ARTICLES

Bit-String Methods for Selective Compound Acquisition

Nicholas Rhodes* and Peter Willett

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

James B. Dunbar, Jr. and Christine Humblet

Parke-Davis Pharmaceutical Research Division,Warner Lambert Company, Ann Arbor, Michigan 48105

Received June 4, 1999

Selective compound acquisition programs need to ensure that the compounds that are chosen do not contain undesirable functionality. This is easy to achieve if a supplier is prepared to provide unambiguous structure representations for the compounds that they have available: this paper discusses selection techniques that can be used when a supplier is prepared to make available only fragment bit-string representations for the compounds in their catalog. Experiments with three databases and three types of bit-string show that a simple *k*-nearest-neighbor searching method provides a surprisingly effective, although far from perfect, way of selecting compounds when only bit-string representations are available. A second approach, based on the use of a fragment weighting scheme analogous to those used in substructural analysis studies, proved to be noticeably less effective in operation.

INTRODUCTION

The need to maximize the structural diversity of the compounds that are submitted to corporate high-throughput screening (HTS) programs has spurred interest in novel sources of compounds that can augment those available from existing in-house repositories. Examples of such sources include collaborations with academic laboratories, natural products and folk medicine, commercial compound suppliers, and intercompany exchanges. The acquisition process normally involves three major stages. First, the potential *supplier* (of whatever type) makes their *catalog* of compounds available in machine-readable form to the *recipient* company. The recipient next applies a range of filters to assess the suitability of each of the compounds that has been offered by the supplier and then finally issues a request for those compounds that pass all of the filters and that thus appear to be appropriate for screening. Several different filtering criteria may be involved:[1−3] a molecule must not be identical with, or very similar to, one that is already available locally; a molecule must have acceptable values for parameters such as molecular weight, hydrophobicity, and number of rotatable bonds; and a molecule must not contain any of the substructures on a *bad list* of undesirable moieties, i.e., molecular fragments (such as toxicophores or highly reactive functional groups) whose presence would lessen that molecule's attractiveness for further development. In this paper, we focus upon the last of these criteria.

It is trivial to implement a bad list filter by checking for the presence of each individual undesirable substructure in the connection table describing a potential acquisition. However, this is not possible when, as is increasingly the case, a supplier will provide only a fragment bit-string representation for each compound that is available for purchase. Bit-strings are very widely used in chemical database processing, e.g., for the screening stage of substructure searching, for similarity searching, and for diversity analysis, but suffer from the fact that they provide an ambiguous representation of a molecule. Specifically, a bit-string records only the presence/absence of substructural features without regard to their precise interconnections and thus describes not one but a whole family of related molecules.[4] It is this feature, of course, that commends fragment bit-strings to suppliers, as they provide a means of hiding from potential recipients the precise nature of the compounds that are available for purchase. It is hence necessary to find some way of matching the explicitly defined set of substructures in a recipient's bad list with the set of substructures that are defined (often implicitly) in a supplier's fragment bit-string. The work reported here sought to achieve this by means of a simple similarity searching procedure.[5]

Assume that a training set is available, containing compounds for which there are full structure representations. This set is searched for the presence of each of the substructures in a bad list, thus enabling the identification of each molecule as a *keeper* or as a *reject*, depending upon whether one would wish to keep or to reject that compound if it was to be offered by a supplier; in practice, of course, the training set will represent a set of compounds already available to the recipient (such as their corporate database). Fragment bit-strings are then generated for each of the compounds in this training set, using the same software package as is used by the supplier to encode the compounds they have available for purchase. Then, when a supplier's catalog is received,

* To whom all correspondence should be addressed. E-mail: n.rhodes@sheffield.ac.uk.

10.1021/ci990428l CCC: $19.00 © 2000 American Chemical Society
Published on Web 01/28/2000

Table 1. Numbers of Test Molecules Having Some Specific Number (*n*) of Keepers in a Set of Nine NNs for Similarity Searching of 1000 Test Molecules from the WDI Database

		obsd	
n	predicted	K	R
0	0	0	83
1	0	1	11
2	0	5	8
3	1	4	8
4	9	8	7
5	42	12	6
6	136	16	12
7	284	18	9
8	344	81	13
9	184	686	12
totals	1000	831	169

the recipient carries out a similarity search for the fragment bit-string describing each compound in the catalog against the fragment bit-strings for the sets of keepers and rejects in the previously encoded training set. A decision as to whether to acquire a particular member of the catalog is then taken on the basis of the resulting sets of similarities; for example, it seems reasonable to select the molecule corresponding to a particular fragment bit-string from the catalog if that fragment bit-string is very similar to the fragment bit-strings for known keepers and very dissimilar to the fragment bit-strings for known rejects. This simple idea forms the basis for the experiments described in the next section.

