CLIP: Similarity Searching of 3D Databases Using Clique Detection[†]

Nicholas Rhodes and Peter Willett*

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield,
Western Bank, Sheffield S10 2TN, United Kingdom

Alain Calvet, James B. Dunbar, and Christine Humblet

Pfizer Global Research and Development, 2800 Plymouth Avenue, Ann Arbor, Michigan 48105

Received September 17, 2002

This paper describes a program for 3D similarity searching, called CLIP (for Candidate Ligand Identification Program), that uses the Bron-Kerbosch clique detection algorithm to find those structures in a file that have large structures in common with a target structure. Structures are characterized by the geometric arrangement of pharmacophore points and the similarity between two structures calculated using modifications of the Simpson and Tanimoto association coefficients. This modification takes into account the fact that a distance tolerance is required to ensure that pairs of interatomic distances can be regarded as equivalent during the clique-construction stage of the matching algorithm. Experiments with HIV assay data demonstrate the effectiveness and the efficiency of this approach to virtual screening.

INTRODUCTION

Virtual screening is being increasingly used to prioritize compounds for biological testing in agrochemical and pharmaceutical lead-discovery programs (see, e.g., the recent study by Gruneberg et al. on inhibitors of carbonic anhydase[1] and references contained therein). Many different methods are available for virtual screening,[2−5] including similarity searching, pharmacophore mapping, machine learning, and ligand docking. It is normal to use a cascade of methods, with rapid methods (such as "rule- of-five" structural filters or similarity searching) being used to identify some small fraction of a data set that can be processed using a more sophisticated screening method (such as flexible pharmacophore searching or ligand docking). Here, we focus on similarity searching.[6] This involves taking a bioactive target structure and then scanning a data set to find other molecules that are structurally similar, and that may hence be expected also to be active in the bioassay of interest. Early approaches to similarity searching were based on 2D fragment substructures, but the last few years have seen an interest in similarity measures that use 3D structural information, typically interatomic distances (see, e.g., refs 7−18).

This brief note describes a program, called CLIP (for Candidate Ligand Identification Program), that has been designed for the virtual screening of large files of 3D structures and that draws upon some of the work referenced above. The principal characteristics of the program are as follows. Molecules are characterized by their sets of constituent pharmacophore points, this representation being chosen on the basis of the very successful results that have been obtained in previous studies of the use of 3-point and 4-point pharmacophores for similarity searching. The similarity between two sets of pharmacophore points is calculated from a graph-theoretic fitting algorithm that identifies the maximum common substructure (MCS) in 3D. The size of the MCS (in terms of the number of matching pharmacophore points) is the principal component of a similarity measure that also involves the distance tolerances used to ensure that pairs of interpoint distances can be regarded as equivalent during the generation of the MCS. CLIP's components are described in the next section, followed by a discussion of its application to a file of structures that had been tested in an HIV-1 protease screen.

PROGRAM COMPONENTS

Search File Creation. CLIP takes as input modified MOL2 files where each pharmacophore point has been typed as donor, acceptor, donor−acceptor, electronegative, electropositive, hydrophobic, aromatic, or other. The searches discussed here used just donor/acceptor information, with these points being defined using the simple scheme described by Pepperrell et al.,[19] who consider nitrogen and oxygen atoms and use five atom-types: donors, acceptors, donor acceptors, electronegative, and a "catch-all" (those atoms falling into none of the preceding categories). Information concerning the nature and positions of the atoms so mapped as well as their relative disposition in 3D space is tagged and written to a modified MOL2 file, and the resulting MOL2 files are then input to a structure generation program, for which we used CONCORD.[20] Each of the resulting sets of coordinates is analyzed to calculate all of the interpoint distances, and the calculated sets of distance matrices then form the search file against which the target structure is matched. There are three sources of target structure. In the absence of any protein structural information, the target structure is a known bioactive molecule of some sort, such as an initial hit from a high-throughput screening (HTS) experiment or a competitor's compound. However, CLIP can also use a bound ligand as the target structure or use a target

[†] This work was first presented at the Sixth International Chemical Structures conference, Noordwijkerhout, June 2002.

