the data set; thus, using the MDL key set (see Experimental Details and Results), each chromosome has 166 elements. The GA seeks to identify those weights that produce the best possible ranking of the molecules in a data set and hence to estimate an upper bound to the effectiveness of virtual screening possible using the substructural analysis approach. The basic idea is illustrated in Figure 1 using a training set containing three molecules, M_{1-3}, each of which is represented by a fingerprint encoding the presence or absence of five fragments, F_{1-5}.

Molecule	F_1	F_2	F_3	F_4	F_5
M_1	0	1	0	0	1
M_2	1	0	0	1	0
M_3	0	0	0	1	1

(a) Fingerprints for three molecules M_{1-3} encoding five different substructural fragments F_{1-5}

Chromosome	W_1	W_2	W_3	W_4	W_5
C_1	6	2	7	0	1
C_2	4	3	1	8	5
C_3	9	9	3	6	7
C_4	1	7	5	1	3
C_5	8	4	8	2	8
C_6	5	8	4	7	2

(b) Six chromosomes C_{1-6} encoding the weights W_{1-5} for F_{1-5}

Chromosome	M_1	M_2	M_3
C_1	3	6	1
C_2	8	12	13
C_3	16	15	13
C_4	10	2	4
C_5	12	10	10
C_6	10	12	9

(c) Sums-of-weights using each chromosome C_{1-6} for each molecule M_{1-3}

Figure 1. Operation of the GA using a population containing six chromosomes and a training set containing three molecules, each of which is described by the presence or absence of five fragments.

Assume that the three fingerprints are as shown in Figure 1a where, for example, M_1 contains the second and fifth fragments, F_2 and F_5. An initial population of possible solutions is generated with the initial weights W_1−W_5 being assigned by a random-number generator that has been primed in this simple example to generate integer weights in the range 0−10. In the example, the population contains six chromosomes, C_{1-6}, and the initial population is shown in Figure 1b. Each chromosome is then used to compute the sum-of-weights for each molecule, as shown in Figure 1c. For example, M_1 contains F_2 and F_5, so its sum-of-weights using C_1 is the sum of W_2 and W_5, i.e., 3, using C_2 the sum is 8, and so on for C_{3-6}. Considering just C_1, the sums-of-weights for M_1, M_2, and M_3 are 3, 6, and 1, respectively, meaning that the application of the weights in this chromosome results in a ranking

$$M_2 > M_1 > M_3$$

of the training set. In like vein, C_2 yields the ranking

$$M_3 > M_2 > M_1$$

and so on for the remaining chromosomes C_3−C_6.

Each chromosome thus corresponds to a particular ranking of the training set, and this ranking can be used to compute a fitness value for that chromosome. The aim of a virtual screening procedure is to cluster active molecules at the top of a database ranking, and hence the fitness function used in our GA is simply the number of training-set actives occurring above some threshold rank position; specifically, the fitness in the experiments reported in later text was the number of actives in the top 1% of the training set. The resulting fitness values for each chromosome then provide the input to the next iteration of the GA, with the standard operations of crossover and mutation being applied to obtain a new population of chromosomes. The procedure is repeated until the fitness values have plateaued or (as was the case in the experiments reported in later text) for a fixed number of iterations.

EXPERIMENTAL DETAILS AND RESULTS

Data Sets. The GA has been evaluated in simulated virtual screening experiments using data sets derived from the MDL Drug Data Report (MDDR) and World of Biological Activity (WOMBAT) databases. The data sets used here are described in detail by Gardiner et al.:[23] the MDDR data set contained 11 activity classes and 102,514 molecules, and the WOMBAT data set contained 14 activity classes and 138,127 molecules. The molecules in these two data sets were characterized by the MDL structural key definitions in the Pipeline Pilot software, i.e., 166-bit fingerprints where each bit describes the presence or absence in a molecule of a particular fragment substructure. These keys were used for all of the experiments reported here.

Table 1. Screening Results Using the GA Described Here and the R4 NBC for the (a) MDDR and (b) WOMBAT Data Sets