EXPERIMENTAL DETAILS AND RESULTS

The training set for our initial experiments contained 41 643 molecules from the *World Drugs Index* database,[6] encoded as SMILES strings, while the bad list contained 125 substructural features that have been used at Parke-Davis for several years for filtering suppliers' catalogues when full structure representations are available. These bad list features were encoded as SMARTS definitions and then searched for in the WDI file, giving totals of 34 512 keepers and 7130 rejects to act as the training data: these sub-files are subsequently referred to as K and R, respectively. Daylight[6] fingerprints were then generated for all of the structures in these two subfiles and subsets created to act as test data for the similarity searching experiments; specifically, we used the default 2048-member, 7-step, bond path strings for most of the experiments, with some (as described further below) using other parametrizations. Each of these test compounds was searched against the WDI file to identify its k ($1 \leq k \leq 100$) nearest neighbors (or NNs), with the similarities being calculated using Daylight Tanimoto coefficient routines. The resulting NN lists were then used to simulate the prediction of the keeper/reject nature of the test compounds.

Since there are 34 512 compounds in K and 7130 in R, the probability of a randomly selected SMILES being in K is 0.828 and in R is 0.172. Use of these values in the standard binomial formula permits the calculation of the number of test molecules that have some specific number, *n*, of keepers in their sets of $k - 1$ NNs (*NB* $k - 1$ as the first NN for a molecule is itself). These calculated numbers can be compared with those observed in practice, as illustrated in Table 1, where the two columns on the right give the observed numbers of compounds with each value of *n*. For example, there are 15 test molecules for which 4 of the top-10 NNs are keepers: of these 15 molecules, 8 of them actually are keepers with the remainder being rejects. There are clearly very large differences between the predicted and observed values, and these differences form the basis of our selection program. The program uses the number of keepers in the k NNs to predict the nature of a test molecule, e.g., all those with ≤ 1 keepers in their top k similarities are predicted to be rejects, or all those with ≤ 10, etc. To illustrate the *modus operandi*, consider the case when $k = 5$, i.e., when the five most similar molecules in WDI are recorded. The most similar (identity) is ignored, leaving it possible for the query molecule to have 0, 1, 2, 3, or 4 NNs in K. Typical results are shown in Table 2, for a file of 1000 test molecules, of which 831 are actually members of K (and thus 169 of R). The table shows that there are 101 members of R in the test set that have no members of K in their top-4 NNs, 116 that have none or one member of K in their top-4 NNs, etc. Thus, if we use, e.g., the presence of ≤ 2 members of K in the top-4 NNs as a way of predicting rejects, then we would correctly predict 135 of the test-set molecules as belonging to R and incorrectly predict 24 of the test molecules that were actually members of K as being members of R. The remaining 841 test molecules that do not have ≤ 2 members of K in their top-4 NNs are then predicted as belonging to K: of these, 807 are actually members of this class while 34 rejects are predicted incorrectly.

Table 2. Prediction of Keeper/Reject Nature Using the Top Five NNs for Each of 1000 Test Molecules (Including Itself)

no. of	prediction of keepers		prediction of rejects		total
top-4 in K	correct	incorrect	correct	incorrect	incorrect
0	829	2	101	68	70
≤1	827	4	116	53	57
≤2	807	24	135	34	58
≤3	762	69	148	21	90
≤4	0	831	169	0	831

Table 3. Minimum Number of Incorrect Predictions using Different Numbers of *k* NNs

$k - 1$	min incorrect	no. of keepers used for min prediction	incorrect K prediction	incorrect R prediction
1	52	≤0	16	36
2	57	≤0	7	57
3	52	≤1	12	40
4	57	≤1	4	53
9	69	≤3	10	59
24	81	≤13 or ≤14	23 or 27	58 or 54
49	83	≤29	25	58
99	94	≤73	46	48

Table 3 summarizes a number of such runs on the same set of 1000 test molecules. These tests sought to predict reject compounds on the basis of 1, 2, 3, ... $k - 1$ of the NNs being in K (in just the same way as was done in Table 2) for $k \leq 100$ NNs. For example, when $k = 10$ (i.e., we consider just the top-9 NNs), then the third column of the table shows the minimal number of incorrect predictions is obtained when a test molecule is assigned to R if it has at most three of its top-9 NNs in K: in this case, 10 keepers and 59 rejects are predicted incorrectly.