* Corresponding author e-mail: p.willett@Sheffield.ac.uk.

10.1021/ci025605o CCC: $25.00 © 2003 American Chemical Society
Published on Web 02/28/2003

that is derived from the geometry of the binding site. In the latter case, which we refer to as *complementary* searching, one is looking for database structures that are compatible with the points in the binding site that interact with the pharmacophore points in the ligand, a rules file being used to specify which pairs of point types are compatible and can match. Four levels of matching are available: identity, e.g., donors match with donors; simple, e.g., donors match with donors or donor−acceptors; complementary, e.g., donors match with acceptors or donor−acceptors; and fuzzy, where the matching criteria are user-defined. The searches discussed here involved simple matching with known ligands or HTS hits as the target structures and with these structures characterized just by their donor and acceptors.

Matching Algorithm. Fragment-based measures of molecular similarity quantify the degree of resemblance between two structures by some function of the number of fragments (whether 2D or 3D) that they have in common. While efficient in operation, this involves losing the relationships (topological in 2D or geometrical in 3D) between the matching features, meaning that it is not normally possible to superpose the common features in the two molecules that are being compared. This problem can be avoided if an appropriate alignment procedure[21] is used in the matching algorithm, and we have adopted an MCS routine based on the Bron-Kerbosch clique-detection algorithm for CLIP.

The Bron-Kerbosch algorithm operates by means of a backtracking tree search. At each level, D, of the tree search, there are two sets, N(D) and C(D), of nodes of the graph which are connected to every node in the set M(D), which consists of the D nodes under consideration for inclusion in the next clique. N(D) contains the nodes which have already been tried in the attempt to enlarge M(D), and C(D) contains those candidate nodes which have yet to be tried. The algorithm moves to the next level of the tree search by moving a candidate node from C(D) to the trial set M(D), which then becomes M(D+1). The sets N(D+1) and C(D+1) are then calculated by removing from C(D) and N(D) those nodes not connected to the candidate node. When backtracking occurs, the node most recently added to M(D+1) is added to N(D) and removed from C(D), and the level of the search becomes D, from its previous value of D+1. A clique is found when both C(D) and N(D) are empty; if only C(D) is empty, then M(D) is a subset of a clique which has already been output. The selection of a candidate node from C(D) is done so as to increase the likelihood of a point in N(D) being connected to all points in C(D). When this happens, further extensions to M(D) from C(D) cannot remove this point from N(D); therefore, N(D) can never become empty by extending M(D), and so backtracking needs to occur. This can be done by choosing that point in N(D) which is connected to the most elements of C(D), and then every time a candidate is taken, selecting a point in C(D) which is not connected with the chosen point. The workings of the algorithm are illustrated by Brint and Willett,[22] and a comparison with several other algorithms is reported by Gardiner et al.[23]

Given the interpoint distance matrices for the target structure and for a database structure, the Bron-Kerbosch routine identifies the largest set of points that is geometrically equivalent, subject to a user-defined tolerance on the distance differences that are acceptable for a match: this tolerance is normally in the range 0.5−1.5 Å. The size of the MCS, in terms of the number of matched points, is then taken to be the similarity between the two matrices, and hence between the target structure and that database structure. The use of MCS size as a similarity measure is obviously biased toward larger molecules, since the larger the molecule, the larger the number of constituent points and the greater the potential for overlap with the target structure. It was hence hardly surprising that early versions of CLIP yielded search outputs in which the larger structures were clustered toward the top of the ranking, with a consequential detrimental effect on the effectiveness of retrieval. This problem was addressed as described below.

