Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms[†]

Valerie J. Gillet,[‡,*] Peter Willett,[‡] and John Bradshaw[§]

Department of Information Studies and Krebs Institute for Biomolecular Research, University of Sheffield,
Western Bank, Shefffield S10 2TN, United Kingdom, and Glaxo Wellcome Research and Development
Limited, Gunnels Wood Road, Stevenage, SG1 2NY, United Kingdom

Received August 28, 1997

A substructural analysis approach is used to calculate biological activity profiles, which contain weights that describe the differential occurrences of generic features (specifically, the numbers of hydrogen-bond donors and acceptors, the numbers of rotatable bonds and aromatic rings, the molecular weights, and the $^2\kappa_\alpha$ shape descriptors) in active molecules taken from the World Drug Index and in (presumed) inactive molecules taken from the SPRESI database. Even with such simple structural descriptors, the profiles discriminate effectively between active and inactive compounds. The effectiveness of the approach is further increased by using a genetic algorithm for the calculation of the weights comprising a profile. The methods have been successfully applied to a number of different data sets.

INTRODUCTION

The recent developments in high-throughput screening technology in drug discovery have resulted in an emphasis on increasing the range of structural types that is available for testing. A number of different methods exists for generating large numbers of compounds; for example, selective compound-acquisition programs[1] and combinatorial synthesis.[2] High-throughput screening can be used at different stages of the drug discovery cycle; for example, in lead generation it may be used to screen as wide a diversity of different structural types as possible, whereas in lead optimization, the aim is to focus on particular structural types. The limiting factor in this approach to drug discovery is the rate at which potential candidates can be screened. Thus, there is much interest in selecting which compounds to screen, both for diverse and focused libraries. In both cases, and especially when the biological target is not known, the effectiveness of high-throughput screening will be maximized if the compounds to be screened exhibit ‘drug-like’ characteristics.

In this paper we seek to characterize ‘drug-like’ by analyzing what is currently known about bioactive molecules, and we define a molecule as ‘drug-like’ if it has characteristics that are similar to known bioactive molecules. The basic assumption is made here that molecules with similar characteristics are likely to exhibit similar activity. It is important to note that in this study we are interested in predicting biological activity across the broad range of known activity classes rather than for particular biological targets. Although it would be of great benefit to be able to predict the biological activity or inactivity of a given molecule directly, this is unlikely to be possible because activity arises from very different combinations of geometric, structural, and chemical properties. Therefore, rather than attempting to say precisely what characteristics a biologically active molecule should have, we have taken a probabilistic approach to the prediction of biological activity. We describe the development of scoring schemes that enable collections of molecules to be ranked according to their likelihood of exhibiting activity. If the compounds are then screened in rank order, the active molecules should be found more rapidly than if they are screened at random. The scoring schemes are developed by analyzing a large body of data that represents the full extent of our publicly available knowledge of molecules about which activity data is known (or presumed). In summary, we are attempting to characterize the full range of activity types that are known and to differentiate molecules with these activities from molecules that are known (or presumed) to be inactive, so that we can make predictions about collections of molecules of unknown activity.

The assumption that screening will be most effective if the compounds that are presented exhibit characteristics that are typical of those that have been previously demonstrated to exhibit activity is clearly an extremely strong assumption because it implies a degree of bias against molecules that are (even moderately) different from those that have been shown to exhibit activity in the past. However, the structure−activity relationships (SAR) that exist in known areas of biological activity are not fully understood and there remains a wealth of information pertaining to bioactivity that is buried within the data. Our work can thus be regarded as a contribution to the growing discipline of “data mining”.

Initially, we focus on methods that are based on substructural analysis techniques that have already been applied in SAR studies. Most existing SAR methods, such as Hansch

[*] Author to whom all correspondence should be addressed. E-mail: V.GILLET@SHEFFIELD.AC.UK.

[†] First presented at Random & Rational: Drug Discovery via Rational Design and Combinatorial Chemistry, Strategic Research Institute, Princeton, NJ, Sept., 1995.

[‡] Department of Information Studies and Krebs Institute for Biomolecular Research.

[§] Glaxo Wellcome Research and Development Limited.

S0095-2338(97)00431-9 CCC: $15.00
Published on Web 02/24/1998

analysis[3] or CoMFA,[4] are designed to optimize activity within a (generally) small series of related molecules. These methods require the calculation or measurement of many detailed descriptors whereupon irrelevant descriptors are identified and removed because they add noise to the system and dilute the signal. These methods are not appropriate for our use because we are not looking for characteristics that fit one particular, SAR rather we are looking to fit many different relationships. The methods also must be applicable to very large sets of compounds, so they must be computationally rapid.

Cramer et al.[5] introduced the SAR method known as substructural analysis, which is designed specifically for the analysis of large, structurally heterogeneous data sets and that may thus be used for lead discovery. The two basic assumptions in a substructural analysis are that it is possible to calculate a weight for a fragment substructure that characterizes its differential occurrence in active molecules and in inactive molecules, and that the overall probability of activity of a molecule may then be calculated by summing (or otherwise combining) the weights for its constituent fragment substructures. There have been many reports of the use of substructural methods, using a range of different types of fragment weight.[6−13]

Substructural analysis is normally applied to data from within an organization's corporate database, typically using the fragment substructures that form the basis for the initial, screening stage of a two-dimensional (2D) chemical substructure search and using activity data from primary biological screening results. However, there is no reason in principle why it cannot also be applied to data extracted from the many large databases of molecules that are now publicly available, using different types of (sub)structural feature and using nonspecific activity classes. We report the use of substructural analysis methods to calculate what we shall refer to as *biological activity profiles*, where a biological activity profile consists of a series of weights that are associated with a number of high-level structural features that are thought to affect (either positively or negatively) the tendency of a molecule to exhibit biological activity. Initially, we have adapted substructural analysis techniques to consider the combination of a range of generalized descriptors of molecules. We have then developed a genetic algorithm as a more effective and flexible means of identifying many different combinations of features that may be of relevance to activity.

EXPERIMENTAL METHODS

The biological activity profiles are derived using structure and activity information extracted from two widely available commercial chemical databases: the SPRESI databases and the World Drug Index (WDI).[15] The version of the SPRESI database used here has 1 711 301 parent molecules that contain only the following elemental types: C, N, O, F, P, S, Cl, Br, and I. The 30 621 molecules in SPRESI that also occur in WDI were removed to leave a set of 1 680 780 molecules that are distinct from those in WDI. Throughout the rest of this paper, SPRESI is used to refer to this set of compounds and, in what follows, it is assumed that the SPRESI file represents the world of inactive molecules. In practice, of course, there may well be SPRESI molecules that have not yet been identified as potential active molecules, but the percentage of these is assumed to be negligible. (According to Young,[16] drug companies typically screen ten thousand molecules to find a novel lead compound. Drug activity, therefore, is a rare event and the chance of finding active compunds in SPRESI is low.) In general, the molecules in the WDI database are assigned to activity classes using key words defined by Derwent.[17] The activity classes were examined and WDI was reduced to 14 861 molecules by removing those molecules with no activity class assigned, molecules that are labeled as "trial-prep", and molecules that belong to the following activity classes: pesticides and plant hormones (except for fungicides), zootoxins, toxins, surfactants, diagnostics, chelators, and absorbents. It is assumed that the remainder of WDI represents a wide variety of active molecules and that WDI is not biased towards any particular class(es) of compound(s), although an inspection of its contents suggests that at least some classes, such as antimicrobials, are overly represented. Both the SPRESI and the WDI data were preprocessed so that only parent compounds were included and, where possible, charges were neutralized by altering the number of hydrogens. This procedure ensured that the molecules were treated consistently with respect to pH and allowed simpler definitions of hydrogen-bond donors and acceptors to be used.

The activity profiles are based on a number of high-level structural features of molecules that have been suggested by Glaxo Wellcome medicinal chemists as being most likely to be involved in a pharmacophore. In addition, some physicochemical properties are also assumed to be of relevance for activity; for example, it is commonly assumed that most active molecules have a molecular weight between 150 and 550 for reasons of solubility and bioavailability.[18] The features included in this study are: the molecular weight (MW), the $^2\kappa_\alpha$ shape indexes,[19] and the numbers of aromatic rings (AR), rotatable bonds (RB), hydrogen-bond donors (HBD), and hydrogen-bond acceptors (HBA) in a molecule. A further feature suggested by the chemists, ClogP, was not used in the WDI/SPRESI experiments because it was not possible to obtain accurate data for sufficient numbers of compounds using the software that was available.

The MW, AR, and $^2\kappa_\alpha$ shape index features were calculated using the Daylight toolkit.[20] An HBD is defined as any heteroatom that carries at least one hydrogen, and an HBA is defined as a heteroatom with no positive charge, excluding the halogens, aromatic oxygen, sulfur, and pyrrole nitrogen and the higher oxidation levels of nitrogen, phosphorus, and sulfur. Note that an atom can be considered as both an HBD and an HBA. The SMARTS definitions of these substructural features are given in Table 1. The features (number of HBDs, HBAs, RBs, AR, MW and $^2\kappa_\alpha$) were calculated for each of the molecules in SPRESI and WDI and the results stored as THOR databases[20] for further analysis.