		actives retrieved		murcko scaffolds retrieved	
activity class	actives	NBC	GA	NBC	GA
	(a)				
5HT reuptake inhibitors	323	34	56	17	20
5HT1A agonists	744	103	140	52	71
5HT3 antagonists	677	138	276	75	122
angiotensin II AT1 antagonists	849	372	404	131	134
cyclooxygenase inhibitors	572	146	161	37	40
D2 antagonists	356	47	60	24	33
HIV protease inhibitors	675	226	326	106	151
protein kinase C inhibitors	408	94	123	34	38
renin inhibitors	1017	620	701	192	205
substance P antagonists	1121	262	325	120	133
thrombin inhibitors	723	226	344	99	147
	(b)				
5HT1A agonists	533	249	285	58	63
5HT3 antagonists	198	79	82	23	25
acetylcholine esterase inhibitors	453	218	230	67	69
angiotensin II AT1 antagonists	652	514	521	105	115
cyclooxygenase inhibitors	869	549	583	32	67
D2 antagonists	819	225	331	61	67
factor Xa inhibitors	758	288	337	75	86
HIV protease inhibitors	1015	398	503	127	143
matrix metalloprotease inhibitors	625	362	391	86	94
phosphodiesterase inhibitors	536	239	259	84	88
protein kinase C inhibitors	128	92	94	15	15
renin inhibitors	427	301	331	73	86
substance P antagonists	502	217	243	52	55
thrombin inhibitors	379	200	212	73	77