Similarity Coefficient. The simplest measure of similarity that is based on an MCS is the size (i.e., the number of nodes) of the common subgraph. This, however, has an inherent bias toward larger database structures, and it is thus common for similarity coefficients to include some form of normalization. In CLIP, we have used a modification of an association coefficient that has been used in information retrieval and that is known as Simpson's coefficient.[24] Let a be the number of points in the MCS, and let b and c be the numbers of points in the target structure and the database structure; then Simpson's coefficient is defined to be

$$\frac{a}{\min(b, c)}$$

For comparison, we also carried out experiments with the graph coefficient of Wallis et al.[25] This is the graph-based equivalent of the well-known Tanimoto coefficient,[6] which has been extensively used in the chemoinformatics applications and which has the form

$$\frac{a}{b + c - a}$$

For two points to be included in the MCS (and thus to contribute to the value of a in the formulae above), their separations must be the same, to within the user tolerance, in the two interpoint distance matrices that are being compared. It seems reasonable that two distances that are very similar should contribute more to the overall degree of resemblance than two distances that only just meet the tolerance value; moreover, if we allow some degree of partial matching in this way we do not need to use a very tight distance tolerance during clique-detection. The approach adopted here is as follows (an analogous approach to the modification of raw MCS sizes has been described very recently by Stahl et al. in their work on proximity scores[26]). Let *DTOL* be the distance tolerance that is acceptable for two distances to match. Then, for an MCS of size a points, there is a total of $a(a-1)$ interpoint distances and hence the maximum possible sum of distance differences, *MAXDIFF*, is given by

$$MAXDIFF = a(a-1) \times DTOL$$

We can then calculate the actual sum of distance differences

$$\sum(d_i d_j - t_k t_l)^2$$

where the difference is calculated for each matched distance

(d_id_j in the database structure and t_kt_l in the target structure) in the MCS. We hence obtain the root-mean-squared deviation, *RMSD*, and the correction factor that is applied to the Simpson coefficient value is then of the form

$$1 - \frac{RMSD}{MAXDIFF}$$

If all of the distances comprising the MCS have been matched perfectly, then *RMSD* is zero and the correction factor has the value of unity; if all of the distances have only just satisfied the tolerance, then the correction factor has the lower-bound value of

$$1 - \frac{1}{\sqrt{a(a-1)}}$$

It is, of course, possible to think of many other ways in which the distance differences could be used for modifying a similarity coefficient, but this approach seemed to work well enough inpractice. The correction factor can either be added to the Simpson or Tanimoto values or multiplied by them.

Our experiments showed that the basic coefficients give a small increase in performance over the use of the raw MCS size and that the corrected coefficients give a noticeable improvement, with little difference between the two types of correction factor. For example, in one of the searches with the bound ligands (as discussed further below) the three top-ranked actives were at rank positions 98, 142, and 185 when the simple Simpson coefficient was used. However, when the multiplicative correction was used in the coefficient, these three structures occurred at positions 2, 5, and 9. The results reported below all involve the use of the multiplicative correction factor and with ±1.5Å as the limiting default for two distances to match.

EVALUATION OF SEARCH PERFORMANCE

CLIP was tested using a file of 4981 molecules (39 actives) that had been screened in an HIV-1 protease assay, together with three known ligands. The CONCORD structures and distance matrices were generated as described previously, with the molecules being represented just by their constituent donor and acceptor points.

The known ligands were used as the initial targets, in their bound conformations. The results were poor, retrieving only a very few of the actives. However, when the ligands were searched in their CONCORD conformations, the results shown in Table 1 were obtained. The figures in this table are enrichment factors. Assume that a search retrieves n_A active molecules in the top-*n* rank positions and that there is a total of N_A actives molecules in the entire, *N*-molecule data set. Then the enrichment factor at rank *n*, E_n, is given by