The distribution of each feature in a database is represented by a set of bins with a total of 20 bins per feature. The structural features HBD, HBA, RB, and AR are represented by counts, and the bin size is set to unity. Thus, for HBD, the first bin represents the number of molecules in the database that have no donors, the second bin represents the

Table 1. SMARTS Definitions for Substructural Features

feature	SMARTS
HBD	[!#6;!H0]
HBA	[$([!#6;+0]);!$([F,Cl,Br,I]);!$([o,s,nX3]);!$([Nv5,Pv5,Sv4,Sv6])]
RB	[! $([NH]!@C(=O))&!D1&!(*#*)]&!@[!$([NH]!@C(=O))!D1&!(*#*)]

Table 2. χ^2 Values to Compare the Distribution of Features in 14 861 WDI Molecules With Their Distribution in 16 807 SPRESI Molecules With υ Degrees of Freedom[a]

feature	$\chi^{2\,a}$	υ
HBD	5231	19
HBA	2739	19
MW	2099	19
$^2\kappa_\alpha$	1441	19
RB	855	19
AR	715	10

[a] The critical value of χ^2 is 43.82 ($\upsilon = 19$, $p \leq 0.001$).

number of molecules with exactly one donor, and so on with the final bin representing the number of molecules with 19 or more donors. The physicochemical properties are also represented by bins, but in these cases the bins represent ranges of values. For example, the first bin for $^2\kappa_\alpha$ represents the number of molecules with $^2\kappa_\alpha$ values in the range 0.00−1.99, the second bin represents the number of molecules with values in the range 2.00−3.99, and so on. The bins representing the distribution of MW have a range of 75, so that the bins represent the ranges 0.00−74.99, 75.00−149.99, ... and ≥1425.00.

Limited computing resources prevented the whole of SPRESI (~1.7 million compounds) from being processed in one unit; however, Daylight provide a utility for extracting compounds from a database. The utility was used to divide SPRESI into 10 subsets by taking the first of every 10 molecules throughout the database; then the second of every 10; and so on. The distributions of the structural features in each of the 10 subsets of SPRESI were compared using the χ^2 test, and there were no significant differences between them. One of the subsets was then taken and further divided into 10 subsets using the same Daylight utility. Again the differences in distributions were compared and no significant differences were found. Thus, it was concluded that when the molecules are described by the features identified here, a subset of SPRESI consisting of 16 807 compounds can be used as representative of the whole database. This subset was then used to represent SPRESI in all of the experiments reported (except for the predictive experiments) to minimize processing costs.

The distributions for the occurrences of the various features in the two databases are shown in Figure 1 where the WDI (14 861 molecules) and SPRESI (16 807 molecules) distributions are shown by dashed and solid lines, respectively. The significance of the differences in the pairs of distributions were calculated using the χ^2 test and the resulting statistic values are listed in Table 2. All of these values are significant ($p \leq 0.001$), with the occurrence of the HBD feature providing the greatest discrimination. The fact that there are differences between the distributions of the features between active molecules and inactive molecules is a finding that agrees with chemists' intuition.

Creation of Profiles Using Substructural Analysis. The weighting schemes that are traditionally used in substructural analysis describe the contributions of substructural fragments to the overall activity or inactivity of molecules that contain them. The weights are calculated from the relative frequencies of fragment occurrences in known active molecules and in known inactive molecules and they therefore provide a measure of the likelihood that a molecule having a particular fragment will be active or inactive. For example, a fragment that occurs in a large proportion of the active molecules but that occurs relatively infrequently in the inactive molecules will be assigned a high weight. A high weight, therefore, indicates an increased likelihood that a molecule containing the fragment will be an active molecule rather than an inactive molecule. However, it is still possible that the molecule is inactive even if it does contain the fragment. In substructural analysis, weights are assigned to many fragments and a molecule is scored according to the contributions of its constituent fragments. The ability of the scoring schemes to predict activity can be measured by scoring and ranking a set of molecules of known activity or inactivity. The molecules are ranked in descending order of score: molecules at the top of the resulting ranking will contain fragments with high weights and, if the weighting scheme is effective, they will have a high probability of being active. An effective scoring scheme can then be applied predictively to molecules of unknown activity.

The results reported in this paper are based on use of the SAF weight, which is that described by Cramer et al.[5] in the first publication on substructural analysis. The SAF weight for the *I*th fragment is given by:

$$SAF(I) = \frac{ACT(I)}{ACT(I) + INACT(I)}$$

where $ACT(I)$ of the active molecules and $INACT(I)$ of the inactive molecules contain the fragment.

We have calculated similar weights to reflect the relative frequency distributions of the various generalized features described earlier to see if they are effective in discriminating between WDI structures and SPRESI structures. A random sample of 1000 WDI molecules was used as representative of WDI. The distributions of the features are determined for each file, and stored in 20 bins for each feature, as described previously. The frequency information is used to calculate weights using one of two weighting schemes, each of which seeks to quantify the differential occurrences of particular structural features or ranges of property values in active molecules and in inactive molecules.

In this case, each feature is represented by a series of weights according to the differential distributions of different occurrences of the features in the two databases. For example, the feature HBD is represented by 20 weights, $HBDw_0$, $HBDw_1$, ...$HBDw_{19}$, which are the weights for exactly zero HBDs in a molecule, exactly one HBD, and up to >19 HBDs in a molecule, respectively. The weights are

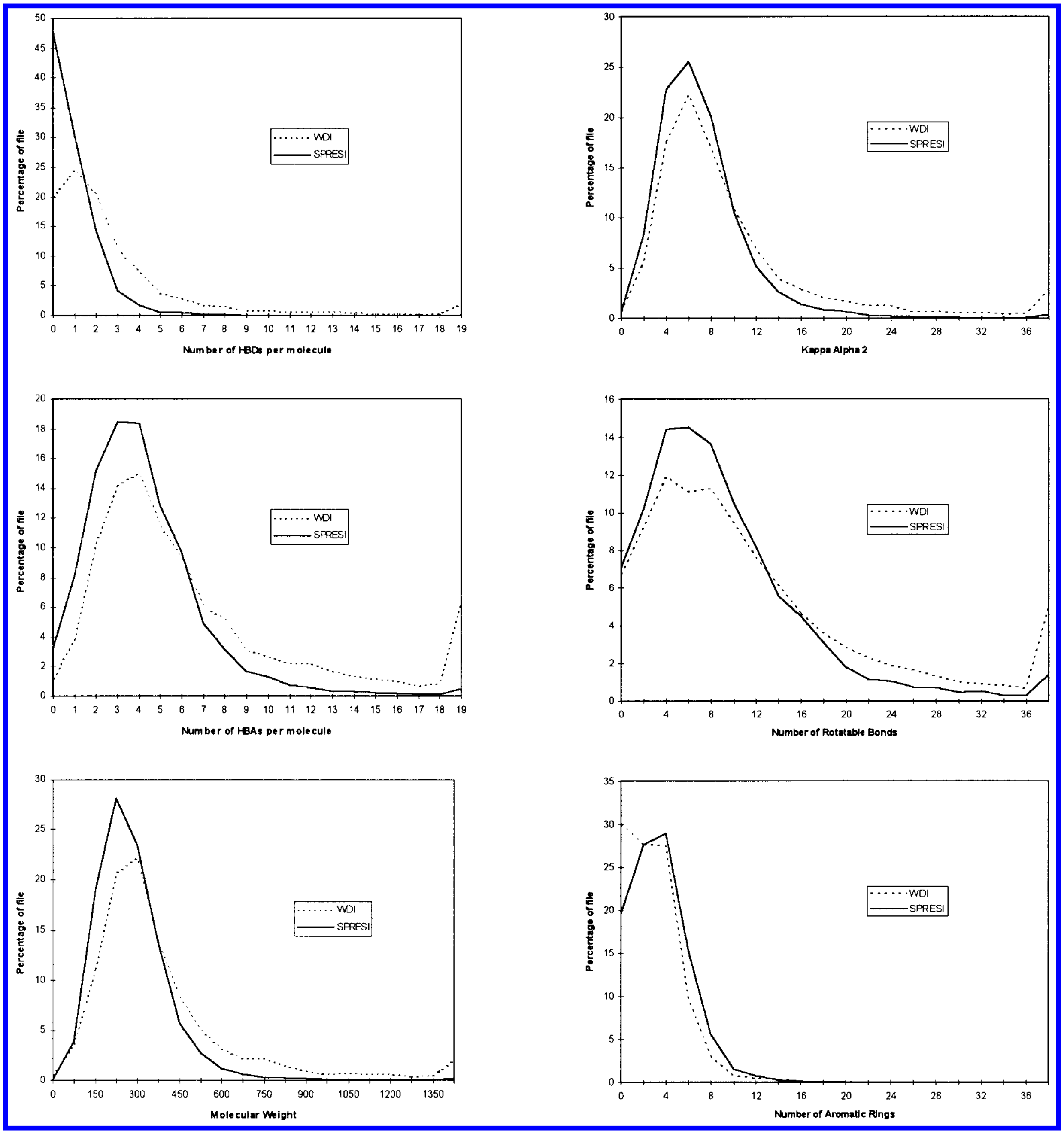


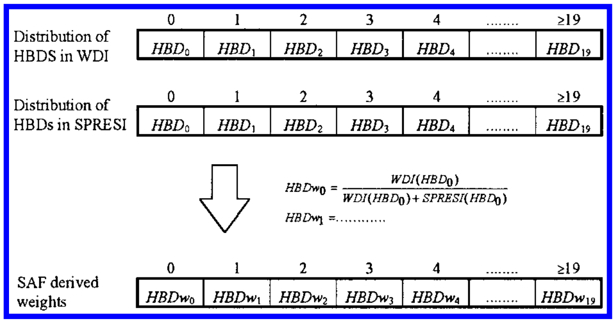
Figure 1. Relative distributions of features in WDI and SPRESI.

calculated using the following formula:

$$HBDw_n = \frac{WDI\,(HBD_n)}{WDI\,(HBD_n) + SPRESI\,(HBD_n)}$$

where $HBDw_n$ is the weight for n occurrences of HBDs; WDI (HAD_n) is the number of molecules with exactly n occurrences of HBDs in the WDI file; $SPRESI$ (HBD_n) is the number of molecules with exactly n occurrences of HBDs in the SPRESI file; and n is in the range 0−19 (i.e., to represent 20 weights per feature). The calculation of the HBD weights is illustrated in Figure 2. Series of weights for each of the other features are calculated in a similar way.

