ELSEVIER

Journal of Molecular Graphics and Modelling 20 (2002) 447–462

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

dbtop: Topomer similarity searching of conventional structure databases

Richard D. Cramer*, Robert J. Jilek, Katherine M. Andrews

Tripos Inc., 1699 S. Hanley Road, St. Louis, MO 63144, USA

Abstract

A new topomer-based method for 3D searching of conventional structural databases is described, according to which 3D molecular structures are compared as sets of fragments or topomers, in single rule-generated conformations oriented by superposition of their fragmentation bonds. A topomer is characterized by its CoMFA-like steric shape and now also by its pharmacophoric features, in some novel ways that are detailed and discussed.

To illustrate the behavior of topomer similarity searching, a new *dbtop* program was used to generate a topomer distance matrix for a diverse set of 26 PDE4 inhibitors and 15 serotonin receptor modulators. With the best of three parameter settings tried, within the 210 shortest topomer distances (of 1460), 94.7% involved pairs of compounds having the same biological activity, and the nearest neighbor to every compound also shared its activity. The standard similarity metric, Tanimoto coefficients of "2D fingerprints", could achieve a similar selectivity performance only for the 108 shortest distances, and three Tanimoto nearest neighbors had a different biological activity. Topomer similarity also allowed "lead-hopping" among 22 of the 26 PDE4 inhibitors, notably between rolipram and cipamfylline, while "2D fingerprints" Tanimotos recognized similarity only within generally recognized structural classes.

In 370 searches of authentic high-throughput screening (HTS) data sets, the typical topomer similarity search rate was about 200 structures per s.

Keywords: Topomers; 3D database searching; Pharmacophoric features; Similarity searching; HTS data analysis, PDE4 inhibitors, serotonin receptor modulators; *dbtop*

1. Introduction

Topomer shape similarity has been found very effective in prospectively identifying structures likely to be biologically similar to a query molecule [1] [1], within truly vast "virtual libraries" [2]. In the uncertain earliest stages of drug discovery, speed and low cost are at least as important as a promising forecast in determining which compounds get tested, and so typically virtual libraries contain only readily synthesizable structures [3]. Nonetheless, it is the compounds that someone has already synthesized which are by far the most accessible. Here, we describe methodology developed to perform topomer similarity searching of the "heterogeneous" structural databases which reference such existing compounds.

Heterogeneous databases present the special challenge that topomer similarity is directly defined only for molecular fragments [4]. (With topomers, the mutual alignment necessary for any 3D comparison is achieved by overlapping the free valences of these fragments). In contrast to virtual libraries, which are created as sets of predefined molecular fragments, each structure in a heterogeneous database must be fragmented before any topomer comparisons can be performed. Within virtual libraries $m + n$ topomer generations and comparisons allow selection among $m \times n$ candidate products (where m and n are the number of variations at two sites). But within a heterogeneous database, at least $2p$ topomer generations and comparisons are required for just one structural similarity calculation (where p is the number of fragmentations). Thus in heterogenous databases, the combinatorics of topomer similarity searching are unfavorable rather than highly favorable. This significant performance challenge has required development of additional methodology. [2]

The topomer methodology has also been ignoring "pharmacophoric features", such as charges or hydrogen bonding functionalities, despite their centrality in most 3D structure searching approaches. While adding this capability, we recognized that the single conformation characteristic of topomers affords some interesting and useful departures

* Corresponding author.

[1] It may be mentioned that topomer similarity searching has recently achieved even more impressive successes (5/6 successes in prospective "lead hops" to different scaffolds), in collaborative discovery research carried out by Tripos Receptor Research (Cornwall, UK), but unfortunately, details will be unavailable for some time.

[2] Patent applications describing all of these methodological developments are pending.

1093-3263/02/$ – see front matter
PII: S1093-3263(01)00146-2

from familiar pharmacophoric feature concepts, also discussed later.

Validation presents severe challenges for similarity searching methodologies [5], especially, those intended for analysis of HTS data sets, which publicly are almost non-existent. In pursuing this work, we were fortunate in establishing a collaboration that provided access to a few such sets for validation. However, the usual restrictions on descriptions of such data sets conflict with the disclosure standards of scientific publication. Therefore, for illustrative purposes, from various reviews [6–9], we assembled a much smaller though diverse data set, consisting of 26 PDE4 inhibitors and 15 serotonin inhibitors. Within this data set, containing only active structures, similarity searching methods may be compared in how effectively and consistently they separate PDE4 inhibitors from serotonin inhibitors. More specifically:

- in general, pairs of structures which are scored as similar to each other should have the same biological activity;
- for each individual structure, the one most similar structure (even if it is quite dissimilar) should have the same biological activity.

For comparison purposes, we have also analyzed the similarities among these structures using a standard methodology, Tanimoto coefficients of "2D fingerprints".

It is particularly valuable to detect any possibility of biological similarity between compounds that to a chemist appear structurally dissimilar. Generally, similarity in the eyes of a chemist is a rather good predictor of biological similarity (otherwise medicinal chemistry would be ineffective). Pharmaceutical patent applications take fullest advantage of this tendency, so obviously any ability to identify structures outside the scope of a competitive patent while retaining biological activity can be of the greatest practical value. Furthermore, such a successful "lead-hop™" may also escape other hazards, such as toxicity, insolubility, or biological instability. For these reasons, we have also divided the PDE4 and serotonin inhibitors into structurally distinct subclasses, hoping to find some topomer similarities that bridge these subclasses and thereby illustrate the exceptional ability of topomer searching to detect potential "lead-hops".

2. Methodology

2.1. Overview of topomer similarity searching

This approach differs from other shape comparison methods [10] in several fundamental ways.

- Topomer shape comparisons consider all the atoms, in contrast to pharmacophore-based "3D searching" approaches where "shape" comparison focuses on a small subset of atom-like features.
- Shapes are compared as a combination of fragment-to-fragment differences, rather than by some operation on complete structures.
- Shape comparison usually involves only one "topomer" conformation for each fragment, rather than the indefinitely large variety of conformations that most fragments are capable of achieving.
- Shapes resemble one another mostly to the extent that the same spatial volume or "field" elements are occupied or avoided by the corresponding fragments (in their topomer conformations), not to the extent that achievable geometries among assumed critical features are shared.
- The topomeric difference between two molecules is the minimum value found among the many fragment set comparisons that are usually possible.

Topomer similarity searching of virtual libraries is an extremely efficient process (see Fig. 1 in reference 2). The topomer descriptions of all the synthons that compose a library are computed and saved in a relational database (RDB) as the library is constructed. A query structure is fragmented in every way that might match entries within the virtual libraries to be searched, typically at all single acyclic bonds and at all pairs of single acyclic bonds such that all the resulting fragments contain at least four heavy atoms. During the search itself, the only calculation results needed are the steric field differences between a synthon and a query fragment (and now their feature differences also, as detailed later).

However, in the topomer searching of conventional databases, each structure becomes a unique hit candidate, to be fragmented multiple times in exactly the same fashion as the query structure is fragmented in every topomeric search. The RDB-supported architecture used in virtual library searching was rejected for *dbtop*, in favor of the increased algorithmic and environmental flexibility provided by a self-contained application. Therefore, during a topomeric search of a conventional database, the following three processes are to be carried out for every candidate hit.

- The topomer conformations are generated for each of the fragments from each of the reasonable candidate fragmentations.
- The topomer descriptors are generated for each of these topomer conformations.
- Each set of candidate topomer descriptors must be compared against each set of query topomer descriptors, until the most similar comparison, the desired result, has been identified.

The computational demands are about the same among these three processes, so considerable thought and experimentation was devoted to the optimization of each.

2.2. Generating topomer conformations

In general, the topomer for a fragment is based on a concord-generated [11] 3D structure of a whole molecule.

Each bond to be fragmented is positioned in Cartesian space, and then, proceeding away from that bond, the chiralities and dihedral angles within that fragment are adjusted in strict accordance with prioritization rules. The general effect of these prioritizations is to place the "most important" attachment farthest away from the fragmentation bond and the second "most important" attachment "to the right" of the main chain as viewed away from the fragmentation bond. "Most important" had meant highest molecular weight, but with the introduction of features, the *dbtop* topomer prioritization accords roughly the same importance as three additional carbon atoms to any hydrogen-bonding group.

As would be expected, recoding the existing SPL scripted procedure [11] for topomer generation into C also improved its speed by many orders of magnitude.