Results similar to those in Table 3 were obtained in experiments involving four other files of 1000 test molecules, and also in an extended series of experiments that varied

both the length of the fingerprint (512−8192 bits) and the sizes of the fragments (3−7 atoms *per* fragment) that are encoded in a fingerprint. None of these variations had any appreciable effect on predictive performance; for example, using five-atom fragments with $k = 5$ and with fingerprints containing between 512 and 8192 bits, the minimum number of incorrect predictions only varied between 59 and 61, i.e., an error rate of just 5.9−6.1%. No improvement was observed when the prediction routines were modified to take account of the magnitude of a NN's similarity coefficient with a test molecule, rather than considering just the keeper/reject nature of that NN.

All of the experiments thus far have employed Daylight fingerprints, which use a superimposed coding procedure in which each algorithmically defined fragment in a molecule results in the setting of several bits in the 512−8192-member fragment bit-string. Additional sets of runs were carried out using Barnard Chemical Information (BCI) fingerprints[7] and UNITY fingerprints.[8] The BCI fingerprints are based on a fragment dictionary that encodes 1119 predefined substructural types and that allocates just a single bit to each such fragment, thus resulting in a less ambiguous bit-string representation than the superimposed coding techniques used for Daylight fingerprints. The UNITY encoding method is intermediate between the two extremes exemplified by the Daylight and BCI approaches, as the 988-bit-string is based on both algorithmically-based and dictionary-based encoding. These additional experiments also involved two further databases: the *Available Chemicals Directory* (ACD),[9] containing 184 498 keepers and 28 766 rejects; and the file resulting from merging the AIDS and cancer databases distributed by the National Cancer Institute,[10] this containing a total of 30 284 keepers and 5983 rejects.

Five sets of queries, each containing ca. 1000 randomly selected query molecules, were searched against each database, using each of the three types of bit-string representation and using values of $k - 1$ in the range 1−30. The results of these experiments are shown in Figure 1, which details the minimum numbers of errors obtained when averaged over the five sets of query molecules in each case. The results are fairly consistent, the error rates being smallest when just a very few NNs are used for the prediction and when the molecules are represented by UNITY bit-strings. It will be seen that the ACD results for the Daylight and Unity bit-strings are noticeably better (i.e., exhibit lower minimum error rates) than for the other two databases. This we ascribe to the more heterogeneous nature of the ACD file, which has a cosine centroid diversity index[11] of 0.724 compared to 0.644 and 0.631 for the NCI and WDI databases, respectively. However, the constitution of the database is not the only factor affecting performance, as is demonstrated by the observation that the BCI fingerprints give comparable results for all three databases. This we attribute to the fact that we are using the default, dataset-independent BCI fingerprints for all three files, whereas it is possible to select an optimized fragment dictionary using statistical criteria that results in a coding mechanism analogous to that underlying the UNITY fingerprints.

Bit-strings have many limitations when used for similarity searching applications,[12] but the results shown in Figure 1 demonstrate that we can calculate the keeper/reject nature of a test molecule with an error rate of less than about 5%