Ormerod et al.[10,11] reported a detailed comparison of the many different weighting schemes that have been suggested for use in substructural analysis studies. They concluded that a weight that was originally used for text retrieval, referred to as *R2*, was the most effective. The *R2* weighting scheme was also adapted for features to give a different set



	0	1	2	3	4	....	≥19
Distribution of HBDS in WDI	HBD_0	HBD_1	HBD_2	HBD_3	HBD_4	....	HBD_{19}
Distribution of HBDs in SPRESI	HBD_0	HBD_1	HBD_2	HBD_3	HBD_4	....	HBD_{19}

$$HBDw_0 = \frac{WDI(HBD_0)}{WDI(HBD_0)+SPRESI(HBD_0)}$$

$HBDw_1$ =............

	0	1	2	3	4	....	≥19
SAF derived weights	$HBDw_0$	$HBDw_1$	$HBDw_2$	$HBDw_3$	$HBDw_4$	....	$HBDw_{19}$

Figure 2. Deriving SAF weights for a feature. Each feature is represented by 20 weights.

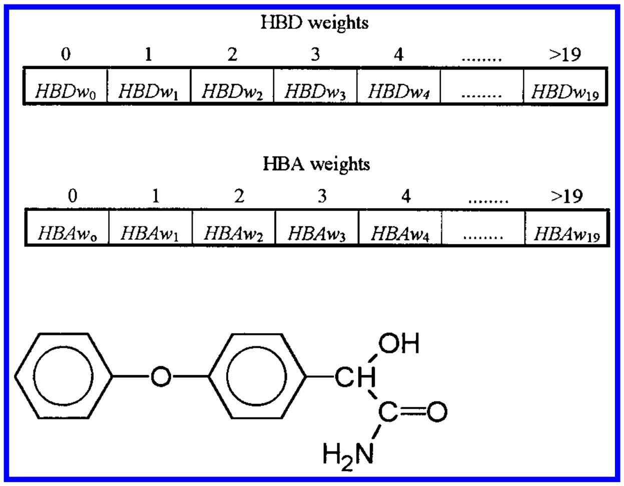


Figure 3. A molecule is scored by summing weights from the different features under consideration (HBD and HBA in this example). HBDs are represented by the 20 weights *HBDw*$_0$..*HBDw*$_{19}$ and HBAs are represented by the 20 weights *HBAw*$_0$..*HBAw*$_{19}$. The occurrences of each feature in the molecule are noted, and the appropriate weights then summed. The molecule shown here has two donors (represented by weight *HBDw*$_2$) and four acceptors (represented by weight *HBAw*$_4$), and therefore has the score *HBDw*$_2$ + *HBAw*$_4$ using the weight vectors shown in the top part of the figure.

of weights for each feature using the following formula:

$$HBDw_n = \log e\frac{WDI\ (HBD_n)/NWDI}{SPRESI\ (HBD_n)/NSPRESI}$$

using the notation given previously and where *NWDI* and *NSPRESI* are the total numbers of molecules in the WDI and SPRESI files, respectively. In fact, we report results only for the SAF weights because there was normally very little difference in performance between these two weighting schemes.

A series of weights was calculated for each feature using the available WDI and SPRESI data, and each molecule was then scored for the individual features and for different combinations of features. To score a molecule for a given feature, the number of occurrences of the feature in the molecule was found and the appropriate weight was retrieved. Molecules were scored for a combination of features (*e.g.*, HBD and HBA) by first finding the appropriate weight for each feature, and then calculating the score as the sum of the weights. This summation implicitly assumes that the features are statistically independent of each other. The process of scoring a molecule is illustrated in Figure 3.

Measurement of Performance. Once the molecules had been ranked, the performance of the weighting schemes was determined using performance measures adapted from Kearsley et al.[21] These measures are based on a simulated screening experiment on a database of molecules that contains some number of active molecules. The molecules are scored, ranked, and then "assayed" in order of descending score. The total number of active molecules found can be plotted against the total number of molecules tested (or the position in the ranked list). The two measures used here are the *initial enhancement* and the *global enhancement.*

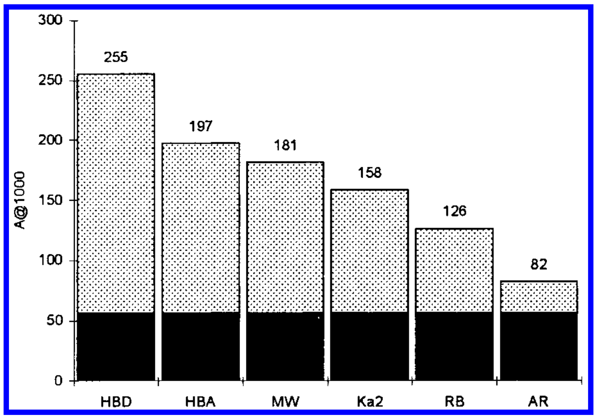


Figure 4. The number of WDI molecules ranking in the top 1000 positions after scoring and ranking 1000 WDI molecules and 16 807 SPRESI molecules using each feature is shown. The number of WDI molecules that would be expected by chance is shown in black.

If A@1000 is the number of active molecules found after testing the first 1000 molecules in the ranked list (when there is a total of 1000 active molecules), then the initial enhancement is how many more active molecules are found than is expected by chance; that is,

$$\text{Initial Enhancement} = \frac{\text{Actual A@1000}}{\text{A@1000 for random case}}$$

In the experiments reported here, there are 1000 WDI molecules (assumed active) and 16 807 SPRESI molecules (assumed inactive). A@1000 for the random case is calculated from:

$$\frac{1000}{1000 + 16\,807} \times 1000$$

that is, we would expect to find 56 WDI molecules in the top 1000 positions of the ranked list because of random effects.

In general, there can be more than one molecule with the same score so that after ranking, the molecules fall into score bands where all the molecules in a given score band have the same score. Within a particular score band, the molecules are ranked in arbitrary order. When the 1000th molecule is followed by other molecules with the same score, the percentage of WDI molecules relative to SPRESI structures in the whole score band was used in the calculation of the initial enhancement. For example, if the 1000th molecule in the ranked list is the 20th of 100 molecules with the same score, and if 25 of those 100 molecules are WDI molecules and 75 are SPRESI molecules, then 5 WDI molecules from this score band are added to A@1000.

If A50 is the number of molecules that must be tested to find half the active molecules (a measure analogous to the IC$_{50}$'s obtained in binding assays), then the global enhancement is the ratio of the actual A50 to the A50 expected for the random case:

$$\text{Global Enhancement} = \frac{\text{A50 for random case}}{\text{Actual A50}}$$

In terms of a compound selection program, the initial enhancement is a useful measure when only a small number of compounds can be selected, whereas the global enhancement is more relevant when large numbers of compounds can be selected.

Rankings Based on Single Features. Figure 4 shows the results of ranking 16 807 SPRESI molecules and 1000

Table 3. The Initial Enhancements and Global Enhancements Are Shown When SAF Weights for Each Feature Are Used To Score and Rank the 1000 WDI Molecules and the 16 807 SPRESI Molecules

feature	A@1000	initial enhancement	A50	global enhancement
HBD	255	4.6	4027	2.2
HBA	197	3.5	9246	1.4
MW	181	3.2	5751	1.5
$^{2}\kappa_{\alpha}$	158	2.8	7444	1.2
RB	126	2.3	7567	1.2
AR	82	1.5	6672	1.3

WDI molecules for each feature in isolation using the SAF weighting scheme (with 20 weights for each feature). The number of WDI compounds found in the first 1000 positions in the ranked lists (A@1000) is shown as a grey bar for each feature. If the WDI compounds were distributed randomly throughout the list (i.e., if there was no association between the weights and activity), then we would expect to find 56 compounds in the first 1000 positions of the list (shown by the area shaded black in each bar). The initial enhancements and global enhancements are also tabulated in Table 3.

In all cases, more of the WDI structures are ranked at the top of the list than would be expected if there was no association between the weights and activity, although the numbers are (hardly surprisingly) much less than the value of 1000 that would be obtained for perfect differentiation. An initial enhancement of 4.6 (or 255 WDI compounds found in the top 1000 positions of the list) is achieved by ranking the compounds using HBDs alone. This value is ~5 times as many WDI molecules as would be expected by chance. Therefore, we conclude that just the number of occurrences of a single high-level structural feature provides a surprisingly high level of discrimination between active molecules and inactive molecules, with the relative effectiveness of the individual features being in agreement with the χ^2 tests discussed previously.

Validation of Rankings. We have validated the results in Table 3 in two ways: by randomly assigning WDI labels to molecules and by using the weights predictively.

In the first set of experiments, the origins of the molecules were scrambled. The original WDI and SPRESI labels were ignored and 1000 molecules were selected at random and assigned WDI labels. These molecules, which could have originated in either database, are referred to as pseudo-WDI molecules. The molecules were scored and ranked in the normal way and the number of pseudo-WDI molecules in the top 1000 positions noted. By chance it is expected that 56 of the molecules that are given pseudo-WDI labels are true WDI molecules. Also by chance, some of these true WDI molecules would be expected to rank in the top 1000 positions of the list; in fact, on average, 3.1 of the true WDI molecules that are given pseudo-WDI labels should rank high. The number of true WDI molecules assigned as pseudo-WDI molecules is recorded along with the number of these that rank in the top 1000 positions. This process of scrambling the labels was repeated 1000 times and the results are shown in Table 4. The average number of pseudo-WDI molecules found was 54.4, with a standard deviation of 7.2. The average number of true WDI molecules assigned as pseudo-WDI label was 56.1, with a standard deviation of 7.3. The average number of true WDI molecules assigned as pseudo-WDI molecules that are found in the top 1000 positions was 16.6, with a standard deviation 3.9, which represents 5.4 times the 3.1 molecules that would be expected by chance. This result agrees with the previous findings of a near fivefold increase on chance of being able to distinguish WDI molecules from SPRESI molecules. It is concluded that the nonscrambled WDI molecules give significantly better results ($p < 0.0001$ in the Z test) than do the pseudo-WDI molecules.