2.3. Topomer descriptor generation

The steric similarity of two topomers is the squared sum of differences in steric field values over corresponding pairs of lattice points. Thus the computing times and memory requirements for this and the final process are essentially proportional to the volume of the steric field lattice. On the other hand, with the fixed lattice dimensions used in virtual library searching, occasionally the topomer conformation does make a few atoms sterically invisible, by placing them outside the lattice. For both these reasons, considerable effort went into making the lattice dimensions in *dbtop* depend on the maximum extent of the query fragment topomers—as big as necessary but no bigger. In practice, the lattice goes two grid points in all directions beyond this maximum extent. A constant penalty is then applied to the steric similarity for each fragment candidate atom located outside the query lattice.

The accompanying numerical approximation—setting the steric field intensity to zero wherever its value falls below 0.2—has a negligible effect on the steric similarity values. Furthermore, this approximation allows run-length encoding of the steric fields, highly useful in both minimizing memory and accelerating field comparisons, since the vast majority of lattice points become zero-valued.

2.4. Topomer descriptor comparisons

For each candidate hit structure, the number of comparisons needed to find the greatest similarity to a query is given by 2kmn, where $k = 2$ for a two-piece fragmentation and $k = 4$ for a three-piece fragmentation; m and n are the number of fragmentations possible for the query and candidate structure; and *2* results from the need to compare "both ways" unless the query or candidate are symmetric. Values of m and n, the number of possible fragmentations for query or candidate hit structure, vary greatly, from 1 to several 100, but 15 is representative. Because each of those roughly 1000 possible comparisons itself involves traversal/summation over about a 1000 lattice points, there has been a continuing search for criteria that might rapidly prioritize them. The most powerful of these have been feature comparison (see later) and difference in heavy atom counts.

2.5. Pharmacophoric features

The fundamental topomer restriction to a single (or at most a very few) conformation(s) per fragment fixes the absolute position of features as well as the steric shape. Thus the degree of feature similarity between two topomeric fragments reduces to two questions for each query feature. Is a pharmacophoric feature of the same-class present in the hit candidate? If so, how many angstroms separate (characteristic atoms in) the corresponding feature locations? A maximum penalty, scaled to yield typical total feature differences of a magnitude comparable to steric field differences, is added if there is not a feature of the same-class or if characteristic atoms in the nearest same-class features are too distant (by default, more than 1.5 Å). No penalty is assessed if the characteristic atoms are separated by 0.5 Å or less, and intermediate distances are handled by linear interpolation. A single feature in the candidate can "match" any number of query features, and also the feature difference penalties are attenuated by rotatable bonds in the same manner as steric shape differences.

The five feature classes, as listed in Table 1, will be familiar, and the substructural definitions used for feature recognition are also conventional and so not shown. (We did find it worthwhile to refine the standard Unity/DISCO [11] definitions by several rounds of manual inspection of around 100 structures whose atoms were color-coded by feature class.) Note, however, in column 2 of Table 1, that the maximum feature penalty varies among the different feature classes. Another novel aspect of topomeric pharmacophore features are additional penalties for any "extra features" in the target, ones that were not "consumed" by matching features in the query. These much lower penalty values appear in column 3 of Table 1. All of these values and behaviors are adjustable at run time.

This pharmacophoric feature model has also been fully implemented in the ChemSpace™ software for topomer searching of virtual libraries.

Table 1
Topomer difference penalty values for various classes of features

Feature class	Maximum penalty values	
	Not in candidate	Not in query
Aromatic	20	2
Positive charge	200	20
Negative charge	200	20
HB acceptor	100	10
HB donor	100	10

2.6. Steric scaling: the "pivot factor"

The initial *dbtop* studies, searches within HTS data sets using the confirmed actives as queries, suggested that effectiveness might be improved if the steric shape similarity radius was effectively varied with query molecule size. Smaller structures seemed to possess a smaller steric similarity radius than did larger structures. Therefore, the following optional behavior was implemented. A "typical" number of heavy atoms in a query, by default 30, is denoted as the "pivot value" p. If the number of heavy atoms in the query, h, is less than p but greater than $(2/3)p$ (thus by default between 20 and 30), the usual steric difference is multiplied by the factor p/h. (This multiplication increases the effective steric differences and so reduces the absolute steric difference acceptable for a hit). If h is less than $(2/3)p$, then a maximum pivot factor of 1.5 is used. On the other hand, if h is greater than the pivot value, then the pivot factor becomes $(2 + (h/p))/3$, an expression whose value evidently approaches 2/3 as p increases.

2.7. Selection and structural classification of PDE4 and serotonin inhibitors

Figs. 1–3 show the 41 structures chosen, numbers **1**–**26** in Figs. 1 and 2 being the PDE4 inhibitors and **27**–**41** in Fig. 3 the serotonin inhibitors. The PDE4 inhibitors all act at the same receptor subclass, while the serotonin inhibitors act on different receptor subclasses. Common names for many of the structures are listed in column 2 of Table 2.

To provide an objective and independent structural classification as a basis for assessing "lead-hopping", the maximum common substructure program Distill [11] was used with all default settings. Its results are shown in column 3 of Table 2, as a cluster membership for each structure, in the form of a numeric primary code and occasionally a letter secondary code.

2.8. Topomeric distance matrices and their tabulation

For each of three standard settings, a complete topomeric distance matrix among the 41 compounds was generated, by performing 41 *dbtop* searches with a sufficiently large search radius. Default values were used for all parameters except for these settings; however, charged features were turned off, so as not to excessively bias the experiment toward the desired outcome (serotonin inhibitors all are basic amines, hence positively charged, while PDE4 inhibitors are either neutral or negatively charged). Note that these topomer distance matrices are not symmetric about their diagonal, in particular because the feature difference and the pivot factor both depend on which one of a compound pair is designated as the query.

These three distance matrices were tabulated with an SPL script that reported, for a given range of distances and each compound, the identity of its neighbors. The script also summarized, for that distance range and the entire matrix, how many values either "correctly" or "incorrectly" linked structures having the same or different biological effects. The distance range was increased and the script rerun until any tendency for the additional neighbors to share biological activities disappeared. These three sets of summaries appear in Table 3. To allow comparison, the fourth block of Table 3 presents the results from applying the same procedures to the same structures with the most widely used "diversity descriptor", the Tanimoto coefficients of (unity) "2D fingerprints".

To explore the possibilities of "lead-hopping", i.e. using topomer similarity to move between structural classes

Table 2
Some attributes of the 41 compounds studied[a]

ID number	Common name	Cluster ID[b]
1	Rolipram	1
2	Piclaminast	1
3	CP-353164	1
4	GW3600	1
5		1
6	Filaminast	1
7	Ariflo	1
8	Atizoram	1
9	RO_20_1724	2
10	Benzylrolipram	1
11	ORG_20241	2
12	Zardaverine	3
13		3
14	Cipamfylline	8
15	KF-19514	6b
16	Nitroquazone	6b
17		6c
18	Denbufylline	8
19	Pentoxyfylline	8
20	Arofylline	8
21		3
22		6b
23		4a
24	AH_21_132	3
25	D22288	9
26	NCS-613	5a
27	Benzotiazione	7
28	GR65630	7
29		7
30	ICS-205-9	7
31	Ondansetrone	7
32	Quipazine	6c
33	Seratonin	7
34		7
35	Zacopride	5b
36	BMY7378	6a
37	GR127935	6a
38	Ketanserin	7
39	SB271046	6a
40	Spiperone	4b
41	WAY100635	6a

[a] Structures are shown in Figs. 1–3.

[b] Clustering as produced by Distill, a maximum common substructure detection program. A numbers indicates a major cluster, possibly followed by a letter indicating a subcluster of the major cluster.