DOI: 10.1021/ci500540s
J. Chem. Inf. Model. 2015, 55, 214−221

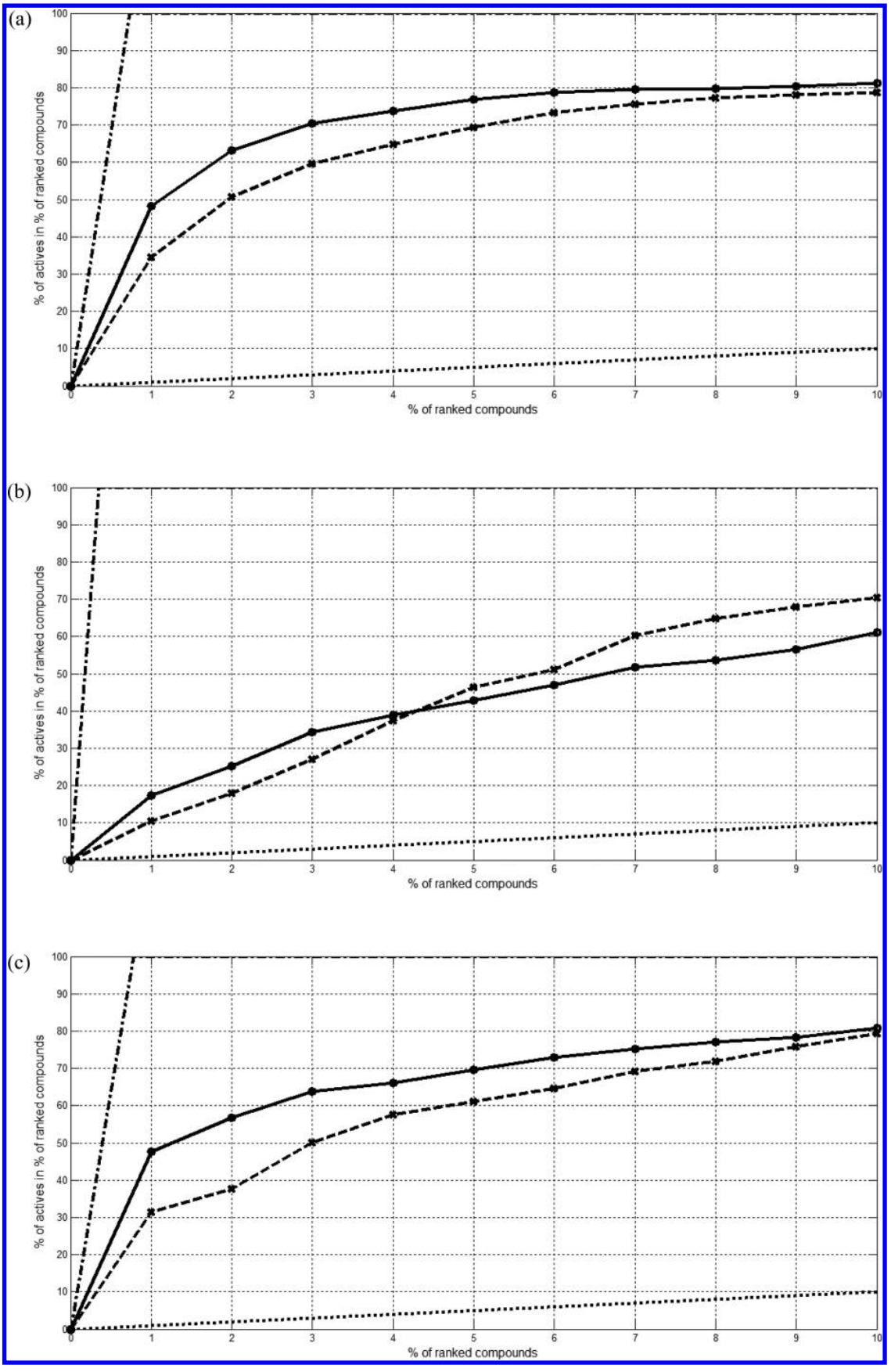


Figure 2. continued

DOI: 10.1021/ci500540s
J. Chem. Inf. Model. 2015, 55, 214−221

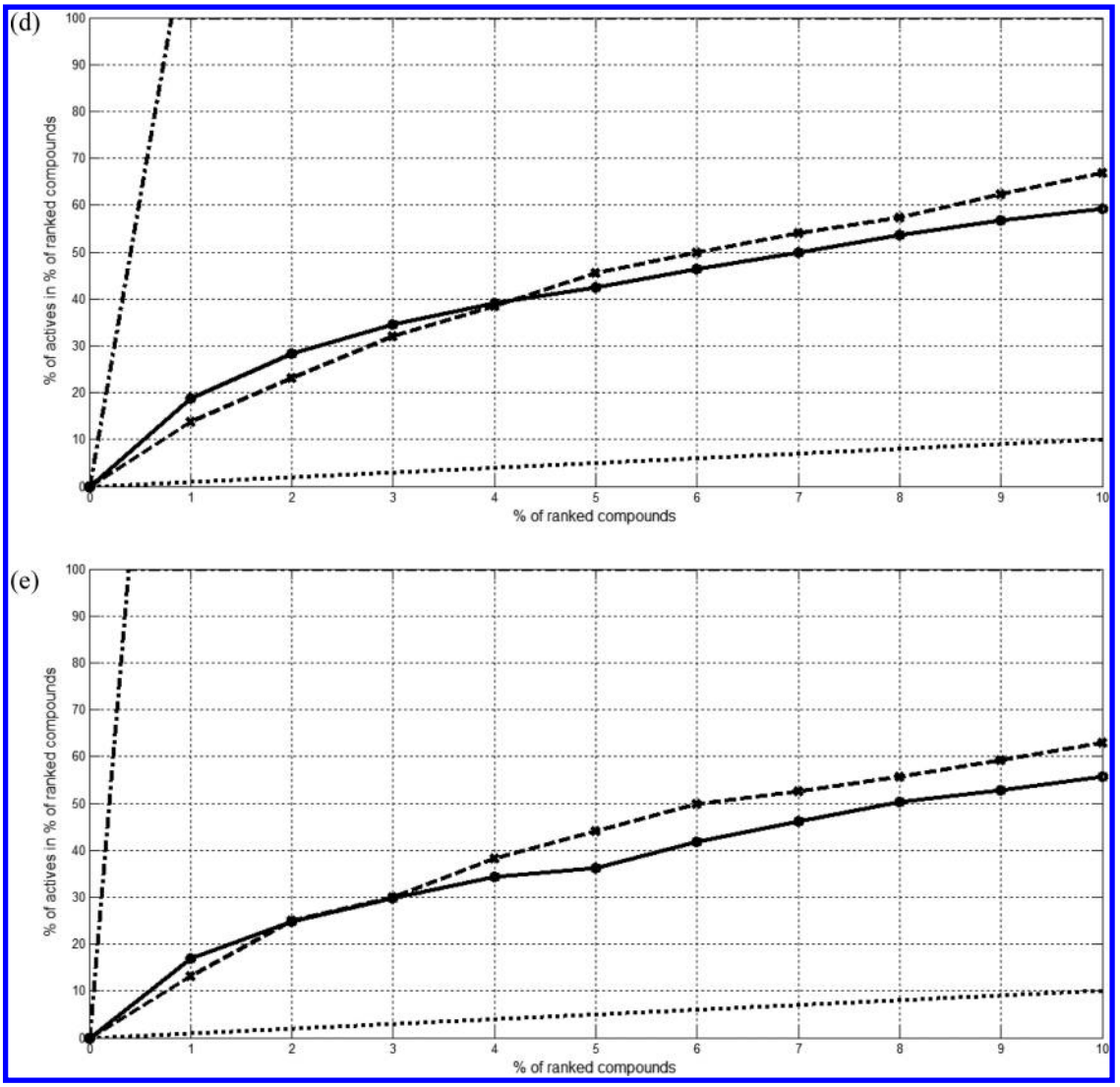


Figure 2. Enrichment curves for simulated virtual screening of MDDR activity classes: (a) HIV protease inhibitors; (b) 5HT reuptake inhibitors; (c) thrombin inhibitors; (d) 5HT1A agonists; (e) D2 antagonists. The continuous line in each case is for the GA and the dashed line is for the NBC.