Table 4. Results of Scrambling Labels[a]

feature	no. pseudo-WDI hits	no. pseudo WDI correctly assigned	no. correctly assigned WDI found as hits	OBS
HBD	54.4 (7 2)	56.1 (7.34)	16.6 (3.9)	255

[a] The column labeled no. pseudo-WDI hits gives the A@1000 values (mean and standard deviation in brackets) when the pseudo-WDI molecules are used to calculate the weights, averaged over 1000 runs. The column labeled no. pseudo-WDI correctly assigned gives the number of pseudo-WDI molecules that are true WDI molecules, and the next column gives the number of these that are found in the top 1000 positions of the list. By chance, we would expect three of the true WDI molecules that are labeled pseudo-WDI to be found. These results agree with the previous findings of a near fivefold increase on chance in being able to distinguish between WDI and SPRESI molecules. OBS gives the A@1000 obtained when the real WDI molecules are used to calculate the weights.

Table 5. Predictive Use of HBD and HBA Weights[a]

feature	% training set	A@1000	initial enhancement	A50	global enhancement
HBD	100	255	4.6	4027	2.2
	75	255	4.6	4027	2.2
	50	252	4.6	3836	2.3
	25	243	4.3	4070	2.2
HBA	100	197	3.5	6246	1.4
	75	197	3.5	6246	1.4
	50	181	3.2	6239	1.4
	25	184	3.3	7541	1.2

[a] The training set is the set of molecules that was used for the calculation of the weights and is given as a percentage of the complete files. The other columns show the performance achieved by applying the weights derived for the training sets to the larger files.

The prediction experiments used a leave-*n*-out procedure. Here, training sets were created by removing increasing numbers of molecules from the 1000 WDI molecules and the 16 807 SPRESI molecules; the weights were calculated for the molecules in the training set and then applied to the remaining molecules in the prediction set. The experiments were repeated a number of times using different percentages of the whole data to calculate the weights, with the same percentage of structures being removed at random from both WDI and SPRESI. The results are shown in Table 5. Even the smallest training sets yielded weights that are little inferior to those resulting from analysis of the larger files.

Rankings by Combining Features. The next set of experiments investigated the effect of combining the weights for different features. Table 6 shows the results as successive features are combined by summing the corresponding weights. It is evident that the effectiveness of the rankings is not increased as more features are combined, with the initial enhancement falling to 4.3 when all six features are combined compared with a value of 4.6 when just the HBD

Table 6. Initial Enhancements and Global Enhancements When the 1000 WDI Molecules and 16 807 SPRESI Molecules Are Scored and Ranked Using the Substructural Analysis Sum-of-Weights Method to Combine the Weights for Different Combinations of Features

feature	A@000	initial enhancement	A50	global enhancement
HBD	255	4.6	4027	2.2
HBD + HBA	258	4.6	3866	2.3
HBD + HBA + MW	245	4.4	3904	2.3
HBD + HBA + MW + $^2\kappa_\alpha$	237	4.2	4139	2.2
HBD + HBA + MW + $^2\kappa_\alpha$ + RB	236	4.2	4162	2.2
HBD + HBA + MW + $^2\kappa_\alpha$ + RB + AR	239	4.3	3858	2.1

feature is employed. The global enhancement remains about the same regardless of the number of features included in the scores.

Although it is clear from these experiments that the weights have substantial predictive power, they also have clear limitations in that the use of more than a single feature in the profiles often resulted in no increase, or even a decrease, in the hit rate. The theoretical model underlying several of the substructural analysis weights that have been described assumes that the features in a molecule are statistically independent of each other. The independence assumption allows the overall probability of activity for a molecule to be calculated as the product of the individual features' probabilities, and hence for the logarithms of these probabilities to be added together, as occurs with the *R2* weight. However, the independence assumption is known to be incorrect for substructural features,[22] and it is clear from the results in Table 6 that the use of increasing numbers of features does not result in any consistent improvements in performance; indeed, there is generally a decrease in performance. Therefore, we decided to use an alternative strategy to calculate weights that did not involve a specific assumption of feature independence. Specifically, a *genetic algorithm* (GA) was developed that would lead to weights that would, hopefully, encompass at least some of the interfeature dependencies.

Creation of Profiles Using a Genetic Algorithm. A GA is a computational analogue of Darwinian evolution that can identify good, although not necessarily optimal, solutions to a wide range of combinatorial optimization problems.[23-25] Potential solutions to a problem are encoded in a population of chromosomes that are usually linear representations of the problem that is to be solved and that are scored using a fitness function. New populations are developed by performing genetic-like operations of mutation and crossover on some members of the existing population. Higher scoring individuals have a higher probability of passing their genes into the new populations. The new chromosomes are scored and the GA iterates, usually until it has converged on a solution. GAs have previously been applied to many problems in computational chemistry,[26] including the design of combinatorial libraries targeted at one particular biological assay[27] and the selection of diverse subsets of molecules.[28] Here, we have developed a GA for the identification of weights in substructural analysis using a procedure analogous to that described by Jones et al.[29] for the calculation of term-relevance weights in text retrieval. The aim of the GA is to identify sets of weights that will maximize the discrimination between active and inactive molecules when the molecules are represented by more than one feature.

The chromosomes in the GA encode the weights assigned to the features; that is, to given numbers of occurrences of a substructural feature and to different values of a physicochemical property. For example, a chromosome of length *n* can be used to encode the occurrences of different numbers of HBDs in a molecule, with the elements representing the occurrence of 0 donors, exactly 1 donor, ... exactly $n - 1$ donors. In the substructural analysis studies, each feature was represented by 20 bins, giving a total of 120 bins for the six features that were tested there. Experiments were performed with chromosomes of this size to compare the GA results with the previous studies. The initial population of chromosomes was assigned random weights.

A molecule is scored by counting the occurrence of each structural feature in the molecule and by calculating the value of the physicochemical properties. The appropriate weight for each feature is retrieved from the chromosome, the weights summed, and the sum-of-weights noted; the complete set of molecules is then ranked in order of decreasing sums-of-weights. Two slightly different functions can be used to evaluate these rankings, though both of them seek to maximize the discrimination between actives and inactives. The first fitness function, called *Top Rank*, counts the number of active molecules occurring in the top *N* molecules, where *N* is the total number of active molecules in the file (c.f., the initial enhancement described previously), with the GA attempting to maximize this number. The second fitness function, called *Average Rank*, calculates the average position of an active molecule in the ranking, with the GA attempting to minimize this function (c.f., the global enhancement described previously). The calculated fitness values are normalized by ranking the chromosomes in order of decreasing fitness and then creating fitnesses that begin with a constant value and decrease linearly. This linear normalization overcomes two potential problems: the existence of a super-fit chromosome can dominate the population and drive out less-fit members, thus causing premature convergence; and the lack of sufficient differentiation when there are several chromosomes with fitness values that are very close together.

The two standard genetic operators are mutation and crossover. Mutation consists of changing one or more elements of a single parent chromosome to produce a single new child chromosome. Mutation is implemented here such that an element (or elements) in a chromosome is assigned a new value at random, subject to it being within the permitted range of values. Mutation requires one parent chromosome, which is selected using roulette-wheel selection. The number of elements changed in each mutation is a parameter of the GA. Three different kinds of crossover are implemented: one-point crossover; two-point crossover; and uniform crossover. The crossover operators mix the information between two parent chromosomes to produce two new child chromosomes. In one-point crossover, a point along the parent chromosomes is chosen at random and the elements following the crossover point are exchanged in each parent; in two-point crossover, two points along the parent chromosomes are chosen at random and the elements between the two points are exchanged in each parent; and in uniform crossover, for each element a decision is made at random whether to exchange the element in the parents. A steady-state replacement strategy is used, in which a fixed percentage of the lowest-fitness chromosomes are replaced by new chromosomes in each generation. The type of operation performed in each iteration is based on user-defined operator weights.

RESULTS USING THE GENETIC ALGORITHM

It will be clear from the brief description already given that our GA has many parameters that can be altered over a series of runs. These parameters include the operator weights, the population size, the chromosome length, the selection pressure, the number of iterations, the relative weights of the various genetic operators, and the mutation rate. Extensive studies showed that varying the parameters could affect the rate of convergence but had little effect on the final results that were obtained, thus demonstrating the robustness of the profiles resulting from use of our GA.

In a GA, the fitness function is applied very frequently (i.e., each time a new chromosome is generated). The same files as used in the previous substructural analysis experiments were used as training sets (i.e., the 1000 WDI molecules and the 16 807 SPRESI molecules), with the weights derived from them subsequently being applied to larger files in the predictive experiments.

The initial runs sought to derive optimum weights for the HBD feature. Using a chromosome containing 20 bins and the fitness function Top Rank, the GA consistently finds sets of weights that provide exactly the same level of discrimination (an initial enhancement of 4.6) between the active molecules and the inactive molecules as do the SAF weights in the substructural analysis experiments. This result suggests that this value is, indeed, the best possible that can be found with this single feature. However, different sets of weights are produced in each run and they give rise to different hit lists, which implies that there is more than one optimal solution (also some of the weights may have no effect on the solution).