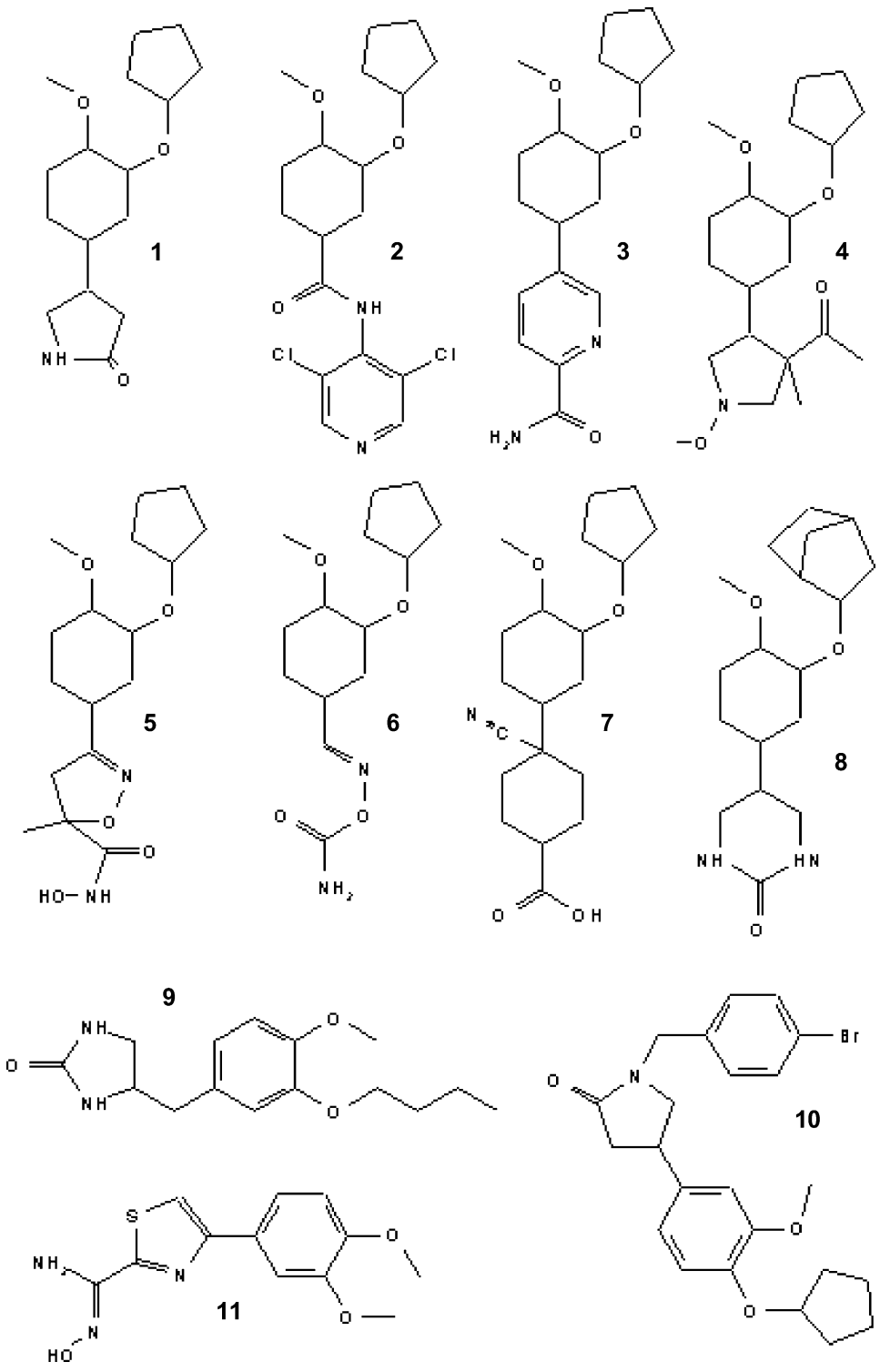

Fig. 1. Structures of the rolipram-like PDE4 inhibitors.

while retaining biological activity, near neighbor lists were generated for various structures by sorting the distance matrix. Cutoffs to each of these four sorted near neighbor lists were established by inspection of the results in Table 3, as the largest difference (topomer or Tanimoto) that still strongly grouped together structures having the same type of activity. (In the case of the Tanimoto coefficient, the critical data bin was split, to optimize this control descriptor's performance and thus again minimize bias in favor of the preferred outcome.) This selection of four cutoff values is discussed further in the Section 3.

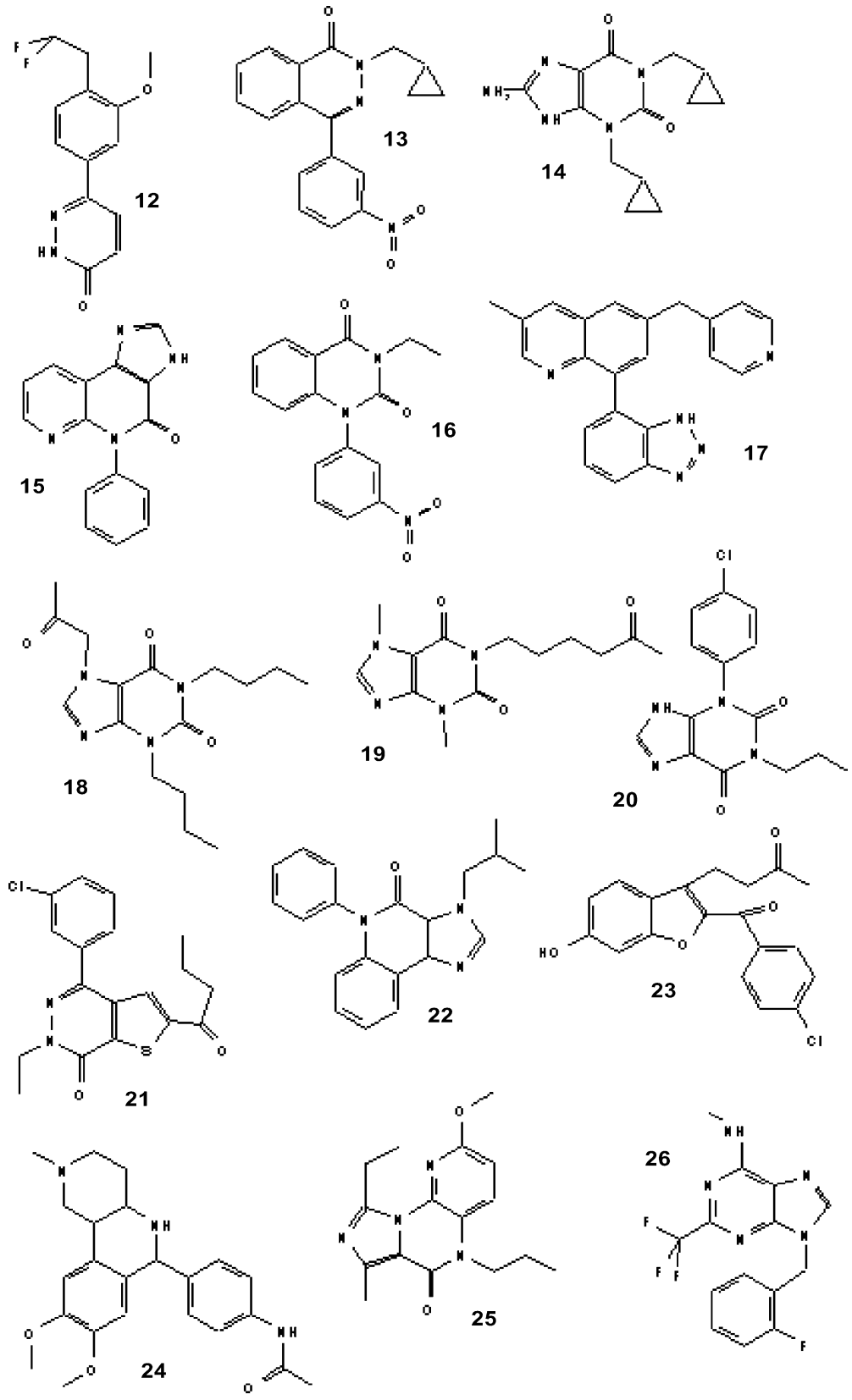

Fig. 2. Structures of the additional PDE4 inhibitors.