$$\frac{n_A/n}{N_A/N}$$

i.e., the extent to which the search output exceeds the output from a purely random selection. It will be seen that the results in Table 1 show a fair level of enrichment using both similarity coefficients, at least for the first two ligands. There is no obvious difference between the performance of the two types of similarity coefficient (Simpson and Tanimoto). We then used each of the hits from the assay in turn as the target structure, with the results shown in Table 2. A cluster analysis of the 39 known actives showed that they comprised a single large cluster of 36 molecules and three singleton clusters, and we hence provide four sets of results: first the numbers of hits for the three singletons and then an averaged number of hits for the 36-member cluster. The first two singleton-based searches are very poor, but the third is able to identify at least some of the other actives in the search file, and the searches for the members of the large cluster are very successful. The search results can be illustrated diagrammatically by a cumulative recall plot, which shows the increase in the total numbers of actives retrieved as the size of the output is increased, i.e., the variation of n_A with *n*. Examples of these plots, for ligand-1 in Table 1 and for a typical molecule in cluster-4 of Table 2, are shown in Figures 1 and 2, respectively.

Table 1. Enrichment Factors for Searches Using Known Ligands as the Target Structures

	top-*n* nearest neighbors					
	Simpson's (normalized)			Tanimoto (normalized)		
target structure	*n* = 50	*n* = 100	*n* = 250	*n* = 50	*n* = 100	*n* = 250
ligand-1	20.5	23.1	18.5	48.7	35.9	16.9
ligand-2	46.2	24.4	13.8	30.8	14.0	15.9
ligand-3	0	0	5.6	0	1.3	0.5

Table 2. Enrichment Factors for Searches Using HTS Hits as Target Structures

	top-*n* nearest neighbors					
	Simpson's (normalized)			Tanimoto (normalized)		
cluster (size)	*n* = 50	*n* = 100	*n* = 250	*n* = 50	*n* = 100	*n* = 250
1 (1)	2.6	1.3	2.1	5.1	2.6	1.0
2 (1)	5.1	2.6	1.1	10.3	5.1	2.6
3 (1)	2.6	1.3	0.5	2.6	1.3	0.5
4 (36)	53.6	34.1	18.2	54.6	33.3	17.1

There are many different types of computational tool that can be used for virtual screening, and we have hence compared CLIP with three other tools: the UNITY flexible 3D substructure searching system, the UNITY 2D fingerprint similarity search, and the GOLD docking program.[20,27]

A substructure search retrieves those molecules that satisfy the structural constraints specified in the query, rather than a ranking, and there was hence a need to ensure that an appropriate comparison was being carried out. The approach taken was to reduce the UNITY distance constraint to a figure that gave an output of approximately the same size as the standard ranking thresholds; for example, taking 250 hits as the ideal was found to correspond to a very tight tolerance of 0.07 Å in a search for the first bound ligand. If this was not done, then very large outputs were obtained, e.g., one of the bound ligands with a tolerance of ±1.5 Å retrieved some 77% of the entire search file. While this will ensure very high recall, it will inevitably mean that very little useful filtering would take place in a practical virtual-screening application. A further complication is that one may need to carry out several searches to ensure that sufficient hits are obtained to populate the hit-list and to ensure that the full range of possible matches has been considered. For example,