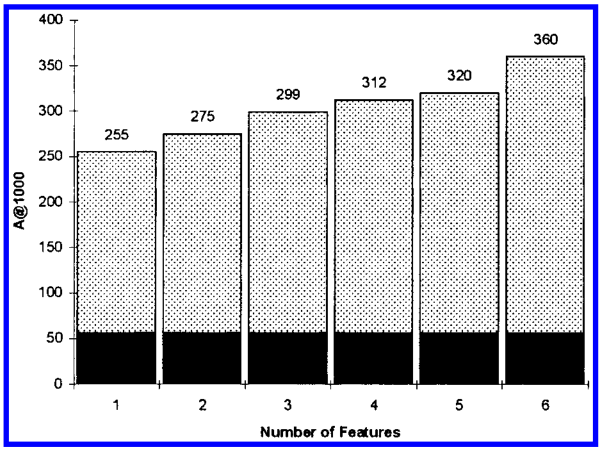


Figure 5. The numbers of WDI molecules ranking in the top 1000 positions of the ranked list increase as more features are included in the GA.

Figure 5 illustrates the effect on A@1000 of including additional features in the GA. The features are included in the same order as for the SAF experiments and the fitness function is Top Rank. It will be seen that the best initial enhancement achieved increases each time that an additional feature is included, thus demonstrating that the GA is better able to combine discriminatory information from different features than are the SAF weights used previously. The best result is obtained when all of the features are included: in this case, 360 WDI molecules are found in the top 1000 positions, representing an initial enhancement of 6.4. The experiments were repeated using the fitness function Average Rank. The results for both fitness functions are given in Table 7, and Figure 6 shows the total number of WDI structures retrieved at each position in the ranked lists using six features for both fitness functions. The dashed line labeled *random* in Figure 6 shows the plot that would be expected if the WDI compounds were distributed randomly throughout the list (i.e., if there was no association between the weights and activity). This line has a slope of:

$$\frac{1000}{1000 + 16\,807} = 0.056$$

and we would expect to find 56 WDI molecules in the top 1000 positions in the list. The dashed line labeled *ideal* shows the plot that would be expected if the weights were able to differentiate perfectly between WDI and SPRESI compounds (i.e., the ideal line has a slope of one for the first 1000 positions in the ranked list and it then has a slope of zero); all 1000 WDI molecules would be found in the top 1000 positions in the ranked list. The Top Rank function maximizes the number of WDI structures in the top 1000 positions in the ranked list and takes no account of where the remaining WDI structures appear in the list. Hence, there is a rapid drop off in WDI structures retrieved after 1000 positions as they tend to a random distribution. This phenomenon is also reflected in Table 7 where although the initial enhancement increases with additional features, the same trend is not seen in the global enhancement. In the case of Top Rank, the initial enhancement is 6.4, with a global enhancement of 2.5. The Average Rank function produces a smoother curve and tends to maximize the global enhancement. Thus, in this case, the initial enhancement is 5.2, with a global enhancement of 3.1. In this case, ∼50% of the WDI structures occur in the top 17% of the list or, in other words, 50% of the active compounds are found by assaying 17% of the total compounds.

The nondeterministic nature of the GA has the effect that different runs often correspond to different sets of weights although they may converge on the same solution in terms of initial or global enhancement (i.e., there is not a single unique set of weights that gives the best solution).

The GA weights derived from the reduced WDI and SPRESI data sets using the average rank fitness function were then applied to larger data sets, specifically 10 000 WDI structures and 168 071 SPRESI structures (again selected at random). The distributions of WDI structures throughout the ranked lists for the prediction data and the training data are shown in Figure 7 and the results are tabulated in Table 8. It can be seen that the weights perform almost as well

Table 7. Initial Enhancements and Global Enhancements When the GA Is Trained to Score and Rank 1000 WDI Molecules and 16 807 SPRESI Molecules with Increasing Numbers of Features

feature	A@1000	initial enhancement	A50	global enhancement
	(a) Top Rank[a]			
HBD	255	4.6	4858	1.8
HBD + HBA	275	4.9	4375	2.0
HBD + HBA + MW	299	5.3	4748	1.9
HBD + HBA + MW + $^2\kappa_\alpha$	312	5.6	4524	2.0
HBD + HBA + MW + $^2\kappa_\alpha$ + RB	316	5.7	4830	1.8
HBD + HBA + MW + $^2\kappa_\alpha$ + RB + AR	360	6.4	3634	2.5
	(b) Average Rank[b]			
HBD	253	4.5	3508	2.5
HBD + HBA	259	4.6	3387	2.6
HBD + HBA + MW	270	4.8	3201	2.8
HBD + HBA + MW + $^2\kappa_\alpha$	259	4.6	3089	2.9
HBD + HBA + MW + $^2\kappa_\alpha$ + RB	264	4.7	3151	2.8
HBD + HBA + MW + $^2\kappa_\alpha$ + RB + AR	290	5.2	2842	3.1

[a] The fitness function optimized by the GA is Top Rank. [b] The fitness function optimized by the GA is Average Rank.

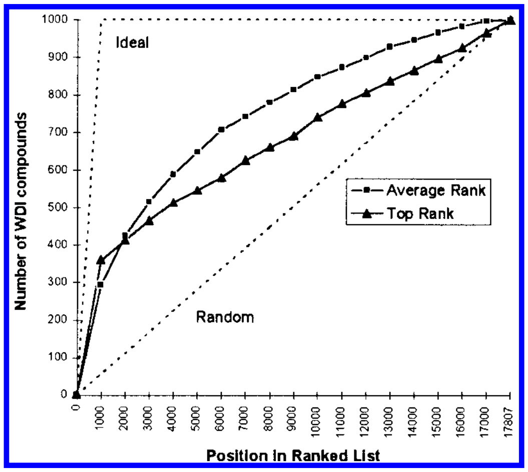


Figure 6. The positions of the WDI molecules within the ranked list are shown for the fitness functions Top Rank and Average Rank. Top Rank maximizes the number of WDI molecules found in the first 1000 positions in the ranked list but takes no account of where the rest of the WDI molecules occur; hence, the curve is steep initially and then tails off. Average Rank minimizes the average position of the WDI compounds in the ranked list; hence, the curve is smooth.

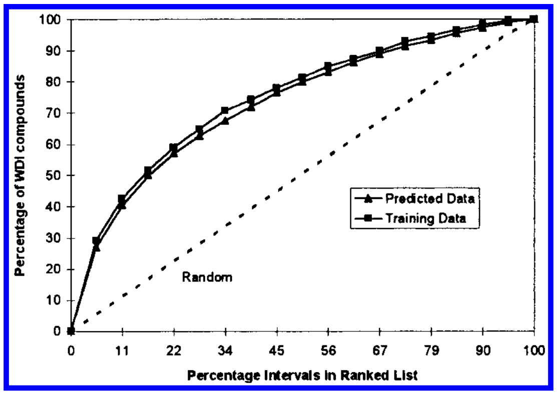


Figure 7. The GA weights produced using the Average Rank fitness function and trained on the 1000 WDI molecules and 16 807 SPRESI molecules perform almost as well when applied predictively to 10 000 WDI molecules and 168 072 SPRESI molecules.

Table 8. The Distribution of WDI Molecules in the Training Set (1000 WDI Molecules and 16 807 SPRESI Molecules) Compared with the Prediction Set (10 000 WDI Molecules and 168 071 SPRESI Molecules)

data	A@x[a]	initial enhancement	A50	global enhancement
		(a) Top Rank[b]		
training set	360	6.4	3434	2.5
predicted set	2814	5.0	36515	2.4
		(b) Average Rank[c]		
training set	290	5.2	2842	3.1
predicted set	2687	4.8	30021	3.0

[a] A@x for the training set is the number of WDI molecules found in the top 1000 positions of the ranked list and A@x for the predicted set is the number of WDI molecules found in the top 10 000 positions of the ranked list. [b] The results for ranking the predicted data sets using the best weights obtained using the fitness function Top Rank. [c] The result for ranking the predicted data sets using the best weights obtained using the fitness function Average Rank.

when applied predictively, thus demonstrating their power. The fitness function Average Rank results in a higher global enhancement when training the GA and when the GA is applied predictively and it produces a more even distribution of the active molecules throughout the ranked lists. Therefore, this function is used in the remaining GA experiments. Similarly, when the WDI/SPRESI weights are used predictively, they are the weights that have been derived using the function Average Rank.

Figure 8 shows a histogram of the scores for the entire ranked list of 16 807 SPRESI molecules and 1000 WDI molecules produced using Average Rank. The molecule scores have been scaled to fit in the range 0..100. The histogram approximates a normal distribution, which indicates a diverse range of molecules with no clumping effects. Figure 9 shows the histogram of scores for the WDI molecules (given as the percentage of the WDI file and shown as dark bars) separated from the histogram of scores obtained for the SPRESI molecules (given as the percentage of the SPRESI file and shown as grey bars). A clear separation between the databases can be seen, demonstrating the ability of the GA to push the WDI structures towards the top of the ranked list. It can be seen that the GA weighting scheme is very successful at ranking high some of the WDI molecules, whereas it is less able to rank other WDI molecules highly. The next section investigates whether the rankings are better for certain therapeutic classes of molecules.