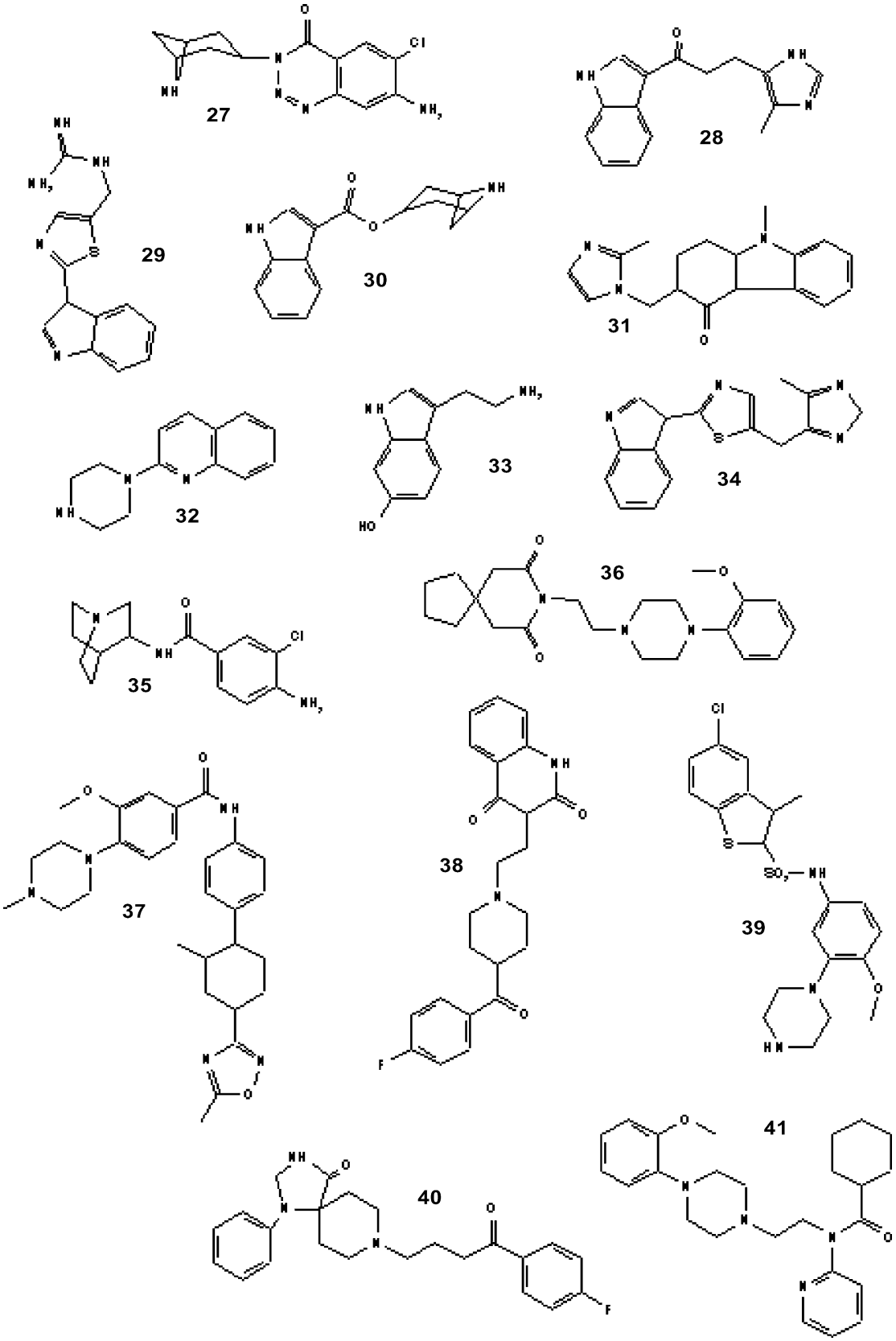

Fig. 3. Structures of the serotonin receptor modulators.

Table 3
Tabulations summarizing the distance matrices from the 41 compounds with two types of activity, for each of the four descriptor classes[a]

Descriptor class (distance range)	Distance tabulation summary						Compound-by-compound		
	Incremental			Cumulative			Number of singletons (no neighbor)	Nearest has different type of activity	
	Numbers of neighbors with wrong activity			Number of neighbors with wrong activity				Number	Structure IDs[b]
	Same	Different	Same (%)	Same	Different	Same (%)			
A. Topomers with features									
<200	82	0	100.0	82	0	100.0	15	0	
200–209	37	3	92.5	119	3	97.5	8	2	**23** > **35**, **32** > **12**
210–219	28	3	90.3	147	6	96.1	7	0	
≫220–229	34	9	79.1	181	15	92.3	5	0	
230–239	53	29	64.6	234	44	84.2	5	0	
240–249	85	46	64.9	319	90	78.0	3	0	
250–259	115	57	66.9	434	147	74.7	2	0	
260–269	115	85	57.5	549	232	70.3	2	0	
>270								0	**24** (294), **39** (273)
B. Topomers without features									
<120	47	0	100.0	47	0	100.0	22	0	
120–129	25	5	83.3	72	5	93.5	16	3	**19** > **28**; **32**> **12**; **33** > **19**
130–139	37	11	77.1	109	16	87.2	12	1	**25** > **33**
140–149	50	12	80.6	159	28	85.0	7	2	**23** > **35**; **26** > **33**
150–159	69	26	72.6	228	54	80.9	3	0	
160–169	93	71	56.7	321	125	72.0	3	0	
>170								1	**24** (176); **37** (184); **39** (174) > **2**
C. Topomers with steric scaling (and features)									
<230	84	0	100.0	84	0	100.0	18	0	
230–239	20	1	95.2	104	1	99.0	15	0	
240–249	24	1	96.0	128	2	98.5	9	0	
250–259	36	1	97.3	164	3	98.2	7	0	
≫260–269	50	9	84.7	214	12	94.7	6	0	
270–279	64	26	71.1	278	38	88.0	4	1	**26** > **33**
280–289	68	44	60.7	346	82	80.8	2	1	**32** > **12**
								1	**24** (297); **33** (341) > **19**
D. Tanimoto coefficients of "2D fingerprints" (unity)									
>0.7	12	0	100.0	12	0	100.0	31	0	
0.60–0.70	10	2	83.3	22	2	91.7	25	1	**16** > **38**
0.50–0.60	44	2	95.7	66	4	94.3	11	1	**22** > **40**
≫0.45–0.50	50	4	92.6	116	8	93.5	8	1	**37** > **24**
0.40–0.45	62	34	64.6	178	42	80.9	6	0	
0.30–0.40	205	261	44.0	383	303	55.8	0	0	

[a] The text contains a full description of this table.

[b] An expression such as "**23**> **35**" indicates that the nearest neighbor to **23** is **35**, the activities of the two compounds are different, and the distance is within the range shown for the row. An expression such as "**24** (294)" indicates that the nearest compound to **24** is 294 units away, a value greater than any tabulated distance range for the descriptor class. Mixed expressions indicate the occurrence of both conditions.

3. Results

The correlation matrix (*r*-values) among the inter-compound distances, as measured by each of the four descriptors, three topomeric and the Tanimoto, appears in Table 4. (Only the distance between a compound and itself was excluded.) There are modest correlations among the three classes of topomeric distance, and practically, no correlation between the Tanimoto and any of the topomeric distances.

Table 3 shows the most important results for the PDE4/serotonin data set. Reading vertically, the four blocks of data compare the results among the various modes of *dbtop* and

Table 4
Correlation matrix (*r*-values) among the 41 structures for the four descriptors

	Topomer (features)	Topomer (no features)	Topomer (pivot)
Topomers features			
Topomers (no features)	0.76		
Topomers (pivot)	0.73	0.55	
Tanimoto	0.28	0.14	0.32

the standard diversity metric, Tanimoto coefficients of "2D fingerprints". (It should be noted that topomeric distances are computed, Euclidean fashion but non-intuitively, as a root sum of the squared individual property differences (lattice point by lattice point for the steric field and feature by feature). Thus a specific structural change that increases a topomeric distance from 100 to 120 will increase a topomeric distance from 200 to 211).

The left distance tabulation section of Table 3 presents the summary distance statistics for various distance ranges, generated as described above, in two triplets of columns. The left triplet contains incremental results, those within the particular distance range shown, and the right contains cumulative results, for all distances less than the upper limit of that range. Within each triplet, the left column records the count of distances between compounds having the same-activity, the middle, the count of distances between compounds of different-activity, and the right, the percentage of the total count between compounds of the same-activity. The larger the percentage, the more discriminating and desirable the underlying molecular descriptor, while percentages around 50% and lower indicate a meaningless descriptor (the exact "meaningless percentage" for this data set would be 52.4%). Of course, no matter the descriptor, a large enough distance range will necessarily include pairs of compounds having different biological activities. However, with a perfect descriptor every one of the same-activity inter-compound distances would be smaller than any of the different-activity inter-compound distances.

The right side of Table 3 considers the topomer distance matrix from a compound-by-compound perspective. In its first column is the count of "singletons", those compounds which do not have any neighboring compound for that descriptor closer than the distance range shown. The two rightmost columns show both the count and the identity of any compounds whose closest neighbor wrongly has the other kind of activity. A right angle bracket points from the compound ID–ID of the neighbor with the nearest activity. Also shown at the bottom of each descriptor block are the IDs of those compounds which remained singletons at the largest distance investigated, with the distance to its nearest neighbor in parentheses, and the angle bracket to the neighbor if there is then an activity class mismatch.