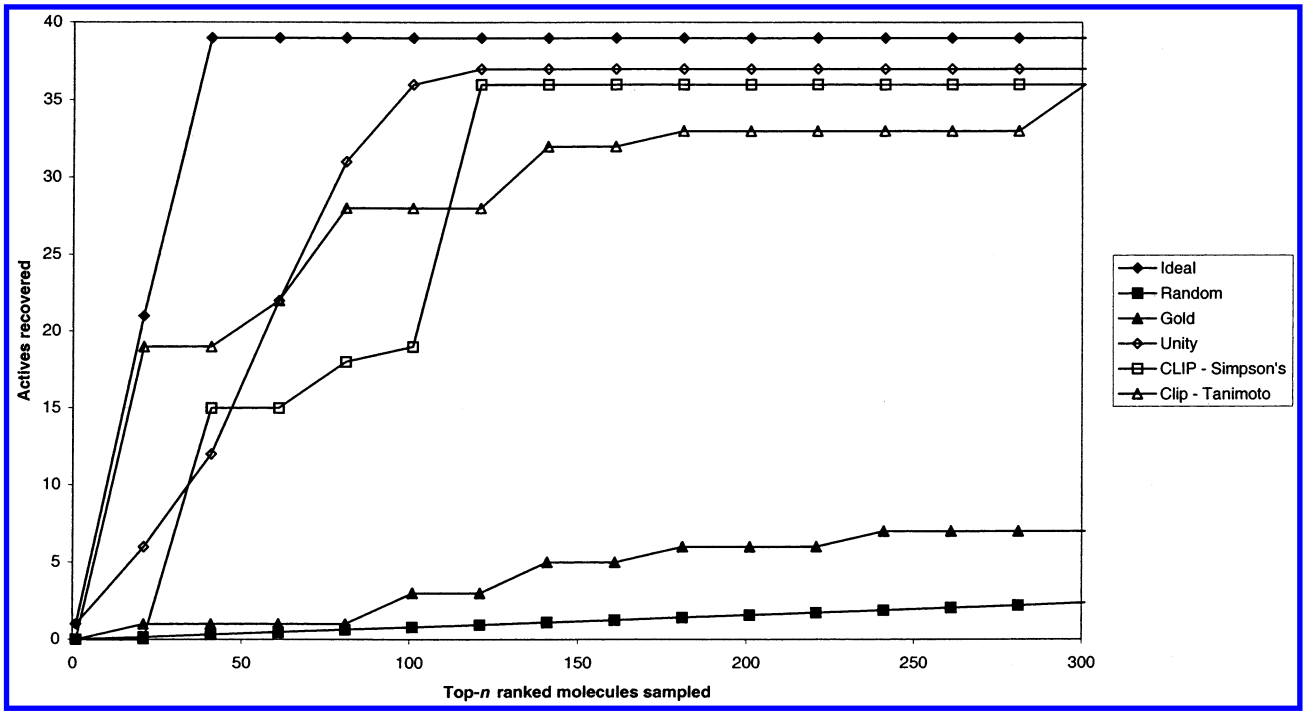

Figure 1. Cumulative recall plot for Ligand-1 in Table 1.