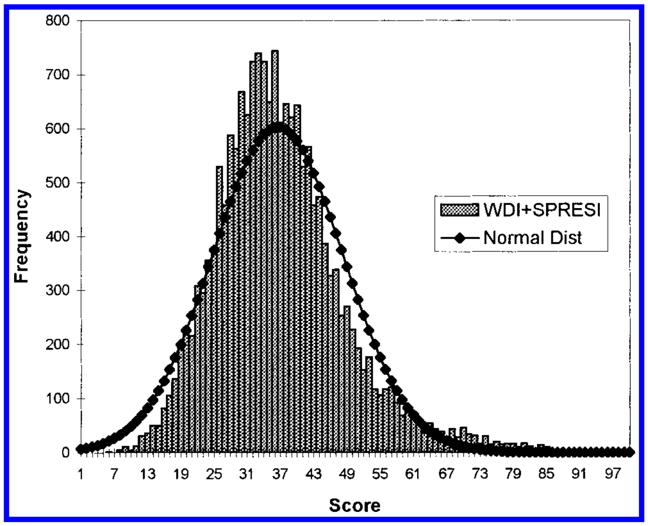


Figure 8. The histogram of scores is plotted for all molecules (16 807 SPRESI and 1000 WDI molecules) for the ranked list produced using the Average Rank fitness function.

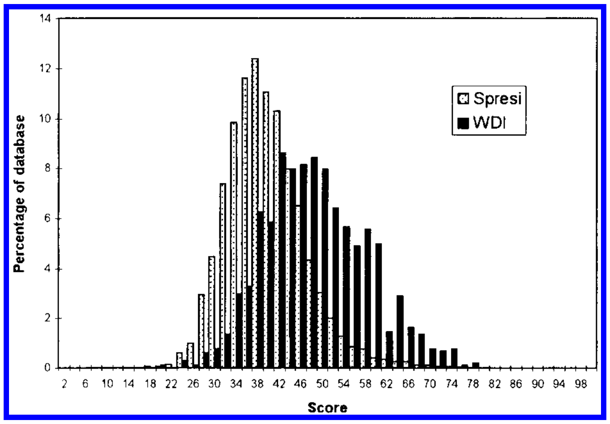


Figure 9. The histogram of scores obtained for the WDI molecules is shown in black and are (given as percentages of the file) separated from the histogram of scores obtained for the SPRESI molecules, shown in grey (also given as percentages of the file).

Although it could be argued that many of the WDI molecules are missed as a result of high-ranking SPRESI molecules, the fact that there is a greater than sixfold increase on chance of finding the WDI (or 'drug-like') molecules is of considerable value in high-throughput screening programs where it is necessary to select subsets of molecules from very large collections, such as combinatorial libraries. However, although these results are highly satisfactory, it must be emphasized that the best set of weights found when all of the features were used is unlikely to be the best possible set of weights, owing to the nonoptimal nature of the GA. Even so, it is clear that this GA approach is substantially superior to those based on the simple SAF (or R2) weights.

Applying the WDI/SPRESI Weights to Specific Therapeutic Classes. The relationship between the molecule scores and the activity classes of the WDI molecules has been investigated further. The molecules in WDI were divided into broad therapeutic classes according to the key words identified by Derwent,[17] as shown in Table 9. The numbers of molecules in each class that occur in the 1000 sample of WDI is also shown (note that some of the molecules belong to more than one class). Figure 10a shows the distribution of scores for the 294 antimicrobials that occur in the 1000 sample of WDI (in black) superimposed on the scores for the entire set of 1000 WDI molecules (in grey). This figure illustrates that the weights are effective in scoring the antimicrobials high within the rank list. In contrast, the weights perform poorly for the class of psychotropics, shown in black in Figure 10b. The weights are extremely effective in scoring the class of antibiotics high within the rank list, as shown in Figure l0c (the antibiotics are a subclass of the antimicrobials class).

The effectiveness of the WDI/SPRESI weights at scoring the therapeutic classes was investigated further. Subsets of 1000 compounds from each category were chosen at random from the 14 861 WDI molecules. The WDI/SPRESI-derived weights were used to score and rank each of these subclasses of WDI against the SPRESI file. The results, tabulated in Table 10, do indeed show that some of the broad therapeutic classes are better discriminated by the weights than others. The molecules in the antimicrobials class perform particularly well, perhaps for two reasons; firstly, there are twice as many molecules in this class as any of the other classes, hence they are well represented in WDI; and secondly, this class contains the subclass of antibiotics that are a structurally closely related set of molecules. At the other extreme, the psychotropics class of molecules perform poorly, presumably because the molecules within the class are not structurally closely related and also because the class as a whole is under-represented in WDI relative to the other therapeutic classes.

Profiles for Specific Therapeutic Classes. In the previous experiments, weights were derived to discriminate between WDI and SPRESI molecules as an attempt at scoring and ranking unknown libraries of molecules according to their likelihood of having biological activity in any therapeutic class. These 'general' weights were then applied to specific therapeutic classes. When libraries of molecules are to be screened for particular assays, it may be more appropriate to develop more specific weights. In the next set of experiments, the GA was used to derive different sets of weights to discriminate between each of the broad therapeutic classes extracted from WDI and the SPRESI molecules. In each case, the aim was to discriminate between the active compounds and the 16 807 SPRESI compounds used in the previous experiments. GA weights were derived for each of the classes using the same methods as used for the WDI compounds, and the molecules were scored and ranked as before.

The results of ranking the therapeutics classes are given in Table 11 and the distribution of actives in the ranked lists is plotted for some of the classes in Figure 11. As expected, in each case, the weights derived for each of the therapeutic classes give better results than when the WDI/SPRESI weights were used. Again, there is a marked difference in performance for the different classes that presumably reflects the degree of structural relatedness within each of the classes. Three of the classes, hormones, antimicrobials, and anticancers, show significant increases in discriminatory power over the broader class of WDI, with the hormones class showing an initial enhancement of 8.3 and with no less than 50% of the actives being found in the top 6% of the ranked list. Therefore, when the therapeutic classes contain related compounds very good discriminatory power is achieved. In other cases (e.g., central nervous system and psychotropic drugs), there is no increase in discriminatory power compared with the WDI compounds, presumably because each of these

Table 9. The WDI Activity Categories Identified by Key Words[a]

WDI activity category	abbreviation	no. structures in 14 861 WDI molecules	no. structures in 1000 WDI molecules
antimicrobials and chemotherapeutics	antimicrobials	4641	294
anticancer drugs and carcinogens	anticancers	2439	171
drugs acting on the blood and cardiovascular system	blood	2092	138
drugs acting on the nervous system	CNS	1924	133
anesthetics and drugs relieving fever, inflammation, and pain	anesthetics	1714	108
hormones and antagonists	hormones	1582	98
psychotropic agents	psychotropics	1139	89

[a] The number of molecules in each category that has >1000 members is shown; the second column lists the abbreviations used within the main body of the paper; the final column gives the number of molecules in each class that is present in the 1000 WDI sample (note that a molecule may be assigned to more than one class).

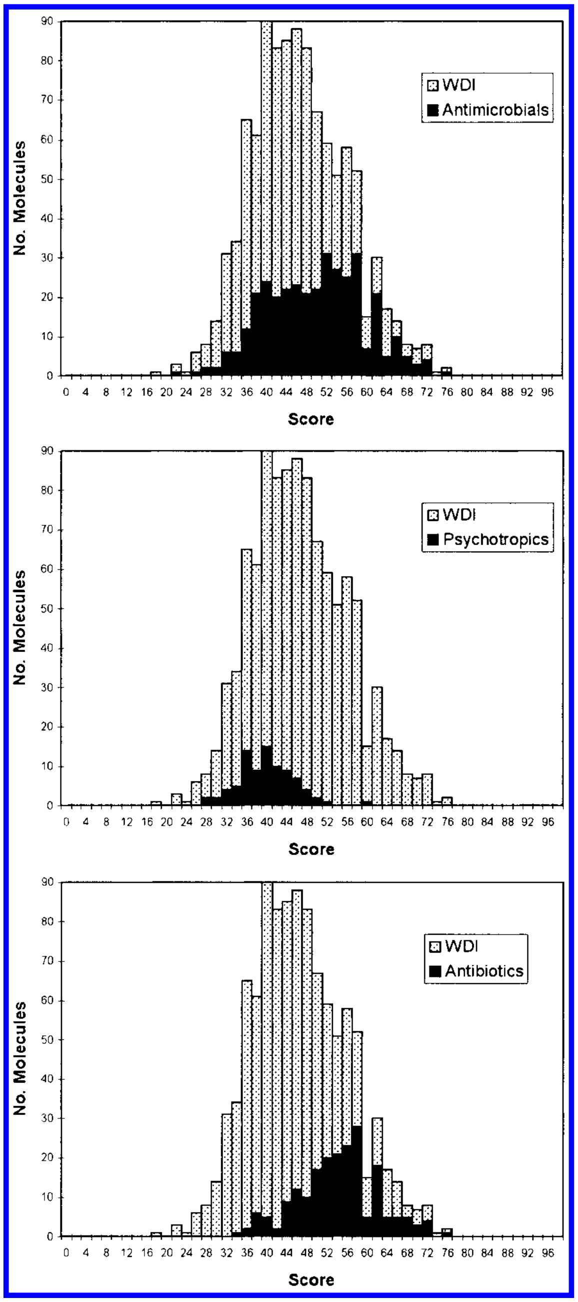


Figure 10. (a) The molecules in the class of antimicrobials tend to appear towards the top of the ranked list. (b) The psychotropics appear low down in the ranked list. (c) The antibiotics, a subclass of the antimicrobials, appear very high in the ranked list.

classes contain widely disparate compounds that operate by different mechanisms.

Table 10. Results of Applying the WDI/SPRESI-Derived Weights to the Specific Therapeutic Classes Identified in WDI[a]

therapeutic class	A@1000	initial enhancement	A50	global enhancement
anticancers	363	6.5	1815	4.9
hormones	335	6.0	1942	4.6
antimicrobials	366	6.5	1973	4.5
WDI	290	5.2	2842	3.1
blood	200	3.6	3893	2.3
anesthetics	212	3.8	4137	2.2
CNS	123	2.2	5141	1.7
psychotropics	68	1.2	6606	1.3

[a] The classes are listed in rank order of global enhancement.