The training set for a particular activity class consisted of 10% of the actives and 10% of the inactives, with the remaining 90% of the database providing the test set for which virtual screening was carried out. The GA was run on the training set, weights determined for each of the 166 fragments comprising a fingerprint, and then these weights were used to rank the molecules in the test set.

Machine learning experiments are often evaluated using area under the curve (AUC) values, i.e., the area under a receiver operating characteristic (or ROC) curve. However, this performance criterion is less appropriate for evaluating virtual screening experiments since it takes the entire ranking of a database into account when calculating the effectiveness of a ranking, where as one is interested here in only that small fraction of the molecules that occur at the top of the ranking, since it is these that may need to be considered subsequently for biological screening.[24,25] Rather than using AUC values, the screening performance was hence measured by the number of actives for the top 1% of the ranked test set and the number of distinct Murko scaffolds[26] in those top-ranked actives was noted to provide a simple measure of structural diversity.

The performance of the GA was compared with that of an existing substructural analysis method, early approaches for which used empirical weighting approaches such as the SAS scheme described by Redl et al.[6] Drawing on previous work by Robertson and Spärck Jones that presented a detailed theoretical rationale for the use of NBCs in information retrieval (where the aim is to rank a text database in order of decreasing probability of relevance to a query),[27] four NBCs have been described that could be used for ligand-based virtual screening.[28,29] One of these, R4, was used here to provide a basis for comparison with the results obtained using the GA. Assume that a particular fragment occurs in a of the A active molecules and i of the I inactive molecules; then R4 is defined to be

$$\mathrm{R4} = \log\left(\frac{a/(A-a)}{i/(I-i)}\right) \quad (1)$$

The numerator of this expression is hence the ratio of the number of active molecules in which the fragment occurs to the number of active molecules in which it does not occur; and the denominator is the corresponding ratio for the fragment's occurrence or nonoccurrence in inactive molecules. Experi-