From almost every point of view, Table 3 shows block C, topomer similarity including both features and steric scaling, to be the most effective descriptor in discriminating PDE4 inhibitors from serotonin inhibitors. One straightforward way to compare the four descriptors is to establish a distance range criterion or "cutoff" for each by establishing a minimum acceptable level of performance. For example, let us say that we will choose the largest distance range within which the activity-matching neighbors make up around 80% of all neighbors. For reference, these distance ranges for each of the four descriptors are marked in Table 3 with double right angle brackets in the first column. At these maximum acceptable distance ranges, topomers with features and steric scaling (C) includes 214 distances, topomers with features (A) includes 181 distances, topomers with shape-only (B) includes 159, but "2D fingerprints" (D) includes only 116 distances. (A perfect descriptor would partition 860 shorter distances between compounds of matching activity ($\mathbf{26} \times \mathbf{26} + \mathbf{15} \times \mathbf{15} - \mathbf{41}$) from 780 longer distances between compounds of mismatching activity ($\mathbf{26} \times \mathbf{15} + \mathbf{26} \times \mathbf{15}$), so there is room for improvement). At the same time, the 94.5% accuracy of the cumulative selection is also highest for topomers with features and steric scaling. Tanimoto at 93.5% are nearly as accurate, albeit for barely half the number of distances and with what may be a fortuitous sub-partitioning of the Tanimoto distance range, followed closely by topomers with features (92.3%) and steric-only topomers (85.0%). Also close are the counts of compounds having no nearest neighbor, where topomers with features and steric scaling (count of 6) are second to features (count of 5), then sterics only (count of 7), and Tanimoto (count of 8) last again. Finally, the number of mistakes, compounds whose nearest neighbor has the other kind of activity, again strongly favors block C, with no mistakes at all. Topomers with features make two mistakes, Tanimotos three, and topomers with sterics only a distant last with six mistakes.

To evaluate the possibilities for topomeric lead-hopping among the PDE4 and the serotonin inhibitors, some agreement is necessary on the distinctive conventional structural themes, such that transitions between themes would exemplify lead-hopping. We have augmented our qualitative comparisons of structure with the objective output of a maximum common substructure program Distill, shown in column 2 of Table 2. In general, both approaches find two major structural themes in both the PDE4 and the serotonin activity classes, as well as some minor themes and individual structures. For PDE4, the two major themes are "rolipram-like", with its 3-cyclopentoxy-4-methoxyphenyl group, including structures **1**–**8** and **10** as cluster 1, and "fylline", alkylated purines including structures **14** and **18**–**20** as cluster 8. For serotonin, the two major structural themes are indole-like, as cluster 7 including **28**–**31**, **33**, **34**, and according to Distill but arguably **27** and **38**, and *ortho*-methoxy-phenylpiperazine-containing, as cluster 6a with **36**–**37**, **39**, and **41**. These four major themes include 25 compounds, almost two-thirds of the 41.

Whenever two members of structurally distinct themes are topomerically separated by a distance that is lower than some topomeric threshold distance or "cutoff", then the possibility of a lead-hop exists. As previously described, from the results in Table 3, the most discriminating topomeric descriptor and threshold found was steric scaling and features with a threshold of 270. With this cutoff, all the potential topomeric lead-hops out of one of the four major themes are listed in Table 5 in two groups, the first including those which land in the other major series and the second those encountering other compounds with the same-activity. (There were also 12 "mistaken" lead-hops, transitions from one class of activity to the other.)

Table 5
Potential "lead-hops", or topomer similarities between structures in different chemical series[a]

Major-to-major[b]		Major-to-minor	
Structure IDs	Topomer distance	Structure IDs	Topomer distance
1 > **14**	261	**8** > **22**	270
4 > **14**	265	**10** > **9**	229
10 > **18**	270	**10** > **26**	247
19 > **1**	264	**10** > **15**	251
		10 > **19**	254
41 > **28**	263	**10** > **11**	259
41 > **30**	254	**10** > **12**	260
41 > **27**	270	**10** > **13**	262
		10 > **22**	265
		14 > **15**	228
		18 > **25**	225
		18 > **15**	250
		20 > **15**	245
		20 > **13**	253
		20 > **22**	265
		29 > **35**	262
		38 > **40**	261
		41 > **40**	250

[a] Series are delineated by MCS cluster membership (column 3 of Table 2). Major and minor series are defined in the text. The descriptor class is topomers with features and steric scaling.

[b] Images of the 3D overlays corresponding to these possible "lead-hops" appear in Figs. 4–10.

Figs. 4–10 show seven overlays of the topomer shape and feature comparisons that underlie the potential lead-hops between major series. Each of the comparisons involves two paired fragments, since it happens that all six best comparisons involved two pieces. The topomeric attachment bond used to align each of the paired fragments is always at the far left of a pair. Every fragment from a query is colored by atom type, and every fragment from a hit is colored yellow.

As these overlays are evaluated, there are several less intuitive properties of topomers to consider. First, the rotatable bond attenuation factor of 0.85^n (where n is the number of rotatable bonds separating an atom from the aligning attachment bond) means that steric and feature differences more distant from the attachment bond have much less effect on the overall topomeric difference. Second, the steric scaling factor accepted much larger steric (but not feature) differences when the query structure was the very large serotonergic **41** (WAY10065).

Unsurprisingly, results were meager from a similar "lead-hopping" experiment with Tanimoto coefficients of "2D fingerprints". No major-to-major transitions occurred. At a cutoff similarity of 0.46 or greater, there were five marginal major-to-minor "lead-hops", three being from the rolipram class to structure **24** and the other two being **27** > **35** and **41** > **32**. There were also eight "false lead-hops".

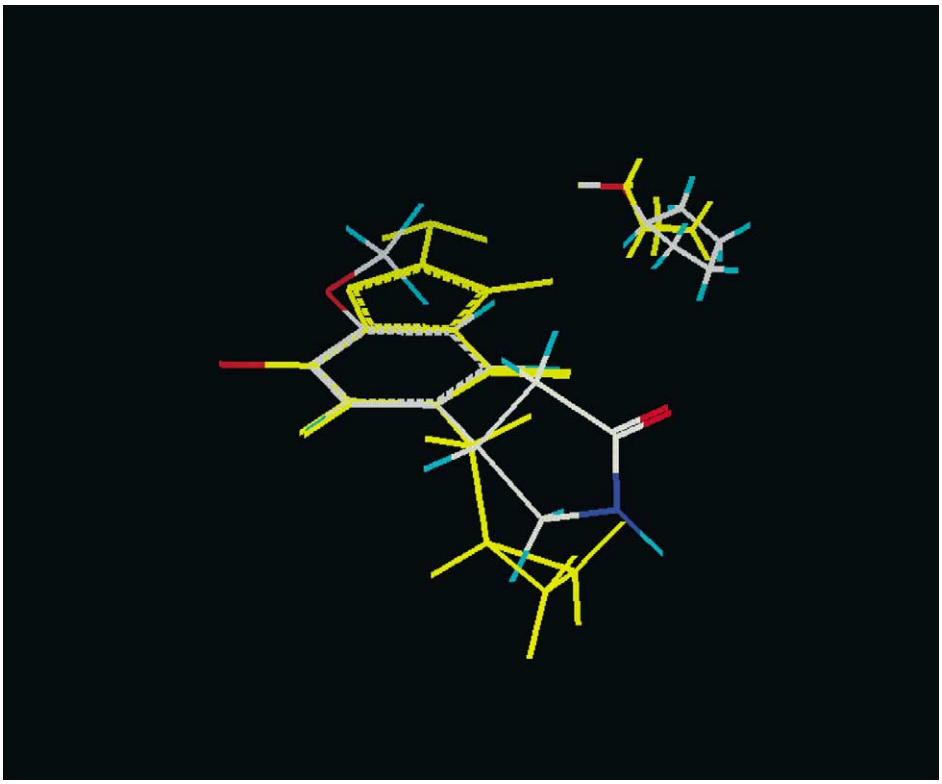

Fig. 4. Topomer similarity overlay of structure **1** (rolipram) and structure **14** (cipamfylline).