Table 11. Results of Applying Weights That Are Derived to Discriminate Between Each of the Therapeutic Classes and the SPRESI Molecules

therapeutic class	A@1000	initial enhancement	A50	global enhancement
hormones	465	8.3	1076	8.3
anticancers	423	7.6	1301	6.8
antimicrobials	425	7.6	1429	6.2
WDI	290	5.2	2842	3.1
anesthetics	230	4.1	3029	2.9
blood	202	3.6	2964	3.0
CNS	216	3.9	3526	2.5
psychotropics	139	2.5	4477	2.0

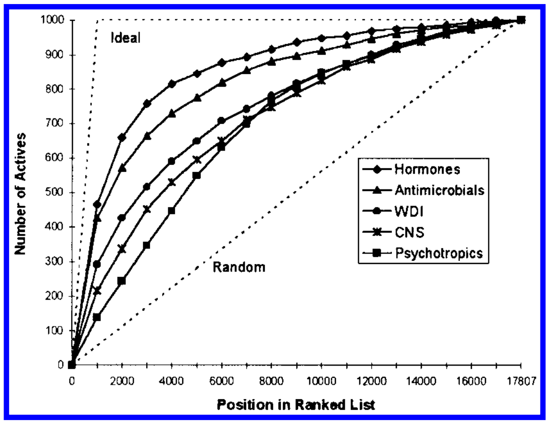


Figure 11. The GA is run to calculate weights that are specific to each of the therapeutic classes. The weights are then used to try to discriminate between each class and SPRESI molecules. The positions of the active molecules in the ranked lists are shown.

Finally, the methods were tested on a more specific set of compounds; that is, 1000 antibiotics extracted from WDI at random. The antibiotics form a subclass of antimicrobials that are, in their turn, a subclass of all the WDI compounds. GA weights were derived to discriminate each of these

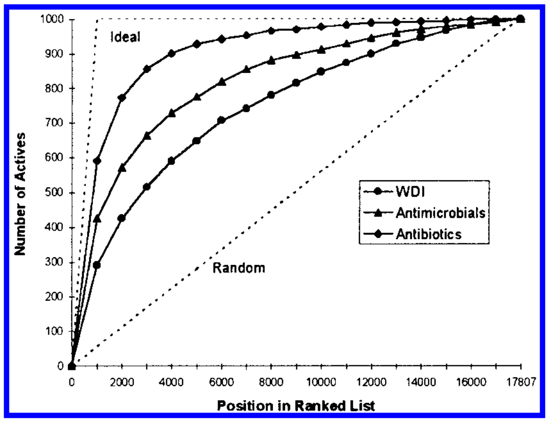


Figure 12. The ability of the GA to discriminate between the class of antibiotics and SPRESI compounds is clearly seen.

Table 12. The Discriminatory Power of the GA-Calculated Weights Increases As the Sets of Compounds Become More Closely Related

class	A@1000	initial enhancement	A50	global enhancement
WDI	290	5.2	2842	3.1
antimicrobials	425	7.6	1429	6.2
antibiotics[a]	590	10.5	758	11.7

[a] The antibiotics are a subclass of antimicrobials, which are, in turn, a subclass of WDI compounds.

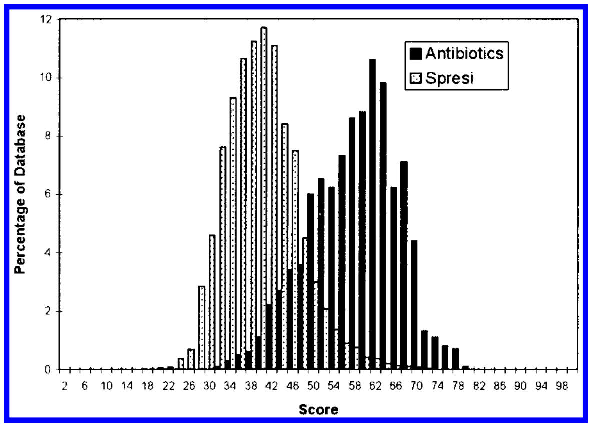


Figure 13. The scores obtained for the antibiotics (as percentages) are shown in black, with the scores obtained for the SPRESI molecules shown in grey.

classes from the 16 807 SPRESI compounds, and the results are shown in Figure 12 and Table 12. The results indicate that as the compounds within a set become more closely related, the GA weights become increasingly effective at discriminating them from inactive compounds, with 50% ofthe antibiotics occurring in just the top 4% of the list. Figure 13 shows the histogram of SPRESI molecule scores (given as a percentage of the SPRESI file) separated from the histogram of scores for the antibiotics (also given as a percentage of the total antibiotics). The effectiveness of the weights in discriminating between these two sets of molecules is clearly demonstrated.

Applying 'Drug-Like' Profiles to Unknowns. In the experiments described so far, the predictive nature of the weights has been tested by applying them to larger collections of compounds that are known to have the same characteristics as the compounds that were used to derive the weights. Weights have also been derived to distinguish subclasses of WDI from the SPRESI molecules. In both cases, there are significant increases by chance in being able to rank the molecules of interest high within the ranked lists. A further interesting test is to see how effective the weights are at distinguishing between previously unknown classes of active and inactive molecules. We tested the weights calculated for the WDI/SPRESI data on two publicly available data sets for which some activity data is known. In addition, the GA was trained for the new data. The data sets are a collection of compounds that has been tested against the DTP AIDS Antiviral Screen and has been made available by the National Cancer Institute (NCI)[30] and the Dictionary of Natural Products.[31]

Table 13. Results for the AIDS Data[a]

data	weights	A@553	initial enhancement	A50	global enhancement
CA/CI	WDI/SPRESI	41	3.4	8332	1.4
CA/SPRESI	WDI/SPRESI	105	6.2	4548	1.9
CA/CI	CA/CI	65	5.0	4496	2.6
CA/SPRESI	CA/SPRESI	122	7.2	2074	4.2

[a] The top two rows show the results of applying the previously calculated WDI/SPRESI weights to discriminate between the confirmed actives (CA) and the confirmed inactives (CI), and the confirmed actives and SPRESI molecules. The bottom two rows show the results of training the GA to discriminate between the confirmed actives (CA) and the confirmed inactives (CI), and the confirmed actives and SPRESI molecules.

AIDS Antiviral Compounds. The NCI's AIDS Antiviral Screen was developed to discover new compounds that are capable of inhibiting HIV. Tens of thousands of compounds have been tested, and the publicly available data (i.e., compounds that are not covered by a confidentiality agreement) consist of 25 058 molecules that have been placed in one of three categories: Confirmed Active (CA); Confirmed Moderately Active (CM); and Confirmed Inactive (CI). The data were preprocessed, as for the SPRESI file (i.e., only molecules containing the common elements were included); that is, where possible, the charges were neutralized and only the parent molecules were used. The CM molecules were considered as active molecules and were relabeled CA and the CI compounds were considered as inactive molecules. After preprocessing, there was a total of 553 CA molecules and 22 734 CI molecules.

The initial test was to apply the WDI/SPRESI-derived weights to attempt to distinguish the CA molecules from the CI molecules. This distinction was accomplished by scoring the molecules using the WDI/SPRESI weights and sorting them in order of descending score. Random effects suggest that we should find 13.5 of the CA molecules in the top 553 positions in the ranked list. In fact, 41 CA molecules were found, giving an initial enhancement of 3.4 (i.e., three times as many CA molecules found as would be expected by chance; see Table 13). This result is surprisingly good because although the CI molecules are inactive for this particular test, there is no reason to assume that they are 'globally inactive'; hence, many of them may be misclassified. In an attempt to address this problem, a second experiment was done where the WDI/SPRESI weights were used to attempt to discriminate between the CA molecules and a subset of the SPRESI file. The 553 CA molecules and the 16 807 SPRESI compounds used previously were pooled. In this case, we would expect to find 17.6 of the

Table 14. Results for the DNP Data[a]

data	weights	A@1000	initial enhancement	A50	global enhancement
DNP_{act}/DNP_{inact}	WDI/SPRESI	129	1.6	5015	1.3
DNP_{act}/SPRESI	WDI/SPRESI	346	6.2	1903	4.7
DNP_{act}/DNP_{inact}	DNP_{act}/DNP_{inact}	190	2.4	3716	1.7
DNP_{act}/SPRESI	DNP_{act}/SPRESI	389	6.9	1623	5.5

[a] The data labeled DNP_{act} is a sample of 1000 molecules selected at random from the 7538 known active molecules in DNP. The data labeled DNP_{inact} is a sample of 11 550 molecules selected from the 57 753 presumed inactive molecules in DNP. The top two rows show the results of using the previously calculated WDI/SPRESI weights to attempt to discriminate between the two DNP classes, and between the DNP known actives and SPRESI molecules. The bottom two rows show the results of training the GA to discriminate between the DNP known actives and the DNP presumed inactives, and between the DNP known actives and SPRESI molecules.

CA molecules in the top 553 positions of the ranked list by random effects. In fact, we find 105 of the CA molecules, which is an initial enhancement of 6.2.

Next, we ran the GA to derive weights that discriminate between the CA and the CI molecules. The weights were then used to score and rank the molecules. This time, 65 of the CA molecules occurred in the top 553 positions of the ranked list, giving an initial enhancement of 5.0. Finally, we ran the GA to derive weights that would distinguish the CA molecules from SPRESI molecules. In this case, we found 122 of the CA molecules in the top 553 positions of the ranked list, which is an initial enhancement of 7.2, and 50% of the CA molecules were found in the top 12% of the ranked list. As expected, when the GA is trained for the specific data, much better results are obtained than when the more general weights are applied and, in both cases, the GA is extremely effective in discriminating between the sets of molecules.