DOI: 10.1021/ci500540s
J. Chem. Inf. Model. 2015, 55, 214−221

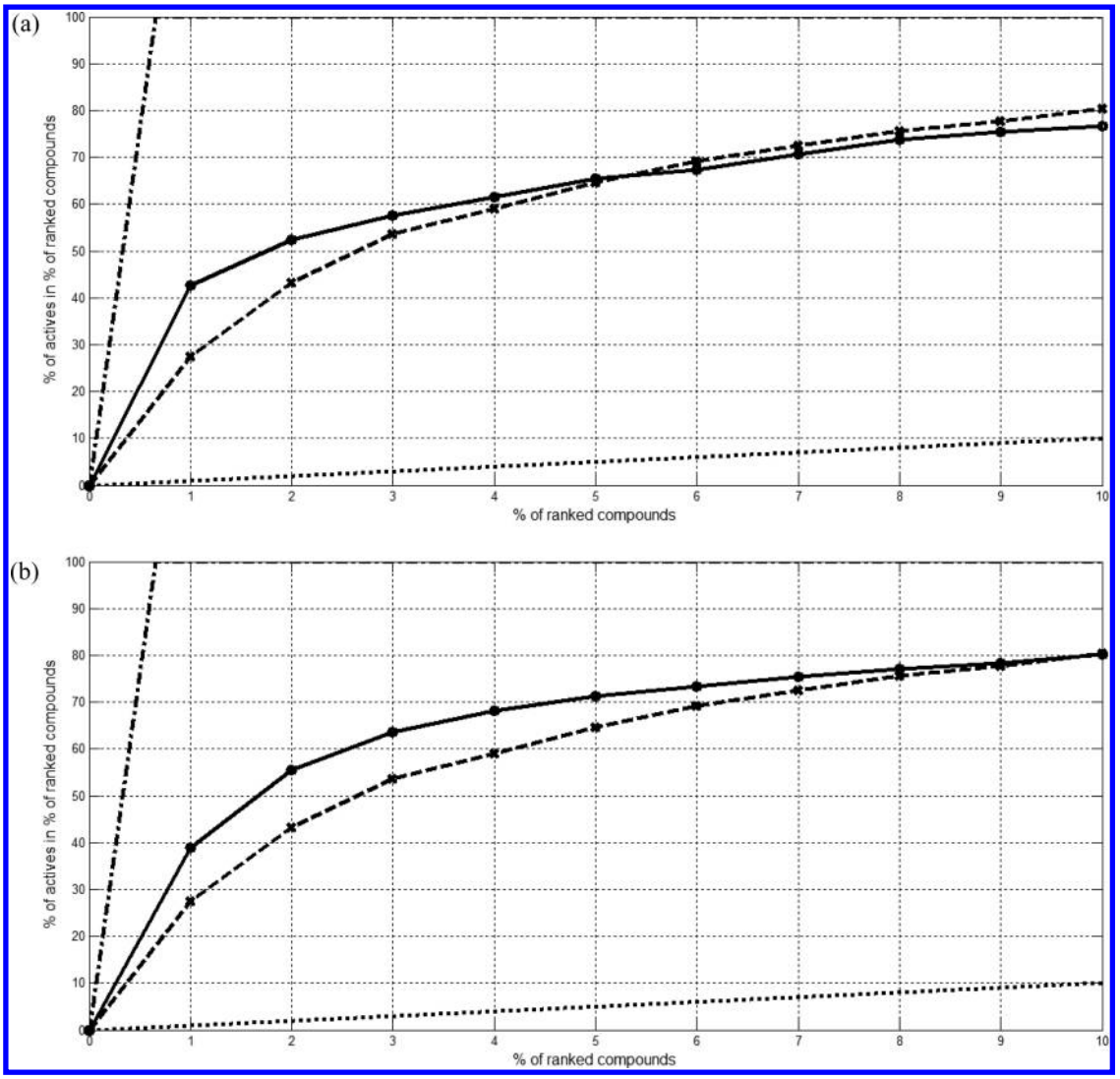


Figure 3. Enrichment curves for simulated virtual screening of the WOMBAT D2 antagonists with the fitness function set to maximize the number of actives in (a) the top 1% of the ranking and (b) the top 10% of the ranking. The continuous line in both cases is for the GA and the dashed line is for the NBC.

ments were also conducted using the NBC in the Pipeline Pilot software;[19−21] however the results obtained were comparable to those with R4, and hence only the latter sets of results are discussed here.

Parametrization of the GA. The GA has been described in general terms in the previous section. An implementation of it requires the specification of the following parameters, where the alternatives that were tested are listed in brackets: parent selection procedure (roulette wheel, tournament, or random); crossover procedure (one-point, two-point, or uniform); crossover rate (0.60−0.95 in steps of 0.05); mutation rate (0.005−0.100); population size (100−500 in steps of 100); number of iterations of the GA (100−1100 in steps of 200). A systematic, detailed series of tests was conducted using the renin and cyclooxygenase activity classes from the MDDR data set. These were chosen since they were the least diverse and the most diverse class respectively of all the 25 sets of actives studied in the experiments (where the diversity was measured by the mean intermolecular similarities for the active molecules in a class when computed using Tripos Unity fingerprints and the Tanimoto coefficient). On the basis of these initial experiments, all of the results presented here were obtained with a GA involving roulette-wheel selection, one-point crossover, a crossover rate of 0.95, a mutation rate of 0.01, a population of 200 chromosomes, and 500 iterations. It was found that more consistent results were obtained when the mutation operator was constrained: if the R4 weight for a particular fragment was calculated to be positive, then only mutations that resulted in a positive weight for that fragment were accepted, and similarly if the R4 weight was negative. The GA was implemented using MATLAB, pseudo-code for which is included in the Supporting Information.

Experimental Results. The basic results are shown in Table 1a,b for the MDDR and WOMBAT data sets, respectively. Each row of one of these sections of the table corresponds to a single activity class and lists the total numbers of actives in the test set, the numbers of active molecules retrieved in the top 1% by the NBC and GA weighting schemes, and then the numbers of distinct Murcko scaffolds in these top-ranked actives. The GA was run three times; the values listed in the two sections of the table are those obtained from the worst run, *viz.*, the run that resulted in the smallest number of top-ranked actives. It will be seen that the GA is consistently, and often markedly, superior to the NBC in terms