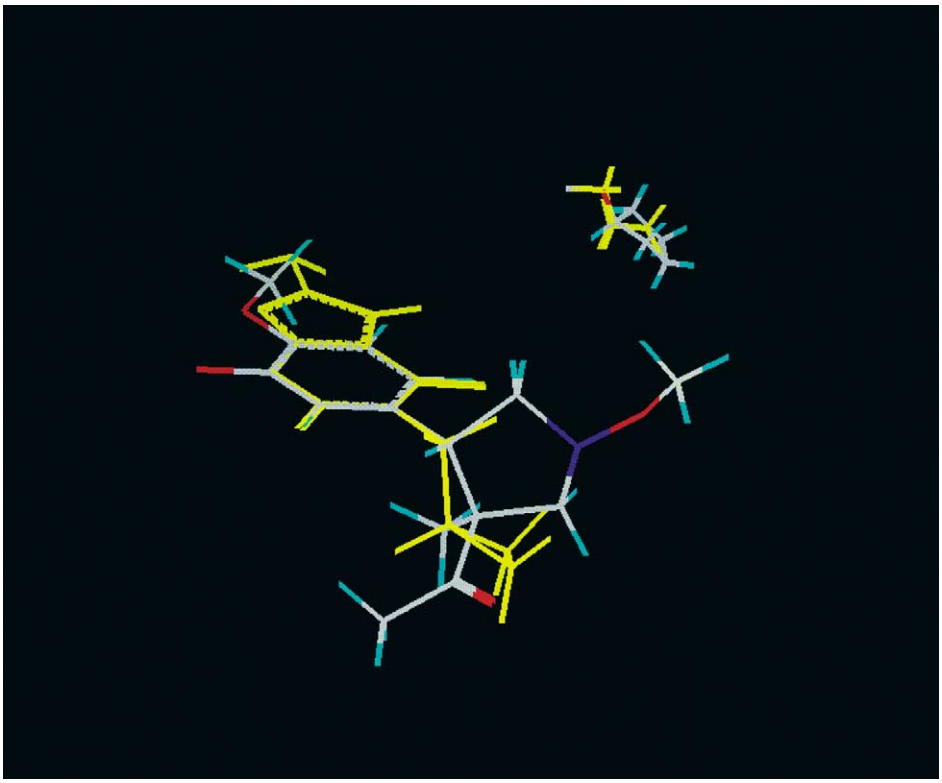

Fig. 5. Topomer similarity overlay of structure **3** (GW3600) and structure **14** (cipamfylline).

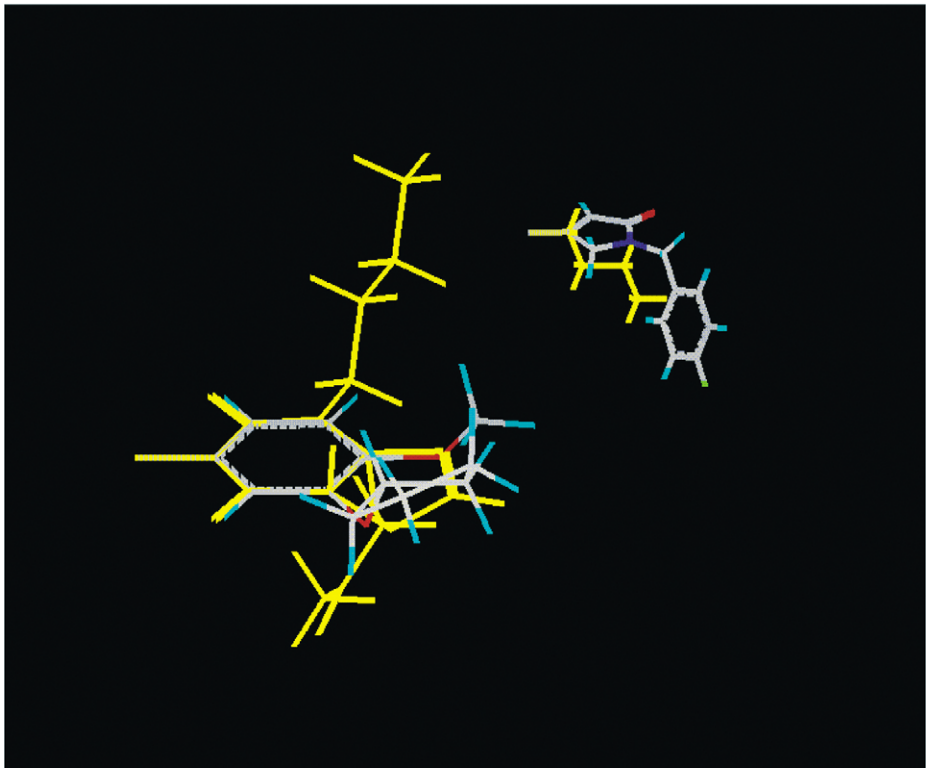

Fig. 6. Topomer similarity overlay of structure **10** (benzylrolipram) and structure **18** (denbufylline).

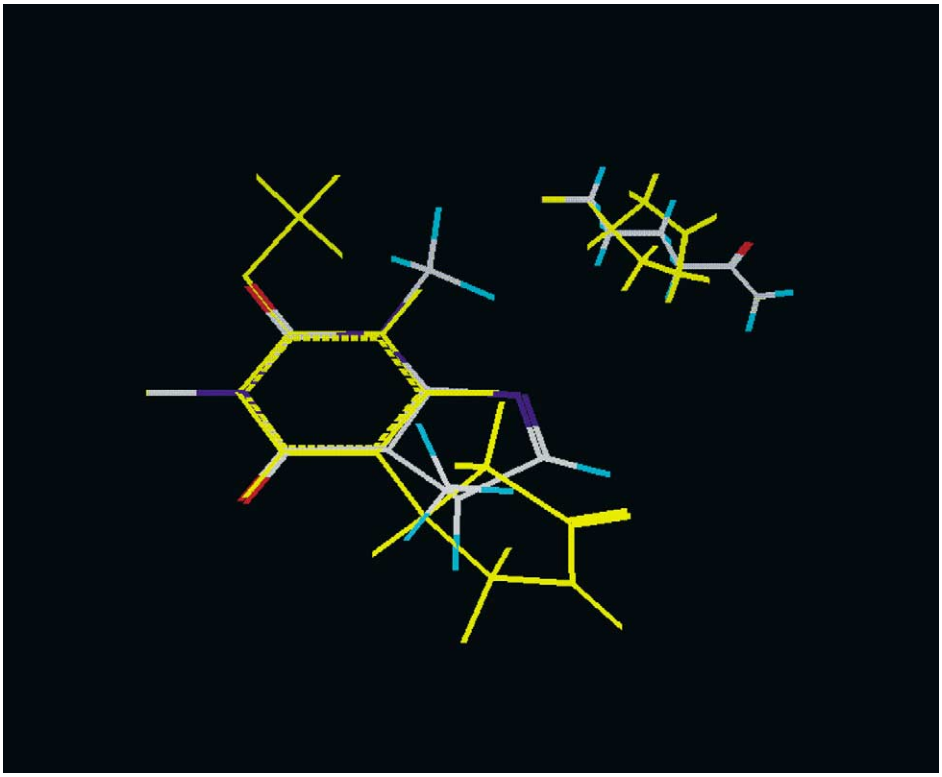

Fig. 7. Topomer similarity overlay of structure **19** (pentoxyfylline) and structure **1** (rolipram).

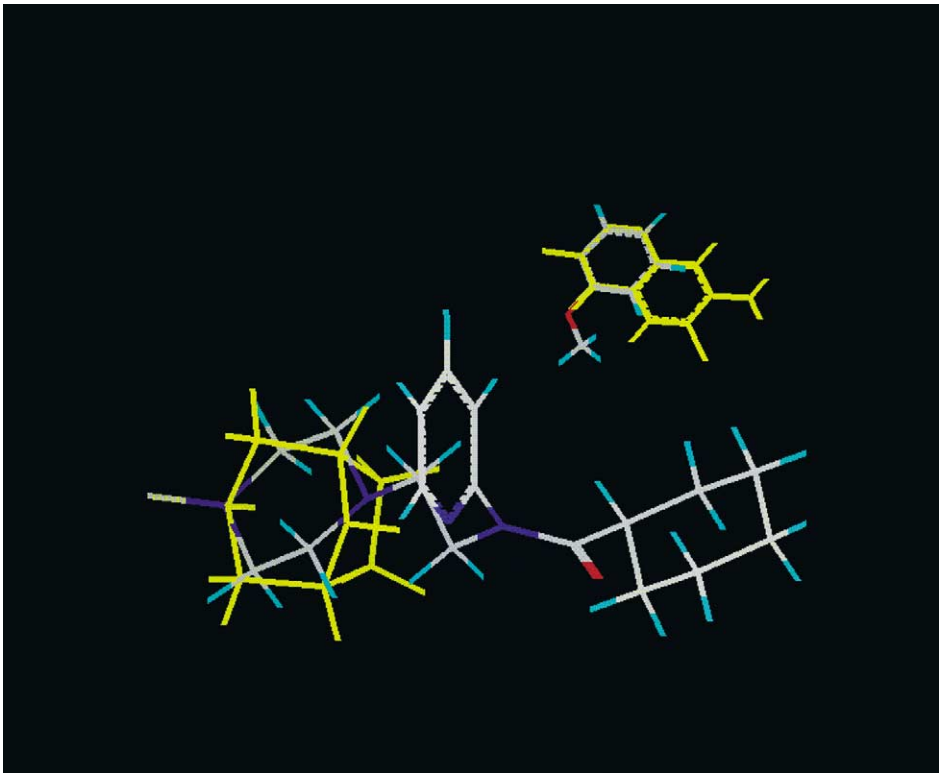

Fig. 8. Topomer similarity overlay of structure **41** (WAY100635) and structure **28** (GR65630).