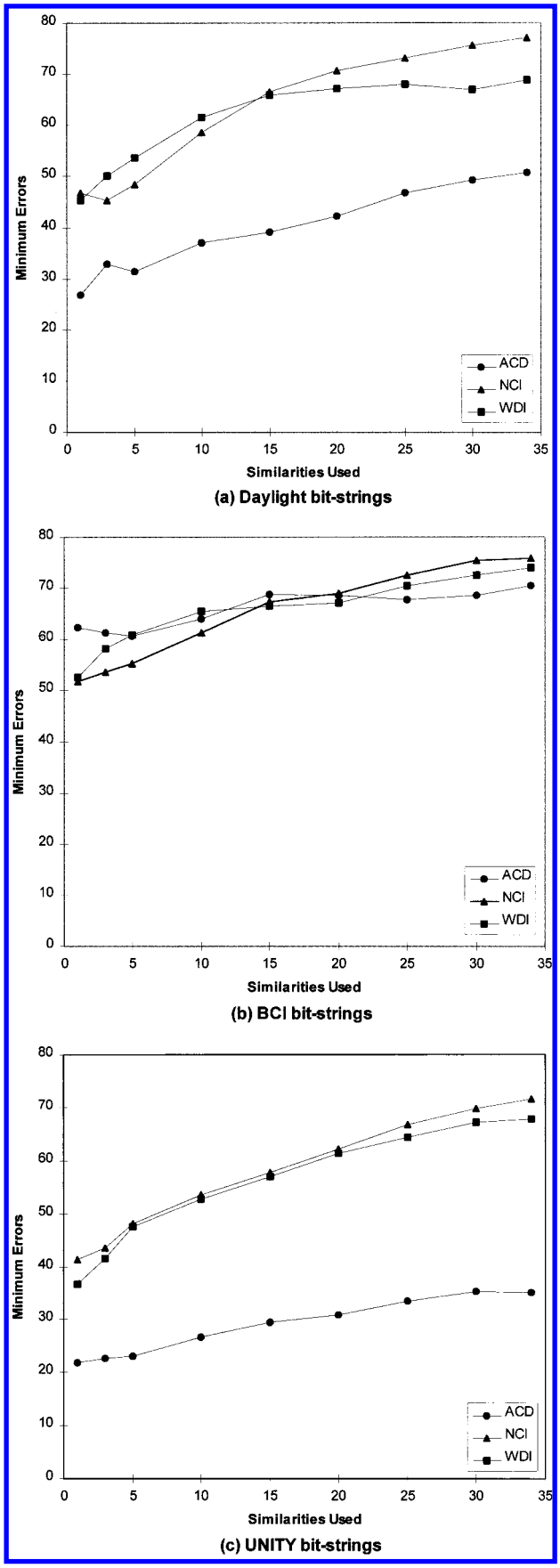

Figure 1. Minimum numbers of errors, averaged over five sets of ca. 1000 query molecules in each case, for searches of the WDI, NCI, and ACD files using bit-strings generated using (a) Daylight, (b) BCI, and (c) UNITY software.

merely by considering a small number of the molecule's NNs. It seems unlikely that these failure rates can be improved substantially using the simple approach adopted here, given the observed patterns of keepers and rejects in sets of NNs. Thus, we have observed reject compounds where all the top-25 NNs are keepers, reflecting the fact that the presence of just a single undesirable functional group on an otherwise eligible candidate molecule is sufficient to preclude it. The converse situation, i.e. a keeper molecule with most of its NNs being rejects, is far scarcer, although one was observed with 22 of the top-25 NNs being rejects (something with an a priori binomial probability of about 2.6×10^{-15}).

We also carried out some preliminary studies of an alternative prediction program that was based on the assumption that some positions in a bit-string would be better able than others to discriminate between keepers and rejects, in much the same way as substructural analysis is based on the idea that different fragments discriminate to different extents between active and inactive molecules.[13] The approach that was developed here counted the numbers of times that each bit was switched on for the two sets of compounds K and R and then used this information to calculate weights analogous to those used for substructural analysis.[14] In this way, it was possible to rank the bit positions in decreasing order of their ability to discriminate between keepers and rejects in the training set. The prediction program took the n top-ranked bit positions (for $n \leq 100$) and then assigned a test molecule to K or R depending on the number of times that these n bits were set. Initial experiments with the ACD file showed that none of the weights gave results that were anywhere near those obtained with the simpler, similarity-searching procedure, and we thus did not consider this second approach any further.

The ACD, NCI, and WDI files do not contain the very large numbers of close analogues that characterize corporate structure databases, and it was thus hardly surprising that the program gave higher error rates when it was applied to Parke-Davis' corporate pharmaceutical and agrochemical files. As a result of these trials the final version of the program at Parke-Davis does not attempt to predict the keepers and rejects correctly, but just those rejects that can be removed from further consideration with high confidence: for example, inspection of the WDI data in Table 1 shows that 50% of the actual rejects can be rejected on the basis of no keepers in the top-9 similarities, while still correctly retaining all of the keepers. Thus the implemented version predicts rejects in order of decreasing confidence so that the user may determine the desired cutoff point as appropriate.