DOI: 10.1021/ci500540s
J. Chem. Inf. Model. 2015, 55, 214−221

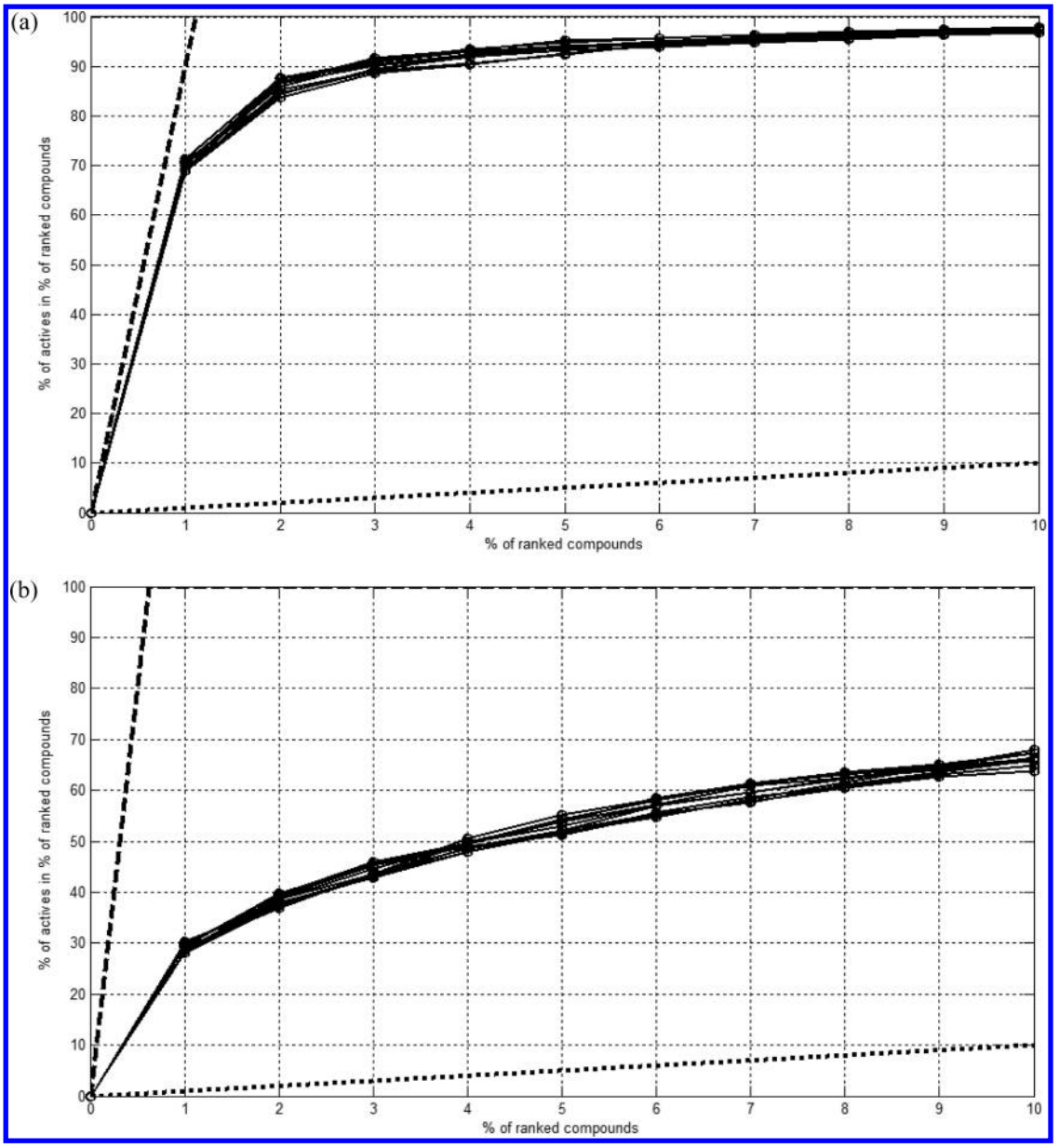


Figure 4. Enrichment curves for 10 separate runs of the GA on two MDDR activity classes: (a) renin inhibitors; (b) cyclooxygenase inhibitors.

Table 2. Numbers of Top-1% Actives in Common in Different Runs of the GA[a]

activity class	run									
renin	701	713	720	725	717	706	705	702	722	717
		653	672	688	779	645	656	639	675	662
cyclooxygenase	167	163	161	174	171	170	164	161	163	169
		150	146	162	150	159	142	141	148	154

[a]The value in the second column of the main body of the table is the number retrievd in the first run; each subsequent entry gives the number retrieved in a run and then below that the number in common with the first run.

of both of the numbers of actives retrieved at the top of the ranking and of the diversity of those actives. The sole exception to this generalization is the WOMBAT protein kinase C inhibitors where the GA and the NBC retrieved the same number of scaffolds. For comparison with the results in Table 1, experiments were carried out with the MDDR renin and cyclooxygenase data in which the active/inactive codes for the training-set molecules were randomly permuted prior to the execution of the GA, and the resulting weights were then applied to the test set. This *y*-randomization procedure was repeated 100 times: the mean and standard deviation of the numbers of actives retrieved in the top 1% were 4.16 and 8.10 (renin) and 5.31 and 6.96 (cyclooxygenase), values that are drastically less than those listed in Table 1 and that demonstrate the robustness of the GA solutions.