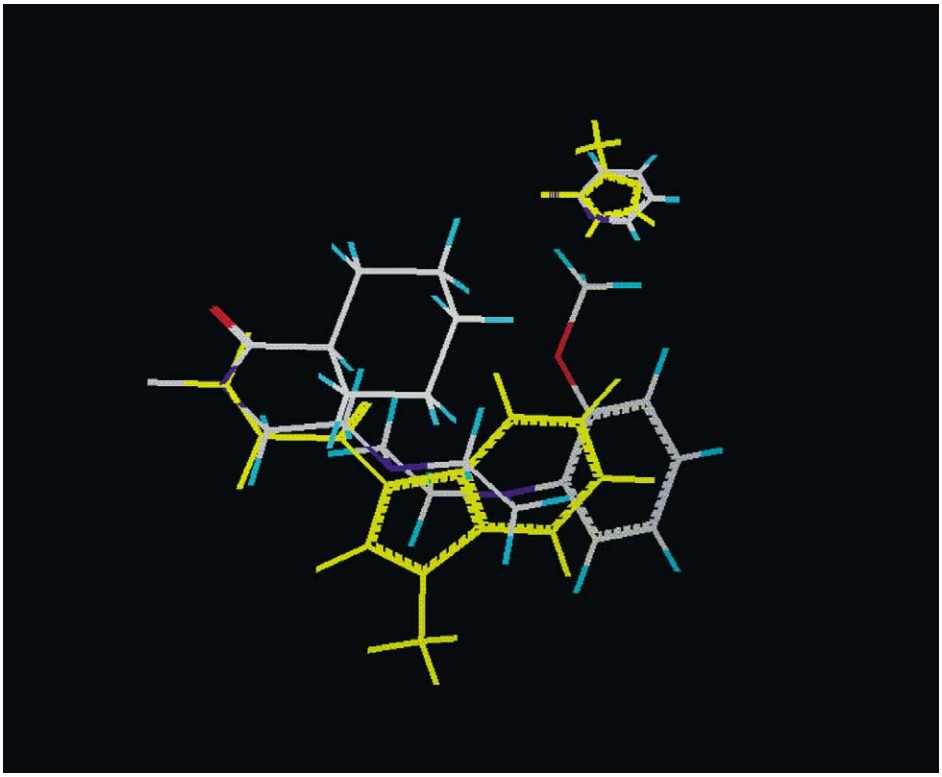

Fig. 9. Topomer similarity overlay of structure **41** (WAY100635) and structure **30** (ICS-205-9).

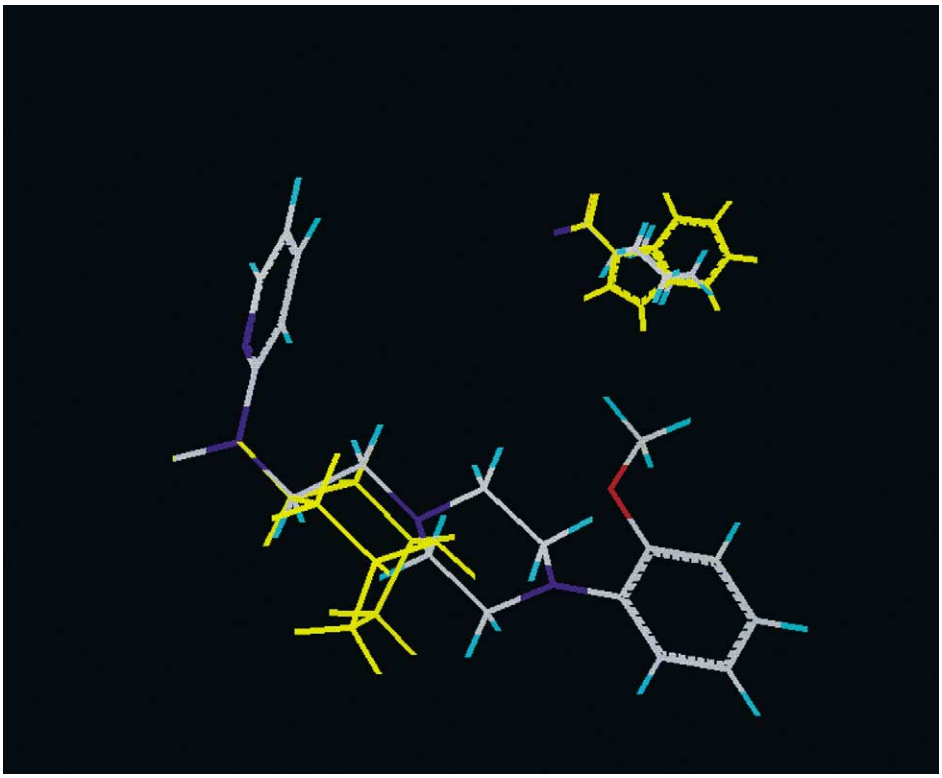

Fig. 10. Topomer similarity overlay of structure **41** (WAY100635) and structure **27** (benzotiazione).

4. Discussion

From the results in Table 3, topomer searching of conventional databases with the *dbtop* program appears quite effective in selecting structures of similar biological activity, particularly when including features and steric scaling (the default behavior). Below the cutoff distance of 270 units, almost all (94.7%) of topomeric distances separated biologically similar compounds, and no compound had a biologically dissimilar structure as its nearest neighbor. (True, three of the six compounds separated by more than 270 units from any other compound did have a biologically dissimilar nearest neighbor. However, this result seems more likely an inadequacy of the density of compound sampling than of the topomer descriptor. For this reason, in comparing descriptor performances, a high average association of short distances with similar activities seems far more important than the biological similarity of nearest neighbors). A more impressive result than the 94.7% correct is that the sample included 202 of the 860 (23%) possible activity-matching differences. These results would have been even better had charged features not been excluded from the topomer searching, as excessively biasing toward the desired outcome.

It may be surprising that such desirable behavior results from comparing the properties of single, perhaps arbitrary-seeming, conformations, that are unlikely to resemble receptor-bound conformations. Our objective in formulating the topomer generation rules has been that fragments whose topomers are shape similar would also be similar in their entire ensemble of energetically accessible states. Yet, as is evident in Figs. 4–10, topomer similarity is not only a consequence of topological similarity.

The standard similarity/diversity metric, Tanimoto coefficients of "2D fingerprints", was not nearly as effective in recognizing biological similarity, despite an ad hoc tuning of the distance range at the critical cutoff (0.05 step size rather than 0.10) to obtain the best possible relative performance. Although the 93.5% correct biological matches at the cutoff of greater than 0.45 similarity is almost as good, the three compounds having a biologically dissimilar nearest neighbor is not as good. However, the most important relative weakness of Tanimotos is the much smaller proportion of the possible activity-matching differences recognized, 108 of 860 activity-matching differences (12%). (Note that because of the lack of symmetry of topomeric distances, full distance matrices excluding the diagonal are the basis for all tabulations reported). It should be noted that other "2D fingerprints" implementations (Daylight, MDL) might have performed better, although the differences among these technologies appear slight [10].

Given these 41 compounds, many workers might not have relied on Tanimotos at all, since the usually accepted Tanimoto "neighborhood radius" value is 0.85 or larger, while in this set, there is no pairing of compounds whatsoever having a Tanimoto similarity greater than 0.75. That lower Tanimoto similarities were actually found useful does suggest some atypical quality about this dataset. One factor is that the dataset is very small, so perhaps there has been some sampling bias. But a certain and probably more important factor is that that the 0.85 Tanimoto cutoff value was derived from a rather different success criterion, that of selecting structures more likely to be active from the great multitude of structures unlikely to have any sort of activity [12,5]. Separating actives from inactives seems a much more demanding task than separating one kind of activity from another, the objective here and in several other similar studies [13,14]. Consistent with this view, the topomer similarity cutoffs used here also are larger than those found in our previous "neighborhood behavior" validation studies of descriptors, although much of this increase arises from the addition of features. (From block B of Table 3 above, the cutoff of 150 for steric-only topomer differences contrasts with a cutoff of 90 from previous work. On the other hand, in actual prospective syntheses, steric-only topomer differences as large as 120 and even 150 have yielded successful "lead-hops"). Considering the enormity of ways that drug candidates can interact with various receptor targets, similarity radii values will always tend to be situationally dependent.