In blind compound acquisition strategies, the initial goal is to determine rapidly if an appropriate number of the library of compounds appears to be sufficiently different from those already on hand and, if possible, suitable for acquisition. This typically establishes the set of compounds for which actual structures would be available at the next level of the acquisition process. Within this initial goal is included the generation of the ordered list of these compounds based on the degree of confidence that a compound is different from those in-house and different from compounds deemed unsuitable. At this point, the priority order cannot be exact, but rather functioning as an enrichment process. So if a compound is on the order of the Tanimoto coefficient equal to 0.70 or less similar to any in-house compounds and is not a close neighbor of any in the undesirable list, it is deemed to be more desirable than one which does have a close neighbor in the undesirable list. As a practical matter, selecting a few close analogues of a very different compound from our in-house collection is considered reasonable and prudent. If this class of compounds were found in a screen, and if those few close analogues also were active, immediate confidence in the hit is established. Singleton hits would require more in terms of retest and synthesis to establish this same level of confidence.

CONCLUSIONS

Selective compound acquisition is widely used to increase the structural diversity of the compounds that are submitted to HTS programs. Acquisition is simple when compound suppliers can provide full structure representations for the compounds that they have available for purchase or exchange: in this paper, we have discussed methods that can be used when only fragment bit-string representations are available. Specifically, experiments with three databases and three types of bit-string demonstrated the general effectiveness of a simple *k*-NN searching procedure for filtering compounds that contain members of a user-defined bad list of undesirable substructural features. Bit-string based procedures have also been described for filtering other types of undesirable compound: thus, Patterson et al.[15] have demonstrated that a threshold Tanimoto coefficient of 0.85 is sufficient to screen structurally similar molecules that might be expected to exhibit comparable activity profiles and that would merely duplicate compound types already available to an organization; and Brown and Martin[16] have shown that bit-strings can be used to predict a range of molecular properties reasonably accurately, thus enabling the filtering of compounds with unacceptable physicochemical characteristics. Combining these procedures with that described in the present paper, it seems that it should be possible to provide a reasonably effective selective compound acquisition program even if compound suppliers are only prepared to make their catalogues available in bit-string form.

ACKNOWLEDGMENT

We thank Parke-Davis for funding, Mark Duffield for providing the software for generating the sets of keepers and rejects, Simon Tyrrell for assistance with the calculation of diversity indices, and John Barnard, Geoff Downs, Val Gillet, and Gareth Wilden for advice on the use of BCI, Daylight, and Tripos fingerprints.

REFERENCES AND NOTES

(1) Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 65−84.

(2) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening−an overview. *Drug Discovery Today* **1998**, *3*, 160−178.

(3) Lipinski, C. A.; Lombard, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(4) Barnard, J. M. Structure representation and searching. In *Chemical Structure Systems*; Ash, J. E., Warr, W. A., Willett, P., Eds.; Ellis Horwood: Chichester, U.K., 1991; pp 9−56.

(5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996.
(6) The *World Drugs Index* database is distributed by Daylight Chemical Information Systems Inc., URL http://www.daylight.com/.
(7) Barnard Chemical Information Ltd., URL http://www.bci1.demon.co.uk/.
(8) The UNITY chemical information management system is distributed by Tripos Inc., URL http://www.tripos.com/.
(9) The *Available Chemical Directory* is distributed by MDL Information Systems Inc., URL http://www.mdli.com/.
(10) The NCI AIDS and cancer databases are distributed by the Developmental Therapeutics Program of the National Cancer Institute, URL http://dtp.nci.nih.gov/.
(11) Turner, D. B.; Tyrrell S. M.; Willett, P. Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 18−22.
(12) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 379−386.
(13) 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.
(14) Ormerod, A.; Willett, P.; Bawden, D. Comparison of fragment weighting schemes for substructural analysis. *QSAR* **1989**, *8*, 115−129.
(15) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighbourhood behaviour: A useful concept for validation of molecular diversity descriptors. *J. Med. Chem*. **1996**, *39*, 3049−3059.
(16) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1−9.

CI990428L