The effectiveness of screening is illustrated diagrammatically by the enrichment curves shown in Figure 2a−e; these curves are for some of the MDDR activity classes, but entirely

DOI: 10.1021/ci500540s
J. Chem. Inf. Model. 2015, 55, 214−221

comparable behavior is observed with the WOMBAT data set. Each such curve shows the percentage of the actives retrieved in the top X% of the ranked test set for $X \leq 10$ (since it is only the top-ranked molecules that are of interest in a virtual screening context). It will be seen that in all cases the GA curve is above that for NBC at the top of the ranking (as indicated by the results in Table 1) but that the NBC curve sometimes approaches (for the thrombin actives in Figure 2c), or crosses (for the 5HT reuptake inhibitor, 5HT1 agonist, and D2 antagonist actives in parts b, d, and e of Figure 2, respectively), the GA curve when a larger percentage of the top-ranked molecules is considered. This is in no way surprising since the fitness function of the GA focuses specifically on clustering the actives in the top 1% of the ranked test set, without any consideration being given to their occurrence in the bottom 99% of the ranking. To illustrate the effect of this focus, consider the WOMBAT D2 plots shown in Figure 3a,b. In the first case, the fitness function is as described previously; i.e., it focuses on the top 1% of the ranking. In the second case, the fitness function has been set to maximize the number of actives in the top 10% of the ranking. In Figure 3a, the plots cross at about 5% of the ranking; in Figure 3b, the separation between the two plots at 1% is less marked, but the GA plot remains above the NBC plot throughout a much larger range of values.

An inherent limitation of any GA is its nondeterministic nature, and it is hence important to assess the degree of variation from one run to the next. Ten runs were carried out on the MDDR renin and cyclooxygenase actives with the results shown in Figure 4, where a high degree of clustering will be seen: the mean and standard deviation of the numbers of actives retrieved in the top 1% were 712.8 and 8.71 (renin) and 166.3 and 4.54 (cyclooxygenase). The level of consistency suggested by these results is shown in more detail in Table 2. Here, the first of the 10 runs for each data set was taken as a standard, and then the set of top-1% actives identified by this run (the first of the 10 columns in the main body of the table) was compared with the corresponding set in each of the nine remaining runs. The upper value for each run in Table 2 is the number of actives retrievd in the top 1% and the lower is the number in common with the first run. For example, the second renin run identified 713 top-1% actives, of which 653 were identical with the actives identified in the first run: as will be seen there is a very high degree of consistency in the identities of the actives returned in different runs. Experiments were also conducted to determine the consistency of the weights that are calculated in different runs of the GA. For these sets of 10 renin and cyclooxygenase runs, the Pearson correlation coefficient was computed between the sets of 166 weights computed for each distinct pair of runs. The mean and standard deviation for the coefficient averaged over the 45 pairs of runs for each activity class were 0.79 and 0.025 (renin) and 0.79 and 0.024 (cyclooxygenase).

CONCLUSION

In this work we have described a GA for the calculation of fragment weights for use in substructural analysis. The GA is extremely simple in concept but effective in operation, since simulated virtual screening experiments using MDDR and WOMBAT data show it to be consistently superior to an NBC in terms of both the numbers of active molecules retrieved and the range of scaffolds within those sets of actives.

Both approaches have their limitations. The NBC approach has a firm theoretical basis for the design of the fragment weights but involves the assumption that the fragment occurrences are statistically independent, an assumption that is known to be incorrect.[30,31] It would be possible to try to relax this assumption, as has been done when the Robertson–Spärck Jones weights (such as R4) have been used in information retrieval.[32] However, the resulting weights are far more complex in nature and have not proved to be any more effective in retrieving relevant documents than the basic approach that assumes independence,[33] and there hence seems little reason to believe that things would be any different in the present context. The GA approach as currently implemented has three limitations. First, its inherently nondeterministic nature that has been discussed previously. Second, it is currently focused purely on maximizing the numbers of retrieved active molecules retrieved at the top of a database ranking. However, it would be easy to implement more sophisticated functions that took account of multiple optimization criteria, e.g., the diversity of the top-ranked molecules and of their computed physicochemical properties in a manner analogous to the library design tool described by Gillet et al.[34] Third, the "black-box" nature of a GA means that there is no algebraic formulation of the weights (such as eq 1 for the R4 weight) that could be used to rationalize the results that have been obtained. We hope to address this in future work using a genetic programming approach to develop weighting equations based on variables such as a, i, A, and I in eq 1.

ASSOCIATED CONTENT

Supporting Information

Text outlining the Metlab pseudo-code for the GA. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: p.willett@sheffield.ac.uk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank a reviewer for comments on an earlier version of this manuscript.

REFERENCES

(1) Goldman, B. B.; Walters, W. P. Machine learning in computational chemistry. *Annu. Rep. Comput. Chem.* **2006**, *2*, 127–140.

(2) Melville, J. L.; Burke, E. K.; Hirst, J. D. Machine learning in virtual screening. *Comb. Chem. High-Throughput Screening* **2009**, *12*, 332–343.

(3) Mitchell, J. B. O. Machine learning methods in chemoinformatics. *WIREs Comput. Mol. Sci.* **2014**, *4*, 468–481.