The possibility of lead-hopping, identifying structurally different but biologically similar structures, is the greatest advantage of any 3D-similarity-searching method. We find the variety of example topomeric lead-hops found among almost all the PDE4 inhibitors to be extremely encouraging. Indeed, as can be decoded from Table 5, only four of the 26 PDE4 inhibitors (**17**, **21**, **23**, and **24**) are topomerically more than one lead-hop away from the two major structural classes. There are also four lead-hops connecting those two structural classes. Perhaps, the most impressive of these is the shape and feature similarity between rolipram and cipamfylline, depicted in Fig. 4. We wonder what proportion of the many medicinal chemists who have sought new PDE4 inhibitors recognized any 3D resemblance between these two landmark structures, given the great dissimilarity in their scaffold structures. It would also be interesting to know how well this proposed 3D correspondence might be confirmed by direct structure determinations.

Lead-hopping among the 15 serotonin inhibitors was less facile, but this is understandable given their smaller number, their greater structural diversity, and in particular, the multiplicity of serotonin target sub-classes that is represented.

As mentioned in the introduction, *dbtop* has been applied to a typical HTS data set supplied to us by a collaborator. Of course, nothing can be said about these results other than to report general consistency with the findings reported here. However, we can be quite specific about search times. For 370 queries (the confirmed actives) of approximately 150,000 structures on typical SGI workstations, the average search time was 2 h (ranging from 0.5 to 6 h depending mostly on the number of fragmentations yielded by the query structure), or roughly 200 candidate structures processed per second. Thus, despite the unfavorable combinatorics for

topomer searching of conventional libraries as compared to virtual libraries, the various refinements summarized in the experimental section have produced a search speed that compares favorably with conventional pharmocophore-centric 3D searching.

Conventional 3D searching also has conceptually inherent weaknesses which are perhaps the less obvious for being universal. The objective of such a search is structures that are conformationally capable of presenting a specified geometric arrangement of a specified set of pharmacophoric features. But the requisite search algorithms have very "hard-edged" behaviors, which seem unfortunate given that surely the actual user objective is to identify "any structure likely to have biological similarity". For example:

- all of the features must be present, no partial credits (algorithms that do accept partial matches are notorious for their slowness);
- the distance tolerances are absolute (again no partial credit) and must not be too large (or the search output and times may be excessive);
- unrequested features in a target structure are ignored, which in practice can favor as hits such unrealistically highly functionalized structures as peptides and sugars;
- all feature classes are equally mandatory, but for actual receptor binding some classes of features are important (charges) more often than are others (hydrophobic);
- structures with many rotatable bonds are strongly favored, despite the increase in the entropic cost of "freezing out" a single binding conformation.

Also, before any conventional 3D search can begin, a sufficiently distinctive query pharmacophore must have somehow been extracted from the increasing flood of less precise screening data.

The topomer approach has allowed the formulation of an alternative pharmacophoric feature model, one that does not have these practical and physicochemical awkwardnesses. Of course, we do not yet have much experience with this new model. The HTS dataset, as well as the example here, does confirm general expectations that topomer performance is improved by adding this feature treatment to the original shape-only description of topomers. The HTS dataset work also suggested that the additional penalty for any unused features in a candidate hit is beneficial. Direct performance comparisons of *dbtop* with conventional 3D searching will be interesting, although evaluation may not be easy considering the subjective aspects of pharmacophoric query formulation.

Given the affiliations of the authors with an organization best known for supplying software for drug discovery, we hope to minimize possible confusion and disappointment by stating that the current plans are not to release *dbtop* for general sale, but to reserve its use for drug discovery collaborative projects. So far this internal use has stimulated a variety of experimental programs and methodological extensions, the latter including directed "building block" and "auxiliary binding site" searches. These extensions and other *dbtop* searching results will be described in due course.

Before concluding, we should emphasize the generic limitation of similarity searching in guiding drug discovery, a limitation that may be especially clear here, where the similarity descriptor is 3D shape. In any type of similarity searching, all changes of a certain magnitude to a query structure are equivalently undesirable. Yet a fundamental objective in early medicinal chemistry, raising potency from a micromolar hit to a nanomolar drug candidate, by definition requires discovery of specific changes to a query structure that are desirable, not undesirable. Thus, in selecting candidate structures, every successful drug discovery program necessarily transitions from similarity-based strategies to SAR-based strategies. Another virtue of topomer similarity, because of its underlying physicochemical model and QSAR heritage, is to make this transition exceptionally rapid, continuous, and simple. Indeed, we are currently finding automatic topomeric alignments to be consistently effective in reproducing a variety of published CoMFA models [15], as well as very convenient in analyzing the SAR from combinatorial libraries.

Acknowledgements

We wish to express special thanks to the anonymous commercial organization that supplied HTS data sets for testing, as well as financial support for the development of *dbtop*. We thank Trevor Heritage, Stefan Guessregen, Michael Lawless, Bernd Wendt, Qian Liu, Tony Cooper, and other colleagues and managers for suggestions, collegiality, and organizational support.

References

[1] R.D. Cramer, M.A. Poss, M.A. Hermsmeier, T.J. Caulfield, M.C. Kowala, M.T. Valentine, Prospective identification of biologically active structures by topomer shape similarity searching, J. Med. Chem. 42 (1999) 3919–3933.
[2] K.M. Andrews, R.D. Cramer, Toward general methods of targeted library design: topomer shape similarity searching with diverse structures as queries, J. Med. Chem. 43 (2000) 1723–1740.
[3] R.D. Cramer, D.E. Patterson, R.D. Clark, F. Soltanshahi, M.S. Lawless, Virtual libraries: a new approach to decision making in molecular discovery research, J. Chem. Inf. Comp. Sci. 6 (1998) 1010–1023.
[4] R.D. Cramer, R.D. Clark, D.E. Patterson, A.M. Ferguson, Bioisosterism as a molecular diversity descriptor: steric fields of single topomeric conformers, J. Med. Chem. 39 (1996) 3060–3069.
[5] D.E. Patterson, R.D. Cramer, A.M. Ferguson, R.D. Clark, L.E. Weinberger, Neighborhood behavior: a useful concept for validation of molecular diversity descriptors, J. Med. Chem. 39 (1996) 3049–3060.
[6] C. Burnouf, M.-P. Prunieuz, C.M. Szilagyi, Phosphodiesterase 4 inhibitors, Ann. Rep. Med. Chem. 33 (1998) 91–110.
[7] J.A. Stafford, P.L. Feldman, Chronic pulmonary inflammation and other therapeutic applications of PDE4 inhibitors, Ann. Rep. Med. Chem. 31 (1996) 31–40.

[8] A.J. Robichaud, B.L. Largent, Recent advances in selective serotonin receptor modulation, Ann. Rep. Med. Chem. 35 (2000) 11–20.
[9] L.M. Gaster, F.D. King, Latest developments in serotonin receptor modulation, Ann. Rep. Med. Chem. 33 (1998) 21–30.
[10] R.D. Brown, Y.C. Martin, Use of structure-activity data to compare structure-based clustering methods for use in compound selection, J. Chem. Inf. Comput. Sci. 36 (1996) 572–584.
[11] Concord is the product of R.S. Pearlman, University of Texas, Austin, and is distributed by Tripos Inc., 1699 S. Hanley Road, St. Louis, MO 63144. Unity, DISCO, and Distill are commercial programs also available from Tripos. Sybyl programming language (SPL) is a feature of SYBYL, another commercial program offered by Tripos.
[12] C. Lemmen, T. Lengauer, Computational methods for the structural alignment of molecules, J. Comput.-Aided Mol. Des. 14 (2000) 215–232.
[13] H. Matter, T. Potter, Comparing 3D pharmacophore triplets and "2D fingerprints" for selecting diverse compound subsets, J. Chem. Inf. Comput. Sci. 39 (1999) 1211–1225.
[14] L. Xue, J.W. Godden, J. Bajorath, Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity, J. Chem. Inf. Comput. Sci. 40 (2000) 1227–1234.
[15] Manuscript in preparation.