Dictionary of Natural Products. The Dictionary of Natural Products (DNP) is supplied by Chapman & Hall[31] and includes detailed coverage of all known natural products and related substances. The version of the database used here consists of ~80 000 compounds. The data were preprocessed as before. The active molecules in DNP were identified as those molecules that also occur in WDI and a few additional compounds that are labeled KA in the TOCN field. This identification gave a total of 7538 molecules. These molecules form a subset of WDI with the characteristic that each compound in the subset occurs in nature and has not been synthesized in a drug discovery experiment. The remaining 57 753 were presumed to be inactive molecules. The WDI/SPRESI weights were applied to discriminate the active molecules from the presumed inactive molecules. By chance, we would expect to find 870 active molecules in the top 7538 positions of the list. In fact, we find 1290, which is an initial enhancement of 1.5. In this experiment, it is assumed that all of the DNP molecules that are not labeled KA or that do not occur in WDI are biologically inactive. However, this classification is likely to be incorrect because the molecules all occur in nature, and it is possible that many of them that have not yet been identified as showing biological activity have been misclassified.

In an attempt to address this problem, the DNP molecules were clustered using the Jarvis Patrick method as implemented in the Daylight software.[20] The parameters used to cluster the molecules were fingerprints of size 2048, and a requirement for 8 out of 14 nearest neighbors in common. Clustering at this level results in 5882 clusters, with an average cluster size of 9.9, and 7758 singletons. The DNP molecules that had been classified as inactive and that score highly were analyzed to see how many of them occur in clusters that also contain known active DNP molecules. It was found that 3509 of the presumed inactive molecules that are in the top 7538 positions of the ranked list occupy clusters that also contain known active molecules. This result suggests that these presumed inactive molecules may well be of therapeutic interest.

The GA was then run to attempt to discriminate between the two classes of DNP molecules. Subsets were chosen at random from each of the sets so the GA would reach convergence in reasonable time. The reduced sets consist of 1000 active molecules and 11 550 inactive molecules (1/5th of the total presumed inactives). By chance, we would expect to find 79 active molecules to be ranked in the top 1000 positions of the ranked list. The GA ranked 190 of the active molecules in the top 1000 positions, which is an initial enhancement of 3.3. Finally, the GA was trained to discriminate between the 1000 DNP known actives and SPRESI molecules. In this case, by chance effects, we would expect to find 56 of the actives in the top 1000 positions of the ranked list and, in fact, the GA ranks 389 of the active molecules high in the ranked list, which represents an initial enhancement of 6.9. These results are summarized in Table 14.

Calculating Profiles for Proprietary Data. The final experiment involves proprietary data and was designed to determine if the discrimination between compounds achieved by the GA is in agreement with chemists' intuition. A sample of 8083 Glaxo Wellcome compounds was analyzed on paper by a panel of chemists and assigned labels of *OK* and *Not OK*: 6074 of the molecules were labeled *OK*; and the remaining 2009 molecules were labeled *Not OK*. The GA was trained with these sets of compounds using the six features described previously and, in addition, ClogP.[32] The results are shown graphically in Figure 14, where a clear separation between the two sets of molecules is evident. Visual inspection of both the low scoring molecules that are labeled *OK* and the high scoring molecules that are labeled *Not OK* indicates that at least some of these are indeed 'grey-area' molecules that may have been misclassified. This experiment demonstrates the potential of the method for ordering compounds for high-throughput screening, especially when the numbers of compounds that could potentially be screened is too large to be examined manually.

CONCLUSIONS

With the increasing importance of combinatorial approaches in drug discovery program, it is important to ensure

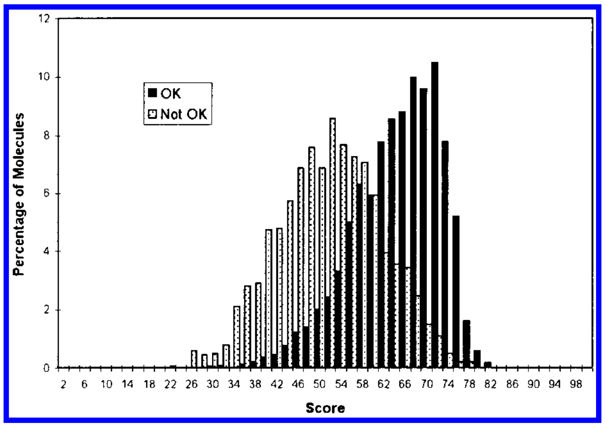


Figure 14. The histogram of scores for the Glaxo Wellcome proprietary data (as percentages). A clear separation can be seen between the 6074 molecules that were labeled *OK* and the 2009 molecules that were labeled *Not OK*.

that the libraries that are synthesized contain compounds that exhibit characteristics typical of bioactive molecules. In this paper, we have described methods for the creation of biological activity profiles that seek to rank a set of molecules in order of decreasing probability of activity on the basis of the relative occurrences of simple structural features and physicochemical properties in large databases of active and inactive compounds. Surprisingly good discrimination between active and inactive compounds is obtained by applying substructural analysis techniques to individual features, and this level of discrimination is further improved by using a GA to combine weights from more than one feature. The weights have considerable predictive power, even when applied to data sets that are noticeably different from the source data used for the calculation of the weights.

There are many ways in which the methods reported here can be developed (e.g., by using other types of high-level descriptor such as the reduced graphs that are used in the representation and searching of generic chemical structures in patents[33]). Even at their current stage of development, our methods provide a simple but effective, way of “data mining” the increasing numbers of data sets for which both structural and bioactivity information are available.

ACKNOWLEDGMENT

We thank GlaxoWellcome Research and Development for funding, Daylight Chemical Information Systems Inc. for software support, and Andrew Poirrette for advice on the GA experiments. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 407−416.

(2) (a) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Application of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251. (b) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A; Gallop, M. A. Application of combinatorial technologies to drug discovery. 2. Combinatorial organic-synthesis, library screening strategies, and future directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(3) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and applications in chemistry and biology*, ACS Professional Reference Book; American Chemical Society: Washington, D.C., 1995.

(4) Cramer, III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

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

(7) Hodes, L. Computer-aided selection of compounds for anti-tumor screening - validation of a statistical-heuristic method. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 128−132.

(8) 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.

(9) Meyer, E.; Sens, E. In *Chemical Structures: the International Language of Chemistry*; Warr, W. E., Ed.; Springer Verlag: Heidelberg, 1988; pp 235−241.

(10) Ormerod, A.; Willett, P.; Bawden, D. Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct.-Act. Relat.* **1989**, *8*, 115−129.

(11) Ormerod, A.; Willett, P.; Bawden, D. Further comparative studies of fragment weighting schemes for substructural analysis. *Quant. Struct.-Act. Relat.* **1990**, *9*, 302.

(12) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A method for automatic generation of novel chemical structures and its potential application to drug discovery. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 527−530.

(13) Sheridan, R. P.; Nachbar, R. B.; Bush, B. L. Extending the trend vector - the trend matrix and sample-based partial least-squares. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 323−340.

(14) The SPRESI database is produced by the All-Union Institute of Scientific and Technical Information of the Academy of Science of the USSR (VINITI) in Moscow, and the Central Information Processing for Chemistry (ZIC) in Berlin. This database consists of data extracted from ∼1000 journals and patents, books, and other sources from 1975 to 1990. SPRESI is distributed by Daylight Chemical Information Systems, Inc., Mission Viejo, CA.

(15) The World Drug Index (WDI) is maintained by Derwent Publications Ltd., London.

(16) Young, S. S.; Sheffield, C. F.; Farmen, M. Optimum utilization of a compound collection or chemical library for drug discovery. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 892−899.

(17) Derwent Information: 14 Great Queen Street, London.

(18) Navia, M. A.; Chaturvedi, P. R. Design principles for orally bioavailable drugs. *Drug Discovery Today* **1996**, *1*, 179−189.

(19) Kier, L. B. Indexes of molecular shape from chemical graphs. *Med. Res. Rev.* **1987**, *7*, 417−440.

(20) *Daylight Programmer’s Manual*. Daylight Chemical Information Systems: Mission Viejo, CA.

(21) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptions. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(22) Adamson, G. W.; Lambourne, D. R.; Lynch, M. F. *J. Chem. Soc., Perkin Trans. 1* **1973**, 2428.

(23) Goldberg, D. E. *Genetic Algorithms in Search, Optimisation, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

(24) Forrest, S. Genetic algorithms - principles of natural-selection applied to computation. *Science* **1993**, *261*, 872−878.

(25) Goldberg, D. E. Genetic and evolutionary algorithms come of age. *Commun. ACM* **1994**, *37*(3), 113−119.

(26) Clark, D. E.; Westhead, D. R. Evolutionary algorithms in computer-aided molecular design. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 337−358.

(27) Sheridan, R. P.; Kearsley, S. K. Using a genetic algorithm to suggest combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 310−320.

(28) Gillet, V. J. ; Willett, P.; Bradshaw, J. The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.

(29) Jones, G.; Robertson, A. M.; Willett, P. An introduction to genetic algorithms and to their use in information retrieval. *Online Rev.* **1994**, *18*, 3−13.

(30) *Developmental Therapeutics Program*, DCTDC; National Cancer Institute: Bethesda, MD.

(31) *Dictionary of Natural Products*; Chapman & Hall: London, U.K.

(32) *CLOGP3 Reference Manual*; Daylight Chemical Information Systems: Mission Viejo, CA.

(33) Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W. Computer storage and retrieval of generic chemical structures in patents. 13. Reduced graph generation. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 260−270.

CI970431+