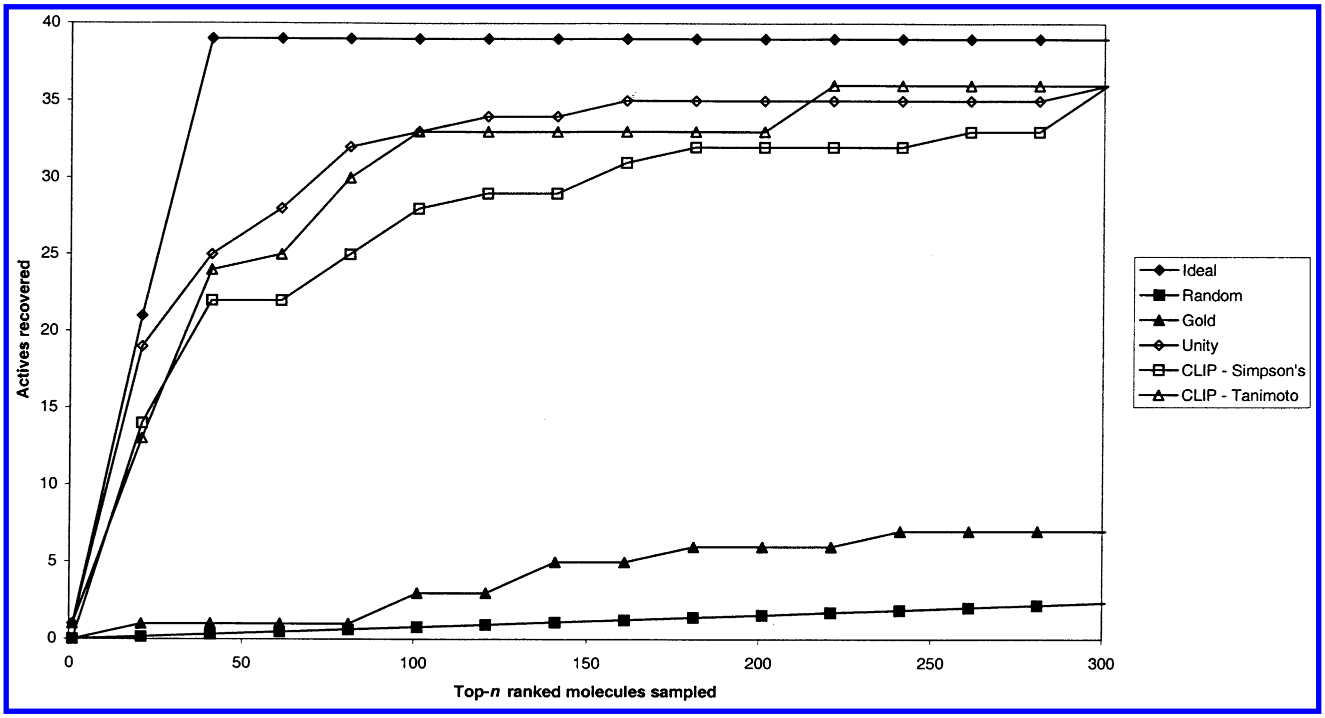

Figure 2. Cumulative recall plot for a typical ligand from Cluster-4 in Table 2.

consideration of all of the 3-point cliques that a 6-point target structure can be involved in implies a need to search for $_3C_6$ (i.e., 20) different substructural patterns. This can be time-consuming if there is a need to generate a ranking in the case of a large target structure, whereas the inherent best-match nature of the similarity search means that just a single pattern, the entire target structure, needs to be considered. Taking these factors into account and using a tolerance of ±0.07 Å to give a hit-list of 242 compounds, the UNITY search retrieved the same set of 36 actives as did the top-250 CLIP search. This is, perhaps, not unexpected given that subgraph isomorphism can be considered as a limiting case of maximum common subgraph isomorphism but does further emphasize the very distinct structural natures of the three singleton actives in this data set.

The other two screening tools, UNITY 2D similarity searching and GOLD flexible docking, both produce a score for each of the database structures, and these scores can hence be used to generate a ranking analogous to that produced by CLIP. The UNITY 2D similarity search was carried out using one of the active hits from the large cluster and one of the bound ligands. The top 200 structures were retrieved and found to contain 35 and 37 actives, respectively; for comparison the CLIP searches at this threshold retrieved 32 and 36 actives, respectively. The GOLD searches used the Protein Data Bank structure 1HVI and were run with the default system settings. This search retrieved just six of the actives in the top 200 positions; that said, it should be emphasized that the scoring function in GOLD is designed to explore the possible binding modes of a single flexible ligand, rather than to rank an entire database. The cumulative recall plots for these two search methods are included in Figures 1 and 2.

Thus far, we have considered only the effectiveness of searching. With regard to efficiency, GOLD took between 1.25 and 3.77 structures per hour (running on a single Silicon Graphics R10000 machine), depending on processor speed and machine load which corresponds to a total run time of ca. 100 CPU days for the data set used here. The CLIP system is written largely in C++, with some additional routines written in the scripting language Python. For a fixed search file, the search times are dependent on the size of the target structure, with the observed scan rates being about 250K structures an hour for a 3-node target and about 150K structures an hour for a 6-node target; for the data set in question CLIP took just under two minutes for the 6-node structure and 72 s for the smaller one. These rates are slower than the UNITY similarity searches but are still well within the design criterion of being able to process 1M structures in an overnight run.