(4) Plewczynski, D.; Spieser, S. A.; Koch, U. Performance of machine learning methods for ligand-based virtual screening. *Comb. Chem. High-Throughput Screening* **2009**, *12*, 358–368.

(5) 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.

(6) Redl, G.; Cramer, R. D.; Berkoff, C. E. Quantitative drug design. *Chem. Soc. Rev.* **1974**, *3*, 273–292.

(7) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. A statistical-heuristic method for automated selection of drugs for screening. *J. Med. Chem.* **1977**, *20*, 469–475.

DOI: 10.1021/ci500540s
J. Chem. Inf. Model. 2015, 55, 214–221

(8) Hodes, L. Computer-aided selection of novel antitumor drugs for animal screening. *ACS Symp. Ser.* **1979**, *112*, 583−602.
(9) Hodes, L. Computer-aided selection of compounds for antitumor screening: Validation of a statistical-heuristic method. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 128−132.
(10) Hodes, L. Selection of molecular fragment features for structure-activity studies in antitumor screening. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 132−136.
(11) Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov, V. Discriminating between drugs and nondrugs by prediction of activity spectra for substances (PASS). *J. Chem. Inf. Comput. Sci.* **2001**, *44*, 2432−2437.
(12) 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.
(13) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers. *J. Chem. Inf. Model.* **2006**, *46*, 193−200.
(14) Klon, A. E.; Glick, M.; Davies, J. W. Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 4356−4359.
(15) Capelli, A. M.; Feriani, A.; Tedesco, G.; Pozzan, A. Generation of a focused set of GSK compounds biased toward ligand-gated ion-channel ligands. *J. Chem. Inf. Model.* **2006**, *46*, 659−664.
(16) Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. PASS: Prediction of activity spectra for biologically active substances. *Bioinformatics* **2000**, *16*, 747−748.
(17) Poroikov, V. V.; Filimonov, D. A.; Borodina-Yu, V.; Lagunin, A. A.; Kos, A. Robustness of biological activity spectra predicting by computer program PASS for non-congeneric sets of chemical compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1349−1355.
(18) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. 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.* **2006**, *46*, 462−470.
(19) 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.
(20) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10*, 682−686.
(21) Hassan, M.; Brown, R. D.; Varma-O'Brien, S.; Rogers, D. Cheminformatics analysis and learning in a data pipelining environment. *Mol. Diversity* **2006**, *10*, 283−299.
(22) Clark, D. E. *Evolutionary Algorithms in Computer-Aided Molecular Design*; Wiley-VCH: Weinheim, Germany, 2000.
(23) Gardiner, E. J.; Gillet, V. J.; Haranczyk, M.; Hert, J.; Holliday, J. D.; Malim, N.; Patel, Y.; Willett, P. Turbo similarity searching: Effect of fingerprint and dataset on virtual-screening performance. *Stat. Anal. Data Mining* **2009**, *2*, 103−114.
(24) Truchon, J.-F.; Bayley, C. I. Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.
(25) Clark, R. D.; Webster-Clark, D. J. Managing bias in ROC curves. *J. Comput.-Aid. Mol. Des.* **2008**, *22*, 141−146.
(26) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(27) Robertson, S. E.; Spärck Jones, K. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.* **1976**, *27*, 129−146.
(28) Ormerod, A.; Willett, P.; Bawden, D. Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct.-Act. Relat.* **1989**, *8*, 115−129.
(29) Cosgrove, D. A.; Willett, P. SLASH: A program for analysing the functional groups in molecules. *J. Mol. Graphics Modell.* **1998**, *16*, 19−32.
(30) Adamson, G. W.; Lambourne, D. R.; Lynch, M. F. Analysis of structural characteristics of chemical compounds in a large computer-based file. Part III. Statistical association of fragment incidence. *J. Chem. Soc., Perkin Trans.* **1972**, *1*, 2428−2433.
(31) Chen, N. G.; Golovlev, V. Structural key bit occurrence frequencies and dependencies in PubChem and their effect on similarity searches. *Mol. Inf.* **2013**, *32*, 355−361.
(32) Van Rijsbergen, C. J. A theoretical basis for the use of co-occurrence data in information retrieval. *J. Doc.* **1977**, *33*, 106−119.
(33) Smeaton, A. F.; van Rijsbergen, C. J. The retrieval effects of query expansion on a feedback document retrieval system. *Comput. J.* **1983**, *26*, 239−246.
(34) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. Selecting combinatorial libraries to optimize diversity and physical properties. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169−177.

DOI: 10.1021/ci500540s
J. Chem. Inf. Model. 2015, 55, 214−221