Given the importance of molecular geometry in bioactivity, it would seem reasonable to expect that 3D descriptors would perform better than 2D descriptors in virtual screening, database clustering, and compound selection procedures. However, in a much cited paper, Brown and Martin[28] carried out an extended comparison of several types of 2D and of single-conformation 3D fragments and demonstrated the general superiority of the 2D fragments. While the UNITY 2D fingerprint search here was indeed superior to the CLIP search, the difference was quite small, suggesting that alignment-based 3D similarity measures may be more competitive than simple 3D fragment descriptors. Experiments with 3D fragments generated from multiple conformations have been more effective,[15,17] and it would hence be of interest to use CLIP in a multiconformer search. This would obviously reduce the speed of scanning, but there are ways available to ameliorate the reduction in search efficiency. Thus, the Carraghan-Pardalos could be employed as an initial screen prior to application of the Bron-Kerbosch algorithm (as described by Gardiner et al.[23]), it would be possible to build an inverted index to the interpoint distances in the database structures, or one could use the flexible 3D version of the RASCAL clique- detection algorithm.[29]

CONCLUSIONS

Similarity searching was originally developed to provide a browsing capability that would complement the traditional, substructure searching approach to accessing databases of 2D chemical structures. The increasing importance of virtual screening in lead-discovery programs has led to a reawakening of interest in similarity searching, focusing particularly on 3D measures of structural resemblance. In this paper, we have described the application of one such measure based on the maximum substructure common to a database structure and the target structure. MCS-based approaches to 3D similarity searching are by no means new[7,30,31]—indeed, they were discussed in the first review of distance-based measures for 3D similarity searching[8]—but there is little quantitative data available on their performance. The CLIP program described here provides an example of the use of this approach to virtual screening: it has been shown to be both efficient and effective in operation when applied to data typical of that being generated on an increasing scale in the search for novel bioactive molecules.

ACKNOWLEDGMENT

We thank the following: Pfizer Global Research and Development for funding; Peter Artymiuk and Eleanor Gardiner for helpful advice; the referees for their comments on an earlier version of this manuscript; and the Royal Society, Tripos Inc., and the Wolfson Foundation for hardware, software, and laboratory support. The Krebs Institute for Biomolecular Research is a Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Gruneberg, S.; Stubbs, M. T.; Klebe, G. Successful Virtual Screening for Novel Inhibitors of Human Carbonic Anhydrase: Strategy and Experimental Configuration. *J. Med. Chem.* In press.

(2) *Virtual Screening for Bioactive Molecules*; Bohm, H.-J., Schneider, G., Eds.; Wiley-VCH: Weinheim, 2000.

(3) *Virtual Screening: an Alternative or Complement to High Throughput Screening*; Klebe, G., Ed.; Kluwer: Dordrecht, 2000.

(4) Good, A. C. Structure-Based Virtual Screening Protocols. *Curr. Opin. Drug Discov. Devel.* **2001**, *4*, 301−307.

(5) Schneider, G.; Bohm, H.-J. Virtual Screening and Fast Automated Docking Methods. *Drug Discov. Today* **2002**, *7*, 64−70.

(6) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(7) Brint, A. T.; Willett, P. Upperbound Procedures for the Identification of Similar Three-Dimensional Chemical Structures. *J. Comput.-Aid. Mol. Design* **1988**, *2*, 311−320.

(8) Pepperrell, C. A.; Willett, P. Techniques for the Calculation of Three-Dimensional Structural Similarity Using Inter-Atomic Distances. *J. Comput.-Aid. Mol. Design* **1991**, *5*, 455−474.

(9) Good, A. C.; Ewing, T. J. A.; Gschwend, D. A.; Kuntz, I. D. New Molecular Shape Descriptors − Applications in Database Screening. *J. Comput.-Aid. Mol. Design* **1992**, *9*, 1−12.

(10) Good, A. C.; Kuntz, I. D. Investigating the Extension of Pairwise Pharmacophore Measures to Triplet-Based Descriptors. *J. Comput.-Aid Mol. Design* **1992**, *9*, 373−393.

(11) Bemis, G. W.; Kuntz, I. D. A Fast and Efficient Method for 2D and 3D Molecular Shape Description. *J. Comput.-Aid. Mol. Design* **1992**, *6*, 607−628.

(12) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New Method for Rapid Characterization of Molecular Shapes: Applications in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.

(13) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: a System to Select Quasi-Flexible Ligands Complementary to a Receptor of Known Three-Dimensional Structure. *J. Comput.-Aid. Mol. Design* **1994**, *8*, 153−174.

(14) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity Searching in Files of Three-Dimensional Chemical Structures: Comparison of Fragment-Based Measures of Shape Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.

(15) Mason J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Structures. *J. Med. Chem.* **1999**, *42*, 3251−3264.

(16) Matter, H.; Potter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.

(17) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. J. A Computational Ensemble Pharmacophore Model for Identifying Substrates of P-Glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737−1740.

(18) Schmitt, S.; Hendlich, M.; Klebe, G. From Structure to Function: A New Approach to Detect Functional Similarity among Proteins

Independent from Sequence and Fold Homology. *Angew. Chem., Int. Ed. Engl.* **2001**, *40*, 3141−3144.

(19) Pepperrell, C. A.; Poirrette, A. R.; Willett, P.; Taylor, R. Development of an Atom-Mapping Procedure for Similarity Searching in Databases of Three-Dimensional Chemical Structures. *Pest. Sci.* **1991**, *33*, 97−111.

(20) The CONCORD and UNITY software is available from Tripos Inc. at http://www.tripos.com.

(21) Lemmen, C.; Lengauer T. Computational Methods for the Structural Alignment of Molecules. *J. Comput.-Aid. Mol. Design* **2000**, *14*, 215−232.

(22) Brint, A. T.; Willett, P. Algorithms for the Identification of Three-Dimensional Maximal Common Substructures. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152−158.

(23) Gardiner, E. J.; Artymniuk, P. J.; Willett, P. Clique-Detection Algorithms for Matching Three-Dimensional Molecular Structures. *J. Mol. Graph. Model.* **1997**, *15*, 245−253.

(24) Ellis, D.; Furner-Hines, J.; Willett, P. Measuring the Degree of Similarity Between Objects in Text Retrieval Systems. *Perspect. Inf Manag.* **1994**, *3*, 128−149.

(25) Wallis, W. D.; Shoubridge, P.; Kraetz, M.; Ray, D. Graph Distances Using Graph Union. *Patt. Recog. Lett.* **2001**, *22*, 701−704.

(26) Stahl, M.; Todorov, N. P.; James, T.; Mauser, H.; Boehm, H.-J.; Dean, P. M. A Validation Study on the Practical Use of Automated *de novo* Design. *J. Comput.-Aid. Mol. Design* **2002**, *16*, 459−478.

(27) The GOLD software is available from Cambridge Crystallographic Data Centre at http://www.ccdc.ac.uk.

(28) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(29) Raymond, J. W.; Willett, P. Similarity Searching in Databases of Flexible 3D Structures Using Smoothed Bounded Distance Matrices. Submitted for publication.

(30) Moon, J. B.; Howe, W. J. 3D Database Searching and *de novo* Construction Methods in Molecular Design. *Tetrahedron Comput. Methodol.* **1990**, *3*, 697−711.

(31) Ho, C. M. W.; Marshall, G. R. Foundation − a Program to Retrieve all Possible Structures Containing a User-Defined Minimum Number of Matching Query Elements from 3-Dimensional Databases. *J. Comput.-Aid. Mol. Design.* **1993**, *7*, 3−22.

CI025605O